JEP-TALN-RECITAL 2012

JEP : Journées d'Études sur la Parole
TALN : Traitement Automatique des Langues Naturelles
RECITAL : Rencontre des Étudiants Chercheurs en Informatique
pour le Traitement Automatique des Langues

Actes de la conférence conjointe JEP-TALN-RECITAL 2012

Volume 2 : TALN

Éditeurs
Georges Antoniadis
Hervé Blanchon
Gilles Sérasset

4 – 8 Juin 2012
Grenoble, France

Des versions imprimées de ces actes peuvent être achetées auprès de :

GETALP-LIG
Laurent Besacier
BP 53
38041 Grenoble Cedex 9
France
Laurent.Besacier@imag.fr

Préface

Pour la quatrième fois, après Nancy en 2002, Fès en 2004, et Avignon en 2008, l'AFCP (Association Francophone pour la Communication Parlée) et l'ATALA (Association pour le Traitement Automatique des Langues) organisent conjointement leurs principales conférences afin de réunir en un seul lieu les deux communautés du traitement de la parole et de la langue écrite pour favoriser les interactions entre nos deux communautés.

Plus précisément, la conférence JEP-TALN-RECITAL'2012 réunit cette année la vingt-neuvième édition des Journées d'Étude sur la Parole (JEP'2012), la dix-neuvième édition de la conférence sur le Traitement Automatique des Langues Naturelles (TALN'2012) et la quinzième édition des Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'2012).

Nous avons souhaité organiser cet événement sur le campus universitaire de l'université de Grenoble, au plus proche des trois laboratoires co-organisateurs (LIG, LIDILEM, GIPSA-Lab). L'université Stendhal-Grenoble 3 (consacrée aux disciplines des humanités) nous accueille dans ses locaux à cette occasion.

Par ailleurs, JEP-TALN-RECITAL'2012 accueille quatre ateliers ; la septième édition du « DÉfi Fouille de Texte » (DEFT), la seconde édition du « DEfi Geste Langue et Signe » (DEGELS), ainsi que deux nouveaux auxquels nous souhaitons longue vie : « Interactions Langagières pour personnes Agées Dans les habitats Intelligents » (ILADI) et « Traitement Automatique des Langues Africaines – écrit et parole » (TALAf). Quatre conférenciers renommés ont accepté notre invitation pour des sessions plénières communes. Nous espérons que leur hauteur de vue et leur ouverture d'esprit permettront des discussions intéressantes et ouvriront des perspectives prometteuses.

Quelques informations sur les processus de sélection pour cette édition sont présentées ci-dessous. Nous remercions tous les relecteurs et membres des différents comités de programme pour leur travail ainsi que nos sociétés savantes : l'AFCP et l'ATALA (avec son comité permanent qui assure la continuité de la forme et du fond entre les diverses éditions).

Nous avons reçu 62 propositions d'articles longs pour TALN, parmi lesquels 24 ont été sélectionnés au moyen d'un processus de relecture consciencieux, soit un taux de sélection de 39 %. 61 articles courts ont été soumis parmi lesquels 29 ont été sélectionnés au moyen d'un processus de relecture identique à celui des articles longs, soit un taux de sélection de 48 %. Comme lors de l'édition précédente de TALN, les articles courts seront présentés sous forme de sessions orales brèves (2 minutes par publication) et de poster. 10 démonstrations seront également présentées au cours d'une session dédiée.

Concernant les JEP, 145 propositions ont été reçues. À l'issue de la réunion du comité de programme qui s'est tenue à Grenoble les 15 et 16 mars, 108 articles ont été sélectionnés (74 %). 28 articles seront présentés en session orale et 80 lors de sessions poster.

La désaffection grandissante des soumissions à RECITAL nous a conduit à proposer plusieurs innovations afin de remobiliser nos jeunes chercheurs. Tout d'abord, l'appel à communication a été étendu pour permettre la soumission de travaux préliminaires, projets de thèse, travaux des premiers mois de recherche (états de l'art, premières pistes...). Ensuite le processus de relecture a été modifié pour offrir à nos jeunes des relectures pédagogiques (encouragements, pistes) et permettre des échanges directs avec les relecteurs (relectures non-anonymes). Ces changements ont été accueillis très favorablement puisque nous avons reçu 42 propositions de communications parmi lesquelles 11 feront l'objet de présentations orales (27 %) et 17 de présentations sous forme de poster (40 %). Nous sommes également revenus à des sessions RECITAL spécifiques qui ne sont pas en parallèle avec des sessions TALN.

En ce qui concerne les actes, nous avons fourni de nouveaux styles optimisés pour une lecture à l'écran. Bien que les habitudes des auteurs aient été changées à cette occasion, nous espérons que les lecteurs nous feront des retours d'usage positifs. Un meilleur référencement des travaux présentés a aussi été l'une de nos préoccupations ; aussi avons-nous choisi de les faire référencer par l'ACL (*Association for Computational Linguistics*) dans l'*ACL Anthology* [1] pour une meilleure visibilité.

Nous vous souhaitons, chers lecteurs, un parcours passionnant et passionné au fil des nombreuses pages de ces actes et, pourquoi pas, des découvertes inattendues grâce au hasard et à votre sagacité ; découvertes qui seront les graines de nouvelles idées pour faire progresser nos champs de recherche.

Laurent Besacier, Président JEP

Hervé Blanchon & Georges Antoniadis, Présidents TALN

Didier Schwab & Jorge Mauricio Molina Mejia, Présidents RECITAL

1. http://www.aclweb.org/anthology/

Le mot de la présidente de l'Association pour le Traitement Automatique des Langues

L'Association pour le Traitement Automatique des Langues (ATALA [2]) soutient depuis 1959 les travaux de recherche fondamentale et appliquée en linguistique informatique.

En complément des travaux sur les modèles informatiques de la langue, il est primordial pour l'ATALA de renforcer ses liens avec des domaines connexes tels que le traitement de la parole ou la représentation des connaissances.

Ceci est d'autant plus important à un moment où, avec l'avènement des technologies de l'Internet et de l'information, les données écrites et parlées, qu'il était jusqu'alors très difficile de recueillir sont devenues, en un laps de temps très court, pléthores et très faciles d'accès. En quelques années seulement, nous sommes passé du rêve, avoir accès à plus de données, au cauchemar, avoir trop de données. L'Internet et l'utilisation généralisée des bases de données sont aujourd'hui la cause principale de la croissance exponentielle et continue des données en ligne.

De nos jours, grâce aux logiciels embarqués la plupart des types de dispositifs électroniques que nous utilisons quotidiennement sont en mesure de fournir des données pérennes. En effet, alors qu'auparavant la plupart des données disparaissaient après avoir été utilisées dans un but précis, les données sont maintenant stockées, fusionnées, distribuées et même revendues pour être analysées et interprétées dans le meilleur des cas, à des fins d'innovation ou d'avancée scientifique.

Dans un contexte en constante mutation, l'organisation conjointe entre l'AFCP et l'ATALA des journées TALN permet aux deux communautés d'échanger leurs méthodes d'analyse et de compréhension de ces données textuelles ou parlées afin de faire progresser la recherche en proposant de nouvelles méthodes et de nouveaux algorithmes sur lesquels s'appuyer pour développer de nouvelles technologies et services dans le domaine de l'analyse intelligente des données.

Frédérique Segond
Présidente de l'ATALA

2. http://www.atala.org/

Le mot de la présidente de l'Association Francophone de la Communication Parlée

Chers collègues,

Après les éditions de 1970 (1ères JEP), 1979 (10èmes JEP), et avec en 2000 un détour à Aussois (23èmes JEP), les Journées d'Etude sur la Parole sont de retour à Grenoble !

L'AFCP (Association Francophone de la Communication Parlée [3]) se réjouit de s'associer de nouveau à l'ATALA (Association pour le Traitement Automatique des Langues) pour l'organisation de cet événement commun que sont les JEP-TALN-RECITAL. Rappelons que depuis 2002, les communautés du traitement de la langue, orale comme écrite, se retrouvent périodiquement en un même lieu afin favoriser les échanges et stimuler l'émergence de projets de recherche commun. Les éditions passées, à Nancy en 2002, à Fès en 2004, à Avignon en 2008, ont été un réel succès et nous gageons que cette édition JEP-TALN-RECITAL'2012 sera de nouveau un moment fort de rencontres et d'échanges fructueux entre les différents acteurs de nos communautés.

Pour ce qui concerne cette 29ème édition des Journées d'Etude sur la Parole, 145 communications ont été soumises, ce qui est très satisfaisant (136 soumissions en 2010 à Mons, 130 en 2008 à Avignon). L'origine variée des soumissions (majoritairement de France, mais aussi de Belgique, de Suisse, du Canada, des Etats-Unis, de Tunisie, du Maroc,...) souligne une fois encore le caractère international de ces journées francophones, qui est une priorité de l'AFCP. Sur ces 145 soumissions, 108 ont été retenues, ce qui donne un taux d'acceptation de 74 % qui est similaire à celui de l'édition précédente. La couverture thématique des papiers retenus est vaste et reflète le dynamisme et la diversité des recherches sur la parole dans la communauté francophone.

Pour rappel, les communications aux JEP sont sélectionnées sur la base d'un article complet. Chaque soumission est évaluée par deux relecteurs. Le comité de programme, constitué des membres du CA de l'AFCP et de membres du comité d'organisation, se réunit pendant deux jours pour examiner les soumissions et leurs évaluations, certaines sont relues par un 3ème lecteur, et la sélection finale est effectuée. Les communications sélectionnées sont alors groupées par thèmes afin de définir les sessions thématiques de la conférence, et pour chaque session, des communications orales sont choisies. Les autres communications, qui seront présentées sous forme de posters, ne sont pas regroupées thématiquement de façon a avoir des sessions poster couvrant un large spectre d'intérêts. Il est donc à noter qu'aux JEP, la sélection entre communication orale et affichée s'effectue principalement sur la base d'un choix thématique pour les sessions orales et ne renvoie donc pas à un critère de qualité.

3. L'Association Francophone de la Communication Parlée (AFCP) est une structure d'animation et de réflexion de la communauté francophone travaillant sur la parole. `http://www.afcp-parole.org/`

Pour ces JEP, outres les traditionnelles bourses proposées aux étudiants et jeunes chercheurs, nous renouvelons notre action d'invitation de jeunes chercheurs appartenant à des laboratoires situés hors de France. Cinq jeunes chercheurs venant de Tunisie et d'Algérie ont été ainsi sélectionnés sur dossier et nous auront le plaisir de les accueillir à ces rencontres. Nous aurons également l'honneur de remettre lors de ces journées les prix de thèse édition 2010 et 2011, à Gwénolé Lecorvé et Juliette Kahn, respectivement.

Pour finir, l'AFCP est ravie de voir cette 29ème édition des Journées d'Etude sur la Parole se tenir à Grenoble. Grenoble est depuis longtemps un haut lieu de la recherche sur la parole et a toujours eu un rôle important dans la structuration et l'animation de notre communauté parole, tant au niveau national, qu'au niveau international. Après des restructurations difficiles du pôle parole grenoblois, nous ne pouvons que nous réjouir que l'ensemble des laboratoires grenoblois, sous l'impulsion du LIG, ait entrepris l'aventure commune qu'est l'organisation de cet événement important pour la communauté francophone. Au nom de l'AFCP, je tiens donc à remercier sincèrement tous les organisateurs de ces Journées, le LIG, le LIDILEM et le GIPSA-Lab et en particulier Laurent Besacier, pour son dynamisme et son investissement dans cette entreprise.

Au nom du comité de programme, je remercie aussi vivement les 114 relecteurs pour leur temps et leur travail fait dans un esprit constructif.

Enfin, je tiens à remercier tous les auteurs, conférenciers, et participants qui sont le moteur de notre communauté scientifique si sympathique.

Je vous souhaite à tous des journées et des rencontres enrichissantes et stimulantes.

Cécile Fougeron
Présidente de l'AFCP
Présidente du Comité de Programme des XXIXèmes JEP

Comité d'organisation de JEP-TALN-RECITAL'2012 :

Georges ANTONIADIS (LIDILEM, Université Grenoble 3)
Véronique AUBERGÉ (Gipsa-Lab, CNRS)
Valérie BELLYNCK (LIG-GETALP, Grenoble INP)
Laurent BESACIER (LIG-GETALP, Université Grenoble 1)
Hervé BLANCHON (LIG-GETALP, Université Grenoble 2)
Francis BRUNET-MANQUAT (LIG-GETALP, Université Grenoble 2)
Emmanuelle ESPERANÇA-RODIER (LIG-GETALP, Université Grenoble 1)
Jérôme GOULIAN (LIG-GETALP, Université Grenoble 2)
Marie-Paule JACQUES (LIDILEM, Université Grenoble 3)
Olivier KRAIF (LIDILEM, Université Grenoble 3)
Alexandre LABADIÉ (LIG-GETALP, CNRS)
Thomas LEBARBÉ (LIDILEM, Université Grenoble 3)
Benjamin LECOUTEUX (LIG-GETALP, Université Grenoble 2)
Mathieu MANGEOT (LIG-GETALP, Université De Savoie)
Jorge Mauricio MOLINA MEJIA (LIDILEM, Université Grenoble 3)
Claude PONTON (LIDILEM, Université Grenoble 3)
François PORTET (LIG-GETALP, Grenoble INP)
Solange ROSSATO (LIG-GETALP, Université Grenoble 3)
Isabelle ROUSSET (LIDILEM, Université Grenoble 3)
Didier SCHWAB (LIG-GETALP, Université Grenoble 2)
Frédérique SEGOND (Pôle Innovation Viseo)
Gilles SÉRASSET (LIG-GETALP, Université Grenoble 1)
Agnès TUTIN (LIDILEM, Université Grenoble 3)
Michel VACHER (LIG-GETALP, CNRS)
Nathalie VALLÉE (Gipsa-Lab, CNRS)
Virginie ZAMPA (LIDILEM, Université Grenoble 3)

Comité de programme de JEP'2012 :

Présidents :

Laurent BESACIER (LIG-GETALP, Université Grenoble 1, France)
Cécile FOUGERON (LPP Paris)
Guillaume GRAVIER, IRISA et CNRS-INRIA Rennes)

Membres :

Gilles ADDA (LIMSI, Paris)
Melissa BARKAT-DEFRADAS (PRAXILING, Montpellier)
Loïc BARRAULT (LIUM, Le Mans)
Philippe BOULA DE MAREUIL (LIMSI, Paris)
Véronique BOULENGER (DDL Lyon)
Elisabeth DELAIS-ROUSSARIE (Lab. Linguistique Formelle, Paris)
Véronique DELVAUX (Univ. Mons, Belgique)

Didier DEMOLIN (Gipsa-Lab, Grenoble)
Laurence DEVILLERS (LIMSI, Paris)
Isabelle FERRANE (IRIT, Toulouse)
Emmanuel FERRAGNE (CLILLAC-ARP, Paris)
Corinne FREDOUILLE (LIA, Avignon)
Bernard HARMEGNIES (Univ. Mons, Belgique)
Fabrice HIRSCH (PRAXILING, Montpellier)
Thomas HUEBER (Gipsa-Lab, Grenoble)
Irina ILLINA (LORIA, Nancy)
David LANGLOIS (LORIA, Nancy)
Georges LINARES (LIA, Avignon)
Hélène LOEVENBRUCK (Gipsa-Lab, Grenoble)
Egidio MARSICO (DDL, Lyon)
Sylvain MEIGNIER (LIUM, Le Mans)
Christine MEUNIER (LPL, Aix en Provence)
Yohann MEYNADIER (LPL, Aix en Provence)
François PELLEGRINO (DDL, Lyon)
Pascal PERRIER (Gipsa-Lab, Grenoble)
François PORTET (LIG-GETALP, Grenoble)
Solange ROSSATO (LIG-GETALP, Grenoble)
Sophie ROSSET (LIMSI, Paris)
Marc SATO (Gipsa-Lab, Grenoble)
Christophe SAVARIAUX (Gipsa-Lab, Grenoble)
Christine SÉNAC (IRIT, Toulouse)
Rudolph SOCK (IPS, Strasbourg)
Annemie VAN HIRTUM (Gipsa-Lab, Grenoble)
Béatrice VAXELAIRE (IPS, Strasbourg)
Chakir ZEROUAL (LPP Paris et Univ. Sidi Mohamed Ben-abdellah, Fes, Maroc)

Relecteurs additionnels :

Martine ADDA-DECKER, LPP et LIMSI Paris)
Régine ANDRE-OBRECHT (IRIT, Toulouse)
Angélique AMELOT (LPP, Paris)
Corine ASTESANO (Univ. Toulouse 2 et LPL, Aix en Provence)
Véronique AUBERGÉ (LIG et GIPSA-Lab, Grenoble)
Nicolas AUDIBERT (LPP, Paris)
Gérard BAILLY (Gipsa-Lab, Grenoble)
Claude BARRAS (LIMSI, Paris)
Denis BEAUTEMPS (Gipsa-Lab, Grenoble)
Nathalie BEDOIN (DDL, Lyon)
Roxane BERTRAND (LPL, Aix en Provence)
Benjamin BIGOT (LIA, Avignon)
Frédéric BIMBOT (IRISA et CNRS-INRIA Rennes)
Anne BONNEAU (LORIA, Nancy)
Hélène BONNEAU-MAYNARD (LIMSI, Paris)
Hervé BREDIN (LIMSI, Paris)

Nathalie CAMELIN (LIUM, Le Mans)
Christian CAVE (LPL, Aix en Provence)
Claire PILLOT-LOISEAU (LPP, Paris)
Lise CREVIER-BUCHMAN (LPP, Paris)
Mariapaola D'IMPERIO (LPL, Aix en Provence)
Paul DELÉGLISE (LIUM, Le Mans)
Christian DICANIO (UC Berkeley, États-Unis)
Cong-Thanh DO (LIMSI, Paris)
Christelle DODANE (PRAXILING, Montpellier)
Driss MATROUF (LIA, Avignon)
Sophie DUFOUR (LPL, Aix en Provence)
Elie EL-KHOURY (LIUM, Le Mans)
Robert ESPESSER (LPL, Aix en Provence)
Yannick ESTÈVE (LIUM, Le Mans)
Martine FARACO (LPL, Aix en Provence)
Jérôme FARINAS (IRIT, Toulouse)
Dominique FOHR (LORIA, Nancy)
Teddy FURON (IRISA et CNRS-INRIA Rennes)
Maeva GARNIER (Gipsa-Lab, Grenoble)
Cedric GENDROT (LPP, Paris)
Alain GHIO (LPL, Aix en Provence)
Antoine GIOVANNI (CHU Marseille et LPL Aix en Provence)
Laurent GIRIN (Gipsa-Lab, Grenoble)
Pierre HALLE (LPP, Paris)
Sophie HERMENT (LPL, Aix en Provence)
Daniel HIRST (LPL, Aix en Provence)
Kathy HUET (Univ. Mons, Belgique)
Stephane HUET (LIA, Avignon)
Denis JOUVET (LORIA, Nancy)
Juliette KAHN (LNE Paris)
Sophie KERN (DDL, Lyon)
Hélène LACHAMBRE (IRIT, Toulouse)
Muriel LALAIN (LPL, Aix en Provence)
Antoine LAURENT (LIUM, Le Mans)
Gwénolé LECORVE (IDIAP Martigny (Suisse))
Thierry LEGOU (LPL, Aix en Provence)
Christophe LÉVY (LIA, Avignon)
Alain MARCHAL (LPL, Aix en Provence)
Odile MELLA (LORIA, Nancy)
Ilya OPARIN (LIMSI, Paris)
Caterina PETRONE (LPL, Aix en Provence)
Myriam PICCALUGA (LPL, Aix en Provence)
Julien PINQUIER (IRIT, Toulouse)
Serge PINTO (LPL, Aix en Provence)
Agnès PIQUARD-KIPFFER (LORIA, Nancy)

Michel PITERMANN (LPL, Aix en Provence)
Rachid RIDOUANE (LPP Paris)
Albert RILLIARD (LIMSI, Paris)
Mickael ROUVIER (LIUM, Le Mans)
Jérémi SAUVAGE (PRAXILING, Montpellier)
Jean-Luc SCHWARTZ (Gipsa-Lab, Grenoble)
Grégory SENAY (LIA, Avignon)
Willy SERNICLAES (ULB Bruxelles, Belgique)
Marion TELLIER (LPL, Aix en Provence)
Michel VACHER (LIG Grenoble)
Nathalie VALLÉE (Gipsa-Lab, Grenoble)
Anne VILAIN (Gipsa-Lab, Grenoble)
Coriandre VILAIN (Gipsa-Lab, Grenoble)
Emmanuel VINCENT (IRISA et CNRS-INRIA Rennes)
Pauline WELBY (LPL, Aix en Provence)

Comité de programme de TALN'2012 :

Présidents :

Georges ANTONIADIS (LIDILEM, Université Grenoble 3, France)
Hervé BLANCHON (LIG-GETALP, Université Grenoble 2, France)

Membres :

Nicholas ASHER (IRIT, CNRS et Université Toulouse 3)
Frédéric BÉCHET (LIF, Aix Marseille Université)
Yves BESTGEN (Université Catholique de Louvain, Louvain-la-Neuve, Belgique)
Philippe BLACHE (LPL, CNRS et Université de Provence)
Christian BOITET (LIG-GETALP, Université Grenoble 1)
Malek BOUALEM (France Telecom Orange Labs, Lannion)
Narjès BOUFADEN (KeaText, Montréal, Canada)
Yllias CHALI (University of Lethbridge, Lethbridge, Canada)
Laurence DANLOS (ALPAGE, Université Paris 7)
Piet DESMET (ITEC, K.U.Leuven et K.U.Leuven KULAK, Belgique)
Mark DRAS (Macquarie University, Sydney, Australie)
Denys DUCHIER (LIFO, Université d'Orléans)
Marc DYMETMAN (XRCE, Grenoble)
Dominique ESTIVAL (University of Western Sydney, Sydney, Australie)
Cédrick FAIRON (Université Catholique de Louvain, Louvain-la-Neuve, Belgique)
Olivier FERRET (CEA LIST, Palaiseau)
Michel GAGNON (École Polytechnique de Montréal, Montréal, Canada)
Claire GARDENT (LORIA, Villers lès Nancy)
Nabil HATOUT (CLLE-ERSS, CNRS et Université Toulouse II)
Sylvain KAHANE (MODYCO-ALPAGE, Université Paris 10)
Laura KALLMEYER (Heinrich-Heine-Universität, Düsseldorf, Allemagne)
Mathieu LAFOURCADE (LIRMM, Université Montpellier 2)
Philippe LANGLAIS (DIRO, Université Montréal, Canada)
Guy LAPALME (RALI, Université Montréal, Canada)

Yves LEPAGE (IPS, Université Waseda, Japon)
Emmanuel MORIN (LINA, Université Nantes)
Adeline NAZARENKO (LIPN, Université Paris 13)
Luka NERIMA (LATL, Université Genève, Suisse)
Alain POLGUÈRE (Université de Lorraine et ATILF CNRS)
Laurent PRÉVOT (LPL, CNRS et Université de Provence)
Violaine PRINCE (LIRMM, Université Montpellier 2)
Jean-Philippe PROST (LIRMM, Université Montpellier 2)
Christian RETORÉ (LaBRI et INRIA, Université Bordeaux 1)
Sophie ROSSET (LIMSI, CNRS)
Didier SCHWAB (LIG-GETALP, Université Grenoble 2)
Holger SCHWENK (LIUM, Université du Maine, Le Mans)
Pascale SÉBILLOT (IRISA, INSA de Rennes)
Gilles SÉRASSET (LIG-GETALP, Université Grenoble 1)
Agnès TUTIN (LIDILEM, Université Grenoble 3)
Anne VILNAT (LIMSI, CNRS et Université Paris Sud)
François YVON (LIMSI, CNRS et Université Paris Sud)
Virginie ZAMPA (LIDILEM, Université Grenoble 3)
Pierre ZWEIGENBAUM (LIMSI, CNRS et INALCO)

Comité Scientifique de TALN'2012 :

Présidents :

Georges ANTONIADIS (LIDILEM, Université Grenoble 3, France)
Hervé BLANCHON (LIG-GETALP, Université Grenoble 2, France)

Membres :

Les membres du comité de programme aidés de . . .
Ramzi ABBES (Techlimed, Lyon)
Stergos AFANTENOS (IRIT, Université de Toulouse)
Salah AIT-MOKHTAR (XRCE, Grenoble)
Maxime AMBLARD (LORIA, Université de Lorraine)
Jean-Yves ANTOINE (LI, Université de Tours et Lab-STICC, CNRS)
Delphine BATTISTELLI (STIH, Université Paris 4)
Denis BECHET (LINA, Université de Nantes)
Patrice BELLOT (LSIS, Université Aix-Marseille)
Delphine BERNHARD (LiLPa, Université de Strasbourg)
Romaric BESANÇON (CEA-LIST, Saclay Nano-Innov)
Brigitte BIGI (LPL, Aix en Provence)
Julien BOURDAILLET (Xerox, États-Unis)
Caroline BRUN (XRCE, Grenoble)
Francis BRUNET-MANQUAT (LIG-GETALP, Université Grenoble 2)
Marie CANDITO (Alpage, Université Paris Diderot)
Thierry CHANIER (LRL, Clermont Université)
Vincent CLAVEAU (IRISA-CNRS, Rennes)
Nathalie COLINEAU (CSIRO ICT Centre, Marsfield, Australie)
Benoît CRABBÉ (Alpage, Paris 7)

Béatrice DAILLE (LINA, Université de Nantes)
Pascal DENIS (Alpage)
Iris ESHKOL-TARAVELLA (LLL, Université d'Orléans)
Cécile FABRE (CLLE-ERSS, Université Toulouse 2)
Benoit FAVRE (LIF, Université Aix-Marseille)
Dominic FOREST (Université de Montréal, Canada)
Karen FORT (INIST et LIPN, Paris 13)
George FOSTER (CNRC, Gatineau, Canada)
Nuria GALA (LIF, Université Aix-Marseille)
Bruno GAUME (CLLE-ERSS, Université Toulouse 2)
Éric GAUSSIER (LIG-GETALP, Université Grenoble 1)
Kim GERDES (LPP, Université Paris 3)
Jérôme GOULIAN (LIG-GETALP, Université Grenoble 2)
Benoît HABERT (ICAR, ENS Lyon)
Najeh HAJLAOUI (Institut de recherche Idiap, Martigny, Suisse)
Thierry HAMON (LimetBio, Université Paris 13)
Marie-Paule JACQUES (LIDILEM, Université Grenoble 1)
Guillaume JACQUET (XRCE, Grenoble)
Christine JACQUIN (LINA, Université de Nantes)
Adel JEBALI (Université Concordia, Montréal, Canada)
Leïla KOSSEIM (Université Concordia, Montréal, Canada)
Olivier KRAIF (LIDILEM, Université Grenoble 3)
Éric LAPORTE (LIGM, Université Paris-Est Marne-la-Valée)
Dominique LAURENT (Synapse, Toulouse)
Thomas LEBARBÉ (LIDILEM, Université Grenoble 3)
Anne-Laure LIGOZAT (LIMSI, ENSIIE)
Cédric LOPEZ (LIRMM, Université Montpellier 2)
Mathieu MANGEOT (LIG-GETALP, Université de Savoie)
Denis MAUREL (LI, Université de Tours)
Aurélien MAX (LIMSI, Université Paris-Sud)
Jasmina MILIĆEVIĆ (OLST, Dalhousie University, Canada)
Laura MONCEAUX (LINA, Université de Nantes)
Richard MOOT (LaBRI et SIGNES, Bordeaux)
Erwan MOREAU (Trinity College Dublin, Irlande)
Fabienne MOREAU (IRISA, Universtié Rennes 2)
Véronique MORICEAU (LIMSI, Université Paris-Sud)
Philippe MULLER (IRIT, Université de Toulouse)
Alexis NASR (LIF, Université Aix-Marseille)
Aurélie NÉVÉOL (NCBI, National Library of Medicine, États-Unis)
Jian-Yun NIE (RALI, Université de Montréal, Canada)
Cécile PARIS (CSIRO ICT Centre, Marsfield, Australie)
Yannick PARMENTIER (LIFO, Université d'Orléans)
Guy PERRIER (LORIA, Université de Lorraine)
Sylvain POGODALLA (LORIA, Vandoeuvre-lès-Nancy)
Thierry POIBEAU (LaTTiCe, Montrouge)
Claude PONTON (LIDILEM, Université Grenoble 3)

Andrei POPESCU-BELIS (Institut de recherche Idiap, Martigny, Suisse)
Carlos RAMISCH (LIG-GETALP, Grenoble)
Mathieu ROCHE (LIRMM, Université Montpellier 2)
Antoine ROZENKNOP (LIPN, Universtité Paris 13)
Benoît SAGOT (Alpage, INRIA Roquencourt)
Djamé SEDDAH (Alpage, Université Paris 4)
Kamel SMAÏLI (LORIA, Université de Lorraine)
Xavier TANNIER (LIMSI, Université Paris-Sud)
Isabelle TELLIER (LaTTiCe, Université Paris 3)
Juan-Manuel TORRES-MORENO (LIA, Université d'Avignon et des Pays de Vaucluse)
François TROUILLEUX (LRL, Université Clermont-Ferrand 2)
Lonneke VAN DER PLAS (IMS, Université de Stuttgart, Allemagne)
Fabienne VENANT (LORIA, Université Nancy 2)
Jacques VERGNE (GREYC, Université de Caen)
Éric VILLEMONTE DE LA CLERGERIE (Alpage, INRIA Roquencourt)
Eric WEHRLI (LATL, Université de Genève, Suisse)
Guillaume WISNIEWSKI (LIMSI, Université Paris-Sud)
Imed ZITOUNI (IBM T.J. Watson Research Center, Yorktown Heights, États-Unis)
Michael ZOCK (LIF, Marseille)
Amal ZOUAQ (Royal Military College of Canada et Athabasca University, Canada)
Mounir ZRIGUI (UTIC, Faculté des Sciences de Monastir, Tunisie)
Sandrine ZUFFEREY (ILC, Université Catholique de Louvain-la-Neuve, Belgique)

Comité de programme de RECITAL'2012 :

Présidents :

Jorge Mauricio MOLINA MEJIA (LIDILEM, Université Stendhal – Grenoble 3)
Didier SCHWAB (GETALP-LIG, Université Pierre Mendes France – Grenoble 2)

Membres :

Vanessa ANDRÉANI (Société CFH et laboratoire ERSS, Université Toulouse 2 – Le Mirail)
Nicolas AUDIBERT (Laboratoire de Phonétique et Phonologie-CNRS, Université Sorbonne-Nouvelle)
Frédéric BÉCHET (Laboratoire d'Informatique Fondamentale de Marseille, Université d'Aix-Marseille)
Patrice BELLOT (LSIS, Université d'Aix-Marseille)
Valérie BELLYNCK (GETALP-LIG, Grenoble INP)
Farah BENAMARA (IRIT, Université Toulouse 3)
Christian BOITET (GETALP-LIG, Université Joseph Fourier – Grenoble 1)
Leila BOUTORA (LPL, Université d'Aix-Marseille, Marseille)
Francis BRUNET-MANQUAT (GETALP-LIG, Université Pierre Mendes France – Grenoble 2)
François-Régis CHAUMARTIN (Société Proxem, Laboratoire Alpage, UMR INRIA, Université Paris 7)
Gaël DE CHALENDAR (CEA LIST, Palaiseau)
Achille FALAISE (GETALP-LIG, Société Floralis, Université Joseph Fourier-Grenoble 1)
Olivier FERRET (CEA LIST, Palaiseau)
Nuria GALA (LIF, Université d'Aix-Marseille)
Jérôme GOULIAN (GETALP-LIG, Université Pierre Mendes France – Grenoble 2)
Thierry HAMON (LIM&BIO, Université Paris 13)
Nicolas HERNANDEZ (LINA, CNRS 6241, Nantes)

Bernard JACQUEMIN (CREM, Université de Haute Alsace, Mulhouse)
Olivier KRAIF (LIDILEM, Université Stendhal – Grenoble 3)
Alexandre LABADIÉ (GETALP-LIG, Grenoble)
Mathieu LAFOURCADE (LIRMM, Université de Montpellier 2)
Guy LAPALME (RALI, Université de Montréal, Canada)
François LAREAU (CLT, Macquarie University, Australie)
Thomas LEBARBÉ (LIDILEM, Université Stendhal – Grenoble 3)
Benjamin LECOUTEUX (LIG-GETALP, Université Pierre Mendès France – Grenoble 2)
Yves LEPAGE (Université Waseda, Japon)
Mathieu LOISEAU (LIDILEM, Université Stendhal – Grenoble 3)
Cédric LOPEZ (LIRMM, Université Montpellier 2)
Denis MAUREL (Université François Rabelais Tours)
Aurélien MAX (LIMSI-CNRS & Université Paris-Sud)
Jean-Luc MINEL (MoDyCO, UMR 7114, Université Paris-Ouest Nanterre La Défense – CNRS)
Emmanuel MORIN (LINA, CNRS 6241, Nantes)
Yayoi NAKAMURA-DELLOYE (LCAO, Université Paris VII)
Claude PONTON (LIDILEM, Université Stendhal-Grenoble 3)
François PORTET (GETALP-LIG, Grenoble INP)
Laurent PREVOT (LPL, Université d'Aix-Marseille, Marseille)
Violaine PRINCE (LIRMM, Université Montpellier 2)
Jean-Philippe PROST (LIRMM, Université Montpellier 2)
Bali RANAIVO-MALANÇON (Universiti Sarawak Malaysia, Malaisie)
Christian RETORÉ (LaBRI, Université Bordeaux 1)
Mathieu ROCHE (LIRMM, Université Montpellier 2)
Solange ROSSATO (GETALP-LIG, Université Stendhal – Grenoble 3)
Azim ROUSSANALY (LORIA, Université de Lorraine)
Isabelle ROUSSET (LIDILEM, Université Stendhal – Grenoble 3)
Fatiha SADAT (Université du Québec à Montréal, Canada)
Tristan VANRULLEN (TVSI, Marseille)
Eric WEHRLI (LATL, Université de Genève, Suisse)
Virginie ZAMPA (LIDILEM, Université Stendhal – Grenoble 3)
Haifa ZARGAYOUNA (LIPN, Université Paris 13)
Michael ZOCK (CNRS-LIF, Marseille)
Mounir ZRIGUI (Faculté des Sciences, Université de Monastir, Tunisie)
Pierre ZWEIGENBAUM (LIMSI-CNRS, Orsay)

Conférenciers invités :

Ian Maddieson (Université de Californie, Berkeley, États-Unis)
Jacqueline Léon (Laboratoire d'histoire des théories linguistiques, CNRS, Paris)
Yoshinori Sagisaka (Université de Waseda, Japon)
Hans Uszkoreit (DFKI, Sarrebruck, Allemagne)

Sponsors :


Inria


upmf
Grenoble
Université Pierre-Mendès-France
Sciences sociales & humaines


UNIVERSITÉ
Stendhal
GRENOBLE 3


quaero


CELIFrance


GRENOBLE-ALPES
METROPOLE


Région
RhôneAlpes


cnrs


UNIVERSITE
JOSEPH FOURIER
SCIENCES TECHNOLOGIE MEDECINE


LNE
Le progrès, une passion à partager


Grenoble
INP


EUROPEAN
ELRA
LANGUAGE
RESOURCES
ASSOCIATION


SyLLabs


Pôle
GRENOBLE
cognition


Synapse
Développement


– LINGUA ET MACHINA –
L'ÉCRIT MULTILINGUE DANS L'ENTREPRISE


MGEN


EDF


VILLE DE
GRENOBLE


ami
ENTERPRISE INTELLIGENCE SOFTWARE


L I G


LIDILEM
UNIVERSITÉ
Stendhal
GRENOBLE 3


gipsa-lab


AFCP


ATALA

Table des matières

Communications orales

Simplification de phrases pour l'extraction de relations
Anne-Lyse Minard, Anne-Laure Ligozat et Brigitte Grau 1

Extraction d'information automatique en domaine médical par projection inter-langue : vers un passage à l'échelle
Asma Ben Abacha, Pierre Zweigenbaum et Aurélien Max 15

Une méthode d'extraction d'information fondée sur les graphes pour le remplissage de formulaires
Ludovic Jean-Louis, Romaric Besançon et Olivier Ferret 29

Traitement automatique sur corpus de récits de voyages pyrénéens : Une analyse syntaxique, sémantique et temporelle
Anaïs Lefeuvre, Richard Moot, Christian Rétoré et Noémie-Fleur Sandillon-Rezer 43

La reconnaissance des mots composés à l'épreuve de l'analyse syntaxique et vice-versa : évaluation de deux stratégies discriminantes
Matthieu Constant, Anthony Sigogne et Patrick Watrin 57

Calcul des cadres de sous catégorisation des noms déverbaux français (le cas du génitif)
Ramadan Alfared, Denis Bechet et Alexander Dikovsky 71

Vectorisation, Okapi et calcul de similarité pour le TAL : pour oublier enfin le TF-IDF
Vincent Claveau 85

TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe
Christophe Benzitoun, Karën Fort et Benoît Sagot 99

Alignement sous-phrastique hiérarchique avec Anymalign
Adrien Lardilleux, François Yvon et Yves Lepage 113

Utilisation de la translittération arabe pour l'amélioration de l'alignement de mots à partir de corpus parallèles français-arabe
Saadane Houda et Semmar Nasredine 127

Compositionnalité et contextes issus de corpus comparables pour la traduction terminologique
Emmanuel Morin et Béatrice Daille 141

Raffinement du Lexique des Verbes Français
Paul Bédaride 155

Étude des manifestations de la relation de méronymie dans une ressource distributionnelle
François Morlane-Hondère et Cécile Fabre 169

Un critère de cohésion thématique fondé sur un graphe de cooccurrences
Clément de Groc, Xavier Tannier et Claude de Loupy 183

Validation sur le Web de reformulations locales : application à la Wikipédia
Houda Bouamor, Aurélien Max, Gabriel Illouz et Anne Vilnat 197

Simplification syntaxique de phrases pour le français
Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat et Thomas François 211

Étude comparative entre trois approches de résumé automatique de documents arabes
Iskandar Keskes, Mohamed Mahdi Boudabous, Mohamed Hédi Maaloul et Lamia Hadrich Belguith 225

Etude sémantique des mots-clés et des marqueurs lexicaux stables dans un corpus technique
Ann Bertels, Dirk De Hertog et Kris Heylen 239

Fouille de graphes sous contraintes linguistiques pour l'exploration de grands textes
Solen Quiniou, Peggy Cellier, Thierry Charnois et Dominique Legallois 253

Une étude en 3D de la paraphrase : types de corpus, langues et techniques
Houda Bouamor, Aurélien Max et Anne Vilnat 267

Détection et correction automatique d'erreurs d'annotation morpho-syntaxique du French TreeBank
Florian Boudin et Nicolas Hernandez 281

Annotation sémantique du French Treebank à l'aide de la réécriture modulaire de graphes
Bruno Guillaume et Guy Perrier 293

Enrichissement du FTB : un treebank hybride constituants/propriétés
Philippe Blache et Stéphane Rauzy 307

Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical
Marie Candito et Djamé Seddah 321

Communications sous forme de poster

ACOLAD Plateforme pour l'édition collaborative dépendancielle
Francis Brunet-Manquat et Jérôme Goulian 335

Extraction de préférences à partir de dialogues de négociation
Anaïs Cadilhac, Farah Benamara, Vladimir Popescu, Nicholas Asher et Mohamadou Seck 343

Détection de conflits dans les communautés épistémiques en ligne
Alexandre Denis, Matthieu Quignard, Dominique Freard, Francoise Detienne, Michael Baker et Flore Barcellini 351

Quel est l'apport de la détection d'entités nommées pour l'extraction d'information en domaine restreint ?
Camille Dutrey, Chloé Clavel, Sophie Rosset, Ioana Vasilescu et Martine Adda-Decker ... 359

Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des sentiments
Egle Eensoo et Mathieu Valette 367

Méthodologie d'exploration de corpus et de formalisation de règles grammaticales pour les langues des signes
Michael Filhol et Annelies Braffort 375

Annotation manuelle de matchs de foot : Oh la la la ! l'accord inter-annotateurs ! et c'est le but !
Karën Fort et Vincent Claveau 383

Etude de différentes stratégies d'adaptation à un nouveau domaine en fouille d'opinion
Anne Garcia-Fernandez et Olivier Ferret 391

Le Lexicoscope : un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques
Olivier Kraif et Sascha Diwersy 399

Analyse des contextes et des candidats dans l'identification des équivalents terminologiques en corpus comparables
Audrey Laroche 407

BiTermEx Un prototype d'extraction de mots composés à partir de documents comparables via la méthode compositionnelle
Emmanuel Planas 415

Combinaison d'approches pour l'extraction automatique d'événements
Laurie Serrano, Thierry Charnois, Stephan Brunessau, Bruno Grilheres et Maroua Bouzid 423

Apprentissage automatique d'un chunker pour le français
Isabelle Tellier, Denys Duchier, Iris Eshkol, Arnaud Courmet et Mathieu Martinet 431

Effacement de dimensions de similarité textuelle pour l'exploration de collections de rapports d'incidents aéronautiques
Nikola Tulechki et Ludovic Tanguy 439

Traduction automatique à partir de corpus comparables : extraction de phrases parallèles à partir de données comparables multimodales
Haithem Afli, Loïc Barrault et Holger Shwenk 447

La reconnaissance automatique de la fonction des pronoms démonstratifs en langue arabe
Yacine Ben Yahia, Souha Mezghani Hammami et Lamia Hadrich Belguith 455

Un annotateur automatique d'expressions temporelles du français et son évaluation sur le TimeBank du français
Andre Bittar et Caroline Hagege 463

Vers le FDTB : French Discourse Tree Bank
Laurence Danlos, Diégo Antolinos-Basso, Chloé Braud et Charlotte Roze 471

Combinaison de ressources générales pour une contextualisation implicite de requêtes
Romain Deveaud et Patrice Bellot 479

Repérage des entités nommées pour l'arabe : adaptation non-supervisée et combinaison de systèmes
Souhir Gahbiche-Braham, Hélène Bonneau-Maynard, Thomas Lavergne et François Yvon 487

Propagation de polarités dans des familles de mots : impact de la morphologie dans la construction d'un lexique pour l'analyse de sentiments
Nuria Gala et Caroline Brun 495

Transitions thématiques : Annotation d'un corpus journalistique et premières analyses
Alexandre Labadié, Patrice Enjalbert et Stephane Ferrari 503

La "multi-extraction" comme stratégie d'acquisition optimisée de ressources terminologiques et non terminologiques
Blandine Plaisantin Alecu, Izabella Thomas et Julie Renahy 511

Une Approche de Recherche d'Information Structurée fondée sur la Correction d'Erreurs à l'Indexation des Documents
Arnaud Renard, Sylvie Calabretto et Béatrice Rumpler 519

Post-édition statistique pour l'adaptation aux domaines de spécialité en traduction automatique
Raphaël Rubino, Stéphane Huet, Fabrice Lefèvre et Georges Linarès 527

Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées
Benoît Sagot, Marion Richard et Rosa Stern 535

Utilisation des fonctions de croyance pour l'estimation de paramètres en traduction automatique
Christophe Servan et Simon Petitrenaud 543

La longueur des tours de parole comme critère de sélection de conversations dans un centre d'appels
Philippe Suignard, Frederik Cailliau et Ariane Cavet 551

Enjeux méthodologiques, linguistiques et informatiques pour le traitement du français écrit des sourds
Tristan Vanrullen, Leïla Boutora et Jean Dagron 559

Simplification de phrases pour l'extraction de relations

Anne-Lyse Minard[1, 2] Anne-Laure Ligozat[1, 3] Brigitte Grau[1, 3]
(1) LIMSI-CNRS, BP 133, 91403 Orsay Cedex
(2) Université Paris Sud, 91400 Orsay
(3) ENSIIE, square de la résistance, 91000 Évry
prenom.nom@limsi.fr

RÉSUMÉ
L'extraction de relations par apprentissage nécessite un corpus annoté de très grande taille pour couvrir toutes les variations d'expressions des relations. Pour contrer ce problème, nous proposons une méthode de simplification de phrases qui permet de réduire la variabilité syntaxique des relations. Elle nécessite l'annotation d'un petit corpus qui sera par la suite augmenté automatiquement. La première étape est l'annotation des simplifications grâce à un classifieur à base de CRF, puis l'extraction des relations, et ensuite une complétion automatique du corpus d'entraînement des simplifications grâce aux résultats de l'extraction des relations. Les premiers résultats que nous avons obtenus pour la tâche d'extraction de relations d'i2b2 2010 sont très encourageants.

ABSTRACT
Sentence simplification for relation extraction

Machine learning based relation extraction requires large annotated corpora to take into account the variability in the expression of relations. To deal with this problem, we propose a method for simplifying sentences, i.e. for reducing the syntactic variability of the relations. Simplification requires the annotation of a small corpus, which will be automatically augmented. The process starts with the annotation of the simplification thanks to a CRF classifier, then the relation extraction, and lastly the automatic completion of the training corpus for the simplification through the results of the relation extraction. The first results we obtained for the task of relation extraction of the i2b2 2010 challenge are encouraging.

MOTS-CLÉS : Extraction de relations, simplification de phrases, apprentissage automatique.

KEYWORDS: Relation extraction, sentence simplification, machine learning.

1 Introduction

Dans le domaine médical, de nombreux documents électroniques sont produits chaque jour, mais ces documents sont sous forme textuelle, et les informations qu'ils contiennent sont donc difficilement exploitables. L'extraction d'information consiste à structurer cette information. Pour une tâche donnée, les documents disponibles ne contiennent cependant pas nécessairement de nombreux exemples d'apprentissage et les corpus peuvent présenter une grande variabilité. Par conséquent, il est nécessaire de pouvoir apprendre à partir de peu d'exemples, éventuellement

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 1–14,
Grenoble, 4 au 8 juin 2012.

très disparates. Dans cet article, nous nous intéressons à une tâche d'extraction de relations médicales et proposons une méthode qui consiste à effectuer une simplification syntaxique préalable des phrases. Cette simplification a pour but de normaliser le corpus en ne gardant que les informations qui sont pertinentes pour l'extraction. Elle est donc guidée par la tâche, et peu d'exemples sont nécessaires pour apprendre la simplification puisque l'extraction de relation est utilisée pour augmenter le corpus annoté.

Après un état de l'art sur le domaine de l'extraction de relations et sur la simplification (section 2), nous présenterons la tâche d'extraction de relations en domaine médical et son application dans le cadre du challenge i2b2 2010 [1], ainsi que le système que nous avons développé (section 3). Ensuite, nous présenterons notre méthode pour l'annotation des simplifications (section 4). Nous détaillerons la méthode originale proposée pour améliorer la simplification grâce à la combinaison du système d'extraction de relations et du classifieur pour l'annotation des simplifications (sections 5 et 6), et terminerons par la présentation des expérimentations que nous avons menées et des résultats obtenus (section 7) [2].

2 État de l'art

De nombreuses méthodes ont été proposées pour l'extraction de relations, les plus courantes étant fondées sur une classification automatique plus ou moins supervisée. Les attributs utilisés pour la classification représentent en général de l'information lexicale, sémantique ou syntaxique. Par exemple (Roberts *et al.*, 2008) proposent une approche fondée sur des SVM pour extraire des relations dans des dossiers de patients atteints d'un cancer. Ils utilisent des attributs lexicaux, sémantiques et morpho-syntaxiques. (Uzuner *et al.*, 2010) utilisent des attributs syntaxiques plus riches puisqu'ils ajoutent les dépendances syntaxiques entre les concepts. Ils les utilisent dans une approche vectorielle fondée sur des SVM pour extraire des relations entre des problèmes, des tests et des traitements dans des comptes-rendus médicaux. Ces informations syntaxiques n'améliorent pas la détection des relations car dans beaucoup de cas il n'existe pas de dépendance entre les deux concepts. (Zhang *et al.*, 2006) incluent également de l'information syntaxique riche dans leur système d'extraction de relations. Pour cela, ils utilisent des arbres syntaxiques avec des tree kernels. Ils ont testé leur système sur le corpus ACE 2003, et ils montrent que les meilleurs résultats sont obtenus en utilisant le plus petit sous-arbre commun aux deux entités. Nous montrons dans (Minard *et al.*, 2011a) que pour l'extraction de relations en domaine médical (sur le corpus i2b2 2010) l'utilisation de l'arbre minimal commun aux deux entités n'est pas suffisant et qu'il est souvent nécessaire d'utiliser l'arbre complet ou tout du moins des éléments de cet arbre.

Pour améliorer l'extraction des relations, nous proposons une méthode de simplification des phrases. Simplifier les phrases consiste alors à supprimer ou à repérer les mots de la phrase qui peuvent gêner le classifieur. Dans notre cas, la simplification ne consiste pas à rendre un texte plus facile à lire, mais à ne garder que les mots permettant de classer une relation.

La simplification de textes a donné lieu à de nombreux travaux, soit en tant que tâche à part entière comme par exemple dans (Woodsend et Lapata, 2011), soit en tant que prétraitement pour d'autres tâches, comme par exemple la génération de questions (Heilman et Smith, 2010). Cette

1. https ://www.i2b2.org/NLP/Relations/
2. Ce travail a été partiellement financé par OSEO dans le cadre du programme Quæro.

simplification est généralement fondée sur des règles syntaxiques. Dans le domaine biomédical, différentes recherches sur la simplification syntaxique pour améliorer l'extraction de relations ont été menées dans le domaine des interactions entre protéines (PPI). (Jonnalagadda et Gonzalez, 2010) ont développé un outil (bioSimplify) qui produit des phrases simples à partir d'une phrase complexe. Leur objectif est d'augmenter le rappel de l'extraction d'information dans le domaine biomédical. Pour cela, ils ont écrit des règles de simplification syntaxique qui s'appliquent au niveau morpho-syntaxique. Leur système produit plusieurs phrases simples et grammaticalement correctes à partir de la phrase d'origine. Aucune sélection de la (des) meilleure(s) phrase(s) simple(s) n'est effectuée, et les règles n'obligent pas la conservation de la paire d'entités candidate. L'évaluation de leur outil pour l'extraction des interactions entre protéines n'est pas assez précise pour en tirer des conclusions. (Miwa *et al.*, 2010) ont également utilisé des règles pour simplifier les phrases. La douzaine de règles qu'ils ont écrites s'appliquent sur la sortie d'un analyseur syntaxique. Elles sont appliquées pour chaque paire de protéines, car leur rôle est de supprimer l'information inutile pour l'extraction des interactions. Ils ont évalué l'impact de la simplification pour l'extraction des interactions entre protéines et montrent que sur 5 corpus différents l'extraction des relations est meilleure. Deux autres travaux portent sur la simplification des arbres de dépendances pour la tâche d'extraction d'interactions entre protéines (Thomas *et al.*, 2011), par suppression ou modification de types de dépendances, et pour la tâche BioNLP'09 [3] (extraction d'événements biologiques) (Buyko *et al.*, 2011), par élagage de l'arbre.

Dans un autre domaine, l'annotation des rôles sémantiques, (Vickrey et Koller, 2008) ont écrit 154 règles s'appliquant à l'arbre de constituants pour supprimer toute l'information en dehors du verbe cible et de ses arguments. Ils proposent une méthode originale pour sélectionner les meilleures règles : ils appliquent les règles de simplification pour produire toutes les phrases simplifiées possibles, puis entraînent leur système d'annotation des rôles sémantiques. La validité de chaque règle est ensuite évaluée en fonction de l'impact de la simplification sur la tâche principale.

La méthode de simplification que nous proposons dans cet article, est fondée sur un apprentissage automatique, contrairement aux travaux que nous venons de présenter. Le mode d'apprentissage semi-supervisé se rapproche de celui développé par (Vickrey et Koller, 2008) pour la tâche d'annotation des rôles sémantiques. En effet, nous proposons d'annoter la simplification en apprenant sur un petit corpus annoté, puis d'évaluer l'annotation selon son impact sur l'extraction des relations, et enfin nous complétons le corpus annoté grâce aux résultats de l'extraction des relations. Cette méthode se rapproche ainsi des travaux en compression de phrases, qui consiste à supprimer certains constituants d'une phrase, considérés comme non essentiels. Les approches en compression de phrases peuvent se fonder sur des règles linguistiques (Yousfi-Monod et Prince, 2006) ou sur un apprentissage(Knight et Marcu, 2000; Waszak et Torres-Moreno, 2008). Cependant, notre tâche s'en distingue par deux aspects : notre objectif n'est pas de simplifier les phrases en fonction des informations saillantes, mais en fonction des informations relatives à l'extraction d'une relation, et par ailleurs, nous souhaitons développer un système qui ne nécessite pas l'annotation d'un grand corpus pour la simplification.

3. http ://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/

3 Extraction de relations en domaine médical

Nous avons développé un système pour l'extraction de relations dans le domaine biomédical. Il utilise un classifieur à base de SVM, avec la bibliothèque LIBSVM (Chang et Lin, 2001). Le système peut être utilisé pour une classification binaire ou multi-classes (avec une approche "un-contre-un"). Il utilise des attributs qui prennent en compte des informations de surface, sur les distances entre mots par exemple, des informations lexicales, comme les mots formant les concepts, des informations syntaxiques, les catégories morpho-syntaxiques des mots, et des informations sémantiques grâce au typage des concepts. Une description détaillée du système est donnée dans (Minard *et al.*, 2011b).

Ce système a été utilisé pour le challenge i2b2 2010, pour le challenge DDI 2011 (extraction d'interactions entre médicaments (Minard *et al.*, 2011c)) et également pour l'extraction d'interactions entre protéines. Dans cet article, les tests sont effectués sur le corpus i2b2 2010 que nous présentons dans la section suivante.

3.1 Corpus de comptes-rendus médicaux

Dans le cadre du challenge i2b2 2010, un corpus annoté composé de rapports cliniques a été fourni aux participants. Les comptes rendus du corpus proviennent de 7 centres médicaux des États-Unis. Ils ont été manuellement anonymisés et annotés. Trois types de concepts ont été annotés : les problèmes médicaux (maladies, syndromes, observations sur l'état psychologique du patient, etc.), les traitements (interventions, médicaments donnés au patient, etc.) et les tests (procédures et examens). Entre ces trois types de concepts, 8 relations peuvent exister :
– un traitement améliore (TrIP), aggrave (TrWP) ou cause (TrCP) un problème médical ;
– un traitement est administré (TrAP) ou pas (TrNAP) pour un problème médical ;
– un test révèle [4] (TeRP) ou est conduit pour examiner (TeCP) un problème médical ;
– un problème médical indique un autre problème médical (PIP).
Le corpus de développement (DEV_I2B2) est composé de 349 documents (4994 relations) et le corpus d'évaluation (EVAL_I2B2) de 477 documents (9070 relations). Nous avons divisé le corpus DEV_I2B2 en deux parties afin de pouvoir entraîner notre système avant d'avoir le corpus d'évaluation : un corpus d'entraînement (TRAIN_I2B2) composé de 295 documents (4515 relations) et un corpus de test (TEST_I2B2) composé de 54 documents (479 relations). Dans le graphique 1, nous avons représenté le nombre d'instances de relations de chaque type dans les différents corpus.

3.2 Résultats obtenus

Les résultats que nous avons obtenus avec ce système sont présentés dans la figure 2. Nous avons également représenté sur le graphique l'accord inter annotateur (IAA) et le nombre d'exemples de chaque relation (nombre normalisé pour être à l'échelle du graphique). Globalement la F-mesure est d'environ 0,7 pour la classification des relations, mais cette classification est moins bonne

4. Cette relation correspond aux cas où le test indique la présence d'un problème, ou bien l'absence d'un problème. Pour ces relations, la présence de négations, très fréquentes en domaine médical, ne sera donc pas prise en compte.

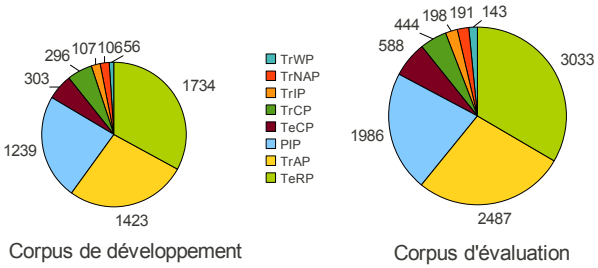


FIGURE 1 – Composition des corpus

pour les relations pour lesquelles peu d'exemples ont été annotés dans le corpus DEV_I2B2 (par exemple pour la relation TrWP ou TrIP).

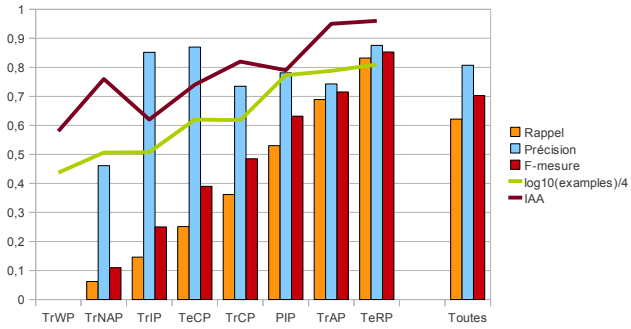


FIGURE 2 – Évaluation du système final

Les types d'erreurs que l'on peut relever proviennent du fait qu'il y a peu d'exemples pour certaines relations, associé à une grande variabilité des expressions, qu'il y a besoin de connaissances externes, et que parfois la classification est discutable car les définitions de certaines relations sont assez proches (Minard *et al.*, 2011b). Le présent travail concerne la réduction de la variabilité syntaxique des expressions par simplification. Ainsi, la relation TeCP entre *pulmonary nodules* et *fu imaging* dans la phrase de l'exemple 1 est mal classée par le système notamment à cause de la présence du verbe *reveal* qui indique généralement une relation TeRP. Dans ce type de construction, il serait intéressant de supprimer la partie de la phrase concernant les premiers tests et problèmes, pour ne conserver que la paire à étudier.

(1) [TEST CTS chest] was negative for [PB PE], however it did reveal

[PB **pulmonary nodules in his RML**] which need [TEST **fu imaging**] in <NUM> months.

Dans l'exemple 2, les deux concepts à annoter *his high right-sided filling pressure* et *his Captopril* sont séparés par la proposition *he was started on Lasix for diuresis* qui n'est pas pertinente. Si les éléments non pertinents pour cette relation étaient supprimés, la phrase pourrait devenir *Given* PROBLEM, TREATMENT *was increased.*, forme qui est fréquente dans le corpus et permet de reconnaître une relation TrAP. Une telle simplification de phrases permettrait bien de réduire la variabilité du corpus pour améliorer la classification des relations.

(2) Given [PB **his high right-sided filling pressure**], he was started on [TREAT Lasix] for [TREAT diuresis] and [TREAT **his Captopril**] was increased for [TREAT greater afterload reduction].

4 Définition du modèle de simplification

4.1 Simplification

Nous définissons la simplification comme une extraction de l'information pertinente pour identifier des relations, qui consiste à ne garder que ce qui est nécessaire à l'identification de la relation, et à supprimer les informations qui ne sont pas en rapport avec la relation ou qui peuvent perturber son identification.

Pour cela, la plupart des travaux ont défini des règles de simplification. Les exemples montrent que celles-ci sont très contextuelles et dépendantes de la tâche, et reposent sur une étude de corpus plutôt que sur une connaissance a priori de la langue. Ainsi, des règles usuellement définies pour la simplification comme la suppression de relatives ne s'appliquent pas dans notre contexte. Une modélisation sous forme de règles nécessiterait d'en redéfinir un grand nombre, ce qui nous a poussé à privilégier une méthode à base d'apprentissage pour annoter dans les phrases les parties à garder et celles que l'on peut supprimer.

Quatre types d'annotation ont été définis. L'annotation «indispensable» permet de caractériser les mots qui portent l'expression de la relation. L'annotation «utile», très proche de «indispensable», indique les mots qui renforcent la relation. Ensuite l'annotation «inutile» est associée aux mots n'apportant pas d'indices pour la classification de la relation, par exemple l'indication du service dans lequel est le patient. L'annotation «gênant» sert à repérer les mots pouvant gêner la bonne classification de la relation. Dans les exemples 3 et 4, les parties de phrase «indispensables» sont soulignées, les parties «inutiles» sont normales, les parties «gênantes» sont barrées et les concepts à mettre en relation sont en gras. Dans l'exemple 3, il s'agit de déterminer la relation TeCP entre *a magnetic resonance imaging study* et *a small vascular malformation*. Dans l'exemple 4, il s'agit d'une relation TrAP entre *the tremendous tumor burden* et *open debulking*.

(3) [TEST **A magnetic resonance imaging study**] will be scheduled as an outpatient in three months to rule out [PB**a small vascular malformation**] if responsible for [PBthe hemorrhage].

(4) The neuro-oncologist felt that because of [PB **the tremendous tumor burden**] ~~that was likely causing his symptoms~~ the patient will require [TREAT **open debulking**] as well as obtaining issue for a pathologic diagnosis.

4.2 Méthode

Nous avons choisi d'utiliser un classifieur à base de CRF («Champs Aléatoires Conditionnels») pour effectuer l'annotation des phrases. Les CRF sont des modèles statistiques qui ont la particularité de modéliser des dépendances entre annotations. Les phrases annotées seront données en entrée du classifieur SVM. Afin de n'annoter que quelques phrases, nous proposons une architecture où la simplification est guidée par la tâche d'extraction de relations, et le corpus d'apprentissage de la simplification est augmenté itérativement en fonction des résultats de la tâche finale. La combinaison des classifieurs est présentée dans la section 6. Cette méthode est donc facilement adaptable à un autre domaine, contrairement aux méthodes à base de règles, qui ne permettent pas toujours une adaptation simple et rapide.

5 Annotation par CRF

5.1 Constitution du corpus d'apprentissage

Nous avons sélectionné 71 phrases provenant du corpus TRAIN_I2B2. Nous avons extrait aléatoirement 14 phrases contenant des paires d'entités qui avaient été correctement classées par notre système d'extraction de relations, 37 paires mal classées et 20 paires qui ne sont pas en relation mais qui avaient été classées comme étant en relation. Une étude de leurs caractéristiques a été menée préalablement à l'annotation.

Cette étude a montré que dans 14 phrases du corpus, la relation est exprimée par un verbe et les deux concepts en relation sont respectivement sujet et complément de ce verbe (exemple 5).

(5) [TEST **An magnetic resonance imaging study**] **showed** [PB **basilar artery disease**], questionable aneurysm.

Dans 14 phrases, les deux concepts en relation sont dans deux propositions différentes (exemple 6). Sept constructions différentes ont été trouvées ; nous en présentons trois dans le tableau 1.

(6) Finger tapping and [TEST **rapid alternating movements**] were slow on the left and she had [PB **trouble isolating individual finger movements**].

Dans 18 phrases, les deux concepts sont reliés par une préposition, et la relation s'exprime au travers de la préposition et du verbe de la proposition (exemple 7).

(7) [...], she had [PB **an acute drop**] **in** [TEST **her systolic blood pressure**] to <NUM> for unclear reasons and without evidence of acute_sepsis.

Prop Conj | Prop Princ	Although **TREAT** were adjusted he continued to be **PB** and there was [...]
Prop Indep | CC | Prop Indep	TEST became **PB** and **TREAT** was held.
Prop Princ | Prop Rel | CC | Prop Indep	He subsequently became PB and **PB** in the Catheterization Laboratory which responded to TREAT, TREAT, **TREAT**, and he was then transferred to the CCU for TEST.

TABLE 1 – Phrases dans lesquelles les concepts en relation sont dans deux propositions différentes

Dans les exemples de non-relation que nous avons dans notre corpus, dans seulement deux phrases les deux concepts sont sujet et objet du même verbe. Dans 8 phrases, les deux concepts sont dans des propositions différentes et dans 9 phrases ils sont reliés par une préposition.

Cette étude fait apparaître l'existence de régularités, que la simplification pourrait dégager.

Les 71 phrases ont été annotées par 3 annotateurs grâce au logiciel Knowtator de Protégé [5]. Les différences ont donné lieu à discussion et accord.

Une phrase pouvant contenir plus d'une paire d'entités, nous les avons annotées pour une paire d'entités définie ; de ce fait certaines phrases sont en double dans le corpus, mais à chaque fois pour une paire d'entités différente.

Dans le tableau 2, nous donnons pour chaque classe de simplification le nombre de mots associés à cette classe dans le corpus TRAIN_SIMP (le corpus annoté obtenu). On remarque que très peu de mots sont annotés «utile», la raison étant la difficulté de distinction entre les classes «indispensable» et «utile». Ces deux classes seront donc regroupées ultérieurement.

étiquette	nombre de mots
indispensable	287
utile	52
inutile	608
gênant	177

TABLE 2 – Étude du corpus annoté

5.2 Application du CRF

Nous avons utilisé le classifieur CRF++ (Kado, 2003) pour apprendre à annoter les simplifications : à chaque mot il attribue une étiquette en fonction de la valeur des attributs pour ce mot.

Les attributs fournis au classifieur sont : le lemme, la catégorie morpho-syntaxique, le nombre de caractères du token, la position du token dans la phrase (position du token/nombre de tokens dans la phrase), le type sémantique si le token fait partie d'une entité et une étiquette indiquant si le token fait partie d'une des entités de la paire étudiée. Ce dernier attribut permet d'avoir une

5. http ://knowtator.sourceforge.net/

annotation dépendante d'un couple particulier de concepts. Les dépendances séquentielles sont calculées, pour chaque type d'attributs, avec un contexte de trois mots avant et trois mots après le token courant.

Nous n'avons pas de corpus annoté pour évaluer la simplification. Pour vérifier que l'annotation de la simplification avec CRF++ était cohérente, nous avons étudié les mots classés dans chacune des trois catégories. Pour chaque lemme, nous avons compté combien de fois il apparaissait dans le corpus TRAIN_I2B2 et combien de fois il était associé à une des trois catégories. Le tableau 3 contient les lemmes les plus fréquents dans chaque catégorie et qui apparaissent au moins 10 fois dans cette catégorie. Nous observons que les lemmes les plus fréquemment étiquetés «gênant» font partie de concepts ; par exemple on retrouve *fluticasone* dans le traitement *fluticasone propionate* ou *fluticasone-salmeterol*. Les lemmes les plus souvent étiquetés «utile» sont principalement des verbes, et ceux étiquetés «inutile» sont des unités (reliées à des dosages), des informations sur le patient (son nom, son âge), etc. Nous avons conclu de cette étude que le classifieur se comporte de manière cohérente pour annoter la simplification.

UTILE		INUTILE		GENANT	
attribute	50 / 56	ml	582 / 582	neutropenia	10 / 10
presence	12 / 17	before	260 / 260	ph	19 / 21
questionable	16 / 23	yo (*year-old*)	219 / 219	thromboplastin	23 / 27
vs	21 / 31	microgram	211 / 211	fluticasone	11 / 13
identify	27 / 40	caution	201 / 201	diskus	11 / 13
demonstrate	130 / 194	mr.	184 / 184	migraine	52 / 62
inaccurate	12 / 18	ask	177 / 177	spiriva	22 / 27
due	314 / 488	asacol	172 / 172	panic	42 / 52

TABLE 3 – Exemple d'annotation de lemmes présents plus de 10 fois dans le corpus

6 Combinaison de classifieurs pour l'extraction des relations

Avec seulement 71 phrases annotées, la simplification obtenue ne permet pas d'améliorer l'extraction des relations. Pour augmenter le corpus TRAIN_SIMP et améliorer la simplification, nous avons combiné les deux classifieurs, et utilisé les résultats de la classification des relations pour augmenter le corpus TRAIN_SIMP. La figure 3 présente de façon simplifiée la méthode développée.

Annotation de la simplification Dans un premier temps, les 71 phrases annotées manuellement sont utilisées comme amorce pour la simplification ; elles forment le corpus d'entraînement TRAIN_SIMP. Elles sont utilisées pour apprendre les simplifications grâce à l'outil CRF++. Ensuite le modèle pour la simplification est appliqué sur la totalité du corpus DEV_I2B2.

Extraction des relations Nous utilisons ensuite ce corpus annoté pour extraire les relations grâce à notre classifieur à base de SVM. Les annotations des simplifications sont utilisées comme

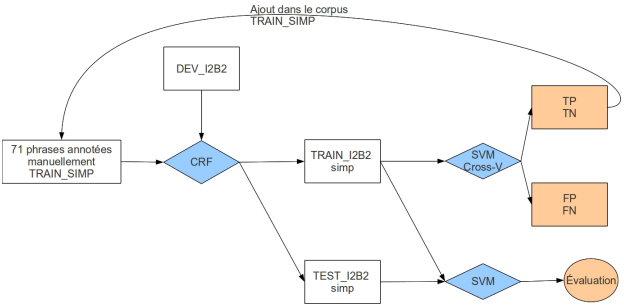


FIGURE 3 – Schéma explicatif de la méthode

des attributs supplémentaires pour la reconnaissance des relations. Un attribut identifie les tokens annotés comme «inutile», un autre pour les tokens «gênants» et un pour les «utiles».

Le corpus DEV_I2B2 a été divisé en deux : corpus TRAIN_I2B2 et corpus TEST_I2B2. Dans un premier temps, nous effectuons l'extraction des relations par validation croisée en 5 parties avec le corpus TRAIN_I2B2. À chaque itération, les phrases contenant des relations correctement extraites (les vrais positifs et les vrais négatifs) et leurs annotations pour la simplification sont ajoutées au corpus TRAIN_SIMP.

Validation de la simplification Une fois la validation croisée terminée sur le corpus TRAIN_I2B2, le corpus DEV_I2B2 est réannoté pour la simplification à l'aide du classifieur à base de CRF et du corpus TRAIN_SIMP augmenté. Ensuite le modèle pour l'extraction des relations est construit à partir du corpus TRAIN_I2B2 et il est appliqué sur le corpus TEST_I2B2. Cette étape permet d'évaluer l'impact de la complétion du corpus TRAIN_SIMP sur l'extraction des relations.

7 Évaluations

Afin d'évaluer les résultats de la simplification, nous avons défini plusieurs protocoles qui permettent de :
– mesurer l'impact de la simplification sur le processus de classification ;
– noter le résultat de la simplification ;
– analyser de manière qualitative les annotations effectuées par le module de simplication.

7.1 Mesure de l'impact de la simplification sur la tâche d'extraction de relations

Pour évaluer l'impact de la simplification sur la classification des relations, nous avons mené plusieurs expérimentations en faisant varier un grand nombre de paramètres. Dans un premier temps, nous pouvons faire varier les attributs utilisés par les CRF pour apprendre la simplification : nous pouvons par exemple ajouter la structure syntaxique de la phrase. Nous pouvons également n'utiliser que deux classes pour la simplification, c'est-à-dire ne pas faire de distinction entre les mots inutiles et gênants, et utiles et indispensables. Deuxièmement, les informations sur la simplification peuvent être prises en compte de 2 manières par le système d'extraction de relation : les mots gênants (voire gênants et inutiles) peuvent être supprimés de la phrase ou des attributs indiquant la classe du mot peuvent être ajoutés. Finalement, nous pouvons faire varier la sélection des paires d'entités correctement classées à ajouter au corpus TRAIN_SIMP. Toutes les paires correctement classées (que les entités soient en relation ou non) peuvent être ajoutées, ou seules les paires qui n'étaient pas bien classées avec le système sans simplification et qui le sont avec la simplification, ou selon le score de décision donné par le classifieur, etc.

Nous présentons ici la configuration donnant les meilleurs résultats. Nous avons appris la simplification en ne donnant que les attributs de base (voir 5.2) et en apprenant 3 classes («utile» et «indispensable» sont regroupées en une classe, «inutile» et «gênant»). Pour prendre en compte la simplification, nous avons donné des attributs supplémentaires au classifieur. Comme il est difficile d'annoter des phrases pour des paires d'entités qui ne sont pas en relation, nous avons modifié le corpus annoté manuellement et nous avons annoté en «inutile» tous les mots de la phrase. Ensuite, après avoir classé les relations par validation croisée sur le corpus TRAIN_I2B2, nous avons ajouté dans le corpus pour la simplification TRAIN_SIMP les phrases contenant des relations correctement classées et dont au moins un des mots avait été annoté «utile», et les phrases contenant des paires qui ne sont pas en relation et qui ont été correctement classées uniquement avec la simplification (elles étaient mal classées par le système n'utilisant pas la simplification). Nous avons exécuté 4 fois le système complet, après quoi nous avons obtenu 589 phrases dans le corpus TRAIN_SIMP, dont 71 qui ont été annotées manuellement. Nous avons appliqué la simplification sur le corpus d'évaluation EVAL_I2B2 afin d'évaluer la classification des relations. Dans le tableau 4, nous donnons les F-mesures obtenues sans simplification, avec la simplification apprise avec les 71 phrases annotées (avant la combinaison des deux méthodes) et avec le corpus TRAIN_SIMP obtenu après 4 itérations.

La F-mesure calculée pour toutes les relations reste stable avec ou sans la simplification même si la F-mesure pour 4 des relations diminue quand nous utilisons la simplification. La différence entre les résultats avec et sans simplification calculée avec le test T de Student est significative ($p<=0,05$), la simplification a donc un effet sur la classification mais ne permet pas encore de l'améliorer.

7.2 Évaluation manuelle de la simplification

Nous n'avons pas de corpus annoté suffisament grand pour pouvoir faire une évaluation automatique de la tâche de simplification. De ce fait, nous avons choisi d'annoter manuellement 41 relations et d'évaluer manuellement la simplification pour ces relations. Nous n'avons étudié que l'annotation des phrases portant sur une paire de concepts en relation. En effet, ainsi que nous

Relations	Sans Simplification	Simplification	
		Corpus TRAIN_SIMP non augmenté	Corpus TRAIN_SIMP augmenté
TrIP	**0,315**	0,266	0,302
TrWP	0,000	0,000	0,000
TrCP	**0,486**	0,464	0,470
TrAP	**0,732**	0,724	0,730
TrNAP	**0,195**	0,151	0,168
PIP	0,625	0,627	**0,630**
TeRP	0,852	0,852	**0,855**
TeCP	**0,452**	0,398	0,408
Toutes les relations	**0,709**	0,704	0,708

TABLE 4 – Évaluation de la classification des relations avec et sans simplification sur le corpus EVAL_I2B2

l'avons déjà mentionné, il est difficile de définir ce qui doit être annoté pour les non relations.

Nous avons donc annoté 41 relations du corpus de test et avons comptabilisé le nombre de relations correctement simplifiées, simplifiées à tort ou partiellement simplifiées à raison. Peu d'informations sont annotées comme gênantes. Aussi, lors de l'évaluation, nous considérons que des informations annotées indispensables sont des informations à garder et que les autres sont des informations à supprimer. Nous avons considéré exacts les cas où le module garde toutes les informations pertinentes, même s'il garde aussi quelques informations que nous jugeons inutiles. Nous avons considéré comme faux les simplifications qui suppriment des informations que nous jugeons indispensables, et partiellement corrects les cas où le module aurait dû garder plus d'informations utiles, mais a gardé quand même les informations indispensables, ou lorsque trop d'informations qu'il aurait dû considérer comme inutiles sont gardées. Avec cette répartition en trois classes, nous obtenons 19 cas exacts, 16 cas faux et 6 cas partiellement corrects.

Dans l'exemple 8, nous considérons que la simplification est correcte mais dans l'exemple 9 le verbe le plus utile à la détection de la relation (*revealed*) est annoté inutile, et l'annotation de la simplification est donc fausse.

(8) He had [TEST **a cardiac catheterization**] performed which revealed [PB **a three vessel coronary artery disease**] with [PB an occluded RCA] , [PB 70%-80% proximal LAD] , and [PB a high grade left circumflex lesion after the OM] with [PB distal left circumflex occlusion] .

(9) He had [TEST **a cardiac catheterization**] performed which revealed [PB a three vessel coronary artery disease] with [PB an occluded RCA] , [PB **70%-80% proximal LAD**] , and [PB a high grade left circumflex lesion after the OM] with [PB distal left circumflex occlusion] .

7.3 Analyse des simplifications

Nous avons tenté d'établir les types de simplifications apprises. Les différentes structures de phrases qui apparaissent sont :
– *concept1 relation concept2* pour lesquelles la partie située entre les concepts doit être conservée, tout ou en partie ; cette structure est généralement bien traitée ;
– *concept1 relation (coordination de concepts) concept2* est généralement mal annotée, et la marque de la relation est souvent supprimée. Ce type de structure peut être reconnu simplement par des règles ;
– *concept1 (structure comportant des concepts) relation concept2* est généralement reconnue et la relation est gardée.

Certains cas nécessitent de garder la partie gauche du premier concept ; cette configuration est mal reconnue. Il en est de même pour les contextes droits du deuxième concept. Ces deux types de structure sont plus rares, et leur traitement nécessite plus d'exemples.

Cette étude nous amène à imaginer des améliorations de notre système. Il serait par exemple intéressant de pouvoir mieux sélectionner les phrases ajoutées au corpus TRAIN_SIMP pour diversifier les exemples de simplification. Une solution serait d'identifier les paires d'entités moins bien classées avec l'utilisation de la simplification que sans, et de les annoter pour apprendre de nouveaux schémas de simplification. Nous pourions également envisager d'utiliser quelques règles pour annoter les cas les plus courants (par exemple pour supprimer les concepts en coordination ou encore les indications de lieux), puis d'utiliser le système à base d'apprentissage.

8 Conclusion

Dans cet article, nous nous sommes intéressées à la simplification de phrases dans le but d'améliorer l'extraction de relations. Nous avons présenté une méthode de simplification guidée par la tâche d'extraction de relations, et nécessitant un petit corpus annoté. Les résultats que nous obtenons sur la tâche finale, à savoir l'extraction de relations, sont significativement différents des résultats de la classification sans simplification, mais la F-mesure finale reste stable. La poursuite de l'étude pourrait porter sur l'amélioration de la sélection des phrases ajoutées au corpus d'apprentissage pour la simplification, par exemple en ne gardant que celles dont le score de confiance du classifieur est élevé. Nous devons également étudier la façon dont nous traitons les cas de non-relation ; devons-nous ajouter des exemples au corpus d'apprentissage ou non, si oui, comment les annoter, etc. Pour finir, un prétraitement à base de règles sur les phrases du corpus pourrait permettre d'annoter les indications temporelles, de lieux (par exemple le nom d'une clinique), l'âge des patients, etc. et ainsi réduire d'avantage la variabilité.

Références

Buyko, E., Faessler, E., Wermter, J. et Hahn, U. (2011). Syntactic simplification and semantic enrichment—trimming dependency graphs for event extraction. 27:610–644.

Chang, C.-C. et Lin, C.-J. (2001). *LIBSVM : a library for support vector machines*. Software available at http ://www.csie.ntu.edu.tw/ cjlin/libsvm.

HEILMAN, M. et SMITH, N. A. (2010). Extracting Simplified Statements for Factual Question Generation. *In Proceedings of the 3rd Workshop on Question Generation*.

JONNALAGADDA, S. et GONZALEZ, G. (2010). Sentence simplification aids protein-protein interaction extraction. *CoRR*, abs/1001.4273.

KADO, T. (2003). CRF++ : Yet another crf toolkit. `http://crfpp.sourceforge.net/`. [consulté le 17/01/2012].

KNIGHT, K. et MARCU, D. (2000). Statistics-based summarization-step one : Sentence compression. *In Proceedings of the National Conference on Artificial Intelligence*, pages 703–710. Menlo Park, CA ; Cambridge, MA ; London ; AAAI Press ; MIT Press ; 1999.

MINARD, A.-L., LIGOZAT, A.-L. et GRAU, B. (2011a). Apport de la syntaxe pour l'extraction de relations en domaine médical. *In Actes TALN 2011*, pages 383–393.

MINARD, A.-L., LIGOZAT, A.-L. et GRAU, B. (2011b). Extraction de relations dans des comptes rendus hospitaliers. *In Actes des 22èmes Journées francophones d'Ingénierie des Connaissances (IC'2011)*.

MINARD, A.-L., LIGOZAT, A.-L., GRAU, B. et MAKOUR, L. (2011c). Feature selection for drug-drug interaction detection using machine-learning based approaches. *In SEPLN'11, Workshop Drug-Drug Interaction*.

MIWA, M., S, R., MIYAO, Y. et TSUJII, J. (2010). Entity-focused sentence simplification for relation extraction. *In Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 788–796, Stroudsburg, PA, USA. Association for Computational Linguistics.

ROBERTS, A., GAIZAUSKAS, R. et HEPPLE, M. (2008). Extracting clinical relationships from patient narratives. *In BioNLP2008 : Current Trends in Biomedical Natural Language Processing*, pages 10–18.

THOMAS, P., PIETSCHMANN, S., SOLT, I., TIKK, D. et LESER, U. (2011). Not all links are equal : Exploiting dependency types for the extraction of protein-protein interactions from text. *In Proceedings of BioNLP 2011 Workshop*, pages 1–9, Portland, Oregen, USA. Association for Computational Linguistics.

UZUNER, O., MAILOA, J., RYAN, R. et SIBANDA, T. (2010). Semantic relations for problem-oriented medical records. *Artificial Intelligence in Medicine*, 50:63–73.

VICKREY, D. et KOLLER, D. (2008). Sentence Simplification for Semantic Role Labeling. *In Proceedings of ACL-08 : HLT*, pages 344–352, Columbus, Ohio. Association for Computational Linguistics.

WASZAK, T. et TORRES-MORENO, J. (2008). Compression entropique de phrases contrôlée par un perceptron. *Journées internationales d'Analyse statistique des Données Textuelles*.

WOODSEND, K. et LAPATA, M. (2011). Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK. Association for Computational Linguistics.

YOUSFI-MONOD, M. et PRINCE, V. (2006). Compression de phrases par élagage de leur arbre morpho-syntaxique. Une première application sur les phrases narratives. *TSI : Revue Technique et Science Informatiques*, 25(4):437–468.

ZHANG, M., ZHANG, J. et SU, J. (2006). Exploring syntactic features for relation extraction using a convolution tree kernel. *In Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 288–295.

Extraction d'information automatique en domaine médical par projection inter-langue : vers un passage à l'échelle

Asma Ben Abacha* Pierre Zweigenbaum* Aurélien Max*
(*) LIMSI-CNRS, BP 133 91403 Orsay cedex
abacha@limsi.fr, pz@limsi.fr, aurelien.max@limsi.fr

RÉSUMÉ

Cette recherche est issue de notre volonté de tester de nouvelles méthodes automatiques d'annotation ou d'extraction d'information à partir d'une langue L1 en exploitant des ressources et des outils disponibles pour une autre langue L2. Cette approche repose sur le passage par un corpus parallèle (L1-L2) aligné au niveau des phrases et des mots. Pour faire face au manque de corpus médicaux français annotés, nous nous intéressons au couple de langues (français-anglais) dans le but d'annoter automatiquement des textes médicaux en français. En particulier, nous nous intéressons dans cet article à la reconnaissance des entités médicales. Nous évaluons dans un premier temps notre méthode de reconnaissance d'entités médicales sur le corpus anglais. Dans un second temps, nous évaluons la reconnaissance des entités médicales du corpus français par projection des annotations du corpus anglais. Nous abordons également le problème de l'hétérogénéité des données en exploitant un corpus extrait du Web et nous proposons une méthode statistique pour y pallier.

ABSTRACT

Automatic Information Extraction in the Medical Domain by Cross-Lingual Projection

This research stems from our willingness to test new methods for automatic annotation or information extraction from one language L1 by exploiting resources and tools available to another language L2. This approach involves the use of a parallel corpus (L1-L2) aligned at the level of sentences and words. To address the lack of annotated medical French corpus, we focus on the French-English language pair to annotate automatically medical French texts. In particular, we focus in this article on medical entity recognition. We evaluate our medical entity recognition method on the English corpus and the projection of the annotations on the French corpus. We also discuss the problem of scalability since we use a parallel corpus extracted from the Web and propose a statistical method to handle heterogeneous corpora.

MOTS-CLÉS : Extraction d'information, projection d'annotation, reconnaissance des entités médicales, apprentissage.

KEYWORDS: Automatic Information Extraction, Annotation Projection, Medical Entity Recognition, Machine Learning.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 15–28,
Grenoble, 4 au 8 juin 2012.

1 Introduction

L'extraction d'information vise à extraire automatiquement à partir de textes des informations structurées pertinentes pour une tâche particulière (Poibeau, 2003). Il y a essentiellement deux types de méthodes utilisées en extraction d'information : les méthodes où une personne (un « expert ») fournit des connaissances (linguistiques ou sur le domaine) [1], et les méthodes dirigées par les données, où ces connaissances sont construites par apprentissage supervisé. Il existe également des méthodes hybrides combinant ces deux techniques. Ces deux types de méthodes ont certaines limitations ((Bach et Badaskar, 2007), (Nadeau et Sekine, 2007), (Ben Abacha et Zweigenbaum, 2011c)) :

– Les méthodes à base de connaissances expertes sont simples à mettre en place mais coûteuses en temps pour ce qui est de la construction des connaissances. Elles ont aussi un potentiel de couverture réduit comparé aux méthodes statistiques.
– Les méthodes par apprentissage peuvent être très robustes si (i) on dispose d'un bon nombre d'exemples d'entraînement et si (ii) le corpus de test est du même type que le corpus d'entraînement. Ces méthodes sont de fait dépendantes des données et des corpus annotés, ressources qui ne sont pas disponibles pour toutes les langues (par exemple, il n'existe pas de corpus médicaux annotés en français) ni pour toutes les tâches (par exemple, reconnaissance des entités médicales, extraction de relations sémantiques, etc.).

Cette observation s'applique aussi au domaine médical : pour l'anglais, plusieurs outils spécialisés d'extraction d'information existent (tels que MetaMap (Aronson, 2001), cTAKES (Guergana K Savova et Chute, 2010)), ainsi que des corpus annotés en entités nommées (tels que i2b2 (Uzuner *et al.*, 2011), Berkeley (Rosario et Hearst, 2004)). En revanche, peu de ressources sont disponibles en français : on ne trouve pas d'outils spécialisés pour l'extraction d'information, ni de corpus médicaux annotés.

L'annotation manuelle d'exemples pour l'entraînement peut être une solution pour les méthodes par apprentissage supervisé ou semi-supervisé. Cependant, cette tâche nécessite des experts du domaine ciblé, au moins pour la validation. D'après nos expériences précédentes portant sur l'annotation manuelle de corpus médicaux en français constitués (i) de résumés d'articles scientifiques et (ii) de textes extraits du corpus EQueR (Ayache, 2005), plusieurs obstacles ont été mis en évidence. Dans une première phase, nous avons annoté manuellement des textes médicaux avec le concours de deux médecins. L'obstacle principal était le fait que la tâche est longue et fastidieuse. Ensuite, et pour accéler la tâche d'annotation, nous avons développé une interface pour l'annotation de phrases (et non pas de textes entiers) permettant à davantage de médecins de prendre part à l'annotation. Le premier inconvénient de cette méthode est la perte du contexte des phrases. Un deuxième inconvénient réside dans le fait que, même si le guide d'annotation est très détaillé, les divergences dans les avis des médecins augmentent avec le nombre de médecins intervenant (par exemple dans l'annotation des symptômes et des relations dans des textes dans le domaine psychiatrique). Ces divergences, portant par exemple sur les types d'entités médicales et les relations à annoter, peuvent ralentir le processus d'annotation manuelle et le rendre moins fiable.

Dans cet article, nous exploitons un autre type de méthode, la *projection d'annotations* d'une langue à une autre (Yarowsky et Ngai, 2001), et testons son application au domaine médical. L'idée générale consiste à transférer des annotations d'une langue L1 (pour laquelle plus de

1. Méthodes souvent appelées improprement « à base de règles ».

ressources sont disponibles) à une langue L2 en utilisant des corpus parallèles et leur alignement au niveau des mots. Cette approche devrait nous permettre d'exploiter, pour l'annotation automatique de textes en français, les ressources disponibles en anglais ainsi que les méthodes d'extraction d'information développées pour cette même langue. Notre premier objectif, présenté à travers cet article, consiste à annoter automatiquement les entités médicales de textes en français par transfert d'entités détectées dans les textes anglais correspondants par des outils existants de reconnaissance d'entités médicales. La table 1 présente un exemple de ce que nous cherchons à obtenir.

Phrase en anglais	The role of carotid endarterectomy in the management of asymptomatic carotid stenosis is much less clear.
Phrase équivalente en français	Le rôle de l'endartériectomie carotidienne dans le traitement d'une sténose carotidienne asymptomatique est beaucoup moins clairement défini.
Alignement au niveau des mots	0-0 1-1 2-2 3-3 3-4 4-3 5-5 6-6 7-7 8-8 9-11 10-8 11-9 11-10 12-12 13-13 14-14 15-15 15-16
Entités médicales (en anglais)	"carotid endarterectomy" 3–4 [treatment] "asymptomatic carotid stenosis" 9–11 [problem]
Résultat final (annotations en français)	"l'endartériectomie carotidienne " 3–4 [treatment] "une sténose carotidienne asymptomatique" 8-11 [problem]

TABLE 1 – Exemple illustratif de l'approche proposée

Après un rappel des travaux similaires (section 2), nous présentons les deux étapes principales de l'approche que nous proposons ici (telle qu'illustrée sur la figure 1) :

1. L'extraction d'information à partir de la partie L1 du corpus parallèle, en utilisant des méthodes déjà développées ou des outils disponibles (section 3). Pour ce faire, nous utilisons une méthode à base de connaissances expertes, MetaMapPlus (section 3.2) que nous adaptons pour traiter des corpus hétérogènes (section 3.3).
2. L'alignement des mots des parties L1 et L2 du corpus (section 4.1) et la projection des entités repérées sur L1 vers la partie L2 en utilisant ces alignements (section 4.2). Nous mettons en place quelques heuristiques pour réparer certaines erreurs et améliorer la précision de la projection en diminuant le bruit des alignements.

Nous évaluons notre approche (section 5) sur une partie du corpus Santé Canada [2] et discutons ses résultats, puis concluons (section 6) sur des perspectives de travaux futurs [3].

2 Travaux similaires

Des travaux sur la projection d'analyses linguistiques ou d'annotations d'une langue à l'autre se sont développés essentiellement à partir des années 2000. Yarowsky et Ngai (2001) ont proposé

2. http://www.hc-sc.gc.ca
3. Ce travail a été partiellement soutenu par OSEO dans le cadre du programme Quæro.

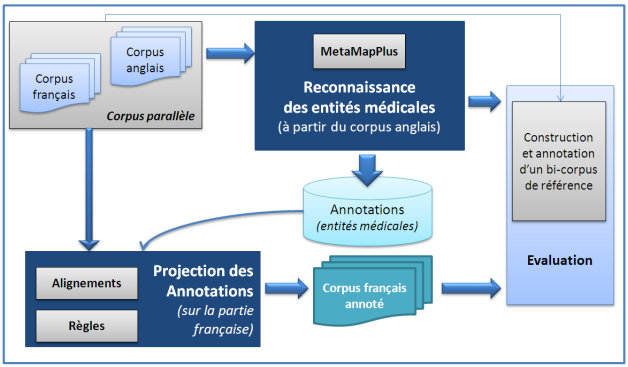


FIGURE 1 – Approche proposée pour l'annotation automatique d'un corpus médical français, utilisant un corpus parallèle et des méthodes d'extraction d'information à partir de textes anglais

d'utiliser des corpus parallèles alignés au niveau des mots pour transférer de façon robuste de l'anglais au français ou au chinois des étiquettes morphosyntaxiques et des frontières de syntagmes nominaux. Lopez *et al.* (2002) ont étudié comment transférer un arbre de dépendances de l'anglais vers le chinois. Lü *et al.* (2002) se sont également intéressés au transfert d'analyses syntaxiques de l'anglais vers le chinois Padó et Pitel (2007) ont traité le problème de l'annotation automatique de rôles sémantiques dans une langue ne disposant pas de lexique FrameNet [4], en s'intéressant au couple de langue (anglais-français).

Dans le domaine médical, plusieurs travaux ont attaqué le transfert de connaissances d'une langue à une autre (Névéol *et al.*, 2005; Deléger *et al.*, 2006). En particulier, Deléger *et al.* (2009) se sont intéressés à l'acquisition de nouvelles traductions de termes issues de trois terminologies différentes («MeSH», «SNOMED CT» et «the MedlinePlus Health Topics»). Ces auteurs se sont basés sur l'alignement des mots à partir d'un corpus parallèle anglais-français.

3 Reconnaissance d'entités médicales dans des textes anglais

Dans cette section, nous présentons la tâche de Reconnaissance des Entités Médicales (REM) (section 3.1). Ensuite, nous décrivons notre méthode à base de connaissances expertes pour la REM à partir de textes anglais (section 3.2). L'application de cette méthode sur des grands volumes de données hétérogènes a révélé certains problèmes liés à l'ambiguité de certains termes. Nous proposons dans la section 3.3 une solution pour pallier cette ambiguité : un filtre statistique utilisé en amont pour améliorer la précision des entités extraites. En effet, en vue de la projection vers un autre corpus à des fins de détection d'entités correctes dans la langue cible, il est fortement souhaitable que les entités du corpus source soient correctes : c'est ce que vise à obtenir le filtrage mis en place, qui privilégie la précision par rapport au rappel.

4. `http://framenet.icsi.berkeley.edu/fndrupal/`

3.1 Description de la tâche de reconnaissance d'entités médicales

La REM est la tâche de base de l'extraction d'information à partir de textes médicaux. Nous désignons par « entité médicale » une instance d'un concept médical ou une catégorie générique (par exemple, *l'Alzheimer* est une instance de la catégorie « Maladie », *la laryngoscopie* est une instance de « Examen »). Cette définition soulève deux questions : (i) quelle est la liste des catégories médicales traitées (Problème médical, Examen, Traitement, etc.) et (ii) quelle est la définition exacte de chaque catégorie (par exemple, les plantes peuvent-elles être considérées comme des traitements ?). Dans cet article nous travaillons sur les trois grandes catégories les plus importantes dans le domaine médical, à savoir : « Problème », « Traitement » et « Examen ». Nous utilisons les types sémantiques de l'UMLS [5] pour définir chaque catégorie (cf. la table 2), en suivant le guide d'annotation i2b2/VA 2010 (Uzuner *et al.*, 2011).

Catégorie	Types sémantiques de l'UMLS correspondants
Problème	Virus, Bacterium, Anatomical Abnormality, Congenital Abnormality, Acquired Abnormality, Sign or Symptom, Pathologic Function, Disease or Syndrome, Mental or Behavioral Dysfunction, Neoplastic Process, Cell or Molecular Dysfunction, Injury or Poisoning
Traitement	Medical Device, Drug Delivery Device, Clinical Drug, Steroid, Pharmacologic Substance, Antibiotic, Biomedical or Dental Material, Therapeutic or Preventive Procedure
Examen	Laboratory Procedure, Diagnostic Procedure.

TABLE 2 – Les catégories médicales traitées

La reconnaissance des entités médicales consiste en (i) le repérage des termes médicaux dans les textes (tels que *beta cell replacement*, *pyogenic liver abscess*, *infection of biliary system*, etc.) et (ii) l'identification de la catégorie sémantique des termes repérés (telles que *Maladie*, *Médicament*, *Examen*, etc.). L'exemple suivant illustre les résultats de la REM pour une phrase extraite d'un résumé MEDLINE. Les termes médicaux sont annotés par des étiquettes *<Treatment>* et *<Disease>*.

<Treatment> Adrenal-sparing surgery </Treatment> is safe and effective , and may become the treatment of choice in patients with <Disease> hereditary phaeochromocytoma </Disease>. [PMID : 10027369]

Ces deux étapes amènent à effectuer des choix dans les catégories médicales à traiter, mais également dans les règles de délimitation des frontières des entités médicales dans le texte. Dans ce travail, nous avons effectué les choix suivants pour la délimitation des frontières : (i) inclure les possessifs, adjectifs, adverbes et chiffres dans les entités nommées (ii) annoter les abréviations séparément et (iii) ne pas annoter une entité médicale incluse dans une autre.

3.2 La méthode MetaMapPlus pour la reconnaissance d'entités médicales

Le domaine médical dispose de grandes bases terminologiques telles que l'UMLS (McCray et Nelson, 1995) ainsi que d'outils qui permettent de détecter les termes médicaux. Un des outils

5. L'UMLS (Unified Medical Language System) comporte (i) le Specialist Lexicon, lexique anglais incluant les termes du domaine ainsi que leurs variations syntaxiques et morphologiques, (ii) le Metathesaurus, vocabulaire de plus de deux millions de concepts (un concept « regroupe » des termes synonymes, acronymes et variantes terminologiques) et (iii) le réseau sémantique qui organise les concepts en 135 « types sémantiques » et définit 54 relations entre ces types.

les plus largement utilisés est MetaMap (Aronson, 2001), un système à base de connaissances qui se fonde sur l'UMLS. MetaMap permet de segmenter les textes médicaux en phrases et syntagmes nominaux qui correspondent à des termes médicaux. L'outil identifie les entités médicales et leurs catégories (concepts et types sémantiques du réseau sémantique UMLS). Cependant l'étude de l'utilisation simple de MetaMap a révélé qu'il présente certains problèmes. Afin d'améliorer la précision des résultats de MetaMap, nous avons proposé la méthode MetaMapPlus (Ben Abacha et Zweigenbaum, 2011a), qui comporte les quatre étapes suivantes :

– Extraire les syntagmes nominaux à l'aide d'un segmenteur (*chunker*). Nous utilisons TreeTagger-chunker qui offre une bonne segmentation et permet de diminuer le bruit de la REM (voir (Ben Abacha et Zweigenbaum, 2011c) pour une comparaison de trois segmenteurs).
– Filtrer les syntagmes candidats avec une liste de *mots vides* en amont de MetaMap.
– Rechercher les termes candidats dans des listes d'entités médicales construites à partir du Web.
– Pour le reste des termes candidats, déterminer leurs catégories avec MetaMap, après un filtrage par une liste des erreurs les plus fréquentes de MetaMap et en contraignant les types sémantiques utilisés par ce dernier.

Les résultats de MetaMapPlus, mesurés sur le corpus i2b2 (rappel de 48,68 %, précision de 56,46 % et F-mesure de 52,28 %), sont significativement meilleurs que ceux de MetaMap (F-mesure de 15,80 %) mais restent limités à cause de la performance du procédé de segmentation. L'approche a cependant permis d'identifier le type correct pour 52,28 % des entités, sachant que seules 60,76 % des entités ont été extraites correctement par le segmenteur (i.e. avec des frontières correctes).

En appliquant la méthode MetaMapPlus sur un grand corpus médical extrait du web, nous avons constaté que des ambiguïtés lexicales apparaissaient plus souvent. En effet, plusieurs termes généraux sont considérés par MetaMap comme des entités médicales. Cette ambiguïté peut être divisée en deux catégories principales : (i) les homonymes (e.g. ten, qui désigne dix en domaine ouvert et la maladie « Toxic Epidermal Necrolysis » en domaine médical) et (ii) les termes généraux ayant un sens qui se spécialise dans le domaine médical (e.g. case, form). Ces ambiguïtés causent du bruit dans la reconnaissance d'entités médicales.

3.3 Traitement des ambiguïtés entre acception générale et spécialisée

Pour résoudre le problème d'ambiguïté lexicale, nous proposons une étape supplémentaire intégrée à la méthode MetaMapPlus. Cette étape (appelée Maxent_SNG) consiste à utiliser un classifieur pour distinguer les termes médicaux et les termes généraux, avant d'appliquer MetaMap. Le but de ce module est de :

1. Réduire le bruit lié à l'ambiguïté lexicale, en éliminant les syntagmes nominaux (SN) « généraux » fréquents en domaine ouvert même lorsqu'ils sont utilisés dans le domaine médical (par exemple « table »).
2. Réduire le volume à traiter par la catégorisation via MetaMap en éliminant une bonne partie des syntagmes nominaux à classifier, ce qui devrait réduire le temps d'exécution.

Les méthodes statistiques à base d'apprentissage supervisé peuvent être très robustes. Cependant, ces méthodes présentent deux inconvénients importants :

1. La dépendance aux données annotées disponibles (cf. (Ben Abacha et Zweigenbaum,

2011c)), ce qui constitue un obstacle à l'utilisation de ce type de méthodes pour des tâches et domaines pour lesquels on ne dispose pas de corpus annotés, considérant en outre que la constitution de ces corpus est une tâche coûteuse.

2. Le problème de portabilité sur des corpus différents de ceux utilisés en entraînement (cf. (Ben Abacha et Zweigenbaum, 2011c)), la dégradation des performances une fois appliquées sur des corpus ayant des caractéristiques différentes de ceux utilisés pour l'entraînement constitue un obstacle pour le passage à l'échelle de ces méthodes.

Ces deux inconvénients constituent un important défi pour la mise en place d'une méthode statistique efficace et portable. Pour différencier les données d'entraînement (ce qui offrira une meilleur *adaptabilité*) et éviter le sur-apprentissage (en apprenant correctement et non pas « par coeur »), nous traitons deux problèmes : (i) comment choisir les exemples d'entraînement ? (ii) et quels sont les attributs à utiliser ?

3.3.1 Sélection des données d'apprentissage

À l'instar des travaux sur l'apprentissage actif (Active Learning) (Thompson *et al.*, 1999; Tomanek et Olsson, 2009) qui sélectionnent des exemples diversifiés et représentatifs à annoter manuellement, nous avons trouvé utile de sélectionner les exemples à utiliser pour « bien » apprendre. Deux questions clés se posent alors :
– le nombre des exemples *positifs* et *négatifs* à utiliser ;
– le choix de ces exemples qui doivent être *représentatifs*.
Pour choisir ces exemples, nous proposons d'utiliser :

1. la fréquence des mots/syntagmes nominaux (positifs et négatifs) dans un même corpus ;
2. la présence des mots/syntagmes nominaux (positifs et négatifs) dans des corpus textuels médicaux de genres différents ;
3. le Web pour collecter des données (des exemples positifs et négatifs).

Plus précisément, pour la construction des données d'apprentissage pour le module qui permet de classifier les syntagmes nominaux (SN) en entités médicales (EM) ou termes généraux (SNG), nous utilisons les exemples positifs et négatifs suivants :

1. Exemples positifs : entités médicales
   – les EM les plus fréquentes dans le corpus i2b2 de textes cliniques ;
   – les EM les plus fréquentes dans le corpus Berkeley d'articles scientifiques (Rosario et Hearst, 2004) ;
   – les EM communes aux deux corpus ;
   – des EM extraites du Web (notamment de Wikipedia [6], HON [7]) ;
2. Exemples négatifs : SN « généraux » (SNG) qui ne correspondent pas à des entités médicales :
   – les SNG les plus fréquents dans le corpus i2b2 ;
   – les SNG les plus fréquents dans le corpus de Berkeley ;
   – les SNG les plus fréquents qui existent dans ces deux corpus ;

6. Différentes listes d'entités médicales ont été extraites à partir de Wikipedia : medical tests, diseases, disorders, treatments, procedures (diagnostiques, therapeutiques, surgicales,..)
7. HON (Health On the Net) : `http://www.hon.ch/`

– des SNG extraits du Web, à partir de sites thématiquement distant du domaine médical. Nous avons choisi des sites d'histoires pour enfants [8] [9]). Notre motivation est d'utiliser des corpus ne contenant pas ou peu d'entités médicales.

La table 3 décrit les types d'exemples positifs et négatifs que nous avons utilisés, selon trois critères : corpus, nombre d'exemples et nombre d'occurrences de chaque exemple.

	Corpus	Nb d'exemples	Fréquence des exemples
Exemples positifs	extraits du Web (Wikipedia, HON, etc.)	1 114	entre 1 et 3
	corpus médical 1 (i2b2 : textes cliniques)	3 974 (sur 26 187 EM)	>= 3 (allant jusqu'à 347 pour « hypertension »)
	corpus médical 2 (Berkeley : articles scientifiques)	391 (sur 2 463 EM)	>=2 (allant jusqu'à 28 pour « chemotherapy »)
	Total	**5 479** entités médicales	
Exemples négatifs	extraits du Web (sites d'histoires pour enfants)	2 127	1 et 2
	corpus médical 1 (i2b2 : textes cliniques)	2 031 (sur 15 882 SN)	>= 3 (allant jusqu'à 855 pour « the patient »)
	corpus médical 2 (Berkeley : articles scientifiques)	1 639 (sur 10 464 SN)	>= 2 (allant jusqu'à 278 pour « patients »)
	Total	**5 797** syntagmes nominaux (généraux)	

TABLE 3 – Classification des syntagmes nominaux en termes médicaux et termes généraux, et sélection des exemples positifs et négatifs selon trois critères : corpus, nombre d'exemples et nombre d'occurrences de chaque exemple.

3.3.2 Attributs utilisés par le classifieur

Pour cette tâche, nous utilisons un classifieur à maximum d'entropie [10]. Pour chaque syntagme nominal (médical ou général), les attributs utilisés par le classifieur sont :
– la longueur du SN, son nombre de tokens ;
– le SN est un mot en majuscules / le SN est en majuscules / le SN contient un mot en majuscules ;
– les mots / lemmes / catégories syntaxiques du SN ;
– la présence et la fréquence des mots du SN dans la liste des mots du corpus général BNC [11] ;
– la présence des mots du SN dans un dictionnaire général (nous avons utilisé le dictionnaire standard du système d'exploitation Linux).

8. `http://www.goodnightstories.com/read/pnkbook1.htm`
9. `http://www.vtaide.com/png/stories.htm`
10. Nous avons utilisé l'implémentation disponible à : `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`.
11. British National Corpus, `http://www.natcorp.ox.ac.uk/`.

4 Projection des annotations sur des textes en français par alignement

4.1 Alignement au niveau des mots

La projection que nous réalisons se fonde sur des alignements calculés au niveau des mots. Pour les obtenir, nous avons utilisé les programmes d'alignement de corpus parallèles du système de traduction statistique MOSES (Koehn *et al.*, 2007), en utilisant le paramétrage par défaut. Celui-ci utilise l'outil GIZA++ (Och et Ney, 2004), qui calcule des modèles statistiques d'alignement de mots de complexité croissante. L'alignement est réalisé dans les deux directions puis ses résultats sont *symétrisés*. Finalement, des *tables de traduction*, qui regroupent l'ensemble des bi-segments pouvant être extraits du corpus, sont construites par application d'heuristiques d'extraction de bi-segments *cohérents*, qui imposent que tout mot d'un segment dans un langue doit être aligné avec au moins un mot du segment dans l'autre langue, mais avec aucun mot en dehors de celui-ci.

4.2 Projection

La figure 2 présente quelques exemples de bi-phrases alignées au niveau des mots.

FIGURE 2 – Exemples : trois bi-phrases alignées au niveau des mots

Pour projeter les annotations, nous utilisons le principe suivant :

Soit $E_1 = \{m_{11}, \ldots, m_{1n}\}$ l'ensemble des mots constituant une entité médicale dans le corpus anglais et $E_2 = \{m_{21}, \ldots, m_{2p}\}$ l'ensemble des mots constituant la projection de E_1 (i.e. l'union des projections de chaque mot dans E_1). En notant $position(m_i)$ la fonction retournant la position d'un mot dans sa phrase, nous considérons que la projection de l'entité médicale anglaise est la séquence ordonnée de mots $E_3 = \{m_{31}, \ldots, m_{3k}\}$ telle que :
– $position(m_{31}) = Min_{m_{2i} \in E_2}(position(m_{2i}))$
– $position(m_{3k}) = Max_{m_{2i} \in E_2}(position(m_{2i}))$

Le bruit produit par l'alignement et les annotations affecte la qualité des annotations projetées. Pour diminuer ce bruit et améliorer la phase de projection, (i) nous définissons des heuristiques (telles que la longueur de l'entité trouvée en français par rapport à l'entité originale en anglais) et (ii) nous utilisons un *antidictionnaire* [12] pour filtrer les entités médicales obtenues et supprimer les « mots vides ».

12. `http://members.unine.ch/jacques.savoy/clef/index.html`

5 Expérimentations et évaluation

Le corpus utilisé pour les expérimentations a été construit à partir du site bilingue « Santé Canada [13] » aligné au niveau des phrases (Deléger *et al.*, 2009). La table 4 présente le corpus parallèle (anglais-français) Santé Canada.

	Corpus anglais	**Corpus français**
Nombre de lignes	395600	395600
Nombre de mots	4 465 672	5 052 543
Nombre de caractères	29 845 733	33 901 471
Nombre de mots par ligne (en moyenne)	11	13
Nombre de caractères par mot (en moyenne)	7	7

TABLE 4 – Le corpus parallèle Santé Canada

5.1 Construction et annotation manuelle d'un bi-corpus de référence

Pour évaluer notre approche, nous avons besoin d'un bi-corpus de référence annoté. Deux éléments sont à déterminer : (i) la taille du corpus de référence et (ii) la manière de choisir ce corpus à partir du corpus initial Santé Canada. Nous nous sommes pour cela basés sur des travaux en statistiques.

Taille du corpus. Pour déterminer la taille (acceptable) du corpus de référence à sélectionner, nous utilisons la formule utilisée en statistiques (Sim et Wright, 2005) pour déterminer la taille d'un échantillon :

$$N = \frac{T^2 P(1-P)}{E^2}$$

$$\begin{cases} N & = & \text{La taille de l'échantillon attendu.} \\ T & = & \text{Niveau de confiance déduit du taux de confiance} \\ & & \text{(traditionnellement 1,96 pour un taux de confiance de 95 \%).} \\ P & = & \text{Proportion estimée de la « population » présentant la caractéristique} \\ & & \text{étudiée dans l'étude. Lorsque cette proportion est ignorée, une pré-} \\ & & \text{étude peut être réalisée ou sinon p = 0,5 sera retenue.} \\ E & = & \text{Marge d'erreur (traditionnellement fixée à 5 \%).} \end{cases}$$

Nous fixons les valeurs suivantes : $P = 0{,}5$, $T = 1{,}96$ et $E = 0{,}05$, ce qui nous donne la valeur : $\mathbf{N = 385}$.

Sélection du corpus. Différentes méthodes sont possibles, telles que l'échantillonnage aléatoire simple (*simple random sampling*) ou l'échantillonnage stratifié (*stratified sampling*). Nous avons choisi d'utiliser l'échantillonnage aléatoire simple et sélectionné aléatoirement 385 phrases, contenant 4 613 mots.

Annotation manuelle du corpus de référence. Nous avons annoté manuellement les 385 phrases sélectionnées avec trois types de catégories médicales : Traitement, Problème et Maladie. Nous avons annoté les deux parties françaises et anglaises du corpus de référence en utilisant le guide d'annotation de i2b2 2010 (tâche 1), en respectant les règles de délimitation des frontières décrites dans la section 3.1.

13. `http://www.hc-sc.gc.ca`

5.2 Évaluation de l'annotation du corpus anglais

Dans cette section, nous évaluons la REM à partir du corpus anglais. Nous différencions le cas où les entités médicales ont été reconnues avec des frontières précises ou exactes et le cas où les frontières ne sont pas précises (par exemple, «as antimicrobial resistance» au lieu de «antimicrobial resistance», «Pap smear» au lieu de «a Pap smear» ou «the Pap smear» dans le texte). Nous utilisons les mesures standard de rappel, de précision et de F-mesure.

Nous avons entraîné le module Maxent_SNG de classification des syntagmes nominaux en entités médicales et entités générales sur le corpus d'entraînement i2b2 et nous l'avons testé sur le corpus de test i2b2. Nous avons obtenu une correction (proportion d'exemple de test correctement classés) de 90,99 % (16 169/17 769). La table 5 présente la contribution à la méthode MetaMapPlus de ce module (Maxent_SNG), entraîné sur le corpus d'entraînement décrit dans la section 3.3.

MetaMapPlus						**Maxent_SNG + MetaMapPlus**					
Frontières strictes			Frontières larges			Frontières strictes			Frontières larges		
R	P	F	R	P	F	R	P	F	R	P	F
61,36	22,37	32,79	82,26	28,18	41,97	50,00	40,23	**44,58**	59,26	45,98	**51,78**

TABLE 5 – Résultats de la méthode MetaMapPlus sans et avec le module Maxent_SNG sur le corpus Santé Canada (partie anglaise).

Comme attendu, ce filtrage améliore sensiblement la précision des entités médicales détectées, et en dépit d'une baisse importante de la valeur de rappel, la F-mesure connaît une nette augmentation.

5.3 Évaluation de l'annotation du corpus français par projection

Dans cette section, nous évaluons la qualité de la projection des entités médicales extraites du corpus anglais par notre méthode (i.e. Maxent_SNG+MetaMapPlus). Dans un premier temps, nous évaluons uniquement la qualité de la projection (indépendamment des erreurs d'extraction). Pour ce faire, nous étudions la qualité de la projection des entités de référence (i.e. annotées manuellement). Dans un second temps, nous évaluons l'ensemble du processus pour la REM en français (comprenant l'extraction automatique des entités médicales en anglais et leur projection). Le tableau 6 présente les résultats de la projection des entités de référence et les résultats de la projection des entités extraites avec la méthode Maxent_SNG+MetaMapPlus.

5.4 Discussion

Nous avons pu améliorer les résultats de la méthode MetaMapPlus en intégrant le module MaxEnt entraîné sur trois types différents de corpus. Notons que les résultats obtenus en exploitant ces trois types de corpus sont meilleurs que ceux obtenus en entraînant le classifieur sur un ou deux corpus uniquement (ce que nous avons testé mais ne pouvons pas détailler dans cet

Annotation du corpus français : projection des entités médicales extraites						**Annotation du corpus français : projection des entités médicales de référence**					
Frontières strictes			Frontières larges			Frontières strictes			Frontières larges		
R	P	F	R	P	F	R	P	F	R	P	F
22,39	22,90	22,64	43,08	42,75	42,91	44,78	57,69	50,42	67,91	87,50	76,47

TABLE 6 – Évaluation de la projection des entités médicales extraites avec la méthode Maxent_SNG+MetaMapPlus et les entités de référence sur le corpus Santé Canada (partie française).

article). Les résultats sont relativement acceptables (51,78 % de F-mesure) étant donné la complexité de la tâche sur un corpus hétérogène extrait du Web.

Pour la projection, nous avons utilisé les alignements au niveau des mots avec une approche simple qui consiste à prendre l'entité correspondante (projetée) la plus large. Nous avons essayé d'améliorer cette projection en utilisant un antidictionnaire pour filtrer les entités obtenues et quelques heuristiques telles qu'une différence maximale entre la longueur de l'entité initiale et celle de l'entité projetée (cf. table 7).

	Frontières larges	**Frontières strictes**
Projection sans filtrage	79,09 %	**47,15 %**
Projection + antidictionnaire	75,52 %	**49,79 %**
Projection + antidictionnaire + heuristiques	76,47 %	**50,42 %**

TABLE 7 – F-mesure de la projection des entités de référence sans et avec filtrage

Les résultats de la projection des entités médicales extraites sont relativement faibles, mais ceci dépend directement de la performance de la méthode d'extraction d'information (51,78 % de F-mesure, avec frontières larges), qui fixe le plafond de performance atteignable en projetant ses résultats sur le corpus français. Par projection, nous perdons tout de même près de 50 % des extractions correctes dans le corpus anglais. Ceci résulte principalement de la qualité des alignements au niveau des phrases puis des mots. L'alignement au niveau des mots est influencé par la qualité de l'alignement des phrases (cf. les exemples 1 et 2 ci-dessous). En effet, dans certains cas, la phrase correspondante en français n'est pas équivalente à celle en anglais (soit elle est beaucoup plus courte et contient moins d'information, soit elle est beaucoup plus longue), dans d'autres cas elle a un contenu complètement différent ou reste formulée en anglais : cela reflète un problème d'alignement de phrases qu'il nous faudra corriger dans la suite de nos travaux.

Exemple 1 :
- Statement on Immunization for <PB>Lyme Disease</PB>, 2000 (*)
- 0-0 1-1 5-2 4-3 5-3 5-4 5-5 2-7 5-9 5-10 5-11 5-12 6-13 7-14
- Déclaration sur <PB>un schéma révisé pour la vaccination des adolescents contre l'hépatite B</PB>, 2000 (*)

Exemple 2 :
- <PB>Lung Cancer</PB> : Guidelines for processing Specimens and Reporting Tumor Stage (2000)
- 0-0 1-0 2-0 1-1 1-2 3-3 4-6 9-13 8-18 6-23 7-24 10-27
- <PB>Utilisation, aux fins </PB> de la surveillance, des renseignements sur les patients atteints de cancer : Examen systématique des lois, des réglements, des politiques et des lignes directrices (2000)

Il semble que ces deux phrases ne soient pas en relation de traduction, qui peut résulter d'un mauvais appariemment de documents ou entre phrases.

6 Conclusion et perspectives

Nous avons proposé dans cet article une approche pour l'annotation automatique de textes médicaux en français par projection depuis l'anglais, et présenté nos premières expérimentations en REM. L'approche présentée utilise un corpus parallèle aligné au niveau des mots pour projeter les annotations obtenues sur la partie anglaise vers la partie française. L'application de notre méthode de REM sur un grand corpus de données hétérogènes extrait du Web a posé une problématique de passage à l'échelle pour laquelle nous avons proposé une solution qui consiste à intégrer un module de filtrage statistique en amont des entités candidates pour améliorer la précision des entités extraites.

Nous envisageons principalement quatre perspectives à ce travail :
– L'annotation automatique des relations sémantiques dans des textes français en reprenant la méthode présentée. Nous avons déjà développé des méthodes à base de patrons et des méthodes statistiques pour l'extraction de relations sémantiques à partir de textes médicaux en anglais (Ben Abacha et Zweigenbaum, 2011b).
– L'utilisation ou la construction d'autres corpus médicaux parallèles de meilleur qualité.
– L'exploitation de corpus français annotés pour la mise en place de méthodes statistiques pour l'extraction d'information à partir de textes en français.
– L'intégration de ces méthodes d'extraction d'information dans un système de questions-réponses translingue.

Références

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS metathesaurus : the MetaMap program. *In AMIA Annu Symp Proc*, pages 17–21.

Ayache, C. (2005). Campagne EVALDA/EQueR – Évaluation en question-réponse, rapport final. Rapport technique, ELDA, Paris. Available at`http://www.technolangue.net/IMG/pdf/rapport_EQUER_1.2.pdf`.

Bach, N. et Badaskar, S. (2007). A Review of Relation Extraction.

Ben Abacha, A. et Zweigenbaum, P. (2011a). Automatic extraction of semantic relations between medical entities : a rule based approach. *Journal of Biomedical Semantics*, 2(Suppl 5):S4.

Ben Abacha, A. et Zweigenbaum, P. (2011b). A hybrid approach for the extraction of semantic relations from MEDLINE abstracts. *In Computational Linguistics and Intelligent Text Processing, 12th International Conference, CICLing 2011*, volume 6608 de *Lecture Notes in Computer Science*, pages 139–150, Tokyo, Japan.

Ben Abacha, A. et Zweigenbaum, P. (2011c). Medical entity recognition : A comparison of semantic and statistical methods. *In Actes BioNLP 2011 Workshop*, pages 56–64, Portland, Oregon, USA. Association for Computational Linguistics.

Deléger, L., Merkel, M. et Zweigenbaum, P. (2006). Contribution to terminology internationalization by word alignment in parallel corpora. *In AMIA Annu Symp Proc.*, pages 185–189, Washington, DC.

Deléger, L., Merkel, M. et Zweigenbaum, P. (2009). Translating medical terminologies through word alignment in parallel text corpora. *Journal of Biomedical Informatics*, 42(4):692–701. Epub 2009 Mar 9.

GUERGANA K SAVOVA, James J Masanz, P. V. O. J. Z. S. S. K. C. K.-S. et CHUTE, C. G. (2010). Mayo clinical text analysis and knowledge extraction system (ctakes) : architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17:507–513.

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A. et HERBST, E. (2007). Moses : Open source toolkit for statistical machine translation. *In Proceedings of ACL*, Czech Republic.

LOPEZ, A., NOSSAL, M., HWA, R. et RESNIK, P. (2002). Word-level alignment for multilingual resource acquisition. *In Actes LREC Workshop on Linguistic Knowledge Acquisition and Representation : Bootstrapping Annotated Data*, Las Palmas, Spain. ELRA.

LÜ, Y., LI, S., ZHAO, T. et YANG, M. (2002). Learning Chinese bracketing knowledge based on a bilingual language model. *In Proceedings of COLING-2002*, pages 591–598.

MCCRAY, A. T. et NELSON, S. J. (1995). The semantics of the UMLS knowledge sources. *Methods of Information in Medicine*, 34(1/2).

NADEAU, D. et SEKINE, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26. Publisher : John Benjamins Publishing Company.

NÉVÉOL, A., MORK, J., ARONSON, A. et DARMONI, S. (2005). Evaluation of French and English MeSH indexing systems with a parallel corpus. *In AMIA Annu Symp Proc.*, pages 565–9, Washington, DC.

OCH, F. J. et NEY, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4).

PADÓ, S. et PITEL, G. (2007). Annotation précise du français en sémantique de rôles par projection cross-linguistique. *In Actes TALN 2007, Toulouse, France*.

POIBEAU, T. (2003). *Extraction automatique d'information : du texte brut au web sémantique*. Hermès science publications.

ROSARIO, B. et HEARST, M. A. (2004). Classifying semantic relations in bioscience text. *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pages 430–437, Barcelona.

SIM, J. et WRIGHT, C. C. (2005). The kappa statistic in reliability studies : Use, interpretation, and sample size requirements. *Physical Therapy*.

THOMPSON, C. A., CALIFF, M. E. et MOONEY, R. J. (1999). Active learning for natural language parsing and information extraction. *In Proceedings of the Sixteenth International Conference on Machine Learning (ICML-99)*, pages 406–414, Bled, Slovenia.

TOMANEK, K. et OLSSON, F. (2009). A web survey on the use of active learning to support annotation of text data. *In Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 45–48, Boulder, Colorado. Association for Computational Linguistics.

UZUNER, O., SOUTH, B. R., SHEN, S. et DUVALL, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*. [Epub ahead of print].

YAROWSKY, D. et NGAI, G. (2001). Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. *In Actes NAACL 2001*.

Une méthode d'extraction d'information fondée sur les graphes pour le remplissage de formulaires

Ludovic Jean-Louis Romaric Besançon Olivier Ferret
CEA, LIST, Laboratoire Vision et Ingénierie des Contenus
F-91191 Gif-sur-Yvette, France
{ludovic.jean-louis,romaric.besancon,olivier.ferret}@cea.fr

RÉSUMÉ

Dans les systèmes d'extraction d'information sur des événements, une tâche importante est le remplissage automatique de formulaires regroupant les informations sur un événement donné à partir d'un texte non structuré. Ce remplissage de formulaire peut s'avérer difficile lorsque l'information est dispersée dans tout le texte et mélangée à des éléments d'information liés à un autre événement similaire. Nous proposons dans cet article une approche en deux étapes pour ce problème : d'abord une segmentation du texte en événements pour sélectionner les phrases relatives au même événement ; puis une méthode de sélection dans les phrases sélectionnées des entités liées à l'événement. Une évaluation de cette approche sur un corpus annoté de dépêches dans le domaine des événements sismiques montre un F-score de 72% pour la tâche de remplissage de formulaires.

ABSTRACT

A Graph-Based Method for Template Filling in Information Extraction

In event-based Information Extraction systems, a major task is the automated filling from unstructured texts of a template gathering information related to a particular event. Such template filling may be a hard task when the information is scattered throughout the text and mixed with similar pieces of information relative to a different event. We propose in this paper a two-step approach for template filling : first, an event-based segmentation is performed to select the parts of the text related to the target event ; then, a graph-based method is applied to choose the most relevant entities in these parts for characterizing the event. Using an evaluation of this model based on an annotated corpus for earthquake events, we achieve a 72% F-measure for the template-filling task.

MOTS-CLÉS : Extraction d'information, segmentation de texte, remplissage de formulaires.

KEYWORDS: Information Extraction, Text Segmentation, Template Filling.

1 Introduction

Le domaine de l'Extraction d'Information couvre toutes les tâches consistant à extraire des informations structurées à partir de textes. Une tâche archétypique de ce domaine est celle définie dans les conférences MUC (*Message Understanding Conferences*) (Grishman et Sundheim, 1996), où les systèmes doivent permettre de remplir de façon automatique des formulaires (ou

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 29–42,
Grenoble, 4 au 8 juin 2012.

templates) concernant des événements. Ces formulaires permettent de mettre en évidence une information spécifique à un type d'événement considéré et d'ignorer tout autre type d'information non pertinente. La figure 1 donne un exemple du remplissage d'un formulaire à partir du texte d'une dépêche de presse.

Texte	***Templates***
EV1Un **séisme** de magnitude **7,2** sur l'échelle de Richter a frappé **samedi** la ville de **Kurihara** (**préfecture de Miyagi**). EV1Le **tremblement** s'est produit à **08H43**, heure locale. EV1La **secousse** a été ressentie jusqu'à **Tokyo**, à 500 kilomètres au sud des préfectures japonaises d'**Iwate** et de **Miyagi**, principales zones touchées. Les **séismes** sont courants au **Japon**, qui est l'une des zones sismiques les plus actives de la planète. EV2En **octobre 2004**, un **séisme** d'une magnitude de **6,8** avait touché la région de **Niigata**, dans le nord du pays.	*EV1* • **ÉVÉNEMENT** : séisme, tremblement, secousse • **DATE** : samedi • **HEURE** : 08h43 • **MAGNITUDE** : 7,2 • **LIEU** : Kurihara *EV2* • **ÉVÉNEMENT** : séisme • **DATE** : octobre 2004 • **HEURE** : / • **MAGNITUDE** : 6,8 • **LIEU** : Niigata

FIG. 1 – Exemple de remplissage de formulaire

Les problèmes soulevés par la réalisation d'un système d'extraction d'information pour le remplissage de formulaire comptent en particulier l'identification des entités nommées ou autres entités spécifiques du domaine, l'établissement des relations entre ces entités, la résolution de la coréférence concernant les entités, le regroupement d'informations dispersées dans le texte, etc. (Turmo *et al.*, 2006).

Il n'existe pas actuellement d'approche considérée comme standard pour le remplissage de formulaire. Néanmoins, la plupart des systèmes d'extraction d'information adoptent une approche en deux temps : des patrons spécifiques au domaine ou des classifieurs sont d'abord utilisés pour extraire au niveau phrastique les informations constitutives du formulaire considéré (dates, lieux, magnitudes et heures dans le cas de la figure 1) en s'appuyant sur les mentions d'événements ; des heuristiques relatives au type d'événement ou de texte considéré sont ensuite appliquées pour fusionner les informations extraites dans des formulaires globaux. Même si ce type d'approche est largement utilisé, elle se heurte à deux problèmes importants : une vision très locale de l'extraction des informations élémentaires et une prise en compte limitée et peu générique des dépendances entre ces informations, en particulier pour le remplissage des formulaires.

La figure 1 illustre clairement le fait que les informations relatives à un événement, ici *EV1*, peuvent être exprimées au-delà de la portée de la phrase. Ce problème pose plus généralement la question de la délimitation des parties de texte relatives à un événement ou un type d'événements donné car les informations d'un événement ne sont pas toujours liées à une mention d'événement proche. Notre approche pour résoudre ce problème s'appuie sur une segmentation discursive des textes sur la base des événements auxquels chaque phrase fait référence. Plus largement, son objectif est de diminuer l'espace textuel à explorer pour faire le lien entre une entité et une mention d'événement et donc *in fine*, pour le remplissage du formulaire associé à un événement donné. Les notions de temps et d'événement étant fortement liées, cette segmentation s'appuie

sur des indices de nature temporelle.

Le second problème évoqué ci-dessus a déjà fait l'objet de quelques travaux assimilant les formulaires à des relations complexes. Dans ce contexte, chaque événement est vu comme une relation *n*-aire dont l'arité est égale au nombre de champs à remplir dans le formulaire (n=5 dans l'exemple précédent). Cette vision a d'abord été appliquée au niveau local pour des phrases contenant plusieurs entités d'intérêt pour le même événement (la première phrase de la figure 1 en contient par exemple quatre) : dans (McDonald *et al.*, 2005), les relations entre la mention de cet événement et les informations qui lui sont liées ne sont ainsi plus considérées indépendamment les unes des autres mais de façon plus globale. Au-delà, plusieurs méthodes ont été proposées pour extraire des relations complexes, parmi lesquelles se distinguent des méthodes à base de graphe (McDonald *et al.*, 2005; Wick *et al.*, 2006) et des méthodes à base d'inférences (Goertzel *et al.*, 2006). Dans cet article, nous présentons une méthode à base de graphe, en commençant par construire un graphe d'entités fondé sur le résultat de la segmentation et en utilisant plusieurs stratégies génériques (*i.e.* indépendantes du domaine considéré) pour la construction de la relation complexe à partir de ce graphe.

2 Motivation et état de l'art

Le remplissage de formulaire est une tâche centrale des systèmes d'extraction d'information et a fait l'objet en tant que telle de nombreuses études. Ainsi, dans le contexte des campagnes d'évaluation MUC (*Message Understanding Conferences*) et ACE (*Automatic Content Extraction*) (Doddington *et al.*, 2004), un des objectifs des systèmes participants était de remplir automatiquement des formulaires prédéfinis avec une structure fixe. Bien que ce soit l'approche la plus répandue, d'autres travaux, comme (Chambers et Jurafsky, 2011), adoptent un point de vue différent et proposent une approche non supervisée pour remplir des formulaires sans connaissance *a priori* sur leur structure. Ils exploitent dans ce cas des techniques de regroupement (*clustering*) pour apprendre la structure des formulaires et des patrons syntaxiques pour en remplir les champs.

Une grande partie des systèmes d'extraction d'information, en particulier ceux fondés sur des approches à base d'apprentissage automatique, s'appuient sur l'idée qu'un événement est souvent décrit dans une seule phrase, ce qui conduit à donner une importance moindre à l'information inter-phrastique. Cette idée est nommée « hypothèse de la phrase seule » (*single sentence assumption*) par (Stevenson, 2006), qui rapporte que seulement 60% des faits mentionnés dans les corpus MUC (MUC 4-6-7) peuvent être identifiés avec cette hypothèse. Ce pourcentage a été confirmé plus récemment par (Ji *et al.*, 2010), montrant qu'environ 40% des relations entre entités nécessitent l'usage de techniques d'inférences inter-phrastiques pour les extraire.

Peu d'approches ont été proposées pour faire de l'extraction d'information à un niveau discursif sans lien étroit avec le domaine abordé. Parmi elles, (Gu et Cercone, 2006) et (Patwardhan et Riloff, 2007) sont les plus proches de l'approche présentée ici. (Gu et Cercone, 2006) définit une approche à base de modèles de Markov cachés, d'une part pour identifier les unités de textes (phrases) pertinentes pour le remplissage de formulaire, et d'autre part pour faire l'extraction des entités dans les phrases retenues. De façon similaire, (Patwardhan et Riloff, 2007) propose tout d'abord d'identifier les phrases pertinentes en utilisant un modèle SVM (*Support Vector Machine*), puis d'appliquer différents niveaux de patrons d'extraction pour remplir les champs du formulaire.

Une des premières approches pour l'extraction de relations *n*-aires vient du domaine biomédical (McDonald *et al.*, 2005) et a ensuite été appliquée dans le domaine des mouvements de personnel dans les entreprises (Afzal, 2009). D'autres travaux s'attaquent au problème des relations complexes dans le contexte de l'extraction de champs pour les bases de données (*database record extraction*), en s'intéressant plus particulièrement à la compatibilité d'un ensemble d'entités données plutôt que d'une paire d'entités, ce qui les amène à prendre en compte des relations inter-phrastiques entre entités (Wick *et al.*, 2006; Mansuri et Sarawagi, 2006; Feng *et al.*, 2007).

3 Description de l'approche

Le cadre applicatif de la méthode d'extraction d'événements présentée dans cet article se situe dans un contexte de veille, dans lequel les utilisateurs ne sont en général intéressés que par les événements les plus récents. Dans ce contexte, notre but est de synthétiser, à partir de dépêches de presse, les informations relatives aux événements récents dans un tableau de bord. Néanmoins, les articles font en général référence à plusieurs événements comparables, en général pour mettre en évidence les similarités ou les différences entre l'événement récent et des événements passés de même nature. Dans notre application spécifique de veille, nous ne nous intéressons pas aux événements passés, que nous considérons comme une source de bruit pour la détection des informations relatives à l'événement principal de l'article. Nous avons donc fait l'hypothèse, comme (Feng *et al.*, 2007), qu'un document est associé à un seul formulaire. Nous utilisons une stratégie en deux étapes pour extraire cette information (Jean-Louis *et al.*, 2011) :

– une segmentation du texte en événements : les informations relatives aux événements peuvent se trouver sur plusieurs phrases. Par conséquent, nous devons découper le texte en segments homogènes du point de vue événementiel. Ces segments regroupent fréquemment des phrases non-contiguës car la structure des articles fait souvent des aller-retours entre l'événement principal et un ou plusieurs événements passés ;
– le remplissage des formulaires : puisque les segments événementiels couvrent plus d'une phrase, la probabilité d'y trouver des relations complexes (impliquant un grand nombre d'entités) est plus forte que dans une seule phrase. Nous devons donc trouver dans ces segments quelles entités sont susceptibles d'être impliquées dans des relations complexes.

4 Segmentation événementielle des textes

L'idée de segmenter des textes en unités homogènes du point de vue événementiel a principalement été abordée selon deux angles : de façon assez liée à un domaine particulier dans des travaux comme (Gu et Cercone, 2006; Patwardhan et Riloff, 2007), avec des méthodes reposant sur des modèles très lexicalisés ; à l'inverse, en ne s'appuyant que sur la logique d'enchaînement des types d'événements dans (Naughton, 2007). Notre approche est intermédiaire : en exploitant des informations de nature temporelle, elle fait appel à des caractéristiques des textes dépassant leur simple appartenance à un domaine donné.

Du point de vue du processus de segmentation, un texte est vu comme une séquence de phrases, chaque phrase étant caractérisée par un statut événementiel. Comme dans la plupart des travaux similaires, nous faisons l'hypothèse, en pratique raisonnablement simplificatrice, qu'une phrase

possède un statut événementiel homogène. Nous distinguons plus précisément trois statuts. **Événement principal** : référence à l'événement principal du texte ; **Événement secondaire** : référence à un événement secondaire du texte, sans distinction de l'événement particulier s'il en existe plusieurs ; **Contexte** : sans référence à un événement.

Dans cette perspective, nous considérons la segmentation événementielle comme une tâche de classification visant à associer à chaque phrase d'un texte un statut événementiel. Néanmoins, une telle segmentation possède un caractère intrinsèquement discursif dans la mesure où les catégories événementielles ne s'enchaînent pas de manière arbitraire. Du point de vue de la classification des phrases, elles sont donc déterminées à la fois par les indices repérables au niveau phrastique mais également par les catégories et les indices des phrases précédentes. Notre approche se focalise ainsi sur la capture des relations entre les changements événementiels et les changements de cadre temporel, manifestées par exemple par le passage du passé composé vers plus-que-parfait accompagnant la transition de l'événement principal à un événement secondaire dans le texte de la figure 2.

Pour la classification des phrases, nous avons donc utilisé un modèle de séquences, en l'occurrence de type CRF linéaire (Champs Conditionnels Aléatoires (Lafferty *et al.*, 2001)), s'appuyant sur les traits de nature temporelle suivants. *Temps des verbes* : un trait binaire est associé à chaque temps possible et activé dès que la phrase contient au moins un verbe du temps correspondant ; *date* : la présence d'une date est souvent le signe de la présence d'un événement différent de celui de la phrase précédente ; *expression temporelle* : ce trait marque la présence d'une expression temporelle, telle que *ces dernières années, au début de l'année*, souvent associée au caractère général d'un propos. Par ailleurs, les dépendances de succession entre les différents statuts événementiels sont prises en compte par le caractère linéaire de notre modèle CRF. Ce modèle est plus amplement détaillé dans (Jean-Louis *et al.*, 2010).

5 Remplissage de formulaires événementiels

Pour le remplissage des formulaires liés aux événements, nous proposons une approche à base de graphe inspirée du paradigme de l'extraction de relations complexes. Sa première étape, dite de *construction du graphe*, détecte d'abord les relations existant entre des paires de mentions d'entités ou d'événements du domaine considéré cooccurrant dans une phrase. Il construit ensuite un graphe d'entités sur la base de la fusion des mentions d'événements et d'entités faisant référence à un même événement ou à une même entité. Sa seconde étape, dite de *remplissage du formulaire*, applique des stratégies génériques à ce graphe pour sélectionner les entités les plus à même de remplir le formulaire correspondant au type d'événement considéré. Ces deux étapes sont détaillées dans les deux sections suivantes.

5.1 Construction du graphe d'entités

Le graphe d'entités que nous construisons dans cette première étape caractérise à l'échelle du document la présence ou l'absence d'une relation entre chaque paire d'entités liées au type d'événements considéré (par exemple, la relation de localisation d'un événement sismique dans notre cas). Il s'agit d'un graphe pondéré dont les nœuds représentent des entités ou des événements et les arcs, les relations qui les unissent. Ces relations étant symétriques, ce graphe

est non dirigé. Le poids associé à chaque arc est un score de confiance prenant ses valeurs dans l'intervalle [0,1] et évaluant le degré de certitude de la présence d'une relation entre les deux entités liées. La figure 2 donne l'exemple d'un tel graphe restreint aux entités liées à l'événement principal du document, compte tenu de notre focalisation applicative.

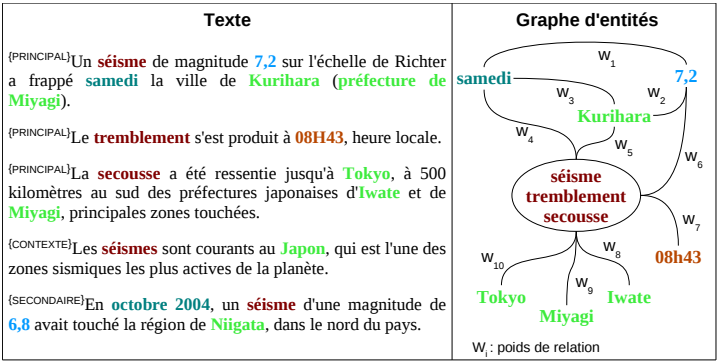


FIG. 2 – Exemple de graphe d'entités

La construction du graphe d'entités d'un texte commence en déterminant si les couples de mentions d'entités ou d'événements apparaissant dans une même phrase sont sous-tendus par une relation propre au type d'événement cible, sans néanmoins préciser cette relation. À l'instar des travaux existants comparables, nous avons réalisé cette détermination par le biais d'un classifieur statistique. Dans ce cadre, l'utilisation d'un ensemble de traits lexicalisés constitue l'approche dominante (Afzal, 2009; Gu et Cercone, 2006; Wick *et al.*, 2006), même si (Liu *et al.*, 2007) se démarque en conjuguant ces traits lexicalisés avec des traits de nature syntaxique. À l'inverse, nous avons construit un modèle n'intégrant que des traits syntaxiques et faisant abstraction des informations lexicales (mots sous forme fléchie ou lemmes) afin de lui conférer un degré de généralité plus important susceptible de rendre son adaptation à un autre domaine plus facile. Pour évaluer l'intérêt relatif des traits syntaxiques et lexicaux, nous avons entraîné différents types de classifieurs avec les trois ensembles de traits détaillés dans le tableau 1.

– *LEXI-BASE* : même ensemble de traits lexicalisés que (Afzal, 2009) ;
– *LEXI-SYN* : ensemble de traits conjuguant traits lexicalisés et traits syntaxiques, dans le prolongement de (Liu *et al.*, 2007)[1] ;
– *NON-LEXI-SYN* : même ensemble de traits que *LEXI-SYN*, à l'exception des traits lexicalisés.

Comme dans (McDonald *et al.*, 2005) et (Liu *et al.*, 2007), le poids associé à chaque relation trouvée est le score de confiance du classifieur l'ayant mise en évidence, ce score étant compris dans l'intervalle [0,1] pour tous les classifieurs expérimentés à la section 6.2.

La seconde étape de construction du graphe d'entités est une forme de condensation résultant de la fusion des mentions d'entités et d'événements identifiées comme faisant référence à une même

[1]Nous n'utilisons pas exactement les mêmes traits que (Liu *et al.*, 2007) car certains d'entre eux ne sont applicables que dans le domaine biomédical.

Traits	LEXI-BASE	LEXI-SYN	NON-LEXI-SYN
Type d'entité de E1 ; type d'entité de E2	✓	✓	✓
Catégories morpho-syntaxiques de E1 ; cat. morpho-synt. de E2	✓	✓	✓
Mots constitutifs de E1 ; mots constitutifs de E2	✓		
Bigrammes de mots de E1 ; bigrammes de mots de E2	✓	✓	
Mots situés entre E1 et E2	✓	✓	
Bigrammes de mots situés entre E1 et E2	✓	✓	
Catégories morpho-syntaxiques situées entre E1 et E2	✓	✓	✓
Nombre de mots situés entre E1 et E2	✓	✓	✓
Bigrammes de catégories morpho-syntaxiques entre E1 et E2		✓	✓
Nombre de relations syntaxiques entre E1 et E2		✓	✓
Chemin syntaxique entre E1 et E2		✓	✓
Position / un événement + catégorie morpho-syntaxique[1]		✓	✓
Nombre d'entités situées entre E1 et E2		✓	✓
Nombre de mentions d'événement entre E1 et E2		✓	✓
Catégorie morpho-syntaxique des deux mots avant/après E1		✓	✓
Catégorie morpho-syntaxique des deux mots avant/après E2		✓	✓

[1] Si E1, respectivement E2, est une mention d'événement, associe la position de E2, respectivement E1, par rapport à elle (avant ou après) et sa catégorie morpho-syntaxique.

TAB. 1 – Traits utilisés pour la classification de relations binaires

entité ou à un même événement. Pour les événements, cette fusion s'appuie sur la segmentation événementielle : toutes les mentions d'événements apparaissant dans un segment étiqueté PRINCIPAL sont supposées faire référence à l'événement principal du document et sont donc fusionnées (cf. fusion de *secousse*, *séisme* et *tremblement* au niveau de la figure 2). Pour les entités, la fusion se fait sur l'égalité de leur forme normalisée dans le cas des dates, heures et magnitudes et sur l'égalité de la forme trouvée dans les textes pour les lieux. Lorsque l'opération de fusion entraîne la présence de plusieurs relations entre deux entités ou entre une entité et l'événement principal, ces relations sont elles-mêmes fusionnées en conservant le poids le plus élevé.

5.2 Remplissage du formulaire

L'étape de remplissage du formulaire a pour objectif de choisir pour chaque rôle de ce formulaire l'entité du graphe construit à l'étape précédente ayant un type compatible avec le type d'entité attendu pour ce rôle et se montrant la plus à même de le remplir. Cette sélection s'accompagne implicitement du choix de ne pas remplir certains rôles du formulaire lorsque les informations correspondantes sont absentes du texte. Ce problème de remplissage de formulaire peut être assimilé au problème de la reconstruction d'une relation complexe tel qu'il est envisagé dans (Afzal, 2009; McDonald *et al.*, 2005). Par exemple, le graphe de la figure 2 comporte une ambiguïté relative à l'entité occupant le rôle de lieu de l'événement et impose un choix entre : *Kurihara, Tokyo, Miyagi ou Iwate*. Dans cette perspective, nous avons testé plusieurs approches :

Position est une heuristique simple mais très efficace dans le contexte considéré qui sélectionne pour chaque type d'entités la première mention apparaissant dans un segment relatif à l'événement principal.

Confiance retient pour chaque type d'entités l'entité liée à l'événement avec le score de confiance (score du classifieur utilisé) le plus grand.

PageRank est une approche exploitant la structure globale du graphe d'entités par le biais de l'algorithme PageRank. Ce dernier permet en l'occurrence d'attribuer un score d'importance à chaque entité en fonction de sa connectivité avec les autres entités et donc de les ordonner. Pour chaque type d'entités, est ainsi retenue l'entité ayant le plus haut score PageRank.

Vote implémente une stratégie de vote majoritaire reposant sur les approches *Position*, *Confiance* et *PageRank*. Pour chaque type d'entités, l'entité ayant été sélectionnée par le plus grand nombre d'approches est ainsi adoptée.

Hybride applique pour chaque type d'entités celle, parmi les stratégies précédentes, donnant le meilleur résultat pour ce type d'entités.

La sortie des approches *Confiance*, *PageRank*, *Vote* et *Hybride* est en outre complétée par l'approche *Position* dans le cas où aucune entité n'est sélectionnée pour un type donné. Il est en effet possible que certaines entités d'un formulaire apparaissent dans un texte sans être associées dans une phrase à une mention d'entité ou d'événement, ce qui interdit leur choix par les approches reposant sur le graphe d'entités.

6 Évaluation

Nous présentons dans cette section une évaluation de notre approche de remplissage de formulairessur un corpus de dépêches de presse concernant les événements sismiques, corpus décrit à la section 6.1. Une évaluation différenciée de chaque étape de notre approche a été menée : les résultats de l'évaluation de la segmentation événementielle de textes sont présentés en détail dans (Jean-Louis *et al.*, 2010) et ont montré que le modèle de segmentation par CRF atteint un F-score de 92,71% pour la détection de l'événement principal, ce qui constitue une bonne base pour l'application des étapes suivantes de notre approche. Les sections 6.2 et 6.3 présentent respectivement l'évaluation de la construction du graphe d'entités et de la sélection des entités. Une évaluation plus ciblée de l'impact de la segmentation en événements sur le résultat final est présentée à la section 6.4 et une analyse des principales erreurs rencontrées et de leur répartition est présentée dans la section 6.5.

6.1 Corpus

Les travaux présentés dans cet article ont été développés dans le cadre d'une application dédiée à la surveillance des événements sismiques à partir de dépêches de presse. Dans ce cadre, un formulaire est associé à un événement sismique et résume ses principales caractéristiques à savoir, la date, l'heure, le lieu, la magnitude, les coordonnées géographiques ainsi que la mention d'événement qui lui est associée (séisme, réplique, etc.)[2]. La figure 1 donne deux exemples illustratifs du formulaire que nous considérons. Notons que dans l'application visée, nous ne cherchons à extraire que les événements principaux et ne sommes donc pas intéressés par l'événement secondaire *EV2* de cette dépêche.

[2]Les dommages liés aux séismes n'ont pas été considérés car leur expression est plus variée et leur identification nécessiterait une analyse linguistique plus profonde.

[POSITIVE] : Cette *secousse*, d'une magnitude de *6,4* sur l'échelle de Richter, est la plus forte enregistrée depuis le tremblement de terre d'une magnitude de 8 qui a ravagé le Sichuan, a précisé un responsable du bureau de sismologie de cette province.

[NEGATIVE] : Cette *secousse*, d'une magnitude de 6,4 sur l'échelle de Richter, est la plus forte enregistrée depuis le tremblement de terre d'une magnitude de *8* qui a ravagé le Sichuan, a précisé un responsable du bureau de sismologie de cette province.

FIG. 3 – Exemples positif et négatif de présence d'une relation entre deux entités

L'ensemble des expériences ont été effectuées à partir d'un corpus composé de 501 dépêches de presse en français concernant le domaine sismique. Ces dépêches ont été collectées entre fin février 2008 et début septembre 2008, en provenance pour partie d'un flux de dépêches AFP (1/3 du corpus), et pour partie de dépêches collectées sur Google Actualités (2/3 du corpus). Le corpus a été manuellement annoté par des analystes du domaine qui ont rempli manuellement les formulaires pour chaque séisme principal d'un document. Au total, les annotateurs ont identifié 2 775 entités, réparties en 6 types d'entités : mention d'événement (18%), lieux (34%), date (17%), heure (12%), magnitude (17%) et coordonnées géographiques (1%)[3].

Concernant l'analyse linguistique des documents, nous avons appliqué la chaîne de traitements linguistiques de l'analyseur LIMA (Besançon *et al.*, 2010) réalisant les étapes de tokenisation, détection des fins de phrases, désambiguïsation morpho-syntaxique, reconnaissance des temps des verbes, reconnaissance des entités nommées et analyse syntaxique.

6.2 Construction du graphe d'entités

La méthode proposée pour la construction du graphe d'entités s'appuie sur un classifieur pour déterminer la présence/absence d'une relation entre deux entités au sein d'une même phrase. Nous avons expérimenté différents types de classifieurs statistiques, utilisant chacun les trois ensembles de traits présentés à la section 5.1 (*LEXI-BASE, LEXI-SYN, NON-LEXI-SYN*). Pour l'annotation des relations binaires entre entités, nous avons considéré un sous-ensemble du corpus composé de 44 dépêches. Sur ce sous-ensemble, nous avons obtenu 5 000 relations binaires, parmi lesquelles 969 relations sont exprimées à l'intérieur de la même phrase. Parmi celles-ci, 43 relations ont été écartées à cause d'erreurs de reconnaissance des entités (par exemple, lorsqu'une entité considérée est en fait incluse dans une entité plus large non reconnue à cause de son type). Les autres relations ont servi pour l'entraînement des classifieurs : 690 pour la catégorie *POSITIVE*, dans laquelle les deux entités font référence au même événement sismique et 236 pour la catégorie *NEGATIVE*, dans laquelle les deux entités sont associées à des événements sismiques différents. La figure 3 illustre des relations pour les deux catégories.

Ce corpus annoté nous a servi à tester trois types de classifieurs[4] : Bayésien Naïf (*NB*), Maximum d'Entropie (*ME*) et Arbres de décision (*DT*). Nous reportons dans le tableau 2 les résultats

[3]La possibilité a été laissée aux annotateurs de retenir plus d'une entité pour le même rôle lorsque plusieurs variantes étaient mentionnées et étaient jugées également pertinentes. Par exemple, pour les lieux, pouvaient être annotés à la fois un nom de ville et un nom de pays.

[4]Nous avons utilisé l'implémentation fournie par l'outil MALLET (`http://mallet.cs.umass.edu`).

obtenus par chaque algorithme, en fonction de l'ensemble de traits utilisé en termes de rappel (*R*), précision (*P*) et F1-mesure (*F*). Les résultats sont obtenus par une validation croisée (4/5 des données servent à l'entraînement et 1/5 pour le test). En complément, nous fournissons pour comparaison les résultats d'une approche basique (*Baseline*) qui attribue la catégorie la plus fréquente (*POSITIVE*) à toutes les relations.

Ensemble de traits	**Classifieur**	**R(%)**	**P(%)**	**F(%)**
LEXI-SYN	ME	96,30	95,92	96,10
LEXI-BASE	ME	91,22	96,09	93,57
NON-LEXI-SYN	ME	91,66	94,99	93,26
LEXI-SYN	DT	89,01	96,45	92,55
LEXI-SYN	NB	93,44	90,69	92,02
NON-LEXI-SYN	DT	91,17	88,74	89,83
NON-LEXI-SYN	NB	89,58	89,23	89,37
LEXI-BASE	DT	84,35	94,70	89,16
LEXI-BASE	NB	86,73	87,86	87,27
Baseline	–	100,00	25,50	40,49

TAB. 2 – Évaluation du classifieur de relations binaires

En premier lieu, les résultats du tableau 2 montrent l'intérêt d'utiliser des traits de nature syntaxique : les scores obtenus à partir de l'ensemble LEXI-SYN dépassent ceux de l'ensemble LEXI-BASE pour les trois modèles. De plus, l'ensemble de traits non lexicalisés NON-LEXI-SYN obtient des scores équivalents à ceux de l'ensemble LEXI-BASE, ce qui est intéressant pour obtenir des modèles plus génériques. Concernant les algorithmes d'apprentissage, les performances s'organisent selon la hiérarchie ME > DT > NB. Notons que (Afzal, 2009) obtient une hiérarchie différente (DT > ME > NB) mais utilise un corpus différent et dans une autre langue, ce qui rend la comparaison difficile. En termes de performances générales, nos résultats sont comparables à ceux présentés par (Afzal, 2009), ses meilleurs scores étant R=95%|P=87%|F=91%, obtenus avec des arbres de décision. Pour la suite de notre démarche, nous avons conservé le modèle Maximum d'Entropie reposant sur l'ensemble de traits NON-LEXI-SYN plutôt que l'ensemble LEXI-SYN. Notre motivation pour ce faire est que l'ensemble NON-LEXI-SYN permet d'obtenir des scores satisfaisants sans être fondé sur des informations fortement liées à un domaine, ce qui n'est pas le cas pour les traits lexicalisés.

6.3 Sélection des entités et remplissage des formulaires

Concernant l'évaluation des stratégies de sélection, l'ensemble des documents du corpus a été utilisé. Nous reportons dans le tableau 3 les scores de remplissage des formulaires pour ces différentes stratégies, agrégés pour l'ensemble des rôles du formulaire, en termes de rappel (*R*), précision (*P*) et F1-mesure (*F*).

Ces résultats confirment en premier lieu que notre méthode de référence *Position* est caractérisée par un niveau déjà très élevé. De plus, cette méthode permet d'obtenir des performances légèrement supérieures à la stratégie *PageRank*, ce qui peut se justifier en partie par le fait que la stratégie *PageRank* repose uniquement sur la structure du graphe, sans tenir compte des poids sur les arcs. Par conséquent, les entités se trouvant dans des zones densément connectées du

Stratégie de sélection	R(%)	P(%)	F(%)
Hybride	77,55	76,87	77,15
Vote	74,93	74,27	74,54
Confiance	74,89	74,16	74,47
Position	73,40	73,06	73,17
PageRank	72,41	71,73	72,01

TAB. 3 – Évaluation du remplissage des formulaires à partir des stratégies de sélection

graphe obtiennent de meilleurs scores que les autres, indépendamment des poids sur les arcs. Ce problème pourrait être, dans une certaine mesure, minimisé en adoptant la version pondérée de l'algorithme PageRank proposée dans (Mihalcea, 2004). D'autre part, les scores du tableau 3 montrent que la meilleure stratégie de sélection est l'approche *Hybride*, ce qui est cohérent avec ses objectifs de faire correspondre à un rôle du formulaire la stratégie qui lui est la mieux adaptée.

6.4 Impact de la segmentation en événements

Dans cette section, nous proposons d'évaluer l'impact de notre approche de segmentation en événements sur la tâche de remplissage des formulaires. Cette segmentation vise à identifier les passages pertinents afin de focaliser le processus d'extraction. Cependant, tous les documents ne mentionnent pas plusieurs événements sismiques et dans le cas des documents ne mentionnant qu'un seul événement, l'usage de la segmentation événementielle se justifie moins (toutes les phrases font *a priori* référence au même événement si elle comporte une mention d'événement). Celle-ci est dès lors susceptible d'apporter essentiellement des perturbations dans la mesure où ses résultats ne sont nécessairement pas parfaits.

Notre but, dans cette section, est donc de mesurer l'impact de la segmentation en événements sur les documents ne faisant référence qu'à un seul événement en comparaison avec ceux faisant référence à plusieurs événements. Notre intuition est que la segmentation devrait avoir un impact limité sur les documents mono-événements et devrait améliorer les scores pour les documents multi-événements. Afin de vérifier cette hypothèse, nous avons manuellement divisé le corpus initial en deux ensembles en fonction du nombre d'événements sismiques mentionnés par les textes. Nous avons ainsi obtenu 227 documents multi-événements (*M*) et 274 documents mono-événements (*S*). Les résultats du remplissage de formulaires pour chaque ensemble, avec ou sans segmentation, sont présentés dans le tableau 4 en termes de F1-mesure et agrégés pour l'ensemble des rôles du formulaire.

Concernant les documents mono-événements, les scores du tableau 4 montrent que les stratégies les plus performantes n'utilisent pas de segmentation, bien que les différences ne soient pas très importantes (+0,71% en moyenne). À l'opposé, les stratégies à base de segmentation sont plus performantes pour les documents multi-événements (+2,74% en moyenne). De plus, notre stratégie la plus performante, approche *Hybride* avec segmentation, obtient de meilleurs scores que notre approche de référence, *Position* sans segmentation, et ce, pour les deux ensembles de documents. Plus généralement, les résultats démontrent que notre segmentation n'introduit qu'une perte limitée pour les documents mono-événements et améliore les performances pour

	Sans segmentation		Avec segmentation	
Stratégie	S(%)	M(%)	S(%)	M(%)
Hybride	79,20	73,61	78,34	75,61
Vote	77,67	68,68	76,89	71,81
Confiance	72,55	66,07	71,79	69,10
Position	73,96	73,16	73,07	73,10
PageRank	70,92	59,72	70,67	65,32

TAB. 4 – Impact de la segmentation sur les documents mono/multi-événements (F1-mesure)

les documents multi-événements.

6.5 Analyse des erreurs

Dans la perspective d'approfondir les évaluations de nos stratégies de remplissage de formulaire, nous avons mené une analyse des erreurs en cherchant à identifier précisément les causes de la présence d'une entité incorrecte (sélection d'une mauvaise entité pour un rôle) ou d'une entité manquante (pas d'entité sélectionnée pour un rôle) dans un formulaire. Dans ce cadre nous avons identifié trois types d'erreurs prépondérants :

– les erreurs de reconnaissance des entités nommées : l'entité n'est pas reconnue lors de l'analyse linguistique du texte ;
– les erreurs de segmentation en événements : l'entité est identifiée lors de l'analyse linguistique mais elle appartient à une phrase qui n'est pas associée à l'événement principal ;
– les erreurs de sélection des entités : l'entité se trouve dans le segment de l'événement principal mais une autre entité a été retenue comme valeur pour le rôle dans le formulaire ;

Le tableau 5 présente la répartition de chaque type d'erreurs, en comparaison avec le nombre d'entités correctement repérées, pour deux approches de construction des formulaires : la première correspond à la sélection à base d'heuristique, sans segmentation (*NonSeg+Position*) ; la seconde s'appuie sur la segmentation en événements et la stratégie *Seg+Hybride*.

Type d'erreurs	NonSeg+Position	Seg+Hybride
Correct	71,6%	75,1%
Sélection d'entités	25,6%	21,2%
Reconnaissance d'entités	2,8%	2,8%
Segmentation	–	0,8%

TAB. 5 – Répartition des erreurs pour le remplissage de formulaires

Le graphe montre que la stratégie de référence *Position* permet d'identifier correctement une part conséquente des entités (*71,6%*) mais qu'un nombre important d'erreurs d'attribution de rôle dans le formulaire (*25,6%*) subsiste. Notre meilleure stratégie réduit ce type d'erreur tout en améliorant le pourcentage d'entités correctes dans les formulaires. De plus, cette stratégie n'induit qu'un nombre très limité d'erreurs dues à la segmentation en événements (*0,8%*).

7 Conclusion

La plupart des approches pour l'extraction d'information s'appuient sur des éléments au niveau phrastique pour remplir automatiquement des formulaires et peu sur des informations au niveau discursif. Dans cet article, nous avons présenté une approche pour le remplissage de formulaires fondée sur une segmentation du texte et une sélection des entités s'appuyant sur un graphe global de relations entre les entités. La segmentation du texte se fait au niveau discursif et utilise des informations temporelles pour segmenter le texte selon les événements présents en utilisant un modèle CRF afin de trouver les phrases les plus pertinentes pour remplir un formulaire donné. Ces phrases sont ensuite utilisées pour construire un graphe d'entités à partir duquel les entités relatives à l'événement d'intérêt sont sélectionnées. Nous avons proposé plusieurs stratégies pour sélectionner les entités (utilisant la position des entités, les scores de confiance des relations ou la structure du graphe, par l'utilisation de PageRank) ainsi que plusieurs façons de combiner ces stratégies (vote majoritaire ou approche hybride).

Nous avons également présenté une évaluation détaillée de notre approche sur un corpus de dépêches de presse concernant les événements sismiques. Cette évaluation a montré que notre approche a permis d'améliorer le remplissage de formulaire par rapport à une heuristique simple (mais efficace) consistant à prendre la première entité du type cherché pour remplir chaque champ du formulaire. Les résultats ont aussi montré que notre approche est particulièrement adaptée pour les documents mentionnant plusieurs événements de même nature. Finalement, une analyse des erreurs a montré que l'on peut encore améliorer ces résultats puisque la part d'erreurs liée à la sélection des entités reste de 21%.

Concernant les perspectives de nos travaux, nous allons expérimenter la généralisation de notre approche de remplissage de formulaires à d'autres contextes, et plus précisément, d'autres langues et d'autres domaines. Nous avons déjà obtenu des résultats prometteurs en testant la segmentation événementielle sur un ensemble de dépêches de presse en anglais, dans le domaine sismique, avec peu d'efforts d'adaptation nécessaires. En ce qui concerne la généralisation à d'autres domaines, nous planifions des expérimentations dans le domaine financier.

Références

Afzal, N. (2009). Complex Relations Extraction. *In Conference on Language & Technology 2009 (CLT'09)*, Lahore, Pakistan.

Besançon, R., de Chalendar, G., Ferret, O., Gara, F. et Semmar, N. (2010). LIMA : A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation. *In 7th Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

Chambers, N. et Jurafsky, D. (2011). Template-Based Information Extraction without the Templates. *In 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 976–986, Portland, Oregon, USA.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S. et Weischedel, R. (2004). The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. *In 4th Conference on Language Resources and Evaluation (LREC 2004)*, pages 837–840, Lisbon, Portugal.

Feng, D., Burns, G. et Hovy, E. (2007). Extracting Data Records from Unstructured Biomedical Full Text. *In EMNLP-CoNLL'07*, pages 837–846, Prague, Czech Republic.

GOERTZEL, B., PINTO, H., HELJAKKA, A., ROSS, M., PENNACHIN, C. et GOERTZEL, I. (2006). Using Dependency Parsing and Probabilistic Inference to Extract Relationships between Genes, Proteins and Malignancies Implicit Among Multiple Biomedical Research Abstracts. *In HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, pages 104–111, New York, USA.

GRISHMAN, R. et SUNDHEIM, B. (1996). Message Understanding Conference-6 : A Brief History. *In 16th International Conference on Computational linguistics (COLING'96)*, pages 466–471, Copenhagen, Denmark.

GU, Z. et CERCONE, N. (2006). Segment-based hidden Markov models for information extraction. *In 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 481–488, Sydney, Australia.

JEAN-LOUIS, L., BESANÇON, R. et FERRET, O. (2010). Using temporal cues for segmenting texts into events. *In 7th International Conference on Natural Language Processing (IceTAL 2010)*, pages 150–161. Springer Berlin / Heidelberg.

JEAN-LOUIS, L., BESANÇON, R. et FERRET, O. (2011). Text segmentation and graph-based method for template filling in information extraction. *In 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 723–731, Chiang Mai, Thailand.

JI, H., GRISHMAN, R. et TRANG DANG, H. (2010). Overview of the TAC 2010 Knowledge Base Population Track. *In Third Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA.

LAFFERTY, J. D., MCCALLUM, A. et PEREIRA, F. C. N. (2001). Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. *In Eighteenth International Conference on Machine Learning (ICML'01)*, pages 282–289, San Francisco, CA, USA.

LIU, Y., SHI, Z. et SARKAR, A. (2007). Exploiting Rich Syntactic Information for Relationship Extraction from Biomedical Articles. *In NAACL-HLT'07, short paper session*, pages 97–100, Rochester, New York.

MANSURI, I. R. et SARAWAGI, S. (2006). Integrating unstructured data into relational databases. *In 22nd International Conference on Data Engineering (ICDE'06)*, pages 29–40, Washington, USA.

MCDONALD, R., PEREIRA, F., KULICK, S., WINTERS, S., JIN, Y. et WHITE, P. (2005). Simple algorithms for complex relation extraction with applications to biomedical IE. *In ACL 2005*, pages 491–498, Ann Arbor, Michigan, USA.

MIHALCEA, R. (2004). Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. *In 42st Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, Barcelona, Spain.

NAUGHTON, M. (2007). Exploiting Structure for Event Discovery Using the MDI Algorithm. *In 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 31–36, Prague, Czech Republic.

PATWARDHAN, S. et RILOFF, E. (2007). Effective Information Extraction with Semantic Affinity Patterns and Relevant Regions. *In EMNLP-CoNLL'07*, pages 717–727, Prague, Czech Republic.

STEVENSON, M. (2006). Fact distribution in Information Extraction. *Language Resources and Evaluation*, 40(2):183–201.

TURMO, J., AGENO, A. et CATALÀ, N. (2006). Adaptive information extraction. *ACM Computer Surveys*, 38(2):1–47.

WICK, M., CULOTTA, A. et MCCALLUM, A. (2006). Learning Field Compatibilities to Extract Database Records from Unstructured Text. *In EMNLP'06*, pages 603–611, Sydney, Australia.

Traitement automatique sur corpus de récits de voyages pyrénéens : Une analyse syntaxique, sémantique et temporelle

Anaïs Lefeuvre[1,2,3], Richard Moot[1,2], Christian Retoré[1,2], Noémie-Fleur Sandillon-Rezer[1,2]
Université de Bordeaux[1], LaBRI-CNRS[2] & INRIA [3]
anais.lefeuvre@labri.fr, moot@labri.fr, retore@labri.fr, nfsr@labri.fr

RÉSUMÉ

Cet article présente notre utilisation de la théorie des types dans laquelle nous nous situons pour l'analyse syntaxique, sémantique et pour la construction du lexique. Notre outil, Grail permet de traiter le discours automatiquement à partir du texte brut et nous le testons sur un corpus de récit de voyages pyrénéens, Itipy. Nous expliquons donc notre usage des grammaires catégorielles et plus particulièrement du calcul de Lambek et la correspondance entre ces catégories et le λ-calcul simplement typé dans le cadre de la DRT. Une flexibilité du typage doit être autorisée dans certains cas et bloquée dans d'autres. Quelques phénomènes linguistiques participant à une forme de glissement de sens provocant des conflits de types sont présentés. Nous expliquons ensuite nos motivations d'ordre pragmatique à utiliser un système à sortes et types variables en sémantique lexicale puis notre traitement compositionnel du temps des évènements inspiré du *Binary Tense* de (Verkuyl, 2008).

ABSTRACT

Processing of a Pyrenees travel novels corpus : a syntactical, semantical and temporal analysis.

In this article, we present a type theoretical framework which we apply to the syntactic analysis and the computation of DRS semantics. Our tool, Grail, is used for the automatic treatment of French text and we use a Pyrenées travel novels corpus, Itipy, as a test case. We explain our use of categorial grammars and specifically the Lambek calculus and its connection to the simply typed λ-calculus in connection with DRT. Flexible typing has to be allowed in some cases and forbidden in others. Some linguistic phenomena presenting some kind of meaning shifts inducing typing conflicts will be introduced. We then present our motivations in the pragmatic field to use a system with sorts and variable types in lexical semantics and then we present how we process events temporality, in the light of Verkuyl's *Binary Tense*(Verkuyl, 2008)

MOTS-CLÉS : compositionalité, interface syntaxe-sémantique, interface sémantique-pragmatique, grammaire catégorielle, théorie des types, récit de voyage.

KEYWORDS: compositionality, syntax-semantics interface, semantics-pragmatics interface, categorial grammar, type theory, travel novel.

Ce travail de recherche a reçu un soutien financier d' INRIA et du Conseil Régional d'Aquitaine dans le cadre du projet Itipy

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 43–56,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Cet article décrit les étapes qui composent notre analyse du discours, en partant du texte brut pour en produire une représentation sémantique dans le cadre de la *Discourse Representation Theory* (DRT) (Kamp et Reyle, 1993). Une chaine complète de traitement est proposée et testée sur le corpus Itipy. Ce corpus de récits de voyage pyrénéens a été rassemblé par la médiathèque de Pau pour mettre en valeur le fond patrimonial de récits de voyage dans sa région.

Une analyse du discours impose de fait une interaction entre la sémantique des unités de langue dont on doit interpréter le sens en discours et la prise en compte de la dimension pragmatique de ce qui est dit (Busquets *et al.*, 2001). Certains phénomènes sémantiques restent difficiles à traiter, certains cas de glissement de sens montrent qu'une flexibilité dans le typage doit être permise, alors que dans les cas les plus courants le typage doit être rigide pour éviter une représentation inappropriée. Nous donnons quelques exemples et proposons, afin d'améliorer les résultats de notre chaîne de traitement, de traiter ces phénomènes par l'affinement des λ-termes du lexique dans le cadre d'un systèmes à sortes et types variables, dans le λ-calcul d'ordre supérieur.

Dans le cadre du projet Itipy, nous nous sommes intéressés à la dimension temporelle des évènements dans le discours, ce qui nous a amené à interroger la compositionnalité de cet aspect. En nous appuyant sur les travaux de (Verkuyl, 2003), nous voulons introduire dans le traitement un système de test puis un lexique propre aux locutions adverbiales de temps. Nous avons introduit les termes en λ-calcul qui permettent de déterminer la valeur temporelle d'un évènement. On compose le terme propre à la phrase que l'on applique aux termes propres des adverbes, puis aux opérateurs perfectif ou imperfectif, postérieur ou simultané, et enfin présent ou passé en fonction de la morphologie du verbe conjugué. Nous expliquerons plus en détails ce système et traiterons deux exemples de notre corpus illustrant ce système.

Nous détaillerons d'abord notre corpus et nos objectifs applicatifs quant à celui-ci, nous présenterons les étapes de traitement du discours, commençant par l'acquisition de la grammaire du français sur corpus annoté, puis l'analyse syntaxique dans le cadre des grammaires catégorielles. Nous expliquerons plus amplement l'interface syntaxe-sémantique dans la théorie des types logiques permettant la construction de nos représentations sémantiques en λ-DRT. Nous présenterons brièvement le système de types variables et notre traitement des phénomènes discursifs en jeu dans l'interaction sémantique-pragmatique, ainsi que le traitement temporel des évènements.

2 Le corpus

Notre corpus de 576 334 mots est une collection de 11 oeuvres classées par la médiathèque de Pau comme récits de voyages pyrénéens du XIXème et début XXème siècle. Concernant les données textuelles de notre corpus, le genre du récit de voyage, implique de fait une hétérogénéité interne reconnue. Certains spécialistes désignent par ailleurs le récit de voyage comme un "genre fragmenté" (Magri-Mourgues, 2009), dans lequel on trouve une myriade de procédés narratifs incluant "le récit métonymique" , "le récit synecdochique", "le récit métaphorique", "le récit de voyage et de découverte du réel", etc. Ajoutons à ceci que le corpus Itipy est constitué de récits écrits par des géologues, des topographes, ou encore des romanciers.

Malgré la diversité des formes de discours qui composent le corpus, sa spécificité réside dans

le récit de l'itinéraire, seul point commun entre tous les textes. La structure narrative du récit de voyage observe une alternance entre la description de l'itinéraire emprunté et d'autres informations telles que des observations sur le relief, le caractère des personnages rencontrés ou encore des considérations introspectives du narrateur sur des domaines variés.

Notre traitement du discours se détache totalement des données du genres, ou encore de la nature des thèmes abordés. Nous intervenons alors sur l'analyse profonde syntaxique et sémantique, mais aussi pragmatique et discursive du discours. En partant d'une automatisation de l'analyse syntaxique, la représentation sémantique est construite automatiquement elle aussi utilisant pour ressource un lexique sémantique saisi préalablement à la main. Afin de produire des représentations du discours satisfaisantes, nous travaillons sur un affinement des λ-termes contenus dans ce lexique.

3 Analyse syntaxique

Grail est un analyseur pour grammaires catégorielles dans la tradition de Lambek (Lambek, 1958) et leurs extensions multimodales (Moortgat, 1997). Dans des travaux récents (Moot, 2010a,b), Grail a été étendu pour l'analyse du français à large couverture. La figure 1 montre les composants de la chaîne de traitement pour le français : il y a un *part-of-speech tagger*, un *supertagger* (Clark et Curran, 2004) pour limiter le nombre de formules que l'analyseur doit traiter et un lexique sémantique qui associe à chaque mot un λ-terme correspondant à son type. Dans le lexique, on utilise des λ-termes produisant des DRS après substitution et normalisation (Muskens, 1994).

3.1 Acquisition du lexique syntaxique et apprentissage des modèles d'entropie maximale

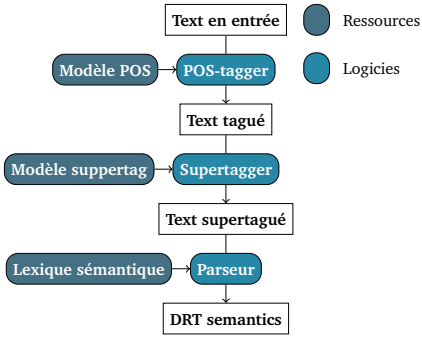


FIGURE 1 – Schéma des ressources et outils de la chaîne de traitement

Le *French Treebank* (Abeillé *et al.*, 2003) a été transformé, en partie automatiquement, en dérivations pour grammaires catégorielles. La complexité résidait dans le fait que les arbres du *French Treebank* sont plats, avec un nombre de fils maximal par noeud non fixé, alors que les arbres de dérivations doivent être binaires. La figure 2 montre une sous-partie d'un arbre du corpus. Des techniques standard pour extraire des grammaires catégorielles à partir d'arbres d'annotation (Buszkowski et Penn, 1990) nécessitent des arbres binaires avec, pour chaque paire de frères, une indication pour la tête, et une pour l'argument : si la tête est à gauche, les formules correspondant au deux frères sont A/B et B, si la tête est à droite ce sont B et $B\backslash A$, où la formule B dépend du syntagme du noeud dans l'arbre d'annotation (eg. NP correspond à la for-

mule atomique np et INF correspond à la formule complexe $np\backslash s_{inf}$). Les arbres du corpus sont alors binarisés et des heuristiques déterminent la tête d'un syntagme, faisant un effort pour rester le plus fidèle possible aux analyses habituelles en grammaire catégorielles ; typiquement le verbe est la tête d'une phrase, le déterminant la tête d'un groupe nominal (pour rester cohérent avec une possible analyse sémantique ultérieure), etc. On calcule ainsi récursivement les formules, de la racine aux feuilles, où les feuilles donnent les formules pour les mots du lexique.

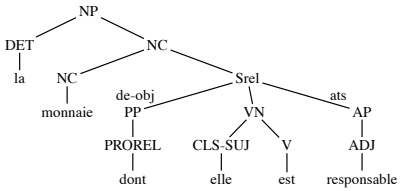


FIGURE 2 – Exemple d'un arbre planaire du *French Treebank*, avant traitement

Une méthode utilisée pour extraire une grammaire catégorielle du *French Treebank*, et donc un lexique représentatif de celui-ci, est l'utilisation d'un transducteur d'arbre. Le principe du transducteur d'arbre, explicité dans (Comon *et al.*, 1997), est de prendre en entrée un arbre tel que montré dans 2 et et restituer un nouvel arbre en sortie. Pour ce travail, nous nous sommes limités à la partie *AB* d'une grammaire de Lambek, qui correspond aux mécanismes les plus basiques d'une langue naturelle. Les grammaires AB sont une référence en inférence grammaticale (Buszkowski et Penn, 1990) et c'est ce qui a motivé notre choix premier. La sortie de ce transducteur [1] est donc une forêt d'arbres de dérivation d'une grammaire AB. A partir de celle-ci, nous pouvons soit extraire un lexique (voir figure 3), qui contient les mots des phrases analysées, les différents types trouvés et leur occurrence, soit une grammaire (voir figure 5) d'arbres qui contient à la fois les règles que nous considérons comme correctes (étant donné que ce sont celles qui apparaissent dans nos arbres de dérivation) et des probabilités sur ces règles. Le lexique peut servir à entrainer le Supertagger.

$$8374 : la \quad \rightarrow \quad 7996 : np/n, \quad 94 : (n\backslash n)/n, \quad 57 : (s\backslash s)/n, \quad 43 : (s/s)/n, ...$$

FIGURE 3 – Extrait du lexique. Le mot "la" est utilisé 8374 fois dans la partie analysée du corpus. La catégorie la plus fréquente correspond à celle d'un déterminant qui attend un nom commun à sa droite pour créer un groupe nominal. Les trois types suivant correspondent à des modificateurs, comme "**La** semaine dernière …" en début de phrase.

Il ne reste ensuite qu'à extraire le λ-terme correspondant à l'arbre. Cette méthode présente cependant des limitations, et certains phénomènes, tels que les traces, ou les ellipses ne se traitent pas avec des grammaires AB. Sans que cela limite le nombre de phrases analysées, les types donnés aux mots s'en trouvent complexifiés.

Ces traces, qui sont des dépendances non-bornées, ne sont pas indiquées dans le corpus et ont été ajoutées à la main dans une phase de nettoyage post-traitement. De plus, une phase de correction manuelle est nécessaire pour éliminer certaines différences entre les analyses choisies pour le corpus et les analyses habituellement utilisées pour les grammaires catégorielles (voir (Moot, 2010a)). Les modifications sont entre autre l'ajout de trace (il y a plus de 500 occurences de "que/qu"), la restructuration des groupes verbaux (l'argument du noyau verbal devint argument uniquement du participe passé lorsqu'il y a lieu).

1. Le transducteur que nous avons mis en place est détaillé dans (Sandillon-Rezer et Moot, 2011).

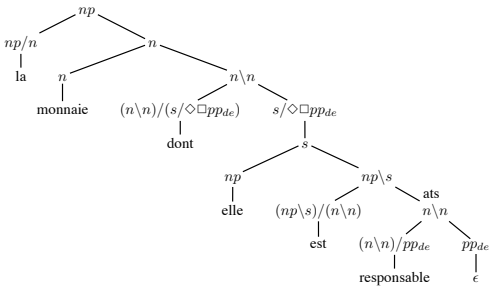


FIGURE 4 – résultat de l'extraction

Elles permettent de réduire le nombre de formules du lexique, qui passent de 5240 à 918, et donc le nombre d'analyses possibles. La figure 4 montre le résultat de l'extraction : les feuilles s'ajoutent au lexique.

Pour l'entrainement des taggers, 11.196 phrases (334.525 mots) sont utilisées et 1.244 phrases (36.504 mots) pour l'évaluation. Les modèles atteignent une précision de 98,4% pour le tagger et de 90,5% pour le supertagger (Moot, 2010a).

s	$\rightarrow np\ np\backslash s$	21,56%
s	$\rightarrow s/s\ s$	7,18%
...		

FIGURE 5 – Extrait de la grammaire. On trouve les règles binaires ayant généré la forêt d'arbres, groupées par racine (ici s) et le pourcentage d'utilisation de celles-ci.

3.2 Analyseur

Comme indiqué dans la figure 1, après une *tokenization* simple, les mots d'une phrase d'entrée sont d'abord étiquetés par le tagger. Au niveau des étiquettes utilisées lors de la *tokenization*, on utilise celles de Treetagger, car elles donnent des informations sur le temps des verbes, ce qui va nous servir pour le traitement du temps, sans pour autant utiliser Treetagger pour l'analyse.

Le supertagger sert surtout comme filtre à l'analyseur propre : en utilisant un facteur β ($\beta \leq 1$), réglable par l'utilisateur, et en se fixant sur le type le plus probable pour un mot, il va sélectionner les types dont les probabilités sont au delà de $\beta.proba\_max$. Ainsi, plus β est petit, plus les types ayant une faible probabilité sont proposés à l'analyseur. Ceci permet d'ajouter plus de formules pour des mots considérés comme difficiles (c'est-à-dire pour lesquels le modèle a relativement peu confiance en son premier choix), mais garde une seule formule pour les mots considérés comme faciles.

Ensuite, l'analyseur se charge des combinaisons des formules selon les règles du calcul de Lambek multimodal, utilisant les formules lexicales les plus probables prioritairement. Pour la grammaire à large couverture, Grail ne garde que la première analyse trouvée. Et cette analyse, qui correspond à un lambda terme simplement typé, va servir pour calculer la sémantique. Remarquons que ce calcul est très simple : on utilise simplement la β-réduction (avec peu ou pas de duplication de termes) pour obtenir une forme normale qui correspond à une formule ou, dans notre cas, à une DRS.

4 Représentation sémantique en λ-DRT

A l'instar de Boxer, développé par (Curran *et al.*, 2007), Grail associe à l'analyse syntaxique, une représentation sémantique dans le style de la DRT. La DRT est une théorie proposant de représenter la sémantique d'un discours grâce à un modèle présenté comme une boîte (*Discourse Representation Structure*) dans laquelle on trouve d'un coté le domaine, composé des référents du discours, quantifiés implicitement par un existentiel, et de l'autre les conditions d'interprétation sémantique de ce modèle. Cette théorie permet de construire ces structures mettant en évidence la valeur sémantique, d'un point de vue logique, des éléments porteurs de sens d'un énoncé au sein d'un contexte. Les descriptions sémantiques en λ-DRT de chaque item constituent un lexique écrit à la main préalablement et qui sert de ressource pour la chaine de traitement.

Abordons maintenant la correspondance entre catégories dans le calcul de Lambek et propriétés calculatoires et logiques des λ-termes. L'isomorphisme Curry-Howard montre que les dérivations en grammaires catégorielles sont des sous-ensembles des dérivations de la logique intuitionniste. Autrement dit, à chaque analyse catégorielle de phrase correspond un λ-terme normal. Les catégories syntaxiques dans le style du calcul de Lambek (*s, sn/n*) permettent donc d'associer une lecture sémantique exprimée par le λ-terme simplement typé fourni par le lexique, dans le style de la DRT, et correspondant à chaque mot taggé. Au sein de la structure syntaxique, on associe à sa catégorie le terme en λ-DRT associé puis on β-réduit l'expression. Les DRS créées se fusionnent les unes avec les autres par l'opération de "*merge*" et permettent d'interpréter les phénomènes de cohérence du discours, comme par exemple la résolution des anaphores pronominales. En dérivant l'analyse sémantique de l'analyse syntaxique, on conserve la bonne formation de la représentation de l'expression correspondant exactement à la catégorie.

4.1 De l'analyse syntaxique à la forme logique

En premier lieu, il convient de définir ce qu'est un type selon la sémantique de Montague. L'analyse syntaxique de type *s* présentée précédemment, est un λ-terme dont les variables libres correspondent aux mots. Le lexique fournit des λ-termes du même type sémantique. En les substituant, et en réduisant le terme obtenu, on obtient un terme normal de type **t** : c'est une formule logique, la représentation sémantique, et dans notre cas la λ-DRS. Néanmoins il faut au minimum partager le type **e**, les individus (aussi appelés entités) [2], en diverses sortes pour que le calcul de la sémantique bloque à juste titre lorsque le type d'un argument ne correspond pas au type attendu par la fonction.

Pour notre système de test, nous avons implémenté un petit jeu d'adverbes temporels comme *dans x heures*, *en x minutes*, *pendant x jours*, etc. Nous souhaitons à terme intégrer les outils tels que (Bittar, 2009), (Parent *et al.*, 2008). Nous précisons que χ (chronos) est la fonction prenant un évènement et renvoyant un intervalle temporel. Prenons deux exemples issus du corpus et le lexique grammatical, syntaxique puis sémantique correspondant à chacun des items : Les formules ici présentées sont des formules de logique partielles dont nous détaillerons la construction dans la section 4.3.

(1) *Le 31, nous sommes partis à six heures du matin.*

2. Nous utilisons les variables e dans les exemples suivant pour désigner des évènements et non des personnes.

item lexical	POS	catégorie	λ-terme
Le 31	ADV	s/s	$\lambda s\lambda e.(s\ e) \wedge \chi e \subseteq jour(31)$
nous	PRO :PER	sn	$nous$
sommes	VER :pres	$(sn\backslash s)/(sn\backslash s_{ppart})$	PRES (PERF)
partis	VER :pper	$sn\backslash s_{ppart}$	$\lambda x\lambda e.partir(e,x)$
à six heures du matin	ADV	$s\backslash s$	$\lambda s\lambda e.(s\ e) \wedge \chi e \circ 6\ :00$

(2) *Dans dix minutes, j'aurai quitté Nohant.*

item lexical	POS	catégorie	λ-terme
Dans dix minutes	ADV	s/s	$\lambda s\lambda e.(s\ e) \wedge distance_{min}(\chi e, n) = 10$
j	PRO :PER	sn	je
aurai	VER :fut	$(sn\backslash s)/(sn\backslash s_{ppart})$	PRES(POST)(PERF)
quitté	VER :pper	$(sn\backslash s_{ppart})/sn$	$\lambda y\lambda x\lambda e.quitter(e,x,y)$
Nohant	NAM	sn	$Nohant$

Les λ-termes manipulant les intervalles χe sont inspirés des préconisations de (Verkuyl, 2003) qui propose un traitement des adverbes. Par exemple pour *dans dix minutes*, on utilise une fonction qui donne la distance la plus courte en minutes entre le moment d'énonciation n et l'intervalle repère de l'évènement (R chez Reichenbach), tout comme pour PRES, POST et PERF, nous reviendrons plus en détail sur les termes associés à ces opérateurs dans la section 4.3.

Pour simplifier la présentation du lexique, *je*, *nous* et *Nohant* sont présentées comme des constantes. *partir* est un prédicat à deux arguments, respectivement un évènement nommé e et un agent, x. On remarque la correspondance entre le sn attendu à gauche dans la catégorie syntaxique et la présence d'un argument agentif dans la représentation sémantique par le λ-terme. *nous* appliqué au terme *partir* une fois β-réduit ne contiendra qu'un seul λe nécessaire à la manipulation temporelle de l'évènement. *quitter* est un prédicat à trois arguments, respectivement un évènement nommé e, un agent, x et une source (argument spatial) y, de la même manière on remarque les deux sn attendu d'abord à droite pour la source puis à gauche pour l'agent.

Le typage du lexique permet de vérifier de manière stricte la bonne formation de la représentation mais ne permet pas d'interpréter certaines expressions du discours parfois plus souples de ce point de vue, c'est pourquoi nous proposons une solution dans la suite des travaux de Pustejovsky (Pustejovsky, 1995) et d'autres issus de la même tradition.

4.2 Un raffinement de la sémantique lexicale : sortes et types variables

Dans un travail initié par Bassac, Rétoré et Mery (Bassac *et al.*, 2010) dédié à la partie sémantique lexicale de l'analyse, et assez proche en surface à la réinterprétation de Markus Egg (Egg, 2002), une nouvelle organisation du lexique a été proposée afin de traiter, dans la sémantique formelle compositionnelle, les glissements de sens, l'accès aux différentes facettes du sens, et la possible coprédication. On peut observer ces phénomènes sur les exemples suivants :

(3) ** La chaise aboie.* Ce genre de composition impossible est rejetée par un système de types plus riche : *aboie* a pour argument un *chien* à la rigueur un *humain* mais jamais un *meuble*.

(4) *Ce livre est volumineux mais intéressant.* Coprédication correcte entre les deux facettes de *livre* : contenu informationnel et objet physique.

(5) ** Grenoble a battu Dax et prévoit d'acquérir pour 474 500 euros d'œuvres d'art.* Coprédication impossible entre le club de rugby et la municipalité.

Pour ces phénomènes courant dans notre corpus, nous avons conçu une structure de lexique et un algorithme qui permettent de calculer les représentations sémantiques de telles phrases, de rendre compte des coprédications correctes (4) d'échouer lorsqu'elles ne le sont pas (5) et de quantifier correctement. Les types de base des λ-termes, qui sont les sortes d'une logique multisorte, servent à éviter les compositions impossibles comme 3. En cas de conflits de type, il est parfois licite de relaxer le typage, ou d'accéder à l'une des facette du sens d'un mot : une ville peut être considérée comme sa municipalité ou son club de rugby, un livre est à la fois un objet physique et informationnel, etc. Pour ce faire, le lexique associe à chaque mot le λ-terme usuel ainsi que plusieurs λ-termes permettant de changer le type du mot, et certaines transformations, comme celle de la ville en club sont déclarées irréversibles, ce qui permet de prédire l'impossibilité de certaines coprédications. Ce genre de phénomènes nécessite, en particulier pour la conjonction à l'œuvre dans les phénomènes de coprédication, des opérations uniformes sur les types, et c'est pour cela que nous nous sommes placés dans le système F (Girard, 1971). Ce formalisme nous permet de manipuler plus finement les types, de quantifier sur eux, d'utiliser la coercion afin de résoudre les cas de coprédication par exemple. Cette organisation du lexique permet de traiter bien des phénomènes, de sémantique lexicale mais aussi de sémantique compositionnelle.

Les conflits se présentent alors sous la forme $(\lambda x^A.u)w^W$: un terme de type A est attendu par la fonction $(\lambda x^A.u)$ mais l'argument fourni est de type W. Pour résoudre ces conflits, lorsque cela conduit à des interprétations licites, on utilise les λ-termes optionnels du lexique qui donnent au mot le type correspondant au sens adéquat, de l'une des deux manières suivantes :

Transformation rigide correspondant aux transformations irréversibles incompatibles avec les autres transformations. Le lexique fournit, pour un mot de u ou pour un mot de w un λ-terme g de type $W \to A$: le terme se résout en $(\lambda x^A.u)(gw)^A$.

Transformation flexible correspondant aux transformations compatibles avec les autres transformations. Les diverses occurrences de x^A dans u sont utilisées avec des types différents $A_1, \ldots, A_n$: on peut utiliser, si le lexique en fournit, des termes différents de types $g_i : W \to A_i$ pour chaque occurrence de x et remplacer comme le veut la β-réduction chaque occurrence de x par $(g_i(w)) : A_i$.

En sémantique lexicale, ce modèle nous permet même de traiter de constructions assez subtiles, comme le *voyageur fictif* où un *chemin* introduit un *voyageur* qui le suivrait.

(6) *Pendant deux heures le chemin descend.(On notera que c'est le circonstanciel qui oblige à considérer un voyageur fictif sinon cela ne serait pas nécessaire.)*

Là encore, c'est le conflit de type : $\left(P^{humain \to \mathbf{t}}\left(u^{chemin}\right)\right)$ $\quad humain \neq chemin$ qui déclenche l'utilisation du λ-terme optionnel associé à chemin : celui-ci transforme la route en un événement, dont l'agent un voyageur fictif (voir (Moot *et al.*, 2011)).

Cette proposition aborde aussi deux questions classiques de sémantique formelle, dont le traitement s'intègrent dans la même proposition, la quantification généralisée et les pluriels : il s'agit ici de construire les formules logiques associées à certains énoncés et non de déterminer leurs conditions de vérité dans tel ou tel modèle.

Voici tout d'abord un exemple de pluriels correspondant à un quantificateur généralisé :

(7) En effet, on est ici voisin de Toulouse ; comme le caractère, le type est nouveau. Les jeunes filles ont des figures fines, régulières, d'une coupe nette, d'une expression vive et gaie. Elles sont petites, elles ont la démarche légère, des yeux brillants, la prestesse d'un oiseau.

Ici, définir les filles de la région comme ayant des "figures régulières", étant "petites" ou encore ayant la "prestesse d'un oiseau" est une comparaison sous entendue à une fille prototypique de

cet ensemble de filles. On conceptualise facilement une idée de la taille comme étant normale pour un spécimen du type "fille". Ainsi il nous faut un opérateur pouvant sélectionner toutes les propriétés telle que la taille d'un type particulier afin de pouvoir l'associer au spécimen de ce type, quelque soit le type concerné. En premier lieu, dans ce système nous rappelons qu'il n'existe qu'un quantificateur peu importe la classe d'objet sur laquelle on quantifie, ce qui permet de quantifier sur tous les ordres. Au lieu d'avoir une constante $\forall_\alpha$ de type $(\alpha \rightarrow \mathbf{t}) \rightarrow \mathbf{t}$ pour chaque type α sur lesquels on voudrait quantifier, le quantificateur est donc $\forall$ de type $(\alpha \rightarrow \mathbf{t}) \rightarrow \mathbf{t}$ pour tout type α et il sera ensuite spécialisé au type désiré. La quantification généralisée "la plupart des" ou "les" est prise en charge par une constante $\measuredangle$, à rapprocher du $\tau x.A$ d'Hilbert : étant donné un type α, notre constante $\measuredangle$ renvoie le spécimen du type α — pour plus de détails voir (Retoré, 2012).

Notre deuxième exemple concerne toujours les pluriels, mais lorsqu'on prédique une propriété d'un ensemble d'individus, ce qui suscite plusieurs interprétations :

(8) Edgar et son guide descendaient toujours ensemble !... Enfin, le groupe allait se briser sur une saillie de roc effrayante, quand Vincent se précipita avec intrépidité au-devant d'eux, enfonçant par un coup désespéré sa hache tout entière dans la neige...

Cet exemple du corpus permet d'observer un phénomène bien connu, où un ensemble d'entités agit collectivement ou au contraire réunit des entités agissant individuellement. Ainsi on peut comprendre dans "le groupe allait se briser sur une saillie de roc effrayante" que la chute sépare les deux individus qui composait le groupe, dans ce cas c'est le groupe qui se brise ou encore que chacun d'entre eux subit les dommages de l'accident, et alors ce sont les individus appartenant au groupe qui sont brisés. Suivant l'interprétation choisie, la valeur de vérité sera vraie pour l'un et fausse pour l'autre. Notre système à sortes et types variables permet de gérer cette difficulté et les deux interprétations grâce à la constante de distributivité présentée ici.

La constante * : $\lambda P^{\alpha \rightarrow t} \lambda Q^{\alpha \rightarrow t} \forall x^{\alpha}.Q(x) \Rightarrow P(x)$ qui peut être spécialisée à n'importe quel type α permet une distributivité de la propriété sur les membres de l'ensemble. Les détails formels concernant cette constante et d'autres gérant la coercion et la distributivité stricte sont décrits dans (Moot et Retoré, 2011)

Cette technique permet de respecter le principe selon lequel la syntaxe guide la composition sémantique de l'énoncé. Ici, le raffinement lexical permet de filtrer les interprétations impossibles et résoudre ces conflits lorsque le lexique le permet : ainsi on rejette les interprétations erronées sans rejeter d'interprétations obtenues pas des glissements de sens ou l'accès à des facettes du sens. Ces mécanismes ont été ajoutés à la sémantique de Grail en λ-DRT — cette dernière a été étendue au λ-calcul du second ordre. L'implantation — sur de petits lexiques — a été réalisée par Emeric Kien (Kien, 2010) et elle est actuellement poursuivie par Samira Kherfellah.

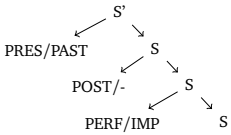
4.3 Le traitement temporel

La temporalité en DRT est traditionnellement (Partee, 1984; Kamp et Reyle, 1993) traitée par des relations entre constantes inspirées des constantes de Reichenbach. Ces modélisations du temps des évènements sont interprétées par des intervalles et des points. Nous nous sommes demandés dans quelle mesure on pouvait retrouver le principe de compositionnalité que Reichenbach ne réussit pas à conserver intégralement. Rappelons simplement les unités utilisées sur l'axe du temps : le point d'énonciation, l'intervalle de l'évènement et le point de repère en cours (respecti-

vement *S*,*E* et *R*) (Reichenbach, 1948). Tout d'abord dans (Verkuyl, 2003), puis dans (Verkuyl, 2008), l'auteur propose une approche complètement compositionnelle, des combinaisons de trois choix entre deux λ-termes complémentaires dont on donne l'abbréviation :
– PAST / PRESENT
– SYNCHRONOUS / POSTERIOR
– PERFECTIF / IMPERFECTIVE
Chaque terme combiné aux deux autres permet de traiter les huit temps verbaux néerlandais à l'indicatif ainsi que les huit temps verbaux anglais correspondant. Le français dispose de six temps supplémentaires dans le mode indicatif dont le passé simple, le passé antérieur et les formes surcomposées du passé et du futur. Toutes ces formes sont traitées par le système. Seule la représentation de la sémantique temporelle de l'évènement est mise en évidence par chacun des termes, n'est pas traité la classe aspectuelle par ces combinaisons. Ils nous permet de procéder à l'analyse temporelle des évènements de manière compositionnelle et conforme à notre analyse en λ-DRT. Nous avons construit la grammaire décrite dans (Verkuyl, 2003) puis nous l'avons intégré dans notre système.

Définissons notre langage inspiré des recommandations de Verkuyl pour traiter de la temporalité des évènements, ce langage étant un sous langage des relations de Allen, il se définit comme ceci : $\langle$ n,i_1,..., i_x, χ, $\circ$, $<$,=, $\subseteq$, $\supset\subset$ $\rangle$
n est une constante
i_1,..., i_x sont des variables
Les lambda termes des opérateurs sont définis et s'appliquent comme suit :

$$\begin{array}{rll} \text{PRES} & =_{def} & \lambda\phi\exists i(\phi\ i)\wedge(\chi i)\circ n \\ \text{PAST} & =_{def} & \lambda\phi\exists i(\phi\ i)\wedge(\chi i)< n \\ \text{POST} & =_{def} & \lambda\phi\lambda i\exists j(\phi\ j)\wedge(\chi i)<(\chi j) \\ \text{PERF} & =_{def} & \lambda\phi\lambda j\exists k(\phi\ k)\wedge(\chi k)<(\chi j) \\ \text{IMP} & =_{def} & \lambda\phi\lambda j\exists k(\phi\ k)\wedge(\chi j)\subseteq(\chi k) \end{array}$$



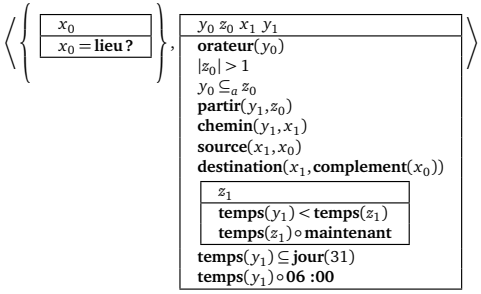
On applique le λ-terme du prédicat évènementiel muni de ses arguments au λ-terme de l'opérateur comme pour n'importe quelle autre unité du lexique. Les adverbes temporels quant à eux doivent intervenir soit avant tout opérateur et donc au plus près du noyau prédicatif, soit entre PERF/IMP et POST/ –. Concernant le typage de ces opérateurs, il est définit sur les valeurs **i**, le type accordé à l'indice chez Verkuyl, qui dans notre cas est le type de l'intervalle [3], et sur la valeur **t** pour la valeur de vérité. Le type des opérateurs sera donc : $\mathbf{i}\rightarrow\mathbf{t}\rightarrow\mathbf{t}$ auquel on applique le prédicat évènementiel avec ses arguments, lui même de type : $\mathbf{i}\rightarrow\mathbf{t}$.
Intuitivement, le n de Verkuyl joue un rôle similaire au S de Reichenbach, le i semble être un R pour le PAST, mais c'est plutôt le j pour le PERF. Une formule de ce langage est donc formée de ces variables, en relation ou non avec la constante, ce qui permet de construire les termes pour les adverbes temporels et pour la temporalité attachée au verbe, on construira la combinaison de plusieurs opérateurs tel que PAST(POST), PRES(POST(PERF)), etc.

Le modèle dans lequel on veut interpréter ce langage est l'ensemble des intervalles tel que l'a définit (Allen, 1983), on donne ici la traduction des relations :

$$\begin{array}{l|l|l|l} \|<\| = \{<,\text{m}\} & \|\subseteq\| = \{\text{s,d,f,=}\} & \|=\| = \{=\} & \text{On remarque que la relation } y\circ n \\ \|>\| = \{>,\text{mi}\} & \|\supseteq\| = \{\text{si,di,fi,=}\} & \|\supset\subset\| = \{\text{m}\} & \text{veut dire } y \text{ di } n. \end{array}$$

3. C'est pourquoi nous avons la fonction χ, soit **temps**( ) dans les DRS prenant un évènement et renvoyant un intervalle dans le lexique, on ne s'étendra pas sur les adverbes qui peuvent affecter deux indices au lieu d'un seul.

Penchons nous désormais sur notre exemple 1 : *Le 31, nous sommes partis à six heures du matin.*



(9)

Pour ce premier exemple, on observe plusieurs choses, tout d'abord, du point de vue spatial, tous les verbes de déplacement doivent être envisagés comme accompagnés d'un chemin défini par une *source* et une *destination*, tous deux liés à la temporalité de l'évènement. La source étant le lieu dans lequel se situe le voyageur au début de l'évènement et la destination, le lieu occupé à la fin de l'évènement de déplacement.

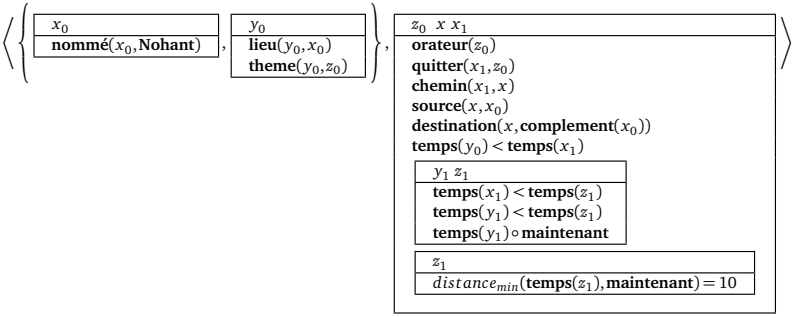
Il y a par ailleurs quelques présuppositions qu'il faut éclairer. On peut inférer de *quitter x* qu'avant de quitter, le voyageur est dans le lieu désigné par x. Par ailleurs, *complément (x)*, argument de *destination* dans le cas de quitter, tout comme dans le cas de partir dans l'exemple ci-après, sera donc la région de l'espace dans laquelle on se trouve une fois avoir quitté x ou être parti de x. On ne peut pas en déduire une destination à proprement dit mais pour le moins on peut désigner par *complément (x)* l'extérieur de x. Nous notons une seconde remarque concernant les présuppositions spatiales profondes nécessaires à l'expression de *quitter* et *partir* selon laquelle, elles résistent aux épreuves de la modalisation induite par *désirer* et de la négation :

(10) Après avoir visité aussi le Vignemale et le Pic du Midi de Bigorre, je désirai ne point quitter les Pyrénées sans avoir fait du moins un effort en faveur de l'ascension de la Maladetta.

Cet exemple ne peut être interprété que comme l'énoncé de quelqu'un situé dans les Pyrénées au moment de l'évènement dénoté par *désirer*.

Il nous faut aussi interroger la nature même de *partir*, considère-t-on la finalité de partir comme le fait de n'être plus là ou bien dans le fait d'être dans le mouvement du départ ? On réfère ici à la distinction entre *accomplishment* et *achievement* de Vendler (Vendler, 1967), typiquement si on dit *je suis presque parti* alors on a affaire à un *achievement* car l'événement sera réalisé selon la condition qu'on ne soit plus dans le lieu d'origine. Si on dit *je partais lorsqu'elle s'est adressée à moi* ici l'essence de l'évènement peut durer et donc être considéré comme un *accomplishment*.

Au sujet d'*à six heures du matin*, la relation ∘ choisie entre l'intervalle *06 :00* et l'intervalle y_1 est plutôt faible et on ne peut véritablement rien inférer d'autre au sujet des deux entités en relation. Concernant *Le 31* il a fallut décider si la valeur accordée à ce syntagme adverbial devait être donnée comme étant dans le futur ou dans le passé. En effet, ici l'interprétation dépend entièrement du contexte. On imagine sans peine un exemple utilisant aussi cet item tel que : "*Le 31 je serai aux Galapagos*" qui impliquerait de fait une interprétation dans le futur. Que faire alors des expressions de l'habitude telles que "*Le vendredi je mange du poisson*" ? Est-ce une résolution, ou une habitude de longue date ? A ce sujet nous pensons intégrer à terme un module traitant de la SDRT dans le système afin d'obtenir les informations propres aux relations discursives pouvant résoudre ce problème. Les expressions telles que *le vendredi*, *le 31* méritent que l'on s'interroge davantage sur leur sémantique en contexte.
Regardons de plus près notre second exemple 2 : *Dans dix minutes, j'aurai quitté Nohant.*



(11)

Pour calculer correctement et respecter l'accessibilité des variables dans la représentation, nous avons fait le choix d'imbriquer les DRS propres à PRES, POST et à PERF décrits plus tôt. Plus exactement si l'on décompose le calcul, *(PRES(POST(PERF (quitter e, je, Nohant))))*, on obtient le terme suivant :
$\exists y_1 \exists z_1 \exists x_1 [quitter(x_1, je, Nohant)] \wedge (\chi x_1) < (\chi z_1) \wedge (\chi y_1) < (\chi z_1) \wedge (\chi y_1) \circ n)]$
interprété comme :
– PRES : par rapport au moment d'énonciation n, il existe un repère y_1 qui est en relation d'$\circ$
– POST : par rapport à ce repère, il existe un intervalle postérieur y_1
– PERF : l' intervalle y_1 est lui même postérieur à la fin de l'intervalle de l'évènement x_1
Dans la figure 6, on donne une représentation graphique possible.

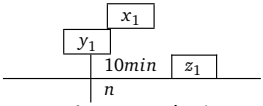


FIGURE 6 – Interprétation possible des variables temporelles pour l'exemple traité.

La transitivité des relations ne nous donne que peu d'informations sur la localisation de l'évènement *quitter*, par exemple x_1 n'a finalement qu'une seule contrainte par rapport à n (maintenant) si ce n'est que tout comme y_1, il doit se situer avant z_1 qui lui même est en relation *di* avec n pour Allen soit $\circ$ ici.

Ce système permet d'exprimer la temporalité des évènements portée par le verbe conjugué ainsi que les modifications que peuvent apporter les adverbes. L'un des atouts de ce système réside dans le fait que l'on ne dit pas plus de choses dans la formule que n'est dit dans le discours. La représentation sémantique de la temporalité de chaque évènement est compositionnelle et propose un contenu précis et approprié au traitement des relations entre les évènements, étape suivante de nos travaux. Les premiers tests opérés semblent prometteurs mais ce travail nécessite un enrichissement du lexique pour les adverbes et des tests à plus grande échelle.

5 Conclusion

Dans cet article nous avons montré les différentes étapes de notre traitement automatique du discours, consistant en l'analyse syntaxique puis en la dérivation sémantique en λ-DRT. L'interface

syntaxe-sémantique dans le cadre de la théorie des types est une base solide permettant de respecter la compositionnalité du sens tout en s'appuyant de l'organisation syntaxique du discours. Le système F quant à lui est approprié pour traiter les phénomènes rencontrés dans le discours et l'interface sémantique-pragmatique justifie un raffinement du lexique par ce système. Pour de futurs travaux, nous envisageons d'enrichir davantage le lexique afin de couvrir plus largement le discours et les phénomènes de glissement de sens ainsi que les modifications temporels. Dans le cadre du projet Itipy, il est nécessaire de développer davantage l'ordonnancement temporel des évènements dans le récit en déterminant les relations appropriées et choisir les composantes temporelles qui doivent être mises en relation. Dans le cadre du genre de discours étudié et de l'objet que nous cherchons à extraire, la classe aspectuelle de chaque prédicat muni de ses arguments doit être déterminée en fonction de son rapport à l'espace afin de relier correctement les évènements qu'ils dénotent. Plus concrètement, nous envisageons en premier lieu de développer une composante permettant la résolution d'anaphores, première étape indispensable à la suite de nos travaux.

Références

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for french. *In Treebanks*. Kluwer.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *In Communications of the ACM*, numéro 26(11), pages 832–843.

Bassac, C., Mery, B. et Retoré, C. (2010). Towards a Type-Theoretical Account of Lexical Semantics. *Journal of Logic Language and Information*, 19(2):229–245.

Bittar, A. (2009). Annotation of events and temporal expressions in french texts. *In The Third Linguistic Annotation Workshop (LAW III) Singapore*.

Busquets, J., Vieu, L. et Asher, N. (2001). La SDRT : Une approche de la cohérence du discours dans la tradition de la sémantique dynamique. *Verbum*, XXIII(1):73–101.

Buszkowski, W. et Penn, G. (1990). Categorial grammars determined from linguistic data by unification. *Studia Logica*, 49:431–454.

Clark, S. et Curran, J. R. (2004). Parsing the WSJ using CCG and log-linear models. *In Proceedings of the 42nd annual meeting of the ACL*, pages 104–111, Barcelona.

Comon, H., Dauchet, M., Jacquemard, F., Lugiez, D., Tison, S. et Tommasi, M. (1997). Tree automata techniques and applications.

Curran, J., Clark, S. et Bos, J. (2007). Linguistically motivated large-scale nlp with c&c and boxer. *In Proceedings of the 45th Annual Meeting of the ACL*, pages 33–36, Prague.

Egg, M. (2002). Semantic contructuction for reinterpretation phenomena. *Linguistics*, 40(3): 579–609.

Girard, J.-Y. (1971). Une extension de l'interprétation de Gödel à l'analyse et son application : l'élimination des coupures dans l'analyse et la théorie des types. *In* Fenstad, J. E., éditeur : *Proceedings of the SLS*, volume 63 de *Studies in Logic and the Foundations of Mathematics*, pages 63–92, Amsterdam. North Holland.

Kamp, H. et Reyle, U. (1993). *From Discourse to Logic*. D. Reidel, Dordrecht.

KIEN, E. (2010). Du sens des mots à l'analyse automatique d'une phrase. Mémoire de stage d'initiation à la recherche, ENS-Cachan & INRIA Bordeaux.

LAMBEK, J. (1958). The mathematics of sentence structure. *American Mathematical Monthly*, 65:154–170.

MAGRI-MOURGUES, V. (2009). *Le voyage à pas comptés. Pour une poétique du récit de voyage au XIXème siècle*. Numéro 9 de Lettres numériques. Honoré Champion.

MOORTGAT, M. (1997). Categorial type logics. *In* van BENTHEM, J. et ter MEULEN, A., éditeurs : *Handbook of Logic and Language*, chapitre 2, pages 93–177. Elsevier/MIT Press.

MOOT, R. (2010a). Semi-automated extraction of a wide-coverage type-logical grammar for French. *In Proceedings of Traitement Automatique des Langues Naturelles (TALN)*, Montreal.

MOOT, R. (2010b). Wide-coverage French syntax and semantics using Grail. *In Proceedings of Traitement Automatique des Langues Naturelles (TALN)*, Montreal.

MOOT, R., PRÉVOT, L. et RETORÉ, C. (2011). Un calcul de termes typés pour la pragmatique lexicale. *In Traitement Automatique du Langage Naturel, TALN 2011*, pages 161–166, Montpellier.

MOOT, R. et RETORÉ, C. (2011). Second order lambda calculus for meaning assembly : on the logical syntax of plurals. *In Coconat 2011*.

MUSKENS, R. (1994). Categorial Grammar and Discourse Representation Theory. *In Proceedings of COLING 94*, pages 508–514, Kyoto.

PARENT, C., GAGNON, M. et MULLER, P. (2008). Annotation d'expressions temporelles et d'évènements en français. *In Traitement automatique des langues naturelles*.

PARTEE, B. H. (1984). Nominal and temporal anaphora. *Linguistics and Philosophy*, 7:243–286.

PUSTEJOVSKY, J. (1995). *The Generative Lexicon*. MIT Press.

REICHENBACH, H. (1948). *Elements of Symbolic Logic*. The Mac millan Company.

RETORÉ, C. (2012). Variable types for meaning assembly : a logical syntax for generic noun phrases introduced by "most". *Recherches linguistiques de Vincennes*, 41:1–18.

SANDILLON-REZER, N.-F. et MOOT, R. (2011). Using tree tranducers for grammatical inference. *Proceedings of Logical Aspects of Computational Linguistics 2011*, pages 233–250.

VENDLER, Z. (1967). *Linguistics in philosophy*. Cornell University Press.

VERKUYL (2003). On the compositionnality of tense : Merging reichenbach and prior. Utrecht University.

VERKUYL, H. (2008). *Binary Tense*. CSLI Publications.

La reconnaissance des mots composés à l'épreuve de l'analyse syntaxique et vice-versa : évaluation de deux stratégies discriminantes

Matthieu Constant[1] Anthony Sigogne[1] Patrick Watrin[2]
(1) université Paris-Est, LIGM, CNRS, 5, bd Descartes 774545 Marne-la-Vallée
(2) Université de Louvain, CENTAL, Louvain-la-Neuve
mconstan@univ-mlv.fr, sigogne@univ-mlv.fr,patrick.watrin@uclouvain.be

RÉSUMÉ

Nous proposons deux stratégies discriminantes d'intégration des mots composés dans un processus réel d'analyse syntaxique : (i) pré-segmentation lexicale avant analyse, (ii) post-segmentation lexicale après analyse au moyen d'un réordonnanceur. Le segmenteur de l'approche (i) se fonde sur un modèle CRF et permet d'obtenir un reconnaisseur de mots composés *état-de-l'art*. Le réordonnanceur de l'approche (ii) repose sur un modèle MaxEnt intégrant des traits dédiés aux mots composés. Nous montrons que les deux approches permettent de combler jusqu'à 18% de l'écart entre un analyseur *baseline* et un analyseur avec segmentation parfaite et jusqu'à 25% pour la reconnaissance des mots composés.

ABSTRACT

Recognition of compound words tested against parsing and vice-versa : evaluation of two discriminative approaches

We propose two discriminative strategies to integrate compound word recognition in a real parsing context : (i) state-of-the-art compound pregrouping with Conditional Random Fields before parsing, (ii) reranking parses with features dedicated to compounds after parsing. We show that these two approaches help reduce up to 18% of the gap between a baseline parser and parser with golden segmentation and up to 25% for compound recognition.

MOTS-CLÉS : Mots composés, analyse syntaxique, champs markoviens aléatoires, réordonnanceur.

KEYWORDS: Multiword expressions, parsing, Conditional random Fields, reranker.

1 Introduction

L'intégration des expressions multi-mots (EMM) dans des applications réelles comme la traduction automatique ou l'extraction d'information est cruciale car de telles expressions ont la particularité de contenir un certain degré de figement. En particulier, elles forment des unités lexicales complexes qui, si elles sont prises en compte, peuvent non seulement améliorer l'analyse syntaxique, mais aussi faciliter les analyses sémantiques qui en découlent. Leur intégration dans un processus d'analyse syntaxique probabiliste a déjà été envisagée dans quelques études. Toutefois, elles reposent pour la majorité sur un corpus au sein duquel l'ensemble des EMMs a

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 57–70,
Grenoble, 4 au 8 juin 2012.

été parfaitement identifié au préalable. Bien qu'artificielles, ces études ont montré une amélioration des performances d'analyse : par exemple, (Nivre et Nilsson, 2004; Eryigit *et al.*, 2011) pour l'analyse en dépendance et (Arun et Keller, 2005; Hogan *et al.*, 2011) pour l'analyse en constituants. Plus récemment, (Green *et al.*, 2011) a intégré la reconnaissance des EMMs au sein de la grammaire et non plus dans une phase préalable. La grammaire est entrainée sur un corpus arboré où les EMMs sont annotées avec des noeuds non-terminaux spécifiques.

Dans cet article, nous nous intéressons à un type d'EMMs : les mots composés. Nous proposons d'évaluer deux stratégies discriminantes d'intégration de ces expressions dans un contexte réel d'analyse syntaxique en constituants : (a) pré-segmentation lexicale au moyen d'un reconnaisseur *état-de-l'art* de mots composés basé sur les champs markoviens aléatoires [CRF] ; (b) analyse basée sur une grammaire incluant l'identification des mots composés, suivie d'une phase de réordonnancement des analyses à l'aide d'un modèle maximum d'entropie intégrant des traits dédiés aux mots composés. (a) est une implémentation réaliste de l'approche classique de pré-regroupement des EMMs. Nous souhaitons évaluer si la reconnaissance automatique des EMMs a toujours un impact positif sur l'analyse syntaxique, c'est-à-dire, si une segmentation lexicale imparfaite ne provoque pas trop d'effets de bord sur les constituants supérieurs. L'approche (b) est innovante pour la reconnaissance des EMMs : nous sélectionnons la segmentation lexicale finale après l'analyse syntaxique afin d'explorer le plus d'analyses possibles (contrairement à la méthode (a)). Cette approche ressemble à celle proposée par (Wehrli *et al.*, 2010) qui reclasse les hypothèses d'analyses générées par un analyseur symbolique en se basant sur la présence ou non de collocations. Les expériences que nous avons menées ont été réalisées sur le corpus arboré de Paris 7 [FTB] (Abeillé *et al.*, 2003) où les mots composés sont marqués. Nous avons utilisé l'analyseur syntaxique de Berkeley (Petrov *et al.*, 2006) qui est fondé sur une stratégie non-lexicalisée et qui obtient les meilleurs résultats en français (Seddah *et al.*, 2009), même en incorporant l'identification des EMMs (Green *et al.*, 2011).

L'article est organisé comme suit : la section 2 présente les problématiques du repérage des EMMs et de leur intégration dans un analyseur syntaxique. La section 3 décrit plus en détail les deux stratégies proposées et les modèles sous-jacents. La section 4 détaille les ressources utilisées pour nos expériences : corpus et lexiques. Nous décrivons ensuite (dans la section 5) l'ensemble des traits dédiés aux EMMs intégrés dans nos deux modèles. Enfin, la section 6 rapporte et analyse les résultats obtenus lors de nos expériences.

2 Les mots composés

Les expressions multi-mots sont des unités lexicales constituées de plusieurs lexèmes qui contiennent un certain degré de figement. Elles couvrent une large gamme de phénomènes linguistiques : les expressions figées et semi-figées, les constructions à verbe support, les entités nommées, les termes, etc. Elles sont souvent divisées en deux classes : les expressions définies au moyen de critères linguistiques et celles définies par des critères purement statistiques (collocations) (Sag *et al.*, 2002). La plupart des critères linguistiques pour déterminer si une combinaison de mots est une EMM sont basés sur des tests syntaxiques et sémantiques comme ceux décrits dans (Gross, 1986). Par exemple, l'expression *caisse noire* est une EMM car elle n'accepte pas de variations lexicales (**caisse sombre*) et elle n'autorise pas d'insertions (**caisse très noire*). De telles expressions définies linguistiquement peuvent coincider en partie avec les collocations

qui constituent des associations habituelles de mots (fondées sur la notion de fréquence). Ces dernières sont souvent identifiées au moyen de mesures statistiques associatives. Dans cet article, nous nous focalisons sur les EMMs continues qui forment des unités lexicales auxquelles on peut associer une partie-du-discours : ex. *tout à fait* est un adverbe, *à cause de* est une préposition, *table ronde* est un nom. Les variations morphologiques et lexicales sont très limitées – e.g. *caisse noire+caisses noires*, *vin (rouge+blanc+rosé+*orange+...)* – et les variations syntaxiques très souvent interdites [1]. De telles expressions sont généralement analysées au niveau lexical. Par la suite, nous utilisons le terme *mot composé* ou *unité polylexicale*.

2.1 Identification

L'identification des EMMs dans les textes est souvent complexe car leur propriété de figement les rend difficilement prédictibles. Elle repose généralement sur des stratégies lexicalisées. La plus simple est fondée sur la consultation de lexiques comme dans (Silberztein, 2000). Le plus grand désavantage est que cette procédure se base entièrement sur des dictionnaires et est donc incapable de découvrir de nouveaux mots composés. L'utilisation d'extracteurs de collocations peut donc s'avérer utile. Par exemple, (Watrin et François, 2011) calcule à la volée pour chaque collocation candidate dans le texte traité, son score d'association au moyen d'une base externe de n-grammes apprises sur un grand corpus brut. L'expression est ensuite étiquetée comme EMM si son score d'association est plus grand qu'un seuil donné. Ils obtiennent d'excellents résultats dans le cadre d'une tâche d'extraction de mots-clés. Dans le cadre d'une évaluation sur corpus de référence, (Ramisch *et al.*, 2010) a développé un classifieur basé sur un séparateur à vastes marges intégrant des traits correspondant à différentes mesures d'associations des collocations. Les résultats sont plutôt faibles sur le corpus GENIA. (Green *et al.*, 2011) a confirmé ces mauvais résultats sur le corpus arboré de Paris 7. Ceci s'explique par le fait que de telles méthodes ne font aucune distinction entre les différents types de EMMs et que les types de EMMs annotés dans les corpus sont souvent limités. Une approche récente consiste à coupler, dans un même "modèle", l'annotation des mots composés avec un analyseur linguistique : un étiqueteur morphosyntaxique dans (Constant *et al.*, 2011) et un analyseur syntaxique dans (Green *et al.*, 2011). (Constant *et al.*, 2011) apprend un modèle CRF intégrant différents traits classiques de l'étiquetage morphosyntaxique et des traits basés sur des ressources externes. (Green *et al.*, 2011) a proposé que l'identification des mots composés soit intégrée dans la grammaire de l'analyseur, qui est apprise sur un corpus arboré où les mots composés sont annotés au moyen de noeuds non-terminaux spécifiques. Ils ont utilisé, avec succès, une grammaire à substitution d'arbres au lieu d'une grammaire probabiliste indépendante du contexte (avec annotations latentes) afin d'apprendre des règles lexicalisées. Les deux méthodes ont l'avantage d'être capables d'apprendre de nouveaux mots composés. Dans cet article, nous exploitons les avantages des méthodes décrites dans cette section en les intégrant comme traits d'un unique modèle probabiliste.

2.2 Intégration dans l'analyse syntaxique

La majorité des expériences d'intégration des EMMs dans un processus d'analyse syntaxique repose sur des corpus au sein desquels les mots composés ont été parfaitement identifiés au

1. De telles expressions acceptent très rarement des insertions, souvent limitées à des modifieurs simples e.g. *à court terme*, *à très court terme*.

préalable. Bien qu'artificielles, ces études ont montré une amélioration des performances d'analyse : par exemple, (Nivre et Nilsson, 2004; Eryigit *et al.*, 2011) pour l'analyse en dépendance et (Arun et Keller, 2005; Hogan *et al.*, 2011) pour l'analyse en constituants. Pour l'analyse en constituants, nous pouvons noter les expériences de (Cafferkey *et al.*, 2007) qui ont essayé de coupler des annotateurs réels de EMMs et différents types d'analyseurs probabilistes pour l'anglais. Ils ont travaillé sur un corpus de référence non annoté en EMMs. Les EMMs sont reconnues et pré-groupées automatiquement à l'aide de ressources externes et d'un reconnaisseur d'entités nommées. Ils appliquent, ensuite, un analyseur syntaxique et réinsèrent finalement les sous-arbres correspondants aux EMMs pour faire l'évaluation. Ils ont montré des gains faibles mais significatifs. Récemment, les travaux de (Finkel et Manning, 2009) et (Green *et al.*, 2011) ont proposé d'intégrer les deux tâches dans le même modèle. (Finkel et Manning, 2009) couple analyse syntaxique et reconnaissance des entités nommées dans un modèle discriminant d'analyse syntaxique basé sur les CRF. (Green *et al.*, 2011) a intégré l'identification des mots composés dans la grammaire. Ils ont, en particulier, montré, pour le français, que le meilleur analyseur syntaxique était toujours l'analyseur de Berkeley (fondé sur une stratégie non-lexicalisée), bien que l'identification des mots composés soit moins bonne qu'avec un analyseur syntaxique fondé sur une stratégie lexicalisée. Enfin, il existe les travaux de (Wehrli *et al.*, 2010) qui reclasse les hypothèses d'analyses générées par un analyseur symbolique en se basant sur la présence ou non de collocations.

3 Modèles discriminants

Nous considérons deux stratégies d'intégration des mots composés dans le processus d'analyse syntaxique : (a) une pré-identification des mots composés, suivie d'une analyse ; et (b) une analyse syntaxique incorporant l'identification des mots composés suivie d'un réordonnanceur intégrant des traits dédiés aux EMMs.

3.1 Pré-identification des mots composés

La reconnaissance de mots composés peut être vue comme une tâche d'annotation séquentielle si l'on utilise le schéma d'annotation IOB (Ramshaw et Marcus, 1995). Ceci implique une limitation théorique : les mots composés doivent être continus. Ce schéma est donc théoriquement plus faible que celui proposé par (Green *et al.*, 2011) qui intègre les mots composés dans la grammaire et autorise des unités polylexicales discontinues. Cependant, en pratique, les mots composés sont très très rarement discontinus et dans la majorité des cas, la discontinuité est due à l'insertion d'un simple modifieur qui peut être incorporé dans la séquence figée : *à court terme*, *à très court terme*. Dans cet article, nous proposons d'associer les composants simples des unités polylexicales à une étiquette de la forme CAT+X où CAT est la catégorie grammaticale du mot composé et X détermine la position relative du token dans le mot composé (soit B pour le début – Beginning–, soit I pour les autres positions –Inside–). Les mots simples sont étiquetés O (outside) : *Jean/O adore/O les/O faits/N+B divers/N+I*.

Pour cette tâche, nous utilisons le modèle des champs aléatoires markoviens linéaires (Tellier et Tommasi, 2011) [CRF] introduits par (Lafferty *et al.*, 2001) pour l'annotation de séquences.

Etant donné une séquence de mots (graphiques) [2] en entrée $x = (x_1, x_2, ..., x_N)$ et une séquence d'étiquettes en sortie $y = (y_1, y_2, ..., y_N)$, le modèle est défini comme suit :

$$P_\lambda(y|x) = \frac{1}{Z(x)}.\sum_{t}^{N}\sum_{k}^{K} log\,\lambda_k.f_k(t, y_t, y_{t-1}, x)$$

où $Z(x)$ est un vecteur de normalisation dépendant de x. Il est basé sur K traits définis par des fonctions binaires f_k dépendant de la position courante t dans x, l'étiquette courante y_t, l'étiquette précédente y_{t-1} et toute la séquence en entrée. Chaque mot x_i de x intègre non seulement sa propre valeur lexicale mais aussi toutes les propriétés du mot (e.g. ses suffixes, ses étiquettes dans un lexique externe, il commence par une majuscule, etc.). Les traits sont activés si une configuration particulière entre t, y_t, y_{t-1} and x est satisfaite (i.e. $f_k(t, y_t, y_{t-1}, x) = 1$). Chaque trait est associé à un poids λ_k. Ces poids sont les paramètres du modèle et sont estimés lors de la phase d'apprentissage. Les traits utilisés pour notre tâche sont décrits dans la section 5. Ils sont générés à partir de patrons qui sont instanciés à chaque position dans la séquence à annoter. Chaque instance x correspond à une fonction binaire f_k qui retourne 1 si l'instance à la position courante correspond à x, 0 sinon.

3.2 Réordonnanceur

Le réordonnancement discriminant consiste à reclasser les n meilleures analyses produites par un analyseur syntaxique de base et à sélectionner la meilleure. Il utilise un modèle discriminant intégrant des traits associés aux noeuds des analyses candidates. (Charniak et Johnson, 2005) a introduit différents traits qui permettent d'améliorer sensiblement les performances d'un analyseur syntaxique. Formellement, étant donné une phrase s, le réordonnanceur sélectionne la meilleure analyse candidate p parmi l'ensemble de tous les candidats $P(s)$ à l'aide d'une fonction de score V_θ :

$$p^* = argmax_{p \in P(s)} V_\theta(p)$$

L'ensemble des candidats $P(s)$ correspond aux n meilleures analyses produites par l'analyseur de base. La fonction de score V_θ est le produit scalaire d'un vecteur de paramètres θ et d'un vecteur de traits f :

$$V_\theta(p) = \theta.f(p) = \sum_{j=1}^{m} \theta_j.f_j(p)$$

où $f_j(p)$ correspond au nombre d'occurences du trait f_j dans l'analyse p. Selon (Charniak et Johnson, 2005), le premier trait f_1 est la probabilité de p fournie par l'analyseur de base. Le vecteur θ est estimé lors de la phase d'apprentissage à partir du corpus arboré de référence et des analyses générées par l'analyseur de base.

Dans cet article, l'utilisation du réordonnanceur est légèrement modifiée par rapport à ce qui se fait traditionnellement. En effet, nous y intégrons des traits chargés d'améliorer la reconnaissance

2. Un mot (graphique) correspond à un token.

des mots composés dans le contexte de l'analyse syntaxique. Ces traits sont décrits dans la section 5 au moyen de patrons qui sont instanciés pour chaque noeud des analyses. L'apprentissage du modèle est réalisé à l'aide de l'algorithme de maximum d'entropie utilisé dans (Charniak et Johnson, 2005).

4 Ressources

4.1 Corpus

Le corpus arboré de Paris 7 [3] [FTB] (Abeillé *et al.*, 2003) est un corpus annoté en constituants syntaxiques. Il est composé d'articles provenant du journal *Le Monde*. Nous avons utilisé la version la plus récente, celle de juin 2010. Elle comporte 15 917 phrases et 473 904 mots graphiques, et utilise 13 catégories syntaxiques pour identifier les constituants. Les mots composés sont marqués et forment au total plus de 5% des unités lexicales (mots simples et composés). Nous avons réalisé nos expériences sur deux instances différentes provenant de cette même version : l'instance issue du prétraitement décrit dans (Green *et al.*, 2011) [FTB-STF] et l'instance issue du prétraitement réalisé par la chaîne de traitement de l'équipe Alpage de Paris 7 [FTB-P7]. FTB-STF possède un jeu de 14 étiquettes morphosyntaxiques et a été utilisé pour avoir des résultats comparables avec (Green *et al.*, 2011) en terme d'identification des mots composés. Les mots composés sont défaits et annotés à l'aide d'un symbole non-terminal spécifique "MWX" où X est la catégorie grammaticale de l'expression. Ils ont une structure plate comme dans la figure 1. Il existe 11 symboles de type EMM. FTB-P7 possède un jeu de 28 étiquettes morphosyntaxiques optimisé pour l'analyse syntaxique et donc très adéquat pour nos expériences. Les composants simples de chaque mot composé sont fusionnés en un seul mot. Pour pouvoir réaliser nos expériences, il a été nécessaire de défaire tous les mots composés et les représenter comme dans l'instance FTB-STF. Les étiquettes morphosyntaxiques des composants simples des unités polylexicales ont été ajoutées à l'aide de l'étiqueteur morphosyntaxique *lgtagger* (Constant et Sigogne, 2011) appris sur la version du FTB où les mots composés ne sont pas défaits. Le partionnement *entraînement/développement/test* correspond au partitionnement officiel : les sections *développement* et *test* sont les mêmes que dans (Candito et Crabbé, 2009), avec 1 235 phrases chacune. La section entraînement comporte 13 347 phrases, soit 3 390 phases en plus que la version généralement utilisée.

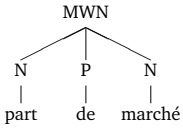


FIGURE 1 – Représentation des mots composés *part de marché* : le noeud MWN correspond à un nom composé ; il a une structure interne plate N P N (nom – préposition – nom)

3. http ://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php

4.2 Ressources lexicales

Il existe de nombreuses resources morphologiques en français incluant les mots composés. Nous avons exploité deux dictionnaires de langue générale : le Dela (Courtois, 2009; Courtois *et al.*, 1997) et le Lefff (Sagot, 2010). Le Dela a été manuellement développé dans les années 80-90 par l'équipe de linguistes du LADL à Paris 7. Nous utilisons la version libre intégrée à la plateforme Unitex[4]. Il est composé de 840,813 entrées lexicales, incluant 104,350 entrées composées (dont 91,030 noms). Les mots composés présents dans la ressource respectent, en général, les critères syntaxiques définis dans (Gross, 1986). Le Lefff[5] est une ressource lexicale qui a été accumulée automatiquement à partir de diverses sources et qui a ensuite été validée manuellement. Nous avons utilisé la version se trouvant dans MeLT (Denis et Sagot, 2009). Elle comprend 553,138 entrées lexicales, incluant 26,311 entrées composées (dont 22,673 noms). Leurs différents modes de construction rendent ces deux ressources complémentaires. Pour toutes les deux, les entrées lexicales possèdent une forme fléchie, un lemme et une catégorie grammaticale. Le Dela possède un trait supplémentaire pour la plupart des mots composés : leur struture interne. Par exemple, *eau de vie* a le code `NDN` car sa structure interne est de la forme nom – préposition *de* – nom.

En terme de collocations, (Watrin et François, 2011) a présenté un système retournant, pour toute phrase, la liste des collocations nominales potentielles accompagnées de leur mesure d'association. Pour le FTB, nous obtenons 17,315 collocations nominales candidates associées à leur log-vraisemblance et leur structure interne.

5 Les traits dédiés aux mots composés

Les deux modèles décrits dans la section 3 nécessitent des traits dédiés aux mots composés. Les traits que nous proposons sont générés à partir de patrons. Dans le but de rendre ces modèles comparables, nous avons mis en place deux jeux comparables de patrons de traits inspirés de (Constant *et al.*, 2011) : l'un adapté à l'annotation séquentielle et l'autre adapté au réordonnancement. Les patrons pour l'annotation séquentielle sont instanciés à chaque position de la séquence en entrée. Les patrons pour le réordonnanceur sont seulement instanciés aux feuilles des analyses candidates, qui sont dominées par un noeud de type EMM (c'est-à-dire qui ont un ancêtre de type EMM). Nous définissons un patron T comme suit :

– Annotation séquentielle : à chaque position n dans la séquence en entrée x,

$$T = f(x,n)/y_n$$

– Réordonnancement : à chaque feuille (à la position n) dominée par un noeud de type EMM m dans l'analyse candidate p,

$$T = f(p,n)/label(m)/pos(p,n)$$

où f est une fonction à définir ; y_n est une étiquette de sortie à la position n ; $label(m)$ est l'étiquette du noeud m et $pos(p,n)$ indique la position relative, dans l'unité polylexicale, du mot à l'indice n : `B` (position initiale), `I` (autres positions).

4. http ://igm.univ-mlv.fr/~unitex
5. http ://atoll.inria.fr/~sagot/lefff.html

5.1 Traits endogènes

Les traits endogènes sont des traits extraits directement des mots eux-mêmes ou d'un outil appris sur le corpus d'apprentissage comme un étiqueteur morphosyntaxique.

n-grammes de mots. Nous utilisons les bigrammes et unigrammes de mots pour apprendre les mots composés présents dans le corpus d'entrainement et pour extraire des indices lexicaux afin d'en découvrir de nouveaux. Par exemple, le bigramme *coup de* est souvent le préfixe d'unités polylexicales comme *coup de pied*, *coup de foudre*, *coup de main*, etc.

n-grammes d'étiquettes morphosyntaxiques. Nous utilisons les unigrammes et bigrammes d'étiquettes morphosyntaxiques dans le but d'apprendre des structures syntaxiques irrégulières souvent caractéristiques de présence de mots composés. Par exemple, la séquence *préposition – adverbe* associée à l'adverbe composé *depuis peu* est très inhabituelle. Nous avons aussi intégré des bigrammes mélangeant mots et étiquettes morphosyntaxiques.

Traits spécifiques. Chaque type de modèle intègre des traits particuliers car chacun s'attèle à des tâches différentes. On incorpore dans le CRF des traits spécifiques pour gérer les mots inconnus et les mots spéciaux (nombres, traits d'union, etc.) : le mot en lettres minuscules ; les préfixes et suffixes de taille 1 à 4, l'information si un mot commence par une majuscule, s'il contient un chiffre, si c'est un trait d'union. Nous ajoutons en plus les bigrammes des étiquettes de sortie. Les modèles liés au réordonnanceur intègrent des traits associés aux noeuds de type EMM, dont les valeurs sont les formes lexicales des mots composés correspondants.

5.2 Traits exogènes.

Les traits exogènes sont des traits qui proviennent totalement ou en partie de données externes (dans notre cas, nos ressources lexicales). Les ressources lexicales peuvent être utiles pour découvrir de nouvelles expressions. Généralement, les mots composés, en particulier les noms, suivent un schéma régulier, ce qui les rend très difficilement repérables à partir de traits endogènes uniquement. Nos ressources lexicales sont appliquées au corpus à l'aide d'une analyse lexicale qui produit, pour chaque phrase, un automate fini qui représente l'ensemble des analyses possibles. Les traits exogènes sont calculés à partir de cet automate.

Les traits basés sur un lexique. Nous associons à chaque mot l'ensemble des étiquettes morphosyntaxiques trouvées dans notre lexique morphologique externe. Cet ensemble forme une classe d'ambiguité ac. Si un mot appartient potentiellement à une unité polylexicale dans son contexte d'occurrence, l'étiquette correspondante au mot composé est aussi intégrée à la classe d'ambiguité. Par exemple, le mot *de* dans le contexte *eau de vie* est associé à la classe *det_nom+I_prep*. En effet, le mot simple *de* est soit un déterminant (det) soit une préposition (prep). Par ailleurs, il se trouve dans une position interne (I) du nom *eau de vie*. Ce trait est directement calculé à partir de l'automate généré par l'analyse lexicale. Nous utilisons également cet automate afin de réaliser une segmentation lexicale préliminaire en appliquant un algorithme du plus court chemin pour favoriser les analyses polylexicales. Cette segmentation préliminaire est source d'indices pour la segmentation finale, donc source de nouveaux traits. On peut associer à tout mot appartenant à un segment composé différentes propriétés : l'étiquette morphosyntaxique mwt du segment, ainsi que sa structure interne mws ; sa position relative $mwpos$ dans le segment ('B' ou 'I').

Traits basés sur les collocations. Dans notre ressource de collocations, chaque candidat du FTB est accompagné de sa structure syntaxique interne et de son score d'association (log-vraisemblance). Nous avons divisé ces candidats en deux classes : ceux qui ont un score supérieur à un certain seuil et les autres. Ainsi, tout mot du corpus peut être associé à un certain nombre de propriétés s'il appartient à une collocation candidate : la classe de la collocation c ainsi que sa structure interne cs, la position relative $cpos$ du mot dans la collocation ('B' ou 'I'). Nous avons manuellement fixé le seuil à une valeur de 150 après une phase de réglage sur le corpus de développement.

Tous les patrons de traits sont donnés dans la table 1.

Traits endogènes
$w(n+i)$, $i \in \{-2,-1,0,1,2\}$
$w(n+i)/w(n+i+1)$, $i \in \{-2,-1,0,1\}$
$t(n+i)$, $i \in \{-2,-1,0,1,2\}$
$t(n+i)/t(n+i+1)$, $i \in \{-2,-1,0,1\}$
$w(n+i)/t(n+j)$, $(i,j) \in \{(1,0),(0,1),(-1,0),(0,-1)\}$
Traits exogènes
ac(n)
mwt(n)/mwpos(n)
mws(n)/mwpos(n)
c(n)/cs(n)/cpos(n)

TABLE 1 – Les patrons de traits utilisés à la fois dans l'annotateur séquentiel et le réordonnanceur (n est la position courante dans la phrase) : ils correspondent à la fonction f.

6 Evaluation

6.1 Préliminaires

L'ensemble des expériences décrites ci-dessous ont été réalisées avec l'analyseur syntaxique de Berkeley [6]. Nous notons BKYc l'analyseur dont la grammaire [7] a été apprise sur le FTB où les mots composés sont fusionnés ; BKY l'analyseur dont la grammaire a été apprise sur le FTB où les mots composés sont défaits et annotés par un symbole non-terminal spécial.

Les expériences sont évaluées à l'aide de plusieurs mesures classiques : la F-mesure [F], la mesure UAS (*Unlabeled Attachment Score*) et la mesure LA (*Leaf Ancestors*). F [8] prend en compte le parenthésage et l'étiquetage des noeuds. Le score UAS [9] évalue la qualité des liens de dépendance non typés entre les mots. Finalement, la mesure LA [10] (Sampson et Babarczy, 2003) calcule la similarité entre les chemins allant des noeuds terminaux à la racine de l'arbre et les chemins de référence correspondants. L'identification des mots composés est évaluée par la F-mesure

6. Nous avons utilisé la version adaptée au Français pour la gestion des mots inconnus qui se trouve dans le logiciel Bonsai (Candito et Crabbé, 2009) : `http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html`.

7. Les grammaires sont apprises avec 6 cycles et une graine aléatoire de 8.

8. Cette mesure est calculée au moyen du programme *Evalb* qui est disponible à `http://nlp.cs.nyu.edu/evalb/`. Nous avons aussi utilisé l'évaluation par catégorie implantée dans la classe `EvalbByCat` de l'analyseur de Stanford.

9. On convertit d'abord automatiquement les analyses en consituants, en analyses en dépendances au moyen du logiciel *Bonsai*. Puis la mesure est calculée avec l'outil disponible à `http://ilk.uvt.nl/conll/software.html`.

10. Nous utilisons l'outil disponible à `http://www.grsampson.net/Resources.html`

F(EMM) associée aux noeuds de type EMM. La significativité statistique entre deux expériences d'analyse syntaxique est calculée au moyen du t-test unidirectionnel pour deux échantillons indépendants [11]. La significativité statistique entre deux expériences d'identification de mots composés est établie par le test de McNemar (Gillick et Cox, 1989). Les résultats de deux expériences sont considérés comme statistiquement significatifs si la valeur calculée lors du test est inférieure à 0.01.

6.2 Analyse syntaxique avec pré-identification des mots composés

Nous avons tout d'abord testé l'analyseur BKYc prenant en entrée un texte segmenté par notre reconnaisseur CRF de mots composés (sans les étiquettes). Ce dernier se base sur le logiciel *Wapiti* [12] (Lavergne *et al.*, 2010) qui apprend et applique le modèle CRF. Le logiciel *Unitex* est utilisé pour appliquer les ressources lexicales. L'étiqueteur morphosyntaxique *lgtagger* [13] sert à extraire les traits liés aux *n*-grammes de catégories grammaticales. Notre reconnaisseur intégrant tous les traits atteint 75.9% de F(EMM) sur FTB-P7 (79.1% sans tenir compte des étiquettes). Il est, en pratique, meilleur que celui proposé par (Green *et al.*, 2011) qui a une F(EMM) de 71.1% sur les phrases inférieures à 40 mots de FTB-STF [14] : notre reconnaisseur atteint, sur ce même corpus, 74% pour les traits endogènes (soit un gain absolu de +2.9%) et 77.3% pour tous les traits (soit un gain absolu de +6.2%).

Pour rendre comparables les analyses générées par BKYc couplé au reconnaisseur, avec celles de l'analyseur BKY, nous avons automatiquement transformé les analyses avec mots composés fusionnés en leurs analyses équivalentes avec des noeuds non-terminaux spécifiques pour les unités polylexicales. Les catégories grammaticales des composants internes ont été déterminées à l'aide de l'étiqueteur morphosyntaxique *lgtagger* appris sur notre corpus d'apprentissage sans intégrer de ressources lexicales externes. Les résultats sont synthétisées dans la table 3.

Traits	F	UAS	LA	F(EMM)
BKY	81.13	83.88	92.96	69.3
-	75.85	77.68	91.42	0.0
endo	81.07*	85.01	93.10	73.5
exo+endo	81.14*	85.22	93.11	75.3
gold	84.17	91.29	94.05	93.2

TABLE 2 – Intégration des mots composés dans l'analyse syntaxique par identification préalable. *endo* et *exo* indiquent que le modèle CRF incorpore respectivement les traits endogènes et les traits exogènes. *gold* signifie que la segmentation lexicale avant analyse syntaxique est parfaite. * indique que le résultat n'est pas significatif par rapport à BKY. Les tests sont réalisés sur FTB-P7.

Les résultats montrent un très grand écart de performance entre un analyseur ne tenant pas compte des mots composés [trait -] et un analyseur avec une segmentation lexicale parfaite [gold] : on a $\Delta F = 8.32$, $\Delta UAS = 13.61$, $\Delta LA = 2.69$ et $\Delta F(EMM) = 93.2$. L'analyseur *baseline* BKY permet, en partie, de combler cet écart : 63% de ΔF, 46% de ΔUAS, 59% de ΔLA et 74% de $\Delta F(EMM)$. On constate qu'une reconnaissance préalable des mots composés n'améliore pas

11. Nous utilisons l'outil de Dan Bikel disponible à `http://www.cis.upenn.edu/~dbikel/software.html`.
12. Wapiti est disponible à `http://wapiti.limsi.fr/`. Nous l'avons configuré de la manière suivante : algorithme 'rprop' et valeurs par défaut pour les pénalités L1 et L2, ainsi que le critère d'arrêt.
13. Disponible à `http://igm.univ-mlv.fr/~mconstan/research/software/`
14. (Green *et al.*, 2011) ont évalué leur système sur les phrases inférieures à 40 mots uniquement.

le parenthésage général des analyses (F-mesure) [15]. Par contre, on observe un gain significatif de +1.34 en UAS, soit une réduction relative de 18% de l'écart avec l'analyseur gold pour le système intégrant tous les traits. On remarque également une amélioration significative de la reconnaissance des mots composés de +6.0 en F(EMM), soit une réduction relative de l'écart de +25%. Si l'on analyse les résultats de F-mesure par catégorie, on s'aperçoit que le pré-repérage des EMMs provoque des effets de bord sur les constituants supérieurs comme les relatives et les subordonnées, et même les groupes nominaux. L'un des principaux problèmes vient de l'identification des verbes composés à l'indicatif ou au subjonctif qui est dramatique (F-mesure de l'ordre de 20%). Dans une moindre mesure, le repérage des noms communs composés et des conjonctions de subordination composées pose également des problèmes.

6.3 Analyse syntaxique avec réordonnancement

Nous avons ensuite évalué l'intégration d'un réordonnanceur après l'analyseur BKY. Comme dans (Charniak et Johnson, 2005), le réordonnanceur se base sur un modèle maximum d'entropie dont les paramètres sont déterminés par un algorithme d'optimisation de second ordre appelé Limited Memory Variable Metric. Concrètement, nous utilisons une implémentation de cet algorithme disponible dans les bibliotèques mathématiques PETSc [16] et TAO42 [17]. Dans un premier temps, nous avons appliqué un modèle incorporant uniquement les traits dédiés aux mots composés (cf. section 5). Nous avons ensuite comparé avec un modèle intégrant aussi les traits généraux décrits dans (Charniak et Johnson, 2005) ou (Collins, 20O0) par les patrons suivants : *Rule*, *Word*, *Heavy*, *HeadTree*, *Bigrams*, *Trigrams*, *Edges*, *WordEdges*, *Heads*, *WProj*, *NGramTree* et *Score*. Pour chaque expérience, le réordonnanceur prend, en entrée, les 50 meilleures analyses de Berkeley. Les résultats sont synthétisés dans la table 3.

Analyseur	Traits	F	UAS	LA	F(EMM)
BKY	-	81.13	83.88	92.96	69.3
BKY	endo	81.35^{*}	84.48^{+}	93.03	70.7^{+}
BKY	endo+exo	81.64	84.98	93.12	74.2
BKY	std	81.98	84.40	93.41	70.8
BKY	tous	82.05^{+}	84.45^{+}	93.42	70.2^{+}
BKYc^{+}	std	81.66^{+}	85.70	93.41	74.8

TABLE 3 – Intégration d'un réordonnanceur dans l'analyse syntaxique. Les notations *std* et *tous* correspondent respectivement aux traits généraux et à tous les traits décrits. BKYc^{+} correspond à l'analyseur BKYc couplé au reconnaisseur de mots composés avec tous les traits endogènes et exogènes. * et $^{+}$ indiquent que le résultat n'est pas significatif respectivement par rapport à l'analyseur baseline BKY et à l'analyseur BKY couplé au réordonnanceur avec les traits *std*. Les tests sont réalisés sur FTB-P7.

L'utilisation de tous les traits dédiés aux mots composés permet d'augmenter toutes les mesures par rapport à BKY : +0.51 en F, +1.10 en UAS, +0.16 en LA et +4.9 en F(EMM). Sur la reconnaissance des mots composés, on constate une relative faiblesse par rapport à la méthode par pré-identification : en analysant les analyses oracles selon F, on s'aperçoit que F(EMM) a un

15. Ces résultats sont cependant à mettre en perspective par rapport aux résultats sur le corpus de dévelopement où l'on observe des gains absolus significatifs : entre +0.2 et +0.7.
16. `http://www.mcs.anl.gov/petsc/`.
17. `http://www.mcs.anl.gov/research/projects/tao/`.

potentiel maximum de 76.9% ce qui n'est pas très élevé. Par ailleurs, les traits généraux seuls sont plus efficaces que les traits dédiés aux mots composés pour ce qui concerne le parenthésage (81.98% vs. 81.64%) et le LA (93.41% vs. 93.12%). Par contre, ils dégradent l'UAS (84.40% vs. 84.98%) et la reconnaissance des mots composés (70.8% vs. 74.2%) Le mélange des deux types de traits (*tous*) n'est pas très concluant car on n'observe aucune variation significative de l'écart par rapport à l'analyseur avec les traits généraux, quelle que soit la mesure. Ces résultats montrent qu'il est nécessaire de trouver un autre moyen de combiner ces deux types de traits.

7 Conclusions et Perspectives

Dans cet article, nous avons évalué deux stratégies discriminantes pour intégrer la reconnaissance des mots composés dans un système d'analyse syntaxique probabiliste : pré-identification *état-de-l'art* des mots composés ; repérage final des mots composés après réordonnancement des *n* meilleures analyses. Les différents modèles comprennaient des traits spécifiques aux unités polylexicales. Nous avons montré que le pré-repérage permettait d'améliorer grandement la reconnaissance des mots composés et la qualité des liens de dépendance non typés, alors que la F-mesure tend à se stabiliser. Le réordonnanceur augmente légèrement tous les scores, mais déçoit en terme d'identification de mots composés par rapport à la première méthode. Par ailleurs, l'intégration des traits généraux de (Charniak et Johnson, 2005) rend caducs les traits dédiés aux unités polylexicales et dégrade la qualité des liens de dépendance non typés. Il semble qu'aucune des deux méthodes ne soit entièrement satisfaisante. Mais ces expériences ouvrent de nouvelles perspectives intéressantes. Nous pourrions combiner efficacement ces deux stratégies en permettant au pre-segmenteur de générer l'automate pondéré des segmentations lexicales possibles et de combiner ce dernier avec l'analyseur syntaxique. Nous pourrions également transposer ces deux solutions à l'analyse en dépendance.

Remerciements

Nous souhaitons remercier Marie Candito et Spence Green pour nous avoir mis à disposition leurs versions du corpus arboré de Paris 7.

Références

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for french. *In* Abeillé, A., éditeur : *Treebanks*. Kluwer, Dordrecht.

Arun, A. et Keller, F. (2005). Lexicalization in crosslinguistic probabilistic parsing : The case of french. *In Actes de ACL*.

Cafferkey, C., Hogan, D. et van Genabith, J. (2007). Multi-word units in treebank-based probabilistic parsing and generation. *In Proceedings of the 10th International Conference on Recent Advances in Natural Language Processing (RANLP-07)*.

Candito, M. H. et Crabbé, B. (2009). Improving generative statistical parsing with semi-supervised word clustering. *In Proceedings of IWPT 2009*.

CHARNIAK, E. et JOHNSON, M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. *In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*.

COLLINS, M. (20O0). Discriminative reranking for natural language parsing. *In Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*.

CONSTANT, M. et SIGOGNE, A. (2011). MWU-aware part-of-speech tagging with a CRF model and lexical resources. *In Proceedings of the Workshop on Multiword Expressions : from Parsing and Generation to the Real World (MWE'11)*.

CONSTANT, M., TELLIER, I., DUCHIER, D., DUPONT, Y., SIGOGNE, A. et BILLOT, S. (2011). Intégrer des connaissances linguistiques dans un crf : application à l'apprentissage d'un segmenteur-étiqueteur du français. *In Actes de Conférence sur le traitement automatique des langues naturelles (TALN'11)*.

COURTOIS, B. (2009). Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, 87:1941 – 1947.

COURTOIS, B., GARRIGUES, M., GROSS, G., GROSS, M., JUNG, R., MATHIEU-COLAS, M., MONCEAUX, A., PONCET-MONTANGE, A., SILBERZTEIN, M. et VIVÉS, R. (1997). Dictionnaire électronique DELAC : les mots composés binaires. Rapport technique 56, University Paris 7, LADL.

DENIS, P. et SAGOT, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. *In Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 2009)*.

ERYIGIT, G., ILBAY, T. et ARKAN CAN, O. (2011). Multiword expressions in statistical dependency parsing. *In Proceedings of the IWPT Workshop on Statistical Parsing of Morphologically-Rich Languages (SPRML'11)*.

FINKEL, J. R. et MANNING, C. D. (2009). Joint parsing and named entity recognition. *In Proceedings of NAACL*.

GILLICK, L. et COX, S. (1989). Some statistical issues in the comparison of speech recognition algorithms. *In Proceedings of ICASSP'89*.

GREEN, S., de MARNEFFE, M.-C., BAUER, J. et MANNING, C. D. (2011). Multiword expression identification with tree substitution grammars : A parsing tour de force with french. *In Proceedings of the conference on Empirical Method for Natural Language Processing (EMNLP'11)*.

GROSS, M. (1986). Lexicon grammar. the representation of compound words. *In Proceedings of Computational Linguistics (COLING'86)*.

HOGAN, D., FOSTER, J. et van GENABITH, J. (2011). Decreasing lexical data sparsity in statistical syntactic parsing - experiments with named entities. *In Proceedings of ACL Workshop Multiword Expressions : from Parsing and Generation to the Real World (MWE'2011)*.

LAFFERTY, J., MCCALLUM, A. et PEREIRA, F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *In Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 282–289.

LAVERGNE, T., CAPPÉ, O. et YVON, F. (2010). Practical very large scale CRFs. *In Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513. Association for Computational Linguistics.

NIVRE, J. et NILSSON, J. (2004). Multiword units in syntactic parsing. *In Proceedings of Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA)*.

PETROV, S., BARRETT, L., THIBAUX, R. et KLEIN, D. (2006). Learning accurate, compact and interpretable tree annotation. *In Proceedings of ACL*.

RAMISCH, C., VILLAVICENCIO, A. et BOITET, C. (2010). mwe-toolkit : a framework for multiword expression identification. *In Proceedings of LREC*.

RAMSHAW, L. A. et MARCUS, M. P. (1995). Text chunking using transformation-based learning. *In Proceedings of the 3rd Workshop on Very Large Corpora*, pages 88 – 94.

SAG, I. A., BALDWIN, T., BOND, F., COPESTAKE, A. A. et FLICKINGER, D. (2002). Multiword expressions : A pain in the neck for nlp. *In Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing '02)*, pages 1–15, London, UK. Springer-Verlag.

SAGOT, B. (2010). The lefff, a freely available, accurate and large-coverage lexicon for french. *In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*.

SAMPSON, G. et BABARCZY, A. (2003). A test of the leaf-ancestor metric for parsing accuracy. *Natural Language Engineering*, 9(4).

SEDDAH, D., CANDITO, M.-H. et CRABBÉ, B. (2009). Cross-parser evaluation and tagset variation : a french treebank study. *In Proceedings of International Workshop on Parsing Technologies (IWPT'09)*.

SILBERZTEIN, M. (2000). Intex : an fst toolbox. *Theoretical Computer Science*, 231(1):33–46.

TELLIER, I. et TOMMASI, M. (2011). Champs Markoviens Conditionnels pour l'extraction d'information. *In* Eric GAUSSIER et François YVON, éditeurs : *Modèles probabilistes pour l'accès à l'information textuelle*. Hermès.

WATRIN, P. et FRANÇOIS, T. (2011). N-gram frequency database reference to handle mwe extraction in nlp applications. *In Proceedings of the Workshop on Multiword Expressions : from Parsing and Generation to the Real World (MWE'11)*.

WEHRLI, E., SERETAN, V. et NERIMA, L. (2010). Sentence analysis and collocation identification. *In Proceedings of the Workshop on Multiword Expression : From Theory to Applications (MWE'10)*.

Calcul des cadres de sous-catégorisation des noms déverbaux français (le cas du génitif)

Ramadan Alfared Denis Béchet Alexandre Dikovsky
LINA, Université de Nantes, 2, rue de la Houssinière, 44000 Nantes
Ramadan.Alfared@etu.univ-nantes.fr
Denis.Bechet@univ-nantes.fr
Alexandre.Dikovsky@univ-nantes.fr

RÉSUMÉ

L'analyse syntaxique fine en dépendances nécessite la connaissance des cadres de sous-catégorisation des unités lexicales. Le cas des verbes étant bien étudié, nous nous intéressons dans cet article au cas des noms communs dérivés de verbes. Notre intérêt principal est de calculer le cadre de sous-catégorisation des noms déverbaux à partir de celui du verbe d'origine pour le français. Or, pour ce faire il faut disposer d'une liste représentative de noms déverbaux français. Pour calculer cette liste nous utilisons un algorithme simplifié de repérage des noms déverbaux, l'appliquons à un corpus et comparons la liste obtenue avec la liste Verbaction des déverbaux exprimant l'action ou l'activité du verbe. Pour les noms déverbaux ainsi obtenus et attestés ensuite par une expertise linguistique, nous analysons la provenance des groupes prépositionnels subordonnés des déverbaux dans des contextes différents en tenant compte du verbe d'origine. L'analyse est effectuée sur le corpus Paris 7 et est limitée au cas le plus fréquent du génitif, c'est-à-dire des groupes prépositionnels introduits par *de*, *des*, etc.

ABSTRACT

On Computing Subcategorization Frames of French Deverbal Nouns (Case of Genitive)

Fine dependency analysis needs exact information on the subcategoriziation frames of lexical units. These frames being well studied for the verbs, we are interested in this paper by the case of the noun deverbals. Our main goal is to calculate the subcategoriziation frame of deverbals in French from that of the source verb. However, this task needs a representative list of French deverbal nouns. To obtain such a list, we use a simplified algorithm detecting deverbal nouns in texts. The obtained list attested by linguists is compared with the existing list Verbaction of deverbals expressing the action/activity of French verbs. For these deverbal nouns, we analyse the origin of their subordinate prepositional phrases in different contexts relative to their source verbs. This analysis is carried out over the corpus Paris 7 and is limited to the most frequent cases of the genitive, i.e. to the prepositional phrases headed by the prepositions *de*, *des*, etc.

MOTS-CLÉS : nom déverbal, cadre de sous-catégorisation, groupe prépositionnel, analyse en dépendances.

KEYWORDS: Deverbal Noun, Subcategorization Frame, Prepositional Phrase, Dependency Tree.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 71–84,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Comparons les trois phrases suivantes :

(1) *Le service de livraison est fermé.*
(2) *Le besoin de reconnaissance est un besoin fondamental de tous les êtres humains.*
(3) *La livraison des commandes est assurée directement par le service de livraison du producteur.*
(4) *On attend l'abrogation définitive de cette loi.*

Nous voyons que dans chaque phrase il y a au moins un groupe prépositionnel (GP) subordonné à un nom : e.g. ***de*** *livraison* est subordonné à *service*, ***des*** *commandes* est subordonné à *livraison*, ***du*** *producteur* est subordonné à une autre occurrence de *livraison*. Parmi ces dépendances entre les noms gouverneurs et les prépositions têtes des GP, lesquelles sont *attributives*, lesquelles sont *dépendances d'objet* ? On ne peut pas effectuer une analyse fine en dépendances sans savoir répondre à cette question. En effet, dans la grammaire de dépendances il faut au moins faire la distinction entre les dépendances obligatoires d'objet et les dépendances facultatives de modifieurs en tout genre. Cette différence entre les deux types de dépendances correspond aussi à une différence fondamentale sémantique : dans le cas de dépendances attributives la sémantique du GP définit la valeur d'un attribut de la sémantique du nom gouverneur, tandis que dans le cas de dépendances d'objet la sémantique du nom gouverneur est une fonction appliquée à la sémantique du GP. Mais comment peut-on savoir dans quel cas de figure on se trouve ?

On peut distinguer deux catégories de substantifs :
(c_1**)** ceux dont au moins un GP subordonné est un complément d'objet (c'est-à-dire, correspond à un argument sémantique) [1] ,
(c_2**)** et ceux qui n'en ont aucun.

Les compléments des substantifs de la catégorie c_1 ont le même statut (par rapport au substantif) que celui des éléments du cadre de sous-catégorisation des verbes (par rapport aux verbes). Aussi ces substantifs peuvent avoir des compléments d'objet dans tous les contextes (ils peuvent aussi avoir des GP attributifs). Parmi ces substantifs on trouve les nombres indéterminés exprimant une quantité limitée et indéterminée d'objets, tels *dizaine*, *douzaine*, *trentaine*, etc. On y trouve aussi les noms communs tels *besoin*, *peur*, *marre*, etc. qui avec un verbe léger (e.g. *avoir*) forment un phrasème et qui en héritent les compléments d'objet (au génitif). Mais surtout on y trouve de très nombreux noms déverbaux. On s'intéresse dans cet article surtout aux noms déverbaux parce qu'il existe un certain nombre de règles qui définissent leurs arguments à partir des arguments des verbes d'origine.

Les déverbaux héritent du cadre de sous-catégorisation du verbe d'origine et ils le transforment. E.g. dans les phrases (1)-(4) nous trouvons les déverbaux *reconnaissance*, *livraison*, *service*, *producteur*, *abrogation*.Comme on peut le voir dans les analyses [2] ci-dessous, *service* (catégorie **(**c_2**)**) a un GP attributif *de livraison*, tandis que *besoin* et *livraison* (catégorie **(**c_1**)**) ont un GP d'objet *de reconnaissance* dans la phrase (2), *des commandes* dans la phrase (3). Le déverbal *abrogation* a un GP d'objet *de cette loi* dans la phrase (4). En même temps, *besoin* peut aussi avoir un GP attributif, e.g. *de tous les êtres humains* dans la phrase (2).

1. Comme tous les arguments sémantiques, ces arguments peuvent être liés à des éléments du contexte, ce qui correspond à l'ellipse de surface mais ne les prive pas du statut d'arguments obligatoires.
2. Ces analyses sont obtenues avec l'analyseur syntaxique et en utilisant la grammaire catégorielle de dépendance du français qui font partie du système CDG Lab (Alfared *et al.*, 2011).

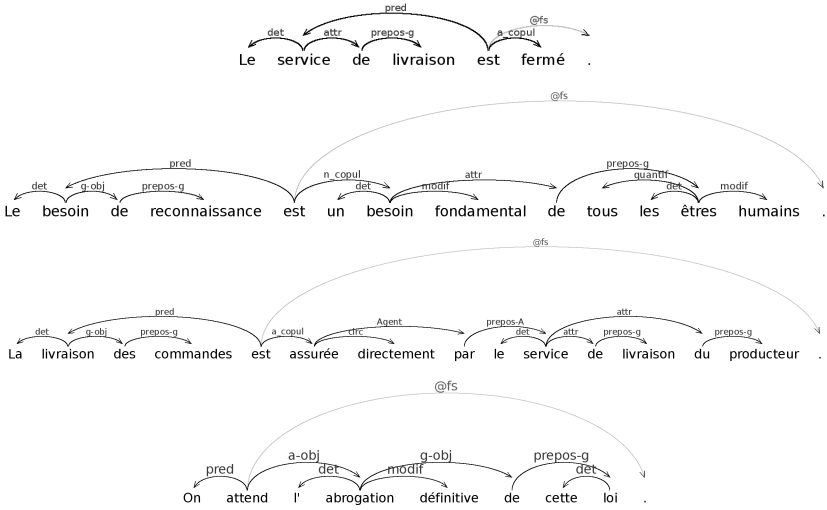


FIGURE 1 – Analyses des phrases de l'introduction

Les nombres indéterminés et les noms composant les phrasèmes ne posent pas de problème parce qu'ils sont peu nombreux. Par contre, la recherche des déverbaux n'est pas une tâche simple (surtout pour les verbes des deuxièmes et troisièmes groupes). Et les difficultés ne s'arrêtent pas là : le vrai problème est de déterminer les cadres de sous-catégorisation des déverbaux à partir des cadres de sous-catégorisation des verbes. Ce sont les deux problèmes abordés dans cet article. Par rapport au système CDG Lab mentionné ci-dessus, ces problèmes se traduisent par des algorithmes de complétion du lexique de la grammaire du français plaçant les déverbaux dans les classes lexicales qui correspondent à leurs cadres de sous-catégorisation. Cela explique notre intérêt pour cette problématique.

L'article est organisé comme suit : la section 2 présente un algorithme de repérage des noms déverbaux ainsi que son évaluation. La section suivante introduit le cadre théorique de l'article sur les liens entre les cadres de sous-catégorisation des noms déverbaux et des verbes d'origine. La section 4 présente nos expériences sur les règles de transformation des cadres (du verbe au déverbal). L'article se termine par une discussion et les perspectives de ce travail.

2 Recherche des déverbaux

En français les noms déverbaux sont formés des verbes selon le schéma :

$$\text{BASE+TERMINAISON} \Rightarrow \text{var(BASE)+SUFFIXE},$$

où var(BASE) est une variation de la base et SUFFIXE est un suffixe formateur (parfois vide : -0).

Le plus simple est d'obtenir la liste plus ou moins complète des suffixes formateurs cf. (Namer, 2009). En ce qui concerne les variations de la base, il s'avère difficile de trouver une source où elles sont toutes présentées systématiquement (en tout cas nous ne l'avons pas trouvée). Heureusement, pour les verbes du premier groupe, auquel appartient la majorité absolue des verbes français, les variations sont très rares et on peut les négliger (ou bien ne considérer que les plus simples). Cela rend possible d'appliquer aux verbes du premier groupe un algorithme simpliste (appelé ci-dessous **proto-déverb**) qui effectue la transformation :

BASE+TERMINAISON ⇒ BASE+SUFFIXE

selon 17 règles citées dans la table 3. Nous avons appliqué **proto-déverb** à deux corpus : un d'arbres syntaxiques : le corpus Paris 7 (Abeillé *et al.*, 2003) [3], un autre textuel : le corpus GEOPO (du CLLE-ERSS(Ho-Dac, 2007)). **Proto-déverb** pose plusieurs problèmes :

Problème de faux déverbaux. L'algorithme **proto-déverb** peut construire des combinaisons BASE+SUFFIXE qui ne sont pas des déverbaux. Par exemple, pour le suffixe *-ence* : *scier* ⇒ *science*, pour le suffixe *-tion* : *roter* ⇒ *rotation*.

Problème de fausse origine. La combinaison BASE+SUFFIXE construit par **proto-déverb** peut être un déverbal, mais provenir d'une autre base. Par exemple, pour le suffixe *-ant* : *garer* ⇒ *garant*, tandis que *garant* est formé du verbe *garantir*.

Problème de fausse direction. Parfois le nom construit par **proto-déverb** n'est pas un déverbal, mais c'est le verbe d'origine qui est dérivé du nom. Par exemple, pour le suffixe *-euse* : *mitrailler* ⇒ *mitrailleuse* (en fait, *mitrailleuse* ⇒ *mitrailler*). Pour ce genre de problème il vaudrait mieux parler de co-déverbaux, mais cette situation est assez rare et on peut aussi la négliger.

2.1 Évaluation de l'algorithme *proto-déverb*

On voit que l'application de l'algorithme **proto-déverb** nécessite une expertise du résultat de son application. Notre démarche d'évaluation consiste ici à réaliser deux expérimentations indépendantes ainsi qu'une comparaison directe avec la liste Verbaction de déverbaux dénotant l'action ou l'activité exprimée par le verbe. La première expérimentation compare la liste des déverbaux produit par **proto-déverb** avec une liste produite par un expert depuis une partie des textes du corpus GEOPO(Ho-Dac, 2007). La seconde expérimentation, effectuée sur le corpus arboré de Paris 7 (Abeillé *et al.*, 2003), qui sera longuement présentée et commentée dans la seconde partie de l'article consacré aux GP des déverbaux, permet d'obtenir, en la restreignant à la notion d'être ou non un déverbal, une évaluation des déverbaux proposés par l'algorithme **proto-déverb**. Finalement, nous avons utilisé la liste des déverbaux de Verbaction (Hathout *et al.*, 2002; Tanguy et Hathout, 2002) développée à l'INaLF/ATILF et à l'ERSS pour la comparer avec la liste produite par **proto-déverb**.

2.2 Évaluation sur des textes du corpus GEOPO

La première expérience vérifie la pertinence de l'algorithme **proto-déverb** pour la construction des couples déverbal/verbe sur des textes du corpus GEOPO. Ce corpus est constitué d'environ 270 000 mots. Il regroupe 32 textes longs qui sont des articles expositifs (informatifs et argumentatifs) proposant des réflexions relatives à notre monde d'aujourd'hui (la crise pétrolière, la guerre contre "l'axe du mal", le terrorisme, l'explosion chinoise, la paix au moyen-orient, etc.) La méthodologie

3. Parmi les 17 règles seulement 12 de la table 2 ont été appliquées au moins une fois dans ce corpus.

utilisée pour l'évaluation consiste à repérer automatiquement les séquences constituées d'un nom commun [4] suivi d'une des formes de la préposition *de* (*de*, *d'*, *du* ou *des*). Ensuite, les séquences repérées sont annotées à la fois par l'algorithme **proto-déverb** et par un expert. L'annotation de **proto-déverb** indique pour un nom commun s'il est un déverbal. Dans ce cas l'algorithme fournit le verbe [5] dont il est dérivé. L'expert indique de même si les séquences repérées correspondent à un déverbal (en particulier, le mot doit être un nom commun dans la phrase). Dans ce cas, le verbe d'origine est fourni par l'expert. Dans le cas contraire, l'annotateur indique la raison pour laquelle le mot n'est pas un déverbal. Cela peut être :

1. le mot n'est pas un nom commun dans le contexte, comme par exemple *parler* dans *plutôt que de parler de centralisation, [...]*,
2. le nom fait partie d'une locution comme *point* dans *ainsi, d'un point de vue législatif [...]*,
3. le nom ne correspond pas (dans le sens de la phrase) à un déverbal comme *priorité* dans *Il s'agit au contraire d'une priorité des pouvoirs publics*,
4. l'expert annotateur n'est pas sûr que le nom soit un déverbal comme le nom *poids* pour le verbe *peser*.

L'algorithme **proto-déverb** ne pouvant pas détecter les situations 1 et 2, nous avons éliminé les séquences correspondantes des statistiques. Les cas 4 étant incertains, nous avons aussi préféré les supprimer. Nous n'avons donc gardé que les séquences du point 3 et celles qui ont été annotées par les experts comme des déverbaux.

Sur les 286 séquences de phrases traitées, 220 ont été retenues par les experts (14 de statut incertain (point 4) et 52 correspondant aux points 1 ou 2).

	Déverbal (proto-déverb)	Non déverbal (proto-déverb)
Déverbal (expert)	28 (12,7%)	63 (28,6%)
Non déverbal (expert)	4 (1,8%)	125 (56,8%)

TABLE 1 – Comparaison sur GEOPO entre **proto-déverb** et les experts annotateurs

Ces résultats suggèrent que les règles utilisées par l'algorithme **proto-déverb** donnent pour la grande majorité des occurrences de vrais déverbaux (28 contre 4) dans ce type de corpus. Par contre, la couverture des règles pourrait être améliorée. En fait, l'algorithme simpliste **proto-déverb** se contente de traiter les verbes du premier groupe. De plus, certaines règles trop peu précises n'ont pas été incluses dans cette version de l'algorithme comme la règle qui supprime le *r* final BASE+er ⇒ BASE+e et qui permet par exemple de générer le couple *contrôle/contrôler*. L'ajout de ce type de règles aurait conduit à une production d'un trop grand nombre de faux couples déverbaux/verbes qui aurait complexifié la tache des experts sur le corpus Paris 7.

2.3 Évaluation sur le corpus arboré Paris 7

Pour compléter ces statistiques, nous pouvons aussi utiliser une partie des résultats de la seconde expérience pour juger de la qualité de l'algorithme **proto-déverb**. Lors de cette seconde

4. En fait, il s'agit des mots qui sont des noms communs dans le lexique Lefff (Sagot, 2010).
5. Exceptionnellement, l'algorithme peut fournir plusieurs verbes correspondant au mot. Par exemple, *parade* correspond à *parer* et à *parader*.

expérience nous avons repéré, dans le corpus d'arbres syntaxiques de Paris 7, les constructions où un nom commun provenant d'une règle de l'algorithme **proto-déverb** possède un groupe prépositionnel (GP) introduit par la préposition *de*. Sur ces éléments, nous avons demandé à des experts annotateurs de repérer les séquences qui contiennent effectivement le déverbal du verbe donné par l'algorithme **proto-déverb**. Nous verrons dans la seconde partie de l'article sur les compléments au génitif des déverbaux que, dans le cas où l'expert pense que le nom est un déverbal, il doit fournir l'éventuel lien entre le GP et un des arguments du verbe d'origine. Par exemple dans la phrase *Il ne faut pas secondariser l'enseignement supérieur ni inciter à un démembrement des universités*, le GP *des universités* du déverbal *démembrement* correspond au complément à l'accusatif (complément d'objet direct) du verbe *démembrer*. En faisant abstraction de ce résultat sur le GP pour l'instant, nous pouvons savoir si la proposition de l'algorithme **proto-déverb** correspond bien à un déverbal dans le contexte d'une phrase. Les résultats en fonction des règles de l'algorithme **proto-déverb** sont les suivants :

Règle	Déverbal (expert)	Non déverbal (expert)	Nombre de couples
er ⇒ ade	2	1	3
er ⇒ age	63	10	73
er ⇒ aison	5	0	5
er ⇒ ant	26	5	31
er ⇒ ation	274	7	281
er ⇒ ement	132	6	138
er ⇒ ence	24	7	31
er ⇒ eur	45	17	62
er ⇒ euse	1	0	1
er ⇒ oir	1	4	5
er ⇒ rice	1	1	2
er ⇒ ure	10	4	14
Total	584 (90,4%)	62 (9,6%)	646

TABLE 2 – Évaluation de **proto-déverb** sur le corpus Paris 7 en fonction des règles

Ce tableau indique pour chaque règle utilisée pour l'analyse du corpus le nombre de couples nom/verbe qui correspondent ou non à un couple déverbal/verbe pour l'expert ainsi que le total. Globalement, le résultat est assez bon mais n'est pas uniforme en fonction de la règle de passage du verbe au nom. La précision globale est de 90,4%.

2.4 Comparaison avec la liste de déverbaux Verbaction

Une troisième expérience consiste à comparer la liste de 5537 couples déverbal/verbe produits par l'algorithme **proto-déverb** à partir du lexique L*efff* avec les 8432 couples du lexique Verbaction qui comprend une liste portant sur des noms morphologiquement apparentés au verbe et qui peuvent être utilisés pour dénoter l'action ou l'activité exprimée par le verbe. Pour cette comparaison nous avons trois possibilités suivant qu'un couple nom/verbe correspond à une ou aux deux listes. La table 3 présente les résultats pour les verbes du premier groupe qui sont produits par l'algorithme **proto-déverb**.

Règle	Déverbal (Verbaction)	Non déverbal (Verbaction)	Nombre de couples
er ⇒ ade	36	33	69
er ⇒ age	994	60	1054
er ⇒ aison	22	15	37
er ⇒ ant		312	312
er ⇒ ante	1	127	128
er ⇒ ation	918	85	1003
er ⇒ ement	778	61	839
er ⇒ ence	19	18	37
er ⇒ ette	6	187	193
er ⇒ eur	1	938	939
er ⇒ euse		538	538
er ⇒ oir		145	145
er ⇒ oire		20	20
er ⇒ rice		17	17
er ⇒ trice		5	5
er ⇒ ure	33	166	199
er ⇒ xion	1	1	2
Total	2809 (50,7%)	2728 (49,2%)	5537

TABLE 3 – Comparaisons entre **proto-déverb** et le lexique Verbaction en fonction des règles

Le nombre de couples avec un verbe du premier groupe du lexique Verbaction qui ne sont pas produits par l'algorithme est de 5623 (sur 8432). En fait, comme nous l'avons déjà vu précédemment, l'algorithme **proto-déverb** ne couvre pas la totalité des verbes du premier groupe. Par contre, il fournit des couples qui ne sont pas dans Verbaction comme les noms de suffixe *ant*, *ante*, *ette*, *eur*, etc qui ne désignent pas directement l'action ou l'activité exprimée par le verbe d'où l'intérêt de cet algorithme dans cette étude.

Pour être complet, nous avons aussi comparé les annotations de nos experts sur le corpus GEOPO avec la liste de Verbaction ce qui a donné les résultats de la table 4. L'adéquation entre nos experts et le lexique Verbaction est assez bonne (85,5%) mais il reste quelques différences comme les couples *structure/structurer*, *position/positionner*, *craintes/craindre* ou *résistant/résister* qui s'explique par le choix de Verbaction de ne s'intéresser qu'aux déverbaux dénotant l'action ou l'activité exprimée par le verbe.

En conclusion, l'algorithme **proto-déverb** est simple et assez précis. Il fournit un grand nombre de déverbaux ayant des liens variés avec le verbe (contrairement à Verbaction qui est spécialisé).

	Déverbal (Verbaction)	Non déverbal (Verbaction)
Déverbal (expert)	69 (31,4%)	22 (10%)
Non déverbal (expert)	10 (4,5%)	119 (54,1%)

TABLE 4 – Comparaison sur GEOPO entre le lexique Verbaction et les experts annotateurs

3 Cadres de sous-catégorisation des déverbaux

Passons maintenant au problème principal, celui de déterminer les cadres de sous-catégorisation des déverbaux à partir des cadres de sous-catégorisation des verbes d'origine. Avant tout il faut expliquer comment nous représentons ces cadres. Notre représentation est basée sur une affectation de cas aux compléments d'objet proposée dans (Dikovsky, 2011). Cette affectation est faite en fonction des clitiques selon la règle suivante (simplifiée ici) :
Soit π un complément d'objet d'un verbe V dans une phrase. Alors π est au cas C si dans la phrase il peut être pronominalisé et transformé en un clitique P au cas C ancré sur V. Enfin, si π ne peut pas être pronominalisé, alors il est au cas *oblique* (o) [6].

En français il n'y a que quatre cas de clitiques : *accusatif* (a), e.g. *me, la, en*, *génitif* (g) : *en*, *datif* (d), e.g. *te, lui, y*, et *locatif* (l) : *y*. Respectivement, selon cette règle chaque complément d'un verbe a un des cinq cas : a, g, d, l ou o [7]. Aussi le cadre de sous-catégorisation du verbe est défini par les fonctions syntaxiques de ses arguments obligatoires (*actants*) : *sujet* $(subj)$, *objet direct* (od), *objet indirect* (oi)), réalisés par les cas qui leur correspondent, e.g. *chanter*$(subj/n)$ (intransitif), ou *chanter*$(subj/n, od/a)$, *donner*$(subj/n, od/a, oi/d)$, *émerger*$(subj/n, oi/g)$, *parler*$(subj/n, od/a)$ ou *parler*$(subj/n, oi1/g, oi2/d)$, *placer*$(subj/n, od/a, oi/l)$, *échanger*$(subj/n, od/a, oi/o)$, etc. Parfois, pour préciser la réalisation du complément direct par une phrase dont la tête est un verbe à l'infinif, nous utilisons l'étiquette inf : *faire*$(subj/n, od/a, oi/inf)$. Le cadre s'étend sur les GP en position d'argument circonstanciel : $circ/C$, quand C est le cas du GP.

L'idée principale de notre calcul des cadres de sous-catégorisation des noms déverbaux est que dans le cas où on sait de quel verbe est formé le nom déverbal, on peut déduire le cadre du second à partir de celui du premier. Il s'agit donc de définir les règles de transformation des cadres [8] des verbes les cadres des noms dérivés. Ce calcul des cadres par réduction rend possible de contourner le problème difficile de définition des cadres de sous-catégorisation des verbes (cf. (Fillmore, 1968, 1977; Grimshaw, 1990; Dowty, 1991; Van Valin, 1997; Mel'čuk, 2004)). Les cadres de sous-catégorisation des verbes sont alors donnés a priori et sont représentés par les listes des réalisations des actants : fonction syntaxique/cas définies ci-dessus : $subj/n$, od/a, oi/d, oi/g,oi/l,oi/o, oi/inf. Les cadres des noms déverbaux sont représentés par les réalisations des objets (sauf od/a), mais aussi des attributs : $attr/C$. Quand un argument circonstanciel d'un verbe ou un arguments attributif d'un nom n'est pas réalisé par un cas nous allons noter cette réalisation par $circ/\perp$ ($attr/\perp$). Sur cette signature étendue on peut définir les transformations diathétiques de nominalisation des verbes par les séquences de règles du genre $(subj/n\rightarrow 0)$ [9] ou $((od/a\rightarrow oi/g), (subj/n\rightarrow attr/g))$ ou $(oi/l\rightarrow oi/l)$, etc.

Notre hypothèse initiale était que les transformations diathétiques de nominalisation en français sont définies par quatre règles simples :
- le sujet du verbe disparait et est converti en un attribut du nom dérivé : $(subj/n\rightarrow attr/C)$,
- son objet direct $(C = a)$ devient l'objet indirect au génitif $(od/a\rightarrow oi/g)$,
- tous les autres objets indirects sont tout simplement hérités sous réserve de leur présence dans le cadre (e.g $(oi/d\rightarrow oi/d)$ ou $(oi/o\rightarrow oi/o)$,

6. Certes, il faut définir plus précisement cette règle pour les clitiques ambigus : *y* au datif vs. *y* au locatif, *en* au génitif vs. *en* à l'accusatif, etc. et la compléter par le cas *n* (*nominatif*) du sujet.
7. Cf. (van den Eynde et Mertens, 2003)) où on tient compte de tous les pronoms.
8. Diathetic shift rules (ang.).
9. *A*→0 veut dire que *A* est supprimé.

- les GP circonstanciels deviennent GP attributifs ($circ/C_1 \rightarrow attr/C_2$).
Il s'avère que le tableau réel est bien plus complexe. Pour commencer, certains arguments des déverbaux hérités des cadres des verbes d'origine ont tendance à perdre leur statut obligatoire. Il s'agit surtout des compléments locatifs et obliques. Par exemple, le cadre du verbe pronominal *se déplacer* est soit (oi/l) soit (oi/o) (un des deux est obligatoire : cf. *Il se déplace à Bordeaux* ; *Il se déplace en voiture* mais **Il se déplace*). Encore pire est la situation où le verbe d'origine est di-transitif avec le cadre $(subj/n, od/a, oi/g)$, (e.g. *saupoudrer*$(subj/n, od/a, oi/g)$). Dans ce cas un des compléments est supprimé. En fonction du contexte le complément d'objet au génitif du déverbal *saupoudrage*(oi/g) peut provenir soit du complément d'objet direct du verbe (cf. *saupoudrage des champs* $(subj/n \rightarrow 0, od/a \rightarrow oi/g, oi/g \rightarrow 0)$), soit de son complément d'objet indirect au génitif (cf. *saupoudrage de réformettes* $(subj/n \rightarrow 0, od/a \rightarrow 0, oi/g \rightarrow oi/g)$. Et enfin, très souvent le GP subordonné à un nom déverbal peut provenir du sujet et même d'un des arguments circonstanciels du verbe d'origine. Par exemple, dans la phrase *L'argumentation du clan pro-inflation consiste à ...* le GP *du clan pro-inflation* est un attribut du nom déverbal *argumentation*(0). Cet attribut provient du sujet du verbe *argumenter*$(subj/n, od/a)$. Ainsi dans ce contexte le cadre de sous-catégorisation du verbe est transformé selon la règle $(subj/n \rightarrow attr/g, od/a \rightarrow 0)$. D'un autre côté, dans la phrase *La débandade de 1982-1983 qui vit le pays...* le verbe *débander*$(subj/n)$ est intransitif. Aussi l'attribut *de 1982-1983* du déverbal *débandade* provient d'un argument circonstanciel du verbe (*débander en 1982-1983*) : $(circ/o \rightarrow attr/g)$.

Sans argument supplémentaire ces exemples peuvent faire croire qu'il faut faire face à un nombre important de transformations de cadres de sous-catégorisation. Heureusement, les actants des noms déverbaux obéissent au postulat suivant (visiblement remontant à la présomption d'invariabilité des actants des mots de (Mel'čuk et Holodovič, 1970)) [10] :

postulat de provenance : les actants des noms déverbaux proviennent d'un des actants du verbe d'origine.

Une conséquence directe de ce postulat est que l'actant du déverbal ne peut pas provenir d'un argument circonstanciel du verbe d'origine mais seulement de son sujet ou d'un des objets présent dans son cadre. Dû à ce postulat notre objectif devient plus précis : on peut établir la provenance des actants des noms déverbaux réalisés par des GN non attributifs seulement à partir des GN non circonstanciels du verbe d'origine. Quoique très restrictif, le postulat de provenance laisse beaucoup de variantes : différents suffixes formateurs, différents rôles syntaxiques des noms déverbaux dans différents contextes et différents actants correspondants du verbe d'origine. Pour simplifier la tâche nous nous limitons dans cet article aux GN au génitif qui réalisent les actants des noms déverbaux. Dans les expériences présentées dans la section suivante nous étudions leur provenance et les règles de conversion qui les concernent.

4 Expériences et évaluation des résultats

Ici, nous nous intéressons à découvrir le rôle des GP des déverbaux pour le verbe qui leur correspond. Nous avons identifié dans un premier temps plusieurs possibilités puis nous avons cherché à comprendre sur des exemples d'un corpus les liens entre les compléments du verbe et les GP introduit par la préposition *de* du déverbal. Notre étude a consisté sur le corpus d'arbres

10. Nos expériences avec le corpus Paris 7 confirment ce postulat (au moins pour les noms déverbaux repérés dans le corpus).

syntaxiques de phrases de Paris 7 (Abeillé *et al.*, 2003), à repérer toutes les occurrences de noms communs qui correspondent à l'algorithme **proto-déverb** et qui possèdent un (ou plusieurs) GP introduit par la préposition *de*. Sur ces occurrences, nous avons demandé à des experts d'indiquer pour ces occurrences si les couples nom/verbe produit par l'algorithme **proto-déverb** correspondent effectivement (dans le contexte des occurrences) à un couple déverbal/verbe. Dans un deuxième temps, pour les occurrences validées comme des déverbaux, les experts devaient indiquer la fonction que pouvaient prendre les GP par rapport au verbe. Voici quelques exemples qui illustrent les différents cas de figure rencontrés par les experts annotateurs :

1. *Le* langage de la formation*, trop souvent fait d'approximations, [...]* avec le couple *langage/langer* : en fait, *langage* n'a rien à voir avec *langer*.
2. *L'alliance des vrais-faux contraires suscite la* parade des supposés amis avec le couple *parade/parer* : le nom *parade* peut être associé à deux verbes suivant son sens. Soit *parer*, soit *parader*. Dans cette phrase, le sens de *parade* correspond à *parader*. Ces deux premiers exemples sont étiquetés **non D** (non déverbal).
3. *Avec une croissance de 3,3% à la même date, les départements d'outre-mer bénéficieront d'un* rattrapage d'un point supplémentaire avec le couple *rattrapage/rattraper* : ici nous avons clairement un déverbal. Le GP *d'un point supplémentaire* joue le rôle du complément à l'accusatif (complément d'objet direct) du verbe *rattraper* car nous pourrions dire *ceci rattrape un point supplémentaire*. Cet exemple est étiqueté **C=a** par les experts.
4. *Reports ou renvois qui n'empêchent pas M. Bérégovoy de juger "injuste" l'*accusation d'immobilisme avec le couple *accusation/accuser* : le GP du déverbal correspond à un complément au génitif du verbe *accuser* (complément d'objet indirect introduit par *de*) car on peut dire *on l'accuse d'immobilisme*. Cet exemple est étiqueté **C=g** par les experts.
5. *Dans l'intervention humanitaire, il n'y a pas d'*obligation de résultats avec le couple *obligation/obliger* : cette fois le GP correspond à un complément au datif (complément d'objet indirect introduit par *à*) car nous pouvons dire *cela nous oblige à des résultats*. Cet exemple est étiqueté **C=d** par les experts.
6. *De plus, la Russie, mastodonte pétrolier empêtré dans des problèmes infinis, paraît incapable de stopper la* dégringolade de sa production*, [...]* avec le couple *dégringolade/dégringoler* : ici, *la production* pourrait être le sujet de dégringoler car nous pouvons dire *la production dégringole*. Cet exemple est étiqueté **subj** par les experts.
7. *Au plan agricole, la délégation de Gdansk a pris contact avec les professionnels de la race bovine limousine, dont les performances zootechniques, notamment en* élevage *extensif* de plein air*, leur ont semblé particulièrement adaptées aux conditions polonaises* avec le couple *élevage/élever* : cette fois, le GP correspond à un argument circonstanciel (optionnel) du verbe *élever* puisque nous pourrions dire *on les élève en plein air*. Cet exemple est étiqueté **circ** par les experts.
8. *Début 1991, autour de Joachim Trautwein, les six "élus" de l'ancien encadrement suivent trois séminaires de deux jours, en présence d'un sociologue est-allemand et d'un* chercheur *français* du CNRS avec le couple *chercheur/chercher* : dans cette phrase, il est difficile de rattacher *le CNRS* au verbe *chercher* car le GP désigne l'appartenance à une organisation de la personne qui cherche. Nous pouvons dire dans ce cas que bien que nous ayons un déverbal, le GP de ce déverbal ne provient pas de la structure des compléments du verbe d'origine. Cet exemple est étiqueté **V/S** (un déverbal dont le GP ne peut pas être subordonné au verbe : $(0 \rightarrow attr/g)$).

Les annotations ont été classées par la règle de l'algorithme **proto-déverb** utilisée pour faire la correspondance entre le déverbal et le verbe. Les séquences ont été extraites d'un sous ensemble du corpus (Abeillé *et al.*, 2003) constitué de 31 fichiers sur 45 (les autres fichiers nous permettant de vérifier nos hypothèses) et comportant 15106 phrases (sur 21776). Le programme d'extraction a identifié 647 séquences *nom + GP* avec *nom*, un nom commun produit par l'algorithme **proto-déverb** (le nom peut être séparé de son GP par d'autres éléments du groupe nominal). Sur ces 647 séquences les experts ont supprimé une phrase pour laquelle le lien entre le GP et le verbe n'était pas clair [11]. Les experts ayant divergé sur l'interprétation de la fonction de certains GP entre **C=a** et les autres cas, nous avons dans un premier temps conservé l'annotation qui n'était pas **C=a**. Toutefois, nous indiquons, dans la colonne **C=a**, avec le symbol + suivi d'un nombre, le nombre de cas **C=a** divergents. En fait, sur les 646 séquences, 18 annotations sont restées différentes même après concertation des annotateurs. Comme nous l'avons vu précédemment, les experts ont repéré 62 séquences ne correspondant pas à un déverbal et 584 correspondant à un déverbal. Voici la table des résultats classés par règle :

Règle	non D	C=a	C=d	C=g	subj	circ	V/S	Total D	Total
er ⇒ ade	1				1	1		2	3
er ⇒ age	10	44 + 3		4	11	4		63	73
er ⇒ aison		5						5	5
er ⇒ ant	5	17 + 2	1		2	5	1	26	31
er ⇒ ation	7	210 + 8	3	3	39	17	2	274	281
er ⇒ ement	6	101 + 4	3	2	11	12	3	132	138
er ⇒ ence	7	5	1		14	4		24	31
er ⇒ eur	17	24 + 1		1	2	10	8	45	62
er ⇒ euse						1		1	1
er ⇒ oir	4					1		1	5
er ⇒ rice	1	1						1	2
er ⇒ ure	4	6	1			2	1	10	14
Total	62	413 + 18	9	10	80	57	15	584	646

TABLE 5 – Analyse sur le corpus Paris 7 de la fonction des GP des déverbaux

Nous voyons que très majoritairement (413 sur 584 soit 70,7%), le GP correspond au complément d'objet direct du verbe. Les deux autres types de compléments d'objet indirect (C=d et C=g) sont bien moins nombreux car les verbes du premier groupe avec ce type de complément sont peu nombreux. Les annotateurs en ont repéré 19 soit 3,3%.

À la lecture des exemples annotés C=d, on s'aperçoit que les déverbaux autorisent de manière générale qu'un GP normalement introduit par la préposition *à* soit introduit par la préposition *de* avec le déverbal. Ainsi, on peut dire *Dans l'intervention humanitaire, il n'y a pas d'obligation* **de** *résultats* à la place de *Dans l'intervention humanitaire, il n'y a pas d'obligation* **à** *des résultats*.

Par contre, les compléments du verbe au génitif (C=g) se traduisent naturellement en un GP du déverbal introduit par *de*. Ainsi, dans la phrase *reports ou renvois qui n'empêchent pas M. Bérégovoy*

11. "*Le* déplacement *en milieu psychiatrique* du traitement de l'acte - *qui ne serait parlé que là, médicalement - exclut l'inculpé de la société : non seulement il est dispensé, mais il est empêché de répondre de son acte devant les parties directement concernées alors que les deux démarches devraient être complémentaires*" avec le couple *déplacement/déplacer*.

de juger injuste l'accusation d'immobilisme, la séquence *accusation d'immobilisme* correspond à *accuser d'immobilisme*.

Le cas des arguments circonstanciels est aussi assez simple puisque les 57 exemples montrent que ces compléments du verbe introduit par des prépositions comme *en*, *pendant*, *vers*, *pour*, *dans*, etc, peuvent se transformer en GP introduit par *de* d'un déverbal. Dans les 57 cas rencontrés par les annotateurs, le GP était toujours optionnel et représentait un attribut plutôt qu'un actant du déverbal.

Le cas de la transformation du sujet du verbe en un GP introduit par *de* du déverbal est plus problématique. Les annotateurs en ont repéré un grand nombre (80 sur 584 soit 13,7%). Toutefois, ils ne sont pas toujours tombés d'accord sur ces séquences car une partie (18 sur 80) a été annotée comme **C=a** par au moins un annotateur. En fait, cela s'explique assez bien car la majorité de ces cas problématiques correspondent soit à des verbes pouvant être pronominaux, soit à des verbes dont le complément d'objet direct peut devenir sujet. Par exemple, nous pouvons dire que *les achats de biens durables redémarrent* mais aussi qu'*on redémarre les achats de biens durables*. Dans ce cas, le GP du déverbal *redémarrage* peut effectivement correspondre au sujet ou à l'objet de *redémarrer*. Dans le même ordre d'idée, si l'on parle de *la précarisation du marché du travail*, on peut considérer soit que *le marché du travail précarise les travailleurs*, soit que *la société précarise le marché du travail*.

Mis-à-part ces problèmes d'ambiguïté, une analyse fine des exemples annotés **subj** rejoint et complète l'analyse que nous avions faite à propos du cas **C=d**. Alors que la forme normale de la transformation du sujet dans le déverbal devrait être un GP introduit par *par*, il semble courant que cette préposition *par* soit remplacée par la préposition *de*. Cela est particulièrement vrai lorsque le verbe est intransitif comme *la dégringolade de sa production* dans le phrase *De plus, la Russie [...] paraît incapable de stopper la dégringolade de sa production [...]*. Dans ce cas très précis, il est même impossible d'utiliser un GP introduit par *par*. Pour les verbes transitifs dont un des compléments est introduit par *de*, les exemples du corpus montrent que le nom déverbal provenant du verbe perd son actant-sujet.

La conclusion de cette étude semble bien indiquer que notre hypothèse d'origine sur les transformations des cadres de sous-catégorisation des verbes vers ceux des déverbaux était globalement fondée. Nous avons toutefois dû revoir le cas du sujet des verbes qui peuvent sous certaines conditions devenir de vrais arguments du déverbal sous la forme d'un GP introduit par *de* et ajouter le cas de la transformation des compléments au datif en un GP au génitif :
- soit le sujet du verbe disparaît et est converti en un attribut optionnel du nom dérivé (en général un GP introduit par la préposition *par*) : $(subj/n \rightarrow attr/\perp)$, soit il est le complément unique du déverbal (en général un GP introduit par la préposition *de*), en particulier si le verbe est intransitif ou bien lorsque le verbe comporte normalement un complément qui est alors élidé : $(subj/n \rightarrow oi/g)$,
- l'objet direct ($C = a$), s'il n'est pas élidé, devient un GP introduit par *de* $(od/a \rightarrow oi/g)$,
- tout autre objet indirect, s'il n'est pas élidé, est tout simplement hérité sous réserve de sa présence dans le cadre (e.g $(oi/d \rightarrow oi/d)$ ou $(oi/o \rightarrow oi/o)$) avec la possibilité pour le datif de se retrouver transformé en un GP introduit par la préposition *de* à la place de la préposition *à* : $(oi/d \rightarrow oi/g)$,
- les GP circonstanciels au génitif sont transformés en GP attributifs au génitif.

Cette analyse complète celle de (Benveniste, 1966) qui ne prévoit que deux alternatives : génitif subjectif / génitif objectif.

Pour terminer ce tableau, nous nous sommes aperçus qu'il y a très peu de cas de déverbaux comportant deux GP introduits par *de* qui correspondent à deux compléments du verbe (contrairement au cas où l'un des deux GP introduit par *de* correspond à un argument circonstanciel du verbe). Un exemple intéressant provient du déverbal *radiation* et de son verbe d'origine *radier* qui comporte deux compléments, un complément d'objet direct $C = a$ et un complément d'objet indirect introduit par *de* $C = g$. *Et semble pencher, en privé, pour une radiation de son groupe du second marché boursier* : les deux GP *de son groupe* et *du second marché boursier* correspondent à *on a radié le groupe du second marché*.

Par contre, nous n'avons trouvé aucun cas de déverbal comportant un GP au génitif provenant du sujet du verbe et un autre GP du déverbal (pas forcément au génitif) provenant d'un complément de ce verbe. Il semble donc que les cadres de sous-catégorisation des déverbaux avec deux actants dont un provient du sujet du verbe transitif, s'ils existent, sont très restreints.

5 Discussion et perspectives

Notre calcul de la provenance des compléments au génitif des noms déverbaux ne donne pas encore une solution complète et définitive au problème du calcul des cadres de sous-catégorisation des déverbaux même s'il fournit une approximation tout-à-fait satisfaisante. En effet :
1) dans les contextes où le nom déverbal a deux compléments d'objet, nous n'avons pas de règle contextuelle définissant la provenance de chacun (nos règles sont individuelles) ;
2) et même dans le cas le plus fréquent où le nom déverbal n'a qu'un seul actant, nos règles ne résolvent pas le choix précis de l'actant du verbe d'origine dont il provient.

Pour arriver à ces règles il faut raffiner notre analyse préliminaire en considérant les noms déverbaux dans les *contextes d'actant*. Il s'agit des contextes où le nom déverbal $N[V_0]$ provenant du verbe V_0 et gouvernant son complément $GP(\mathbf{de})$ est lui-même subordonné à un verbe V_g, c'est-à-dire :

$$V_g \overset{C-obj}{\longrightarrow} N[V_0] \longrightarrow GP(\mathbf{de}).$$

Par exemple dans la phrase *Ce prince de la pensée vêtu en clochard ne semblait pas voir l'admiration [...] de ses élèves*, le déverbal $N[V_0]$=*admiration* du verbe V_0 =*admirer* comporte un GP $GP(\mathbf{de})$ =*de ses élèves* et est subordonné au verbe V_g =*voir*. Dans cet exemple, c'est le verbe *voir* qui fournit le complément d'objet direct de *admirer*. Aussi ce complément sort du cadre de sous-catégorisation du déverbal *admiration* tandis que le GP *de ses élèves* qui est le sujet du verbe d'origine *admirer* devient son attribut au génitif. Il s'agit donc de la transformation : $((subj/n \rightarrow attr/g);(od/a \rightarrow 0))$.

E. Paducheva, dans son travail classique (Paducheva, 1977), décrit pour le russe le calcul complet des réductions des choix de l'actant de V_0 dont provient GN(**de**) en fonction des actants différents de C dans le cadre de sous-catégorisation du verbe gouverneur V_g (c'est-à-dire ceux qui ne sont pas remplis par le déverbal). Pour élaborer les règles similaires pour les noms déverbaux français il nous faudra plus de données et une expertise linguistique additionnelle très fine qui devra regarder au-delà des simples groupes nominaux et de leur GP.

Références

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for French. *In* Abeillé, A., éditeur : *Treebanks : Building and Using Parsed Corpora*, pages 165–188. Kluwer, Dordrecht.

Alfared, R., Béchet, D. et Dikovsky, A. (2011). CDG Lab : a toolbox for dependency grammars and dependency treebanks development. *In* Gerdes, K., Hajicova, E. et Wanner, L., éditeurs : *Proc. of the 1st Intern. Conf. on Dependency Linguistics (Depling 2011)*, Barcelona, Spain. http ://depling.org/proceedingsDepling2011/.

Benveniste, E. (1966). *Problèmes de linguistique générale*. Gallimard, Paris.

Dikovsky, A. (2011). Categorial dependency grammars : from theory to large scale grammars. *In* Gerdes, K., Hajicova, E. et Wanner, L., éditeurs : *Proc. of the 1st Intern. Conf. on Dependency Linguistics (Depling 2011)*, Barcelona, Spain. http ://depling.org/proceedingsDepling2011/.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Fillmore, C. J. (1968). The case for case. *In* Bach, E. et Harms, R. T., éditeurs : *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart and Winston.

Fillmore, C. J. (1977). The case for case reopened. *In* Cole, P. et Sadock, J. M., éditeurs : *Syntax and Semantics 8 : Grammatical Relations*, pages 59–81. New York : Academic Press.

Grimshaw, J. (1990). *Argument Structure*. MIT Press, Cambridge, Mass.

Hathout, N., Namer, F. et Dal, G. (2002). An Experimental Constructional Database : The MorTAL Project. *In* Boucher, P., éditeur : *Many Morphologies*, pages 178–209. Cascadilla, Somerville, Mass.

Ho-Dac, L.-M. (2007). *La position initiale dans l'organisation du discours, une exploration en corpus*. Thèse de doctorat, Université Toulouse-le Mirail, France.

Mel'čuk, I. et Holodovič, A. (1970). To the theory of grammatical voice : (definition, calculus). *In Problemy lingvisticheskoj tipologii i struktury jazyka*. Nauka, Leningrad. (Rus.).

Mel'čuk, I. A. (2004). Actants in semantics and syntax 1,2. *Linguistics*, 42(1,2).

Namer, F. (2009). *Morphologie, lexique et traitement automatique des langues*. Hermes Science, Lavoisier.

Paducheva, E. (1977). O proizvodnyh diatezah otpredikatnyh imen v russkom jazyke. *In Problemy lingvisticheskoj tipologii i struktury jazyka*. Nauka, Leningrad. (Rus.).

Sagot, B. (2010). The lefff, a freely available and large-coverage morphological and syntactic lexicon for French. *In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Tanguy, L. et Hathout, N. (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du web. *In* Pierrel, J.-M., éditeur : *Actes de la 9eConférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2002)*, pages 245–254, Nancy. ATALA.

van den Eynde, K. et Mertens, P. (2003). La valence : l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, 13:63–104.

Van Valin, Jr., R. D. (1997). Generalized semantic roles and the syntax-semantics interface. *In* Corblin, F., Dobrovie-Sorin, C. et Marandin, J.-M., éditeurs : *Empirical Issues in Formal Syntax and Semantics 2 : Selected papers from the Colloque de Syntaxe et Semantique à Paris*, pages 373–388. Peter Lang.

Vectorisation, Okapi et calcul de similarité pour le TAL : pour oublier enfin le TF-IDF

Vincent Claveau
IRISA-CNRS
Campus de Beaulieu, 35042 Rennes, France
vincent.claveau@irisa.fr

RÉSUMÉ

Dans cette prise de position, nous nous intéressons au calcul de similarité (ou distances) entre textes, problématique présente dans de nombreuses tâches de TAL. Nous nous efforçons de montrer que ce qui n'est souvent qu'un composant dans des systèmes plus complexes est parfois négligé et des solutions sous-optimales sont employées. Ainsi, le calcul de similarité par TF-IDF/cosinus est souvent présenté comme « état-de-l'art », alors que des alternatives souvent plus performantes sont employées couramment dans le domaine de la Recherche d'Information (RI). Au travers de quelques expériences concernant plusieurs tâches, nous montrons combien ce simple calcul de similarité peut influencer les performances d'un système. Nous considérons plus particulièrement deux alternatives. La première est le schéma de pondération Okapi-BM25, bien connu en RI et directement interchangeable avec le TF-IDF. L'autre, la vectorisation, est une technique de calcul de similarité que nous avons développée et qui offrent d'intéressantes propriétés.

ABSTRACT

Vectorization, Okapi and computing similarity for NLP : say goodbye to TF-IDF

In this position paper, we review a problem very common for many NLP tasks: computing similarity (or distances) between texts. We aim at showing that what is often considered as a small component in a broader complex system is very often overlooked, leading to the use of sub-optimal solutions. Indeed, computing similarity with TF-IDF weighting and cosine is often presented as "state-of-theart", while more effective alternatives are in the Information Retrieval (IR) community. Through some experiments on several tasks, we show how this simple calculation of similarity can influence system performance. We consider two particular alternatives. The first is the weighting scheme Okapi-BM25, well known in IR and directly interchangeable with TF-IDF. The other, called vectorization, is a technique for calculating text similarities that we have developed which offers some interesting properties.

MOTS-CLÉS : Calcul de similarité, modèle vectoriel, TF-IDF, Okapi BM-25, vectorisation.

KEYWORDS: Calculating similarities, vector space model, TF-IDF, Okapi BM-25, vectorization.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 85–98,
Grenoble, 4 au 8 juin 2012.

1 Introduction

« Okapi, qu'est ce que c'est ? », « Quelle est la différence entre un lemme et un *stem* (racine) ? » Voici deux questions posées à plusieurs reprises dans des conférences de Traitement Automatique des Langues (TAL) pour la première et de Recherche d'Information (RI) pour la seconde. Elles illustrent le relatif cloisonnement des deux communautés concernées, bien que RI et TAL partagent un grand nombre de problématiques, et le langage (écrit ou oral) comme matériau de base. Dans cette prise de position, nous nous intéressons indirectement à la première question et donc à l'une des conséquences de la méconnaissance de certaines techniques de RI dans les applications du TAL. Précisons que le propos n'est pas ici de dresser un état de l'art complet sur les points de rencontre entre TAL et RI (pour un panorama de l'utilisation du TAL en RI, voir par exemple Moreau et Sébillot (2005)).Nous souhaitons simplement mettre en évidence l'importance du calcul de distance (ou de similarité), problématique largement traitée en RI, dans ces applications du TAL[1].

Au travers de quelques travaux, nous montrons dans cet article que la méthode de calcul de similarité, à système identique par ailleurs, influence grandement les résultats. Plus particulièrement, le but de cette communication est de faire valoir quatre points :

1. au sein d'un système de TAL nécessitant des calculs de similarité, le choix de la méthode de calcul doit être soigneusement étudié car il peut changer drastiquement les performances d'un système ;

2. la similarité basée sur le TF-IDF/cosinus n'est pas « état-de-l'art », comme on le lit trop souvent ;

3. des alternatives bien établies et tout aussi simples (par exemple Okapi) donnent de meilleurs résultats en général ;

4. enfin, des propositions récentes, notamment l'approche par vectorisation que nous avons développée, en donnent encore de meilleurs.

Cette position peut paraître d'autant plus facile à prendre qu'elle semble consensuelle. Pour autant, il est frappant de constater combien ces différents points sont loin d'être ancrés dans la communauté : par exemple, dans les actes des quatre dernières éditions de la conférence TALN, la pondération TF-IDF est citée dans 31 articles, contre 4 fois pour Okapi (principalement par les mêmes auteurs de plus).

Cet article est organisé de la manière suivante. La section 2 propose quelques rappels et considérations sur les calculs de distance entre textes, en détaillant notamment les approches TF-IDF et Okapi. Nous présentons ensuite en section 3 une méthode de calcul de distance aux propriétés intéressantes que nous avons développée et utilisée dans plusieurs tâches. Les trois sections suivantes présentent de manière concise trois applications (des références détaillant les approches sont données) dans lesquelles nous rapportons les performances au regard du choix du calcul de similarité.

[1]Ce travail a été réalisé dans le cadre du programme Quæro (`http://www.quaero.org`), financé par OSEO, agence nationale de valorisation de la recherche.

2 Calcul de distance et modèle vectoriel

2.1 Représentation vectorielle et sacs de mots

Dés lors que l'on manipule des données (des textes dans notre cas), il est souvent possible de les décrire avec un ensemble (fini et fixe) de valeurs. Ces valeurs, et donc les objets qu'elles représentent, peuvent alors être interprétées comme des vecteurs formant un espace vectoriel. L'avantage de cette représentation est que l'on sait faire certaines opérations assez facilement dans de tels espaces, notamment des calculs de distance/similarité très rapides.

Dans le cas des textes, ces représentations consistent souvent à considérer le document (ou n'importe quelle donnée textuelle) comme un sac-de-mots, c'est-à-dire un ensemble non structuré, sans information sur la séquentialité des mots dans le texte. Usuellement, on calcule pour chaque mot présent dans le document une valeur reflétant son importance comme descripteur du document (cf. schéma de pondération ci-après). Les mots du vocabulaire (ou de la collection de documents traitée) absents du document ont une valeur nulle. Finalement, le texte est donc décrit comme un vecteur d'un espace ayant pour dimensions tous les mots du vocabulaire. Ces espaces sont donc très grands (par exemple $\mathbb{R}^{100\,000}$ pour une collection moyenne monolingue), mais les vecteurs sont aussi très creux, ce qui permet une représentation compacte, mais surtout de rendre certains calculs très rapides (cf. sous-section suivante).

Cette représentation du texte est très utilisée, avec différentes variantes (sac-de-ngrams ou plus généralement sac-de-*features*...). Cela s'explique par sa simplicité de mise en œuvre, ses manipulations efficaces, et bien que pauvre (elle ne nécessite aucun traitement complexe ou de données externes), elle donne souvent de bons résultats en pratique.

2.2 Distances dans les espaces vectoriels

Dans le modèle vectoriel (Salton *et al.*, 1975), les documents et les requêtes sont représentés par des vecteurs. L'appariement entre un document et une requête se fait en calculant la proximité de leur vecteur, le plus souvent en utilisant une distance de type Lp. Pour deux documents d et q (ou un document et une requête), la distance Lp est définie comme suit (on note V le vocabulaire de la collection, c'est-à-dire, l'ensemble des termes d'indexation de tous les documents) :

$$\delta_{Lp}(q,d) = \sqrt[p]{\sum_{t \in V} |q_t - d_t|^p}$$

Les valeurs les plus courantes sont $p = 1$ (distance L1, dite *manhattan* ou *city-block*), $p = 2$ (distance L2 ou euclidienne), ou $p \to \infty$ (distance de Chebyshev) ; p n'est pas nécessairement un entier mais doit être supérieur à 1 pour que soit respectée l'inégalité triangulaire. Rappelons que la distance basée sur le cosinus habituellement utilisée en RI textuelle est équivalente (i.e. produit le même ordonnancement des documents) à la distance L2 lorsque les vecteurs sont normalisés : $\delta_{L2}(q,d) = \sqrt{2 - 2 * \delta_{cos}(q,d)}$

Les distances Lp avec p pair, et donc la distance L2 et le cosinus en particulier, ont l'avantage d'être rapides à calculer quand les vecteurs sont normés (sous Lp) et creux. En effet, seules

les composantes des vecteurs qui sont non nulles à la fois pour la requête et le document interviennent dans le calcul. Dans le modèle vectoriel standard, cela revient à ne s'intéresser qu'aux mots partagés par la requête et le document puisque les termes absents des documents (ou de la requête) ont une pondération nulle. Cela explique les mise en œuvre par fichiers inversés et la grande rapidité de cette étape d'appariement, notamment lorsque l'on manipule des requêtes de quelques mots.

Ce calcul de distance devient beaucoup plus coûteux, dans le cas où les vecteurs sont de grandes dimensions mais ne sont pas creux. Mais il existe des techniques de calcul rapide des distances dans les espaces vectoriels. De manière générale, ces techniques troquent un fort gain en temps de réponse contre une légère perte de précision, les éléments retournés étant simplement similaires et non *les plus* similaires. Une approche est de découper l'espace des données en portions et de n'effectuer des recherches que sur une ou plusieurs portions (Stein, 2007; Datar *et al.*, 2004) et/ou de calculer des approximations de la distance réelle (Lejsek *et al.*, 2008).

2.3 Pondérations

Les pondérations sont une des caractéristiques majeures du modèle vectoriel introduit par Salton (1975). Elles permettent de caractériser non seulement la présence ou l'absence de termes dans les documents, mais également leur importance relative pour décrire le contenu du document : un poids w_{ij} est attribué à chaque terme t_i du document d_j ; plus ce poids est important, plus t_i est considéré comme pertinent pour décrire d_j.

Le célèbre TF-IDF décompose la pertinence d'un terme selon deux heuristiques. La première est le TF, issue des travaux de Luhn (1958) : plus le terme est fréquent dans le document, plus il est jugé pertinent. La seconde, l'IDF, est souvent attribuée à Spärck Jones (1972) : plus un terme apparaît dans un grand nombre de document, moins il est pertinent. Sa formulation la plus connue est (tf est le nombre d'occurrence ou la fréquence du terme t dans le document considéré, df sa fréquence documentaire, c'est-à-dire le nombre de documents dans lequel il apparaît, N est le nombre total de documents) :

$$w_{TF-IDF}(t,d) = tf(t,d) * \log(N/df(t))$$

Outre le TF-IDF/cosinus, beaucoup de techniques (pondérations et similarités) pour calculer des similarités dans des espaces algébriques existent (Tirilly, 2010, pour une revue). Parmi celles-ci, la pondération Okapi[2] est devenue une référence grâce aux très bons résultats qu'elle permet d'obtenir sur de nombreuses tâches de RI. Cette pondération a initialement été proposée comme modèle de similarité dans un cadre probabiliste (Robertson *et al.*, 1998) ; ce cadre repose sur le principe de classement probabiliste (PRP, *Probability Ranking Principle* que Robertson énonce ainsi : *If retrieved documents are ordered by decreasing probability of relevance on the data available, then the system's effectiveness is the best to be gotten for the data.*

Ce principe généraliste est en pratique décliné en modèles pouvant s'interpréter comme des pondérations dans un modèle vectoriel. Le modèle Okapi peut ainsi être vu comme un

[2]La formule de cette pondération s'appelle en réalité BM-25, mais est souvent appelée Okapi du nom du premier système l'ayant implémenté.

TF-IDF prenant mieux en compte la longueur des documents. Sa définition est donnée dans l'équation 1 qui indique le poids du terme t dans le document d ($k_1 = 2$ and $b = 0.75$ sont des constantes, dl la longueur du document, dl_{avg} la longueur moyenne des documents).

$$\begin{aligned} w_{BM25}(t,d) &= TF_{BM25}(t,d) * IDF_{BM25}(t) \\ &= \frac{tf(t,d) * (k_1+1)}{tf(t,d) + k_1 * (1 - b + b * dl(d)/dl_{avg})} * \log \frac{N - df(t) + 0.5}{df(t) + 0.5} \quad , \end{aligned} \tag{1}$$

La partie TF_{BM25} est dérivée d'un modèle probabiliste de la fréquence des termes dans les documents, le modèle 2-Poisson de Harter (Spärck Jones *et al.*, 2000). Ce modèle représente la distribution des termes dans les documents comme un mélange de deux distributions de Poisson : l'une représentant la fréquence des termes pertinents pour décrire le document, l'autre celle des termes non-pertinents (Harter, 1975). C'est dans ce TF qu'est intégré une normalisation en fonction de la taille du document.
La partie IDF_{BM25} est une simplification d'une formule dérivée du PRP (Spärck Jones *et al.*, 2000), théoriquement optimale mais nécessitant des données d'apprentissage. L'IDF obtenu est très proche de l'IDF standard et confirme le bien-fondé de la formulation empirique.

Enfin, signalons que de nombreuses autres pondérations très performantes ont été proposées en RI, comme les modèles DFR (Divergence from Randomness, proposés par Amati et Van Rijsbergen (2002)) ou les modèles de langues (Ponte et Croft, 1998). Ces deux approches construisent des mesures de similarités basées sur des modèles probabilistes de fréquence des termes. La encore, ces modèles peuvent s'interpréter comme des pondérations dans un modèle vectoriel.

3 Vectorisation ou distance du second ordre

En plus des techniques de pondération exposées précédemment, nous proposons une autre alternative au TF-IDF que nous avons développée. Il s'agit de la vectorisation, dont nous décrivons dans cette section les principes.

3.1 Principe

La vectorisation est une technique de plongement (*embedding*) permettant de projeter un calcul de similarité quelconque entre deux documents (ou un document et une requête pour la RI) dans un espace vectoriel. Son principe est relativement simple : pour chaque document de la collection considérée, il consiste à calculer avec une mesure de similarité, quelle qu'elle soit, des scores de similarité entre ce document et m documents-pivots. Cette similarité est dite de premier ordre dans la suite de cet article. Les m scores obtenus forment ainsi un vecteur de m dimensions représentant le document (cf. figure 1) et place le document dans un nouvel espace vectoriel. Dès lors, la comparaison de deux documents (ou d'un document et d'une requête) peut donc s'effectuer de manière standard dans ce nouvel espace, par exemple en calculant une distance L2. C'est la similarité de second ordre.

Il est important de noter que la vectorisation change l'espace de représentation. Il ne s'agit donc pas seulement d'une réduction de l'espace ou d'une approximation de la distance

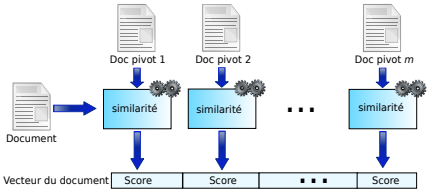


FIGURE 1 – Schématisation du principe de vectorisation d'un document

originelle comme proposée par exemple dans les travaux de (Bourgain, 1985). Il ne s'agit pas non plus d'une orthogonalisation, comme pour les approches LSI/LSA. C'est ce changement d'espace qui est à l'origine de deux propriétés intéressantes.
D'une part, cet *embedding* permet de réduire la complexité quand le calcul de similarité de premier ordre est trop coûteux pour être utilisé en-ligne (Claveau *et al.*, 2010).
D'autre part, la vectorisation permet que deux documents soient considérés comme proches s'ils sont proches des mêmes documents-pivots. Cette comparaison indirecte, ou affinité du second ordre, permet par exemple de mettre en relation des documents textuels qui ne contiennent pourtant aucun mot en commun.

3.2 Constitution des documents-pivots

De nombreuses alternatives sont possibles pour constituer les m documents-pivots, en fonction du problème abordé. Dans le principe, ces pivots ne sont pas nécessairement issus de la collection traitée, même s'il semble plus raisonnable que ce soit le cas ; on a ainsi une assurance plus grande d'une bonne adéquation, notamment du vocabulaire, de ces pivots avec les documents à traiter. Dans les expériences présentées ci-après, nous indiquons pour chaque application comment ces pivots ont été constitués.

3.3 Calcul de distances

Notre processus de vectorisation permet de ramener n'importe quel modèle de similarité à une représentation vectorielle. La comparaison de deux documents se fait donc par un calcul de distance entre vecteurs, comme dans le modèle vectoriel classique. Les vecteurs sont normalisés (en cohérence avec la distance utilisée) ; dans les expériences présentées dans cet article, nous utilisons une distance L2 (les vecteurs sont donc normalisés en L2). À la différence du modèle vectoriel classique dans lequel les vecteurs représentant les documents sont généralement extrêmement creux, les vecteurs obtenus par vectorisation n'ont pas forcément beaucoup de composantes à 0. Comme nous l'avons expliqué précédemment, cela rend le calcul de distance beaucoup plus coûteux s'il est fait de manière exhaustive et exacte. Dans les expériences reportées ci-dessous, les quantités de données manipulées permettent un tel calcul exact ; nous avons donc pris le parti d'évaluer les résultats sans le biais d'une recherche approximative. Les calculs de distance se font donc de manière classique avec la

distance L2.

4 Utilisation en recherche d'information

Dans les systèmes de recherche d'information, les modèles vectoriels sont très largement utilisés pour calculer les distances entre requêtes et textes. Il existe de nombreux articles ont comparé les mesures/pondérations standard, comme le TF-IDF et Okapi (Savoy, 2005, *inter alia*). Dans cette section, nous rapportons quelques expériences (Claveau *et al.*, 2010, pour les détails) les comparant également à la vectorisation.

4.1 Description de la tâche

Pour ces évaluations, nous utilisons deux collections de RI aux caractéristiques très différentes, en français et provenant de la campagne d'évaluation Amaryllis. La première est la collection ELDA, petite collection de 3500 documents issus de questions/réponses de la commission européenne, accompagnée de 19 requêtes. La seconde est la collection INIST composée de 160 000 documents (résumés d'articles de diverses disciplines scientifiques) et de 30 requêtes. Pour ces deux collections, les requêtes sont composées de plusieurs champs : titre, corps, description et concepts associés. Dans les expériences reportées ci-dessous, seuls le titre et le corps sont utilisés. À chaque requête est associée la liste des documents pertinents qui sont attendus en réponse et donc utilisés pour évaluer les performances des systèmes de RI. Dans les expériences rapportées ci-dessous, ces performances sont mesurées en utilisant les mesures classiques de la RI, à savoir la MAP (*Mean Average Precision*) et les précisions obtenues pour divers seuils de documents (5 premiers documents retournés, 10 premiers...) moyennées sur l'ensemble des requêtes.

4.2 Approches proposées

Pour ces expériences, un système de RI vectoriel a été implémenté dont seul le calcul des distances varie. Basé sur ces distances, les documents sont proposés par similarité décroissante avec les requêtes. Pour le calcul par vectorisation, les m documents-pivots sont choisis comme des concaténations aléatoires de documents de la collection formant une partition de l'ensemble des documents. Différents nombres de pivots (et donc la dimension de l'espace résultant) sont testés, et les similarités de premier ordre (servant à construire les vecteurs) sont calculées avec Okapi.

4.3 Résultats

Les tableaux 1 et 2 présentent les résultats de ce système, selon les différentes méthodes de calcul de distances. En prenant Okapi comme base de comparaison, on indique les améliorations statistiquement significatives (t-test avec $p = 0.05$) en gras et les dégradations significatives en italiques. Plusieurs éléments en ressortent. D'une part, le système basé sur

BM-25 domine très largement et dans tous les cas celui s'appuyant sur TF-IDF. Ce résultat est conforme à l'ensemble des expériences de ce type dans la littérature. Les résultats de la vectorisation sont moins tranchés en terme de performances globales (MAP), puisque les résultats sont très fortement améliorés pour la petite collection, mais comparable pour la collection INIST. En revanche, il apparaît clairement et dans tous les cas la propriété de la vectorisation de trouver plus de documents pertinents (précision améliorée pour des seuils hauts). C'est cette propriété, reposant sur la capacité de juger de la similarité de deux documents mêmes s'ils ne partagent pas de termes communs, qui est particulièrement utile pour certaines tâches de TAL.

	TF-IDF	Okapi	Vectorisation ($m = 1\,750$)	Vectorisation ($m = 3\,500$)
MAP	*28.36 (-21.7 %)*	36.22	39.01 (+7.7 %)	**43.46 (+20 %)**
P@10	*36.18 (-24.1 %)*	47.67	**51.67 (+8.4 %)**	**54.67 (+14.7 %)**
P@50	*26.37 (-12.1 %)*	30.00	29.07 (-3.1 %)	**33.53 (+11.8 %)**
P@100	*19.36 (-6.6 %)*	20.73	19.87 (-4.2 %)	21.50 (+3.7 %)
P@500	6.04 (-0.9 %)	5.99	5.71 (-4.8 %)	6.17 (+2.9 %)
P@1000	3.16 (-0.4 %)	3.15	3.15 (0 %)	**3.27 (+3.9 %)**
P@3000	1.08 (+0.4 %)	1.07	**1.17 (+9.0 %)**	**1.18 (+10 %)**

TABLE 1 – Performances des systèmes sur la collection ELDA

	TF-IDF	Okapi	Vectorisation ($m = 10\,000$)
MAP	*10.62 (-26.9 %)*	14.52	14.26 (-1.8 %)
P@10	*24.00 (-29.4 %)*	34.00	*29.00 (-14.7 %)*
P@50	*14.40 (-22.0 %)*	18.47	18.00 (-2.5 %)
P@100	*10.93 (-8.9 %)*	12.00	**13.47 (+12.2 %)**
P@500	4.16 (-2.6 %)	4.27	**4.93 (+15.3 %)**
P@1000	2.40 (-2.4 %)	2.46	**2.89 (+17.3 %)**
P@3000	1.00 (-1 %)	1.01	**1.12 (+10.9 %)**

TABLE 2 – Performances des systèmes sur la collection INIST

5 Utilisation pour la fouille de texte

Comme nous le soulignions dès l'introduction, calculer des distances entre textes n'est pas utile que pour la RI, mais aussi pour beaucoup de tâches du TAL. Dans cette section, nous illustrons l'importance du choix de la similarité dans un contexte de fouille de textes.

5.1 Description de la tâche

Cette tâche était l'une de celles proposées dans le cadre du Défi Fouille de Texte (DeFT) 2011 (Grouin et Forest, 2011). Elle consistait à retrouver l'année de parution d'extraits d'articles de journaux OCRisés publiés entre 1801 et 1944. Les participants disposaient de

données d'apprentissage (extraits d'articles avec leur date de parution), et de données de test (extraits d'articles pour lesquels il faut fournir la date de parution). Deux sous-tâches étaient proposées : l'une avec des extraits de 300 mots et l'autre avec des extraits de 500 mots.

La difficulté de ce défi tenait d'une part à la qualité très dégradée des textes OCRisés et au grand nombre de classes possibles (144 années). La mesure d'évaluation mise en place par les organisateurs permettait de pondérer les erreurs de datation en fonction de leur distance à l'année réelle. Cette mesure va de 0 pour un document éloigné de plus de 15 ans, à 1 quand la prédiction de date est exacte.

5.2 Approches proposées

L'approche que nous avons proposée lors de notre participation repose sur un apprentissage paresseux (*lazy-learning*), à savoir les k-plus proches voisins (k-ppv), qui se veut souple et adapté à la tâche. Dans cette approche, une instance inconnue est classée en trouvant les k instances connues les plus similaires et en lui assignant la classe majoritaire de ces instances. Il n'y a donc pas à proprement parler d'apprentissage, d'où le nom de *lazy-learning*, mais l'induction repose sur la calcul de similarité, qui permet de trouver les plus proches voisins, et la mise en œuvre du vote (Beyer *et al.*, 1999, pour les autres paramètres pouvant intervenir).

Ce calcul de similarité est donc central. Lors du défi, nous avons utilisé la pondération Okapi, et cette simple approche nous a permis d'être classés premiers. Dans la sous-section suivante, nous avons repris ces expériences et nous présentons en plus les résultats obtenus avec un TF-IDF et par vectorisation. Pour ces expériences, les documents-pivots sont simplement des concaténations aléatoires des articles de l'ensemble d'entraînement. La seule contrainte est que les documents concaténés doivent avoir la même année de parution. Chaque dimension de notre nouvel espace vectoriel correspond donc à une année (et éventuellement, plusieurs dimensions peuvent porter sur la même année). Comme précédemment, chaque document de l'ensemble d'entraînement est décrit à l'aide de ces pivots : sa distance (similarité de premier ordre) à chacun des pivots est calculée en utilisant Okapi, ce qui forme l'ensemble de ses coordonnées dans le nouvel espace. Il est fait de même pour les documents test. Finalement, les plus proches voisins d'un document test sont donnés par une distance $L2$.

5.3 Résultats

Le tableau 3 recense les résultats de l'approche k-ppv avec une distance de type Okapi tels que publiés, ainsi que la même approche utilisant cette fois la distance par vectorisation. À des fins de comparaison, nous indiquons également les résultats du même système qui utiliserait cette fois un TF-IDF standard avec une similarité cosinus, ainsi que ceux obtenus par le LIMSI, deuxième système le plus performant.

Il apparaît encore une fois l'intérêt de l'utilisation d'une pondération BM-25 comparée au TF-IDF standard. Celle-ci a permis d'obtenir les meilleurs résultats avec un algorithme simple de vote lors du challenge, alors que le TF-IDF aurait fait apparaître le système comme moins adapté que la proposition du LIMSI. La vectorisation permet en plus de dépasser ces résultats ; grâce au choix des documents-pivots, les documents de même année de parution peuvent avoir des représentations vectorielles proches, même s'ils ne partagent pas de mots communs.

	extraits de 300 mots	extraits de 500 mots
système LIMSI	0.378	0.452
k-ppv TF-IDF	0.364	0.398
k-ppv Okapi	0.430	0.472
k-ppv Vectorisation	**0.466**	**0.505**

TABLE 3 – Résultats sur la tâche de datation de DeFT 2011 d'un système k-ppv avec différents calculs de similarité et d'une *baseline* (système du LIMSI)

6 Utilisation pour la segmentation thématique

Outre la fouille de textes, le calcul de similarité est aussi utile dans certains systèmes dédiés à d'autres applications classiques du TAL, comme par exemple la segmentation thématique. Comme dans les sections précédentes, nous illustrons l'influence du choix du calcul de similarité dans un tel système de segmentation reposant sur ce calcul (Claveau et Lefèvre, 2011).

6.1 Description de la tâche

La segmentation thématique est une tâche classique du TAL consistant à diviser un texte ou un flux textuel en parties thématiquement cohérentes. De nombreuses méthodes ont été proposées dans la littérature, que l'on peut diviser en deux familles. Il y a d'une part des approches s'appuyant sur des propriétés de formatage des documents, ou sur la détection de marqueurs discursifs (Christensen *et al.*, 2005). L'autre grande famille d'approches s'appuie sur le contenu des documents pour détecter les changements de thème. C'est dans cette famille que s'inscrit notre approche et beaucoup des systèmes existants, tels que SEGMENTER (Kan *et al.*, 1998), l'approche Utiyama et Isahara (Utiyama et Isahara, 2001; Guinaudeau *et al.*, 2010), DOTPLOTTING (Reynar, 2000), C99 (Choi, 2000), TEXT-TILING (Hearst, 1997).

Ces différentes approches de l'état de l'art ont été comparées, que ce soit sur l'anglais (Choi, 2000) ou le français (Sitbon et Bellot, 2004). À des fins de comparaison, nous réutilisons ces données[3] sur le français pour évaluer l'importance du calcul de similarité dans ce contexte. Celles-ci ont été constituées artificiellement en segmentant et mélangeant des articles du journal *Le Monde* de plusieurs catégories (Sports, Arts...) et des extraits de la bible. Pour répondre à la critique d'artificialité de ces données, nous utilisons également deux autres jeux, composés respectivement de transcriptions de journaux TV et d'émissions de reportage développés par Guinaudeau *et al.* (2010). La segmentation de référence a été effectuée indépendamment en considérant qu'un changement de thème a lieu à chaque changement de reportage. Cette définition de la rupture thématique a l'avantage de correspondre à un besoin applicatif réel et bien défini. Les bandes-son de ces deux corpus ont été transcrites automatiquement par le système de reconnaissance de la parole IRENE (Huet *et al.*, 2010).

[3]Nous remercions L. Sitbon pour la mise à disposition de ces données et des systèmes de l'état-de-l'art.

6.2 Approche proposée

Notre technique de segmentation thématique cherche à détecter les ruptures thématiques en comparant (i.e. en calculant une distance) le contenu avant et après chaque segment. Plus la distance est importante, plus la rupture est probable. D'un point de vue technique, elle adapte un principe utilisé en segmentation d'image : la ligne de partage des eaux (*watershed*). Celle-ci consiste à représenter l'image à segmenter comme un relief (ou surface topographique) en calculant un gradient de l'image pour faire ressortir les zones de fortes variations (par exemple de luminance d'un pixel). Une inondation progressive du relief par ses minima est alors simulée et des digues sont (virtuellement) construites pour séparer les différents bassins associés à chaque minimum. À l'issue du processus, ces digues représentent les lignes de partage des eaux, ou autrement dit les frontières des régions.

Dans le cas de texte, il n'y a qu'une dimension à considérer, et le gradient est vu comme une mesure de distance textuelle. Un gradient est donc calculé entre chaque phrase (ou groupe de souffle dans le cas de transcriptions) : on calcule la similarité entre les phrases précédentes et suivantes de chaque frontière potentielle (Claveau et Lefèvre, 2011, pour une présentation plus détaillée). Il faut noter qu'on ne compare pas seulement le groupe de souffle précédent au groupe de souffle suivant, mais on considère les n précédents et les n suivants. Là encore, ce calcul de similarité peut se faire en utilisant les outils standard de RI comme le TF-IDF, ou bien Okapi, ou encore par vectorisation.

6.3 Résultats

Pour évaluer la qualité des segmentation sur nos jeux de données, nous indiquons la mesure WindowDiff (WD), habituellement utilisée pour l'évaluation de ce type de tâche, et qui peut être vu comme un taux d'erreurs (Pevzner et Hearst, 2002, pour une présentation détaillée). À des fins de comparaison avec des systèmes n'utilisant pas le WindowDiff, nous utilisons aussi la F-mesure (F1).

	News		*Reports*	
Méthodes	F1 (%)	WD	F1 (%)	WD
Utiyama (Utiyama et Isahara, 2001)	59.44	-	51.09	-
Guinaudeau (Guinaudeau *et al.*, 2010)	61.41	-	62.92	-
DotPlot (Reynar, 2000)	36.42	0.4472	49.49	0.2125
c99 (Choi, 2000)	50.25	0.3646	57.42	0.1893
TextTiling (Hearst, 1997)	38.73	0.313	23.38	0.3456
TF-IDF + Watershed	43.04	0.3833	60.12	0.1844
Okapi + Watershed	60.04	0.2571	69.22	0.1428
Vectorisation + Watershed	**64.4**	**0.2269**	**73.31**	**0.1181**

TABLE 4 – Performances des systèmes de segmentation sur la collection *News* et *Reports*

Sur cette application, les résultats sont encore une fois très homogènes, quel que soit le corpus d'évaluation. À système équivalent, le choix de la mesure a une très grande influence, et on constate comme précédemment que le TF-IDF est surpassé par Okapi, lui-même légèrement dépassé par l'approche par vectorisation. On note encore qu'à architecture similaire, le

Méthodes	Sous-corpus de test				
	Sciences	Éco	Sports	Arts	Bible
Meilleur système	0.2132	0.2243	0.2839	0.2811	0.3139
(d'après Sitbon et Bellot (2004))	C99	C99	C99	C99	DOTPLOT
TF-IDF + Watershed	0.2964	0.2996	0.3205	0.3560	0.3702
Okapi + Watershed	0.2135	0.2177	0.2654	0.2729	0.2981
Vectorisation + Watershed	**0.1967**	**0.1836**	**0.2582**	**0.2587**	**0.2931**

TABLE 5 – Performances (WindowDiff) des systèmes de segmentation sur la collection de Sitbon et Bellot (2004)

système avec TF-IDF aurait été écarté, ses performances étant largement inférieures à l'état-de-l'art.

7 Conclusion

Le calcul de similarité entre textes est une problématique importante pour le TAL. Pour autant, ce qui n'est souvent qu'un composant dans des systèmes plus complexes est parfois négligé et des solutions sous-optimales sont employées. Dans les expériences précédentes relevant du cas très fréquent de la représentation vectorielle, on a montré combien un simple changement de pondération (de TF-IDF vers Okapi) pouvait modifier les performances finales d'un tel système. Ainsi, il convient de s'interroger si l'emploi de ces calculs de similarité sous-optimaux n'a pas condamné des systèmes qui auraient été par ailleurs jugés performants à condition d'utiliser un module de similarité plus état-de-l'art.

Outre l'utilisation de pondérations plus actuelles comme Okapi, nous avons montré que la vectorisation offre d'excellentes performances et des propriétés intéressantes. D'un point de vue technique, elle conserve le cadre vectoriel qui permet au besoin l'emploi d'outils de calcul de distances efficaces. Une autre propriété intéressante pour le TAL est la possibilité de juger de la similarité de textes même lorsqu'ils ne partagent pas de mots communs.

Quelques remarques s'imposent en complément de ces résultats. Tout d'abord, la représentation sac-de-mots n'est pas la représentation la plus adaptée à tous les problèmes de similarité. Les approches par modèles de langue, outils communs du TAL, sont par exemple plus adaptées dès lors que l'aspect séquentiel du texte est important (Ebadat *et al.*, 2011, pour un exemple). Il en va de même pour les similarités entre arbres syntaxiques ou graphes sémantiques.

D'un point de vue plus général, la relative séparation des communautés RI et TAL explique sans doute le relatif manque d'importance accordée aux calculs de similarité dans certaines tâches du TAL. Cela milite pour des rapprochements thématiques ponctuels entre ces communautés, par exemple par le biais de tutoriels, numéros spéciaux de journaux, organisation de conférences jointes... Enfin, cela peut passer aussi par des enseignements dédiés ; on remarque en effet que les descriptifs des principales formations TAL, lorsque disponibles, font rarement état de modules abordant l'indexation ou la recherche d'information.

Références

AMATI, G. et VAN RIJSBERGEN, G. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*.

BEYER, K., GOLDSTEIN, J., RAMAKRISHNAN, R. et SHAFT, U. (1999). When is "nearest neighbor" meaningful ? *In Proceedings of the Int. Conf. on Database Theory*, pages 217–235.

BOURGAIN, J. (1985). On Lipschitz embedding of finite metric spaces in hilbert space. *Israel Journal of Mathematics*, 52(1).

CHOI, F. Y. Y. (2000). Advances in domain independent linear text segmentation. *In Proceedings of the 1st metting of the North American Chapter of the Association for Computational Linguistics*, USA.

CHRISTENSEN, H., KOLLURU, B., GOTOH, Y. et RENALS, S. (2005). Maximum entropy segmentation of broadcast news. *In Proceedings of the 30th IEEE ICASSP*.

CLAVEAU, V. et LEFÈVRE, S. (2011). Topic Segmentation of TV-streams by mathematical morphology and vectorization. *In Proceedings of InterSpeech*, pages 1105–1108, Italie.

CLAVEAU, V., TAVENARD, R. et AMSALEG, L. (2010). Vectorisation des processus d'appariement document-requête. *In 7e conférence en recherche d'informations et applications, CORIA'10*, pages 313–324, Sousse, Tunisie.

DATAR, M., IMMORLICA, N., INDYK, P. et MIRROKNI, V. (2004). Locality-sensitive hashing scheme based on p-stable distributions. *In Proc. of the 20th ACM Symposium on Computational Geometry*, Brooklyn, New York, USA.

EBADAT, A. R., CLAVEAU, V. et SÉBILLOT, P. (2011). Using shallow linguistic features for relation extraction in bio-medical texts. *In Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles, TALN'11*, volume 2, pages 125–130, Montpellier, France.

GROUIN, C. et FOREST, D., éditeurs (2011). *Actes de l'atelier Défi Fouille de Textes (DeFT'11)*, Montpellier, France.

GUINAUDEAU, C., GRAVIER, G. et SÉBILLOT, P. (2010). Improving ASR-based topic segmentation of TV programs with confidence measures and semantic relations. *In Proc. Annual Intl. Speech Communication Association Conference (Interspeech)*.

HARTER, S. (1975). A probabilistic approach to automatic keyword indexing. *Journal of the american society for information science*, 26(6):197–206.

HEARST, M. (1997). Text-tiling : segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

HUET, S., GRAVIER, G. et SÉBILLOT, P. (2010). Morpho-syntactic post-processing with n-best lists for improved French automatic speech recognition. *Computer Speech and Language*, 24(4):663–684.

KAN, M.-Y., KLAVANS, J. L. et MCKEOWN, K. R. (1998). Linear segmentation and segment significance. *In Proceedings of the 6th International Workshop of Very Large Corpora (WVLC-6)*.

LEJSEK, H., ASMUNDSSON, F., JÓNSSON, B. et AMSALEG, L. (2008). Nv-tree : An efficient disk-based index for approximate search in very large high-dimensional collections. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 99(1).

LUHN, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal on Research and Development*, 2(2).

MOREAU, F. et SÉBILLOT, P. (2005). Contributions des techniques du traitement automatique des langues à la recherche d'information. Rapport technique 1690, IRISA.

PEVZNER, L. et HEARST, M. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*.

PONTE, J. M. et CROFT, W. B. (1998). A language modeling approach to information retrieval. *In Proc. of SIGIR*, Melbourne, Australie.

REYNAR, J. C. (2000). *Topic Segmentation : Algorithms and applications*. Thèse de doctorat, University of Pennsylvania.

ROBERTSON, S. E., WALKER, S. et HANCOCK-BEAULIEU, M. (1998). Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive. *In Proceedings of the 7th Text Retrieval Conference, TREC-7*, pages 199–210.

SALTON, G. (1975). *A Theory of Indexing*. Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphy.

SALTON, G., WONG, A. et YANG, C. S. (1975). A vector space model for automatic indexing. *Comm. of the ACM*, 18(11).

SAVOY, J. (2005). Comparative study of monolingual and multilingual search models for use with asian languages. *ACM Transactions on Asian Languages Information Processing*, 4(2).

SITBON, L. et BELLOT, P. (2004). Évaluation de méthodes de segmentation thématique linéaire non supervisées après adaptation au fran cais. *In Actes de la conférence Traitement automatique des langues*, Fez, Tunisie.

SPÄRCK JONES, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1).

SPÄRCK JONES, K., WALKER, S. G. et ROBERTSON, S. E. (2000). Probabilistic model of information retrieval : Development and comparative experiments. *Information Processing and Management*, 36(6).

STEIN, B. (2007). Principles of hash-based text retrieval. *In Proc. of SIGIR*, Amsterdam, Pays-Bas.

TIRILLY, P. (2010). *Traitement automatique des langues pour l'indexation d'images*. Thèse de doctorat, Université de Rennes 1.

UTIYAMA, M. et ISAHARA, H. (2001). A statistical model for domain-independent text segmentation. *In Proceedings of the 9th conference of the ACL*.

TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe

Christophe Benzitoun[1] Karën Fort[2,3] Benoît Sagot[4]
(1) ATILF, Nancy Université & CNRS, 44, avenue de la Libération, BP 30687, 54063 Nancy cedex
(2) INIST-CNRS, 2 allée de Brabois, 54500 Vandoeuvre-lès-Nancy
(3) LIPN, Université Paris 13 & CNRS, 99 av. J.B. Clément, 93430 Villetaneuse
(4) Alpage, INRIA Paris–Rocquencourt & Université Paris 7, Rocquencourt, France
christophe.benzitoun@atilf.fr, karen.fort@inist.fr, benoit.sagot@inria.fr

RÉSUMÉ

Nous présentons dans cet article un travail portant sur la création d'un corpus de français parlé spontané annoté en morphosyntaxe. Nous détaillons la méthodologie suivie afin d'assurer le contrôle de la qualité de la ressource finale. Ce corpus est d'ores et déjà librement diffusé pour la recherche et peut servir aussi bien de corpus d'apprentissage pour des logiciels que de base pour des descriptions linguistiques. Nous présentons également les résultats obtenus par deux étiqueteurs morphosyntaxiques entrainés sur ce corpus.

ABSTRACT

TCOF-POS : A Freely Available POS-Tagged Corpus of Spoken French

This article details the creation of TCOF-POS, the first freely available corpus of spontaneous spoken French. We present here the methodology that was followed in order to obtain the best possible quality in the final resource. This corpus already is freely available and can be used as a training/validation corpus for NLP tools, as well as a study corpus for linguistic research. We also present the results obtained by two POS-taggers trained on the corpus.

MOTS-CLÉS : Etiquetage morpho-syntaxique, français parlé, ressources langagières.

KEYWORDS: POS tagging, French, speech, language resources.

1 Introduction

L'annotation automatique du français parlé est généralement réalisée par le biais de prétraitements de corpus ou d'adaptation d'outils existant pour le texte (Dister, 2007; Blanc *et al.*, 2008). Une autre solution peut consister à masquer certains phénomènes tels que les "disfluences" (répétitions, amorces de mots, etc.) (Valli et Véronis, 1999). Pourtant, l'utilisation d'étiqueteurs automatiques élaborés pour et à partir de données écrites n'est pas une solution optimale étant données les particularités des corpus oraux par rapport à l'écrit. Même si l'étiquetage de corpus oraux ne représente pas un problème spécifique (Benzitoun, 2004), l'utilisation de modèles entraînés sur des données écrites donne des résultats médiocres. Ainsi, nous avons testé TreeTagger (Schmid, 1997), avec son modèle standard pour le français, sur un échantillon de 3 007 tokens extraits du corpus de référence décrit dans cet article et nous avons obtenu une précision de seulement 83,1 %.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 99–112,
Grenoble, 4 au 8 juin 2012.

Un corpus du français parlé annoté en morphosyntaxe librement disponible serait donc utile, non seulement pour les logiciels d'annotation en morphosyntaxe, mais également pour améliorer les systèmes de transcription automatique (Huet *et al.*, 2006) ou d'autres outils. Cependant, il n'existe pas encore, à notre connaissance, de corpus de français parlé spontané annoté en morphosyntaxe (parties du discours et/ou lemmes) qui soit diffusé librement. Parmi les corpus annotés mais non diffusés librement, on peut citer les projets elicop (Mertens, 2002), C-ORAL-ROM (Campione *et al.*, 2005), Valibel (Dister, 2007), Corpus de Français Parlé Parisien (Branca-Rosoff *et al.*, 2010) ou bien encore ESLO (Eshkol *et al.*, 2010).

Notre objectif est donc de développer et diffuser librement à l'ensemble de la communauté scientifique un corpus pré-annoté automatiquement puis corrigé manuellement, dont la qualité aura été précisément évaluée. Il pourra servir notamment de corpus d'apprentissage spécifique au français parlé et plus largement de corpus exploitable pour des recherches en linguistique ou en Traitement Automatique des Langues (TAL).

Nous présentons tout d'abord le corpus de français parlé TCOF (Traitement des Corpus Oraux du Français), puis la méthodologie utilisée pour l'annotation manuelle, les différentes évaluations réalisées pendant la campagne et enfin les résultats obtenus par les étiqueteurs morphosyntaxiques entrainés sur une partie du corpus annoté TCOF-POS.

2 Présentation du corpus

Le corpus d'origine que nous avons annoté est celui du projet TCOF (André et Canut, 2010), librement disponible sur le site du CNRTL [1]. Ce corpus est constitué de transcriptions de données orales recueillies dans des contextes aussi naturels que possible. Il comporte une partie d'interactions entre adultes et une autre entre adultes et enfants. En ce qui concerne la partie adulte (la seule que nous ayons exploitée jusqu'à présent), elle est composée :

– d'interactions sollicitées, dans lesquelles au moins deux locuteurs sont engagés dans des récits de vie, d'événements ou d'expériences, ou dans des explications sur un savoir-faire professionnel ou technique ;
– de conversations à bâtons rompus ou portant sur des thématiques spécifiques ;
– de données non sollicitées dans des situations publiques ou professionnelles : réunions publiques, activités professionnelles diverses.

De ce corpus, nous avons extrait un échantillon de 22 240 tokens [2], soit 11 transcriptions différentes. Cet échantillon contient des conversations, des réunions professionnelles, ainsi que des extraits d'une Assemblée Générale à l'Université. L'intégralité des paroles prononcées a été scrupuleusement retranscrite en orthographe standard, sans artifice ou aménagement orthographique (donc sans ponctuation), suivant en cela les recommandations de (Blanche-Benveniste et Jeanjean, 1987) [3] largement diffusées et utilisées. Elles sont au format généré par le logiciel Transcriber (trs - XML).

1. http ://cnrtl.fr/corpus/tcof/
2. Notre conception de la notion de token est assez élémentaire (aucune insertion possible).
3. Les conventions de transcription sont disponibles sur le site suivant : http ://cnrtl.fr/corpus/tcof/TCOFConventions.pdf

Ces transcriptions sont automatiquement converties en texte brut à l'aide d'une feuille de style XSLT (qui élimine l'intégralité des balises XML), puis d'une série d'expressions régulières qui supprime les informations non désirées, telles que les pauses. Le texte final contient les mentions des locuteurs (L1, L2, etc.), l'intégralité des paroles prononcées, ainsi que les multi-transcriptions. Il s'agit donc de transcriptions brutes non retouchées, dont voici un exemple :

> *L1 et puis je crois que c'est en je crois je crois même que c'est en zone industrielle*
> *L2 ouais ouais je pense aussi ça doit pas être en ville*
> *L1 oui mais*
> *L2 en Belgique aussi il y a des trucs euh un genre de grand tr- enfin un genre de grande galerie en Belgique et puis c'est que des magasins de fringues aussi*

Les transcriptions ont été faites dans le cadre d'un cours de deuxième année de Sciences du Langage à l'Université Nancy 2, puis revues par des enseignants de Sciences du Langage. L'anonymisation, quant à elle, a été réalisée manuellement par des étudiants-vacataires. A la lecture, ils devaient repérer les toponymes, anthroponymes, etc. puis les remplacer par un symbole et insérer un son dans la portion de signal sonore correspondante.

3 Méthodologie

L'annotation totalement manuelle de corpus étant très coûteuse, nous avons procédé, comme décrit dans (Marcus *et al.*, 1993), à une correction manuelle de corpus pré-annotés automatiquement. La nature du pré-annotateur, ainsi que les modalités de la correction manuelle diffèrent selon les étapes du processus, comme nous allons le voir dans cette section. Toutefois, toutes les pré-annotations ont été produites par différentes instances du système TreeTagger (Schmid, 1997), qui fournit pour chaque token d'entrée une étiquette morphosyntaxique et un lemme.

Comme indiqué en introduction, l'utilisation comme pré-annotateur pour un corpus de parole spontanée transcrite, d'un étiqueteur morphosyntaxique entraîné sur un corpus écrit n'est pas adaptée. Parmi les phénomènes qui posent problème, lesquels ne sont pas totalement absents des corpus écrits mais y sont bien plus rares (Benzitoun, 2004), on peut citer :
– les répétitions de mots ou de groupes de mots (*ça ça redevient ça redevient le bordel comme ça*),
– les reformulations (*peut-être séparer complètement euh junior euh homme enfin euh adulte*),
– les ruptures de construction (*ouais ouais que de la gueule que de la*),
– les amorces de mots (*moi j'aurais p- j'aurais pas mis de pantalon*),
– les incises (*euh on considérait que former les hommes et c'est toujours euh en en en vigueur ça hein former les les en- les les enfants d'aujourd'hui c'est aussi former les hommes de demain*),
– les formes non conventionnelles (*tu sais genre trop vénère ; avoir du matos en entrée de mag*),
– les particules discursives (*hein, eh ben*, etc.)...
Pour ne prendre que deux exemples, la version standard de TreeTagger pour le français considère que *bon* est systématiquement un adjectif et *quoi* un pronom, alors qu'ils sont majoritairement des particules discursives. De plus, nos transcriptions ne sont pas segmentées en « phrases » (Blanche-Benveniste et Jeanjean, 1987), ce qui peut également poser des problèmes aux outils. Par exemple, l'étiquette SENT (pour *sentence*) indiquant une frontière de phrase doit obligatoirement être présente dans le lexique servant pour l'apprentissage de TreeTagger (même s'il ne s'en sert pas par la suite). En conséquence, nous avons procédé, dès que possible, à l'entraînement de versions de TreeTagger à partir des annotations déjà obtenues sur notre corpus.

La méthodologie retenue, décrite en détail dans cette partie, peut être résumée comme suit :

1. Définition de critères de tokenisation et d'identification des composés, puis tokenisation automatique ;
2. Définition d'un jeu d'étiquettes adapté à la parole spontanée transcrite ;
3. Création d'un corpus de référence C_{ref} de 22 240 tokens par correction d'une pré-annotation automatique, effectuée par deux experts linguistes :
   – Les 10 000 premiers tokens de C_{ref} ont été pré-annotés avec la version standard de TreeTagger ;
   – Les 12 240 tokens suivants de C_{ref} ont été pré-annotés avec une version de TreeTagger entraînée sur les 10 000 premiers ;
4. Ré-annotation par deux étudiantes d'environ 7 500 tokens du corpus de référence C_{ref} (suivie d'une phase d'adjudication), afin d'évaluer la qualité des annotations dans deux configurations distinctes :
   – environ 6 000 tokens ont été pré-annotés par la version standard de TreeTagger ;
   – environ 1 500 tokens ont été pré-annotés avec une version de TreeTagger entraînée sur les 16 312 premiers tokens de C_{ref} ;

   L'objectif était ici de mesurer l'impact de la différence de qualité entre pré-annotateurs en termes de vitesse d'annotation et de précision du résultat de l'étape manuelle ;
5. Application de cette méthodologie à un plus grand nombre d'étudiants pour en valider la robustesse ;
6. Annotation par deux étudiantes d'un corpus additionnel C_{add} de 80 000 nouveaux tokens, pré-annotés avec la version de TreeTagger entraînée sur la totalité de C_{ref}.

Nous avons appliqué pour cette campagne les bonnes pratiques actuelles en annotation manuelle de corpus, qui consistent à évaluer le plus tôt possible l'accord inter-annotateurs et de mettre à jour le guide d'annotation (Bonneau-Maynard *et al.*, 2005). La répétition régulière de ce processus conduit à ce qu'on appelle maintenant l'annotation agile (Voormann et Gut, 2008).

3.1 Tokenisation et gestion des composés

Le corpus ayant fait l'objet d'une pré-annotation (voir section 3.3), nous avons pris comme base la tokenisation par défaut de TreeTagger, qui repose notamment sur un fichier de composés. Mais ce dernier s'est avéré insuffisant (par exemple, *parce que* reste découpé en deux tokens distincts mais *puisqu'ils* en un seul token). Nous l'avons donc complété au fur et à mesure, en respectant le critère suivant : toute séquence dans laquelle il est possible d'insérer un élément est découpée en plusieurs tokens, afin d'exclure les unités discontinues. Ainsi, *un peu* est découpé en deux tokens (car on peut trouver *un tout petit peu*).

3.2 Un jeu d'étiquettes adapté à la parole spontanée transcrite

Afin de bénéficier au mieux des ressources développées pour l'écrit et de limiter le travail de correction, tout en prenant en considération les phénomènes spécifiques à la parole spontanée cités ci-dessus, nous avons décidé d'utiliser un jeu d'étiquettes basé au départ sur les étiquettes par défaut fournies par TreeTagger. Nous l'avons complété à l'aide de (Abeillé et Clément, 2006).

Les répétitions, reformulations, etc. n'ont pas fait l'objet de traitements spécifiques, chacun des tokens a la catégorie qu'il a habituellement (ex : le[DET] le[DET] le[DET] chat). Au final, même si les identifiants des étiquettes sont différents, les catégories retenues sont quasiment identiques à (Abeillé et Clément, 2006), avec toutefois un peu moins de sous-catégories (notamment aucune pour les adverbes et les adjectifs) et l'ajout de la catégorie « auxiliaire » ainsi que de trois étiquettes spécifiques à l'oral : *MLT* (multi-transcription), *TRC* (amorce de mot) et *LOC* (locuteur) (cf. tableau 1). Afin de nous aider dans la rédaction du manuel d'annotation, nous nous sommes d'ailleurs inspirés de (Abeillé et Clément, 2006). Notre jeu d'étiquettes comprend 62 étiquettes.

En outre, il a été affiné durant la phase de constitution du corpus servant de référence. En effet, nous voulions que les étiquettes soient apposées de manière aussi systématique que possible pour que nos choix soient réversibles et que les modifications soient automatisables, autant que faire se peut. De ce fait, même si cela peut paraître discutable d'un point de vue théorique, nous avons privilégié les choix qui potentiellement génèrent le moins de fluctuations entre annotateurs. Par exemple, la distinction entre participe passé et adjectif n'est pas aisée et plutôt que d'obtenir une annotation de qualité moindre, nous avons préféré neutraliser celle-ci. Ainsi, chaque fois que la forme verbale existe (sauf cas de changement notoire de sens), nous avons annoté « verbe ». Dans le cas contraire, nous avons annoté « adjectif ».

Nous avons également décidé d'essayer de limiter les cas de transferts d'une catégorie vers une autre (trans-catégorisation). En effet, ceux-ci auraient artificiellement été limités aux cas rencontrés dans le corpus à annoter, sans possibilité d'avoir une vision globale du phénomène. De plus, cela aurait complexifié la tâche de correction. Ainsi, dans *rouler tranquille*, *tranquille* est considéré comme un adjectif et non comme un adverbe (ce qui, de toute façon, est discutable d'un point de vue théorique). Enfin, il n'a pas été possible d'exclure totalement les cas d'étiquettes limitées à un mot unique. Ainsi, l'étiquette « particule interrogative » ne s'utilise que pour *est-ce qu-e/i* et « prédéterminant » uniquement pour *tous*.

3.3 Création du sous-corpus de référence

Comme indiqué ci-dessus, la création de la première tranche de 10 000 tokens du corpus de référence C_{ref} de 22 240 tokens a été réalisée en utilisant comme pré-annotateur la version standard de TreeTagger, entraînée sur un corpus écrit. Nous (L. Bérard et C. Benzitoun) avons ensuite corrigé ces pré-annotations en plusieurs passes. Nous avons tout d'abord effectué des remplacements automatiques, lorsque les modifications étaient systématiques ou que l'étiquette majoritaire n'était pas celle apposée par défaut par le logiciel (ce qui est le cas pour *bon* (ADJ/INT) et *quoi* (PRO :int/INT), par exemple). Ensuite, nous nous sommes répartis les données à corriger et, après les avoir intégralement traitées, nous nous les sommes échangées pour révision. Nous avons ensuite discuté des cas où nous n'étions pas en accord jusqu'à trouver des solutions. Nous avons effectué ces étapes plusieurs fois, jusqu'à obtenir des annotations fiables. Le guide d'annotation était mis à jour à chaque étape.

Nous avons par ailleurs généré automatiquement des fichiers de fréquences, afin de faciliter le repérage des erreurs. A ainsi été calculée la fréquence de chaque étiquette pour un même lemme ou un même token, ce qui nous a permis d'identifier et de corriger quelques erreurs supplémentaires. Par exemple, *C.E.* ayant été annoté 1 fois *NAM* et 3 fois *NOM :sg* (pour un même lemme *C.E.*), cette dernière étiquette a été attribuée aux 4 occurrences. De même, cela nous a permis de corriger deux occurrences de *du*, indûment annotées *DET :ind*.

ADJ	adjectif	NUM	numéral
ADV	adverbe	PRO :clo	clitique objet
AUX :cond	auxiliaire au conditionnel	PRO :cls	clitique sujet
AUX :futu	auxiliaire au futur	PRO :clsi	clitique sujet impersonnel
AUX :impe	auxiliaire à l'impératif	PRO :dem	pronom démonstratif
AUX :impf	auxiliaire à l'imparfait	PRO :ind	pronom indéfini
AUX :infi	auxiliaire à l'infinitif	PRO :int	pronom interrogatif
AUX :pper	auxiliaire au participe passé	PRO :pos	pronom possessif
AUX :ppre	auxiliaire au participe présent	PRO :rel	pronom relatif
AUX :pres	auxiliaire au présent	PRO :ton	pronom tonique
AUX :simp	auxiliaire au passé simple	PRP	préposition
AUX :subi	auxiliaire au subjonctif imparfait	PRP :det	préposition/déterminant
AUX :subp	auxiliaire au subjonctif présent	PRT :int	particule interrogative (est-ce que)
DET :def	déterminant défini	SYM	symbole
DET :dem	déterminant démonstratif	TRC	amorces de mots
DET :ind	déterminant indéfini	VER	verbe sans flexion (voilà)
DET :int	déterminant interrogatif	VER :cond	verbe au conditionnel
DET :par	déterminant partitif (du)	VER :futu	verbe au futur
DET :pos	déterminant possessif	VER :impe	verbe à l'impératif
DET :pre	pré-déterminant (tout (le))	VER :impf	verbe à l'imparfait
EPE	épenthétique	VER :infi	verbe à l'infinitif
ETR	mots étrangers	VER :pper	verbe au participe passé
FNO	forme noyau (oui, non, d'accord, etc.)	VER :ppre	verbe au participe présent
INT	interjection et particules discursives	VER :pres	verbe au présent
KON	conjonction	VER :simp	verbe au passé simple
LOC	locuteur	VER :subi	verbe au subjonctif imparfait
MLT	multi-transcription	VER :subp	verbe au subjonctif présent
NAM	nom propre	NOM :trc	nom commun tronqué
NOM	nom commun	NAM :trc	nom propre tronqué
NOM :sig	sigle	VER :trc	verbe tronqué
NAM :sig	sigle	ADJ :trc	adjectif tronqué

TABLE 1 – Jeu d'étiquettes du corpus TCOF-POS

Nous avons ensuite appliqué les résultats obtenus par Fort et Sagot (2010) sur l'intérêt d'une pré-annotation avec un outil de qualité moyenne. Ainsi, une fois la première tranche de 10 000 tokens annotés, nous avons ré-entraîné TreeTagger sur ce sous-corpus (mais sans utiliser de lexique externe) et avons pré-annoté les transcriptions suivantes de C_{ref} (12 240 tokens) avec ce nouvel outil. La même méthodologie que celle utilisée pour corriger les 10 000 premiers tokens nous a permis de finaliser le corpus de référence C_{ref} de 22 240 tokens.

3.4 Création du sous-corpus additionnel

Le corpus diffusé est composé pour une part du sous-corpus de référence C_{ref} et pour une autre part d'un autre sous-corpus additionnel C_{add} corrigé par deux étudiantes (de L3 et M2 de Sciences du Langage) recrutées spécifiquement pour cette tâche. Dans un premier temps, afin d'évaluer *a priori* la méthodologie prévue pour l'annotation de C_{add}, nous avons mené une campagne de tests en nous servant de C_{ref} comme référence. Pour ce faire, les deux étudiantes ont eu 15 fichiers extraits de C_{ref} d'environ 500 tokens chacun [4] à corriger dans un ordre contraint. Les 12 premiers fichiers avaient été pré-annotés par la version standard de TreeTagger. Afin de mesurer l'impact de la qualité du pré-annotateur, les 3 derniers fichiers avaient été pré-annotés par une version de TreeTagger ré-entraînée à partir d'un extrait de 16 312 tokens de la référence C_{ref} (et sans lexique externe). Naturellement, ces tokens forment un sous-ensemble de C_{ref} disjoint des 15 fichiers à ré-annoter. Les étudiantes avaient l'interdiction d'échanger des informations durant la phase d'annotation.

La correction a été effectuée dans un tableur, les cellules contenant les étiquettes étant munies d'une liste déroulante se limitant au jeu d'étiquettes défini ci-dessus. La saisie était donc contrainte. Une fois la correction terminée, les fichiers annotés en parallèle ont été comparés automatiquement. Les cas de divergence entre les deux annotateurs ont ainsi été repérés automatiquement et corrigés par un expert [5].

Dans le cadre de ce travail, les mesures suivantes ont été effectuées :
– le temps mis par les étudiantes pour annoter chaque fichier ;
– la précision de chaque fichier par rapport à la référence ;
– l'accord inter-annotateurs des étudiantes (Kappa de Cohen (Cohen, 1960)) ;
– la précision après fusion et adjudication.
L'évaluation de leurs annotations sur ces 15 fichiers est reproduite ci-dessous (figures 1 et 2 et tableau 2). Elle tient compte de la lemmatisation et des parties du discours.

1e	2e	3e	4e	5e	6e	7e	8e	9e	10e	11e	12e	13e	14e	15e
107	71	80	67	60	60	57	65	50	50	52	47	32	32	31

TABLE 2 – Temps d'annotation (en minutes)

Entre le 12^e et le 13^e fichier, la différence de temps est vraisemblablement imputable au changement de pré-annotation par TreeTagger.

4. Cette taille a été retenue car nous avons observé qu'elle permet une correction rapide et une attention soutenue sans être obligé de s'interrompre en cours d'annotation.

5. Pour des raisons pratiques, il n'a pas été possible de confier cette phase à un expert externe. La personne qui l'a réalisée a également collaboré à la réalisation du corpus servant de référence, ce qui peut représenter un biais méthodologique.

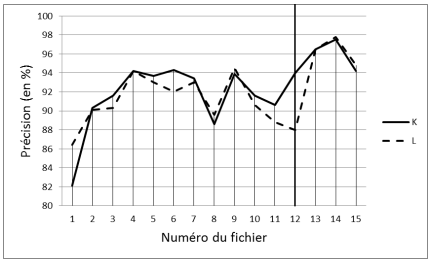


FIGURE 1 – Évolution de la précision des deux étudiantes

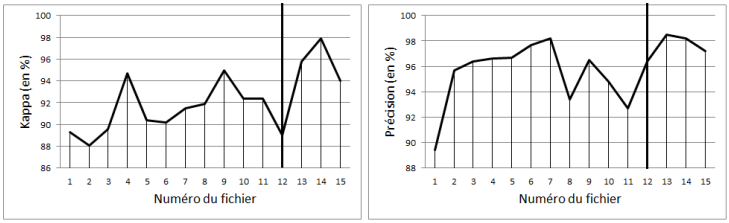


FIGURE 2 – Évolution du Kappa (à gauche) des 2 étudiantes et de la précision après adjudication (à droite)

La qualité des corrections après ré-entraînement et pré-étiquetage ainsi que le faible temps de correction (pour les 3 derniers fichiers, donc) nous ont paru suffisants pour valider notre méthodologie et ainsi poursuivre l'élaboration du corpus. Sur les 3 derniers fichiers, la précision moyenne est de 98,03 % en ne tenant compte que des étiquettes. Nous sommes donc passés à l'annotation par les deux étudiantes du corpus C_{add}. Elles ont ainsi reçu le même jeu de 160 nouveaux fichiers de 500 tokens chacun, pré-annotés par la version ré-entraînée de TreeTagger. En 60 heures, elles ont corrigé 80 000 tokens chacune, ce qui fait une moyenne d'un peu plus de 21 minutes par fichier de 500 tokens. Sur l'ensemble, l'accord inter-annotateurs (Kappa de Cohen (Cohen, 1960)) est en moyenne de 96,5 % et le temps moyen consacré à l'adjudication de 2 min. 45 sec par fichier.

4 Élargissement de l'évaluation

Afin d'évaluer le caractère robuste de notre méthodologie, nous avons élargi l'évaluation à plus d'étudiants. En effet, nous comptons augmenter de manière importante la quantité de fichiers corrigés dans les années à venir et nous voulons vérifier si notre méthodologie donne des résultats comparables quels que soient les correcteurs. Pour ce faire, nous avons adopté

la même méthodologie que celle décrite ci-dessus, à savoir une double-annotation de chaque fichier puis une adjudication pour les cas de divergence uniquement. Cette évaluation a porté sur les corrections fournies par 10 étudiants en Sciences du Langage à l'Université Nancy 2 (L3 et M2) dans le cadre de deux enseignements. A chaque binôme, nous avons donné 6 fichiers (4 fichiers pré-annotés avec le TreeTagger de base et 2 fichiers avec le TreeTagger ré-entraîné) à corriger dans un ordre contraint. Dans cette expérience, comme dans la précédente, les étudiants devaient corriger les lemmes en plus des étiquettes. Dans la suite de ce travail, les mesures que nous présentons tiennent compte à la fois des lemmes et étiquettes (sauf précision contraire).

4.1 Temps d'annotation et accord inter-annotateurs

En ce qui concerne le temps d'annotation, nous avons observé une diminution systématique avec une nette différence entre les 4 premiers fichiers et les deux derniers (voir tableau 3).

1e annot.	2e annot.	3e annot.	4e annot.	5e annot.	6e annot.
110,1	101,8	79,2	72,5	41,3	39,7

TABLE 3 – Temps d'annotation (en minutes)

Au-delà de la diminution du temps de correction inhérente à une meilleure maîtrise des étudiants, il paraît difficile d'expliquer la diminution du temps entre le quatrième et le cinquième fichier par un autre facteur que le basculement entre le TreeTagger standard et la version ré-entraînée. Le même phénomène peut être observé concernant l'accord inter-annotateurs (*cf.* figure 3). Le coefficient d'accord inter-annotateurs présenté ici est, comme précédemment, le κ de Cohen (Cohen, 1960).

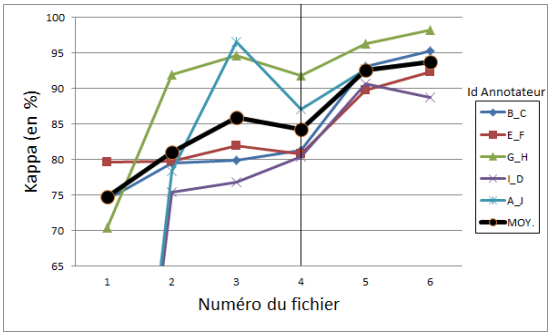


FIGURE 3 – Évolution de l'accord inter-annotateurs (kappa) des étudiants

Dans la figure 3, la courbe noire représente l'évolution de la moyenne des accords inter-annotateurs, de même que dans les graphiques suivants.

4.2 Précision

Outre une diminution significative du temps d'annotation et une augmentation de l'accord inter-annotateurs, nous avons également constaté une importante augmentation de la précision en moyenne pour chaque étudiant (*cf.* figure 4). On observe encore une fois une nette augmentation entre le quatrième et le cinquième fichier, et ce chez tous les étudiants. La figure 5 indique la précision de chaque fichier après fusion et adjudication.

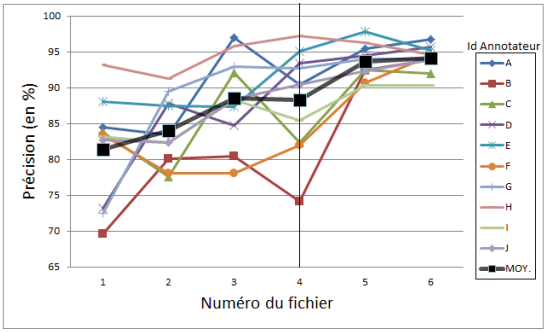


FIGURE 4 – Evolution de la précision de chaque étudiant

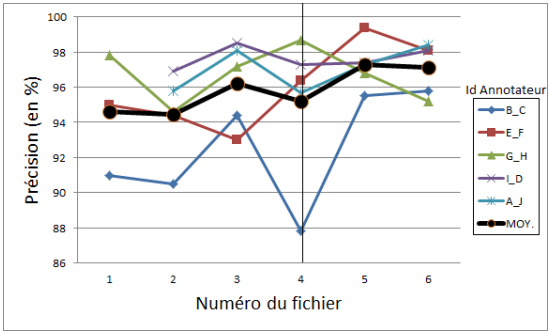


FIGURE 5 – Évolution de la précision de chaque fichier corrigé après fusion et adjudication

Finalement, sur les deux derniers fichiers annotés, le taux de précision moyen est respectivement de 97,28 % et 97,12 % en évaluant les erreurs portant à la fois sur les lemmes et les étiquettes. Si l'on prend en compte seulement les étiquettes, le taux de précision moyen est alors respectivement de 97,42 % et de 97,8 % pour ces mêmes fichiers, ce qui est légèrement inférieur à ce qui a

été relevé précédemment pour les deux étudiantes. Cependant, nous estimons que cela permet d'affirmer que les principes adoptés permettent d'obtenir des corpus annotés de qualité proche, et ce quelle que soit la personne qui corrige.

5 Premiers résultats de l'annotation automatique

Pour effectuer les premiers tests concernant l'apprentissage automatique, notre choix s'est porté sur deux étiqueteurs morphosyntaxiques : MElt (Denis et Sagot, 2009, 2012), étiqueteur état de l'art pour le français, et TreeTagger, utilisé comme pré-annotateur pour constituer le corpus, et largement utilisé bien qu'il ne soit pas celui qui donne les meilleurs résultats à l'heure actuelle (Denis et Sagot, 2009; Eshkol *et al.*, 2010). Tous deux sont librement disponibles et multi-plateformes.

Notre objectif était d'étudier la courbe d'apprentissage, et ce sous plusieurs angles : la précision de l'étiqueteur entraîné augmente-t-elle avec la taille du corpus d'entraînement ? L'utilisation d'un lexique externe augmente-t-elle de façon significative la précision de l'étiqueteur ? Avec quelle taille de corpus d'entraînement obtient-on le meilleur étiqueteur ? Quelle est sa précision ? À partir de quelle taille de corpus d'entraînement l'étiqueteur obtenu peut-il être utilisé comme pré-annotateur dans une campagne d'annotation manuelle qui consiste en la correction manuelle de l'annotation automatique ? Les deux systèmes d'étiquetage, MElt et TreeTagger, conduisent-ils à des résultats similaires concernant les questions précédentes ?

Nous avons donc procédé à l'entraînement de MElt et de TreeTagger sur 10 sous-corpus successifs du corpus de référence C_{ref}, dont la taille croît de 2 000 à 20 000 tokens. Nous avions préalablement mis de côté trois tranches de 500 tokens afin de servir d'échantillons de test. Pour rendre nos résultats comparables avec ceux présentés dans d'autres campagnes, l'évaluation de la précision s'est limitée aux seules étiquettes.

Pour l'entraînement de TreeTagger, nous avons utilisé comme lexique externe le lexique Morphalou 2.0 (Romary *et al.*, 2004). Nous avons dû convertir Morphalou au format attendu par TreeTagger puis le fusionner, pour chacun des 10 sous-corpus d'apprentissage, avec le lexique qui en est extrait. Nous avons également effectué des tests sans Morphalou, uniquement avec un lexique endogène. Pour l'entraînement de MElt, nous avons utilisé la version du lexique L*efff* (Sagot, 2010) utilisée pour l'entraînement de la version standard de l'étiqueteur MElt pour le français. Le lexique externe étant utilisé par MElt comme une source de traits pour le modèle d'étiquetage, nous avons pu conserver le jeu de catégories du lexique externe bien qu'il soit différent des catégories du corpus d'entraînement. Pour MElt, le lexique externe reste distinct du lexique extrait du corpus d'entraînement.

Premier constat : à l'exception du premier étiqueteur entraîné sur 2 000 tokens, les étiqueteurs obtenus avec MElt sont systématiquement meilleurs que ceux obtenus avec TreeTagger. Les meilleurs scores, obtenus à partir du corpus de 20 000 tokens, sont respectivement 96,9 % avec MElt et 94,9 % avec TreeTagger. Comparés aux 85–90 % annoncés par (Eshkol *et al.*, 2010) et aux 80 % obtenus par A. Dister (c.p. du 24 janvier 2008 [6]), nos résultats constituent donc une amélioration significative. Mais cela masque des variations d'un échantillon à un autre, ainsi qu'au niveau de la moyenne (voir figure 6), mais surtout des différences entre jeux d'étiquettes.

6. Diaporama disponible à l'adresse : http ://rhapsodie.risc.cnrs.fr/docs/Dister_Syntaxe_240108.pdf

Deuxième constat, sans surprise : l'utilisation du lexique externe améliore la précision de l'étiqueteur. Par exemple, sur le corpus de 2 000 tokens, TreeTagger n'atteint que 78,4 % de précision sans lexique externe contre 90,9 % avec Morphalou. Pour ce même corpus, la différence est moins importante avec MElt, mais elle est significative : la précision passe de 84,9 % à 88,1 %. Avec 20 000 tokens, l'utilisation du lexique externe permet à la précision de l'étiqueteur entraîné par MElt de passer de 95,5 % à 96,9 %. On note que MElt, sans lexique externe, donne des résultats supérieurs à TreeTagger avec lexique externe dès que le corpus d'entraînement fait plus de 12 000 tokens.

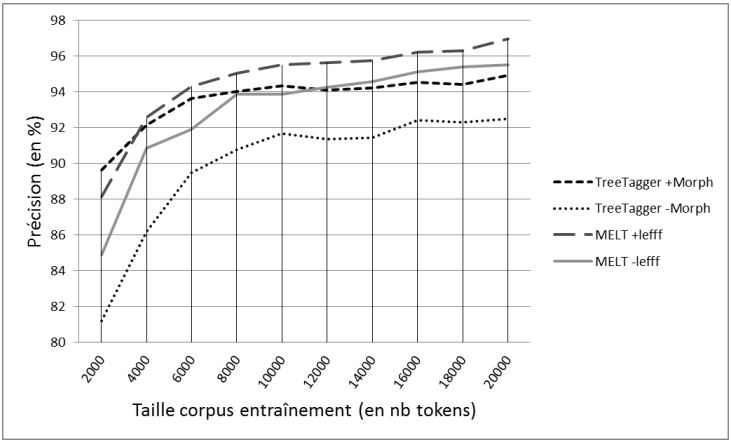


FIGURE 6 – Évolution de la précision de l'annotation automatique par tranche de 2 000 tokens

La figure 6 montre qu'à partir de 6 000 tokens, les résultats commencent à progresser de manière moins marquée, que ce soit avec MElt ou TreeTagger. Les précisions obtenues avec cette taille de corpus (94,3% avec MElt) sont suffisantes pour lancer une campagne de correction telle que nous la décrivons ci-dessus, après ré-entraînement. Il n'est pas indispensable que le corpus d'apprentissage soit plus volumineux. En tout cas, au vu de nos résultats, il n'est pas utile d'aller au-delà de 10 000 tokens si l'on utilise TreeTagger. L'utilisation de MElt semble toutefois préférable, avec des résultats qui continuent à croître jusqu'à 20 000 tokens.

6 Conclusion et perspectives

Le corpus TCOF-POS ($C_{\text{ref}} + C_{\text{add}}$) est disponible sur le site du CNRTL[7] sous licence Creative Commons BY-NC-SA 2.0[8], héritée du corpus TCOF. Il contient un peu plus de 100 000 tokens, dont un peu plus de 20 000 tokens de référence et 80 000 tokens obtenus grâce à la double-annotation (par les deux étudiantes recrutées) puis adjudication par un expert linguiste (C.

7. http ://cnrtl.fr/corpus/perceo/
8. http ://creativecommons.org/licenses/by-nc-sa/2.0/fr/

Benzitoun). Les meilleurs modèles d'étiquetage pour TreeTagger et pour MElt, qui ont une précision respectivement de 94,9 % et 96,9 % à ce stade de développement du corpus, seront également mis à disposition sous peu sur ce site (pour l'instant, seul le fichier paramètre pour TreeTagger est téléchargeable). Le lexique fusionné avec Morphalou est également disponible à cette même adresse. La version ré-entraînée de TreeTagger a déjà été utilisée par les concepteurs du Corpus de Français Parlé Parisien [9] pour annoter leurs données.

Dans le cadre d'une campagne de correction d'une pré-annotation automatique, nous avons mis en évidence le seuil de 6 000 tokens comme base de départ minimale pour ré-entraîner le logiciel. A ce stade, on obtient de bons résultats (94,3 % pour MElt et 93,6 % pour TreeTagger) et la précision progresse de manière moins marquée. Mais cette recommandation est valable lorsque le logiciel d'étiquetage est couplé à un lexique externe. Or, dans le cadre de notre campagne d'évaluation des corrections manuelles, nous n'avons pas utilisé de lexique externe pour ré-entraîner TreeTagger. Il faudra donc tester si les résultats sont de meilleure qualité lorsque l'on ajoute ce paramètre, travail que nous effectuons à l'heure actuelle.

Remerciements

Nous tenons à remercier les 12 étudiants ayant collaboré à ce projet et plus particulièrement M. Salcedo et M. Paquot, recrutées spécifiquement pour faire l'annotation. De même, L. Bérard, E. Jacquey, V. Meslard, S. Ollinger et E. Petitjean ont apporté une contribution significative à ce projet. Nous souhaitons également remercier l'ATILF pour son soutien financier dans le cadre d'un projet interne. La participation de K. Fort a été financée dans le cadre du programme Quæro [10], financé par OSEO, agence nationale de valorisation de la recherche. Celle de B. Sagot entre dans le cadre du projet ANR EDyLex (ANR-09-CORD-008).

Références

Abeillé, A. et Clément, L. (2006). *Annotation morpho-syntaxique. Les mots simples - Les mots composés Corpus Le Monde*.

André, V. et Canut, E. (2010). Mise à disposition de corpus oraux interactifs : le projet TCOF (traitement de corpus oraux en francais). *Pratiques*, 147/148:35–51.

Benzitoun, C. (2004). L'annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique ? *In Actes de la conférence RECITAL*, pages 13–22, Fès, Maroc.

Blanc, O., Constant, M., Dister, A. et Watrin, P. (2008). Corpus oraux et chunking. *In Journées d'étude sur la parole (JEP)*, Avignon, France.

Blanche-Benveniste, C. et Jeanjean, C. (1987). *Le Francais parlé. Transcription et édition*. Didier Érudition, Paris, France.

Bonneau-Maynard, H., Rosset, S., Ayache, C., Kuhn, A. et Mostefa, D. (2005). Semantic Annotation of the French Media Dialog Corpus. *In InterSpeech*, Lisbonne, Portugal.

9. http ://cfpp2000.univ-paris3.fr/search-transcription-tt/
10. `http://www.quaero.org`

BRANCA-ROSOFF, S., FLEURY, S., LEFEUVRE, F. et PIRES, M. (2010). Discours sur la ville. corpus de francais parlé parisien des années 2000 (CFPP2000). Rapport technique.

CAMPIONE, E., VÉRONIS, J. et DEULOFEU, J. (2005). *C-ORAL-ROM, Integrated Reference Corpora for Spoken Romance Languages*, édité par E. Cresti et M. Moneglia, chapitre 3. The French corpus, pages 111–133. John Benjamins, Amsterdam, Hollande.

COHEN, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

DENIS, P. et SAGOT, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. *In Proceedings of PACLIC 2009*, Hong Kong, Chine.

DENIS, P. et SAGOT, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*. À paraître.

DISTER, A. (2007). *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelle orale VALIBEL*. Thèse de doctorat, Université de Louvain, Belgique.

ESHKOL, I., TELLIER, I., TAALAB, S. et BILLOT, S. (2010). étiqueter un corpus oral par apprentissage automatique à l'aide de connaissances linguistiques. *In 10th International Conference on statistical analysis of textual data (JADT 2010)*, Rome, Italie.

FORT, K. et SAGOT, B. (2010). Influence of Pre-annotation on POS-tagged Corpus Development. *In Proc. of the Fourth ACL Linguistic Annotation Workshop*, Uppsala, Suède.

HUET, S., GRAVIER, G. et SÉBILLOT, P. (2006). Peut-on utiliser les étiqueteurs morphosyntaxiques pour améliorer la transcription automatique. *In Actes des 26èmes Journées d'Études sur la Parole (JEP)*, Dinard, France.

MARCUS, M., SANTORINI, B. et MARCINKIEWICZ, M. A. (1993). Building a large annotated corpus of english : The penn treebank. *Computational Linguistics*, 19(2):313–330.

MERTENS, P. (2002). Les corpus de francais parlé ELICOP : consultation et exploitation. *In* BINON, J., DESMET, P., ELEN, J., MERTENS, P. et SERCU, L., éditeurs : *Tableaux Vivants. Opstellen over taal-en-onderwijs aangeboden aan Mark Debrock*, pages 101–116. Universitaire Pers, Leuven, Belgique.

ROMARY, L., SALMON-ALT, S. et FRANCOPOULO, G. (2004). Standards going concrete : from LMF to Morphalou. *In Workshop on Electronic DictionariesWorkshop on Electronic Dictionaries, Coling 2004*, Genève, Suisse.

SAGOT, B. (2010). The L*efff*, a freely available and large-coverage morphological and syntactic lexicon for french. *In 7th international conference on Language Resources and Evaluation (LREC 2010)*, La Vallette, Malte.

SCHMID, H. (1997). *New Methods in Language Processing, Studies in Computational Linguistics*, édité par D. Jones et H. Somers, chapitre Probabilistic part-of-speech tagging using decision trees, pages 154–164. UCL Press, Londres.

VALLI, A. et VÉRONIS, J. (1999). étiquetage grammatical de corpus oraux : problèmes et perspectives. *Revue Francaise de Linguistique Appliquée*, IV(2):113–133.

VOORMANN, H. et GUT, U. (2008). Agile corpus creation. *Corpus Linguistics and Linguistic Theory*, 4(2):235–251.

Alignement sous-phrastique hiérarchique avec Anymalign

Adrien Lardilleux[1] François Yvon[1,2] Yves Lepage[3]
(1) LIMSI-CNRS
(2) Université Paris-Sud
(3) Université Waseda, Japon
adrien.lardilleux@limsi.fr, francois.yvon@limsi.fr

RÉSUMÉ

Nous présentons un algorithme d'alignement sous-phrastique permettant d'aligner très facilement un couple de phrases à partir d'une matrice d'alignement pré-remplie. Cet algorithme s'inspire de travaux antérieurs sur l'alignement par segmentation binaire récursive ainsi que de travaux sur le clustering de documents. Nous évaluons les alignements produits sur des tâches de traduction automatique et montrons qu'il est possible d'atteindre des résultats du niveau de l'état de l'art, affichant des gains très conséquents allant jusqu'à plus de 4 points BLEU par rapport à nos travaux antérieurs, à l'aide une méthode très simple, indépendante de la taille du corpus à traiter, et produisant directement des alignements symétriques. En utilisant cette méthode en tant qu'extension à l'outil d'extraction de traductions Anymalign, nos expériences nous permettent de cerner certaines limitations de ce dernier et de définir des pistes pour son amélioration.

ABSTRACT

Hierarchical sub-sentential alignment with Anymalign

We present a sub-sentential alignment algorithm that aligns sentence pairs from an existing alignment matrix in a very easy way. This algorithm is inspired by previous work on alignment by recursive binary segmentation and on document clustering. We evaluate the alignments produced on machine translation tasks and show that we can obtain state-of-the-art results, with gains up to more than 4 BLEU points compared to our previous work, with a method that is very simple, independent of the size of the corpus to be aligned, and can directly produce symmetric alignments. When using this method as an extension of the translation extraction tool Anymalign, our experiments allow us to determine some of its limitations and to define possible leads for further improvements.

MOTS-CLÉS : corpus parallèle ; alignement sous-phrastique ; traduction automatique statistique.

KEYWORDS: parallel corpus ; sub-sentential alignment ; statistical machine translation.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 113–126,
Grenoble, 4 au 8 juin 2012.

1 Introduction

L'alignement sous-phrastique consiste à identifier des traductions d'unités textuelles à partir d'un corpus parallèle aligné en phrases, c'est-à-dire dont les phrases ont été préalablement mises en correspondance avec leur traduction. Cette tâche constitue la première étape du processus d'entraînement de la plupart des systèmes de traduction automatique fondée sur les données (traduction statistique ou par l'exemple). L'approche la plus répandue est actuellement la traduction automatique statistique par segments (n-grammes de mots), où le modèle central prend la forme d'une table de traductions, obtenue à partir de correspondances sous-phrastiques. Cette table consiste en une liste pré-calculée de couples de segments (*source*, *cible*), à chacun desquels est associé un certain nombre de scores reflétant la probabilité que *source* se traduise par *cible*.

Le problème de l'identification d'associations sous-phrastiques à partir de textes parallèles, entre mots isolés ou n-grammes de mots par exemple, est bien connu, et de nombreuses propositions ont été faites pour le résoudre. On peut grossièrement classer ces méthodes en deux catégories. La première, l'approche *probabiliste*, introduite par Brown *et al.* (1988), considère le problème d'identifier des *liens* entre mots ou groupes de mots dans des phrases parallèles. Cette approche consiste à définir un modèle probabiliste du texte parallèle, dont les paramètres sont estimés par un processus de maximisation global qui considère toutes les associations possibles du corpus en même temps. Le but est de déterminer le meilleur ensemble de liens d'alignement entre les mots source et cible de chaque couple de phrases parallèles. Les plus connus dans cette catégorie sont les modèles IBM (Brown *et al.*, 1993), permettant d'aligner des mots isolés, et qui ont donné lieu à une impressionnante liste de variantes et d'améliorations (voir par exemple les travaux de Vogel *et al.* (1996); Wu (1997); Deng et Byrne (2005); Liang *et al.* (2006); Fraser et Marcu (2007); Ganchev *et al.* (2008), pour ne citer qu'eux). La généralisation des modèles d'alignement de mots à l'alignement de segments s'avère être un problème bien plus difficile, et au vu des déficiences des propositions de Marcu et Wong (2002) et Vogel (2005), de tels alignements sont généralement produits en combinant des alignements de mots 1–n asymétriques (« orientés ») dans les deux directions à l'aide d'heuristiques (Koehn *et al.*, 2003; DeNero et Klein, 2007). Une fois l'ensemble de ces liens d'alignement constitué, il est possible d'attribuer des scores à chacun des couples de segments extraits.

La seconde approche, *associative* (qualifiée d'*heuristique* par Och et Ney (2003)), a été introduite par Gale et Church (1991). Celle-ci ne nécessite pas de modèle d'alignement : pour détecter des traductions, elle repose sur des mesures d'indépendance statistique telles que, par exemple, le coefficient de Dice, l'information mutuelle (Gale et Church, 1991; Fung et Church, 1994), ou le rapport de vraisemblance (Dunning, 1993) — voir aussi les travaux plus récents de Melamed (2000) et Moore (2005). On limite généralement les tests à une liste d'associations candidates pré-calculée à partir de motifs et de filtres, en se concentrant par exemple uniquement sur les n-grammes de mots les plus fréquents. Dans cette approche, on utilise un processus de maximisation locale, où chaque segment est traité indépendamment des autres. Cette approche permet généralement d'extraire directement des couples de traductions. Dans ce courant, on trouve par exemple les travaux de Gale et Church (1991), qui ont été depuis étendus aux corpus non strictement parallèles (Fung et Church, 1994; Fung et Yee, 1998), de Dagan et Church (1994); Gaussier et Langé (1995); Smadja *et al.* (1996) pour apprendre des associations de segments ou de termes, ou encore des travaux ayant recours à diverses mesures d'association, telles que le G^2 (Gale et Church, 1991) ou le ϕ^2 (Dunning, 1993; Moore, 2004, 2005). Dans

un second temps, on peut induire des liens d'alignement à la façon des méthodes probabilistes, comme l'a proposé Melamed (2000) avec le *competitive linking*.

L'approche probabiliste est la plus répandue, principalement du fait de sa bonne intégration avec la traduction automatique statistique, dont elle constitue un fondement depuis l'introduction des modèles IBM (Brown *et al.*, 1993). Les deux approches présentent des forces et faiblesses complémentaires, comme l'ont montré par exemple les travaux de Johnson *et al.* (2007), où les associations extraites à partir d'alignements de mots sont ensuite filtrées selon des mesures d'association.

Nous avons récemment proposé une méthode d'extraction de traductions de segments sous-phrastiques (Lardilleux *et al.*, 2011a), nommée *Anymalign*, qui s'attaque à un certain nombre de problèmes souvent négligés dans le domaine. En particulier, cette méthode permet le traitement d'un nombre quelconque de langues simultanément, ne fait aucune distinction entre source et cible, est massivement parallélisable, passe facilement à l'échelle, et est très simple à implémenter. Cette méthode, qui s'inscrit dans le courant des méthodes associatives, est meilleure que l'état de l'art sur des tâches de constitution de lexiques bilingues. Les résultats obtenus lorsqu'on l'utilise pour construire des modèles de traductions statistiques s'avèrent toutefois inférieurs aux méthodes standard (Lardilleux *et al.*, 2011b).

Une des hypothèses que nous avons précédemment émises pour expliquer ces résultats contrastés est qu'Anymalign ne comporte pas de phase d'alignement à proprement parler. Cette méthode ne produit donc pas de *liens* à la manière des méthodes probabilistes, mais directement des tables de traductions avec leurs scores associés. Ces tables ont des profils très différents de celles extraites à partir d'alignements produits par les méthodes probabilistes, principalement en termes de distribution des n-grammes (Luo *et al.*, 2011). En particulier, malgré de récentes améliorations (Lardilleux *et al.*, 2011b), la quantité de traductions de longs n-grammes est relativement faible comparée aux tables de traductions obtenues à partir des méthodes probabilistes. Dans cet article, nous proposons une extension à notre méthode lui permettant de produire des liens d'alignement, à la manière des approches probabilistes, tout en conservant le caractère local de la recherche des traductions propre aux approches associatives. Notre but principal n'est pas ici de proposer une nouvelle méthode d'alignement destinée à améliorer les outils de l'état de l'art, mais d'essayer de mieux comprendre les limitations actuelles d'Anymalign, en l'utilisant ici de manière non plus directe, mais détournée, pour construire le modèle de traduction. La méthode pour construire des alignements, très simple, est donc indépendante d'Anymalign et pourrait être remplacée par tout autre procédé équivalent.

Cet article est organisé comme suit : la section 2 présente en détail chacune des étapes qui compose notre méthode d'alignement, la section 3 présente une évaluation de la méthode sur des tâches de traduction automatique et une analyse des résultats obtenus, et la section 4 conclut ces travaux.

2 Description de la méthode

En un mot, notre méthode consiste à segmenter chaque couple de phrases d'un corpus parallèle de façon binaire, déterminer parmi les deux segments cible obtenus lequel est la bonne traduction de chacun des deux segments source (traduction monotone ou inversée), et recommencer

récursivement sur chacun des deux couples de segments obtenus.

Ces travaux s'inspirent fortement de ceux de Wu (1997) et Deng *et al.* (2006). Les premiers présentent des grammaires de transduction inversibles où les parties source et cible d'un couple de phrases alignées sont analysées simultanément selon un arbre de dérivation binaire dont la particularité est de permettre l'inversion des constituants d'une langue à l'autre à n'importe quel niveau de l'arbre (approche *bottom-up*). On retrouve un concept similaire dans les seconds, où on extrait des bi-segments plus ou moins grossiers à partir de textes parallèles non préalablement alignés en phrases en appliquant une segmentation binaire de façon itérative selon le principe « diviser pour régner » (approche *top-down*).

Nos travaux se rapprochent davantage de ces derniers en ce sens que nous ne nous intéressons qu'à une procédure simple ne reposant que sur des décomptes au niveau lexical, plutôt que sur une grammaire telle qu'utilisée par Wu. Néanmoins, alors que Deng *et al.* produisent des alignements de segments plus ou moins grossiers à partir d'un bi-texte non préalablement aligné en phrases, dans le but de simplifier des tâches subséquentes d'alignement sous-phrastique par exemple, notre but est plus classiquement d'aligner directement le grain le plus fin possible, ici le mot typographique, à partir de textes préalablement alignés en phrases. Le critère que nous utilisons pour décider de la segmentation d'un couple de phrases est adapté en conséquence.

2.1 Matrice d'alignement

Notre point de départ se compose :
– d'un bi-texte préalablement aligné en phrases ;
– d'une fonction w associant à chaque couple de mots (*source*, *cible*) du bi-texte un score reflétant la force du lien de traduction entre *source* et *cible*.

Plusieurs définitions de w sont possibles ; il est néanmoins naturel de la définir de façon endogène à partir des occurrences des mots sur l'ensemble du bi-texte. En ce qui nous concerne, les scores que nous utiliserons seront dans un premier temps obtenus à partir des sorties d'Anymalign. Nous verrons par la suite que ceux-ci mènent à de meilleurs résultats que d'autres scores obtenus à partir de modèles plus répandus, principalement du fait de la grande redondance des sorties d'Anymalign, qui permet de renforcer les scores de traductions se produisant dans des contextes variés.

Par la suite donc, le score $w(s,c)$ entre un mot source s et un mot cible c sera défini comme le produit des deux probabilités de traduction orientées $\mathrm{p}(s|c) \times \mathrm{p}(c|s)$, celles-ci étant calculées à partir des décomptes associés aux traductions produites par Anymalign :

$$\begin{aligned} w(s,c) &= \mathrm{p}(s|c) \times \mathrm{p}(c|s) \\ &= \frac{\sum_{n=1}^{N} [\![(s,c) \in (S_n, C_n)]\!] k_n}{\sum_{n'=1}^{N} [\![s \in S_{n'}]\!] k_{n'}} \times \frac{\sum_{n=1}^{N} [\![(s,c) \in (S_n, C_n)]\!] k_n}{\sum_{n'=1}^{N} [\![c \in C_{n'}]\!] k_{n'}} \\ &= \frac{\left(\sum_{n=1}^{N} [\![(s,c) \in (S_n, C_n)]\!] k_n\right)^2}{\left(\sum_{n'=1}^{N} [\![s \in S_{n'}]\!] k_{n'}\right) \times \left(\sum_{n'=1}^{N} [\![c \in C_{n'}]\!] k_{n'}\right)} \end{aligned}$$

avec :
– $[\![x]\!] = 1$ si x est vrai, 0 sinon ;

S_n	C_n	k_n
pays	countries	151 190
pays	*country*	17 717
pays tiers	third countries	10 865
les *pays*	countries	6 284
mon *pays*	my *country*	4 057
ces *pays*	these countries	3 742
pays .	*country* .	2 007
état	*country*	122

$$
\begin{aligned}
w(pays, country) &= \mathrm{p}(pays|country) \times \mathrm{p}(country|pays) \\
&= \frac{17\,717 + 4\,057 + 2\,007}{151\,190 + 17\,717 + 10\,865 + 6\,284 + 4\,057 + 3\,742 + 2\,007} \\
&\quad \times \frac{17\,717 + 4\,057 + 2\,007}{17\,717 + 4\,057 + 2\,007 + 122} \\
&\approx 0,121
\end{aligned}
$$

FIG. 1 – Exemple de calcul de score entre le mot source *pays* et le mot cible *country* sur un sous-ensemble d'une table de traductions produite par Anymalign à partir des parties française et anglaise du corpus parallèle Europarl (Koehn, 2005).

– N le nombre d'entrées (couples de segments source–cible) dans la table de traductions produite par Anymalign ;
– S_n (resp. C_n) le segment source (resp. cible) d'une entrée de la table de traductions ;
– k_n le décompte associé au couple (S_n, C_n) dans la table de traductions. Ce nombre n'est pas en soi un indicateur de la qualité de l'entrée ; il s'agit simplement du nombre de fois où le couple a été produit par Anymalign (voir détails dans (Lardilleux *et al.*, 2011a)).

La figure 1 donne un exemple.

En pratique, ce que nous faisons ici revient à partir d'une table de traductions pour aller vers des liens d'alignements — pour retourner ultimement vers une nouvelle table de traductions. Cela va à rebours des usages du domaine, qui construisent la table de traductions à partir de l'ensemble des liens d'alignements calculés sur un corpus parallèle. Cette particularité ouvre de nouvelles pistes pour l'amélioration de la qualité des liens d'alignements et d'une table de traductions, l'amélioration des uns pouvant avoir des répercussions sur l'autre, et vice-versa, de façon itérative, à la manière des approches probabilistes reposant par exemple sur l'algorithme Espérance Maximisation. Cela sort néanmoins du cadre de cet article, et nous nous consacrons pour l'instant au passage de la table de traductions vers les liens d'alignements.

2.2 Critère de segmentation

Le critère de segmentation décrit ci-après est issu des travaux de Zha *et al.* (2001) sur le clustering de documents. Leur problème consiste à partitionner de façon optimale un graphe biparti représentant les occurrences d'un ensemble de termes au sein d'un ensemble de documents. Nous le transposons à la recherche du meilleur alignement entre l'ensemble des mots d'une

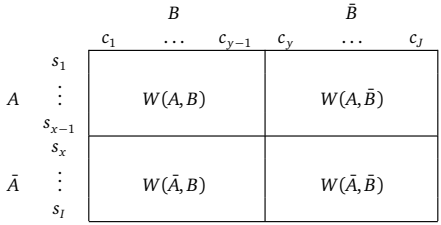


FIG. 2 – Représentation schématique de la segmentation d'un couple de phrases $S = A\,.\,\bar{A}$ et $C = B\,.\,\bar{B}$.

phrase source et l'ensemble des mots d'une phrase cible.

Pour cela, nous considérons un couple de phrases (S,C) du corpus parallèle, où la phrase source S est constituée de I mots source et la phrase cible C est constituée de J mots cible : $S = [s_1 \dots s_I]$ et $C = [c_1 \dots c_J]$. Nous considérons par ailleurs des indices de coupure x et y définissant une segmentation binaire des phrases source et cible (le symbole « . » désigne la concaténation de chaînes de mots) :

$$\begin{array}{lllll} S = A\,.\,\bar{A} & \text{avec} & A = [s_1 \dots s_{x-1}] & \text{et} & \bar{A} = [s_x \dots s_I] \\ C = B\,.\,\bar{B} & \text{avec} & B = [c_1 \dots s_{y-1}] & \text{et} & \bar{B} = [c_y \dots c_J] \end{array}$$

Le choix de x et y sera guidé par la somme W des scores d'association entre chacun des mots source et cible d'un couple de segments $(X,Y) \in \{A,\bar{A}\} \times \{B,\bar{B}\}$:

$$W(X,Y) = \sum_{s \in X, c \in Y} w(s,c)$$

On retrouve l'ensemble des notations utilisées dans la figure 2, qui donne une représentation schématique de la segmentation d'un couple de phrases.

On définit alors :

$$\text{cut}(X,Y) = W(X,\bar{Y}) + W(\bar{X},Y)$$

Notons que $\text{cut}(X,Y) = \text{cut}(\bar{X},\bar{Y})$. Dans notre cas, une valeur faible indique que les scores d'association entre les mots de X et $\bar{Y}$ d'une part, et entre ceux de $\bar{X}$ et Y d'autre part, sont faibles également, autrement dit que ces deux couples de segments ont peu de chances d'être de bonnes traductions, (X,Y) et $(\bar{X},\bar{Y})$ constituant alors *éventuellement* de bonnes traductions. Idéalement donc, nous désirons déterminer le couple (x,y) qui mène à la plus petite valeur de $\text{cut}(X,Y)$ possible. Zha *et al.* (2001) pointent néanmoins le fait que cette quantité tend à produire des segments (clusters de documents dans leur cas) déséquilibrés du fait de l'absence de normalisation, et en proposent par conséquent une version normalisée :

$$\text{Ncut}(X,Y) = \frac{\text{cut}(X,Y)}{\text{cut}(X,Y) + 2 \times W(X,Y)} + \frac{\text{cut}(\bar{X},\bar{Y})}{\text{cut}(\bar{X},\bar{Y}) + 2 \times W(\bar{X},\bar{Y})}$$

Cette variante permet de rajouter une contrainte de densité sur (X,Y) et $(\bar{X},\bar{Y})$, ce qui est partiellement satisfait par l'introduction des dénominateurs dans l'expression ci-dessus. Sa valeur est comprise entre 0 et 2.

```
procédure aligner(S, C) :
    si longueur(S) = 1 ou longueur(C) = 1 :
        lier chacun des mots de S avec chacun des mots de C
        arrêt procédure
    minNcut = 2
    (X, Y) = (S, C)
    pour chaque (i, j) ∈ {2...I} × {2...J} :
        si Ncut(A, B) < minNcut :
            minNcut = Ncut(A, B)
            (X, Y) = (A, B)
        si Ncut(A, B̄) < minNcut :
            minNcut = Ncut(A, B̄)
            (X, Y) = (A, B̄)
    aligner(X, Y)
    aligner(X̄, Ȳ)
```

FIG. 3 – Algorithme d'alignement récursif.

Notre problème consiste finalement à déterminer le couple (x, y) qui minimise Ncut. Bien que des méthodes de recherche performantes existent et sont couramment utilisées en théorie des graphes, nos « graphes » (couples de phrases) sont petits en pratique : environ 30 mots par phrase en moyenne dans le corpus Europarl que nous utilisons pour la suite de nos expériences. Nous nous contentons donc par la suite de déterminer la meilleure segmentation en testant toutes les coupures possibles.

2.3 Algorithme d'alignement

À partir du critère défini précédemment, nous pouvons segmenter et aligner un couple de phrases de façon récursive. À chaque étape, nous testons tous les couples (x, y) possibles afin de déterminer le plus faible Ncut. Le pire des cas se produit lorsque la matrice est coupée de la façon la plus déséquilibrée possible ; la complexité de l'algorithme est donc cubique (de l'ordre de $I \times J \times \min(I, J)$). Pour un couple (x, y) donné, nous calculons deux valeurs : l'une correspondant à un alignement monotone (Ncut(A, B)) et l'autre à une inversion des deux segments (Ncut$(A, \bar{B})$). Le processus est alors appliqué sur chacun des couples de segments correspondant au Ncut minimal. Il s'arrête lorsqu'un segment ne comporte qu'un seul mot : les alignements produits sont tous de multiplicité 1–n ou n–1, et il en résulte que tous les mots sont nécessairement alignés. Des variantes où le processus récursif s'arrête plus tôt sont envisageables, en fixant un seuil sur Ncut par exemple, auquel cas les alignements produits seraient de multiplicité m–n. Nous gardons cette possibilité pour des recherches futures.

La figure 3 présente l'algorithme complet, et la figure 4 illustre le processus sur deux exemples réels. Dans la suite de l'article, nous ferons référence à cet algorithme sous le nom « Cutnalign ».

L'algorithme en lui-même est indépendant de la taille du corpus parallèle à aligner, car chaque couple de phrases est traité indépendemment des autres. On peut donc très facilement paralléliser l'alignement d'un corpus : le temps d'alignement total est divisé par le nombre de processeurs à disposition. Un autre avantage est que les alignements produits sont symétriques tout au long du processus, contrairement à des modèles plus répandus comme les modèles IBM qui produisent de

	the	level	of	budgetary	implementation	;
le	**0,037**	ϵ	0,001	ϵ	ϵ	ϵ
niveau	ϵ	**0,591**	ϵ	ϵ	ϵ	ϵ
d'	ϵ	ϵ	**0,003**	ϵ	ϵ	ϵ
exécution	ϵ	ϵ	ϵ	ϵ	**0,060**	ϵ
budgétaire	ϵ	ϵ	ϵ	**0,659**	ϵ	ϵ
;	ϵ	ϵ	ϵ	ϵ	ϵ	**0,287**

	finally	,	what	our	fellow citizens	are	demanding	is	the	right	to	information	.
enfin	**0,607**	0,001	ϵ	ϵ	0	ϵ	0	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ
,	0,001	**0,445**	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	0,001	ϵ	0,001	ϵ	0,001
c'	ϵ	ϵ	0,001	ϵ	ϵ	ϵ	0	**0,036**	0,001	ϵ	ϵ	ϵ	ϵ
est	ϵ	ϵ	0,001	ϵ	ϵ	ϵ	0	**0,223**	0,016	ϵ	0,001	ϵ	0,001
un	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	0,005	ϵ	ϵ	ϵ	ϵ	ϵ
droit	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	0	ϵ	ϵ	**0,084**	ϵ	ϵ	ϵ
à	ϵ	ϵ	ϵ	ϵ	ϵ	0,001	ϵ	0,001	0,003	0,001	**0,018**	ϵ	ϵ
l'	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	0,002	0,009	ϵ	**0,002**	ϵ	ϵ
information	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	**0,499**	ϵ
que	ϵ	ϵ	0,002	ϵ	ϵ	0,001	ϵ	0,002	0,001	ϵ	0,001	ϵ	ϵ
réclament	0	0	ϵ	ϵ	ϵ	ϵ	**0,152**	ϵ	ϵ	0	0	0	ϵ
nos	ϵ	ϵ	ϵ	**0,171**	0,004	0,001	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ
concitoyens	0	ϵ	ϵ	0,001	**0,323**	**0,009**	ϵ	ϵ	ϵ	0	ϵ	0	ϵ
.	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	0,001	0,001	ϵ	ϵ	ϵ	ϵ	**0,954**

FIG. 4 – Deux exemples de segmentation-alignement. Les nombres dans les cellules correspondent à la valeur de la fonction w, avec ϵ une valeur non nulle inférieure à 0,001. Une valeur de 0 indique que les deux mots n'apparaissent jamais ensemble dans la table de traductions. Les points d'alignement retenus par l'algorithme, correspondant au niveau de récursion maximal, sont indiqués en gras. Dans le premier cas, la traduction est monotone à l'exception de l'inversion de l'ordre du nom et de l'adjectif (*exécution budgétaire/budgetary implementation*), la plupart des liens d'alignement se situent donc sur la diagonale. Le second cas, plus complexe, rend compte de l'inversion de l'ordre de propositions au sein de la phrase.

meilleurs résultats lorsqu'exécutés dans les deux sens de traduction puis leurs sorties combinées à l'aide d'heuristiques.

3 Évaluation

3.1 Description des expériences

Nous évaluons notre méthode d'alignement en tant que premier module d'un système de traduction automatique statistique par segments. Nous utilisons pour cela le système de traduction Moses (Koehn *et al.*, 2007), et des données constituées d'un échantillon du corpus parallèle Europarl (Koehn, 2005), couvrant trois couples de langues : finnois–anglais (langue agglutinante–langue isolante), français–anglais, et portugais–espagnol (langues très proches). Pour chacun, nous utilisons un jeu d'entraînement de 350 000 couples de phrases (30 mots par phrase en moyenne en anglais), et des jeux de développement et de test de 2 000 couples de phrases chacun. L'optimisation des systèmes est réalisée à l'aide de la procédure MERT (Och, 2003). Sauf mention contraire, un modèle de réordonnancement lexicalisé est utilisé.

Nous comparons quatre approches :

MGIZA++ (Gao et Vogel, 2008), implémentant les modèles IBM (Brown *et al.*, 1993) et le modèle caché de Markov de Vogel *et al.* (1996). Intégré à Moses, il s'agit toujours de la référence du domaine. Nous l'utilisons avec ses paramètres par défaut, en enchaînant 5 itérations de chacun des modèles IBM1, HMM, IBM3 et IBM4. Une table de traductions est ensuite produite à partir des alignements à l'aide des outils de Moses.

Anymalign (Lardilleux *et al.*, 2011a), produisant directement des tables de traductions. Cet outil pouvant être arrêté à tout moment, nous fixons son temps d'exécution de façon à ce qu'il soit exécuté pendant la même durée que MGIZA++. Nous répétons la même expérience en faisant varier son paramètre « -i », permettant de contrôler la longueur des segments qu'il produit en sortie, de 1 à 4 (voir détails dans (Lardilleux *et al.*, 2011b)). Nous y faisons référence par la suite sous les noms « Anymalign-1 » à « Anymalign-4 ». Le modèle de réordonnancement utilisé dans cette configuration n'est qu'un simple modèle basé sur la distance entre mots, car Anymalign seul ne peut fournir l'information nécessaire à un modèle de réordonnancement lexicalisé.

Anymalign + Cutnalign : nous appliquons l'algorithme décrit dans la section précédente à chacune des quatre tables de traductions produites par Anymalign-1 à Anymalign-4. Les alignements obtenus sont utilisés pour construire de nouvelles tables de traductions à l'aide du jeu d'outils de Moses.

Simples probabilités + Cutnalign : cette configuration permet d'évaluer non pas l'algorithme proposé précédemment, mais le choix de la fonction w, qui sert de base à l'algorithme. Nous utilisons pour cela un score d'association très simple : la probabilité qu'un mot source et un mot cible soient traductions l'un de l'autre (produit des deux probabilités de traduction), cette probabilité étant calculée à partir de leurs occurrences dans le corpus d'entraînement. La définition de w est donc ici la même qu'à la section 2.1, à deux différences près :
- les décomptes ne sont pas effectués sur une table de traductions produite par Anymalign, mais directement sur le bi-texte d'entraînement ;
- $k_n = 1, \forall n$.

Les traductions sont évaluées selon les mesures BLEU (Papineni *et al.*, 2002) et TER (Snover *et al.*, 2006, contrairement à BLEU, des scores faibles sont meilleurs).

3.2 Résultats

Les résultats sont présentés dans le tableau 1. Sur chacune des trois tâches, Anymalign (version « de base ») est plus ou moins en retrait par rapport à MGIZA++. L'utilisation du paramètre « -i » permet de réduire cet écart de moitié environ, à l'exception notable du couple finnois–anglais (langue agglutinante–langue isolante), ce qui est conforme aux résultats présentés dans (Lardilleux *et al.*, 2011b).

L'ajout de Cutnalign mène à un gain considérable dans toutes les configurations : de 1,6 à 4,6 points BLEU (fr–en, Anymalign-1 + Cutnalign), avec un gain moyen de 2,6 points BLEU et 2,7 points TER. Anymalign+Cutnalign est toujours en retrait de 1,1 à 1,6 point BLEU en finnois–anglais par rapport à MGIZA++, mais produit des résultats de même qualité, voire meilleurs mais de façon non significative, en français–anglais et portugais–espagnol.

L'approche « simples probabilités + Cutnalign » produit des résultats de qualité intermédiaire,

Tâche	Système	BLEU (%)	TER (%)	Entrées (millions)	Long. des entrées	Liens	Long. des blocs extraits
fi–en	MGIZA++	**22,27**	**62,92**	22,2	3,24	26	1,16
	Anymalign-1	18,68	67,30	11,8	1,87		
	Anymalign-2	17,86	68,60	4,4	2,09		
	Anymalign-3	18,06	68,13	3,0	2,32		
	Anymalign-4	18,06	68,53	2,1	2,42		
	Anymalign-1 + Cutnalign	21,14	63,74	7,7	3,26	62	1,45
	Anymalign-2 + Cutnalign	21,14	64,69	7,5	3,27	69	1,48
	Anymalign-3 + Cutnalign	20,83	64,18	7,3	3,29	73	1,50
	Anymalign-4 + Cutnalign	20,64	64,52	7,1	3,29	78	1,53
	Simples prob. + Cutnalign	19,09	67,09	5,5	3,23	74	1,78
fr–en	MGIZA++	29,65	55,25	25,6	4,29	31	1,17
	Anymalign-1	25,10	59,36	6,1	1,27		
	Anymalign-2	26,60	58,16	6,3	1,99		
	Anymalign-3	27,02	57,96	3,9	2,29		
	Anymalign-4	26,85	58,00	2,6	2,42		
	Anymalign-1 + Cutnalign	29,65	**55,22**	12,9	4,21	50	1,49
	Anymalign-2 + Cutnalign	**29,69**	55,44	13,1	4,22	48	1,48
	Anymalign-3 + Cutnalign	29,26	55,49	13,0	4,23	50	1,49
	Anymalign-4 + Cutnalign	29,16	55,46	12,8	4,23	52	1,51
	Simples prob. + Cutnalign	27,97	56,85	10,2	3,95	54	1,62
pt–es	MGIZA++	38,53	48,46	32,2	4,30	30	1,09
	Anymalign-1	35,20	50,89	5,7	1,26		
	Anymalign-2	36,80	49,60	5,9	1,99		
	Anymalign-3	36,82	49,67	3,7	2,26		
	Anymalign-4	36,96	49,80	2,4	2,37		
	Anymalign-1 + Cutnalign	37,35	49,55	17,9	4,30	50	1,32
	Anymalign-2 + Cutnalign	**38,96**	**48,04**	18,0	4,30	48	1,32
	Anymalign-3 + Cutnalign	38,55	48,40	17,7	4,31	50	1,33
	Anymalign-4 + Cutnalign	38,56	48,37	17,3	4,31	54	1,35
	Simples prob. + Cutnalign	37,71	49,04	13,9	4,09	50	1,41

TAB. 1 – Récapitulatif des résultats obtenus dans nos expériences. Les deux premières colonnes de nombres donnent les scores obtenus en traduction automatique. Les deux colonnes du milieu présentent les caractéristiques des tables de traductions : nombre d'entrées et longueur de celles-ci en nombre de mots. Les deux dernières colonnes présentent les caractéristiques des alignements avant production de la table de traductions : nombre moyen de liens d'alignements par couple de phrases d'entraînement et longueur moyenne de la partie source des blocs minimaux extraits (après détermination des segments alignés cohérents avec les liens d'alignement).

généralement entre Anymalign « de base » et Anymalign + Cutnalign. Cela montre que le choix de la fonction w a une grande influence sur le comportement de la méthode d'alignement que nous avons proposée. En admettant que la fonction définie dans ces expériences est une des plus simples qui soient, nous pouvons anticiper que de nombreuses améliorations sont possibles, comme le montrent les résultats obtenus en initiant la méthode à partir des tables de traductions d'Anymalign.

3.3 Regard sur les alignements

Comme précisé en introduction, l'une des raisons pour laquelle nous avons proposé cette méthode d'alignement est que, malgré de récentes améliorations, Anymalign peine toujours à extraire suffisamment de traductions de longs n-grammes. Dans cette section, nous étudions quelques caractéristiques des alignements produits par la méthode que nous avons proposée. Elles sont présentées dans le tableau 1.

En ce qui concerne les tables de traductions d'abord, on constate que celles qui sont obtenues à partir de Cutnalign contiennent un nombre beaucoup plus important d'entrées que les tables correspondantes produites par Anymalign seul[1] (trois fois plus en moyenne), à l'exception notable d'Anymalign-1 en finnois–anglais. Elles sont néanmoins toujours beaucoup plus petites que les tables obtenues à partir de MGIZA++ et contiennent deux fois moins d'entrées en moyenne. La longueur moyenne de ces entrées est en outre quasiment égale à celles des tables de traductions de MGIZA++, alors que celles produites par Anymalign sont beaucoup plus courtes : la production d'une table de traductions à partir de liens d'alignement permet bien de combler le manque de longs n-grammes comme nous le désirions.

Dans un second temps, nous étudions plus en détail les liens d'alignement à proprement parler, tels qu'ils sont avant la production des tables de traductions. La colonne « Liens » du tableau 1 montre que le nombre de liens d'alignement produits par notre méthode est bien supérieur à celui de ceux produits par MGIZA++ : entre 1,5 et 3 fois plus selon la tâche. La dernière colonne en donne la principale raison : les blocs d'alignement extraits par notre méthode, c'est-à-dire les rectangles obtenus au niveau de récursion maximal, sont toujours plus longs que les blocs minimums obtenus à partir des alignements de MGIZA++ (+ 26 % en moyenne). Comme nous alignons systématiquement tous les mots source avec tous les mots cible d'un tel rectangle, et tous les mots d'un couple de phrases étant par conséquent nécessairement alignés, le nombre total de liens produits est naturellement élevé. Cela explique également le fait que le nombre d'entrées dans les tables de traductions est toujours beaucoup plus faible que dans celles obtenues à partir de MGIZA++, ce dernier produisant des alignements de multiplicité 0–1 qui sont à l'origine de l'extraction de très nombreux segments lors de la constitution de la table par Moses (heuristique *grow-diag-final-and* par défaut) (Ayan et Dorr, 2006). Malgré cela, les alignements produits par notre méthode permettent d'atteindre des scores identiques à l'état de l'art dans deux tâches de traduction automatique sur trois dans nos expériences.

[1] Ces tables ont été produites en exécutant Anymalign pendant un temps identique dans les quatre configurations, ce qui explique pourquoi de plus grandes valeurs de l'option « -i » mènent à de plus petites tables — voir détails dans (Lardilleux *et al.*, 2011b).

4 Conclusion

Nous avons présenté une méthode d'alignement sous-phrastique fondée sur un découpage récursif binaire de la matrice d'alignement entre une phrase source et sa traduction. Inspirée des travaux de Wu (1997) et Deng *et al.* (2006) sur l'alignement et de Zha *et al.* (2001) sur le clustering de documents, nous avons montré qu'en dépit de sa simplicité, cette méthode produit des résultats du niveau de l'état de l'art dans deux tâches sur trois dans nos expériences. Couplée à Anymalign, elle permet des gains conséquents (jusqu'à 4,6 points BLEU en français–anglais) par rapport à l'utilisation d'Anymalign seul. Nos expériences ont confirmé que le principal handicap d'Anymalign concerne bien les traductions de longs n-grammes. Une étape complémentaire d'alignement au sens strict du terme se révèle donc souhaitable pour améliorer ses résultats en traduction automatique, car elle permet de combler la plupart de ses manques en termes de traductions de longs segments. La méthode d'alignement proposée ici est relativement simple, symétrique du point de vue du sens de la traduction, et le caractère local du calcul des alignements lui permet de passer facilement à l'échelle. Dans l'optique d'améliorer les alignements, de multiples enrichissements de la méthode sont possibles, comme par exemple l'intégration des valeurs seuils lors de la recherche du meilleur découpage de la matrice afin d'arrêter le processus d'alignement à des blocs plus larges et plus sûrs, ou encore l'examen d'un découpage ternaire plutôt que binaire afin de rendre compte de constructions linguistiques plus complexes générant des constituants non connexes.

Remerciements

Ces travaux ont été financés par le projet Cap Digital SAMAR.

Références

Ayan, N. F. et Dorr, B. J. (2006). Going beyond AER : an extensive analysis of word alignments and their impact on MT. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 9–16, Sydney, Australie.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R. et Roossin, P. (1988). A statistical approach to language translation. *In Proceedings of the 12th International Conference on Computational Linguistics (Coling'88)*, pages 71–76, Budapest.

Brown, P., Della Pietra, S., Della Pietra, V. et Mercer, R. (1993). The mathematics of statistical machine translation : Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Dagan, I. et Church, K. (1994). Termight : identifying and translating technical terminology. *In Proceedings of the fourth conference on Applied natural language processing*, pages 34–40, Stuttgart.

DeNero, J. et Klein, D. (2007). Tailoring word alignments to syntactic machine translation. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, pages 17–24, Prague.

DENG, Y. et BYRNE, W. (2005). HMM word and phrase alignment for statistical machine translation. *In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 169–176, Vancouver, British Columbia, Canada.

DENG, Y., KUMAR, S. et BYRNE, W. (2006). Segmentation and alignment of parallel text for statistical machine translation. *Natural Language Engineering*, 13(3):235–260.

DUNNING, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

FRASER, A. et MARCU, D. (2007). Getting the structure right for word alignment : LEAF. *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 51–60, Prague.

FUNG, P. et CHURCH, K. (1994). K-vec : A new approach for aligning parallel texts. *In Proceedings of the 15th International Conference on Computational Linguistics (Coling'94)*, volume 2, pages 1096–1102, Kyōto.

FUNG, P. et YEE, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. *In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 1, pages 414–420, Montreal.

GALE, W. et CHURCH, K. (1991). Identifying word correspondences in parallel texts. *In Proceedings of the fourth DARPA workshop on Speech and Natural Language*, pages 152–157, Pacific Grove.

GANCHEV, K., GRAÇA, J. et TASKAR, B. (2008). Better alignments = better translations ? *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (ACL-08 : HLT)*, pages 986–993, Columbus, Ohio.

GAO, Q. et VOGEL, S. (2008). Parallel implementations of word alignment tool. *In Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus (Ohio, USA).

GAUSSIER, E. et LANGÉ, J.-M. (1995). Modèles statistiques pour l'extraction de lexiques bilingues. *Traitement Automatique des Langues*, 36(1-2):133–155.

JOHNSON, H., MARTIN, J., FOSTER, G. et KUHN, R. (2007). Improving translation quality by discarding most of the phrasetable. *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 967–975, Prague.

KOEHN, P. (2005). Europarl : A parallel corpus for statistical machine translation. *In Proceedings of the tenth Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket.

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A. et HERBST, E. (2007). Moses : Open source toolkit for statistical machine translation. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 177–180, Prague.

KOEHN, P., OCH, F. et MARCU, D. (2003). Statistical phrase-based translation. *In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, pages 48–54, Edmonton.

LARDILLEUX, A., LEPAGE, Y. et YVON, F. (2011a). The contribution of low frequencies to multilingual sub-sentential alignment : a differential associative approach. *International Journal of Advanced Intelligence*, 3(2):189–217.

LARDILLEUX, A., YVON, F. et LEPAGE, Y. (2011b). Généralisation de l'alignement sous-phrastique par échantillonnage. *In Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011)*, volume 1, pages 507–518, Montpellier.

LIANG, P., TASKAR, B. et KLEIN, D. (2006). Alignment by agreement. *In Proceedings of the Human Language Technology Conference of the NAACL*, pages 104–111, New York City.

LUO, J., LARDILLEUX, A. et LEPAGE, Y. (2011). Improving sampling-based alignment by investigating the distribution of n-grams in phrase translation tables. *In Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, pages 150–159, Singapour.

MARCU, D. et WONG, D. (2002). A phrase-based, joint probability model for statistical machine translation. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 133–139, Philadelphie.

MELAMED, D. (2000). Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

MOORE, R. (2004). On log-likelihood-ratios and the significance of rare events. *In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 333–340, Barcelona.

MOORE, R. (2005). Association-based bilingual word alignment. *In Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 1–8, Ann Arbor.

OCH, F. (2003). Minimum error rate training in statistical machine translation. *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 160–167, Sapporo.

OCH, F. et NEY, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19–51.

PAPINENI, K., ROUKOS, S., WARD, T. et ZHU, W.-J. (2002). BLEU : a method for automatic evaluation of machine translation. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318, Philadelphie.

SMADJA, F., HATZIVASSILOGLOU, V. et MCKEOWN, K. (1996). Translating collocations for bilingual lexicons : A statistical approach. *Computational Linguistics*, 22(1):1–38.

SNOVER, M., DORR, B., SCHWARTZ, R., MICCIULLA, L. et MAKHOUL, J. (2006). A study of translation edit rate with targeted human annotation. *In Proceedings of the 7th Conference of the Association for Machine Translation of the Americas (AMTA 2006)*, pages 223–231, Cambridge.

VOGEL, S. (2005). PESA : Phrase pair extraction as sentence splitting. *In Proceedings of the tenth Machine Translation Summit (MT Summit X)*, pages 251–258, Phuket.

VOGEL, S., NEY, H. et TILLMAN, C. (1996). Hmm-based word alignment in statistical translation. *In Proceedings of the 16th International Conference on Computational Linguistics (Coling'96)*, pages 836–841, Copenhague.

WU, D. (1997). Stochastic inversion transduction grammar and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404.

ZHA, H., HE, X., DING, C., SIMON, H. et GU, M. (2001). Bipartite graph partitioning and data clustering. *In Proceedings of the tenth international conference on Information and knowledge management*, pages 25–32, Atlanta.

Utilisation de la translittération arabe pour l'amélioration de l'alignement de mots à partir de corpus parallèles français-arabe

Houda Saadane[1] *Nasredine Semmar*[2]
(1) LIDILEM, Université de Grenoble, 38400 Grenoble Cedex 9
(2) CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, 91191 Gif-sur-Yvette Cedex
houda.saadane@e.u-grenoble3.fr, nasredine.semmar@cea.fr

RESUME

Dans cet article, nous nous intéressons à l'utilisation de la translittération arabe pour l'amélioration des résultats d'une approche linguistique d'alignement de mots simples et composés à partir de corpus de textes parallèles français-arabe. Cette approche utilise, d'une part, un lexique bilingue et les caractéristiques linguistiques des entités nommées et des cognats pour l'alignement de mots simples, et d'autre part, les relations de dépendance syntaxique pour aligner les mots composés. Nous avons évalué l'aligneur de mots simples et composés intégrant la translittération arabe en utilisant deux procédés : une évaluation de la qualité d'alignement à l'aide d'un alignement de référence construit manuellement et une évaluation de l'impact de cet alignement sur la qualité de la traduction en faisant appel au système de traduction automatique statistique Moses. Les résultats obtenus montrent que la translittération améliore aussi bien la qualité de l'alignement que celle de la traduction.

ABSTRACT

Using Arabic transliteration to improve word alignment from French-Arabic parallel corpora

In this paper, we focus on the use of Arabic transliteration to improve the results of a linguistic word alignment approach from parallel text corpora. This approach uses, on the one hand, a bilingual lexicon, named entity and cognates linguistic properties to align single words, and on the other hand, syntactic dependency relations to align compound words. We have evaluated the word aligner integrating Arabic transliteration using two methods: A manual evaluation of the alignment quality and an evaluation of the impact of this alignment on the translation quality by using the statistical machine translation system Moses. The obtained results show that Arabic transliteration improves the quality of both alignment and translation.

MOTS-CLES : Translittération, alignement de mots, construction de dictionnaires multilingues, traduction automatique, recherche d'information interlingue.
KEYWORDS : Transliteration, word alignment, multilingual lexicons construction, machine translation, cross-language information retrieval.

1 Introduction

La translittération consiste à substituer à chaque graphème un système d'écriture, un autre graphème ou un groupe de graphèmes d'un autre système d'écriture,

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 127–140,
Grenoble, 4 au 8 juin 2012.

indépendamment de la prononciation.

La translittération connait un essor important en raison du caractère de plus en plus multilingue de l'Internet et des besoins exponentiels dans le domaine de la recherche d'information interlingue. Cela est d'autant plus vrai pour la recherche d'entités nommées (noms de personnes, de lieux, de sociétés, d'organisations, etc.), mais ces dernières présentent une pluralité de formes écrites, d'orthographes et de transcriptions selon les langues et les pays. Le cas des noms propres en arabe illustre cette situation complexe et multiforme. Le meilleur exemple pour montrer cette pluralité est le nom معمر القذافي (Mouammar Kadhafi) qui est transcrit en latin par plus de 60 formes, parmi lesquelles : Muammar Qaddafi, Mo'ammar Gadhafi, Muammer Kaddafi, Moammar El Kadhafi, etc.

Cet article décrit un système de translittération automatique de noms arabes en écriture latine et montre une utilisation concrète de la translittération arabe en alignement de mots à partir de corpus de textes parallèles dans le but d'améliorer la qualité des lexiques bilingues ainsi construits.

Nous présentons dans la section 2 un résumé de l'état de l'art dans le domaine de la translittération. Dans la section 3, nous décrivons les approches que nous avons utilisées pour développer notre système de translittération automatique des noms arabes voyellés et non voyellés vers les différentes transcriptions possibles en écriture latine. Nous montrons dans la section 4 comment cette translittération est utilisée pour améliorer les résultats d'un outil d'alignement de mots. Nous présentons en section 5 les résultats que nous avons obtenus en précisant les taux d'amélioration de la qualité de l'alignement et de la traduction. La section 6 conclut notre étude et présente nos travaux futurs.

2 État de l'art

Le problème de la translittération a intéressé les spécialistes dans plusieurs langues, mais cet intérêt est relativement récent et lié au développement croissant de l'utilisation de l'Internet. De nombreux travaux ont été réalisés pour aligner automatiquement les translittérations à partir de corpus de textes multilingues en vue de l'enrichissement de lexiques bilingues indispensables pour la recherche d'information interlingue et la traduction automatique. Citons notamment (Yaser et Knight, 2002) et (Sherif et Kondrak, 2007), qui ont travaillé sur l'alignement arabe-anglais, (Tao et al., 2006) qui ont travaillé sur l'arabe, le chinois et l'anglais ainsi que (Shao et Ng, 2004) qui utilisent l'information apportée par les translittérations sur la base de leur prononciation. Ils combinent l'information apportée par le contexte des traductions avec l'information apportée par les translittérations entre l'anglais et le chinois. L'intérêt de ce travail réside dans le fait qu'il permet l'alignement de mots très spécifiques mais rares.

On trouve ainsi des propositions de systèmes visant à attribuer une seule translittération à un nom donné : c'est le cas du modèle génératif proposé pour les noms d'origine anglaise écrits en japonais (Katakana) vers le système d'écriture latin (Knight et Graehl, 1997).

Cette approche a été adaptée par (Stalls et Knight, 1998) à la façon dont un nom anglais écrit en arabe est transcrit en anglais. Le système de génération de translittérations

s'appuie sur un dictionnaire d'apprentissage et ne prend pas en compte les prononciations non répertoriées ou inconnues du dictionnaire.

Cela a conduit certains chercheurs à pallier cette carence par un recours à la technique statistique. C'est le cas du système de translittération des noms anglais vers l'arabe proposé par (Abduljaleel et Larkey, 2003). Mais celui-ci a montré également ses limites parce qu'il est basé sur le calcul de la forme la plus probable, censée être la forme correcte, ce qui n'est pas vrai pour tous les pays arabes ni pour tous les dialectes.

Pour contourner la difficulté de la prononciation et le problème des variantes dialectales, (Alghamdi, 2005) a proposé un système de translittération en écriture anglaise des noms arabes voyellés. Ce système est basé sur un dictionnaire de noms arabes dans lequel la prononciation est réglée au moyen de voyelles ajoutées aux noms répertoriés, avec indication en vis à vis de leur équivalent en écriture anglaise. Mais cette approche cumule les inconvénients des deux précédentes : non seulement elle ne prend pas en compte les prononciations non répertoriées dans le dictionnaire, mais en plus elle est normative par le fait qu'elle ne propose qu'une seule translittération pour un nom donné. L'objectif de l'auteur semblerait être de favoriser l'adoption d'un standard de translittération, mais cela ne peut être le résultat d'une initiative individuelle et isolée.

En réalité, l'état actuel de la recherche dans ce domaine ne rend pas compte de la complexité du problème de la transcription et de la translittération, lequel touche autant à l'oralité qu'à la scripturalité dans deux ou plusieurs systèmes linguistiques en même temps. En effet, transcrire un nom ou un prénom d'un système linguistique source vers un système d'écriture cible, est une tâche délicate qui nécessite un certain nombre d'opérations exigeant de prendre en considération un ensemble de propriétés morphologiques, phonologiques et sémantiques. Ces opérations sont nécessaires pour assurer un processus de translittération robuste, notamment pour des applications de sécurité, de vérification d'identité, ou encore de recherche d'informations sur Internet.

Or, très peu d'études prennent en considération le lien :

- entre phonologie comparée et transcription interlingue;
- entre graphématique comparée et translittération multilingue;
- entre dialectologie arabe et systèmes de translittération latins.

Les rares études qui proposent une solution prenant en compte partiellement l'une de ces problématiques, sont dédiées à l'identification automatique de l'origine du locuteur à partir de son dialecte. C'est le cas notamment des travaux de (Guidère, 2004) et de (Barkat-Defradas et al., 2004).

3 Translittération en caractères latins des noms écrits en arabe standard

Le système d'écriture de la langue arabe standard est constitué d'un alphabet de 28 lettres, dont 25 consonnes et 3 voyelles, celles-ci pouvant être courtes ou longues en fonction du mot.

Il existe également des phénomènes morphologiques et phonologiques particuliers dont il faut tenir compte dans la translittération tels que le dédoublement des consonnes,

parfois matérialisé dans l'écriture arabe par la «shadda», et le redoublement des voyelles, parfois matérialisé dans l'écriture arabe par le «tanwin». Mais l'écriture arabe moderne présente la particularité de ne pas marquer dans les textes –de manière générale– ni le dédoublement ni les voyelles courtes, ce qui constitue l'une des principales sources d'ambiguïté pour les systèmes de translittération.

3.1 Méthodologie de construction du translittérateur

Nous avons choisi une méthodologie ascendante pour la construction de notre translittérateur. En d'autres termes, nous avons commencé par faire un recensement des translittérations existantes pour chaque lettre de l'alphabet arabe standard à partir des normes et des usages observés sur Internet. Cette investigation empirique est basée sur un corpus de textes qui a été recueilli dans les différentes langues cibles visées par le translittérateur. Elle a permis de constituer une librairie des équivalents graphématiques actuellement en usage dans les écrits utilisant l'alphabet latin.

Nous faisons figurer dans le tableau suivant quelques équivalences graphématiques établies à partir de cette étude sur corpus :

Lettre arabe	Équivalent en écriture latine	Lettre arabe	Équivalent en écriture latine
ء	a	غ	Gh, gh, Ǧ, ǧ, ḡ
ا	A, a, ä, â, á, ā, e, ê	ف	F, f, ph
ب	B, b	ق	Q, q, C, c, K, k
ت	T, t	ك	K, k, C, c
ث	Th, th, t, t̲	ل	L, l

TABLE 1 – Exemples d'équivalences graphématiques entre les alphabets arabe et latin

L'étude sur corpus a également permis de constater que certaines lettres arabes, sans équivalent graphématique dans l'écriture latine, étaient transcrites par le biais de chiffres arabes dans les textes écrits en caractères latins. Ce type de translittération est particulièrement utilisé dans les messages téléphoniques (SMS) et dans les sites web sociaux en Europe et au Moyen Orient. Le tableau suivant récapitule ces équivalences alphanumériques pour les lettres concernées de l'alphabet arabe :

Lettre	ء	ح	خ	ص	ض	ط	ظ	ع	غ	ق
Équivalence alphanumérique	2	7	7'	9	9'	6	6'	3	3'	8

TABLE 2 – Équivalences alphanumériques dans les textes écrits en alphabet latin

Ainsi, en combinant ces deux types de représentation symbolique, on peut rencontrer dans les textes des translittérations qui illustrent ces différentes équivalences pour des noms et des prénoms courants dans le monde arabe :

Nom en arabe	منى	عدنان	حنان	طارق
Exemple d'équivalents en écriture latine	Mouna ou Mona…	Adnane ou 3adnan…	Hanane ou 7anan…	Tarek ou 6ariq…

TABLE 3 – Exemples de noms et prénoms arabes

Cette variation dans les usages translittérationnels, source d'ambiguïté lors du traitement automatique et de la recherche d'information, s'explique par trois types de raisons :

Tout d'abord, des raisons historiques puisque certains pays arabes ont été colonisés ou placés sous mandat français ou britannique pendant une période plus ou moins longue selon les pays et ont, par conséquent, gardé de cette période des traces dans leur vocabulaire, dans leur prononciation et dans la manière dont ils ont tendance à translittérer les noms et les prénoms. Ainsi, l'influence du système linguistique et graphématique du français est perceptible dans les usages translittérationnels des pays du Maghreb, de manière plus ou moins forte selon les pays. Il en est de même des pays du Proche et du Moyen-Orient par rapport à l'influence britannique ou américaine.

Ensuite, pour des raisons politiques puisqu'il n'existe pas de norme commune ni de stratégie unifiée dans le domaine de la translittération pour ce qui est de la langue arabe. Cela a conduit chaque écrivain ou scripteur à s'appuyer sur la prononciation dialectale qui lui était la plus familière pour transcrire les noms arabes. L'exemple le plus célèbre est celui de Laurence d'Arabie qui, pour transcrire le nom de la ville de Djeddah (جدة) en Arabie Saoudite, utilise : 25 fois l'orthographe « Jeddah », 6 fois l'orthographe « Jidda », et 1 fois l'orthographe « Jedda », et cela dans le même ouvrage (1926). Laurence d'Arabie justifie cette variation dans la translittération de la manière suivante : « On ne peut pas transcrire correctement et de la même façon un nom arabe à cause des consonnes qui diffèrent des consonnes latines et des voyelles dont la prononciation diffère d'une région à une autre. » (Alsalman et al., 2007). Cela est d'autant plus vrai que les différentes orthographes données par Laurence d'Arabie diffèrent de l'usage actuel en Arabie Saoudite pour la transcription du nom de cette même ville : « Jaddah ».

Enfin, pour des raisons dialectologiques puisqu'il existe une telle variété de parlers régionaux et locaux dans le monde arabe qu'il est impossible de retrouver la même prononciation d'un pays à l'autre et d'une région à l'autre. Ainsi par exemple, l'un des prénoms arabes les plus répandus, celui du Prophète Muhammad (محمد) – transcrit en français Mahomet depuis l'époque moderne – possède une dizaine de prononciations – et donc de transcriptions – différentes. Citons notamment : Mohamed, Mouhammad, Muhamed, Mhamed, M'Hamed, Muhammad, etc. Même lorsque ce prénom est voyellé (مُحَمَّدٌ), il présente plusieurs translittérations dans les textes : Muhamad, Mouhamad, Mohamad, Mehammad, Mehammade.

Cette variation dans les translittérations possibles selon les dialectes est parfois accompagnée par l'utilisation de caractères spéciaux dans certaines régions ou pays

arabes. Citons comme exemples les noms suivants qui présentent des formes non conventionnelles en écriture latine : Mu`ammar, Mabrūk, Muṣṭafá, Ismā`īl, Hâdî.

Tous ces phénomènes nécessitent une observation fine en amont du traitement pour identifier les cas problématiques et construire des règles efficaces permettant l'automatisation du processus de translittération des noms arabes en temps réel.

3.2 Fonctionnement du translittérateur de l'arabe vers le latin

Le module de translittération de l'écriture arabe vers l'écriture latine est fondé sur les automates d'états finis. Cela signifie qu'il est constitué d'états et de transitions. Son fonctionnement est déterminé par la nature du mot fourni en entrée : l'automate passe d'état en état suivant les transitions, à la lecture de chaque lettre arabe de l'entrée.

A l'issue de la lecture, l'automate produit une réponse « oui » ou « non », c'est-à-dire qu'il accepte (oui) ou rejette (non) l'entrée en question : voyellée ou non-voyellée. Ensuite, il traite l'entrée de la manière suivante : si voyellée, il supprime les voyelles avant de translittérer le nom; si non-voyellée, il procède directement à la translittération du nom. Enfin, le module produit en sortie une liste triée de noms arabes écrits en caractères latins.

Le cœur du système de translittération est constitué de règles contextuelles. Ces règles visent à rendre compte de la manière la plus précise possible des formes observées en entrée : s'agit-il d'une « kunya » ? d'un nom précédé d'un article ? ou bien d'un prénom seul ?

On sait à cet égard que le nom d'une personne contient plusieurs éléments en arabe. Il est constitué en principe de quatre composants principaux :

1. La « Kunya » (particule d'usage) : généralement composée de « Abou » (père de…), suivi du nom d'un enfant ou bien de « Oum » (mère de + nom d'un enfant de la famille). Exemple : « Abou Omar » (Père d'Omar), «Oum Mohamed» (Mère de Mohamed), etc.
2. Le « Ism » (Prénom) : par exemple, Omar, Ali, Mohamed, Khaled, Abdallah, etc. Il indique parfois l'origine ethnique ou confessionnelle de celui qui le porte : par exemple, « Omar » est un prénom typiquement sunnite ; « Rustam » est un prénom typiquement iranien ; « Arslan » est typiquement turc, etc.
3. Le « Nasab » (particule généalogique) : chaque nom est précédé par « Ibn » ou «Bin/Ben» («Bint/Bent» pour les femmes). Il indique la filiation généalogique exacte de l'individu concerné. Les Arabes remontent parfois très loin dans l'indication des ancêtres pour éviter les confusions entre personnes : ex. Muhammad Bin Abdallah Bin Salih Bin Said, etc.
4. La « Nisba » (suffixe d'origine) : ce suffixe renvoie en principe à la tribu ou au clan dans la généalogie ancienne mais aujourd'hui, il désigne surtout le lieu de naissance des individus : Maghribi (né au Maroc), Libi (né en Libye), Masri (né en Égypte), etc. La « Nisba » est toujours précédée de l'article [Al-] et se termine par le suffixe [i]. Elle indique la résidence territoriale initiale des personnes, ou encore leur nationalité.

Selon la forme d'entrée, on applique d'abord des règles adéquates pour transcrire la

partie qui ne constitue pas le nom à proprement parler (particules), puis on applique les règles pour la translittération des noms eux-mêmes.

Les règles pour la translittération des noms s'appliquent à leur tour selon le nombre de consonnes du nom considéré, et dans un ordre de priorité déterminé. Par exemple, Si le mot est composé par Abd (عبد) + Al (ال) + Nom (رحيم), le système procède de la manière suivante :

- Translittération de la particule عبد « Abd »;
- Translittération de l'article ال « Al »;
- Concaténation de la particule « Abd » et de l'article « Al » en les reliant au nom par un trait d'union ou en insérant un blanc entre les deux : Abd Al-Rahim (عبد الرحيم) ;
- Génération de toutes les formes de translittération possibles pour ces trois éléments :

Nom propre arabe	Translittérations
عبد الرحيم	Abd Al-Rahim
	Abd Al Rahim
	Abd al-Rahim
	Abd al Rahim
	Abd El-Rahim
	Abd El Rahim
	Abd el-Rahim
	Abd el Rahim
	Abd Ar-Rahim
	Abd Ar Rahim
	Abd Ar-Rahîm
	Abd ar-Rahim

TABLE 4 – Quelques formes de translittération pour le nom propre عبد الرحيم

Une étape intermédiaire s'ajoute afin de procéder à d'autres traitements, pour ne pas occulter l'un des problèmes très difficile de la transcription, comme la transcription de certains noms propres qui changent totalement phonétiquement pour des raisons religieuses ou autres : c'est le cas de Moussa qui est traduit par Moïse, Yussuf par Josef,

Yaakoub par Jackoub, Hawa par Eve, etc. Cette étape consiste à fournir ces transcriptions dans une liste.

Une fois générée la liste triée des noms translittérés, on procède à deux types de traitements :

- Normalisation de la liste des noms en écriture latine : cette phase consiste à effectuer certains traitements sur la sortie du nom en écriture latine tels que la suppression des caractères spéciaux (diacritiques et chiffres) et l'ajout de la majuscule au début de nom propre, étant donné que les majuscules n'existent pas dans l'écriture arabe des noms. Cette notion de majuscule est conservée seulement dans le cas d'une utilisation dans des bases de données, mais elle n'est pas ajoutée pour les moteurs de recherche usuels, qui ne considèrent pas la casse comme pertinente;
- Pondération de la liste des noms en écriture latine : cette étape consiste à attribuer un poids aux règles qui ont servi à la génération de la liste, afin de pouvoir afficher les résultats en sortie du plus probable vers le moins probable, ou inversement. Pour réaliser cette pondération, nous utilisons le moteur de recherche Google en notant à chaque fois le nombre d'occurrences pour chaque forme générée du nom propre : par exemple pour le prénom arabe جمال (jamal), le système génère trois translittérations distinctes et attestées dans les textes (Djamel, Jamel, Gamel) et le calcul de fréquences fournit les résultats suivants :

Forme translittérée du nom en écriture latine	Nombre moyen d'occurrences du nom sur le moteur de recherche Google
Djamel	4000000
Jamel	5500000
Gamel	500000

TABLE 5 – Résultats pour les formes translittérées du prénom جمال

Du point de vue de la pondération, cet exemple permet de constater que la lettre arabe (ج) est transcrite, en termes de fréquence, majoritairement par la lettre (J), puis par la graphie (Dj), puis par la lettre (G).

Cette procédure a été appliquée à toutes les formes de translittération des caractères arabes. Elle a permis d'établir une liste d'équivalences pondérée au niveau des graphèmes, qui sert à afficher les résultats en sortie du plus probable vers le moins probables.

4 Utilisation de la translittération en alignement de mots

L'outil d'alignement de mots simples et composés utilisé dans cette étude est décrit dans (Semmar et Laib, 2010). Cet outil utilise les ressources et les modules suivants :

- un lexique bilingue français-arabe composé de 124581 entrées. Les entrées de ce

lexique sont utilisées comme des points d'ancrage pour réduire l'espace de recherche des mots à aligner dans les phrases source et cible ;
- un module pour l'appariement des entités nommées présentes dans les phrases source et cible ;
- un module permettant d'apparier les catégories grammaticales des mots composant les phrases source et cible. Ce module utilise les positions des mots à apparier par rapport aux entrées du lexique bilingue et des entités nommées déjà alignées ;
- un module pour l'appariement des mots composés identifiés à partir des mots simples déjà alignés et les relations de dépendance syntaxique entre ces mots.

Les entrées de cet outil d'alignement sont les sorties (résultats) d'une analyse morpho-syntaxique effectuée à l'aide de la plate-forme d'analyse linguistique LIMA (Besançon et al., 2010) sur le corpus de textes parallèles. Cette plate-forme fournit pour chaque couple de phrases source et cible :

- les lemmes et les formes fléchies des mots ainsi que leur position dans la phrase,
- les catégories grammaticales des mots,
- les entités nommées,
- les relations de dépendance syntaxique entre les mots,
- les mots composés.

Nous avons constaté lors de l'alignement de mots à partir de corpus de textes parallèles anglais-arabe ou français-arabe que beaucoup de noms arabes ne sont pas reconnus comme entités nommées par la plate-forme LIMA. Cela vient du fait que cette plate-forme utilise des listes ainsi que des règles de déclencheurs pour reconnaître des entités telles que les noms de personnes, d'organisations, de lieux… mais ces listes sont limitées et plus particulièrement pour les langues peu dotées comme l'arabe. C'est pour cette raison que nous avons ajouté un module supplémentaire à notre outil d'alignement de mots. Ce module est utilisé pour permettre l'appariement des cognats présents dans les phrases source et cible. Nous considérons comme cognats les mots dont les quatre premiers caractères sont identiques.

Cette étape utilise la translittération des noms propres et permet de détecter, par exemple, que le nom propre « Jackson » et la translittération du mot arabe « جاكسون » («jackson») sont des cognats. En revanche, cet algorithme ne permet pas de détecter des couples de mots comme « blair » et « bleer » (translittération du mot arabe «بْلِير»). Pour ce faire, nous avons défini une similarité basée sur le nombre de lettres en commun. Ceci permettra de détecter les couples de mots cités précédemment ainsi que les noms propres et les expressions numériques. L'algorithme de détection de cognats a été adapté pour ne sélectionner que les mots de taille proche et avec un nombre important de caractères en commun sans tenir compte de l'ordre de ces caractères. L'adaptation de cet algorithme a été réalisée en ajoutant les deux paramètres « Ratio_mots » et « Ratio_cognats » définis comme suit :

$$Ratio\_mots = (Nombre\ de\ caractères\ du\ mot\ court)/(Nombre\ de\ caractères\ du\ mot\ long)$$

$$Ratio\_cognats = (Nombre\ de\ caractères\ en\ commun)/(Nombre\ de\ caractères\ du\ mot\ court)$$

Deux mots sont cognats si Ratio_mots est supérieur à 0,8 et Ratio_cognats est supérieur à

0,5. Les valeurs de ces deux paramètres ont été fixées empiriquement.

Cette adaptation permet certes d'identifier comme cognats le mot « blair » et la translittération « bleer » mais il génère aussi des erreurs comme c'est le cas du couple de mots « mohamed » et la translittération « mahmoud ». Pour réduire le taux d'erreurs de ce module, nous avons ajouté un critère supplémentaire relatif aux positions des deux mots dans les phrase source et cible.

Le tableau 6 présente le résultat de l'alignement des mots simples et composés de la phrase source « M. Blair a imposé des frais d'inscription élevés à l'université qui ont introduit une sélection par l'argent. » et sa traduction en langue cible « فرض بلير رسوم تسجيل مرتفعة في الجامعة مما ادى الى اختيار الطلاب على قاعدة المال.».

Lemmes des mots simples et composés de la phrase source	Lemmes des mots simples et composés de la phrase cible
Blair	بْلير
imposer	فْرَضَ
frais	رَسْم
inscription	تَّسْجِيل
élevé	مُرْتَفِع
université	جَامِعَة
introduire	أدَّى
sélection	إخْتِيَار
argent	مَال
frais_inscription	رَسْم_تَّسْجِيل

TABLE 6 – Résultats de l'alignement de mots simples et composés

Le mot « Blair » a été aligné à l'aide de l'appariement de cognats après translittération, les mots « frais », « élevé » et « introduire » ont été alignés à l'aide de l'appariement de catégories grammaticales et les autres mots existent dans le lexique bilingue. Le mot composé « frais_inscription » a été aligné en utilisant les alignements de ces composants « frais » et « inscription ». Notons que le mot arabe « قاعدة » (base) n'a pas de mot qui lui correspond dans la phrase en langue source (français).

5 Résultats expérimentaux

Pour illustrer l'apport de la translittération sur la qualité de l'alignement de mots simples

et composés, nous avons évalué les résultats de l'alignement selon deux approches différentes :

- une évaluation manuelle comparant les résultats de notre aligneur de mots par rapport à un alignement de référence ;
- une évaluation automatique en intégrant les résultats de notre aligneur de mots dans le corpus d'apprentissage du modèle de traduction du système Moses (Koehn et al., 2007).

L'évaluation manuelle de l'aligneur de mots a été réalisée sur une partie composée de 283 phrases du corpus MD (Monde Diplomatique) français-arabe de la campagne ARCADE II (Veronis et al., 2008). Le choix d'une telle taille de corpus s'explique par le fait que la constitution de l'alignement de référence est une tâche coûteuse puisque l'identification des alignements des mots simples et composés est réalisée manuellement sur les 283 phrases. Pour les métriques d'évaluation, nous avons utilisé celles du protocole défini lors de la conférence HLT/NAACL 2003 (Mihalcea et Pedersen, 2003).

Le tableau 7 résume nos résultats en termes de précision et de rappel selon que l'aligneur de mots utilise ou non l'appariement de cognats avec la translittération de noms propres arabes. Ces résultats montrent que l'utilisation de la translittération arabe permet d'augmenter aussi bien la précision que le rappel.

Alignement de mots	Précision	Rappel	F-mesure
sans l'appariement de cognats	0,85	0,80	0,82
avec l'appariement de cognats	0,88	0,85	0,86

TABLE 7 – Résultats de l'évaluation de l'alignement de mots

Certes, la taille insuffisante du corpus utilisé pour l'évaluation de notre aligneur de mots ne permet pas de mesurer quantitativement l'apport de la translittération mais les résultats obtenus indiquent clairement qu'il y a une amélioration de la qualité de l'alignement.

La non disponibilité d'un alignement de référence d'une taille significative pour les mots simples et composés ne nous permet pas de comparer notre approche avec les différents travaux de l'état de l'art. C'est la raison pour laquelle, nous avons décidé d'étudier l'apport de l'utilisation de la translittération en alignement de mots en intégrant les résultats de notre aligneur de mots dans le corpus d'apprentissage du modèle de traduction du système Moses. Le modèle de traduction utilisé est appris sur les lemmes des mots composant le corpus parallèle d'apprentissage et les lemmes des mots produits par notre aligneur (Koehn et Hoang, 2007).

Le corpus initial d'apprentissage est composé de 10000 paires de phrases français-arabe issues du corpus ARCADE II auquel nous avons ajouté environ 10000 paires de mots simples et composés correspondant aux résultats de l'aligneur de mots intégrant l'appariement de cognats à l'aide de la translittération sur 500 paires de phrases français-arabe. Nous avons aussi spécifié un modèle de langue pour la langue cible en utilisant la

totalité des phrases arabes du corpus ARCADE II.

La performance du système de traduction statistique Moses est évaluée à l'aide du score BLEU sur un corpus de test composé de 250 paires de phrases. Pour chaque phrase source, une seule phrase de référence en langue cible est considérée. Les résultats de traduction obtenus sont regroupés dans le tableau 8.

Corpus d'apprentissage	BLEU
sans les résultats de l'appariement de cognats (sans translittération)	12,50
avec les résultats de l'appariement de cognats (avec translittération)	12,82

TABLE 8 – Résultats de traduction selon le score BLEU

Ces résultats montrent que l'intégration dans le corpus d'apprentissage du modèle de traduction des alignements obtenus par le module d'appariement de cognats utilisant la translittération a permis d'obtenir un gain de $+0{,}32$ points BLEU.

Il est difficile de dire à ce stade si ce gain en score BLEU induit une amélioration significative de la qualité de la traduction au vu de la faible valeur de ce score liée à la taille des corpus d'apprentissage utilisés (uniquement 10000 paires de phrases pour l'apprentissage du modèle de traduction et environ 11000 phrases pour l'apprentissage du modèle de la langue cible). Nous pourrons conclure tout de même que la translittération améliore la performance de l'aligneur de mots, quelle que soit la manière d'évaluer les résultats, manuellement ou automatiquement.

6 Conclusion

Dans cet article, nous avons décrit un système de translittération des noms propres de l'écriture arabe vers l'écriture latine. Ce système a été utilisé dans un processus d'alignement de mots à partir de corpus de textes français-arabe. Ce processus se déroule en deux phases : d'abord, les mots simples sont alignés en utilisant un lexique bilingue et certaines propriétés de ces mots (positions, catégories grammaticales, entités nommés et cognats), ensuite les mots composés sont alignés en les identifiant à l'aide des relations de dépendance syntaxique reliant leurs composants. Ce processus donne des résultats très satisfaisants lorsque la translittération arabe est utilisée pour apparier les noms propres présents dans les phrases source et cible. Nos travaux futurs s'orientent, d'une part, vers une évaluation à une large échelle de notre outil d'alignement en vue de consolider les résultats déjà obtenus, et d'autre part, vers une translittération géo-localisée pour identifier comment les différentes translittérations peuvent fournir des indications sur l'origine et/ou sur le profil de celui qui les utilise (francophone ou anglophone, du Maghreb ou du Macherek, du nord ou du sud...).

Références

ABDULJALEEL, N. et LARKEY, L. (2003). Statistical transliteration for English-Arabic Cross Language Information Retrieval. *In Proceedings of the Twelfth ACM International Conference on Information and Knowledge Management,* New Orleans, Louisiana, pages 139–146.

ALGHAMDI, M. (2005). Alghorithms for Romanizing Arabic names. *In Journal of King Saud University: Computer Sciences and Information*, n° 17, 2005, Riyadh, pages 1–27.

ALSALMAN, A., ALGHAMDI, M., ALHUQAYL, K. et ALSUBAI, S. (2007). Romanization System for Arabic Names. *In Proceedings of The First International Symposium on Computer and Arabic Language (ISCAL – 07)*, Riyadh, pages 214–227.

BARKAT-DEFRADAS, M., HAMDI, R. et PELLEGRINO, F. (2004). De la caractérisation linguistique à l'identification automatique des dialectes arabes. *In Proceedings of MIDL 2004*.

BESANÇON, R., DE CHALENDAR, G., FERRET, O., GARA, F., LAIB, M., MESNARD, O. et SEMMAR, N. (2010). Lima: A multilingual framework for linguistic analysis and linguistic resources development and evaluation. *In Proceedings of LREC 2010*, Malta.

GUIDERE, M. (2004). Le traitement de la parole et la détection des dialectes arabes. *In Langues stratégiques et défense nationale, Publications du CREC*, Saint-Cyr, pages 53–75.

KNIGHT, K. et GRAEHL, J. (1997). Machine transliteration. *In Journal version Computational Linguistics, 24(4), 1997*, pages 599–612.

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORGAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A. et HERBST, E. (2007). Moses: Open source toolkit for statistical machine translation. *In Proceedings of ACL 2007, demo session*, Prague.

KOEHN, P. et HOANG, H. (2007). Factored Translation Models. *In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (ACL 2007),* Prague, pages 868–876.

MIHALCEA, R. et PEDERSEN, T. (2003). An evaluation exercise for word alignment. *In Proceedings of the Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, pages 10–10.

SEMMAR, N. et LAIB, M. (2010). Using a Hybrid Word Alignment Approach for Automatic Construction and Updating of Arabic to French Lexicons. *In Proceedings of LREC 2010: Workshop on Language Resources and Human Technologies for Semitic Languages,* Malta.

SHAO, L. et NG, H. T. (2004). Mining new word translations from comparable corpora. *In Proceedings of the 20th International Conference on Computational Linguistics (COLING'04),* Stroudsburg, pages 618–624.

SHERIF, T. et KONDRAK, G. (2007). Bootstrapping a stochastic transducer for Arabic-English transliteration extraction. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007),* Prague, pages 864–871.

STALLS, B. et KNIGHT, K . (1998). Translating names and technical terms in arabic text. *In Proceedings of the COLING/ACL Workshop on Computational Approches to Semitic Languages,* Montreal.

TAO, T., YOON, S. Y., FISTER, A., SPROAT, R. et ZHAI, C. (2006). Unsupervised named entity transliteration using temporal and phonetic correlation. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'06),* Sydney, pages 250–257.

VERONIS, J., HAMON, O., AYACHE, C., BELMOUHOUB, R., KRAIF, O., LAURENT, D., NGUYEN, T. M. H., SEMMAR, N., STUCK, F. et ZAGHOUANI, W. (2008). Arcade II Action de recherche concertée sur l'alignement de documents et son évaluation. *In Chapitre 2, Editions Hermès*.

YASER, A. O. et KNIGHT, K. (2002). Translating named entities using monolingual and bilingual resources. *In Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL'02),* Philadelphia, pages 400–408.

Compositionnalité et contextes issus de corpus comparables pour la traduction terminologique

Emmanuel Morin Béatrice Daille
Université de Nantes, LINA UMR CNRS 6241
2, rue de la Houssinière, BP 92208
F-44322 Nantes cedex 03
{emmanuel.morin,beatrice.daille}@univ-nantes.fr

RÉSUMÉ

Dans cet article, nous cherchons à mettre en correspondance de traduction des termes extraits de chaque partie monolingue d'un corpus comparable. Notre objectif concerne l'identification et la traduction de termes spécialisés. Pour ce faire, nous mettons en œuvre une approche compositionnelle dopée avec des informations contextuelles issues du corpus comparable. Notre évaluation montre que cette approche améliore significativement l'approche compositionnelle de base pour la traduction de termes complexes extraits de corpus comparables.

ABSTRACT

Compositionality and Context for Bilingual Lexicon Extraction from Comparable Corpora

In this article, we study the possibilities of improving the alignment of equivalent terms monolingually acquired from bilingual comparable corpora. Our overall objective is to identify and to translate highly specialised terminology. We applied a compositional approach enhanced with pre-processed context information. Our evaluation demonstrates that our alignment method outperforms the compositional approach for translationally equivalent term discovery from comparable corpora.

MOTS-CLÉS : Corpus comparable, compositionnalité, information contextuelle, lexique bilingue.

KEYWORDS: Comparable Corpora, compositionality, context information, bilingual lexicon.

1 Introduction

L'alignement lexical à partir de corpus comparables s'intéresse tout particulièrement aux domaines de spécialités, en particulier relevant de domaines scientifiques. Les domaines de spécialités sont caractérisés par des ressources textuelles réduites en comparaison à la langue générale et par une grande proportion de vocabulaire spécifique qui n'est pas présent dans les dictionnaires monolingues ou bilingues de langue générale. Un vocabulaire relevant d'un domaine de spécialité recense des termes simples, *i.e.* des mots simples, ou des termes complexes, *i.e.* des composés syntagmatiques, ces derniers étant particulièrement productifs (Sag *et al.*, 2002). Un terme est l'expression d'un concept dans un domaine de spécialité : par exemple, dans le domaine médical, *cancer* est un terme simple et *cancer du sein* un terme complexe.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 141–154,
Grenoble, 4 au 8 juin 2012.

Les corpus comparables qui rassemblent : « *des documents textuels dans des langues différentes qui ne sont pas des traductions les uns des autres* [1] » (Bowker et Pearson, 2002, p. 93) apparaissent comme une solution viable pour résoudre le manque de ressources linguistiques des domaines de spécialités : d'une part, leur statut de production monolingue garantit la qualité de leur vocabulaire spécialisé et, d'autre part, l'aspect multilingue du web garantit une disponibilité de ressources textuelles dans un grand nombre de langues et pour de nombreux domaines de spécialités. La comparabilité du corpus doit bien entendu être vérifiée à l'aide de caractéristiques communes partagées par les différentes langues lors de sa compilation (McEnery et Xiao, 2007). Pour les domaines de spécialités, le domaine et le sous-domaine sont des caractéristiques partagées obligatoires tout comme l'intention communicative et le genre textuel de manière à obtenir des traductions fiables (Bowker et Pearson, 2002). Cette comparabilité peut aussi être évaluée de manière quantitative en termes de degré de comparabilité et d'homogénéité du corpus comme dans Li et Gaussier (2010).

Pour construire des lexiques relevant de domaines de spécialités, les termes sont extraits de chaque partie monolingue du corpus comparable. Pour collecter les mêmes termes dans deux langues différentes, il est important d'utiliser un programme d'extraction terminologique adoptant la même méthode dans les deux langues. Les termes complexes constituant environ 80 % d'un lexique de domaine de spécialité, comme constaté par Nakagawa et Mori (2003) pour le japonais, il est essentiel de pouvoir les traduire.

Notre objectif est d'identifier pour un terme complexe dans une langue, sa bonne traduction au sein d'un ensemble de termes complexes candidats dans une autre langue. Une méthode triviale consiste à traduire chacun des éléments du terme complexe à l'aide d'un dictionnaire bilingue de la langue générale, à générer la combinatoire de toutes les traductions trouvées dans le dictionnaire, puis de ne conserver que les termes complexes apparaissant soit dans la liste des termes complexes candidats (Morin et Daille, 2010), soit dans le corpus comparable (Robitaille *et al.*, 2006), ou encore directement sur le web (Grefenstette, 1999). Cette méthode ne fonctionne que pour les termes complexes partageant une sémantique compositionnelle : Baldwin et Tanaka (2004) ont constaté que c'était le cas de 48,7 % des termes complexes de structure N N pour la paire de langue anglais/japonais.

Dans cet article, nous proposons d'améliorer cette méthode fondée sur une traduction compositionnelle par l'utilisation de contextes extraits d'un corpus comparable. Ces informations contextuelles seront utilisées lorsqu'un ou plusieurs éléments du terme complexe à traduire n'apparaîtront pas dans le dictionnaire bilingue. Nous démontrons que l'utilisation du contexte permet de produire un nombre important de traductions correctes pour des termes complexes ne pouvant pas être traduit par la méthode compositionnelle.

Dans la suite de cet article, nous présentons en section 2 les problèmes de traductions rencontrés avec les termes complexes. La section 3 détaille la méthode fondée sur une traduction compositionnelle pour l'obtention de traductions de termes complexes. Notre nouvelle approche associant traduction compositionnelle et contextes issus de corpus comparables est introduite en section 4. La section 5 décrit les différentes ressources textuelles et dictionnairiques utilisées pour nos expériences. La section 6 évalue l'impact de notre approche mixte sur la qualité des lexiques bilingues ainsi obtenus. La section 7 examine quelques travaux similaires à l'approche proposée avant de conclure.

1. « *sets of texts in different languages, that are not translations of each other* ».

2 Traduction des termes complexes

Si les termes complexes sont moins polysémiques (Savary et Jacquemin, 2003) et plus représentatifs (Nomura et M., 1989; Nakagawa et Mori, 2003) d'un domaine de spécialité que les termes simples, le repérage de leurs traductions pose un certain nombre de difficultés comme la fertilité, la non compositionnalité ou encore la variation terminologique [2] :

Fertilité Elle correspond à un problème de différence de longueur entre le terme complexe de la langue source et celui de la langue cible (Brown *et al.*, 1993). Par exemple, le terme complexe français *dépistage du cancer du sein* (trois mots pleins) est traduit en anglais par le terme complexe *breast screening* (deux mots pleins).

Non compositionnalité Elle s'exprime lorsqu'un terme complexe de la langue cible n'est pas typiquement composé de la traduction des parties du terme de la langue source (Melamed, 2001). Par exemple, le terme complexe français *curage axillaire* est traduit en anglais par le terme *axillary dissection* où le mot anglais *dissection* n'est pas la traduction du mot français *curage*.

Variation terminologique Cela fait référence à un terme complexe qui apparaît dans des documents sous différentes formes reflétant des différences morphologiques, syntaxiques ou sémantiques. Par exemple, les termes complexes français *cancer du sein* et *cancer mammaire* sont traduits en anglais par le même terme complexe *breast cancer*. Les termes complexes source et cible peuvent apparaître dans différentes structures syntaxiques. Ansi, le terme complexe français *prolifération tumorale* de structure N Adj est traduit en anglais par le terme complexe *tumour proliferation* de structure N N où l'adjectif français *tumorale* est lié par dérivation morphologique à la traduction française du nom anglais *tumour*. La variation terminologique peut aussi impliquer une variation paradigmatique quand un élément du terme complexe est remplacé par un synonyme ou un hyperonyme tel que *tumour size* → *diameter tumour* en anglais et non en français *taille tumorale*. Ce dernier type de variante n'est généralement pas traité par les programmes d'extraction terminologique.

Il est assez difficile de concevoir un cadre général qui puisse répondre à l'ensemble de ces problèmes (Robitaille *et al.*, 2006). Par exemple, le problème de fertilité doit probablement être réglé en premier pour éviter des alignements incomplets entre les termes complexes des langues source et cible. En ce qui concerne le problème de variation terminologique, celui-ci pourrait être en partie résolu lors de l'extraction terminologie monolingue par le regroupement de toutes les variantes morphologiques et syntaxiques du terme complexe dans les langues source et cible. Un terme complexe est ainsi vu comme un ensemble de séquences de termes reflétant une forme de base ou une variante. Ce regroupement peut être interprété comme une normalisation terminologique de la même manière que la lemmatisation au niveau morphologique.

3 Approche compositionnelle

La compositionnalité est définie comme la propriété où « *le sens du tout est fonction du sens des parties* [3] » (Keenan et Faltz, 1985, p. 24-25) : une *poêle à frire* est en effet une *poêle* destinée à *frire*.

2. Les exemples de termes français et anglais sont extraits d'un corpus comparable médical décrit en section 5.
3. « *the meaning of the whole is a function of the meaning of the parts* ».

La mise en œuvre du principe de traduction compositionnelle à partir de corpus comparables repose sur les étapes suivantes (Grefenstette, 1999; Tanaka, 2002; Robitaille *et al.*, 2006) :

Traduction du terme complexe de la langue source Pour un terme complexe de la langue source à traduire, chaque mot plein composant le terme complexe est traduit à l'aide d'un dictionnaire bilingue. Généralement, lors de cette phase de projection, la catégorie grammaticale des composants du terme complexe n'est pas utilisée. Par exemple pour le terme complexe français *examen clinique*, nous avons six traductions en anglais pour *examen* (*consideration*/N, *examen*/N, *examination*/N, *inspection*/N, *review*/N et *test*/N) et deux traductions pour *clinique* (*clinic*/N et *clinical*/Adj).

Génération des traductions candidates Toutes les combinaisons sont générées sans tenir compte de l'ordre des mots avec un total de $O(n! \prod_{i=1}^{p} t_i)$ combinaisons possibles (où t_i est le nombre de traductions du mot plein i et n le nombre total de mots pleins). 24 combinaisons sont obtenues avec le précédent exemple.

Sélection des traductions candidates À partir de l'ensemble des traductions candidates, les traductions les plus probables sont ordonnées en fonction de leur fréquence d'apparition dans la langue cible. Pour l'exemple précédent, les traductions candidates sont les termes complexes de la langue cible identifiés par le système d'extraction de terminologie.

Le principe de traduction compositionnelle est restrictif. Pour palier cette difficulté, Robitaille et al. (2006) proposent d'utiliser une méthode de repli : s'il n'y a pas suffisamment de données dans le dictionnaire bilingue pour traduire un terme de longueur n (avec $n > 2$ mots pleins) alors ce terme sera décomposé en toutes les combinaisons de termes de longueur inférieure ou égale à n. Cette approche permet de pouvoir traduire directement une sous partie du terme complexe s'il est présent dans le dictionnaire bilingue. Par exemple, pour le terme complexe français *technique du ganglion sentinelle* quatre combinaisons seraient générées : (i) [technique du ganglion sentinelle], (ii) [technique du ganglion] [sentinelle], (iii) [technique] [ganglion sentinelle] et (iv) [technique] [ganglion] [sentinelle]. Morin et Daille (2010) ont proposé quant à eux une méthode compositionnelle étendue qui comble le fossé entre termes complexes de différentes structures syntaxiques en exploitant des liens morphologiques. En s'appuyant sur une liste de règles morphologiques de codage/décodage associée à un système d'extraction de terminologie, leur méthode est plus efficace que la méthode compositionnelle de base. Pour 859 termes complexes français de structure N Adjr (où Adjr est un adjectif relationnel), ils retrouvent dans un dictionnaire bilingue français/japonais 30 termes complexes et traduisent 8 termes complexes avec une précision de 62 % avec la méthode compositionnelle de base et 128 termes complexes avec une précision de 88 % avec leur approche compositionnelle étendue.

L'approche compositionnelle est aussi appelée « approche par sacs de mots équivalents [4] » par Vintar (2010) lorsque le dictionnaire bilingue est construit à partir d'un corpus parallèle et qu'il contient tous les mots qui apparaissent dans le corpus avec leurs équivalences de traduction accompagnées d'un score de probabilité. Le nombre de traductions généré peut être réduit en utilisant des structures syntaxiques de traductions entre les termes des langues source et cible. Par exemple, Tanaka et Baldwin (2003) utilisent les structures suivantes pour filtrer les candidats de traduction : un terme complexe japonais de structure N_1 N_2 est traduit en anglais par un terme complexe de structure N_1 N_2 (dans 33,2 % des cas), par Adj_1 N_2 (28,4 %) ou encore par N_2 of (the) N_1 (4,4 %).

4. « *bag-of-equivalents approach* ».

4 Approche compositionnelle enrichie par des informations contextuelles

L'approche compositionnelle de base qui propose des traductions pour des termes complexes est facile à mettre en œuvre, mais elle échoue lorsque :
– les termes complexes ne partagent de propriété compositionnelle, *i.e.* 50 % des situations (Baldwin et Tanaka, 2004) ;
– l'un des composants du terme complexe ne peut être traduit directement par un dictionnaire bilingue ;
– les combinaisons proposées de termes candidats ne sont pas présentes dans la liste des termes complexes extraits de la langue cible ou plus généralement dans la langue cible du corpus comparable.

Pour palier cette difficulté, une première solution serait de trouver des synonymes dans la langue source. Pour les mots de basse fréquence, Pekar et al. (2006) prédisent des valeurs de cooccurrences absentes en s'appuyant sur des mots similaires dans la même langue. Pour les traductions jugées plus difficiles, Sharoff et al. (2009) identifient des mots similaires dans la langue source pour produire une similarité plus fiable. Dans notre cas, nous avons déjà réalisé un regroupement de synonymes en exploitant un ensemble de variantes du terme au lieu d'un terme unique (voir la section 5).

L'approche proposée pour identifier les traductions d'un terme complexe à partir d'un corpus comparable s'appuie sur l'exploitation du contexte des composants du terme à traduire lorsque l'approche compositionnelle de base échoue. Nous référons ici aux deux parties monolingues du corpus comparables comme le corpus source et cible. Ce modèle se décompose en quatre étapes :

Calcul du contexte des termes complexes Pour un terme complexe à traduire défini par $C_{s1}C_{s2}\cdots C_{sk}$ (où k est le nombre de mots pleins du terme), nous recherchons chaque composant C_{si} dans le dictionnaire bilingue. Ici, chaque composant non traduit par le dictionnaire est remplacé par des informations de cooccurrence. Plus précisément, nous calculons les mots qui cooccurrent avec C_{si} dans une fenêtre de w mots autour de C_{si} dans le corpus source. L'information mutuelle comme le rapport de vraisemblance sont deux bonnes mesures pour déterminer la relation de cooccurrence entre deux mots. Ces informations de cooccurrence, normalisées avec l'une des précédentes mesures, sont représentées sous la forme d'un vecteur de contexte (V_{si}). À titre d'exemple, considérons le terme complexe français *antécédent familial* ($C_{s1}C_{s2}$). Si le premier composant *antécédent* (C_{s1}) n'est pas présent dans le dictionnaire bilingue alors ce composant est remplacé par son vecteur de contexte (V_{s1}) (voir la figure 1).

Transfert des termes complexes Pour chaque vecteur de contexte V_{si}, ses éléments sont projetés dans la langue cible en utilisant le dictionnaire bilingue et le vecteur de contexte transféré devient V'_{si}. Si le dictionnaire bilingue propose plusieurs traductions pour un élément, elles sont toutes utilisées, mais chaque traduction est pondérée en fonction de la fréquence de l'élément dans la langue cible. Si un élément n'est pas trouvé dans le dictionnaire bilingue, il est alors écarté. En revanche, quand le composant C_{si} est présent dans le dictionnaire, nous calculons les informations de cooccurrence de chaque traduction dans le corpus cible et les sauvegardons dans un vecteur de contexte, V'_{si}. Par exemple, si nous avons trouvé deux traductions anglaises pour le composant *familial* (C_{s2}) telles que *fami-*

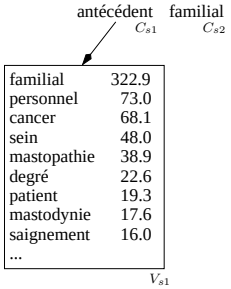


FIGURE 1 – Calcul du contexte d'un terme complexe

lial et *family*, nous retenons alors deux vecteurs de contexte $V_{s2_1}^{'}$ et $V_{s2_2}^{'}$ dans le corpus cible (voir la figure 2).

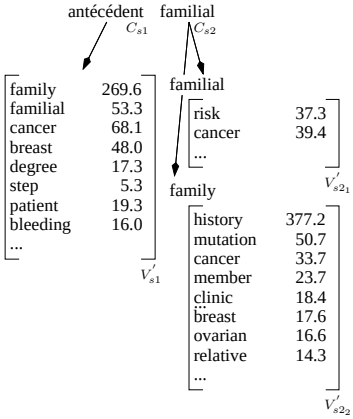


FIGURE 2 – Projection dans la langue cible

Génération des traductions candidates Chaque terme complexe de la langue cible, pour lequel chaque composant C_{ti} est décrit par son vecteur de contexte V_{ti}, est ensuite comparé au vecteur de contexte transféré à travers une mesure de distance vectorielle comme le Cosinus ou le Jaccard pondéré. Pour un terme complexe en langue cible composé de deux vecteurs de contexte V_{t1} et V_{t2} et un terme complexe transféré composé de deux vecteurs de contexte $V_{s1}^{'}$ et $V_{s2}^{'}$, deux paires de scores de similarité correspondantes aux différentes combinaisons possibles seraient calculées : $sim(V_{t1}, V_{s1}^{'})$ avec $sim(V_{t2}, V_{s2}^{'})$ et $sim(V_{t1}, V_{s2}^{'})$ avec $sim(V_{t2}, V_{s1}^{'})$. Le score terminal pour chaque paire est quant à lui défini comme la moyenne géométrique de chaque score de similarité : $\sqrt{sim(V_{t1}, V_{s1}^{'}).sim(V_{t2}, V_{s2}^{'})}$ et

$\sqrt{sim(V_{t1}, V_{s2}^{'}).sim(V_{t2}, V_{s1}^{'})}$ (voir la figure 3).

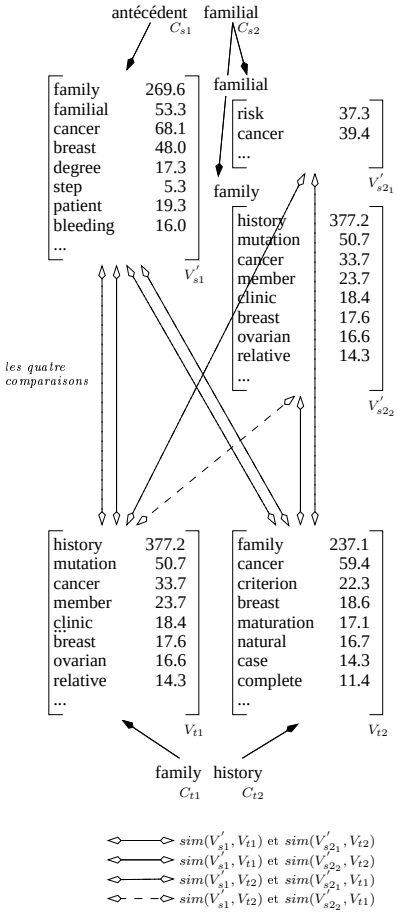


FIGURE 3 – Comparaison entre un terme complexe à traduire et un terme complexe de la langue cible (une paire de flèches correspond à une comparaison entre les composants du terme source et du terme cible - par exemple la paire de flèches pleines correspond à la comparaison entre $V_{s1}^{'}$ et V_{t1} et entre $V_{s2_2}^{'}$ et V_{t2})

Sélection des traductions candidates Les traductions candidates sont finalement ordonnées en fonction du score d'association (voir la figure 4).

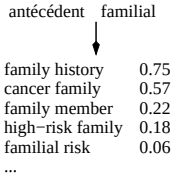


FIGURE 4 – Liste ordonnée des traductions candidates

5 Ressources

Dans cette section, nous décrivons les différentes ressources utilisées pour nos expériences, à savoir le corpus comparable, le dictionnaire bilingue et la liste de référence.

5.1 Corpus comparable

Le corpus comparable spécialisé français/anglais utilisé dans cette étude relève du domaine médical et plus précisément du sous-domaine du cancer du sein. Les documents composant ce corpus ont été sélectionnés automatiquement à partir du site d'Elsevier [5] en sélectionnant les articles scientifiques publiés sur la période 2001 et 2008 pour lesquels le titre ou les mots-clés des articles contiennent les termes « *cancer du sein* » en français et « *breast cancer* » en anglais. La compilation de ce corpus comparable remplit les exigences d'un corpus comparable de spécialité en termes de domaine, sous-domaine, de paramètres de communication (experts-à-experts) et de genre textuel qui sont des caractéristiques communes à travers les langues. Nous avons ainsi automatiquement collecté 130 documents pour le français et 118 pour l'anglais ce qui représente environ 530 000 mots par langue. L'ensemble des documents ont été nettoyés et normalisés à travers les traitements suivants : segmentation en occurrences de formes, étiquetage morpho-syntaxique [6], lemmatisation [7] et extraction terminologique [8]. Enfin, les mots outils et les hapax ont été supprimés dans les parties française et anglaise. Le corpus comparable fournit finalement 4 000 mots simples et 5 100 mots composés en français et 4 000 mots simples et 4 100 mots composés en anglais.

5.2 Dictionnaire bilingue

Le dictionnaire français-anglais nécessaire à l'étape de transfert a été construit à partir de différentes ressources disponibles sur le web. Il comporte, après normalisation, 22 300 mots pour le français avec en moyenne 1,6 traductions par entrée. Il s'agit d'un dictionnaire de langue générale qui ne contient que peu de termes en rapport avec le domaine médical.

5. `http://www.elsevier.com`
6. Pour le français et l'anglais, nous avons utilisé l'étiqueteur de Brill (Brill, 1994).
7. Pour le français, nous avons utilisé le lemmatiseur FLEMM : `http://www.univ-nancy2.fr/pers/namer/` et pour l'anglais un lemmatiseur construit à partir de la base CELEX.
8. Nous avons choisi d'utiliser ACABIT (Daille, 2003), un outil ouvert qui permet de traiter des corpus volumineux et dont la conception est fondamentalement multilingue, avec des implémentations pour le français et l'anglais.

5.3 Liste de référence

La liste de référence contient une liste de termes complexes extraits automatiquement dans l'une des parties monolingues du corpus comparable. Cette liste est utilisée pour comparer la couverture et la précision des trois différentes approches : la simple projection dans un dictionnaire bilingue, l'approche compositionnelle et l'approche compositionnelle enrichie par des informations contextuelles. Les unités terminologiques qui sont extraites par ACABIT sont des termes complexes dont les structures syntaxiques correspondent soit à un structure canonique soit à structure de variation. Les structures sont exprimées en utilisant des étiquettes syntaxiques [9]. Pour le français les principales structures sont N N, N Prep N et N Adj et pour l'anglais N N, Adj N et N Prep N. Les variantes prises en compte sont morphologiques et syntaxiques pour les deux langues. ACABIT considère comme une variante morphologique la modification morphologique de l'un des composants de la forme de base, comme une variante syntaxique l'insertion d'un autre mot dans les composants de la forme de base. Par exemple, dans la partie française du corpus comparable le terme candidat *cancer du sein* apparaît sous les formes suivantes :

- **forme de base** de structure N Prep N : *cancer du sein* ;
- **variante flexionnelle** : *cancers du sein* ;
- **variante syntaxique** (par insertion d'un modifieur dans la forme de base) : *cancer primitif du sein* ;
- **variante syntaxique** (par expansion par coordination de la forme de base) : *cancer des ovaires et du sein*.

Pour la liste de référence, nous avons sélectionné les termes complexes français extraits par ACABIT ayant un nombre d'occurrences supérieur ou égal à 5. Cette liste de référence est composée de 976 termes complexes français. Dans cette liste, 90 % des candidats termes fournis par le processus d'extraction terminologique après regroupement sont composés de deux mots pleins.

6 Expériences et résultats

Dans cette section, nous évaluons les performances des différentes approches en fonction de la qualité des traductions obtenues.

6.1 Projection du dictionnaire et approche compositionnelle

Dans un premier temps, nous commençons par compter le nombre de termes qui peuvent être directement traduits par le dictionnaire bilingue. Nous évaluons ensuite la qualité des traductions fournies par l'approche compositionnelle. La table 1 présente les résultats obtenus pour la traduction du français vers l'anglais. La première colonne indique le nombre de termes complexes français qui sont traduits. Puisque l'approche compositionnelle, comme la traduction dictionnairique, peut donner plusieurs traductions pour un terme à traduire, la colonne suivante indique le nombre de termes complexes français pour lesquels une ou plusieurs traductions sont obtenues en anglais. La troisième colonne indique quant à elle le nombre de bonnes traductions en anglais. Enfin, les deux dernières colonnes indiquent la précision aux rangs 1 (Top_1) et 5 (Top_5)

9. Les symboles sont Adj (Adjectif), N (Nom), Prep (Préposition).

pour chaque stratégie. La précision au rang n (Top_n) représente le nombre de traductions correctes trouvées dans la liste ordonnée des n premières traductions candidates. Ici, les traductions candidates sont ordonnées selon leur fréquence d'apparition dans la partie anglaise du corpus comparable. Les résultats de cette première expérience montrent que sur les 976 termes complexes de la liste de référence, 51 termes complexes sont présents dans le dictionnaire et 140 termes complexes sont traduits au moyen de l'approche compositionnelle avec une précision de 79,1 % pour le Top_5 (les termes complexes présents dans le dictionnaire ne sont pas utilisés par l'approche compositionnelle). Ici, nous sommes incapables de proposer une traduction pour 785 termes complexes de la liste de référence.

	# termes français	# termes anglais	# traductions correctes	Top_1	Top_5
projection du dictionnaire	51	69	69	100 %	100 %
approche compositionnelle	140	172	136	73,2 %	79,1 %

TABLE 1 – Projection du dictionnaire et approche compositionnelle

6.2 Approche compositionnelle enrichie par des informations contextuelles

Nous appliquons maintenant l'approche compositionnelle enrichie par des informations contextuelles sur les 785 termes non traduits de la liste de référence. Dans cette expérience, les paramètres utilisés sont les suivants : la taille de la fenêtre contextuelle w est fixée à 3 (c'est-à-dire une fenêtre de sept mots), la mesure d'association est l'information mutuelle et la mesure de distance vectorielle est le Cosinus. D'autres combinaisons de paramètres ont été évaluées, mais les précédents paramètres sont ceux qui donnent les meilleurs résultats. La table 2 présente le pourcentage de termes français pour lesquels la bonne traduction est obtenue parmi les $Top_{1, 5, 10, \text{ et } 20}$ traductions candidates pour une traduction du français vers l'anglais. En partant des 785 termes complexes non traduits, nous traduisons 514 termes complexes français par la méthode compositionnelle enrichie avec une précision de 33,6 % pour le Top_1 et 51,6 % pour le Top_{20}. Ces résultats indiquent que la majorité des termes complexes correctement traduits sont en fait obtenus pour le Top_5.

# traductions	Top_1	Top_5	Top_{10}	Top_{20}
514	33,6 %	48,9 %	50,7 %	51,6 %

TABLE 2 – Précision des traductions pour la méthode compositionnelle enrichie par des informations contextuelles

En ce qui concerne les termes complexes correctement traduits, nous trouvons une grande majorité de termes complexes français impliquant un adjectif relationnel. Par exemple, le terme complexe français *dépistage mammographique* n'est pas traduit par l'approche compositionnelle

de base puisque l'adjectif relationnel français *mammographique* n'est pas trouvé dans le dictionnaire bilingue. En revanche, la traduction attendue *mammographic screening* est trouvée pour le Top_3 avec l'approche basée sur le contexte dans la mesure où nous avons associé le vecteur de contexte français de *mammographique* avec le vecteur de contexte anglais de *mammographic* et la paire français/anglais *dépistage/screening* est bien présente dans le dictionnaire. Les autres termes complexes correctement traduits sont principalement des termes avec une structure compositionnelle pour lesquels un élément n'est pas trouvé dans le dictionnaire comme : *amélioration significative/significant benefit* (Top_1) et *analyse multivariée/multivariate analysis* (Top_4) ou sans structure compositionnelle comme *curage axillaire/axillary dissection* (Top_{11}). Pour ce qui est des termes complexes mal traduits, nous identifions principalement deux situations. D'une part, nous trouvons comme traductions candidates des termes sémantiquement proches du terme à traduire comme *retrospective study* pour *étude comparative*. D'autre part, nous ne proposons parfois qu'une sous-partie du terme complexe anglais tel que *node dissection* pour *curage ganglionnaire* (*lymph node dissection*). Cette dernière situation nécessite la prise en compte de la fertilité pour pouvoir être traitée.

7 Travaux connexes

La plupart des travaux d'alignement lexical à partir de corpus comparables qui s'appuient sur le contexte traitent uniquement les mots simples (Fung, 1998; Rapp, 1999; Chiao et Zweigenbaum, 2002; Gaussier *et al.*, 2004; Laroche et Langlais, 2010, parmi d'autres).

Les travaux portant sur la traduction des termes complexes adoptent plutôt l'approche compositionnelle simple comme celle présentée en section 3 ou améliorée par des propriétés morphologiques ou du repli (Grefenstette, 1999; Tanaka, 2002; Robitaille *et al.*, 2006; Vintar, 2010).

Les cooccurrents sont au cœur de toutes les tâches de désambiguïsation contextuelle. Pour la traduction automatique statistique, Koehn et Knight (2002) construisent un lexique bilingue à partir de corpus comparables composé de mots possédant une graphie proche dans les deux langues. L'utilisation de cooccurrents permet d'améliorer la précision de ce lexique bilingue de 15 %. Une approche similaire est utilisée par Haghighi et al. (2008) qui se contentent de réduire l'espace de recherche et par Ismail et Manandhar (2010) qui s'appuient sur des cooccurrents spécifiques au domaine.

Munteanu et Marcu (2006) extraient des segments de texte parallèles à partir de corpus comparables. À l'aide d'un corpus parallèle, ils extraient pour chaque mot du texte source ses traductions candidates et obtiennent ainsi un premier dictionnaire bilingue où chaque traduction candidate est associée à un poids. Un segment de texte parallèle rencontré dans un corpus comparable sera une suite de mots apparaissant dans un texte source pour laquelle chacun des mots a une traduction dans le lexique pondéré. Cette méthode correspond à la méthode compositionnelle de base où est substitué à un dictionnaire bilingue un dictionnaire construit à partir d'un corpus aligné. Cette méthode diffère de plus sur le type et la taille du corpus utilisé ainsi que sur la nature des segments textuels qui ne correspondent pas forcément à une entrée lexicale.

Enfin, Shezaf et Rappoport (2010) filtrent à l'aide de cooccurrents les entrées d'un dictionnaire bilingue bruité construit par pivot à l'aide d'une lingua franca. L'utilisation de ces cooccurrents recueillis sur un corpus comparable augmente la qualité du dictionnaire bilingue de 20 %.

8 Conclusion

Dans cet article, nous avons proposé une méthode mixte pour aider à la construction de terminologies bilingues à partir de corpus comparables. Nous avons montré que la méthode compositionnelle utilisée par les programmes d'alignement de termes complexes pouvait être largement améliorée à l'aide de cooccurrents collectés à partir de corpus comparables. Cette méthode permet de traduire de nombreux termes complexes dont les traductions ne pouvaient pas être trouvées par la seule méthode compositionnelle. Prochainement, nous souhaitons généraliser cette méthode pour prendre en compte d'autres types de termes complexes qui existent dans d'autres langues tels que les composés morphologiques allemands ou chinois. Nous étudierons comment éviter de générer des traductions incomplètes et comment résoudre ainsi le problème de fertilité pour les termes complexes. Enfin, nous envisageons de modifier le protocole d'évaluation en acceptant plusieurs traductions possibles dans le cas de synonymes ou de traductions proches qui n'ont pas été détectés lors de l'extraction terminologique.

Remerciements

Ce travail a bénéficié de l'aide du septième programme cadre de la Commission européenne (FP7/2007-2013) (Grant Agreement no 248005).

Références

BALDWIN, T. et TANAKA, T. (2004). Translation by Machine of Complex Nominals : Getting it Right. *In Proceedings of the ACL 2004 Workshop on Multiword Expressions : Integrating Processing*, pages 24–31, Barcelona, Spain.

BOWKER, L. et PEARSON, J. (2002). *Working with Specialized Language : A Practical Guide to Using Corpora*. Routledge, London/New York.

BRILL, E. (1994). Some Advances in Transformation-Based Part of Speech Tagging. *In Proceedings of the 12th National Conference on Artificial Intelligence (AAAI'94)*, pages 722–727, Seattle, WA, USA.

BROWN, P., DELLA PIETRA, S., DELLA PIETRA, V. et MERCER, R. (1993). The Mathematics of Statistical Machine Translation : Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

CHIAO, Y.-C. et ZWEIGENBAUM, P. (2002). Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora. *In Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1208–1212.

DAILLE, B. (2003). Terminology Mining. *In* PAZIENZA, M. T., éditeur : *Information Extraction in the Web Era*, pages 29–44. Springer.

FUNG, P. (1998). A Statistical View on Bilingual Lexicon Extraction : From Parallel Corpora to Non-parallel Corpora. *In* FARWELL, D., GERBER, L. et HOVY, E., éditeurs : *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, pages 1–16, Langhorne, PA, USA.

GAUSSIER, E., RENDERS, J.-M., MATVEEVA, I., GOUTTE, C. et DÉJEAN, H. (2004). A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora. *In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL'04)*, pages 526–533, Barcelona, Spain.

GREFENSTETTE, G. (1999). The World Wide Web as a Resource for Example-Based Machine Translation Tasks. *In ASLIB'99 Translating and the Computer 21*, London, UK.

HAGHIGHI, A., LIANG, P., BERG-KIRKPATRICK, T. et KLEIN, D. (2008). Learning Bilingual Lexicons from Monolingual Corpora. *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, pages 771–779, Columbus, Ohio, USA.

ISMAIL, A. et MANANDHAR, S. (2010). Bilingual lexicon extraction from comparable corpora using in-domain terms. *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 481–489, Beijing, China.

KEENAN, E. L. et FALTZ, L. M. (1985). *Boolean Semantics for Natural Language*. D. Reidel, Dordrecht, Holland.

KOEHN, P. et KNIGHT, K. (2002). Learning a translation lexicon from monolingual corpora. *In Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, pages 9–16, Philadelphia, Pennsylvania, USA.

LAROCHE, A. et LANGLAIS, P. (2010). Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 617–625, Beijing, China.

LI, B. et GAUSSIER, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. *In Proceedings of the 23rd International Conference on Computational Linguistics*, COLING'10, pages 644–652, Beijing, China.

MCENERY, A. et XIAO, Z. (2007). Parallel and comparable corpora : What is happening ? *In* ANDERMAN, G. et ROGERS, M., éditeurs : *Incorporating Corpora : The Linguist and the Translator*, Multilingual Matters. Clevedon.

MELAMED, I. D. (2001). *Empirical Methods for Exploiting Parallel Texts*. MIT Press, Cambridge, MA, USA.

MORIN, E. et DAILLE, B. (2010). Compositionality and Lexical Alignment of Multi-word terms. *In Language Resources and Evaluation*, volume 44, pages 79–95. Springer.

MUNTEANU, D. S. et MARCU, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'2006)*, Sydney, Australia.

NAKAGAWA, H. et MORI, T. (2003). Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2):201–219.

NOMURA, M. et M., I. (1989). *Gakujutu Yogo Goki-Hyo*. National Language Research Institute, Tokyo.

PEKAR, V., MITKOV, R., BLAGOEV, D. et MULLONI, A. (2006). Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20(4):247–266.

RAPP, R. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 519–526, College Park, MD, USA.

ROBITAILLE, X., SASAKI, X., TONOIKE, M., SATO, S. et UTSURO, S. (2006). Compiling French-Japanese Terminologies from the Web. *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 225–232, Trento, Italy.

SAG, I. A., BALDWIN, T., BOND, F., COPESTAKE, A. A. et FLICKINGER, D. (2002). Multiword Expressions : A Pain in the Neck for NLP. *In Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'02)*, pages 1–15, Mexico City, Mexico.

SAVARY, A. et JACQUEMIN, C. (2003). Reducing Information Variation in Text. *In* GREFENSTETTE, G., éditeur : *Text- and Speech-Triggered Information Access*, Lecture Notes in Computer Science, pages 141–181. Springer Verlag.

SHAROFF, S., BABYCH, B. et HARTLEY, A. (2009). 'Irrefragable answers' using comparable corpora to retrieve translation equivalents. *Language Resources and Evaluation*, 43(1):15–25.

SHEZAF, D. et RAPPOPORT, A. (2010). Bilingual lexicon generation using non-aligned signatures. *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, pages 98–107, Uppsala, Sweden.

TANAKA, T. (2002). Measuring the Similarity between Compound Nouns in Different Languages Using Non-parallel Corpora. *In Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1–7, Taipei, Taiwan.

TANAKA, T. et BALDWIN, T. (2003). Noun-Noun Compound Machine Translation : A Feasibility Study on Shallow Processing. *In Proceedings of the ACL 2003 workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, pages 17–24, Sapporo, Japan.

VINTAR, S. (2010). Bilingual term recognition revisited : The bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16(2):141–158.

Raffinement du Lexique des Verbes Français

Paul Bédaride
Université de Stuttgart
paul.bedaride@gmail.com

RÉSUMÉ

Nous présentons dans cet article les améliorations apportées à la ressource « Les Verbes Français » afin de la rendre plus formelle et utilisable pour le traitement automatique des langues naturelles. Les informations syntaxiques et sémantiques ont été corrigées, restructurées, unifiées puis intégrées à la version XML de cette ressource, afin de pouvoir être utilisée par un système d'étiquetage de rôles sémantiques.

ABSTRACT

Resource Refining : « Les Verbes Français »

This paper introduce the impovements we made to the resource « Les Verbes Français » in order to make it more usable in the field of natural language processing. Syntactic and semantic information is corrected, restructured, unified and then integrated to the XML version of this resource, in order to be used by a semantic role labelling system.

MOTS-CLÉS : ressource, lexique, verbes, raffinement, étiquetage de rôles sémantiques.

KEYWORDS: resource, lexicon, verbs, refinement, semantic roles labeling.

1 Introduction

Le domaine du traitement automatique des langues naturelles nécessite à la fois des ressources représentant les particularités des langues et de leurs sémantiques, ainsi que des théories d'analyse utilisant ces ressources. Si les théories sont la plupart du temps rapidement adaptables d'une langue à l'autre, il n'en est pas de même pour les ressources. En effet, les ressources devant représenter la richesse d'une langue, il est nécessaire de réaliser une nouvelle analyse des cas problématiques ou une nouvelle annotation de corpus. Si l'on peut dire que l'anglais est une langue riche en ressources linguistiques (e.g. : PropBank, FrameNet, WordNet, ...), il n'en est pas de même pour le français. Il existe des équivalents pour certains types de ressources (e.g. : FrenchTreeBank, EuroWordNet, Wolf, ...), mais ils ont souvent une moins grande couverture et une moins bonne qualité. Si les ressources linguistiques ne sont pas très développées, c'est parce qu'elles exigent des analyses d'experts du domaine et des annotations nécessitant une quantité considérable de temps et de travail, qui engendre un fort coût de production. D'un autre coté, il existe des ressources linguistiques méconnues et sous-utilisées car nécessitant des efforts conséquents pour être adaptées au traitement automatique des langues. Plutôt que de laisser stagner ces ressources et de créer de nouvelles ressources à partir de rien, nous avons choisi d'améliorer l'une d'entre elles – « Les verbes français » (Dubois et Dubois-Charlier, 1997),

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 155–168,
Grenoble, 4 au 8 juin 2012.

avec pour objectif de la rendre utilisable pour la tâche d'étiquetage de rôles sémantiques. Nous allons maintenant introduire cette ressource en évoquant ses qualités et faiblesses, ainsi que les améliorations déjà réalisées, puis nous expliquerons pourquoi cette ressource nous semble être un bon choix pour l'étiquetage de rôles sémantiques. Dans les sections suivantes nous présenterons notre objectif de ressource, les traitements réalisés pour atteindre cet objectif, ainsi qu'une évaluation de ces améliorations. Nous conclurons cet article en résumant les gains que nos travaux ont apportés à cette ressource.

2 Les Verbes Français

2.1 Historique

Le *LVF*, « *Les Verbes Français* », est une ressource lexicale réalisée par Jean Dubois et Françoise Dubois-Charlier, dont l'objectif est de fournir une description linguistique des verbes, basée sur l'adéquation entre schèmes syntaxiques et interprétation sémantique (Levin, 1993). Cette ressource a été confiée dans un premier temps à un industriel qui n'a pas su l'utiliser, ainsi qu'à un éditeur qui ne l'a pas publié (elle fut distribuée sous la forme de photocopies). Comme ces deux entreprises détenaient les droits d'utilisation, cette ressource ne pouvait pas être diffusée afin d'être largement utilisée et améliorée. Ils ont cependant accepté après un certain temps, de restituer aux auteurs le droit de diffuser leur ressource comme ils le souhaitaient.

Depuis sa libération en 2007 sous la forme d'un fichier Excel (*eLVF*), un nombre croissant de personnes se sont intéressées à cette ressource. Un des premiers constats réalisé fut que Jean Dubois et Françoise Dubois-Charlier ont développé le *eLVF* sans se soucier des problèmes que les informaticiens pourraient avoir pour utiliser leur ressource dans le traitement automatique de la langue française. Une grande partie des problèmes est cependant due aux limites d'espace de stockage existantes lors de la création du *LVF*. Un grand nombre de mots ont ainsi été tronqués ou abrégés. La représentation de la ressource sous la forme d'un tableau limite sa structuration, et les formats utilisés pour certains champs ne sont pas assez formels. Enfin, l'ouvrage « Les verbes français » est obligatoire pour comprendre tous les codes et formats utilisés dans la version Excel du *LVF*(*eLVF*).

2.2 Description

Le *eLVF* est composé de 25 609 entrées représentant les différents sens de 12 310 verbes. Il y a 4 188 verbes à plusieurs entrées et un verbe peut avoir jusqu'à 61 entrées (e.g. :pour le verbe *« passer »*). Une entrée est composée des onze champs suivants :

- MOT : entrée du verbe à l'infinitif
- DOMAINE : code donnant l'emploi principal (géologie, psychologie, . . .)
  et le niveau de langue (familier, vieux, littéraire, . . .)
- CLASSE : code définissant la classe syntactico-sémantique (appartenant à une hiérarchie)
- OPÉRATEUR : définition syntactico-semantique de l'entrée
- SENS : synonymes et définitions abrégées
- PHRASE : exemples d'utilisation de ce sens

M	DOM	CLA	OPER	SENS	PHRASE	C	CONST	DER	N	L
amasser 01	OBJ	L3b	*lc.qp qc+pl e amas*	accumuler	On a~des documents,des livres. Les preuves s'a~contre P.	1aZ	T1801 P8001	-1- -D ---- --	3*	2
amasser 02	MON	L4b	*lc.qp arg e tas*	accumuler	On a de l'argent,de l'or.. On a~sans cesse.L'argent s'a~	1aZ	T1301 P3001	--- -- ---- --	-	5
amasser 03(s)	LOC	U1a	*(qn+pl)li.simul qp*	se grouper	La foule s'a~sur la place.	1aZ	P7001	--- -- ---- --	-	5

TABLE 1 – Entrées du *eLVF* pour le verbe *« amasser »*

- CONJUGAISON : codes permettant de conjuguer le verbe
- CONSTRUCTIONS : codes pour obtenir les schèmes de construction syntaxique
- DÉRIVATIONS : codes pour produire les adjectifs verbaux et les dérivés nominaux
- NOM : code pour produire le mot dont le verbe est dérivé
- LEXIQUE : code pour obtenir le type de dictionnaire où l'entrée est répertoriée

La table 1 recense les entrées du *eLVF* représentant les différents sens du verbe *« amasser »* (i.e. : « accumuler des objets », « accumuler de l'argent », « se grouper quelque part »). Le *eLVF* donne les informations suivantes sur le premier sens de ce verbe : il est utilisé dans le DOMAINE des objets ; sa CLASSE sémantique est *mettre quelque chose quelque part, dans/sur un lieu, autour de quelque chose* ; son OPÉRATEUR signifie *être ou mettre quelque part plusieurs choses en amas* ; il a pour synonyme le verbe *« accumuler »*, il est réalisé dans les PHRASES d'exemple *« On amasse des documents. »* et *« Les preuves s'amassent contre Pierre. »* ; il fait partie des verbes du premier groupe avec un auxiliaire *« avoir »* pour le transitif et *« être »* pour le pronominal ; il se réalise dans un cadre transitif avec un sujet humain, un objet choses, et un circonstant locatif (où on est) ainsi que dans un cadre pronominal avec sujet choses et un circonstant locatif ; l'adjectif *« amassable »* et le déverbal *« amas »* en sont des dérivés ; et il provient d'un dictionnaire de base de 15 000 mots. Une bonne connaissance des encodages est clairement nécessaire pour dériver ces informations à partir de l'entrée du *eLVF*. Nous n'allons pas décrire plus précisément tous les champs[1], car cela serait trop long, mais nous allons nous focaliser sur les champs les plus utiles pour la tâche d'étiquetage de rôle sémantiques : les champs opérateur et constructions.

Le champ OPÉRATEUR interprète sémantiquement les schèmes syntaxiques. Il est composé d'un prédicat (le premier token qui n'est pas entre parenthèses), et d'un certain nombre d'arguments. Les définitions des prédicats et de certaines abréviations sont accessibles dans la version papier du *LVF*. Le sujet du prédicat, qui est optionnel, est défini entre parenthèses juste avant le prédicat. Les autres arguments suivent le prédicat et peuvent être formés d'un ou plusieurs mots. Les limites des arguments n'étant pas définies, c'est à l'utilisateur d'identifier les différents arguments à l'aide de ses connaissances de la langue et de ses capacités de raisonnement. Il existe deux types d'arguments : les contraintes syntaxico-sémantiques pouvant être réalisés syntaxiquement et les spécifications sémantiques précisant la sémantique du prédicat. Un argument est une contrainte syntaxico-sémantique s'il est le sujet du prédicat, s'il appartient à un certain ensemble d'abréviations (e.g. : *qc*, *qn*, ...) ou s'il est composé d'une préposition en capitales ; et est une spécification sémantique dans le reste des cas. Par exemple, dans la première entrée du verbe *« amasser »*, *qc+pl* représente une contrainte alors que *e amas* représente une spécification. Des opérateurs de disjonction (i.e. : */*,) permettent d'associer plusieurs contraintes à un même emplacement sémantique.

Le champ CONSTRUCTIONS liste les différents schémas syntaxiques pouvant être réalisés. Un schéma syntaxique commence par une lettre en capitale, qui définit son type et celui de ses

1. Pour une description complète de la ressource : `http://margaux.philosophie.uni-stuttgart.de/lvf/`

A	intransitif	sujet + circonstant
N	transitif indirect	sujet + complément prépositionnel
T	transitif direct	sujet + obj. direct + cpl. prép. + circonstant
P	pronominal	sujet + obj. direct + cpl. prép. + circonstant

TABLE 2 – Type de CONSTRUCTION avec arguments associés

arguments. La table 2 donne les quatre types de schèmes existant dans le *LVF* avec le type de leurs arguments. Chaque argument encode une contrainte syntaxique ou sémantique par un chiffre ou une lettre. Un sujet ou un objet ayant pour valeur *1* représente une contrainte sémantique sur le domaine de l'*humain*, un circonstant avec la valeur *1* représente une contrainte sémantique sur le domaine *locatif (où on est)* et un complément prépositionnel ayant pour valeur *b* représente une contrainte syntaxique sur un complément prépositionnel avec la préposition *« de »*. Nous donnerons uniquement la signification des codes que nous utiliserons dans cet article, et nous vous renvoyons à la version papier de *LVF* ou à notre wiki [1] si vous souhaitez étudier les autres codes.

L'espace de stockage étant nettement moins problématique de nos jours, il serait intéressant de décoder la ressource pour la rendre plus lisible et utilisable. C'est ce qu'ont commencé à faire Hadouche et Lapalme (Hadouche et Lapalme, 2010) en transformant le *eLVF* au format XML ainsi qu'une interface de consultation [2], comme nous allons le voir dans la sous-section suivante. Dans leur article présentant la version XML du *eLVF* (*xLVF*), ils ont comparé le *eLVF* avec des ressources existantes pour l'anglais (*VerbNet*(Schuler, 2005), *FrameNet*(Baker *et al.*, 1998), *WordNet*(Fellbaum, 1998)) et le français (*Dicovalence*(Mertens, 2010)). Pour résumer ce comparatif, nous pouvons dire que les approches utilisées par les ressources sont variées (pronominale pour *Dicovalence*, distributionnelle et transformationnelle pour le *LVF*, à base de cadres sémantique pour *FrameNet*, et d'ensembles de synonymes pour *WordNet*), mais qu'elles ont toutefois un certain nombre de points communs dans leur représentation de la description des unités lexicales, et de la hiérarchisation des données. D'après cette comparaison le *eLVF* est une bonne ressource linguistique car il intègre une grande partie des informations contenues dans les autres ressources, comme la hiérarchisation du sens des verbes avec les CLASSES, la description de la sémantique avec le champ OPÉRATEUR, et une liste de SYNONYMES pour chaque entrée. Il manque cependant certaines informations comme la gestion des rôles thématiques, mais le *eLVF* propose des informations que les autres ressources ne contiennent pas comme la gestion du sens figuré des verbes.

2.3 La version XML

Nous allons maintenant parler de la version XML du *eLVF* développée par Hadouche et Lapalme. Cette version du *eLVF* a pour objectif de rendre la ressource plus accessible, utilisable et extensible. Pour cela ils ont informatisé la description des différents codes contenus dans la version papier du *LVF* sous la forme de fichiers XML (voir figure 1). Ils ont ensuite généré un fichier XML représentant les données du *eLVF* décompressées. La figure 2 nous montre la version XML du verbe *« amasser »*. Les informations d'origine ont été conservées dans le fichier à l'intérieur des balises et les codes décompressés sont représentés par des attributs associés à ces balises. Toutes les informations n'ont pas été complètement décompressées, comme pour le code de CLASSE *L3b* dont nous savons qu'il représente la classe générique *Locatif* avec un sujet *non-animé propre*

2. http://rali.iro.umontreal.ca/Dubois/

```
<classes>
  <generique code="C" desc="communication">
    <semantico-syntaxique code="C1" desc="s'exprimer par un son, une parole">
      <sous-classe-syntaxique code="C1a" desc="émettre un cri, humain ou animal"/>
      <sous-classe-syntaxique code="C1b" desc="émettre un chant, humain"/>
      ...
    </semantico-syntaxique>
    <semantico-syntaxique code="C2" desc="dire/demander qc">
      <sous-classe-syntaxique code="C2a" desc="dire que, dire qc à qn"/>
      <sous-classe-syntaxique code="C2b" desc="dire que, donner un ordre à qn"/>
      ...
    </semantico-syntaxique>
    ...
  </generique>
  ...
</classes>
```

FIGURE 1 – Codes du *xLVF* pour les CLASSES

mais pas qu'il représente *mettre quelque chose quelque part, dans/sur un lieu, autour de quelque chose*. Certains champs ont aussi été restructurés comme les PHRASES, les CONSTRUCTIONS et les DÉRIVATIONS où les informations ont été séparées. Il est dommage que Hadouche et Lapalme aient uniquement intégré la description des codes au XML, car des traitements comme la dérivation des adjectifs, des adverbes et des noms auraient aussi pu être intégrés à cette nouvelle version.

2.4 Le *LVF* pour l'étiquetage de rôles sémantiques

Le *LVF* n'a pas été conçu pour l'étiquetage de rôles sémantiques, mais il contient néanmoins des informations pertinentes pour cette tâche. Les champs OPÉRATEUR, CONSTRUCTIONS et DOMAINE donnent des informations sur la syntaxe, la sémantique et l'utilisation des différentes entrées associées à un verbe. L'exploitation de ces informations devrait permettre l'identification du sens utilisé et de projeter les arguments syntaxiques sur une représentation sémantique (i.e. : le champ opérateur). Un système utilisant cette ressource serait différent de ceux existants, basés essentiellement sur de l'apprentissage automatique appliqué à de grand corpus annotés, car il utiliserait les contraintes syntaxiques et sémantiques définies manuellement par Jean Dubois et Françoise Dubois-Charlier. Le champ PHRASE pourrait être utilisé comme corpus d'exemples permettant une première évaluation d'un système d'étiquetage de rôles sémantiques. Dans un premier temps, nous allons définir notre objectif de restructuration de la ressource, puis nous décrirons les différents traitements effectués pour l'atteindre et nous terminerons sur une évaluation de la ressource obtenue.

3 Objectif

Pour qu'un système puisse utiliser le *xLVF* pour faire de l'étiquetage de rôles sémantiques, il doit être capable d'en extraire les informations nécessaires. Il est aussi important de pouvoir faire le lien entre les différents types d'information de la ressource, ainsi qu'avec des informations contenues dans d'autres ressources (e.g. : *Wolf*, *Disco*, *French TreeBank*). Il est donc nécessaire de restructurer et d'uniformiser un certain nombre d'informations du *xLVF*. De plus, il serait intéressant d'utiliser les phrases d'exemple afin de concevoir un corpus annoté, même si celui-ci n'est pas forcément représentatif. La figure 3 montre ce que l'on souhaite obtenir pour la première

```
<verbe mot="amasser" nb="3" id="amasser">
  <entree>
    <mot no="1">amasser 01</mot>
    <domaine nom="objet">OBJ</domaine>
    <classe generique="locatif" semantico-syntaxique="non-animé propre"
            construction-syntaxique="b">L3b</classe>
    <operateur>lc.qp qc+pl e amas</operateur>
    <sens>accumuler</sens>
    <phrases>
      <phrase>On <lexie-ref>a~</lexie-ref> des documents,des livres.</phrase>
      <phrase>Les preuves s'<lexie-ref>a~</lexie-ref> contre P.</phrase>
    </phrases>
    <conjugaison auxiliaire="avoir (sauf si pronominal ou entrée en être)"
   groupe="1" sous-groupe="chanter">1aZ</conjugaison>
    <construction>
      <scheme type="transitif direct" sujet="humain" objet="pluriel chose"
              circonstant="locatif (ou on est)">T1801</scheme>
      <scheme type="pronominal" sujet="pluriel chose"
              circonstant="locatif (ou on est)">P8001</scheme>
    </construction>
    <derivation der-e="positif seul"
     der-ment="il n'y a pas de nom en -ment mais il y a un déverbal">
      -1- -D ---- --
    </derivation>
    <nom nb="3">3*</nom>
    <lexique desc="dictionnaire de base" nbmots="15000">2</lexique>
  </entree>
  ...
</verbe>
```

FIGURE 2 – Entrée du *xLVF* pour le verbe *« amasser »*

entrée du verbe *« amasser »*.

Il est nécessaire d'uniformiser la ressource pour deux raisons. Premièrement pour qu'un système utilisant la ressource ne traite pas différemment une information qui aurait deux représentations distinctes, comme les abréviations *lgt* et *lgts* pour *« longtemps »* ou *poissons* et *poisson+pl* pour représenter le pluriel de *« poisson »*. La seconde raison est de rendre la ressource plus interopérable avec d'autres ressources. Pour identifier les arguments d'un prédicat, les contraintes sémantiques du *xLVF* peuvent être exploitées, mais il est nécessaire d'utiliser des ressources comme *Wolf* (Sagot et Fišer, 2008) ou *Disco* (Kolb, 2009) pour associer un type sémantique aux arguments (e.g. : humain, chose, ...). Ces ressources n'ayant pas connaissance des abréviations utilisées par le *xLVF*, il est nécessaire de remplacer les abréviations et les mots tronqués par les lemmes les représentant et de préférer la représentation du pluriel par l'utilisation de l'attribut *« +pl »*.

L'étiquetage de rôles sémantiques a pour but d'identifier les arguments du prédicat et leur associer des rôles sémantiques. L'utilisation de contraintes syntaxiques et sémantiques est une des solutions envisageables pour réaliser cette tâche. Pour cela, il est important d'avoir des informations syntaxiques et sémantiques structurées et inter-connectées pour chaque sens de chaque verbe. Ces informations doivent permettre de projeter la syntaxe sur la sémantique et inversement. Il est aussi important de bien séparer les informations syntaxiques et sémantiques, afin de permettre une meilleure cohérence de la ressource. Dans le *xLVF*, ces informations sont mélangées, et le même type d'information peut se retrouver à différents endroits, ce qui peut mener à des incohérences. Le champ OPÉRATEUR sera utilisé comme base pour la sémantique et le champ CONSTRUCTEUR comme base pour la syntaxe.

Le cadre sémantique est défini comme un prédicat auquel sont associés des arguments et des contraintes sémantiques. Les contraintes sémantiques ont des identifiants permettant d'associer

ses arguments à ceux des cadres syntaxiques. Le prédicat a pour valeur un des différents prédicats du champ OPÉRATEUR (e.g. : *r.d* : rendre/devenir tel). Les arguments sémantiques précisent la sémantique du prédicat et ne peuvent pas être réalisés syntaxiquement. Les contraintes sémantiques permettront de définir le type des différents arguments du prédicat (e.g. : un humain, une chose, un animal) qui pourront être réalisés syntaxiquement.

Les cadres syntaxiques sont ceux utilisés par le champ CONSCRUCTIONS (i.e. : intransitif, transitif indirect, transitif direct, pronominal). Les arguments sont définis par un type, un complément selon le cas (e.g. : une préposition) ainsi que des liens vers les arguments sémantiques réalisés. Un argument syntaxique peut réaliser plusieurs arguments sémantiques. Ainsi, dans la phrase « Marie se maquille », on a une variation syntaxique pronominale réfléchie et *Marie* endosse les rôles d'agent et d'expérient.

Le corpus d'exemples sera composé de phrases simples mettant en avant les différentes façons de réaliser les différents sens de chaque verbe. Ces phrases seront analysées en dépendances, et posséderont des annotations permettant de savoir quel cadre syntaxique leur est associé et à quels arguments sémantiques correspondent leurs arguments syntaxiques.

4 Transformation

Nous allons maintenant décrire les différents traitements appliqués au *xLVF* pour atteindre notre objectif. La production de cette nouvelle ressource est composée de sept étapes : l'uniformisation du champ OPÉRATEUR, sa structuration, la récupération de données à partir de la version papier du *LVF*, l'alignement des CONSTRUCTIONS, la liaison des champs OPÉRATEUR et CONSTRUCTIONS, la répartition de l'information et enfin la construction du corpus d'exemples.

L'étape d'uniformisation du champ OPÉRATEUR permet de corriger les abréviations et les mots tronqués. Un ensemble de mots suspects n'apparaissant pas dans les noms et adjectifs de *Morphalou* (Romary *et al.*, 2004) a été récupéré automatiquement (947 mots). Les occurrences de chaque mot ont été examinées manuellement afin d'identifier s'il s'agissait d'une abréviation, d'un mot tronqué ou mal orthographié, ou encore d'un mot technique n'existant pas dans *Morphalou*. Une définition ou une correction a ensuite été associée à chaque mot. Les différentes orthographes d'un mot ont été homogénéisées dans la définition afin qu'il y ait une représentation unique par sens. La définition *longtemps* a ainsi été associée aux abréviations *lgt* et *lgts*. Les mots ayant plusieurs orthographes (e.g. : acuponcture, acupuncture) ont été unifiés à l'aide de *Morphalou* et du *Wiktionnaire*, pour que les verbes ayant des orthographes différentes aient là même sémantique (e.g. : « *acuponcturer* » et « *acupuncturer* »). La gestion du pluriel a été uniformisée en prenant le singulier des mots et en leur ajoutant l'attribut *+pl*. Le mot *plantes* est ainsi devenu *plante+pl*.

La seconde étape consiste à traiter le champ OPÉRATEUR afin d'obtenir une structure proche de celle du cadre sémantique défini précédemment. Pour cela, nous avons dû identifier le prédicat et ses arguments, déterminer pour chaque argument s'il représentait une contrainte syntaxico-sémantique ou une spécification sémantique, et discerner la portée des disjonctions. Nous avons choisi d'utiliser des méthodes d'analyse de surface car elles sont robuste et facilement maintenable. Ces aspects sont importants, car le champ OPÉRATEUR comporte un grand nombre de cas particuliers que nous avons dû gérer au fur et à mesure de leur rencontre. Dans un premier

```
<CadreSemantique>
  <Contrainte cidx="0" desc="humain" />
  <Predicat desc="être ou mettre quelque part" />
  <Contrainte cidx="1" pluriel="True" desc="chose" />
  <Semantique sidx="0" prep="en" desc="amas" />
  <Contrainte cidx="2" desc="locatif (où on est)" />
</CadreSemantique>

<CadresSyntaxiques>
  <CadreSyntaxique type="transitif direct">
    <Argument type="sujet">
      <LienSemantique cidx="0" />
    </Argument>
    <Argument type="objet">
      <LienSemantique cidx="1" />
    </Argument>
    <Argument type="circonstant">
      <LienSemantique cidx="2" />
    </Argument>
  </CadreSyntaxique>
  <CadreSyntaxique type="pronominal">
    <Argument type="sujet">
      <LienSemantique cidx="1" />
    </Argument>
    <Argument type="circonstant">
      <LienSemantique cidx="2" />
    </Argument>
  </CadreSyntaxique>
</CadresSyntaxiques>

<Phrases>
  <Phrase text="On  a~ des documents.">
    <Dep dep="root" wid="2" form="a~" lemma="a~" pos="V" srl="transitif direct" >
      <Dep dep="suj" wid="1" form="On" lemma="on" pos="CL" cidx="0" />
      <Dep dep="obj" wid="4" form="documents" lemma="document" pos="N" cidx="1">
        <Dep dep="det" wid="3" form="des" lemma="un" pos="D" />
      </Dep>
      <Dep dep="ponct" wid="5" form="." lemma="." pos="PONCT" />
    </Dep>
  </Phrase>
  <Phrase text="On  a~ des livres.">
    <Dep dep="root" wid="2" form="a~" lemma="a~" pos="V" srl="transitif direct">
      <Dep dep="suj" wid="1" form="On" lemma="on" pos="CL" cidx="0"/>
      <Dep dep="obj" wid="4" form="livres" lemma="livre" pos="N" cidx="1">
        <Dep dep="det" wid="3" form="des" lemma="un" pos="D" />
      </Dep>
      <Dep dep="ponct" wid="5" form="." lemma="." pos="PONCT" />
    </Dep>
  </Phrase>
  <Phrase text="Les preuves s' a~ contre P.">
    <Dep dep="root" wid="4" form="a~" lemma="a~" pos="V" srl="pronominal">
      <Dep dep="suj" wid="2" form="preuves" lemma="preuve" pos="N" cidx="1">
        <Dep dep="det" wid="1" form="Les" lemma="le" pos="D" />
      </Dep>
      <Dep dep="aff" wid="3" form="s'" lemma="il" pos="CL" />
      <Dep dep="mod" wid="5" form="contre" lemma="contre" pos="P" cidx="0">
        <Dep dep="obj" wid="6" form="Pierre" lemma="pierre" pos="N" />
      </Dep>
      <Dep dep="ponct" wid="7" form="." lemma="." pos="PONCT" />
    </Dep>
  </Phrase>
</Phrases>
```

FIGURE 3 – Objectif d'amélioration

1	f.ire/PRD abs/ABR SR/PP	faire aller quelque chose d'abstrait sur	focaliser l'attention sur
2	abda/PRD chemin/MOT abs/ABR A/PP qn/ABR	enlever chemin abstrait à quelqu'un	couper la route du succès
3	ict/PRD total/MOT soi/ABR abs/ABR	frapper totalement soi abstrait	se suicider
4	abda/PRD pr/PP soi/ABR abs/ABR	obtenir pour soi quelque chose d'abstrait	acquérir de l'expérience
5	loq/PRD AV/PP qn/ABR D/PP //DIS SR/PP prix/	parler avec quelqu'un de prix, sur le prix	marchander
6	dat/PRD A/PP qc/ABR ,/DIS A/PP +inf/ATT	donner à quelque chose, à (infinitif)	contribuer
	contrainte syntaxico-sémantique argument sémantique disjonction		

TABLE 3 – Structuration du champ OPÉRATEUR

temps nous avons étiqueté les tokens pour abstraire les règles de l'analyse de surface. L'étiquette utilisée par défaut représente un mot (MOT). Les autres étiquettes définissent les prédicats (PRD), les prépositions (PP), les abréviations (ABR), les attributs (ATT), et les différents symboles (e.g : (),/-). L'étiquetage a été réalisé grâce à des lexiques définis manuellement et en fonction de l'emplacement des tokens. L'analyse de surface a ensuite été réalisée à partir d'un ensemble de règles générales, basées sur les étiquettes, et d'un ensemble de règles spécifiques, basées sur les mots. Les trois gros problèmes de la structuration de ce champ furent le regroupement de tokens pour former les arguments, la gestion de la portée des disjonctions, et le typage des arguments. Des règles générales permettent de regrouper des suites de mots et des prépositions avec le groupe à leur droite. Il existe cependant des cas où les abréviations et les mots peuvent être regroupés entre eux comme les exemples 2 et 3 de la table 3. La tâche n'est pas triviale car le regroupement de certains tokens est dépendant du contexte. L'abréviation *abs* peut être reconnue comme un argument canonique désignant un concept abstrait (e.g. : exemple 1, *« une idée »*), où être associée à un autre token pour l'abstraire (e.g. : exemple 2, *« le chemin du succès »*. Son association dépend de son emplacement, du prédicat de l'OPÉRATEUR, et du token auquel elle peut se lier. Des règles spécifiques utilisant des lexiques ont été mises en œuvre pour gérer ces cas particuliers. Les exemples 5 et 6 montrent que les disjonctions peuvent avoir des portées plus ou moins grandes. Le choix de la portée des arguments se fait en fonction des étiquettes des groupes adjacents à la disjonction. La portée courte se fait en priorité sur les disjonctions ayant des tokens adjacents avec les mêmes étiquettes (e.g. : exemple 5). La portée longue se fait à la fin en prenant les groupes adjacent à la disjonction (e.g. : exemple 6). Pour la dernière étape, consistant à typer les arguments, nous avons utilisé la casse des préposition, les étiquettes des tokens, ainsi que le prédicat et l'emplacement de l'argument par rapport à celui-ci. La table 3 nous montre différents exemples d'identification du type des arguments.

Un premier essai d'alignement des CONSTRUCTIONS nous a révélé que les informations contenues dans le *xLVF* n'étaient pas suffisantes. Il est nécessaire, entre autre, de savoir si le verbe est factitif et de connaitre le type des constructions pronominales (i.e. : subjectif, réfléchi, réciproque, passif) pour lever les ambigüités existantes. Ces informations apparaissant dans la version papier du *LVF*, nous avons entrepris de les extraire à partir d'une version PDF du *LVF*. Le PDF a tout d'abord été converti en HTML[3] pour avoir un format plus facilement analysable. Des expression régulières basées sur les balises HTML et sur différents mots-clés ont permis d'identifier les données utiles. Une analyse de surface a été effectuée pour donner du volume à cette suite d'éléments afin de générer un fichier XML (le *xoLVF*, voir figure 4) contenant les informations de la version papier du *LVF* dans un format univoque et structuré. En plus de ces informations, la description complète de la classification des verbes a été récupérée, ajoutant deux nouveaux niveaux de classification.

3. grâce à pdftohtml

```
<LVF>
  <ClasseGenerique nom="C" nombre="2039" desc="de communication">
    <ClasseSemantique nom="C1" nombre="1059" desc="exprimer par cri, parole, son">
      <Classe nom="C1a" nombre="232" desc="émettre un cri, humain ou animal">
        <SousClasse nom="1" desc="émettre le cri spécifique de l'espèce animale">
          <Const nom="A20" desc="intransitifs" intran="True">
            <Entree nom="aboyer 01" oper="(canis)f.cri espèce"
                    sens="émettre aboiement" phrase=" Le chien a~ ."
                    deriv="aboi,-">
                  ...
          </Const>
          ...
        </SousClasse>
        <SousClasse nom="2" desc="émettre une des diverses formes de parler ou
                                  d'écrit spécifiques à l'humain">
          <Const nom="A16" desc="intransitifs" intran="True">...</Const>
          <Const nom="P1006" desc="pronominaux" prono="subjectif">...</Const>
          ...
        </SousClasse>
        ...
      </Classe>
      ...
    </ClasseSemantique>
    ...
  </ClasseGenerique>
  ...
</LVF>
```

FIGURE 4 – Extrait du *xoLVF*, la version structurée de l'ouvrage *LVF* accessible sur notre wiki[1]

À l'aide des informations complémentaires extraites du *LVF*, l'alignement des CONSTRUCTIONS associées à une entrée a pu être effectué. Les contraintes syntaxico-sémantiques des constructions ont été utilisées afin d'identifier les arguments compatibles. Un ensemble de contraintes limitant les associations possibles entre les arguments des différentes constructions a été défini. Une première contrainte, gérant l'identité, permet d'associer des arguments ayant des codes identiques (e.g. : sujet humain et objet humain). Une autre contrainte permet d'aligner un argument pluriel avec deux arguments singuliers du même type (e.g. : sujet humain pluriel). Les informations extraites du *LVF* interdisent les associations entre certains emplacements (e.g. : le sujet d'un transitif et le sujet d'un pronominal passif), empêchent l'alignement de certains emplacements (e.g. : sujet des factitifs), et permettent de lier un élément à plusieurs du même type (e.g. : pronominaux réciproques et réfléchis). D'autres contraintes plus spécifiques ont été rajoutées après une analyse des premiers résultats, comme celle permettant de lier un argument de type humain pluriel avec un argument de type humain et un argument prépositionnel en « à ». Cet ensemble de contraintes ne permettant pas de lever toutes les ambigüités, l'alignement donne la priorité aux premiers arguments. Ainsi, pour les CONSTRUCTIONS T1100 et A10 (« applaudir »), le sujet de l'intransitif est lié au sujet du transitif (et pas à son objet direct).

La liaison des champs OPÉRATEUR et CONSTRUCTIONS peut être accomplie maintenant qu'ils ont été uniformisés et structurés. Pour cela, la redondance des informations contenue dans ces deux champs est utilisée. Les arguments des différentes CONSTRUCTIONS sont liées aux arguments de l'OPÉRATEUR en prenant soin de lier les arguments des CONSTRUCTIONS ayant été associés précédemment, au même argument de l'OPÉRATEUR. L'opération est similaire à l'étape de liaison des CONSTRUCTEURS, mais est plus complexe car l'OPÉRATEUR a un vocabulaire plus varié. Comme précédemment, des règles avec différentes priorités permettant de lier les éléments entre eux ont été utilisées. Des règles basées sur les informations syntaxiques permettent de lier les contraintes similaires, comme celles ayant des prépositions identiques. Des règles basées sur l'emplacement

Verbe	Sens	Const	Phrase
expirer 04	« faire sortir de l'air de soi »	A 1^0 0	« Pierre expire »
		T 1^0 3^1 0 0	« Pierre expire de l'air »
préoccuper 03	« être inquiet »	A 1^0 0	« Pierre est préoccupé »
		T 3^1 1^0 0 0	« L'avenir de son fils préoccupe Pierre »
accoucher 01	« enfanter »	N 1^0 b^1	« Marie accouche de Pierre »
		A 1^0 0	« Marie accouche »
		T† 1^2 1^0 0 8^3	« Le médecin accouche Marie »
libérer 04	« affranchir »	T 1^0 1^1 b^2 0	« On libère Pierre de l'emprise de sa mère »
		P‡ $1^{0,1}$ 0 b^2 0	« On se libère de l'influence de Pierre »

0 :rien, 1 :humain, 3 :chose, 8 :instrumental/moyen, b :préposition « de »
† : factitif, ‡ : pronominal réflechi

TABLE 4 – Exemples de CONSTRUCTIONS

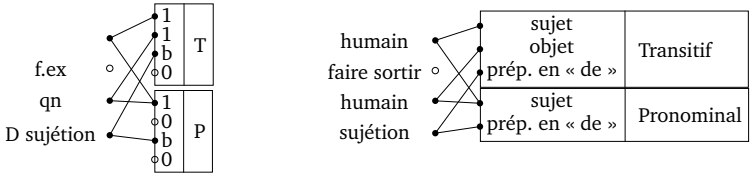


FIGURE 5 – Alignement et redistribution des champs OPÉRATEUR et CONSTRUCTIONS de *« libérer 04 »*

permettent d'associer le sujet de l'OPÉRATEUR avec le sujet de la première CONSTRUCTION. Enfin, le dernier type de règle est basé sur la sémantique et permet à des contraintes sémantiques similaires d'être liées. Si un argument d'une CONSTRUCTION ne peut être lié à aucun argument de l'OPÉRATEUR (e.g. : si l'OPÉRATEUR n'a pas de sujet), un argument factice est créé. La partie gauche de la figure 5 nous donne un exemple d'alignement, montrant les projections des constructions syntaxiques transitive et pronominale réfléchie de l'entrée *« libérer 04 »* sur son opérateur.

L'étape finale consiste à redistribuer les informations sémantiques et syntaxiques afin d'atteindre l'objectif défini dans la section 3. L'OPÉRATEUR est utilisé comme base pour le cadre sémantique et les CONSTRUCTIONS comme base pour les cadres syntaxiques. Les informations syntaxiques des contraintes de l'OPÉRATEUR ont été transférées au niveau des CONSTRUCTIONS et les contraintes sémantiques des CONSTRUCTIONS ont été transférées au niveau de l'OPÉRATEUR. Les arguments sémantiques de l'OPÉRATEUR ont été laissés tel quel. La partie droite de la figure 5 nous montre la redistribution et la transformation des codes pour l'entrée *« libérer 04 »*. Le résultat final correspond bien à l'objectif de la figure 3.

Pour la création du corpus d'exemples, nous avons utilisé les PHRASES associées aux différentes entrées. Le premier problème est que ces exemples peuvent représenter plusieurs phrases (e.g. : « On a∼des documents,des livres. », « amasser »). Nous avons donc dû décomposer ces exemples en phrases canonique à l'aide d'outils d'analyse de surface. La décomposition n'est pas toujours possible comme pour certains verbes avec des sujets pluriel et singulier (e.g. : *« Ses forces, la chance ont a∼ Pierre . »*, « abandonner »). Le second problème est que le verbe est représenté par son initiale suivie du symbole ∼, et donc un analyseur syntaxique normal ne pourra pas analyser

correctement ces exemples. Pour résoudre ce problème, nous avons entrainé un analyseur syntaxique (Bohnet *et al.*, 2010) sur une version modifiée du FrenchTreeBank(Abeillé *et al.*, 2000). Nous avons remplacé tous les verbes en tête de phrase par leur initiale suivie du symbole ∼. Nous n'avons pas remplacé tous les verbes car il existe des entrées avec des compléments syntagmatiques où le verbe est conjugué (i.e. : « On v∼qu'il soit heureux. », « vouloir »). L'autre intérêt d'avoir uniquement transformé les verbes en tête de phrase est que cela pousse l'analyseur syntaxique à prendre le verbe abrégé comme tête de la phrase. L'association des rôles sémantiques aux arguments syntaxiques a été effectuée à l'aide de règles de réécriture permettant d'identifier les types de constructions syntaxiques utilisées. Les informations sémantiques n'ont pas été nécessaires car nous connaissions le sens du verbe et les réalisations syntaxiques utilisées dans le *xLVF* sont assez limitées.

5 Évaluation

L'évaluation de la qualité des améliorations réalisées a été effectuée à l'aide de scripts permettant de vérifier la cohérence de la ressource obtenue, ainsi que par une analyse manuelle d'un échantillon représentatif.

La vérification automatisée de la cohérence de la ressource est importante car elle est peu couteuse et permet d'éviter nombre d'erreurs. Ce contrôle a été fait à l'aide de scripts vérifiant que :
– un OPÉRATEUR a un sujet, un prédicat et un certain nombre d'arguments complémentaires,
– le *xoLVF* contient le bon nombre d'entrées pour chaque classe et sous-classe,
– les arguments du CONSTRUCTEUR sont tous liés à ceux de l'OPÉRATEUR
Pour l'évaluation manuelle, un échantillon représentatif de 100 entrées a été extrait puis examiné afin de vérifier si le résultat obtenu était celui souhaité. Chaque entrée a été contrôlée sur la correction des abréviations, la structuration de l'OPÉRATEUR, la liaison des CONSTRUCTEURS, et la liaison de l'OPÉRATEUR avec les CONSTRUCTEURS. Sur les 100 entrées analysées 84 sont bonnes, 13 ont des erreurs dues à notre analyses et 3 ont des erreurs dues au *xLVF*. Parmi les erreurs de nos analyses, il y en a 3 de structuration de l'OPÉRATEUR, 1 de liaison des CONSTRUCTEURS et 9 de liaison OPÉRATEUR-CONSTRUCTEUR. Les erreurs sont dues à certaines entrées manquantes dans les lexiques. Par exemple, *org mvs* (i.e. : un mauvais organe) n'est pas regroupé en un seul terme lors de la structuration de l'OPÉRATEUR, et les contraintes du CONTRUCTEUR sur le domaine des *choses* ne sont pas liées avec les arguments *coup* et *viande* des OPÉRATEURS (car ces deux mots ne sont pas considérés comme des choses). La plupart des erreurs sont donc facilement corrigibles en ajoutant des entrées aux lexiques. Les erreurs dues au *xLVF* sont dues à certaines prépositions qui ne sont pas mises en majuscules et qui sont donc considérés comme des arguments uniquement sémantiques, comme pour l'entrée *« ressaisir 01 »* où l'argument P SOM de l'OPÉRATEUR GRP+RE QN P SOM () devrait être associé au circonstant de manière du CONSTRUCTEUR P1108.

Nous avons analysé 100 phrases annotées afin de vérifier si les différentes étapes de l'annotation se sont bien déroulées. Cette analyse indique que 70 phrases ont été bien annotées, et possèdent les bons rôles sémantiques. Les problèmes rencontrés dans les 30 autres phrases sont de natures diverses. Le problème le plus apparent est le mauvais filtrage des structures en dépendances (15 phrases). En effet, les tournures ne correspondant pas à celles que nous avions identifié n'ont pas été étiquetées. Cela est toutefois facilement corrigible en intégrant ces nouvelles tournures à

nos règles de filtrage. Le second problème rencontré est la mauvaise analyse des exemples par l'analyseur syntaxique (6 phrases). Par exemple, il arrive que des groupes prépositionnels soient rattachés à un argument plutôt qu'au verbe. Un autre problème provient de la mauvaise liaison entre les constructions et les opérateurs. Nous nous retrouvons donc à associer les arguments syntaxiques aux mauvais arguments sémantiques. Le dernier problème est la mauvaise séparation des différents exemples (3 phrases). Il arrive ainsi d'avoir une phrase avec plusieurs occurrences du verbe ou d'un de ses arguments. Les résultats de cette analyse sont donc positifs étant donné que la plupart des phrases ont été bien annotées et qu'une grande parties des erreurs est corrigible.

6 Conclusion

Nous avons présenté des améliorations du *xLVF* qui donnent à cette ressource une nouvelle dimension. Les informations son structurées, moins ambigües et plus uniformes, permettant ainsi l'utilisation du *xLVF* pour faire de l'étiquetage de rôles sémantiques. Ces améliorations vont aussi permettre de nouvelles améliorations du *xLVF* comme identifier quelles sont les entrées d'un verbe qui sont SYNONYMES d'une entrée (pour le moment les SYNONYMES sont des verbes et non des entrées). Ainsi, on sait que l'entrée *« humilier 01 »* a pour SYNONYME le verbe *« abaisser »*, mais on ne sait pas si c'est l'entrée *« abaisser 01 »* (« On abaisse le rideau. ») ou l'entrée *« abaisser 06 »* (« On abaisse Pierre. ») qui est SYNONYME. Cette nouvelle version pourra aussi être utilisée pour effectuer des recherches plus précises sur le *xLVF*, permettant par exemple à des linguistes d'identifier des verbes ayant certaines caractéristiques syntaxiques et sémantiques. Les deux fichiers XML obtenus sont en libre accès sur notre wiki `http://margaux.philosophie.uni-stuttgart.de/lvf/`. Nous pensons prochainement associer cette ressource avec *Disco* et *Wolf* pour annoter le French TreeBank.

Références

ABEILLÉ, A., CLÉMENT, L. et KINYON, A. (2000). Building a treebank for french. *In In Proceedings of the LREC 2000*.

BAKER, C. F., FILLMORE, C. J. et LOWE, J. B. (1998). The berkeley framenet project. *In Proceedings of the COLING-ACL*, Montreal, Canada.

BOHNET, B., WANNER, L., MILLE, S. et BURGA, A. (2010). Broad coverage multilingual deep sentence generation with a stochastic multi-level realizer. *In Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 98–106, Stroudsburg, PA, USA. Association for Computational Linguistics.

DUBOIS, J. et DUBOIS-CHARLIER, F. (1997). *Les verbes français*. Larousse-Bordas.

FELLBAUM, C., éditeur (1998). *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London.

FRANÇOIS, J., PESANT, D. et LEEMAN, D. (2007). *Le classement syntactico-sémantique des verbes français*. Langue française. Larousse.

HADOUCHE, F. et LAPALME, G. (2010). Une version électronique du LVF comparée avec d'autres ressources lexicales. *Langages*, 10(179-180):193–220. Mise en page différente que celle parue dans la revue.

KINGSBURY, P. et PALMER, M. (2003). Propbank : The next level of treebank. *In Proceedings of Treebanks and Lexical Theories*, Växjö, Sweden.

KOLB, P. (May 2009). Experiments on the difference between semantic similarity and relatedness. *In Proceedings of the 17th Nordic Conference on Computational Linguistics - NODALIDA '09*, Odense, Denmark.

LEVIN, B. (1993). *English verb classes and alternations : a preliminary investigation*. University of Chicago Press.

MERTENS, P. (2010). Restrictions de sélection et réalisations syntagmatiques dans dicovalence. *In Actes TALN 2010*, Montreal, Canada.

ROMARY, L., SALMON-ALT, S. et FRANCOPOULO, G. (2004). Standards going concrete : from LMF to Morphalou. *In Workshop Enhancing and Using Electronic Dictionaries*, page 7 p, Geneva, Switzerland. none. Colloque avec actes et comité de lecture. internationale.

SAGOT, B. et FIŠER, D. (2008). Building a free French wordnet from multilingual resources. *In OntoLex*, Marrakech, Morocco.

SCHULER, K. K. (2005). *Verbnet : a broad-coverage, comprehensive verb lexicon*. Thèse de doctorat, , Philadelphia, PA, USA. AAI3179808.

Étude des manifestations de la relation de méronymie dans une ressource distributionnelle

François Morlane-Hondère Cécile Fabre
CLLE-ERSS, Université de Toulouse - Le Mirail, 5, allées Antonio Machado - Toulouse Cedex 9
francois.morlane@univ-tlse2.fr, cecile.fabre@univ-tlse2.fr

RÉSUMÉ

Cette étude vise à étudier les manifestations de la relation de méronymie dans une ressource lexicale générée automatiquement à partir d'un corpus de langue générale. La démarche que nous adoptons consiste à recueillir un jeu de couples de méronymes issus d'une ressource externe que nous croisons avec une base distributionnelle calculée à partir d'un corpus de textes encyclopédiques. Une annotation sémantique des mots qui entrent dans ces couples de méronymes montre que la prise en compte de la nature sémantique des mots composant les couples de méronymes permet de mettre au jour des inégalités au niveau du repérage de la relation par la méthode d'analyse distributionnelle.

ABSTRACT

Study of meronymy in a distribution-based lexical resource

In this paper, we study the way meronymy behaves in a distribution-based lexical resource. We address the question of the evaluation of such resources through a semantic-based approach. Our method consists in collecting meronyms from a resource which we cross with a distribution-based lexical resource made from an encyclopedic corpus. Meronyms are then sub-categorized manually : firstly following the sub-relation they bear (STUFF/OBJECT, MEMBER/COLLECTION, etc.), then following the semantic class of their members. Results show that distributional analysis identifies meronymic relations in different proportions according to the semantic classes of the words involved in the meronymic pairs.

MOTS-CLÉS : analyse distributionnelle, sémantique lexicale, méronymie, évaluation.

KEYWORDS: distributional analysis, lexical semantics, meronymy, evaluation.

1 Introduction

Cette étude s'inscrit dans la problématique générale de l'évaluation des méthodes d'analyse distributionnelle (dorénavant *AD*) pour l'acquisition d'informations sémantiques (Baroni et Lenci, 2010). Ces méthodes de sémantique distributionnelle consistent à mesurer le degré de proximité sémantique entre mots sur la base du recouvrement de leurs contextes syntaxiques. La qualité des résultats fournis par ces méthodes s'avère difficile à mesurer du fait de la grande quantité de couples de mots générée par l'analyse et de la diversité des relations mises au jour (Sahlgren, 2006). Le typage des relations sémantiques calculées par l'AD est donc un enjeu pour optimiser l'utilisation des ressources distributionnelles dans des applications de TAL (van der Plas, 2008). Différents travaux ont cherché à mesurer l'efficacité des méthodes distributionnelles pour le repérage des relations lexicales en employant des méthodes impliquant des ressources de

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 169–182,
Grenoble, 4 au 8 juin 2012.

référence (Lin, 1998; Turney, 2008; Baroni et Lenci, 2011). En particulier, Baroni et Lenci ont soumis les résultats du calcul distributionnel à un large éventail de tâches sémantiques. Pour le français, les études ont été principalement consacrées au repérage de la synonymie (Bourigault et Galy, 2005; Ferret, 2010; Muller et Langlais, 2011), en particulier du fait de la disponibilité de lexiques de synonymes. Dans une étude précédente, nous nous sommes intéressés au cas de l'antonymie (Morlane-Hondère et Fabre, 2010). Tous ces travaux confirment la grande diversité des relations que peut détecter l'AD, et montrent aussi la nécessité de mieux comprendre sous quelles conditions le critère distributionnel opère, autrement dit, quels types d'informations distributionnelles doivent être pris en compte selon la nature de la tâche sémantique que l'on veut réaliser.

Dans cet article, nous nous focalisons sur le cas d'une relation lexicale particulière, la relation de méronymie (ou relation partie/tout), pour étudier la façon dont elle est repérée par un programme d'analyse distributionnelle. La relation de méronymie est intéressante à plusieurs titres : tout d'abord, elle constitue l'une des relations visées par les méthodes d'acquisition de ressources lexicales et terminologiques, au même titre que les relations plus souvent étudiées que sont l'hyperonymie et la synonymie. Ensuite, elle a la particularité de recouvrir un ensemble varié de relations (LIEU/ZONE, CONSTITUANT/OBJET, ÉTAPE/ACTIVITÉ, MEMBRE/COLLECTION, etc.), ce qui offre un terrain d'observation particulièrement riche pour étudier les modalités d'application de l'AD. Enfin, contrairement à la synonymie et à l'hyperonymie, la méronymie ne relie pas des mots relevant systématiquement de la même classe sémantique : c'est le cas par exemple du couple de mots *tête* et *enfant*, le premier étant une partie du corps, le second un être humain. Un tel cas de figure semble *a priori* défavorable au repérage par l'AD. C'est un des points que nous cherchons à examiner dans cet article.

La démarche que nous avons adoptée s'appuie sur un jeu de couples de méronymes issu du réseau JeuxDeMots (désormais *JDM*) (Lafourcade, 2007). Nous avons croisé ces données avec une base distributionnelle construite à partir d'un corpus issu de l'encyclopédie en ligne Wikipédia. Après une présentation de la relation de méronymie et des sous-relations qui la composent (2), nous décrivons les deux ressources que nous avons utilisées (3). Nous présentons ensuite la phase d'annotation, qui a donné lieu à deux procédures successives (4). Dans la section consacrée aux résultats (5), les couples ainsi annotés sont analysés du point de vue distributionnel, ce qui nous permet de dégager des classes de relations selon leur propension à être détectées par l'AD, et d'analyser ces différences.

2 La relation de méronymie : définition et typologie

La relation de méronymie est la relation qui s'établit entre une *partie* et son *tout*. Elle est asymétrique et sa réciproque, la relation entre un tout et l'une de ses parties, est l'holonymie. C'est une relation qui opère principalement entre deux noms, bien que Winston *et al.* (1987) proposent une relation FEATURE/ACTIVITY pour les couples désignant une étape dans un processus comme *paying/shopping*. La définition que donne Cruse (1986) de la méronymie est la suivante "X is a meronym of Y if and only if sentences of the form *A Y has Xs / an X* and *An X is a part of a Y* are normal when the noun phrases *an X, a Y* are interpreted generically." Ainsi, *La main est une partie du bras* est vrai, et ce même s'il existe des bras dont la main a été coupée.

La relation de méronymie est prise en compte dans la construction de thésaurus et d'ontologies (Van Campenhoudt, 1996; Keet et Artale, 2008). Elle se décline en plusieurs sous-relations.

Winston *et al.* (1987) définissent six sous-types de méronymes en s'appuyant sur trois critères :
– la *fonctionnalité* : la partie a-t-elle une fonction vis-à-vis du tout ? Par exemple, *poignée* est fonctionnel vis-à-vis de *porte*, mais pas vis-à-vis de *maison*.
– l'*homéomérité* : la partie et le tout sont-ils matériellement identiques (comme *tranche/gâteau*, contrairement à *arbre/forêt*) ?
– la *séparabilité* : la partie et le tout sont-ils séparables ? C'est le cas de *anse* et *tasse*, mais pas d'*acier* et *vélo*.

Relation	Exemple	Critères		
		Fonct.	Homéo.	Sépar.
ÉLÉMENT/OBJET	anse/tasse	+	−	+
MEMBRE/COLLECTION	arbre/forêt	−	−	+
PORTION/MASSE	tranche/gâteau	−	+	+
CONSTITUANT/OBJET	acier/vélo	−	−	−
ÉTAPE/ACTIVITÉ	payer/magasiner	+	−	−
LIEU/ZONE	oasis/désert	−	+	−

TABLE 1 – Sous-types de la relation de méronymie définis par Winston *et al.* (1987).

La combinaison de ces trois critères leur permet de dégager les relations rapportées au tableau 1. En marge de cette première série, Winston *et al.* (1987) décrivent des relations qui s'apparentent à de la méronymie sans en être tout à fait. Ces relations sont les suivantes :
– l'inclusion topologique : l'holonyme est un contenant (*prisonnier/cellule*), une zone (*Berlin Ouest/Allemagne de l'Est*) ou exprime une durée temporelle (*réunion/matin*).
– l'inclusion de classe : il s'agit ici de la relation d'hyponymie (*rose/fleur*, *peur/émotion*, etc.).
– la relation d'attribution : il s'agit d'une relation de type modifieur entre un mot et un adjectif (*tour/haute*, *blague/drôle*, etc.).
– la relation d'attachement : elle porte sur deux objets attachés l'un à l'autre (*boucle d'oreille* et *oreille*).
– la relation d'appartenance : elle relie des mots comme *millionnaire* et *argent* ou *auteur* et *copyright* et peut être confondue avec la méronymie à cause de l'ambiguïté du patron *X a Y*, qui peut exprimer l'appartenance (*Camille a un vélo* vs. *Un vélo a des roues*).

La différence entre certaines de ces relations et la méronymie *stricto sensu* est parfois assez fine. Nous verrons à la section 4 que beaucoup des paires annotées relèvent de l'une ou l'autre de ces relations pseudo-méronymiques.

3 Description des données

Cette étude repose sur la confrontation de deux types de données : les Voisins de Wikipédia, qui est une base lexicale générée automatiquement par analyse distributionnelle à partir de corpus, et une ressource lexicale construite de manière collaborative, JeuxDeMots.

3.1 Les voisins de Wikipédia

La base distributionnelle utilisée dans cette étude a été calculée à partir d'un corpus constitué de l'intégralité des articles de l'encyclopédie en ligne Wikipédia dans une version datant d'avril 2007. Ce corpus compte environ 194 millions de mots. L'analyse syntaxique du corpus a été

effectuée par le programme Syntex (Bourigault, 2007) à partir d'une version du corpus Wikipédia précédemment étiquetée morpho-syntaxiquement par TreeTagger. L'analyse distributionnelle a été réalisée par l'outil Upery développé par Didier Bourigault (2002) [1]. À partir des relations de dépendance syntaxique calculées par Syntex, le programme Upery extrait dans un premier temps des triplets de structure (mot1, RELATION, mot2). Les relations syntaxiques prises en compte sont les relations sujet, objet, complément prépositionnel, modification adjectivale. Les mots sont des unités simples ou des syntagmes, sous une forme lemmatisée, par exemple : *utiliser*, OBJET, *voiture*. Ces triplets servent de base au calcul distributionnel, qui rapproche les couples d'éléments qui partagent les mêmes contextes syntaxiques. Ces éléments sont de deux types : prédicats ou arguments. L'argument correspond au mot régi par la relation (ex : *voiture*). Le prédicat résulte de l'association du mot recteur et de la relation (ex : *utiliser*_OBJ). Upery rapproche donc les prédicats qui partagent les mêmes arguments, ainsi que les arguments qui partagent les mêmes prédicats. Par exemple, *voiture* est rapproché de *véhicule* parce que ces deux mots partagent, en position argument, les contextes suivants : *loueur*_DE, *pneu*_DE, *garer*_OBJ, *percuter*_SUJ, etc. La mesure de similarité utilisée est basée sur l'indice de Lin (1998). Le score de similarité de deux prédicats ou arguments varie – de 0 à 1 – en fonction de plusieurs facteurs : le nombre de contextes partagés, le nombre de triplets différents dans lesquels chacun de deux mots apparaît (indice de productivité), le degré de spécificité du contexte qui permet d'effectuer le rapprochement (se reporter à (Bourigault, 2002) pour les détails de la procédure de calcul). La base de voisins distributionnels de Wikipédia compte 2 441 118 paires de mots.

3.2 JeuxDeMots

Le jeu de couples que nous avons utilisé est issu de la base JeuxDeMots, qui est un réseau lexical enrichi de façon collaborative (Lafourcade, 2007) : des utilisateurs – experts et non-experts – se connectent à une interface en ligne [2] et ont pour tâche de proposer un ensemble de mots pour une relation et un mot-cible donnés. Les réponses communes à deux joueurs sont ajoutées au réseau, et si le lien était déjà présent, alors il est renforcé selon un système de pondération. Les relations proposées par le jeu incluent aussi bien des relations lexicales classiques que des relations moins usuelles (CHOSE/LIEU, ACTION/INSTRUMENT, etc.). Nous nous sommes tournés vers cette ressource car elle est librement accessible et est une des rares en français à inclure des couples portant la relation partie-tout.

Nous avons donc récupéré de JDM (dans sa version du 10/05/2011) les couples de noms entretenant une relation de méronymie ou d'holonymie. Ces derniers ont été produits par des joueurs auxquels il était demandé de "Donner des TOUT/PARTIES" des mots-cibles qui leur étaient proposés. Cette consigne étant relativement floue, nous verrons que les couples produits se trouvent souvent à la frontière de ce que l'on considère comme de la méronymie au sens strict (section 4.1).

3.3 Croisement des deux ressources

Nous avons croisé la base de méronymes avec les voisins distributionnels afin de repérer les paires de méronymes qui ont été repérées par l'AD. Pour cela, nous avons dans un premier temps éliminé de la base de méronymes les paires dont un mot au moins était absent du vocabulaire

1. La constitution du corpus et l'application de Syntex et Upery à ce corpus ont été réalisées par Franck Sajous, qui en a également assuré la mise en ligne : `http://redac.univ-tlse2.fr/voisinsdewikipedia/`.
2. `http://www.jeuxdemots.org/jdm-accueil.php`

des voisins. Nous avons ensuite symétrisé les couples de méronymes : pour tout couple A/B où A est méronyme de B nous avons généré un couple B/A où B est holonyme de A. La base obtenue compte 15 912 couples dont 34 % (5380) sont présents parmi les voisins. Ce taux de recouvrement est comparable à celui que nous avons pu observer pour d'autres relations (synonymie, antonymie).

Dans un deuxième temps, nous avons calculé le rapport de productivité (cf. section 3.1) entre les deux membres de chaque couple contenu dans la base de méronymes afin de ne conserver que les couples dont les deux membres ont des productivités comparables. En effet, de nombreux couples de mots ne sont pas repérés par l'AD parce que leurs productivités sont trop déséquilibrées. Cette étape vise donc à atténuer les effets liés au calcul des voisins en ne conservant que les couples qui sont potentiellement repérables par l'AD. Le seuil du rapport de productivité a été fixé à 0,60, ce qui signifie que, dans un couple, un mot ne pourra pas avoir une productivité 40 % plus élevée ou plus basse que l'autre mot. La base obtenue (désormais $\text{JDM}_{méro}$) compte 1520 paires dont 55 % (829) sont captées par l'AD.

4 Phase d'annotation

La phase d'annotation doit permettre de prendre en compte parmi les couples de $\text{JDM}_{méro}$ la diversité des sous-relations qu'inclut la méronymie. Nous nous sommes appuyés dans un premier temps sur la typologie de Winston *et al.* (1987), décrite à la section 2.

4.1 Typologie de Winston *et al.* (1987)

Cette annotation étant entièrement manuelle, nous n'avons dans un premier temps annoté qu'une partie de la base, soit 481 couples de $\text{JDM}_{méro}$ en prenant pour critère un seuil de productivité supérieur ou égal à 0,85. La répartition des paires dans les catégories a été rapportée au tableau 2.

Les relations décrites comme méronymiques dans la typologie figurent dans la partie haute du tableau, les relations pseudo-méronymiques apparaissent en bas. On peut constater que c'est la relation ÉLÉMENT/OBJET qui prévaut (elle englobe plus d'un tiers de l'ensemble des couples). Elle correspond à un vaste éventail de cas où la partie a un rôle fonctionnel vis-à-vis de son tout : *pince/crabe*, *clavier/piano*, Les autres relations sont nettement minoritaires. Dans la partie basse du tableau, les relations pseudo-méronymiques représentent 44 % des couples annotés. La dernière relation, identifiée dans le tableau par un point d'interrogation, regroupe tous les couples qui, selon nous, portent une relation autre que toutes celles qui ont été identifiées par (Winston *et al.*, 1987) : *chemin/voyage*, *corps/femme*, *électricité/fil*, *activité/temps*, etc. Ces couples sont au nombre de 151, ce qui représente 31,4 % de l'ensemble. Certains d'entre eux passent le test de l'insertion dans un patron de type *X a Y* (*chauffeur/taxi*, *billet/montant*, *carte/couleur*). Beaucoup des paires appartenant à cette catégorie sont constituées de noms exprimant des concepts abstraits pour lesquels les critères de fonctionnalité, d'homéomérité et de séparabilité sont difficilement applicables comme dans *calcul/chiffre* ou *lumière/univers*. Malgré le caractère périphérique d'une partie des relations, nous prenons comme objet d'étude le jeu de couples dans son ensemble, dans la mesure où ils ont été produits par des locuteurs qui les ont perçus comme relevant de la relation partie/tout.

Le bilan que l'on peut tirer de cette première annotation est que la typologie montre ses limites

	Relation	Fréq.	Proportion
Relations méronymiques	ÉLÉMENT/OBJET	177	36,8 %
	CONSTITUANT/OBJET	35	7,3 %
	LIEU/ZONE	34	7,1 %
	MEMBRE/COLLECTION	14	2,9 %
	ÉTAPE/ACTIVITÉ	6	1,2 %
	PORTION/MASSE	1	0,2 %
Autres relations	?	151	31,4 %
	INCLUSION TOPOLOGIQUE	31	6,4 %
	CLASSE	18	3,7 %
	SYNONYMIE	11	2,3 %
	APPARTENANCE	3	0,6 %

TABLE 2 – Résultats de l'annotation basée sur la typologie de Winston *et al.* (1987).

lorsqu'elle est confrontée aux données de JDM, pour deux raisons :
– une seule relation – ÉLÉMENT/OBJET – concentre près de 66 % des couples considérés comme relevant strictement de la méronymie. Cette classe apparaît manifestement comme trop englobante dans la mesure où elle porte sur des couples de nature hétérogène.
– 31,4 % des couples ne relèvent pas d'une des relations définies dans la typologie de référence, même en l'augmentant avec la série des relations pseudo-méronymiques.
Nous avons donc décidé de délaisser une typologie préétablie et d'adopter une approche *bottom-up* : nous nous focalisons cette fois sur le sens des mots composant les paires de méronymes afin de faire émerger des combinaisons de classes sémantiques.

4.2 Annotation en classes sémantiques

La deuxième procédure d'annotation consiste à attribuer une classe sémantique à chaque mot des couples de méronymes, afin de mettre au jour de nouvelles configurations distributionnelles. Elle s'inspire du point de vue de Murphy (2003), qui rejette l'idée selon laquelle il existerait plusieurs déclinaisons de la méronymie et qui considère que la seule chose qui change entre les différents sous-types est la nature des mots sur lesquels porte la relation.

Les couples que nous avons utilisés sont ceux de la base $\text{JDM}_{méro}$ (section 3.2) : la méthode d'annotation étant cette fois semi-automatisée, nous avons utilisé un ensemble plus élevé de paires que dans la section précédente. Suite aux résultats obtenus lors de l'annotation effectuée à partir de la typologie de Winston *et al.* (1987), nous avons choisi de retirer manuellement les couples d'hyperonymes et de synonymes. $\text{JDM}_{méro}$ compte désormais 1334 paires dont 53 % (711) sont détectés par l'AD (contre 1520 paires dont 55 % de voisins dans sa version précédente).

En guise de classe sémantique, nous avons associé chaque mot à l'un de ses hyperonymes *de haut niveau* dans WordNet (Fellbaum, 1998). La raison pour laquelle nous avons utilisé cette ressource plutôt que les hyperonymes de JDM est que les relations d'hyperonymie y sont présentes de façon plus systématique que dans JDM (tous les mots de notre sous-ensemble n'ont pas forcément d'hyperonyme dans JDM). Cette démarche a, dans un premier temps, consisté à

traduire le lexique de JDM$_{méro}$ en anglais[3]. Nous avons ensuite associé chaque mot à l'ensemble des hyperonymes de sa traduction anglaise dans le réseau, et ce pour chacune de ses acceptions recensées dans WordNet. Ainsi, *église* est associé aux quatre chemins suivants :

ENTITY>ABSTRACT_ENTITY>ABSTRACTION>EVENT>HUMAN_ACTIVITY>ACTIVITY>CEREMONY

ENTITY>ABSTRACT_ENTITY>ABSTRACTION>SOCIAL_GROUP>ORGANIZATION>INSTITUTION>RELIGION

ENTITY>ABSTRACT_ENTITY>ABSTRACTION>SOCIAL_GROUP>GATHERING>BODY

ENTITY>PHYSICAL_ENTITY>PHYSICAL_OBJECT>WHOLE>ARTIFACT>STRUCTURE>EDIFICE>PLACE_OF_WORSHIP

L'étape suivante consiste à procéder à un *élagage* de l'arborescence. Par défaut, la granularité de WordNet est bien trop fine pour nous permettre d'obtenir des classes de taille satisfaisante (par exemple, dans nos données, *église* est le seul mot à figurer en position hyponyme de CEREMONY). L'élagage vise à obtenir une arborescence moins complexe. Ainsi, nous avons choisi de couper les noms abstraits (hyponymes de ABSTRACT_ENTITY) au troisième niveau de profondeur et les noms concrets (hyponymes de PHYSICAL_ENTITY) au cinquième. Ce choix se justifie par un nombre plus important de noms concrets dans nos données. Dans le cas de *église*, cela entraîne la disparition de la nuance entre les deux acceptions du mot en tant que groupe social :

ENTITY>ABSTRACT_ENTITY>ABSTRACTION>EVENT

ENTITY>ABSTRACT_ENTITY>ABSTRACTION>SOCIAL_GROUP

ENTITY>PHYSICAL_ENTITY>PHYSICAL_OBJECT>WHOLE>ARTIFACT>STRUCTURE

Les mots ainsi annotés sont ensuite désambiguïsés manuellement en fonction du mot avec lequel ils entretiennent une relation de méronymie. Dans l'exemple précédent, cette démarche consiste à associer *église* à un type de bâtiment (STRUCTURE) dans le couple *église/village* et à un groupe social (SOCIAL_GROUP) dans le couple *fidèle/église*. Toujours dans la même optique, nous avons procédé à différents ajustements consistant à opérer des regroupements entre certaines classes de mots. Cette étape s'est faite de façon empirique en fonction notamment du nombre d'éléments contenus dans chaque catégorie : par exemple, les éléments appartenant à des catégories de moins de 10 membres ont été systématiquement déplacés sous l'hyperonyme de niveau supérieur. Dans l'exemple ci-dessous, le premier chemin correspond à celui de *doigt*, le deuxième à celui de *nez* :

ENTITY>PHYSICAL_ENTITY>THING>PIECE>BODY_PART>EXTERNAL_BODY_PART>MEMBER>DIGIT

ENTITY>PHYSICAL_ENTITY>THING>PIECE>BODY_PART>ORGAN>SENSORY_RECEPTOR>CHEMORECEPTOR

Après regroupement, les deux mots se retrouvent au même niveau dans la hiérarchie : ils sont directement subordonnés à BODY_PART. La répartition finale des mots de JDM$_{méro}$ après désambiguïsation et élagage de la hiérarchie a été rapportée à la figure 1 pour les noms concrets et à la figure 2 pour les noms abstraits. Sur ces figures, on constate clairement que la profondeur de la hiérarchie varie selon les classes. La classe ABSTRACT_ENTITY n'a qu'un niveau de profondeur. Elle se divise en quatre classes : les événements (EVENT : *exposition*, *procès*), les groupes sociaux (SOCIAL_GROUP : *peuple*, *famille*), les collections (COLLECTION : *flotte*, *galaxie*) et *autres* (OTHER : *trou*, *valeur*). La catégorie *autres* est un ajout de notre part, elle contient des mots appartenant à des classes comme les jours de la semaine, les unités monétaires ou les notes de musique qui contiennent trop peu de membres pour avoir une existence autonome dans notre classification. La classe des noms concrets regroupe un nombre de noms beaucoup plus important que la classe des noms abstraits (2082 contre 460). Elle est structurée de façon plus complexe et possède trois niveaux de profondeur. Le premier niveau distingue les parties du corps (LIVING_PART, qui

3. Cette étape a été facilitée par l'utilisation de Google Traduction (`http://translate.google.fr/`). Les traductions ont ensuite été vérifiées manuellement. La trentaine de cas d'ambiguïtés liés à la traduction – *plat* traduit *flat* plutôt que *dish*, ou *car* traduit *because* au lieu de *bus* – ont été également désambiguïsés.

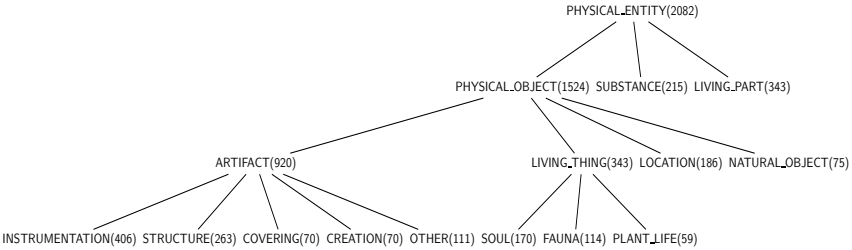


FIGURE 1 – Répartition des mots de JDM$_{méro}$ dans la classe PHYSICAL_ENTITY.

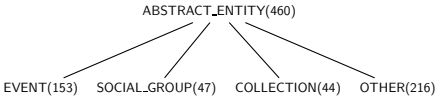


FIGURE 2 – Répartition des mots de JDM$_{méro}$ dans la classe ABSTRACT_ENTITY.

regroupe en fait les parties de corps humain, animal, et les parties de végétaux : *langue*, *patte*), les substances ou matières (SUBSTANCE : *fer*, *laine*), et les objets physiques. Cette dernière classe, la plus volumineuse, regroupe les objets naturels (NATURAL_OBJECT : *torrent*, *volcan*), les lieux (LOCATION : *grotte*, *quartier*), les entités vivantes (LIVING_THING) et les artéfacts (ARTIFACT). Ces deux dernières classes possèdent enfin un dernier niveau de profondeur. La classe des entités vivantes se subdivise en trois sous-classes regroupant les noms se rapportant à des humains (SOUL : *joueur*, *pompier*), des animaux (FAUNA : *canard*, *requin*) et des végétaux (PLANT_LIFE : *olive*, *rose*). La classe des artéfacts comprend les noms d'instruments (INSTRUMENT : *lampe*, *pneu*), de bâtiments (STRUCTURE : *lycée*, *magasin*), de vêtements (COVERING : *cape*, *chapeau*), les créations (CREATION : cette classe englobe les productions littéraires, artistiques comme *fresque* ou *roman*) et une catégorie *autres* (OTHER : de la même façon que la catégorie éponyme dans la classe des entités abstraites, elle regroupe des objets de nature hétérogène comme *brique* ou *savon*).

L'abandon d'une typologie préétablie au profit de classes sémantiques va nous permettre de mener des analyses plus précises : dans la section suivante, nous analysons les propriétés distributionnelles des couples de classes les plus fréquents dans notre jeu de méronymes.

5 Analyse des couples

À ce stade de l'étude, nous nous intéressons à deux propriétés des 1334 couples de la base JDM$_{méro}$:
– chaque couple a été catégorisé selon qu'il a été capté par l'AD ou non (section 3.2),
– chaque membre de chaque couple de méronymes est associé a une étiquette sémantique qui lui est propre (section 4.2).

classe méro.	classe holo.	nb. de couples	% voisins
SOUL	SOCIAL_GROUP	54	96,2
LOCATION	LOCATION	19	90,0
EVENT	EVENT	29	85,3
STRUCTURE	LOCATION	94	82,6
STRUCTURE	STRUCTURE	12	82,5
SOUL	STRUCTURE	25	66,7
INSTRUMENTATION	STRUCTURE	11	51,9
SUBSTANCE	SUBSTANCE	27	51,2
INSTRUMENTATION	INSTRUMENTATION	82	41,5
LIVING_PART	PLANT_LIFE	76	41,2
LIVING_PART	LIVING_PART	62	33,9
SUBSTANCE	INSTRUMENTATION	17	33,3
LIVING_PART	SOUL	54	18,4
LIVING_PART	FAUNA	41	14,6

TABLE 3 – Couples de classes qui englobent au moins 10 couples dans $\text{JDM}_{méro}$.

Nous croisons à présent ces deux aspects afin de mettre au jour des couples de classes de mots et d'expliquer pourquoi certaines sont mieux captées par l'AD que d'autres. Nous avons rapporté au tableau 3 les couples de classes représentés par au moins 10 paires de méronymes dans la base (nous avons supprimé la classe OTHER à cause du caractère hétérogène des relations qu'elle englobe). Ils sont classés par ordre de proportion de voisins décroissante. On peut constater qu'il y a de fortes disparités entre les différentes combinaisons : alors que 96,2 % des couples dont le méronyme est un humain et l'holonyme un groupe social sont repérés par l'AD, cela n'est vrai que de 14,6 % des couples dont le méronyme est une partie du corps et l'holonyme un animal. Dans cette section, nous nous focalisons sur l'observation des propriétés distributionnelles de ces différents types de classes afin d'expliquer pourquoi certaines sont plus compatibles que d'autres. Les différentes combinaisons de classes sont analysées en deux temps. La première section est consacrée à l'étude des couples constitués de deux mots appartenant à des classes identiques – couples dits *homogènes*. Les couples constitués de deux mots relevant de deux classes différentes – couples *hétérogènes* – sont analysés à la deuxième section. Ce découpage est motivé par le fait que, contrairement aux autres relations classiques, la méronymie possède la particularité de pouvoir associer deux mots qui ont des natures sémantiques différentes (*vache/troupeau*, *métal/épée*). Or, on sait que l'AD basée sur l'analyse des contextes syntaxiques présente une tendance à rapprocher des mots qui sont sémantiquement similaires. Les données dont nous disposons nous donnent la possibilité de mettre au jour les conditions dans lesquelles se principe se vérifie ou ne se vérifie pas.

5.1 Couples homogènes

Parmi les 14 types de couples rapportés au tableau 3, 6 sont homogènes. Leur proportion de voisins moyenne est de 64,1 %, ce qui est un peu plus élevé que celle des couples hétérogènes, qui est de 50,6 %.

5.1.1 Les classes les mieux repérées

Les couples composés de deux éléments appartenant aux classes LOCATION, EVENT ou STRUCTURE sont repérés par l'AD dans des proportions allant de 82,5 % à 90 %. Les couples dont les deux membres appartiennent à la classe LOCATION expriment une relation entre deux lieux, l'un étant localisé dans un second de taille supérieure (*Allemagne/Europe*, *place/village*). Ce sont les couples homogènes qui sont le mieux repérés par l'AD. Les mots qui les composent partagent la propriété d'exprimer des entités localisées spatialement. De fait, ils partagent des contextes comme la position objet de verbes de localisation – *(se) situer*, *se trouver* – *via* des prépositions complexes comme AU SUD DE ou AU CENTRE DE, ou encore la position complément du nom, quand le nom exprime un point cardinal (NORD DE, SUD DE, etc.). En plus de partager ce faisceau de contextes, les mots exprimant des lieux se distinguent par des contextes spécifiques qui permettent de distinguer des sous-classes le lieux. Par exemple, beaucoup des couples de lieux expriment différents niveaux de subdivisions administratives (*commune/canton*, *village/département*, etc.). Ils partagent des contextes comme *administration*_DE, *communauté*_DE, *population*_DE ou *territoire*_DE. De la même façon, l'analyse des contextes du couple *propriété/parc* montre qu'ils ont été rapprochés à la fois grâce au fait qu'ils sont des objets localisés dans l'espace (*limite*_DE, *s'étendre*_SUR, *superficie*_DE), mais aussi parce qu'il apparaissent comme des biens que l'on peut posséder (*revendre*_OBJ, *acheter*_OBJ, *gérer*_OBJ). Ces contextes spécifiques viennent renforcer la proximité distributionnelle entre les différents sous-ensembles de la classe des noms de lieux.

Le cas des couples d'événements est assez similaire si ce n'est que les mots expriment des valeurs temporelles et non plus spatiales : l'événement méronyme prend place dans un processus de plus grande ampleur exprimé par le second membre de la paire (*bataille/campagne*, *départ/course*, *victoire/combat*). Les noms exprimant une durée s'emploient dans des contextes comme *avoir lieu*_SUJ, *prendre fin*_SUJ et *se terminer*_SUJ, par l'intermédiaire de prépositions comme LORS DE, AU COURS DE, etc. Ici aussi, certains types d'événements se distinguent du fait, par exemple, que certains ont un aspect duratif alors que d'autres sont plus ponctuels (*mission* vs. *victoire*). Comme ça été le cas pour les noms de lieux, les contextes exprimant la localisation temporelle d'un événement sont associés à d'autres contextes exprimant des caractéristiques liées au sous-type d'événement.

Enfin, la classe STRUCTURE relie des noms de bâtiments ou de parties de bâtiments qui se situent au sein d'un autre bâtiment (*tour/château*, *hall/immeuble*, *salle/lycée*). Il semblerait que la classe des bâtiments ait une distribution moins bien circonscrite que celle des lieux et des événements. En effet, l'étude des paires de cette catégorie ne fait apparaître que peu de contextes transversaux, qui s'appliquent à l'ensemble des mots appartenant à la classe des bâtiments. Le contexte *construire*_OBJ en est un : il peut virtuellement s'appliquer à tout type de bâtiment mais ne permet pas, par exemple, de rapprocher la paire *salon/appartement*. De la même façon, le contexte *habiter*_OBJ est assez répandu mais ne s'applique, par définition, qu'aux structures destinées à être habitables (ce contexte n'apparaît pas dans les contextes communs de la paire de voisins *pièce/musée*, par exemple). Ainsi, la classe des bâtiments apparaît de façon assez floue, dans la mesure où l'emploi qui est fait des noms de bâtiments dans le corpus met l'accent sur leur aspect fonctionnel. Il semblerait que les classes qui émergent se situent à un niveau de granularité inférieur. Par exemple, les couples *appartement/immeuble* et *chambre/hôtel* possèdent en commun des contextes comme *habiter*_OBJ, *louer*_OBJ ou *se installer*_DANS. Ces contextes définissent un type de bâtiment bien particulier, à savoir les bâtiments destinés au logement. De la même façon, les couples *tour/château* et *fortification/fort* partagent des contextes comme

*protéger*_OBJ, *détruire*_OBJ ou *attaquer*_OBJ qui permettraient de dessiner les contours de la classe des bâtimens militaires.

Ainsi, les mots qui appartiennent à ces trois types de couples de classes sont particulièrement bien repérés par l'AD du fait que les propriétés sémantiques qu'ils partagent se répercutent sur le plan distributionnel. Nous avons vu que les distributions des couples de lieux et d'événements se caractérisaient par un ensemble de contextes compatibles avec la plupart des mots appartenant à chacune de ces classes. Ce constat se vérifie dans une moindre mesure sur la classe des bâtiments, pour laquelle les classes qui émergent se situent à un niveau plus fin.

5.1.2 Classes repérées en quantités moindres

Les couples composés de deux éléments appartenant aux classes SUBSTANCE, INSTRUMENTATION ou LIVING_PART sont captés par l'AD dans des proportions allant seulement de 33,9 % à 51,2 %. Nous avons donc affaire à des couples de mots qui, tout en possédant la même étiquette sémantique, se caractérisent par des propriétés distributionnelles différentes.

Dans le cas de la classe SUBSTANCE, les mots reliés désignent deux substances ou matières (au sens large) dont l'une entre dans la composition de l'autre : *carbone/diamant*, *crême/beurre*, *éthanol/rhum*. Nous identifions deux phénomènes expliquant la raison pour laquelle ces couples de mots sont mal repérés par l'AD. Le premier est que leurs membres ne sont pas forcément employés comme des substances dans le corpus. *Rhum*, par exemple, apparaît comme un produit fini et non pas comme un ingrédient (sauf dans le contexte *baba*_À). Le second est que, même dans les – rares – cas où les deux mots sont employés comme des composants, ils n'entrent pas forcément dans la composition du même type d'objets : pour le couple *carbone/diamant*, les contextes comme *collier*_DE sont incompatibles avec *carbone*. Un couple comme *crème/beurre* fait exception à la règle. *Crème* et *beurre* ont été détectés comme voisins, ils partagent les contextes *mélanger*_OBJ, *incorporer*_OBJ, *verser*_OBJ, etc. Ces deux mots ont en commun qu'ils apparaissent comme des ingrédients de cuisine. Dans le cas de *carbone/diamant*, les contextes se recoupent moins dans la mesure où on a affaire, d'un côté, à un élément chimique et, de l'autre, à un minéral. Cette différence sémantique semble suffisamment importante pour qu'elle soit perceptible au niveau de leurs distributions respectives et que ces deux mots ne soient donc pas repérés comme des voisins.

Les couples dont les deux membres appartiennent à la classe INSTRUMENTATION sont repérés par les voisins à hauteur de 41,5 %. La notion d'*instrument* est à prendre au sens large, et les couples appartenant à cette classe expriment une relation où un élément fait partie d'un dispositif ou un système de plus grande ampleur : *écran/ordinateur*, *pédale/bicyclette*, *pneu/autobus*. Dans la plupart des cas, les distributions entre les deux mots sont trop éloignées pour que l'analyse permette de les rapprocher. Par exemple, les contextes dans lesquels apparaît le méronyme *réservoir* (*volume*_DE, *servir*_DE, *placer*_OBJ) diffèrent complètement de ceux dans lesquels apparaissent ses holonymes *automobile* et *moto* (*accident*_DE, *conduire*_OBJ, *modèle*_DE). Le cas du méronyme *moteur*, en revanche, illustre une situation où la distribution du méronyme et de l'holonyme se recoupent : les 7 paires dans lesquelles il prend place sont toutes repérées par les voisins. Il apparaît en position méronyme de *avion*, *bateau*, *machine*, *navire*, *train*, *véhicule* et *voiture*. L'analyse des contextes communs fait apparaître une certaine symbiose entre le moteur et la machine qu'il équipe, dans la mesure où ils partagent un éventail de contextes relativement étendu comme *panne*_DE, *bruit*_DE, *consommer_suj*, *puissance*_DE, *se arrêter*_SUJ, *fonctionner*_SUJ, etc. On pourrait analyser certains de ces contextes comme des cas de métonymie : le bruit produit

par l'avion est en fait le bruit du moteur, de même que la puissance de la voiture est celle de son moteur.

Les couples de mots appartenant tous deux à la classe LIVING_PART sont les couples homogènes les moins bien identifiés par l'AD. Ils relient deux parties du corps (corps humain, animal ou partie d'un végétal), dont l'une est elle-même une partie de l'autre : *chair/doigt*, *muscle/bras*, *peau/visage*. Une des raisons expliquant les différences de distribution parmi les parties du corps est que, dans la plupart des cas, l'on affaire à des sous-classes de parties du corps dont les fonctionnements diffèrent radicalement. Ainsi, le fait que les couples *nerf/jambe* ou *nerf/doigt* ne sont pas captés s'explique par le fait que *jambe* et *doigt* sont des membres du corps. Ils peuvent par conséquent apparaître en position objet de verbes comme *lever*, *croiser* ou *replier*, soit autant de contextes dans lesquels ne peut pas apparaître *nerf*.

Ainsi, nous pouvons conclure de l'analyse de ces trois types de couples que les catégories *substance*, *instrumentation* et *living_part* s'avèrent peu pertinentes du point de vue distributionnel. Elle sont constituées de mots dont les distributions sont particulièrement dissemblables. Ainsi, si on postule *a priori* l'existence d'une classe sémantique des parties du corps, l'analyse du corpus montre que les mots *jambe*, *bras*, *doigt*, etc. entrent en fait dans un paradigme différent de celui de *veine*, *nerf* ou *os*. Il y a donc un décalage entre les classes sémantiques que l'on pourrait dégager intuitivement et les classes distributionnelles qui émergent de l'analyse du texte.

5.2 Couples hétérogènes

Nous avons auparavant évoqué la tendance qu'a l'AD à faire émerger des rapprochements relevant de la similarité sémantique, c'est-à-dire des mots qui sont "le même genre de choses" (van der Plas, 2008). De ce fait, on pouvait s'attendre à ne pas trouver de couples hétérogènes parmi les voisins distributionnels. Les résultats montrent toutefois que certains couples de catégories différentes sont quasi-intégralement repérés par l'AD.

C'est notamment le cas des couples de mots dont le méronyme appartient à la classe des humains (SOUL) et l'holonyme à celle des groupes sociaux : *capitaine/marine*, *fils/famille*, *musicien/orchestre*. Ces couples sont repérés à 96,2 %. Cela s'explique par le fait que les mots de la classe SOCIAL_GROUP ont des distributions similaires à ceux de la classe SOUL. Ils partagent par exemple la propriété d'apparaître en position sujet des verbes d'actions. Ainsi, le couple *directeur/entreprise* a été rapproché sur la base de contextes comme *détenir*_SUJ, *conseiller*_OBJ ou *affirmer*_SUJ, qui sont clairement destinés à être employés avec des animés. Il en va de même pour *joueur* et *équipe*, qui ont été rapprochés *via* les contextes *se entraîner*_SUJ, *affronter*_SUJ ou *se qualifier*_SUJ. Nous avons ici aussi affaire à un fonctionnement de type métonymique, dans la mesure où l'ensemble est employé pour désigner les membres.

Les couples dont le méronyme est un bâtiment et l'holonyme un lieu – *château/canton*, *école/commune*, *immeuble/métropole* – sont également bien captés par l'AD (c'est le cas de 82,6 % d'entre eux). Cela peut s'expliquer par l'ambiguïté des noms de bâtiments, qui peuvent aussi bien être employés comme des noms de lieux. Ainsi, le recouvrement entre ces deux classes implique une certaine similarité au niveau des distributions de leurs membres.

À l'autre extrémité du spectre, on remarque que la catégorie LIVING_PART apparaît en position méronyme dans trois des configurations hétérogènes les moins bien repérées par l'AD. Cette classe est successivement associée à PLANT_LIFE (*pétale/marguerite*, *tige/rose*, *tronc/chêne*), SOUL (*bras/citoyen*, *doigt/bébé*, *main/professeur*) et FAUNA (*bec/canard*, *patte/chat*, *queue/loup*). Dans

les trois cas, le fait que les couples relevant de ces classes ne soient que peu repérés s'explique par le fait qu'ici, les *touts* sont des êtres animés, contrairement à leurs parties. La conséquence en est que leurs propriétés distributionnelles sont radicalement opposées à celles de leurs méronymes. Cela semble être un peu moins flagrant pour les végétaux (ce qui explique que les couples LIVING_PART/PLANT_LIFE sont mieux repérés que les couples LIVING_PART/SOUL et LIVING_PART/FAUNA). On est donc dans le cas attendu de mots relevant de sens différents et par conséquent dissemblables sur le plan distributionnel.

5.3 Conclusion

Dans le cadre de cette étude consacrée à l'acquisition de relations sémantiques par des techniques d'analyse distributionnelle, nous nous sommes concentrés sur le cas de la relation de méronymie. Nous avons adopté une méthode d'évaluation qualitative reposant sur l'annotation sémantique de couples de méronymes. Sur le plan méthodologique, cette étude a montré que la typologie habituellement utilisée pour décrire les différents types de relations méronymiques était peu adaptée pour catégoriser nos données. Une approche consistant à typer sémantiquement les couples de méronymes permet de mieux rendre compte de la diversité des relations qu'ils expriment. Sur le plan des résultats, nous avons montré que si la méronymie, considérée globalement, est repérée dans des proportions comparables à d'autres relations (environ 1/3 des méronymes de JDM sont détectés par le programme d'AD que nous avons utilisé), elle n'est pas repérée par l'AD de manière homogène : la nature sémantique des mots qui entrent dans la relation de méronymie constitue un facteur décisif pour leur détection par l'AD. Tout d'abord, nous avons constaté que l'AD privilégie le repérage des couples de méronymes dont les membres relèvent de la même classe sémantique. Ensuite, nous avons vu que certaines configurations étaient identifiées dans des proportions beaucoup plus fortes que d'autres. C'est le cas des paires associant deux lieux, deux événements ou deux structures, ou associant un humain et un groupe social ou un lieu et un bâtiment. D'autres relations méronymiques, comme celles impliquant les parties du corps, sont mal détectées par l'AD car elles mettent en jeu des termes qui ne fonctionnent pas de la même manière sur le plan distributionnel. Cette étude contribue donc à préciser les conditions d'application du critère distributionnel au repérage d'une relation sémantique donnée. À ce stade, elle laisse cependant ouverte la question de l'influence du corpus de test sur la prédominance de certaines configurations distributionnelles des résultats.

Références

BARONI, M. et LENCI, A. (2010). Distributional memory : A general framework for corpus-based semantics. *Computational Linguistics*, 36.

BARONI, M. et LENCI, A. (2011). How we BLESSed distributional semantic evaluation. *GEMS 2011*, pages 1–10.

BOURIGAULT, D. (2002). UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *In Actes de la 9econférence sur le Traitement Automatique de la Langue Naturelle*, pages 75–84, Nancy.

BOURIGAULT, D. (2007). *Un analyseur syntaxique opérationnel :* SYNTEX. Mémoire d'habilitation à diriger des recherches. Université Toulouse II – Le Mirail.

BOURIGAULT, D. et GALY, E. (2005). Analyse distributionnelle de corpus de langue générale et synonymie. *In 4esJournées de la linguistique de corpus*, Lorient.

CRUSE, D. A. (1986). *Lexical semantics*. Cambridge University Press.

FELLBAUM, C., éditeur (1998). *WordNet : an electronic lexical database*. MIT Press, Cambridge.

FERRET, O. (2010). Similarité sémantique et extraction de synonymes à partir de corpus. *In Actes de TALN 2010 Traitement Automatique des Langues Naturelles - TALN 2010*.

KEET, C. et ARTALE, A. (2008). Representing and reasoning over a taxonomy of part–whole relations. *Applied Ontology*, 3(1):91–110.

LAFOURCADE, M. (2007). Making people play for lexical acquisition. *In 7th Symposium on natural Language Processing*.

LIN, D. (1998). An information-theoretic definition of similarity. *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304.

MORLANE-HONDÈRE, F. et FABRE, C. (2010). L'antonymie observée avec des méthodes de TAL : une relation à la fois syntagmatique et paradigmatique ? *In Actes de TALN 2010 Traitement Automatique des Langues Naturelles - TALN 2010*.

MULLER, P. et LANGLAIS, P. (2011). Comparaison d'une approche miroir et d'une approche distributionnelle pour l'extraction de mots sémantiquement reliés. *In Traitement Automatique des Langues Naturelles (TALN), Montpellier*, volume 1, pages 235–246.

MURPHY, M. L. (2003). *Semantic Relations and the Lexicon*. Cambridge University Press, New York.

SAHLGREN, M. (2006). *The Word-Space Model : using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* Thèse de doctorat, Stockholm University.

TURNEY, P. D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. *In COLING*, pages 905–912.

VAN CAMPENHOUDT, M. (1996). Recherche d'équivalences et structuration des réseaux notionnels : le cas des relations méronymiques. *Terminology*, 3(1):53–83.

van der PLAS, L. (2008). *Automatic lexico-semantic acquisition for question answering*. Thèse de doctorat, Université de Groningen (Pays-bas).

WINSTON, M. E., CHAFFIN, R. et HERRMANN, D. (1987). A taxonomy of part-whole relations. *Cognitive Science*, 11(4):417–444.

Un critère de cohésion thématique fondé sur un graphe de cooccurrences

Clément de Groc[1, 2] Xavier Tannier[2, 3] Claude de Loupy[1]
(1) Syllabs, 15 rue Jean-Baptiste Berlier, 75013 Paris
(2) Univ. Paris-Sud, 91403 Orsay Cedex
(3) LIMSI-CNRS, B.P. 133, 91403 Orsay Cedex
cdegroc@limsi.fr, xtannier@limsi.fr, loupy@syllabs.com

RÉSUMÉ
Dans cet article, nous définissons un nouveau critère de cohésion thématique permettant de pondérer les termes d'un lexique thématique en fonction de leur pertinence. Le critère s'inspire des approches *Web as corpus* pour accumuler des connaissances exogènes sur un lexique. Ces connaissances sont ensuite modélisées sous forme de graphe et un algorithme de marche aléatoire est appliqué pour attribuer un score à chaque terme. Après avoir étudié les performances et la stabilité du critère proposé, nous l'évaluons sur une tâche d'aide à la création de lexiques bilingues.

ABSTRACT
Topical Cohesion using Graph Random Walks

In this article, we propose a novel metric to weight specialized lexicons terms according to their relevance to the underlying thematic. Our method is inspired by Web as corpus approaches and accumulates exogenous knowledge about a specialized lexicon from the web. Terms cooccurrences are modelled as a graph and a random walk algorithm is applied to compute terms relevance. Finally, we study the performance and stability of the metric and evaluate it in a bilingual lexicon creation context.

MOTS-CLÉS : Cohésion thématique, graphe de cooccurrences, marche aléatoire.

KEYWORDS: Thematic relevance, cooccurrence graph, random walk.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 183–195,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Les lexiques et les terminologies sont des éléments essentiels du traitement automatique des langues. Ils sont utilisés dans une grande variété de tâches, allant de la catégorisation de textes à l'analyse d'opinions. Dans cet article, nous nous intéressons plus particulièrement aux lexiques dits thématiques ou spécialisés, c'est-à-dire composés de termes pertinents pour un domaine particulier. La Table 1 présente un extrait de lexique thématique sur le domaine de l'astronomie.

soleil	étoile	rayon gamma	étoile à neutron	masse solaire	...
planète	disque d'accrétion	naine blanche	proto-étoile	pulsar	...
astronome	quasar	astronomie	trou noir	neutron	...

TABLE 1 – Extrait de lexique thématique sur l'astronomie

La construction manuelle de tels lexiques est une tâche laborieuse et coûteuse. C'est pourquoi l'utilisation du Web ou de traducteurs automatiques comme appui pour la création de lexiques et de corpus spécialisés est une idée maintenant largement répandue (Baroni et Bernardini, 2004; Groc *et al.*, 2011; Kilgarriff et Grefenstette, 2003; Wan, 2009). Bien que l'utilité de telles approches ne soit plus à démontrer, une étape de validation manuelle reste requise.

Dans cet article, nous proposons un nouveau critère de *cohésion thématique* permettant de pondérer les termes d'un lexique thématique en fonction de leur pertinence pour le thème. Nous utilisons le Web comme source de corpus spécialisés sur les termes d'un lexique thématique. Nous modélisons ensuite les cooccurrences entre les termes du lexique sous la forme d'un graphe orienté où les sommets sont les termes du lexique et les arcs dénotent la cooccurrence de ces termes. Ce graphe peut être perçu comme un graphe de recommandation où l'apparition de deux termes dans un même document signifie qu'ils se recommandent l'un l'autre. Cette observation nous amène naturellement à utiliser un algorithme de marche aléatoire (*random walk* (Cohen, 2010; Page *et al.*, 1999)) attribuant une pertinence globale à chaque sommet du graphe.

Ce critère de cohésion thématique peut avoir de multiples applications. Dans le cadre de lexiques spécialisés construits automatiquement (Baroni et Bernardini, 2004; Groc *et al.*, 2011), ordonner les éléments du lexique par leur score de cohésion peut réduire la charge de validation manuelle ou même limiter la dispersion au fil des itérations. Dans le cadre de la traduction assistée, une valeur de cohésion peut représenter un score de confiance utile pour le traducteur. C'est d'ailleurs par cette dernière application que nous choisissons d'évaluer notre critère dans cet article.

L'article présente tout d'abord brièvement les travaux en *Web as corpus* dont notre approche découle, ainsi que ceux centrés sur les graphes de cooccurences et les algorithmes de marche aléatoire (section 2). Dans une 3ème section, nous présentons le modèle de graphe et l'algorithme de marche aléatoire utilisés pour le calcul du critère de cohésion thématique. Nous évaluons ensuite ce dernier sur une tâche de filtrage de lexiques thématiques traduits automatiquement (section 4). Nous concluons enfin (section 5) en suggérant plusieurs pistes envisagées.

2 Travaux liés

L'utilisation du Web comme source de documents (Kilgarriff et Grefenstette, 2003) est une idée maintenant largement répandue. Pour accéder aux documents Web, deux approches sont couramment mises en œuvre : soumettre un ensemble de requêtes à un moteur de recherche (Baroni et Bernardini, 2004; Ghani *et al.*, 2005) ou parcourir directement le Web (*crawling*) (Baroni et Ueyama, 2006). Le parcours du Web permet une meilleure spécification du besoin mais nécessite un investissement important. De plus, les efforts des moteurs de recherche pour garantir des résultats de qualité doivent être reproduits (filtrage des pages de spam). Au contraire, l'utilisation d'un moteur de recherche grand public comme point d'entrée au Web offre un accès simple et peu coûteux pour la communauté de Traitement Automatique des Langues. Nous adoptons ici cette approche afin de constituer un ensemble de connaissances exogènes sur les lexiques thématiques fournis en entrée.

Dans cet article, nous définissons un critère basé sur les cooccurrences des termes d'une même thématique pour déterminer leur lien avec le thème. Ces travaux partagent donc certaines hypothèses avec les travaux en similarité sémantique et notamment les analyses distributionnelles (Pereira *et al.*, 1993; Baker et McCallum, 1998; Rajman *et al.*, 2000) ou la désambiguïsation sémantique *via* des réseaux de cooccurrences (Dorow et Widdows, 2003; Ferret, 2004). En effet, notre graphe de cooccurrences modélise explicitement les cooccurrences de premier ordre mais l'application d'un algorithme de propagation d'importance de type PageRank permet la prise en compte de cooccurrences d'ordres supérieurs.

Enfin, l'algorithme TextRank (Mihalcea et Tarau, 2004) est étroitement lié à nos travaux. Les auteurs modélisent la cooccurrence des mots dans une fenêtre de taille N sous forme de graphe non-orienté et appliquent un algorithme de marche aléatoire afin de détecter les mots-clés saillants. Nous nous démarquons cependant de ces travaux en au moins deux points : nous considérons les cooccurrences au niveau du document (*snippet* dans nos évaluations) et modélisons ces dernières par un graphe orienté (plus de détails en Section 3).

3 Un critère de cohésion thématique

Étant donné un lexique thématique $\mathscr{L}_T$ composé de N termes, $\mathscr{L}_T = (t_1, t_2, \ldots, t_N)$, nous voulons calculer un vecteur de poids $\mathbf{w}_{\mathscr{L}_T} = (w_1, w_2, \ldots, w_N)$ où chaque poids w_i mesure la pertinence du terme t_i pour la thématique T.

3.1 Recueil de connaissances exogènes

Dans ces travaux, nous adoptons une approche *Web as corpus*, qui nous permet de créer rapidement des corpus spécialisés en nous appuyant sur un moteur de recherche généraliste. Nous proposons dès lors de constituer, pour chaque terme t_i, un corpus C_i correspondant au M meilleurs résultats renvoyés par un moteur de recherche pour la requête `"<t_i>"`.

L'unité d'information que nous considérons dans le cadre de cet article est le *snippet*, le court extrait de page Web renvoyé par le moteur de recherche. En effet, si prendre en compte le document entier permettrait en théorie de bénéficier d'un contexte plus large et plus riche, cela

pose surtout de nombreux problèmes. D'une part, télécharger les documents renvoyés par le moteur de recherche rallonge considérablement le temps de calcul. D'autre part, il est ensuite indispensable d'opérer un nettoyage des pages Web, et en particulier de supprimer les menus, les publicités ou les balises HTML, dans le but de ne conserver que le minimum de contenu non informationnel [1]. L'évaluation finale dépend donc beaucoup de la qualité de ce nettoyage, ce qui la rend plus difficilement interprétable. Enfin, le caractère local des *snippets* peut permettre de réduire le bruit pouvant apparaître dans les pages Web.

Nous avons utilisé le moteur de recherche Bing [2] comme source de *snippets*. Ces derniers sont composés de portions de textes de 155 caractères en moyenne issus du corps des pages Web et contenant les termes de la requête.

3.2 Cohésion thématique et graphe de cooccurrences

Étant donné un lexique thématique $\mathscr{L}_T$, nous proposons une première définition de notre critère comme suit : le poids w_i d'un terme t_i est égal au nombre de termes du lexique (t_i exclu) cooccurrant avec t_i dans le corpus C_i. Plus formellement, le poids w_i d'un terme t_i est défini par :

$$w_i = \sum_{t_j \in \mathscr{L}_T^\dagger} n_{t_j, C_i} \tag{1}$$

où n_{t_j,C_i} est le nombre d'occurrences du terme t_j dans l'ensemble des documents du corpus C_i et $\mathscr{L}_T^\dagger = \mathscr{L}_T \setminus \{t_i\}$, c'est-à-dire l'ensemble des termes du lexique $\mathscr{L}_T$, t_i exclu.

Cette même définition peut être modélisée sous la forme d'un graphe (Figure 1). Soit un graphe orienté $G =< V, E >$, où V est l'ensemble des sommets ($V = \mathscr{L}_T$) et E l'ensemble des arcs. Chaque arc $e(t_i, t_j)$ symbolise l'apparition du terme t_i dans le corpus C_j de t_j. Les arcs sont pondérés en fonction du nombre d'occurrences du terme t_i dans C_j. Notre approche *Web as corpus* nous différencie des précédents travaux (Mihalcea et Tarau, 2004) visant à modéliser les cooccurrences sous forme de graphe non-orienté : en effet, pour deux termes t_i et t_j et leurs corpus associés C_i et C_j, l'apparition du terme t_i dans le corpus C_j constitue un indice du "vote" de t_i pour t_j. Cependant, cette relation n'est pas symétrique puisque les corpus C_i et C_j sont distincts. En conséquence, nous optons pour un modèle de graphe orienté.

Le poids d'un terme tel que défini par l'équation 1 est alors équivalent au degré entrant de ce terme dans le graphe, c'est-à-dire la somme des poids des arcs entrants.

Cette nouvelle modélisation graphique nous amène à intégrer les poids du voisinage entrant d'un sommet dans le calcul du poids de celui-ci. Nous rectifions alors la première définition de notre critère et proposons la définition suivante : le poids w_i d'un terme t_i est égal à la somme des poids des termes du lexique cooccurrant avec t_i dans le corpus C_i. De plus, nous normalisons cette somme afin que le poids d'un terme soit réparti entre tous les termes auxquels il est lié.

1. Ce problème est d'ailleurs un thème de recherche à part entière fédéré par la campagne d'évaluation CLEANEVAL (`http://cleaneval.sigwac.org.uk`).
2. `http://www.bing.com`

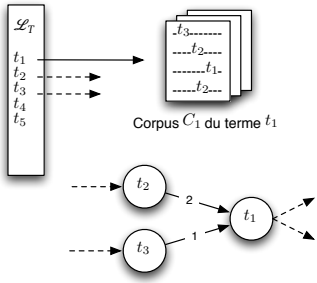


FIGURE 1 – Modélisation des cooccurrences sous forme de graphe orienté

Formellement, nous définissons le poids w_i d'un terme t_i comme :

$$w_i = \frac{\sum_{t_j \in \mathscr{L}_T^\dagger} n_{t_j,C_i} \cdot w_j}{\sum_{t_j \in \mathscr{L}_T^\dagger} n_{t_j,C_i}} \tag{2}$$

Cette nouvelle définition est "récursive" dans le sens où la pertinence d'un terme du lexique est définie en fonction de la pertinence des autres termes du lexique apparaissant dans son corpus. Il est ainsi possible de voir la pertinence d'un terme t_i défini en fonction de la pertinence d'un terme t_j, elle-même défini en fonction la pertinence de t_i. D'un point de vue graphique, ce phénomène se traduit alors simplement par un cycle dans le graphe de cooccurrences.

L'équation 2 est en fait proche de l'algorithme de marche aléatoire PageRank (Page *et al.*, 1999) et peut être résolue par un algorithme itératif sous certaines conditions. Cette version naïve de l'algorithme pose cependant deux problèmes :

1. L'algorithme ne gère pas correctement les sommets sans arcs sortant (appelés "dangling nodes" ou "rank sink" dans la littérature) : il n'est pas souhaitable qu'un terme ne renvoyant que des documents extérieurs à la thématique obtienne un poids important. Par exemple, si un terme de notre lexique $\mathscr{L}_T$ est un mot outil ("et"), il possèdera de nombreux liens entrant mais potentiellement aucun lien sortant : il accumulera itérativement un poids important.
2. La convergence vers une unique solution n'est pas garantie pour notre graphe (Langville et Meyer, 2005; Farahat *et al.*, 2006).

Pour résoudre le premier problème, nous ajoutons un lien des sommets sans arcs sortant vers un sommet virtuel et un lien de ce sommet virtuel vers tous les sommets du graphe. Le poids des sommets sans arcs sortant est ainsi redistribué à tous les sommets du graphe. Concernant le second problème, nous appliquons la solution de Page et Brin (Page *et al.*, 1999) et ajoutons une probabilité de téléportation uniforme à chaque itération de l'algorithme ce qui garantit la forte connectivité du graphe et la convergence vers une solution unique (équation 3).

$$w_{i,n+1} = \frac{(1-\alpha)}{N} + \alpha \cdot \frac{\sum_{t_j \in \mathscr{L}_T^{\dagger}} n_{t_j,C_i} \cdot w_{j,n}}{\sum_{t_j \in \mathscr{L}_T^{\dagger}} n_{t_j,C_i}} \tag{3}$$

où N est le nombre de sommets du graphe (c'est à dire le nombre de termes du lexique) et α est un facteur d'amortissement traditionnellement fixé à 0.85 (Page *et al.*, 1999; Mihalcea et Tarau, 2004). Nous utilisons également cette valeur de α pour nos expériences.

L'algorithme 1 récapitule l'intégralité du calcul du critère de cohésion thématique.

Algorithme 1 Critère de cohésion thématique

Entrées : $\mathscr{L}_T$: termes $t_i, i \in [1, N]$
M : nombre de documents téléchargés par requête
α : facteur d'amortissement

// Téléchargement du corpus
Pour tout terme $t_i \in \mathscr{L}_T$ **faire**
  Soumettre t_i à un moteur de recherche
  Télécharger M documents comme corpus C_i
Fin Pour

// Initialisation
Pour tout terme $t_i \in \mathscr{L}_T$ **faire**
  $w_{i,1} = 1/N$
Fin Pour

// Procédure itérative de calcul des poids
$n = 1$
Tant que (non convergence) **faire**
  Pour tout terme $t_i \in \mathscr{L}_T$ **faire**
    $w_{i,n+1} = \frac{(1-\alpha)}{N} + \alpha \cdot \frac{\sum_{t_j \in \mathscr{L}_T^{\dagger}} n_{t_j,C_i} \cdot w_{j,n}}{\sum_{t_j \in \mathscr{L}_T^{\dagger}} n_{t_j,C_i}}$
  Fin Pour
  Normalisation des poids : $\sum_i w_{i,n+1} = 1$
  $n = n + 1$
Fin Tant que
Retourner w_n

Pour améliorer la correspondance entre les lexiques et les documents issus du Web, une série de normalisations supplémentaires est appliquée : conversion des termes en minuscules, racinisation (*stemming*) et normalisation des caractères unicode (accents, ...).

Notons que le critère proposé traite les phénomènes d'ambiguïtés graphiques de la langue (homographie) de la façon souhaitée. Par exemple, si le terme "jaguar" est soumis à un moteur de recherche actuel, il est fort probable que ce dernier renvoie des résultats diversifiés à propos de l'animal mais également de la marque de voiture, de la console de jeu ou du système d'exploitation MacOS. Le nombre de cooccurrences avec les termes du lexique sera donc plus limité, conduisant à un score plus faible, soulignant ainsi qu'un terme ambigüe contribue moins à la cohésion thématique.

4 Évaluation

4.1 Tâche

Dans cet article, nous évaluons l'apport de notre critère pour l'aide à la création de lexiques thématiques bilingues à partir de lexiques monolingues. Comme mentionné dans l'introduction, nous envisageons de nombreuses applications pour le critère proposé dont notamment le *boots-trapping* de lexiques thématiques monolingues. Cependant, la tâche de création de lexiques thématiques bilingues, claire et facilement reproductible, fournit une évaluation objective de notre critère de cohésion.

Partant de lexiques thématiques monolingues, une approche commune pour la création de lexiques thématiques bilingues est d'utiliser un outil de traduction automatique en ligne tel que Google Translate [3]. Cependant, ces outils ne permettent pas d'intégrer une notion de *contexte thématique* dans le processus de traduction simplement. Ainsi, le terme simple "avocat" non intégré à une phrase, par exemple, sera traduit par ces outils "avocado" ou "lawyer" en anglais, indifféremment du fait qu'il appartient à un lexique juridique ou culinaire. Une validation manuelle laborieuse est donc nécessaire pour supprimer les traductions erronées.

Nous proposons d'appliquer notre critère de cohésion thématique aux lexiques traduits, attribuant ainsi à chaque traduction un score de confiance. Le tri des lexiques en fonction de ce score permet alors de réduire le temps nécessaire à leur validation.

4.2 Données

Nous avons utilisé trois lexiques bilingues français/anglais spécialisés sur trois thèmes différents :
– Astronomie (*The Astronomy Thesaurus* [4]) ;
– Médical (*Unified Medical Language System* - UMLS [5]) ;
– Statistiques (*International Statistical Institute* [6]).
Un exemple de termes issus de chaque lexique bilingue est donné Table 2.

3. http://translate.google.com
4. http://msowww.anu.edu.au/library/thesaurus/
5. http://www.nlm.nih.gov/research/umls/
6. http://isi.cbs.nl/glossary/

Astronomie		Statistiques	
Anglais	*Français*	*Anglais*	*Français*
afterglow	rémanence	Birnbaum's inequality	inégalité de Birnbaum
celestial coordinates	coordonnee celeste	geometric mean	moyenne géométrique
asteroids	asteroide	K-test	test K de Mann
dwarf stars	etoile naine	invariant	invariant
bow shocks	onde de choc en form d'arc	cross spectrum	spectre croisé

Médical	
Anglais	*Français*
wandering spleen	rate flottante
dimethoxyphenylethylamine	diméthoxyphényléthylamine
wolman disease	maladie de wolman
antimalarials	antipaludiques
optical illusions	illusions optiques

TABLE 2 – Extrait des lexiques thématiques bilingues utilisés pour l'évaluation

Une série de traitements a été appliquée à chaque lexique dans le but d'en améliorer la qualité ou l'utilisation pour notre évaluation. Ainsi, nous avons supprimé les termes apparaissant entre crochets ou parenthèses dans les lexiques Astronomie et Statistiques. Le lexique Médical présentant des termes trop ambigus pour être nettoyés automatiquement (par exemple *3-pyridinecarboxylic acid, 1,4-dihydro-2,6-dimethyl-5-nitro-4-(2-(trifluoromethyl)phenyl)-, methyl ester*), nous avons simplement supprimé les termes contenant une parenthèse ou une virgule.

Nous avons ensuite traité le cas des traductions multiples de la manière suivante : lorsqu'un terme de la langue source possédait plusieurs traductions dans la langue cible, nous n'avons conservé que le terme le plus proche (au sens de la distance de Damerau-Levenshtein (Damerau, 1964; Levenshtein, 1966)) de la traduction automatique [7]. Ainsi dans l'exemple "Afterglow" ⇒ "Postluminescence ou Remanence" issu du lexique "Astronomie", nous n'avons conservé que le terme "Remanence" car il est le plus proche de la traduction automatique trouvée : "rémanence".

Le lexique UMLS comprenant plus de 19 000 termes, nous avons choisi de ne travailler que sur un échantillon de ce dernier. Nous avons donc tiré aléatoirement deux séries de 2 000 termes que nous désignons comme lexiques Médical-1 et Médical-2.

Nous obtenons enfin les lexiques suivants :
– Astronomie (2 940 termes) ;
– Statistiques (2 752 termes) ;
– Médical-1 (2 000 termes) ;
– Médical-2 (2 000 termes).

4.3 Méthode

Nous traduisons chaque lexique thématique d'une langue source vers une langue cible (par exemple Astronomie fr → en ou Astronomie en → fr) à l'aide du moteur de traduction Google

7. voir section 4.3 pour la méthode de traduction automatique

Translate. Le lexique résultant est alors pondéré avec le critère de cohésion thématique puis ordonné et comparé avec la référence.

Le choix de Google Translate est justifié par le fait que ce moteur propose une très large couverture, ce qui en fait un candidat idéal pour traiter nos lexiques spécialisés. De plus, les récentes évaluations du NIST ont montré que l'outil de Google propose des performances état de l'art quant à la qualité des traductions produites (NIST, 2005, 2008).

La comparaison entre les termes traduits et les termes des lexiques originaux est réalisée à l'aide d'une mesure ad hoc incluant la suppression des déterminants en début de terme ("le bleu de bromothymol" ⇒ "bleu de bromothymol") et une distance d'édition de Damerau-Levenshtein (Damerau, 1964; Levenshtein, 1966). Nous avons considéré un terme comme valide s'il est au plus à une distance d'édition de 1 du terme de référence, autorisant ainsi une légère marge d'erreur due au moteur de traduction ou à la référence (singuliers transformés en pluriels, espace remplacé par un tiret, . . .).

La mesure de précision moyenne non-interpolée (*uninterpolated average precision* - UAP (Manning et Schütze, 1999)) est employée pour évaluer la validité de l'ordre des termes traduits.

4.4 Évaluation du critère

Nous évaluons notre critère *Cohésion-RW* comparativement à une baseline *Hasard*, obtenue par le simple tri aléatoire de la liste de traduction, et à la première version du critère *Cohésion-DEG* (équation 1). La baseline *Hasard* est obtenue en calculant une macro-moyenne sur dix tris aléatoires successifs. Nous fixons le nombre de *snippets* téléchargés pour chaque requête (la valeur M de l'algorithme 1) à 200 documents. Les résultats sont présentés à la Table 3.

Thème		Hasard	Cohésion-DEG	Cohésion-RW
Astronomie	en → fr	0.429	0.494	**0.512**
	fr → en	0.553	0.664	**0.678**
Statistiques	en → fr	0.382	0.663	**0.711**
	fr → en	0.488	0.667	**0.705**
Médical-1	en → fr	0.530	0.683	**0.735**
	fr → en	0.620	0.707	**0.718**
Médical-2	en → fr	0.522	0.662	**0.699**
	fr → en	0.638	0.739	**0.750**

TABLE 3 – Précision moyenne non-interpolée (UAP) pour le classement des termes des lexiques traduits.

Nous constatons que l'algorithme de marche aléatoire fournit les meilleurs résultats avec gain sur la baseline *Hasard* allant de 15,8 % (Médical-1 fr → en) à 86,1 % (Statistiques en → fr). La baseline fournit une idée de la qualité des traductions produites par le moteur de traduction. Ainsi, il semble que les lexiques Médical-1 et Médical-2 soient les mieux traduits. Au contraire le lexique Statistiques semble être le plus difficile à traduire. Le coefficient de corrélation de Pearson entre la précision du Hasard et le gain obtenu vaut -0,61 ce qui semble signifier qu'ils sont fortement corrélés négativement : plus la traduction est de bonne qualité et plus le gain est

Thème		50	100	150	200
Astronomie	en → fr	0.500	0.507	0.507	**0.512**
	fr → en	0.672	0.676	**0.680**	0.678
Statistiques	en → fr	0.666	0.695	0.704	**0.711**
	fr → en	0.678	0.691	0.702	**0.705**
Médical-1	en → fr	0.710	0.719	0.726	**0.735**
	fr → en	0.677	0.693	0.706	**0.718**
Médical-2	en → fr	0.672	0.678	0.687	**0.699**
	fr → en	0.702	0.723	0.735	**0.750**

TABLE 4 – Précision moyenne non-interpolée (UAP) pour le classement des termes des lexiques traduits avec le critère Cohésion-RW pour différentes valeurs de NB_DOCS.

faible. Cependant ce résultat n'est pas statistiquement significatif (la *p-value* vaut 0,106).

Nous évaluons ensuite l'influence du nombre de *snippets* téléchargés par requête sur les résultats du critère proposé (Table 4). Nous constatons que la précision moyenne augmente avec le nombre de documents téléchargés. Cela est probablement dû au fait que la qualité du graphe de cooccurrences augmente avec le nombre de documents et que la précision moyenne en est directement impactée.

4.5 Stabilité du critère

Le poids d'un terme t_i est défini en fonction des cooccurrences du terme t_i avec les autres termes du lexique (équation 3). Il semble donc légitime de s'interroger quant à la stabilité des poids des termes en fonction de la taille du lexique. La somme des poids étant égale à 1, le poids absolu de chaque terme est donc lié à la taille du graphe, il va diminuer avec l'augmentation du nombre total de termes. La question est donc de savoir ce qu'il en est du poids relatif, c'est-à-dire du rang des termes en fonction de la taille des lexiques.

Pour évaluer l'évolution des rangs des termes, nous avons de nouveau utilisé les lexiques traduits. Pour chaque lexique, nous avons sélectionné aléatoirement 20 termes, puis avons augmenté itérativement la taille du lexique de 20 termes. Chaque lexique (de 20, 40, 60, ... termes) a ensuite été ordonné par le critère de cohésion thématique. Enfin, le coefficient de corrélation de Spearman a été employé pour mesurer l'évolution des rangs des termes entre lexiques successifs (20-40, 40-60, ...).

Afin de réduire l'influence du hasard, nous avons répété la procédure décrite précédemment 10 fois et avons calculé une macro-moyenne des coefficients de corrélation. Dans un soucis de clarté, nous ne présentons que les résultats sur les lexiques anglais (Figure 2). Toutefois, les résultats obtenus sur les lexiques français sont équivalents.

Nous constatons que le coefficient de corrélation augmente pour tous les lexiques au fur et à mesure que la taille des lexiques augmente. Déjà forte avec une corrélation de plus de 0,70, le coefficient de corrélation s'approche du maximum à partir de 300 termes. Autrement dit, passée cette limite, l'ajout de nouveaux termes ne modifie presque pas l'ordre déjà établi entre les autres termes, ce qui nous permet de conclure que la mesure est stable à partir de ce seuil.

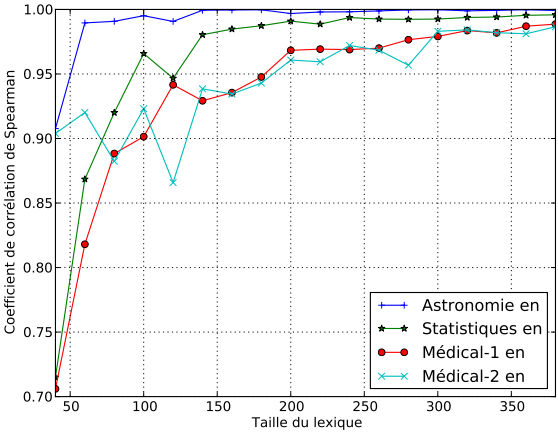


FIGURE 2 – Étude de l'évolution des rangs des termes en fonction de la taille des lexiques

5 Conclusion

Dans cet article, nous avons présenté un nouveau critère de cohésion thématique fondé sur un graphe de cooccurrences et un algorithme de marche aléatoire.

Nous avons évalué ce critère par une tâche d'aide à la création de lexiques bilingues car il s'agit d'une tâche claire, facilement reproductible et évaluable de façon objective. Cependant, comme nous l'avons indiqué, les applications possibles sont diverses, que ce soit pour réduire la charge de validation manuelle ou pour mieux sélectionner les termes automatiquement pour de la recherche d'information, de la collecte de corpus ou la mise en œuvre de techniques de *bootstrapping*. Les résultats obtenus sont encourageants et montrent la pertinence de notre approche.

Nous prévoyons d'analyser plus en détail le comportement du critère proposé en évaluant, par exemple, sa robustesse à la présence de termes non-pertinents dans les lexiques thématiques. Nous comptons également évaluer l'apport de quelques annotations manuelles en intégrant ces annotations dans l'algorithme de marche aléatoire sous forme d'un vecteur de personnalisation (Haveliwala, 2003).

Remerciements

Nous voudrions remercier Pierre Zweigenbaum de nous avoir fourni les lexiques nécessaires à l'évaluation de notre méthode et l'International Statistical Institute de nous avoir autorisé à utiliser le glossaire de termes statistiques multilingue. Nous remercions également Javier Couto pour ses conseils avisés sur la première version de ce manuscrit ainsi que les relecteurs anonymes pour leurs remarques et conseils. Ce travail s'inscrit dans le cadre des projets METRICC (ANR-08-CORD-013) et TTC (FP7/2007-2013 GA n°248005).

Références

BAKER, L. et MCCALLUM, A. (1998). Distributional clustering of words for text classification. *In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 96–103.

BARONI, M. et BERNARDINI, S. (2004). BootCaT : Bootstrapping Corpora and Terms from the Web. *In Proceedings of the LREC 2004 conference*, pages 1313–1316.

BARONI, M. et UEYAMA, M. (2006). Building general-and special-purpose corpora by web crawling. *In Proceedings of the 13th NIJL international symposium, language corpora : Their compilation and application*, pages 31–40.

COHEN, W. W. (2010). *Graph Walks and Graphical Models*. Carnegie Mellon University, School of Computer Science, Machine Learning Dept.

DAMERAU, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

DOROW, B. et WIDDOWS, D. (2003). Discovering corpus-specific word senses. *In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2*, pages 79–82.

FARAHAT, A., LOFARO, T., MILLER, J., RAE, G. et WARD, L. (2006). Authority rankings from hits, pagerank, and salsa : Existence, uniqueness, and effect of initialization. *SIAM Journal on Scientific Computing*, 27(4):1181–1201.

FERRET, O. (2004). Discovering word senses from a network of lexical cooccurrences. *In Proceedings of the 20th international conference on Computational Linguistics*, pages 1326–1332.

GHANI, R., JONES, R. et MLADENIC, D. (2005). Building Minority Language Corpora by Learning to Generate Web Search Queries. *Knowl. Inf. Syst.*, 7(1):56–83.

GROC, C. d., TANNIER, X. et COUTO, J. (2011). GrawlTCQ : Terminology and Corpora Building by Ranking Simultaneously Terms , Queries and Documents using Graph Random Walks. *In Proceedings of the TextGraphs-6 Workshop, Association for Computational Linguistics*, pages 37–41.

HAVELIWALA, T. (2003). Topic-sensitive pagerank : A context-sensitive ranking algorithm for web search. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4):784–796.

KILGARRIFF, A. et GREFENSTETTE, G. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29(3):333–347.

LANGVILLE, A. et MEYER, C. (2005). A survey of eigenvector methods for web information retrieval. *SIAM review*, pages 135–161.

LEVENSHTEIN, V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707–710.

MANNING, C. et SCHÜTZE, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

MIHALCEA, R. et TARAU, P. (2004). TextRank bringing order into text. *In Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 404–411.

NIST (2005). Nist 2005 machine translation evaluation official results. `http://www.itl.nist.gov/iad/mig/tests/mt/2005/doc/mt05eval_official_results_release_20050801_v3.html`. [consulté le 23/01/2012].

NIST (2008). Nist 2008 open machine translation evaluation (mt08) - official evaluation results. `http://www.itl.nist.gov/iad/mig/tests/mt/2008/doc/mt08_official_results_v0.html`. [consulté le 23/01/2012].

PAGE, L., BRIN, S., MOTWANI, R. et WINOGRAD, T. (1999). The PageRank Citation Ranking : Bringing Order to the Web. Rapport technique, Stanford InfoLab.

PEREIRA, F., TISHBY, N. et LEE, L. (1993). Distributional clustering of english words. *In Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 183–190.

RAJMAN, M., BESANÇON, R. et CHAPPELIER, J. (2000). Le modèle dsir : Une approche à base de sémantique distributionnelle pour la recherche documentaire. *Traitement automatique des langues*, 41(2):549–578.

WAN, X. (2009). Co-training for cross-lingual sentiment classification. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 1*, pages 235–243.

Validation sur le Web de reformulations locales: application à la Wikipédia

Houda Bouamor Aurélien Max Gabriel Illouz Anne Vilnat
LIMSI-CNRS
Univ. Paris Sud 11
Orsay, France
prenom.nom@limsi.fr

RÉSUMÉ
Ce travail présente des expériences initiales en validation de paraphrases en contexte. Les révisions de Wikipédia nous servent de domaine d'évaluation : pour un énoncé ayant connu une courte révision dans l'encyclopédie, nous disposons d'un ensemble de réécritures possibles, parmi lesquelles nous cherchons à identifier celles qui correspondent à des paraphrases valides. Nous abordons ce problème comme une tâche de classification fondée sur des informations issues du Web, et parvenons à améliorer la performance de plusieurs techniques simples de référence.

ABSTRACT
Assisted rephrasing for Wikipedia contributors through Web-based validation

This works describes initial experiments on the validation of paraphrases in context. Wikipedia's revisions are used : we assume that a set of possible rewritings are available for a given phrase that has been rewritten in the encyclopedia's revision history, and we attempt to find the subset of those rewritings that can be considered as valid paraphrases. We tackle this problem as a classication task which we provide with features obtained from Web data. Our experiments show that our system improves performance over a set of simple baselines.

MOTS-CLÉS : paraphrase, Wikipédia, aide à la rédaction.

KEYWORDS: paraphrasing, Wikipedia, authoring aids.

1 Introduction

Il existe plusieurs scénarios dans lesquels il est souhaitable de pouvoir faire produire du texte par la machine. Ce problème a traditionnellement été abordé comme une tâche de génération de texte à partir de concepts. Toutefois, ces besoins s'appliquent parfois à des cas où un nouveau texte devrait être dérivé de certains textes existants, par exemple lorsqu'il s'agit de transformer un texte afin qu'ils aient certaines propriétés souhaitables pour un usage particulier (Zhao *et al.*, 2009). Par exemple, on peut souhaiter qu'un texte soit condensé (Cohn et Lapata, 2008), adapté à certains profils de lecteur (Zhu *et al.*, 2010), conforme à certaines normes spécifiques (Max, 2004), voire même simplement plus adapté pour des tâches de traitement automatique ultérieures.

Le mécanisme de réécriture de texte doit donc produire un texte dont le sens est compatible avec la définition de la tâche à accomplir, tout en garantissant que celui-ci demeure grammatical.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 197–210,
Grenoble, 4 au 8 juin 2012.

La complexité de la *génération texte-à-texte*, par opposition à la génération *concepts-à-texte*, provient essentiellement du fait que la correspondance sémantique entre deux textes est difficile à contrôler, car les réécritures mises en jeu sont très dépendantes du contexte. En effet, la grande diversité des techniques d'acquisition de paraphrases *sous-phrastiques* (Madnani et Dorr, 2010), la polysémie de ces unités linguistiques ainsi que les contraintes pragmatiques associées à leur substitution font qu'il est impossible de garantir que des paires de paraphrases candidates seront substituables quel que soit le contexte de réécriture. Ce problème a été déjà décrit au niveau lexical (Zhao *et al.*, 2007; McCarthy et Navigli, 2009) ; la validation automatique en contexte de reformulations de segments demeure une question fondamentale pour la réécriture de texte.

Dans ce travail, nous abordons le problème sous l'angle d'un *paraphrasage ciblé* [1], défini comme la réécriture d'un segment d'un énoncé. Bien que ce problème soit plus simple que la réécriture d'une phrase complète, son étude se justifie par la nécessite de bien comprendre ce niveau moins complexe avant d'aborder la réécriture d'unités plus étendues, ce qui en outre facilite la tâche complexe de l'évaluation.

Nous présentons ici un scénario en *révision interactive de textes* dans lequel des paraphrases sous-phrastiques doivent être proposées en tenant compte du contexte. Les paraphrases candidates considérées sont obtenues à partir d'un répertoire existant, et sont validées en contexte à l'aide d'informations obtenues sur le Web. Les expériences que nous avons menées ciblent plus particulièrement les contributeurs de l'encyclopédie Wikipédia dans leurs tâches de révision des articles. Nous avons pour cela utilisé un ensemble de segments ayant fait l'objet de réécritures dans l'historique des articles de Wikipédia, que nous substituons par des paraphrases connues à l'avance. Étant donné la grande variété de segments possibles et de leurs paraphrases, nous ne nous appuyons pas sur des modèles de substituabilité préétablis, mais nous les construisons *à la volée* à partir du Web.

Dans cet article, nous allons tout d'abord décrire la tâche de révision de texte sous forme de paraphrasage ciblé (section 2). Nous passerons ensuite en revue les principaux travaux précédents portant sur l'acquisition de paraphrases sous-phrastiques et décrirons les sources de connaissances que nous avons utilisées dans ce travail (section 3). Nous détaillerons ensuite notre méthode de calcul des modèles de substitution de segments en contexte exploitant des informations issues du Web (section 4). Les expériences menées pour valider les paraphrases contenues dans le répertoire existant et leurs résultats seront finalement présentés (section 5). Notre article se conclura par une discussion de ces résultats et une présentation des principales voies de recherche (6).

2 Paraphrasage ciblé pour la révision de texte

La reformulation d'un énoncé, ou d'un segment plus précis, est une activité importante en révision de texte. Certaines modifications locales ont ainsi vocation à améliorer sa qualité générale, en le rendant par exemple plus facile d'accès (Zhu *et al.*, 2010) ou en l'adaptant au niveau d'expertise de ses lecteurs (Deléger et Zweigenbaum, 2009). Les modifications de ce type, qui n'altèrent pas le sens des textes, incluent non seulement la synonymie lexicale mais également des transformations lexico-syntaxiques plus complexes.

1. Ce terme est utilisé par Resnik *et al.* (2010) pour décrire l'obtention (manuelle) de paraphrases pour des segments jugés difficiles à traduire.

On trouve notamment de telles reformulations dans les historiques de révision de textes, qui sont désormais disponibles en grandes quantités avec l'émergence de ressources collaboratives sur le Web telles que l'encyclopédie Wikipédia. L'historique des révisions des articles de cette ressource constitue en effet une source importante de phénomènes de *réécriture naturelle*. L'étude de Dutrey *et al.* (2011) a notamment montré que cet historique contient une variété importante de phénomènes de reformulation, dont de nombreuses paraphrases. Cette étude a également montré, au travers d'une tentative d'identification automatique à base de règles, les difficultés pour parvenir à une bonne couverture de l'ensemble des phénomènes paraphrastiques présents.

Peu de travaux, ont, à notre connaissance, porté sur l'utilisation du paraphrasage contextuel dans le cadre de l'aide à la rédaction. Max et Zock (2008) présentent une méthode proposant aux rédacteurs des paraphrases sous-phrastiques candidates pour les segments qu'ils souhaitent reformuler. L'approche utilisée pour la génération des paraphrases est fondée sur l'équivalence de traduction (Bannard et Callison-Burch, 2005). Les travaux de Bernstein *et al.* (2010) portent eux sur l'externalisation de diverses tâches d'édition de texte, dont la révision, *via* le *crowdsourcing*.

Par ailleurs, la réécriture d'un texte peut être destinée plus spécifiquement à une application automatique. Dans (Resnik *et al.*, 2010), des reformulations pour des segments jugés difficiles à traduire sont acquises *via le crowdsourcing* : des contributeurs monolingues de la langue source proposent ainsi des reformulations en contexte pour ces unités [2]. Les reformulations collectées sont ensuite utilisées en entrée dans un système de traduction automatique, qui peut ainsi bénéficier de la variété d'expressions pour produire de meilleures traductions (Schroeder *et al.*, 2009). Par exemple, le segment *une optique festive* dans *L'usage intervient alors dans* **une optique festive** peut être réécrit en : 1) *un cadre festif*, 2) *une perspective de fête*. Ces réécritures sont grammaticalement correctes et ont des significations raisonnablement proches de la formulation d'origine.

Outre la reformulation des segments de texte, la réécriture d'énoncés a aussi été à l'origine de plusieurs travaux (Barzilay et Lee, 2003; Quirk *et al.*, 2004; Zhao *et al.*, 2010; Ganitkevitch *et al.*, 2011). Cependant, celle-ci pose de nombreux autres défis, notamment au niveau de l'évaluation des reformulations produites. Le jugement par des humains devient alors encore plus complexe et n'autorise pas des distinctions fines ni des accords inter-annotateurs satisfaisants. La génération de paraphrases d'énoncés peut toutefois être évaluée indirectement dans le cadre de leur utilisation dans une application plus complexe. Par exemple, Madnani *et al.* (2008) parviennent à améliorer les performances d'un système de traduction automatique statistique en fournissant des paraphrases automatiques des traductions de référence lors de l'apprentissage des paramètres du système. Cependant, les améliorations observées n'indiquent pas clairement les liens avec la qualité des paraphrases utilisées.

Nous abordons dans ce travail la tâche plus modeste de paraphrasage sous-phrastique appliqué à la révision de texte. Afin d'éviter tout biais, nous utilisons des réécritures *écologiques* (que nous entendons ici comme : produites naturellement) extraites d'une mémoire de rédaction des articles de Wikipédia. Nous utilisons pour cela le corpus WiCoPaCo (Max et Wisniewski, 2010), qui contient de nombreux phénomènes de réécriture, dont de nombreuses instances de reformulations lexicales, syntaxiques et sémantiques (Dutrey *et al.*, 2011). Ce dernier type de reformulation est illustré dans l'exemple suivant, où le remplacement du segment *un mode d'expression* par sa paraphrase possible *une figure de rhétorique* permet de préciser et d'affiner le

2. Nous notons toutefois que les contributeurs ne reçoivent aucune indication directe de l'utilité des reformulations qu'ils proposent.

sens voulu par le contributeur initial :

L'antiphrase est ***[un mode d'expression → une figure de rhétorique]*** *consistant à dire le contraire de ce que l'on pense.*

Ce corpus est pertinent à plusieurs titres pour la tâche que nous visons. Tout d'abord, le fait qu'il contienne des réécritures obtenues hors du cadre d'expériences offre une source riche et intéressante d'unités textuelles réécrites en contexte. De plus, les instances de réécriture où le sens n'a pas été modifié offrent directement une paraphrase candidate qui peut être considérée comme *correcte*, donnée pouvant s'avérer utile pour l'apprentissage automatique du processus de validation en contexte.

3 Acquisition de paraphrases sous-phrastiques

La disponibilité grandissante de masses de données textuelles a rendu possible un grand nombre de travaux en acquisition et en génération de paraphrases (Madnani et Dorr, 2010). Les techniques proposées apparaissent néanmoins assez fortement liées aux types de ressources auxquelles elles s'appliquent. Les types de corpus utilisés pour sont principalement :

- des paires de paraphrases d'énoncés (**corpus monolingues parallèles**), qui permettent d'obtenir des paraphrases précises, mais en faible quantité (Barzilay et McKeown, 2001; Pang *et al.*, 2003; Cohn *et al.*, 2008; Bouamor *et al.*, 2011) ;
- des paires d'énoncés en relation de traduction (**corpus multilingues parallèles**), qui permettent de générer de nombreuses paraphrases candidates (Bannard et Callison-Burch, 2005; Kok et Brockett, 2010) ;
- des paires d'énoncés en relation partielle (**corpus monolingues parallèles**), qui permettent sur le principe d'acquérir de nombreuses paraphrases (Barzilay et Lee, 2003; Pasça et Dienes, 2005; Bhagat et Ravichandran, 2008; Deléger et Zweigenbaum, 2009).

Bien que la précision de ces techniques d'acquisition peut se mesurer sur la base d'une référence attendue portant sur une collection de paires d'énoncés (Cohn *et al.*, 2008), il est plus utile de pouvoir la mesurer au travers de la question de *substituabilité en contexte*, laquelle a déjà été abordée au niveau lexical (Connor et Roth, 2007; Zhao *et al.*, 2007) où elle a fait l'objet de campagnes d'évaluation (McCarthy et Navigli, 2009). Celle-ci pose des défis supplémentaires, dûs au fait que les segments sont plus rares que les mots en corpus.

4 Validation contextuelle sur le Web

4.1 Cadre d'évaluation

Le présent travail porte sur la tâche de validation automatique de paraphrases sous-phrastiques en contexte. Pour cela, nous avons eu recours à un répertoire existant de paires de paraphrases. Comme décrit plus haut, nous avons utilisé le corpus WiCoPaCo comme corpus de reformulations sous-phrastiques naturelles. La réécriture contenue dans cette ressource peut être utilisée comme paraphrase potentielle. Afin d'obtenir d'autres paraphrases candidates de différentes qualités, nous avons utilisé deux autres méthodes d'acquisition, qui fourniront des paraphrases aux

instances extraites de WiCoPaCo qui ne seront pas nécessairement substituables en contexte : *a*) une traduction automatique par pivot, et *b*) une acquisition manuelle de paraphrases.

La génération de paraphrases par traduction s'effectue simplement en traduisant automatiquement un segment dans une langue pivot, puis en le rétrotraduisant dans la langue d'origine, et en retenant la première hypothèse différente du segment d'origine. Si cette technique n'offre aucune garantie sur la qualité des résultats, elle est aisée à mettre en œuvre et produit des résultats variés. En outre, l'utilisation d'une langue pivot proche de la langue d'origine augmente la probabilité d'obtenir de bonnes paraphrases (ceci sera étudié lors de nos expériences, décrites dans la section 5).

Nous avons défini l'acquisition manuelle de paraphrases de la façon suivante : un corpus d'extraits de documents du Web contenant les segments à réécrire est tout d'abord constitué, en s'assurant que ce corpus ne contient pas de données provenant de Wikipédia. Pour chaque segment à réécrire dans ce corpus, des locuteurs natifs du français proposent *via* une interface web une réécriture possible. Ainsi, les contextes utilisés pour faire l'acquisition de paraphrases des segments sont possiblement différents de ceux, extraits de WiCoPaCo, sur lesquels portera l'évaluation : notre système de validation en contexte aura donc à considérer des paraphrases *potentiellement* valides [3] mais qui ne le sont pas dans le contexte d'une réécriture particulière.

Ces deux méthodes, dont la mise en œuvre est aisée, nous permettent de simuler la disponibilité d'un répertoire existant de paraphrases sous-phrastiques, qui nous servira pour l'évaluation de la performance de notre technique de validation en contexte.

4.2 Classification automatique de réécritures en contexte

Nous décrivons maintenant l'approche que nous proposons pour réaliser une validation de réécritures en contexte, fondée sur une classification binaire exploitant des modèles calculés à partir d'informations du Web. Le recours au Web semble indispensable : seule une telle échelle nous permet d'accéder à des exemple en nombre suffisants pour certains segments. En outre, il a été montré qu'un certain nombre d'applications de Traitement Automatique des Langues peuvent être améliorées grâce à l'exploitation de fréquences de *n*-grammes sur le Web (Lapata et Keller, 2005).

Considérant un ensemble de contextes de réécritures pour des segments ainsi qu'un répertoire existant contenant des paraphrases pour ces segments, notre tâche consiste à classer (i.e. *paraphrase* vs. *pas paraphrase*) chaque paraphrase possible pour chaque contexte original. Une instanciation concrète possible de cette tâche est la proposition de Max et Zock (2008), où de telles reformulations candidates sont présentées dans un ordre décroissant de pertinence à un utilisateur d'un éditeur de texte, et donc éventuellement lors de la révision d'un article de Wikipédia.

La tâche d'identification automatique de paraphrases a été déjà abordée par classification automatique dans des travaux précédents, en utilisant des modèles calculés sur des corpus collectés (Brockett et Dolan, 2005) et sur des documents issus du Web (Zhao *et al.*, 2007). Cependant, ces travaux se sont limités à l'identification de paraphrases lexicales (McCarthy et Navigli, 2009). Une difficulté importante est que certains mots sont absents ou très peu fréquents

3. On les suppose ici valides parce que obtenues par réécriture manuelle d'un segment dans un texte. Ceci repose cependant fortement sur la capacité de nos contributeurs natifs à bien réaliser la tâche demandée.

dans les index des moteurs de recherche, et *a fortiori* dans des corpus spécialisés, difficulté qui s'amplifie lorsque l'on considère des segments. [4]

De façon analogue aux travaux de Brockett et Dolan (2005), nous considérons l'identification de paraphrases comme une tâche de classification : étant donné un segment d'origine s dans le contexte d'une phrase p, nous cherchons à déterminer si une paraphrase candidate s' serait une paraphrase *grammaticale* de s dans le contexte de p. Nous avons abordé ce problème avec un classifieur de type séparateur à vaste marge (SVM) exploitant les traits décrits ci-dessous.

Distance d'édition Les approches les plus répandues en calcul de pertinence d'un document relativement à une requête exploitent des mesures de similarité de surface, qui peuvent dans certains cas être de bons indicateurs de proximité sémantique. Un coût de transformation entre chaînes de caractères peut par exemple être celui donné par la mesure TER (Snover *et al.*, 2010), initialement développée pour mesurer la similarité entre une hypothèse de traduction et une traduction de référence. Cette mesure se base sur des opérations d'édition (substitution, déplacement, insertion, suppression) plus informatives que les méthodes basées sur des intersections lexicales [5]. Nous effectuons en outre ce calcul sur les lemmes plutôt que sur les formes de surface, que nous avons obtenus à l'aide du TreeTagger (Schmid, 1994) [6]. Nous retenons donc le score suivant, calculé entre un segment d'origine seg_{orig} et une paraphrase seg_{para}, où la fonction Lem produit une forme lemmatisée de son argument :

$$h_{edit} = \text{TER}(Lem(seg_{orig}), Lem(seg_{para})) \quad (1)$$

Il convient de noter que, contrairement aux autres modèles, celui-ci ne dépend pas d'informations provenant du Web.

Score de modèle de langue La vraisemblance d'une phrase peut être un relativement bon indicateur de sa grammaticalité locale (Mutton, 2006). Les probabilités données par un modèle de langue peuvent désormais être obtenues à l'aide de comptes provenant du Web. Nous avons pour cela utilisé le Service Web N-gram de Microsoft (Wang *et al.*, 2010) dans sa déclinaison à des fins de recherche [7]. Afin de pouvoir utiliser correctement ce service sur des textes français, nous avons dû supprimer tous les diacritiques : un examen précis des paraphrases candidates classées a montré que cette transformation, bien qu'abérante, nous a permis d'obtenir des résultats cohérents. [8]

4. Nous faisons cependant l'hypothèse que des segments absents ou très peu fréquents sur le Web présentent un intérêt moindre pour la réécriture, et n'accordons donc pas pour cette étape de nos travaux d'attention particulière à ce problème. Il est toutefois possible d'argumenter que ces segments pourraient être *mal écrits* (par exemple, par un locuteur non natif, un apprenant, voire une machine) et donc possiblement non connus des moteurs de recherche, pour lesquels une assistance à la réécriture serait tout à fait pertinente. Cela représente néanmoins une problématique en soi.

5. Il faut noter que les opérations de racinisation et de correspondance sémantique utilisant WordNet n'ont pas été prises en compte car nos expériences portent sur le français.

6. Ce calcul de lemmatisation se fait, pour le segment original et sa paraphrase, dans le contexte de la substitution testée : il est toutefois possible que la lemmatisation produise des erreurs.

7. `http://research.microsoft.com/en-us/collaboration/focus/cs/web-ngram.aspx`

8. La description du service n'était pas très explicite lorsque nous l'avons utilisé : il semblerait que l'intention de son fournisseur était avant tout de proposer un service pour l'anglais.

Un simple score de modèle de langue pour un énoncé après réécriture n'est toutefois pas suffisant, car il ne tient pas compte de l'énoncé d'origine. Nous avons donc utilisé le rapport entre le score de modèle de langue de l'énoncé paraphrasé phr_{para} et le score de modèle de langue de l'énoncé d'origine phr_{orig}, normalisé par la longueur des énoncés (Onishi *et al.*, 2010) :

$$h_{ML} = \frac{ML(phr_{para})^{1/longueur(phr_{para})}}{ML(phr_{orig})^{1/longueur(phr_{orig})}} \tag{2}$$

Scores de similarité thématique hors contexte Les techniques mises en œuvre pour calculer une notion de *similarité* entre unités textuelles sont fréquemment fondées sur le calcul de représentations des contextes d'occurrences de ces unités sur lesquelles sont calculées des mesures de similarité. Nous avons suivi ce type d'approche pour mesurer une similarité thématique entre paraphrases entre profils de mots cooccurrents. Nous construisons tout d'abord des profils *hors contexte* de la manière suivante : un moteur de recherche est interrogé afin de récupérer les N premiers extraits de documents (*snippets*) pertinents pour le segment seg_{orig}. La fréquence des mots pleins présents dans ces extraits est alors calculée et est utilisée pour obtenir les valeurs de chaque dimension d'un vecteur de profil lexical T, dont la valeur pour un mot m est définie ainsi :

$$T_{orig}[m] = \frac{freq(seg_{orig}, m)}{freq(seg_{orig})} \tag{3}$$

Pour le calcul des fréquences, $freq(u)$ correspond au nombre d'extraits de documents retournés contenant l'unité u, et $freq(u, v)$ au nombre d'extraits de documents rapportés contenant les deux simultanément. Nous construisons de façon analogue un profil thématique pour chaque paraphrase possible seg_{para}, en se limitant aux dimensions du vecteur pour le segment d'origine :

$$T_{para}[m] = \frac{freq(seg_{para}, m)}{freq(seg_{para})} \tag{4}$$

Enfin, nous mesurons la similarité entre le profil du segment d'origine et de chacune de ses paraphrases possibles à l'aide du cosinus entre les vecteurs de leur profil thématique :

$$h_{them} = \frac{T_{orig} \cdot T_{para}}{||T_{orig}|| * ||T_{para}||} \tag{5}$$

Pour l'ensemble de nos expériences, nous avons utilisé le service Web Yahoo ! Search BOSS [9] pour obtenir le nombre de documents du Web indexés contenant une expression littérale (typiquement, un segment d'intérêt) ainsi que les extraits de documents à partir desquels nous construisons les vecteurs de profils thématiques. En supposant que la distribution des mots pleins cooccurrents n'est pas biaisée par l'ordre des résultats du moteur de recherche, notre modèle mesure donc un certain type de similarité thématique entre seg_{orig} et seg_{para}.

9. `http://developer.yahoo.com/search/boss/`

Scores d'un modèle thématique contextuel Nous définissons également un modèle thématique contextuel de la façon suivante : considérant $cont_{orig}$, constituée des deux sous-chaînes de phr_{orig} privée de seg_{orig}, nous construisons un vecteur de profil T^{cont} ayant pour dimension uniquement pour les mots pleins du contexte de la phrase où a lieu la réécriture. Les valeurs associées à chaque dimension correspondent à des rapports de fréquence obtenus comme précédemment par interrogation du moteur de recherche. La similarité thématique contextuelle utilisée est finalement définie par :

$$h^{cont}_{them} = \frac{T^{cont}_{orig} \cdot T^{cont}_{para}}{||T^{cont}_{orig}|| * ||T^{cont}_{para}||} \quad (6)$$

5 Expériences et résultats

Dans cette section, nous détaillons les expériences que nous avons menées afin d'évaluer les performances de l'approche de validation automatique de paraphrases en contexte.

5.1 Description des données utilisées

Nous avons extrait aléatoirement 150 énoncés en français du corpus WiCoPaCo et leur réécriture pour des exemples annotés comme "paraphrases" lors d'une annotation manuelle réalisée par une étudiante en linguistique francophone. Un sous-ensemble de 100 énoncés a été utilisé comme corpus d'apprentissage, les 50 énoncés restants ayant servi pour l'évaluation. Les segments originaux ainsi que leur paraphrase dans le corpus d'évaluation sont décrits dans la figure 1.

taille segment	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
# segments originaux	0	3	29	8	6	2	2	0
# paraphrases	39	64	74	36	21	10	5	1

FIGURE 1 – Répartition du nombre de segments par taille (nombre de tokens) dans le corpus d'évaluation

Nous disposons finalement de 5 paraphrases par segment d'origine :

- **WICOPACO** : la paraphrase associée au segment dans le corpus WiCoPaCo ;
- **HUMAIN** : deux paraphrases candidates proposées par des contributeurs humains pour d'autres contextes issus du Web ;
- **PIVOT$_{ES}$** and **PIVOT$_{ZH}$** : deux paraphrases candidates obtenues par traduction par pivot. Nous avons utilisé le système de traduction automatique sur le Web GOOGLE TRANSLATE [10], avec une langue proche du français comme pivot (l'espagnol), et une autre plus distante (chinois).

La partie évaluation de nos expériences a impliqué 4 évaluateurs humains [11], tous francophones. Ceux-ci ont participé à la collecte manuelle des paraphrases (HUMAIN) pour la moitié du corpus d'apprentissage et d'évaluation. Afin d'évaluer le caractère approprié de l'utilisation des

10. http://translate.google.com
11. La personne ayant réalisé l'annotation originelle de WiCoPaCo n'a pas pris part à ce nouveau travail.

paraphrases issues des paraphrases collectées dans les contextes de réécriture sélectionnés, les phrases d'origine et leurs différentes paraphrases ont été présentées dans un ordre aléatoire aux deux évaluateurs ayant initialement travaillé sur l'autre moitié des corpus. Une interface sur le Web, illustrée sur la figure 2, permet alors aux évaluateurs d'indiquer quelles substitutions sont acceptables, à la fois au niveau de la conservation du sens et de la grammaticalité du nouvel énoncé.

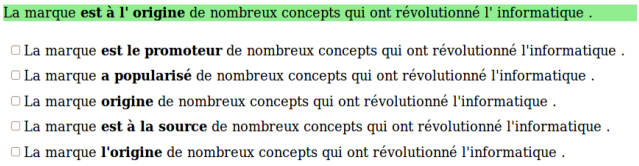


FIGURE 2 – Exemple d'une phrase d'origine (sur fond vert) et de ses 5 paraphrases candidates (présentées dans un ordre aléatoire). Le segment en gras dans la phrase d'origine, *est à l'origine*, est ici paraphrasé par *est le promoteur* , *a popularisé* , *origine* , *est à la source* et *l'origine*.

La valeur d'accord inter-annotateur [12] sur l'ensemble des énoncés annotés est de $\kappa = 0,65$, ce qui correspond à un accord *fort* selon les grilles de Landis et Koch (1977). Nous pensons que le fait d'aborder tout d'abord des tâches relativement certaines du point de vue de l'accord entre humains comme celle-ci est nécessaire avant de s'attaquer à des problèmes plus complexes, tels que l'identification de paraphrases d'énoncés ou encore l'identification d'implications textuelles.

Notre technique de validation étant très dépendante de la fréquence des segments considérés sur le Web, nous avons décidé dans ces premières expériences de ne conserver que les segments ayant une fréquence minimale de 10 occurrences pour le moteur de recherche utilisé. Le nombre d'exemples du corpus d'apprentissage a ainsi été réduit de 750(=150*5) à 434, et celui du corpus d'évaluation de 250(=50*5) à 215. L'atténuation de cette limitation devra bien évidemment faire partie de la suite de nos travaux.

Nous détaillerons nos résultats pour les 3 conditions suivantes :

- **Possibles** : les exemples annotés comme "paraphrases" par au moins l'un des juges sont utilisés : l'ensemble d'évaluation correspondant comprend 116 cas positifs et 99 cas négatifs.
- **Sûres** : les exemples que les deux juges n'ont pas annotés comme "paraphrases" ou "non paraphrases" ne sont pas retenus : l'ensemble d'évaluation correspondant comprend 76 cas positifs et 139 cas négatifs.
- **Sûres++** : seuls les exemples pour lesquels les deux juges proposent la même annotation sont retenus. Ceci réduit nos ensembles d'apprentissage et d'évaluation à respectivement 287 et 175 exemples, ce qui ne permet pas une comparaison directe avec les deux autres conditions. L'ensemble d'évaluation correspondant comprend 76 cas positifs et 99 cas négatifs.

12. Nous avons utilisé le logiciel R (`http://www.r-project.org`) pour calculer la valeur de κ de Cohen. Cette valeur est calculée sur l'ensemble des données, chaque moitié étant annotée par les deux mêmes annotateurs.

5.2 Techniques de référence

Nous présentons ici brièvement les techniques de référence auxquelles nous comparerons notre système.

Fréquence sur le Web Les deux premières techniques sont fondées sur des calculs de fréquences sur le Web. La première, ML_WEB considère un énoncé comme paraphrase d'un énoncé d'origine si son score de modèle de langue issu du Web est plus élevé que celui de l'énoncé d'origine. La deuxième technique, ML_FRONTIÈRES, considère qu'un énoncé est paraphrase d'un énoncé d'origine si la fréquence sur le Web des bigrammes traversant les frontières gauche et droite après substitution est supérieure à 10.

Conservation de dépendances syntaxiques Lors de la réécriture d'une partie d'un énoncé, la conservation des dépendances syntaxiques entre un segment d'origine et son contexte d'une part, et sa paraphrase avec le même contexte d'autre part, peut renseigner sur la substituabilité grammaticale des deux segments (Zhao *et al.*, 2007; Max et Zock, 2008). Nous avons calculé les dépendances syntaxiques pour les deux segments à l'aide de la version française (Candito *et al.*, 2010) de l'analyseur probabiliste de Berkeley (Petrov et Klein, 2007). Nous considérons donc le sous-ensemble des dépendances qui existent entre les mots du segment d'origine et son contexte (Dep_{orig}) et entre les mots de la paraphrase et ce contexte (Dep_{para}). Cette technique, DEPCONT, retient la paraphrase candidate si et seulement si $Dep_{para} = Dep_{orig}$.

5.3 Résultats et analyse

Nous avons utilisé un séparateur à vastes marges (SVM) avec les traits décrits dans la section 4 [13] Les performances des différentes techniques sur les 3 conditions décrites précédemment sont données dans la figure 3.

	ML_WEB	LM_FRONTIÈRES	DEPCONT	CLASSIFIEUR
POSSIBLES	**62,79**	54,88	48,53	57,67
SÛRES	68,37	36,27	51,90	**70,69**
SÛRES++	56,79	51,41	42,69	**62,85**

FIGURE 3 – Résultats de la performance de la classification (*exactitude*) pour les 3 techniques de référence et notre classifieur sur le corpus d'évaluation et les 3 conditions. Il convient de noter que la condition SÛRES++ n'est pas directement comparable aux autres conditions puisque les tailles des corpus d'apprentissage et d'évaluation sont différentes à celles des deux autres conditions.

La première observation que nous pouvons faire est que la tâche de classification de paraphrases est une tâche difficile : la meilleure performance (*exactitude*) obtenue par l'un des systèmes est de 70,69 pour la condition SÛRES. En outre, il existe une variation importante entre les

13. Nous avons utilisé l'implémentation LIBSVM (Chang et Lin, 2001).

différentes conditions testées avec un résultat faible pour notre classifieur de 57,67 dans la condition POSSIBLES (cas de désaccord entre annotateurs, où un seul reconnaît le statut de paraphrase).

D'une manière plus générale, la technique ML_WEB et notre classifieur sont plus performants que les autres techniques de références. ML_FRONTIÈRES et DEPCONT ne modélisent que des contraintes grammaticales locales, ce qui fait qu'il n'est pas surprenant que ces informations ne permettent pas la reconnaissance de variations sémantiques licites entre paraphrases candidates. WEBLM, qui se limite à la comparaison de scores de modèles de langue dérivé du Web, apparaît donc comme une technique relativement compétitive [14], mais sa performance est peu élevée (56,79) pour la condition SURE++. Puisque cette condition ne prend en compte que les annotations consensuelles pour l'apprentissage et l'évaluation, nous considérons cette condition comme la plus utile pour l'interprétation des résultats de ces travaux préliminaires. Ici, notre système obtient la meilleur performance, avec un avantage de 6,06 points par rapport à WEBLM. Ceci montre que la seule utilisation d'un modèle de langue, aussi bien estimé soit-il, est trop limitée pour rendre compte correctement de l'ensemble des phénomènes de paraphrases présents dans notre corpus d'évaluation, ce qui confirme des résultats précédents où les modèles de langue n'étaient pas issus de comptes du Web (Bannard et Callison-Burch, 2005).

Finalement, la figure 4 détaille les performances atteintes par chacune des méthodes d'acquisition de paraphrases pour chacune des 3 conditions. Il n'est tout d'abord pas surprenant que les reformulations extraites de WICOPACO soient largement identifiées comme de bonnes paraphrases en contexte, en particulier dans les conditions POSSIBLES et SÛRES++. Ces paraphrases sont le résultat de reformulations par des contributeurs de Wikipédia dans le contexte d'évaluation, et avaient déjà été reconnues comme telles par une première annotatrice.

	WICOPACO	HUMAIN	PIVOT_{ES}	PIVOT_{ZH}
POSSIBLES	**89,33**	67,00	47,33	20,66
SÛRES	**64,00**	44,50	31,33	10,66
SÛRES++	**86,03**	57,34	37,71	12,60

FIGURE 4 – Performance (valeurs d'*exactitude*) de nos différentes méthodes d'acquisition pour nos trois conditions d'évaluation.

Les paraphrases obtenues par collecte manuelle sur des contextes issus du Web, donc d'un contexte possiblement différent de celui de l'évaluation, obtiennent une performance relativement acceptable. Les résultats confirment cependant le fait attendu que la substituabilité des paraphrases dépend fortement du contexte. Par exemple, la substitution du segment *de l'éditeur* par *publiée par les éditions* dans le contexte de l'énoncé "*Neopolis est une collection de bandes dessinées* ***de l'éditeur*** *Delcourt.* [15]" permet de conserver le sens d'origine ainsi que la grammaticalité de l'énoncé. *A contrario*, la substitution par le segment *du logiciel* n'est pas adaptée à ce contexte.

Finalement, les paraphrases obtenues automatiquement par traduction par pivot ne sont pas de bonne qualité. Nous notons cependant que la proximité de la langue pivot avec la langue

14. Une explication peut résider dans le fait que nos méthodes d'acquisition de paraphrases utilisant Google Translate commme un traducteur automatique par pivot ont tendance à produire des segments ayant une forte valeur de probabilité dans le modèle de langue utilisé, qui est certainement assez comparable à celui utilisé dans nos expériences.

15. Une réécriture est extraite de l'historique de révision de l'article "Neopolis" sur Wikipédia accessible sur : `http://fr.wikipedia.org/w/index.php?title=Neopolis&diff=45811975&oldid=2017149`.

de réécriture joue un rôle important : l'utilisation de l'espagnol mène ainsi à de bien meilleurs résultats que l'utilisation du chinois [16].

6 Conclusions et perspectives

Nous avons présenté dans cet article une approche de paraphrasage en contexte appliqué à la révision de texte, un scénario soutenu par les données extraites des réécritures contenues dans la Wikipédia francophone. La méthode d'identification que nous avons proposée prend en entrée un répertoire existant de paraphrases sous-phrastiques, et détermine par classification automatique exploitant des données issues du Web si les paraphrases connues peuvent se substituer à un segment dans un contexte particulier. Nous avons simulé différents niveaux de qualité pour les paraphrases existantes, en exploitant des paraphrases provenant de Wikipédia, des contributions humaines acquises dans d'autres contextes, et des paraphrases obtenues par traduction automatique par pivot.

Nos expériences ont montré que la version actuelle de notre classifieur est plus performante que les différentes techniques de référence utilisées lorsque l'on ne considère que les paraphrases obtenant des jugements consensuels dans la référence utilisée. Bien que ces premières expériences soient positives, nous sommes conscients que leurs résultats peuvent être améliorés sur différents aspects. Tout d'abord, il est possible d'élargir l'exploration des différentes caractéristiques que nous mettons en jeu dans le classifieur. Nous comptons intégrer d'autres traits, dont des modèles mettant en jeu des dépendances syntaxiques calculées sur des données du Web. Nous allons également analyser plus finement nos résultats afin d'identifier les cas problématiques, dont certains ne peuvent pas être modélisés sans avoir recours à des connaissances du monde, ce qui suggérera notamment l'intégration de connaissances du domaine, éventuellement dérivées de méta-informations provenant des articles Wikipédia concernés. L'ensemble de ces expériences pourra être conduit en plusieurs langues, les données utilisées et les méthodes employées pouvant facilement être transposées. Finalement, nous sommes également intéressés par le fait d'utiliser l'approche décrite ici comme un cadre pour l'évaluation des systèmes d'acquisition de paraphrases.

Références

BANNARD, C. et CALLISON-BURCH, C. (2005). Paraphrasing with bilingual parallel corpora. *In Actes de ACL*, Ann Arbor, USA.

BARZILAY, R. et LEE, L. (2003). Learning to paraphrase : an unsupervised approach using multiple-sequence alignment. *In Actes de NAACL-HLT*, Edmonton, Canada.

BARZILAY, R. et MCKEOWN, K. (2001). Extracting paraphrases from a parallel corpus. *In Actes de ACL*, Toulouse, France.

BERNSTEIN, M. S., LITTLE, G., MILLER, R. C., HARTMANN, B., ACKERMAN, M. S., KARGER, D. R., CROWELL, D. et PANOVICH, K. (2010). Soylent : a word processor with a crowd inside. *In Proceedings of the ACM symposium on User interface software and technology*.

16. Bannard et Callison-Burch (2005) ont montré que l'utilisation simultanée de plusieurs langues pivots permettait de diminuer de façon importante les phénomènes de bruit.

BHAGAT, R. et RAVICHANDRAN, D. (2008). Large scale acquisition of paraphrases for learning surface patterns. *In Actes de ACL-HLT*, Columbus, États-Unis.

BOUAMOR, H., MAX, A. et VILNAT, A. (2011). Monolingual alignment by edit rate computation on sentential paraphrase pairs. *In Proceedings of ACL, Short Papers session*, Portland, USA.

BROCKETT, C. et DOLAN, W. B. (2005). Support vector machines for paraphrase identification and corpus construction. *In Proceedings of The 3rd International Workshop on Paraphrasing IWP*, Jeju Island, South Korea.

CANDITO, M., CRABBÉ, B. et DENIS, P. (2010). Statistical French dependency parsing : treebank conversion and first results. *In Proceedings of LREC*, Valletta, Malta.

CHANG, C.-C. et LIN, C.-J. (2001). *LIBSVM : a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

COHN, T., CALLISON-BURCH, C. et LAPATA, M. (2008). Constructing corpora for the development and evaluation of paraphrase systems. *Comput. Linguist.*, 34(4).

COHN, T. et LAPATA, M. (2008). Sentence compression beyond word deletion. *In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Manchester, UK.

CONNOR, M. et ROTH, D. (2007). Context sensitive paraphrasing with a single unsupervised classifier. *In Proceedings of ECML*, Warsaw, Poland.

DELÉGER, L. et ZWEIGENBAUM, P. (2009). Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. *In Proceedings of the 2nd Workshop on Building and Using Comparable Corpora : from Parallel to Non-parallel Corpora*, Singapore.

DUTREY, C., BOUAMOR, H., BERNHARD, D. et MAX, A. (2011). Paraphrases et modifications locales dans l'historique des révisions de wikipédia. *In Actes de TALN 2011*, Montpellier, France.

GANITKEVITCH, J., CALLISON-BURCH, C., NAPOLES, C. et VAN DURME, B. (2011). Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. *In Proceedings of EMNLP*, Edinburgh, UK.

KOK, S. et BROCKETT, C. (2010). Hitting the Right Paraphrases in Good Time. *In Proceedings of NAACL*, Los Angeles, USA.

LANDIS, J. et KOCH, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

LAPATA, M. et KELLER, F. (2005). Web-based Models for Natural Language Processing. *ACM Transactions on Speech and Language Processing*, 2(1):1–31.

MADNANI, N. et DORR, B. J. (2010). Generating Phrasal and Sentential Paraphrases : A Survey of Data-Driven Methods . *Computational Linguistics*, 36(3).

MADNANI, N., RESNIK, P., DORR, B. et SCHWARTZ, R. (2008). Are multiple reference translations necessary ? investigating the value of paraphrased reference translations in parameter optimization. *In Proceedings of AMTA*, Waikiki, Hawai'i.

MAX, A. (2004). From controlled document authoring to interactive document normalization. *In Proceedings of COLING*, Geneva, Switzerland.

MAX, A. et WISNIEWSKI, G. (2010). Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History. *In Proceedings of LREC*, Valletta, Malta.

MAX, A. et ZOCK, M. (2008). Looking up phrase rephrasings via a pivot language. *In Proceedings of the COLING Workshop on Cognitive Aspects of the Lexicon*, Manchester, United Kingdom.

McCarthy, D. et Navigli, R. (2009). The english lexical substitution task. *Language Resources and Evaluation*, 43(2).

Mutton, A. (2006). *Evaluation of sentence grammaticality using Parsers and a Support Vector Machine*. Thèse de doctorat, Macquarie University.

Onishi, T., Utiyama, M. et Sumita, E. (2010). Paraphrase lattice for statistical machine translation. *In Proceedings of the ACL 2010 Conference, Short Paper session*, Uppsala, Sweden.

Pang, B., Knight, K. et Marcu, D. (2003). Syntax-based alignement of multiple translations : Extracting paraphrases and generating new sentences. *In Actes de NAACL-HLT*, Edmonton, Canada.

Pasça, M. et Dienes, P. (2005). Aligning Needles in a Haystack : Paraphrase Acquisition Across the Web. *In Proceedings of IJCNLP*, Jeju Island, South Korea.

Petrov, S. et Klein, D. (2007). Improved inference for unlexicalized parsing. *In Proceedings of NAACL-HLT*, Rochester, USA.

Quirk, C., Brockett, C. et Dolan, W. B. (2004). Monolingual machine translation for paraphrase generation. *In Proceedings of EMNLP*, volume 149, Barcelona, Spain.

Resnik, P., Buzek, O., Hu, C., Kronrod, Y., Quinn, A. et Bederson, B. B. (2010). Improving translation via targeted paraphrasing. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Schroeder, J., Cohn, T. et Koehn, P. (2009). Word Lattices for Multi-Source Translation. *In Proceedings of EACL*, Athens, Greece.

Snover, M., Madnani, N., Dorr, B. J. et Schwartz, R. (2010). TER-Plus : paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, 23(2-3).

Wang, K., Thrasher, C., Viegas, E., Li, X. et Hsu, B.-j. P. (2010). An Overview of Microsoft Web N-gram Corpus and Applications. *In Proceedings of the NAACL-HLT Demonstration Session*, Los Angeles, USA.

Zhao, S., Lan, X., Liu, T. et Li, S. (2009). Application-driven Statistical Paraphrase Generation. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Suntec, Singapore.

Zhao, S., Liu, T., Yuan, X., Li, S. et Zhang, Y. (2007). Automatic acquisition of context-specific lexical paraphrases. *In Proceedings of IJCAI*, Hyderabad, India.

Zhao, S., Wang, H., Liu, T., et Li, S. (2010). Leveraging multiple mt engines for paraphrase generation. *In Proceedings of COLING*, Beijing, China.

Zhu, Z., Bernhard, D. et Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. *In Proceedings of COLING*, Beijing, China.

Simplification syntaxique de phrases pour le français

Laetitia Brouwers[1, 2] Delphine Bernhard[1, 3]
Anne-Laure Ligozat[1, 4] Thomas François[2, 5]
(1) LIMSI-CNRS, 91403 Orsay, France
(2) Université catholique de Louvain, Belgique
(3) LiLPa, Université de Strasbourg, France
(4) ENSIIE, Evry, France
(5) University of Pennsylvania, USA

RÉSUMÉ

Cet article présente une méthode de simplification syntaxique de textes français. La simplification syntaxique a pour but de rendre des textes plus abordables en simplifiant les éléments qui posent problème à la lecture. La méthode mise en place à cette fin s'appuie tout d'abord sur une étude de corpus visant à étudier les phénomènes linguistiques impliqués dans la simplification de textes en français. Nous avons ainsi constitué un corpus parallèle à partir d'articles de Wikipédia et Vikidia, ce qui a permis d'établir une typologie de simplifications. Dans un second temps, nous avons implémenté un système qui opère des simplifications syntaxiques à partir de ces observations. Des règles de simplification ont été décrites afin de générer des phrases simplifiées. Un module sélectionne ensuite le meilleur ensemble de phrases. Enfin, nous avons mené une évaluation de notre système montrant qu'environ 80% des phrases générées sont correctes.

ABSTRACT

Syntactic Simplification for French Sentences

This paper presents a method for the syntactic simplification of French texts. Syntactic simplification aims at making texts easier to understand by simplifying the elements that hinder reading. It is based on a corpus study that aimed at investigating the linguistic phenomena involved in the manual simplification of French texts. We have first gathered a parallel corpus of articles from Wikipedia and Vikidia, that we used to establish a typology of simplifications. In a second step, we implemented a system that carries out syntactic simplifications based on these corpus observations. We described simplification rules in order to generate simplified sentences. A module subsequently selects the best subset of sentences. The evaluation of our system shows that about 80% of the sentences produced by our system are accurate.

MOTS-CLÉS : simplification automatique, lisibilité, analyse syntaxique.

KEYWORDS: automatic simplification, readability, syntactic analysis.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 211–224,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Dans la majorité de nos activités quotidiennes, la capacité de lire rapidement et efficacement constitue un atout certain, voire un pré-requis. Willms (2003) souligne ainsi une corrélation entre ces compétences et le statut socio-économique des individus. Pourtant, une tranche non négligeable de la population n'est pas capable de traiter efficacement les données textuelles auxquelles ils sont confrontés. Richard *et al.* (1993) rapportent une expérience où, sur 92 demandes d'allocation de chômage remplies par des personnes avec un faible niveau d'éducation, pas moins de la moitié des informations requises (dont certaines étaient cruciales pour le traitement de la demande) manquaient, notamment à cause de problème de compréhension. Dans un contexte légèrement différent, à savoir la pharmacologie, Patel *et al.* (2002) parviennent à un constat similaire : la plupart de leurs sujets ont rencontré des problèmes importants dans la compréhension des différentes étapes à réaliser pour la bonne administration du médicament testé.

Ces problèmes de compréhension s'expliquent souvent par une trop grande complexité des textes, en particulier au niveau du lexique et de la syntaxe. Ces deux facteurs sont connus comme étant des causes importantes des difficultés de lecture (Chall et Dale, 1995), en particulier chez les jeunes enfants, les apprenants d'une langue étrangère ou les personnes présentant des déficiences intellectuelles.

Dès lors, la simplification automatique de textes apparaît comme un moyen susceptible d'aider ces personnes à accéder plus facilement au contenu des documents écrits auxquels ils sont confrontés. Il s'agit d'un domaine du traitement automatique des langues (TAL) visant à rendre des textes plus abordables tout en garantissant l'intégrité de leur contenu et en veillant à en respecter la structure. Dès lors, il faut déterminer d'une part quelles informations sont secondaires afin de les supprimer et de rendre les informations primordiales plus visibles et d'autre part quelles sont les constructions syntaxiques qui peuvent poser problème pour les simplifier.

Parmi les premiers efforts en ce sens, citons (Carroll *et al.*, 1999) et (Inui *et al.*, 2003), qui ont proposé des outils pour produire des textes plus abordables pour les personnes atteintes d'un handicap langagier tel que l'aphasie ou la surdité. Cependant, l'aide à la lecture ne s'adresse pas qu'aux lecteurs présentant des handicaps, mais aussi à ceux qui apprennent une langue (première ou seconde). Ainsi, Belder et Moens (2010) se sont intéressés à la simplification pour des enfants de langue maternelle anglaise, tandis que Siddharthan (2006), Petersen et Ostendorf (2007) et Medero et Ostendorf (2011) ont étudié la simplification pour les apprenants d'une langue seconde. La plupart de ces travaux concernent la langue anglaise, à l'exception de (Inui *et al.*, 2003) qui traitent également le japonais.

Parallèlement, la simplification automatique a également été utilisée comme un pré-traitement visant à augmenter l'efficacité d'opérations postérieures effectuées sur des textes. Les premiers, Chandrasekar *et al.* (1996) ont considéré que les phrases longues et complexes constituaient un obstacle pour l'analyse syntaxique ou la traduction automatique et que leur simplification préalable pouvait conduire à de meilleures analyses. Plus récemment, Heilman et Smith (2010) ont montré, quant à eux, qu'un texte simplifié produit de meilleurs résultats dans un contexte de génération automatique de questions. Du coté du biomédical, Lin et Wilbur (2007) et Jonnalagadda *et al.* (2009) ont optimisé l'extraction de données en simplifiant les textes lors d'un prétraitement.

La majorité des méthodes de simplification syntaxique proposées reposent sur un ensemble de règles de transformation définies manuellement pour être appliquées aux phrases. La simplification semble toutefois naturellement se prêter à l'utilisation de méthodes issues de la traduction automatique ou de l'apprentissage automatique, dont les modèles sont construits à partir de corpus comparables de textes complexes et simplifiés (Zhu *et al.*, 2010; Specia, 2010; Woodsend et Lapata, 2011). Les données utilisées dans ce cas sont notamment issues de Wikipédia en anglais et de Simple English Wikipedia, destinée aux enfants et aux locuteurs non natifs. L'encyclopédie Simple English Wikipedia compte à ce jour plus de 75 000 articles.

Il existe des projets comparables pour le français, Vikidia (voir Section 2.1) et Wikimini, mais ils ne sont pas aussi fournis que leur homologue anglophone. Par ailleurs, les différentes versions d'un article de Wikipédia ne sont pas strictement parallèles, ce qui complique encore l'apprentissage automatique. La méthode proposée dans cet article repose donc sur un ensemble de règles de simplification automatique qui ont été définies manuellement (voir Section 2.3), après étude de corpus. Nous utilisons la technique de la sur-génération, qui consiste à produire dans un premier temps un nombre important de simplifications possibles, avant de procéder à une sélection optimale des meilleures simplifications produites, à l'aide de la programmation linéaire en nombre entiers (PLNE, en anglais *Integer Linear Programming – ILP*). La PLNE permet de définir des contraintes qui régissent le choix du résultat fourni par l'outil de simplification automatique. Cette méthode a notamment été appliquée à la simplification de textes en anglais par (Woodsend et Lapata, 2011), (Belder et Moens, 2010), ainsi que par (Gillick et Favre, 2009) pour le résumé automatique.

Les apports de cet article sont les suivants : l'étude des procédés de simplification en français, et notamment la constitution d'un corpus de phrases parallèles, et une typologie des simplifications ; l'utilisation de critères originaux de sélection des phrases, tels que la liste orthographique de base de Nina Catach ou les mots-clés d'un texte. Nous présenterons tout d'abord le processus de constitution du corpus (Section 2.1), puis la typologie des simplifications observées (Section 2.2). Nous détaillerons ensuite le fonctionnement du système mis en œuvre, qui procède en deux temps : une surgénération de phrases simplifiées (Section 2.3.1), et une sélection des phrases correspondant à des critères de lisibilité (Section 2.3.2). Enfin, nous évaluerons cette simplification du point de vue de la correction des phrases générées, et analyserons les causes d'erreurs (Section 3).

2 Méthodologie

2.1 Présentation du corpus

Pour établir une typologie des règles de simplification, une étude sur corpus a été réalisée. Puisqu'il s'agit de déterminer les stratégies utilisées pour passer d'une phrase complexe à une phrase simplifiée, un corpus de phrases parallèles a été construit à partir d'articles des encyclopédies en ligne Wikipédia [1] et Vikidia [2]. Cette dernière est destinée aux jeunes de huit à treize ans et rassemble des articles plus accessibles, tant au niveau de la langue que du contenu. Afin de constituer ce corpus, nous sommes partis des articles de Vikidia et avons utilisé l'API MediaWiki

1. http://fr.wikipedia.org
2. http://fr.vikidia.org

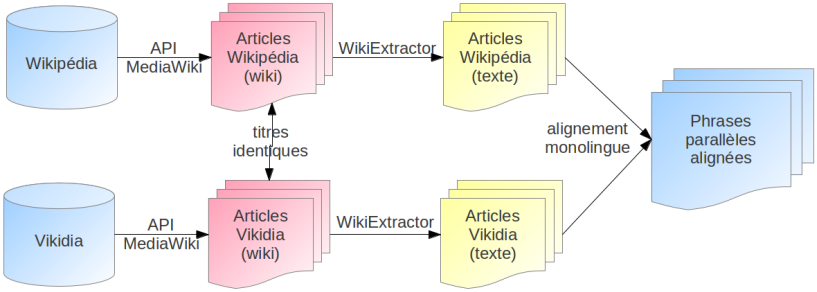


FIGURE 1 – Constitution du corpus de phrases parallèles

pour récupérer les articles de Wikipédia et Vikidia de mêmes titres. Le programme WikiExtractor [3] a ensuite été appliqué à ces articles afin d'en extraire les textes bruts (c'est-à-dire sans la syntaxe wiki). Le corpus ainsi constitué comprend 13 638 fichiers (dont 7 460 de Vikidia et 6 178 de Wikipédia, certains articles de Vikidia n'ayant pas d'équivalent direct dans Wikipédia).

Ces articles ont ensuite été analysés afin de repérer des phrases parallèles (phrase de Wikipédia ayant un équivalent simplifié dans Vikidia). Cet alignement a été effectué en partie manuellement et en partie automatiquement grâce à l'algorithme d'alignement monolingue décrit dans (Nelken et Shieber, 2006), qui se fonde sur une similarité cosinus entre phrases, avec un *tf.idf* adapté pour la pondération des mots. Ce programme fournit en sortie des alignements entre phrases, avec un score de confiance associé. La figure 1 résume le processus de constitution de ce corpus.

Parmi ces fichiers, vingt articles ou extraits d'articles de Wikipédia et leur équivalent dans Vikidia ont été sélectionnés, ce qui nous donne respectivement 72 phrases et 80 phrases. Les extraits suivants - correspondant à l'entrée «archipel» - ont par exemple été sélectionnés :

> (1a) Wikipédia : *Un archipel est un ensemble d'îles relativement proches les unes des autres. Le terme «archipel» vient du grec ancien "Archipelagos", littéralement «mer principale» (de "archi" : «principal» et "pélagos" : «la haute mer»). En effet, ce mot désignait originellement la mer Égée, caractérisée par son grand nombre d'îles (les Cyclades, les Sporades, Salamine, Eubée, Samothrace, Lemnos, Samos, Lesbos, Chios, Rhodes, etc.)*.
> (1b) Vikidia : *Un archipel est un ensemble de plusieurs îles, proches les unes des autres. Le mot «archipel» vient du grec "archipelagos", qui signifie littéralement «mer principale» et désignait à l'origine la mer Égée, caractérisée par son grand nombre d'îles.*

Notons que les deux articles présentent les mêmes informations globalement, mais de manière différente. Il y a simplification lexicale, sémantique et syntaxique. En effet, dans Vikidia, il n'y a que deux phrases, qui contiennent l'essentiel de l'explication (information nécessaire) tandis que dans Wikipédia, trois phrases détaillent la signification et l'origine du terme de manière plus précise (informations secondaires, par exemple mises entre parenthèses).

3. `http://medialab.di.unipi.it/wiki/Wikipedia_Extractor`

2.2 Typologie de simplifications

Les observations réalisées sur ce corpus ont permis d'établir une typologie articulée selon trois grands niveaux de transformations : lexical, sémantique et syntaxique. Dans les travaux réalisés, la simplification est communément considérée comme composée de deux catégories, lexicale et syntaxique (Carroll *et al.*, 1999; Inui *et al.*, 2003; Belder et Moens, 2010). Le domaine de la sémantique quant à lui n'est pas cité. Ces trois grands niveaux peuvent être à leur tour divisés en sous-catégories, comme le montre la table 1.

Lexique	**Sémantique**	**Syntaxe**
Synonyme ou hyperonyme Traduction	Réorganisation Suppression Ajout	Temps Suppression Modification Division Regroupement

TABLE 1 – Typologie

En ce qui concerne le lexique, deux phénomènes sont observés. D'une part, les termes considérés comme difficiles sont remplacés par un synonyme ou un hyperonyme. Dans l'exemple (1), *terme* a été remplacé par *mot* qui est plus courant. D'autre part, les concepts utilisés dans leur langue d'origine dans Wikipédia sont traduits en français dans Vikidia.

Au niveau sémantique, les auteurs de Vikidia prêtent une attention particulière à l'organisation de l'information qui doit être claire et synthétique. Dans cette optique, il arrive que des propositions soient interverties, afin d'assurer une meilleure présentation de l'information. De plus, le contenu considéré comme secondaire à la compréhension est supprimé tandis que des explications ou des exemples sont ajoutés pour plus de clarté. Ainsi, dans l'exemple (1), la décomposition de la signification du mot *archipel* est explicitée dans Wikipédia, mais pas dans Vikidia.

Enfin, du point de vue syntaxique, qui nous intéresse prioritairement ici, cinq types de changements sont observés : les modifications de temps, la suppression, la modification, la division et le regroupement. Les deux derniers types peuvent être envisagés ensemble dans la mesure où ce sont deux phénomènes opposés. Cette classification peut se rapprocher de celle de (Medero et Ostendorf, 2011) qui reprend trois catégories - la division, la suppression et l'extension - ou de (Zhu *et al.*, 2010) (composée de la division, la suppression, la réorganisation et la substitution).

– Tout d'abord, les temps utilisés dans Vikidia sont plus quotidiens et moins littéraires que ceux utilisés dans Wikipédia. Ainsi, le présent et le passé composé sont préférés au passé simple.
– Ensuite, les informations secondaires ou redondantes, telles que certains compléments circonstanciels, qui sont en général considérées comme supprimables au niveau syntaxique, ne sont pas reprises dans les articles de Vikidia. Dans l'exemple (1), l'adverbe *relativement* qui précédait *proches les unes des autres* a ainsi été supprimé dans Vikidia. L'adverbe n'ajoutait effectivement rien au niveau informationnel.
– De plus, si certaines structures plus complexes ne sont pas supprimées, elles sont alors déplacées ou modifiées pour plus de clarté. Dans Vikidia, par exemple, une construction affirmative est préférée à une forme négative :

> (2a) Wikipédia : *Les personnes qui ont voté blanc ou nul ne sont généralement pas considérées comme abstentionnistes mais le résultat est identique : leur choix n'est pas*

pris en compte.
(2b) Vikidia : *L'abstention est différente du vote blanc et du vote nul.*

– Finalement, les auteurs choisissent parfois de diviser des phrases longues ou à l'inverse de réunir plusieurs phrases en une seule. Dans l'exemple (1), les deux dernières phrases ont été regroupées dans Vikidia, car elles ont été simplifiées et sont dès lors devenues beaucoup plus courtes. Il faut d'emblée préciser que le regroupement d'éléments est beaucoup moins utilisé que la division de phrases. Pour scinder une phrase, les auteurs prennent par exemple une proposition secondaire (telle qu'une relative) qu'ils transforment en phrase indépendante.

Parmi les changements observés, certains d'entre eux sont difficilement implémentables. C'est le cas lorsqu'une modification nécessite de recourir à la sémantique, c'est-à-dire qu'il n'est possible de repérer les structures à modifier que par le sens. Il est difficile d'appliquer ce type de stratégies de manière automatique. Par exemple, il est parfois possible de supprimer les éléments qui se rapportent au nom, alors que d'autre fois, ils sont indispensables, sans que cela ne soit marqué typographiquement ou grammaticalement dans la phrase.

D'autres changements syntaxiques doivent s'accompagner de transformations lexicales, difficilement généralisables. Par exemple, la modification d'une phrase négative en une phrase affirmative nécessite de trouver un verbe dont la forme affirmative recouvre le sens de la construction négative à remplacer.

Il y a également des changements qui sont effectués de manière isolée et non systématisable. Ils relèvent plutôt d'un traitement manuel que d'un traitement automatique d'un texte, dans le sens où chaque cas est différent (même s'il s'inscrit dans une règle plus globale). De plus, ils font généralement appel à des informations sémantiques ou lexicales et pas simplement syntaxiques. Il s'agit de changements complexes, qui sont utiles dans certains cas, mais ardus à détecter automatiquement.

Enfin, les changements syntaxiques qui ont un impact sur d'autres parties du texte ou qui concernent des éléments dépendants d'une autre structure demandent des modifications plus globales du texte. Par conséquent, ils sont également difficiles à traiter automatiquement. Ainsi, pour modifier le temps d'un verbe dans une phrase, il faut veiller à ce que la concordance des temps soit respectée dans l'entièreté du texte.

2.3 Système de simplification syntaxique

Nous avons utilisé cette typologie pour mettre en œuvre un système de simplification syntaxique pour le français. La simplification d'un texte y est effectuée en deux étapes : une étape de génération de toutes les simplifications possibles pour chaque phrase du texte, et une étape de sélection du meilleur ensemble de phrases simplifiées. L'architecture de ce système est présentée dans la figure 2.

Le module de surgénération s'appuie sur un ensemble de règles (au nombre de 19), utilisant des informations sur les caractéristiques (morpho-)syntaxiques des mots et sur les relations de dépendance présentes au sein d'une phrase. C'est pourquoi les textes de notre corpus ont été analysés par MElt [4] (Denis *et al.*, 2009) et Bonsai [5] (Candito *et al.*, 2010). Ces textes ont ainsi été

4. https://gforge.inria.fr/projects/lingwb
5. http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html

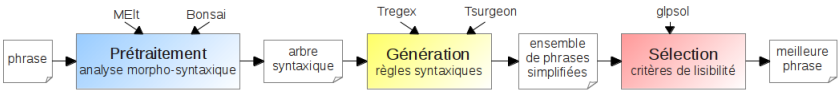


FIGURE 2 – Organisation du système de simplification syntaxique

représentés sous la forme d'arbres syntaxiques, lesquels contiennent un maximum de données utiles à l'application de règles de simplification. Ces dernières peuvent alors être appliquées de manière récursive, jusqu'à ce qu'il n'y ait plus aucune structure à simplifier dans chacune des phrases des textes. Il faut ajouter que toutes les phrases créées à chaque application d'une règle sont enregistrées, produisant un ensemble de variantes. Par la suite, le meilleur ensemble de phrases sera retenu via un modèle de programmation linéaire, en fonction d'une série de critères détaillés par la suite.

2.3.1 Génération de phrases simplifiées

Les règles de simplification syntaxique qui composent notre programme sont respectivement des règles de suppression (12 règles), de modification (3 règles) et de division (4 règles). Notons que, par rapport à la typologie établie, deux types de règles n'ont pas été mises en place. D'une part, les stratégies de regroupement de plusieurs phrases en une n'ont pas été observées de manière assez systématique dans le corpus d'étude. Il est dès lors difficile d'en retirer une règle automatisable. De plus, les règles de regroupement pourraient entrer en conflit avec les règles de suppression, puisqu'elles ont des buts opposés. D'autre part, en ce qui concerne les aspects temporels, nous avions noté que certains temps étaient plus utilisés que d'autres dans l'encyclopédie pour les jeunes, Vikidia. Toutefois, cette stratégie n'a pas été implémentée car elle demandait des changements trop globaux, pouvant toucher au texte entier. En effet, lorsqu'un verbe au passé simple est remplacé par un verbe au présent, il faut veiller à ce que la concordance des temps soit toujours respectée partout, ce qui demande d'examiner tout le texte, ou du moins le paragraphe qui contient la forme verbale modifiée. On risque alors de détruire la cohérence du texte et d'en altérer la qualité.

Pour appliquer ces 19 règles, il convient tout d'abord de repérer les structures concernées par de possibles changements à l'aide d'expressions régulières et grâce à `Tregex`[6] (Levy et Andrew, 2006) qui gère le repérage d'éléments et de relations dans un arbre. Dans un deuxième temps, une série d'opérations sont effectuées par le biais de `Tsurgeon` qui permet de modifier des arbres syntaxiques. Par exemple, pour supprimer une coordonnée introduite par *soit*, il faut repérer une proposition coordonnée, étiquetée `COORD`, qui domine la conjonction de coordination *soit* et lui donner un nom comme `Pcoord`. Ensuite, l'opération `Tsurgeon delete` doit être appliquée à l'ensemble repris sous `Pcoord` :

Repérage (`Tregex`) : `COORD=Pcoord < (CC < /soit/)`
Opération (`Tsurgeon`) : `delete Pcoord`

6. `http://nlp.stanford.edu/software/tregex.shtml`

Cette règle s'appliquerait par exemple à la phrase suivante :

> (3a) Phrase d'origine : *Elle compte, selon les autorités du pays, en 2009, 5 878 609 habitants pour l'agglomération, et 3 796 677 habitants pour la ville, soit 20 % de la population totale du pays.*
> (3b) Phrase après application de la règle : *Elle compte, selon les autorités du pays, en 2009, 5 878 609 habitants pour l'agglomération, et 3 796 677 habitants pour la ville.*

Les opérations varient en fonction du type de règle appliquée :

1. Pour les règles de suppression, il suffit de supprimer tous les éléments concernés (via l'opération `Tsurgeon delete`). Les éléments concernés par les règles de suppression sont les compléments circonstanciels, les ensembles entre parenthèses, une partie des propositions subordonnées, les propositions entre virgules ou introduites par un terme tel que *comme*, *voire* ou *soit*, les adverbes et les compléments d'agent.

2. Pour les règles de modification, il s'agit de combiner plusieurs opérations : la suppression de certains termes (opération `Tsurgeon delete`), le déplacement d'éléments (opération `Tsurgeon move`) et l'ajout d'étiquettes (opération `Tsurgeon insert`) qui signalent un traitement éventuel par la suite. En effet, certaines règles demandent que les formes verbales se conjuguent à un autre temps, un autre mode, etc. Dans ce cas, des étiquettes sont ajoutées autour du verbe pour indiquer qu'il doit être modifié. Il sera, dans un traitement postérieur, conjugué à la forme voulue grâce au système de conjugaison le `Verbiste`[7]. Ainsi, pour passer d'une structure passive à une structure active, il faut modifier le mode, mais aussi parfois la personne, pour que le verbe s'accorde correctement avec le complément d'agent devenu sujet. Trois règles de modification ont été mises en place : le déplacement à l'initiale des compléments circonstanciels, le passage à l'actif des formes passives et la transformation d'une clivée en non clivée.

3. Pour les règles de division, le processus se déroule en deux étapes. La proposition secondaire est d'abord supprimée et la nouvelle phrase enregistrée telle quelle. Ensuite, la phrase d'origine est reprise et la proposition principale est cette fois supprimée avant que la proposition secondaire soit transformée de manière à devenir indépendante. En général, il faut veiller à modifier la forme verbale de cette proposition secondaire pour qu'elle puisse fonctionner comme un verbe principal. Par ailleurs, le pronom qui régit la proposition doit être remplacé par son antécédent et le sujet doit être inséré lorsqu'il est manquant. Par exemple, pour transformer une relative en une proposition indépendante, le pronom relatif doit être remplacé par son antécédent et il est important de tenir compte de la fonction du pronom pour savoir où insérer l'antécédent. Signalons que les phrases sont scindées quand elles contiennent des propositions secondaires introduites par deux points, des coordonnées, des participiales ou des relatives.

Ces règles de simplification sont appliquées de manière récursive à une phrase jusqu'à ce que toutes les variantes possibles aient été générées. Plusieurs résultats sont donc régulièrement obtenus pour une même phrase. Dès lors, il convient de déterminer la phrase, parmi toutes celles produites, qui est la plus appropriée pour remplacer celle d'origine. Ce processus est décrit à la section suivante.

7. Le programme est disponible à l'adresse `http://sarrazip.com/dev/verbiste.html` sous licence GNU (page consultée le 6 novembre 2011). Il a été créé par Pierre Sarrazin.

2.3.2 Sélection de phrases simplifiées

Étant donné un ensemble de phrases simplifiées possible pour un texte, notre objectif est de sélectionner le meilleur sous-ensemble de phrases simplifiées, c'est-à-dire celui qui maximise une mesure de lisibilité. Cette mesure de lisibilité se traduit par différents critères. Pour résoudre ce genre de problèmes, la programmation linéaire en nombres entiers constitue une technique appropriée.

Dans notre cas, quatre critères ont été pris en compte pour choisir la phrase adéquate : la longueur de la phrase, la longueur des mots, la familiarité du vocabulaire et la présence de termes-clés, c'est-à-dire récurrents dans le texte. La longueur de la phrase est exprimée en nombre de mots tandis que la longueur des mots est donnée en nombre de caractères. En ce qui concerne la familiarité des mots, la liste de Catach [8] (Catach, 1985) a été utilisée pour calculer le poids de chaque terme. Il s'agit d'une liste des 3000 mots les plus fréquents, dont il convient d'enseigner l'orthographe en priorité aux élèves de primaire. Les termes-clés ont été définis, quant à eux, comme les mots qui apparaissaient deux fois ou plus dans un texte.

Ces critères sont combinés grâce à la formule suivante au sein du module de programmation linéaire [9] :

$$\begin{array}{ll} \text{Il s'agit alors de maximiser :} & h_w + h_s + h_a + h_c \\ \text{Où :} & h_w = \text{wps} \times \sum_i s_i - \sum_i l_i^w s_i \\ & h_s = \text{cpw} \times \sum_i l_i^w s_i - \sum_i l_i^c s_i \\ & h_a = \text{aps} \times \sum_i s_i - \sum_i l_i^a s_i \\ & h_c = \sum_j w_j c_j \end{array} \quad (1)$$

Nous avons défini les paramètres et variables suivants pour la formulation du problème :
– wps : le nombre moyen de mots par phrase souhaité
– cpw : le nombre moyen de caractères par mot souhaité
– aps : le nombre moyen de mots absents de la liste de Catach souhaité
– s_i un indicateur de la présence de la phrase i dans la simplification de texte finale
– c_j un indicateur de la présence du mot-clé j dans la simplification de texte finale
– l_i^w la longueur en mots de la phrase i
– l_i^c la longueur en caractères de la phrase i
– l_i^a le nombre de mots absents de la phrase i
– w_j le nombre d'occurrences du mot-clé j

`wps`, `cpw` et `aps` sont des paramètres constants dont les valeurs ont été fixées respectivement à 10, à 5 et à 2 pour cette étude. Toutefois, il s'agit de paramètres susceptibles de varier en fonction du contexte d'utilisation et du public cible, puisqu'ils déterminent directement le niveau de difficulté des phrases simplifiées retenues.

Pour illustrer ce processus, prenons le texte de départ pour l'article de Wikipédia intitulé *Abel*. Il comprenait 25 phrases, à partir desquelles 67 phrases simplifiées ont été produites. Parmi le texte simplifié, nous observons que ce sont les phrases 3 de l'exemple (4b) qui remplacent la phrase du texte original (exemple (4a)) :

> (4a) Phrase d'origine (Phrase 1) : *Caïn, l'aîné, cultive la terre et Abel (étymologie : de l'hébreu " souffle ", " vapeur ", " existence précaire ") garde le troupeau.*

8. Elle est notamment disponible sur le site `http://www.ia93.ac-creteil.fr/spip/spip.php?article2900`.
9. Le module repose sur `glpk` qui est disponible à l'adresse suivante : `http://www.gnu.org/software/glpk/`

(4b) Simplifications possibles :
Phrase 2 : *Caïn, l'aîné, cultive la terre et Abel garde le troupeau.*
Phrases 3 : *Caïn, l'aîné, cultive la terre. Abel garde le troupeau.*
Phrase 4 : *Caïn, l'aîné, cultive la terre.*
Phrase 5 : *Abel garde le troupeau.*
Phrases 6 : *Caïn, l'aîné, cultive la terre. Abel (étymologie : de l'hébreu " souffle ", " vapeur ", " existence précaire ") garde le troupeau.*
Phrase 7 : *Abel (étymologie : de l'hébreu " souffle ", " vapeur ", " existence précaire ") garde le troupeau.*
(4c) Simplification sélectionnée (Phrase 3) : *Caïn, l'aîné, cultive la terre. Abel garde le troupeau.*

Les valeurs des paramètres pour chaque phrase de l'exemple (4) sont données dans la table 2.

	Longueur de la phrase	**Longueur des mots**	**Familiarité des mots**	**Termes clés**
Valeurs souhaitées	10 mots	5 caractères	2 mots absents	
Phrase 1	19 mots	6,1 caractères	11 mots absents	5 termes
Phrase 2	11 mots	4,3 caractères	5 mots absents	5 termes
Phrases 3	5 mots	4,6 caractères	2 mots absents	5 termes
Phrase 4	6 mots	4,5 caractères	3 mots absents	3 termes
Phrase 5	4 mots	4,7 caractères	2 mots absents	2 termes
Phrases 6	9 mots	6,3 caractères	5 mots absents	5 termes
Phrase 7	12 mots	7,3 caractères	8 mots absents	2 termes

TABLE 2 – Valeurs des paramètres pour les phrases de l'exemple (4)

3 Évaluation

La simplification syntaxique implique des modifications importantes au sein de la phrase aussi bien au niveau du contenu que de la forme. C'est pourquoi il est important de vérifier que l'application d'une règle ne provoque pas des erreurs qui rendraient les phrases produites incompréhensibles ou agrammaticales. Une évaluation manuelle de notre système de génération de phrases simplifiées a donc été réalisée dans ce but. Elle repose sur un nouveau corpus composé de neuf articles de Wikipédia, c'est-à-dire de 202 phrases. Les résultats obtenus sont détaillés à la table 3 et dans la Section 3.1. Nous y détectons deux grands types d'erreurs, à savoir les erreurs d'analyse (morpho-)syntaxique et les erreurs de simplification. Celles-ci sont discutées dans les Sections 3.2 et 3.3.

3.1 Données obtenues

Sur les 202 phrases qui composent le corpus d'évaluation, 113 d'entre elles (56%) ont subi une ou plusieurs simplifications. Ces 113 phrases auxquelles des règles ont pu être appliquées donnent lieu à 333 variantes susceptibles de comporter des erreurs. C'est effectivement le cas de 71 d'entre elles (21,32%). Parmi celles-ci, il faut distinguer d'emblée les erreurs dues au

programme de simplification et les erreurs provoquées par un pré-traitement (analyse morpho-syntaxique et syntaxique) inexact. Ainsi, parmi les 21,32% de phrases problématiques, il apparaît que 89% des erreurs proviennent de l'analyse (morpho-)syntaxique et seulement 11% d'entre elles sont effectivement dues au système mis en place. Par conséquent, à partir du corpus d'évaluation, seulement 2,4% des phrases produites par notre système posent problème en raison de l'application d'une règle de simplification. Parmi ces rares erreurs de simplification (notre corpus en compte 8), il faut enfin distinguer les erreurs de contenu (25%) et de forme (75%). Dans la suite de cette section, nous revenons plus en détail sur les deux types d'erreurs principales rencontrées : morpho-syntaxiques et de simplification.

<table>
<tr><td colspan="4">Phrases produites par le programme
333 phrases - 100%</td></tr>
<tr><td>Phrases correctes
262 phrases - 78,68%</td><td colspan="3">Phrases incorrectes
71 phrases - 21,32%</td></tr>
<tr><td></td><td>Erreurs dues au pré-traitement
63 phrases - 18,92%</td><td colspan="2">Erreurs dues au simplificateur
8 phrases - 2,4%</td></tr>
<tr><td></td><td></td><td>Syntaxe
6 phrases
1,8%</td><td>Sens
2 phrases
0,6%</td></tr>
</table>

TABLE 3 – Évaluation des règles de simplification

3.2 Erreurs d'analyse (morpho-)syntaxique

La phase de pré-traitement consiste à étiqueter, annoter et structurer des phrases. Dès lors, il peut y avoir des erreurs dans les étiquettes attribuées, les relations identifiées, les regroupements d'éléments et les délimitations de phrases et de parenthèses.

Les erreurs d'étiquette sont les plus fréquentes et concernent des entités nommées, des cas ambigus ou des expressions figées. Par exemple, les mots *ainsi que* peuvent poser problème puisqu'il peut s'agir d'un connecteur ou des termes *ainsi* et *que*. L'analyseur ne parvient pas toujours à différencier les deux cas, ce qui peut provoquer des erreurs lors de la suppression des coordonnées (si les mots *ainsi que* sont identifiés comme un connecteur à tort). La phrase suivante en est un exemple :

> (5) *Les mélodies sont accrocheuses et les arrangements très soignés ; c'est ainsi que "Mamma Mia" et "Fernando" (malgré quelques erreurs de grammaire anglaise) occupèrent la première place des palmarès mondiaux dans le premier semestre de cette même année.*

Puisque *ainsi que* est considéré comme un connecteur et non comme l'adverbe *ainsi* suivi du deuxième terme de la clivée *que*, la règle de suppression des coordonnées produit la phrase suivante qui est agrammaticale : *C'est*.

Au-delà des problèmes d'étiquette, les relations de dépendance posent aussi fréquemment des difficultés à l'analyseur syntaxique. En effet, il est difficile de déterminer où s'arrête un groupe ou une proposition, quels éléments le composent ou de quel élément il dépend, particulièrement lorsque les constructions sont complexes et même emboitées. À nouveau, cela peut poser problème si une règle de simplification s'applique justement à un groupe mal analysé.

Les ponctuations constituent également des éléments difficiles à traiter pour l'analyseur. C'est pourquoi les phrases et les groupes entre parenthèses ne sont pas toujours convenablement délimités. De plus, l'analyseur ne distingue pas les points contenus dans les citations entre guillemets et ceux qui marquent une fin de phrase. Cela peut amener le programme, qui applique des règles, à diviser une phrase en plusieurs propositions de manière erronée. Les parenthèses, quant à elles, ne sont pas toujours marquées à un même niveau, ce qui détruit l'unité de l'ensemble. En effet, tous les composants de l'expression entre parenthèses ne sont pas rassemblés sous un même élément, ce qui signifie qu'une partie de l'expression peut être supprimée sans le reste et inversement.

3.3 Erreurs de simplification

À côté des erreurs issues du pré-traitement, certaines phrases, erronées aux niveaux sémantique et syntaxique, peuvent être produites à la suite de l'application des règles de simplification. Il s'agit évidemment là des erreurs les plus intéressantes, qui se répartissent en deux grandes catégories.

D'une part, les informations véhiculées par la phrase peuvent se trouver modifiées ou amputées. De fait, lors de la suppression de l'infinitive, il arrive qu'une partie du contenu de la phrase soit perdue. Ainsi, dans la phrase suivante, issue de l'article *abbé*, la proposition infinitive, qui explique le terme *abbé*, est supprimée :

(6a) *C'est aussi depuis le XVIIIe siècle le terme en usage pour désigner un clerc séculier ayant au moins reçu la tonsure.*
(6b) *C'est aussi depuis le XVIIIe siècle le terme en usage.*

Par ailleurs, lors de la suppression du complément d'agent, le sens d'une phrase peut être bouleversé par ce type de modification. Il en est ainsi pour la phrase de l'exemple (7b). En effet, le sens de la phrase originale (7a) était tout à fait différent :

(7a) *Ils ne sont pas caractérisés par leur profession comme dans la Bible : l'un pasteur, l'autre agriculteur.*
(7b) *Ils ne sont pas caractérisés : l'un pasteur, l'autre agriculteur.*

D'autre part, la structure de la phrase peut être modifiée de telle façon que la phrase devient syntaxiquement incorrecte. Trois règles de simplification sont concernées. Tout d'abord, les règles de suppression sont sujettes à ce genre de problème puisqu'il s'agit de supprimer une partie de la phrase, qui est normalement secondaire au bon fonctionnement de la phrase. Pourtant, il arrive que l'élément supprimé soit essentiel, comme dans le cas de la suppression du référent d'un pronom. La suppression de la subordonnée ou de l'infinitive peut provoquer ce type de désagrément. De son côté, la division de la phrase à partir de la relative produit un autre genre d'erreurs. Si le verbe est suivi d'un infinitif, le complément direct (l'antécédent du relatif qui est replacé dans la relative lors de la division) peut dépendre du verbe ou de l'infinitif. Le simplificateur n'en tient pas compte et estime dans tous les cas que le complément direct dépend du verbe et non de l'infinitif, parfois de manière erronée comme dans l'exemple (8) :

(8a) *Ils ont sur leurs religieux un droit de juridiction, une autorité qu'il leur est recommandé de n'exercer que par la voie de la patience et de la douceur.*
(8b) *Il leur est recommandé cette autorité de n'exercer que par la voie de la patience et de la douceur.*

4 Conclusion et perspectives

Cet article a décrit un système automatique de simplification syntaxique pour le français à destination des enfants en particulier. Celui-ci repose sur un ensemble de règles obtenues sur la base d'une étude de corpus, laquelle a aussi mené à l'élaboration d'une typologie des simplifications en français. Il serait aisé d'étendre notre typologie à d'autres publics sur base d'autres corpus adéquats. Notre démarche utilise également la technique de la sur-génération, qui permet de retenir le meilleur ensemble de simplifications en fonction de critères de lisibilité. Notons que parmi ceux employés, certains n'avaient pas été considérés précédemment et produisent des résultats intéressants. Enfin, il est apparu que les performances de notre système sont bonnes (environ 80% des phrases générées sont correctes), en particulier si l'on ne tient pas compte des erreurs dues aux outils de prétraitement.

Nous envisageons plusieurs perspectives d'amélioration pour notre système. Tout d'abord, la simplification syntaxique pourrait être complétée par une simplification lexicale, ainsi que cela est fait dans certaines études pour l'anglais (Woodsend et Lapata, 2011). Il s'agit en effet d'une autre source de problèmes pour certains lecteurs. Par ailleurs, notre analyse des erreurs a souligné la nécessité d'ajouter ou de répéter des termes lorsqu'une division de phrase est effectuée. Il serait dès lors utile de développer un outil qui gérerait les référents, afin d'améliorer la qualité du texte simplifié. Enfin, une dernière perspective d'amélioration consisterait à rendre le système de règles modulable en fonction d'un public cible. Cela demanderait d'évaluer la pertinence des différentes transformations et des critères de sélection des meilleures simplifications en fonction des publics visés. Cette perspective nécessiterait d'évaluer l'efficacité des règles au moyen de tests de compréhension portant sur des phrases originales et simplifiés.

Remerciements Nous remercions Antoine Sylvain pour sa participation à la constitution du corpus de textes issus de Wikipédia et Vikidia. Ces travaux ont reçu le soutien financier du projet DOXA du pôle de compétitivité CAP-DIGITAL.

Références

BELDER, J. D. et MOENS, M.-F. (2010). Text Simplification for Children. *In Proceedings of the Workshop on Accessible Search Systems, in conjunction with SIGIR 2010*.

CANDITO, M., NIVRE, J., DENIS, P. et ANGUIANO, E. (2010). Benchmarking of statistical dependency parsers for French. *In Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, pages 108–116. Association for Computational Linguistics.

CARROLL, J., MINNEN, G., PEARCE, D., CANNING, Y., DEVLIN, S. et TAIT, J. (1999). Simplifying Text for Language-Impaired Readers. *In Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 269–270.

CATACH, N. (1985). *Les listes orthographiques de base du français*. Nathan, Paris.

CHALL, J. et DALE, E. (1995). *Readability Revisited : The New Dale-Chall Readability Formula*. Brookline Books, Cambridge.

CHANDRASEKAR, R., DORAN, C. et SRINIVAS, B. (1996). Motivations and methods for text simplification. *In Proceedings of the 16th conference on Computational linguistics*, pages 1041–1044.

Denis, P, Sagot, B. *et al.* (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. *In Proceedings of PACLIC*.

Gillick, D. et Favre, B. (2009). A scalable global model for summarization. *In Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, ILP '09, pages 10–18, Stroudsburg, PA, USA.

Heilman, M. et Smith, N. A. (2010). Extracting Simplified Statements for Factual Question Generation. *In Proceedings of the 3rd Workshop on Question Generation*.

Inui, K., Fujita, A., Takahashi, T., Iida, R. et Iwakura, T. (2003). Text simplification for reading assistance : a project note. *In Proceedings of the second international workshop on Paraphrasing*, pages 9–16.

Jonnalagadda, S., Tari, L., Hakenberg, J., Baral, C. et Gonzalez, G. (2009). Towards Effective Sentence Simplification for Automatic Processing of Biomedical Text. *In Proceedings of NAACL-HLT 2009*.

Levy, R. et Andrew, G. (2006). Tregex and tsurgeon : tools for querying and manipulating tree data structures. *In Proceedings of the fifth international conference on Language Resources and Evaluation*, pages 2231–2234.

Lin, J. et Wilbur, W. J. (2007). Syntactic sentence compression in the biomedical domain : facilitating access to related articles. *Information Retrieval*, 10(4):393–414.

Medero, J. et Ostendorf, M. (2011). Identifying Targets for Syntactic Simplification. *In Proceedings of the SLaTE 2011 workshop*.

Nelken, R. et Shieber, S. (2006). Towards robust context-sensitive sentence alignment for monolingual corpora. *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 161–168.

Patel, V., Branch, T. et Arocha, J. (2002). Errors in interpreting quantities as procedures : The case of pharmaceutical labels. *International journal of medical informatics*, 65(3):193–211.

Petersen, S. E. et Ostendorf, M. (2007). Text Simplification for Language Learners : A Corpus Analysis. *In Proceedings of Speech and Language Technology in Education (SLaTE2007)*, pages 69–72.

Richard, J., Barcenilla, J., Brie, B., Charmet, E., Clement, E. et Reynard, P. (1993). Le traitement de documents administratifs par des populations de bas niveau de formation. *Le Travail Humain*, 56(4):345–367.

Siddharthan, A. (2006). Syntactic Simplification and Text Cohesion. *Research on Language & Computation*, 4(1):77–109.

Specia, L. (2010). Translating from Complex to Simplified Sentences. *In Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language (Propor-2010).*, pages 30–39.

Willms, J. (2003). Literacy proficiency of youth : Evidence of converging socioeconomic gradients. *International Journal of Educational Research*, 39(3):247–252.

Woodsend, K. et Lapata, M. (2011). Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK.

Zhu, Z., Bernhard, D. et Gurevych, I. (2010). A Monolingual Tree-based Translation Model for Sentence Simplification. *In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China.

Étude comparative entre trois approches de résumé automatique de documents arabes

Iskandar Keskes[1,2] *Mohamed Mahdi Boudabous*[1] *Mohamed Hédi Maaloul*[1,3]
Lamia Hadrich Belguith[1]

(1) ANLP Research Group, Laboratoire MIRACL, Route de Tunis Km 10, BP 242, Sfax, Tunisie
(2) Laboratoire IRIT, 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France
(3) Laboratoire LPL, 5 avenue Pasteur, BP 80975, 13604 Aix-en-Provence, France
Keskes@irit.fr, mehdiboudabous@gmail.com
mohamed.maaloul@lpl-aix.fr, l.belguith@fsegs.rnu.tn

RÉSUME

Dans cet article, nous proposons une étude comparative entre trois approches pour le résumé automatique de documents arabes. Ainsi, nous avons proposé trois méthodes pour l'extraction des phrases les plus représentatives d'un document. La première méthode se base sur une approche symbolique, la deuxième repose sur une approche numérique et la troisième se base sur une approche hybride. Ces méthodes sont implémentées respectivement par le système ARSTResume, le système R.I.A et le système HybridResume. Nous présentons, par la suite, les résultats obtenus par les trois systèmes et nous procédons à une étude comparative entre les résultats obtenus afin de souligner les avantages et les limites de chaque méthode. Les résultats de l'évaluation ont montré que l'approche numérique est plus performante que l'approche symbolique au niveau des textes longs. Mais, l'intégration de ces deux approches en une approche hybride aboutit aux résultats les plus performants dans notre corpus de textes.

ABSTRACT

Comparative study of three approaches to automatic summarization of Arabic documents

In this paper, we propose a comparative study between three approaches for automatic summarization of Arabic documents. Thus, we proposed three methods for extracting most representative sentences of a document. The first method is based on a symbolic approach, the second is relied on a numerical approach and the third is based on a hybrid approach. These methods are implemented respectively by the ARSTResume, R.I.A and HybridResume systems. Then, we present the results obtained by the three systems and we conduct a comparative study between the obtained results in order to highlight the advantages and limitations of each method. The evaluation results showed that the numerical approach has better performances than the symbolic approach. But, combining into a hybrid approach achieved the best results for our text corpus.

MOTS-CLES : Résumé automatique, approche symbolique, approche numérique, approche hybride, document arabe.

KEYWORDS: Automatic summarization, symbolic approach, numerical approach, hybrid approach, Arabic document.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 225–238,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Le Traitement Automatique du Langage Naturel (TALN) nous montre que les approches peuvent être convergées pour résoudre le même problème. Chaque approche a ses propres avantages et inconvénients qui peuvent être identifiés par une étude comparative.

Le présent travail présente une étude comparative entre différentes approches de TALN, dans le cadre du résumé automatique de textes.

Ce domaine aide à contribuer à une meilleure compréhension de la façon dont les gens produisent et comprennent la langue, car il peut résoudre les besoins croissants d'information de synthèse dans notre société.

La tâche de résumé semble être intrinsèquement interprétée dans le sens où différentes personnes produisent généralement des résumés très différents pour un texte donné. Ainsi, la qualité des résumés peut être jugée très différemment (Iria et al., 2007).

En matière de résumé automatique, on peut distinguer trois principales approches à savoir, l'approche par compréhension appelée l'approche symbolique, l'approche par extraction appelée l'approche numérique et l'approche qui combine les deux approches précédentes appelée l'approche hybride. L'approche symbolique exploite un savoir purement linguistique, et plus précisément sémantique pour extraire les phrases pertinentes d'un document (Azmi et Al-Thanyyan, 2012). Plusieurs théories entrent dans le cadre de cette approche à savoir : la Théorie de la Structure Rhétorique (RST) (Mann et Thompson, 1988), la Théorie de la Représentation Discursive (DRT) (Kamp, 1981 ; Kamp et Reyle, 1993), la Théorie de la Représentation Discursive Segmentée (SDRT) (Asher, 1993 ; Lascarides et Asher, 1993)… tandis que l'approche numérique repose sur un calcul de poids ou de scores associés à chaque phrase afin d'estimer son degré d'importance dans le texte. On distingue deux grandes techniques à savoir : la technique statistique (mots des titres, position des phrases,…) et la technique d'apprentissage (apprentissage supervisé, apprentissage semi-supervisé et apprentissage non supervisé) (Amini, 2001). L'extrait final contient les unités textuelles qui ont les scores les plus élevés. Concernant l'approche hybride, elle utilise des méthodes linguistiques et numériques pour extraire les phrases du résumé.

Nous proposons dans cet article une étude comparative entre les trois approches (symbolique, numérique et hybride). Cette étude a pour objectif d'évaluer la robustesse de chacune de ces approches ainsi que la mise en relief de leurs avantages et de leurs inconvénients pour le résumé automatique.

La suite de cet article se structure autour de cinq piliers. Le premier pilier présente la méthode symbolique pour le résumé automatique de documents arabes implémentée dans le système ARSTResume. Le deuxième pilier présente la méthode numérique pour le résumé automatique de documents arabes implémentée dans le système R.I.A. Le troisième pilier présente la méthode hybride pour le résumé automatique de documents arabes implémentée dans le système HybridResume. Le quatrième pilier expose le corpus

d'évaluation, l'évaluation de ces trois systèmes et les résultats obtenus. Enfin, le cinquième, montre une étude comparative entre les trois approches.

2 Méthode symbolique proposée

Dans cette section, nous présentons la méthode symbolique que nous proposons pour le résumé automatique de documents arabes, ainsi qu'une description détaillée des différentes étapes de cette méthode (Keskes, 2011).

2.1 Présentation

La méthode symbolique proposée pour le résumé automatique des documents arabes se base principalement sur des techniques d'extraction moyennant des critères linguistiques. Elle repose sur la théorie de la structure rhétorique (RST) (Mann et Thompson, 1988). Il s'agit de détecter les relations sémantiques et les relations intentionnelles qui existent entre les segments d'un document. En effet, l'analyse rhétorique a pour but d'établir les relations et les dépendances ainsi que l'importance relative à des phrases ou propositions les unes par rapport aux autres (Keskes et Maâloul, 2010). Notre méthode se déroule en trois temps. D'abord, le repérage des relations rhétoriques entre les différentes unités minimales du texte dont l'une possède le statut de noyau – qui est le segment de texte primordial pour la cohérence – les autres ayant un statut de noyau ou de satellite, sont des segments optionnels. Ensuite, le dressage et la simplification de l'arbre RST. Enfin, la sélection des phrases noyaux formant le résumé final, selon type de relation rhétorique choisi pour l'extrait.

À l'issue de notre étude du corpus, formé de cent textes en langue arabe annotés par trois linguistes (ces derniers ont sélectionné les phrases pertinentes), nous avons pu repérer des *frames* de relations rhétoriques. Ces *frames* sont des règles rhétoriques formées par des signaux linguistiques. Ces signaux sont principalement des marqueurs linguistiques indépendants d'un domaine particulier pour le repérage des relations rhétoriques (Minel, 2002). Toutefois, ces marqueurs peuvent être répertoriés en deux types : indicateurs déclencheurs et indices complémentaires. Les indicateurs déclencheurs énoncent la présence d'une relation rhétorique. Les indices complémentaires sont recherchés dans un espace défini à partir de l'indicateur (dans le voisinage de l'indicateur). Ils peuvent ainsi agir, dans le contexte, afin de confirmer ou d'infirmer la relation rhétorique énoncée par l'indicateur déclencheur. Ces règles rhétoriques sont appliquées pour construire par la suite l'arbre rhétorique. À partir de notre corpus d'étude, nous avons énuméré vingt relations rhétoriques. La table 1 présente quelques relations :

Liste des relations rhétoriques	Condition / شرط
	Concession / استدراك
	Enumération / تفصيل
	Restriction / استثناء
	Confirmation / توكيد
	Réduction / تقليل
	Joint / ربط

	Evidence / قاعدة
	Négation / نفي

TABLE1 -Exemples de relations rhétoriques

Le *frame* suivant est utilisé pour détecter la relation rhétorique négation:

Nom de relation :	{négation /نفي }
Contrainte sur (1) :	Contient un/des indice(s) complémentaire(s){لكن , لكنني ,لكنهم ,لكنه }
Contrainte sur (2) :	Contient l'indice déclencheur {ليست , ليسوا , ليس , لن , ولم ,لم }
Position de l'indice déclencheur	Milieu
Unité retenue	(1)

TABLE 2 -*Frame* de la relation rhétorique négation

2.2. Description détaillée de la méthode

La mise en œuvre fonctionnelle de notre méthode est représentée par la figure 1. Elle repose sur une segmentation à différents niveaux (titres, sections, paragraphes, phrases) ainsi que sur une recherche basée sur les règles rhétoriques afin de détecter les relations rhétoriques. Ces règles rhétoriques sont utiles pour la construction de l'arbre rhétorique. Enfin, à travers le choix du type de résumé (i.e. résumé indicatif, résumé informatif, …), on précède à la simplification de l'arbre et à la sélection des phrases du résumé.

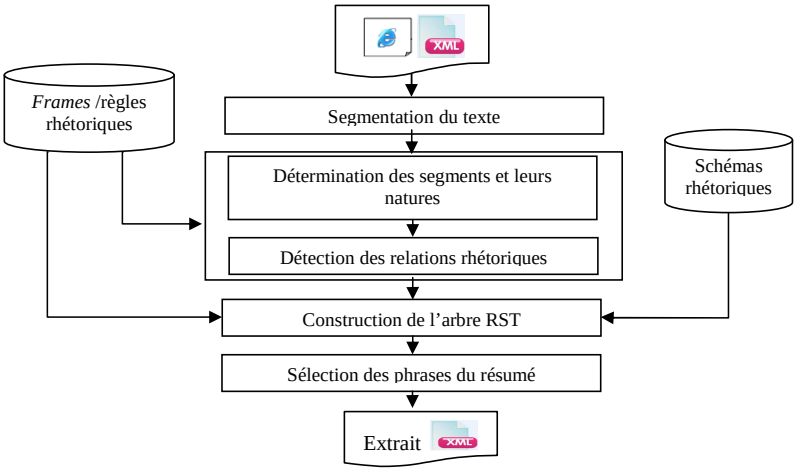


FIGURE1-Principales étapes de la méthode symbolique

2.1.1 Segmentation du document source

La segmentation du document est une étape nécessaire pour la tâche du résumé automatique. Cette étape consiste à hiérarchiser et à structurer le texte source en différentes unités (titres, sections, paragraphes et phrases).

Signalons, à ce niveau de traitement, une grande difficulté. En effet, La segmentation des textes en langue arabe ne peut pas se reposer uniquement sur la ponctuation puisqu'elle utilise certaines particules telles que " و " (waw) et "ف" (fā) et certains mots connecteurs pour séparer entre les phrases (Belguith et al., 2005).

Pour notre corpus constitué de textes en format HTML, nous utilisons une segmentation basée sur les signes de ponctuation et sur un ensemble de balises HTML. Cette étape de segmentation fournit en sortie un texte en format XML enrichi avec des balises encadrant les titres : <عنوان>…</عنوان>, les sections : <جزء>…</جزء>, les paragraphes : <فقرة>…</فقرة > et les phrases : <جملة>…</جملة>.

La deuxième étape de la segmentation est la segmentation des phrases en unités minimales, en utilisant les indicateurs principaux des règles rhétoriques, afin de descendre à un niveau plus bas dans l'analyse et de mieux dégager les relations. Ces dernières sont encadrées par les balises <قطاع>…</قطاع> (Tofiloski et al., 2009).

2.1.2 Application des règles rhétoriques

L'application des règles rhétoriques à un double but : déterminer la nature des segments (noyau ou satellite) et détecter les relations rhétoriques entre ces segments.

2.1.2.1 Détermination du segment Noyau et Satellite

Cette étape consiste à repérer les indicateurs principaux dans les phrases déjà segmentées et à préciser leurs positions dans l'unité minimale afin d'appliquer les règles rhétoriques, en cherchant les indices complémentaires.

Dans cette étape, nous allons donner, pour chaque unité minimale, un statut qui indique l'importance de cette unité par rapport à la phrase ou pour lui donner plus d'importance par rapport à une autre unité minimale. Le statut peut être un noyau ou un satellite.

Le noyau est un segment de texte qui comporte une information très pertinente. C'est un élément essentiel pour comprendre l'intention de l'auteur. Lorsqu'on élimine le noyau, nous ne pouvons pas comprendre le sens de la phrase. De même, un satellite est un segment de texte, mais qui comporte une information moins pertinente que le noyau. Donc, le noyau est un segment de texte primordial pour la cohérence et le satellite est un segment optionnel.

2.1.2.2 Détection des relations rhétoriques

Cette étape consiste à chercher les indices complémentaires de validation au voisinage de l'indicateur principal, c'est-à-dire le segment qui contient l'indicateur principal et le segment qui le précède. C'est l'indicateur principal qui signale la relation rhétorique

entre ces deux segments et c'est le rôle des indices complémentaires de confirmer ou non cette relation et de valider aussi le statut des deux segments.

Cette technique nous permet une analyse plus profonde, en tenant compte de la spécificité de la langue arabe sachant qu'on a des relations qui peuvent donner des sens proches comme les relations "حصر" et "استثناء" et aussi "تفسير" et "تفصيل".

2.1.3 Construction de l'arbre RST

Une fois l'étape de détection du type des unités minimales et des différentes relations rhétoriques existantes est achevée, nous ajoutons à notre technique les schémas rhétoriques (Mann et Thompson, 1988) afin de spécifier la composition structurale du texte et construire l'arbre RST.

Ces schémas rhétoriques décrivent l'organisation structurelle d'un texte, quelque soit le niveau hiérarchique de ce dernier. Ils permettent de lier un noyau et un satellite, deux ou plusieurs noyaux entre eux, et un noyau avec plusieurs satellites (Marcu, 1999).

Ainsi, les schémas rhétoriques se présentent sous la forme de cinq modèles de schémas (figure 2) qui peuvent être utilisés récursivement pour décrire des textes de taille arbitraire.

Généralement, le schéma le plus utilisé est celui liant un satellite unique à un noyau unique représenté dans la figure 3.

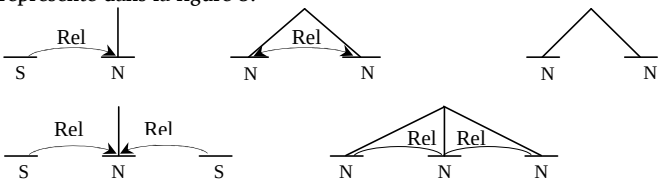


FIGURE 2 -Schéma rhétorique de base de la technique RST (Mann et al., 1988)

En plus des schémas rhétoriques, nous avons utilisé d'autres règles que nous avons dégagées suite à une étude empirique. Ces règles ont été validées par des linguistes. Elles permettent de traiter aussi les cas où nous n'avons pas de relations entre les phrases et assurent ainsi le maximum de couverture de texte que possible.

Afin de déterminer l'arbre RST le plus approprié pour le texte, nous avons essayé d'étudier le texte et la manière dont l'auteur l'a écrit. En effet, les auteurs veulent principalement donner un message aux lecteurs. Ce message est mentionné comme plusieurs faits; cependant, l'étude que nous avons faite sur le corpus prouve que les auteurs tendent à mentionner ces faits dans l'ordre, et chaque fait est suivi par des rapports qui le soutiennent. À travers cette étude empirique, nous avons pu dégager des règles de construction d'arbre RST représentées dans la table 3 :

Règles	Si (Indice principal est au début de la phrase) alors la relation détectée relie cette phrase avec la phrase précédente.
	Si (Indice principal est à la fin de la phrase) alors le segment qui contient cet indice est le seul qui contribue à la définition de la relation.
	Si (on a une ou plusieurs phrases qui n'admettent pas de relation entre elles) et (l'indice principal qui les suit est au début de la phrase) alors La relation relie toutes les phrases qui précèdent cet indice avec la phrase où il se trouve.

TABLE3 –Exemple de règles de construction d'arbres

Prenons par exemple la première règle, elle exprime le fait que s'il existe un marqueur principal, qui déclenche une relation rhétorique, situé au début de la phrase, alors cette relation relie entre le segment qui contient le marqueur principal et la phrase qui la précède. Car, sémantiquement, cette relation doit être subordonnante ou coordonnante de la relation rhétorique qu'elle précède et non pas le segment qu'il précède (Keskes et al., 2010b).

2.1.4 Sélection des phrases du résumé

Une fois l'arbre généré, nous allons faire l'élagage (simplification de l'arbre) selon le type de résumé indicatif ou selon les relations choisies par l'utilisateur tout en tenant compte des segments noyaux.

Tous les noyaux ne sont pas d'égale importance. En effet, l'étape de sélection des unités minimales importantes (noyaux), profite des relations entre les structures de discours pour décider du degré de leur importance. L'extrait final affiche les unités noyaux retenues après la simplification de l'arbre RST.

La simplification de l'arbre, prendra en considération la liste des relations retenues par l'utilisateur. Au cas où ce dernier ne précise aucun choix, le système détermine automatiquement les relations retenues pour le type de résumé indicatif. En effet, la réduction de l'arbre RST se fait par la suppression de tous les descendants qui viennent d'une relation rhétorique non choisie par l'utilisateur (Keskes et al., 2010a).

Cette méthode proposée a été implémentée dans le système ARSTResume.

3 Méthode numérique proposée

Dans cette section nous présentons la méthode numérique proposée pour le résumé automatique de documents arabes, ainsi qu'une description détaillée des différentes étapes de cette méthode.

3.1 Présentation

La méthode numérique pour le résumé automatique, d'articles de journaux en langue arabe, se base sur une technique d'apprentissage. Plus précisément, elle est basée sur la technique d'apprentissage semi-supervisé, qui se compose de deux phases à savoir :

La phase d'apprentissage qui permet au système d'apprendre à extraire les phrases du résumé. Cette phase se compose de deux étapes, une étape de segmentation et d'annotation, et une étape d'apprentissage.

La deuxième phase est la phase d'utilisation qui permet aux utilisateurs de résumer un nouveau document. Cette phase est composée de deux étapes, une étape de segmentation et d'annotation et une étape de classification (Boudabous et al., 2010).

Les différentes phases de notre méthode sont illustrées dans la figure 3.

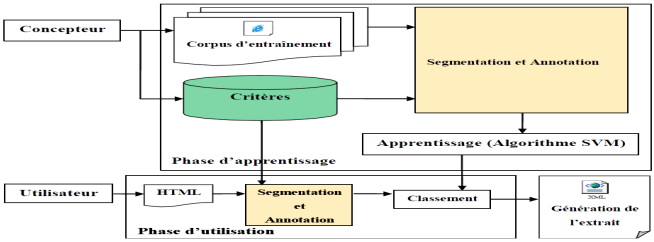


FIGURE 3 -Principales étapes de la méthode numérique

3.2 Description détaillée de la méthode

3.2.1 Phase d'apprentissage

La phase d'apprentissage nécessite l'utilisation d'un corpus d'entraînement ainsi qu'une base de critères d'extraction.

Le corpus d'entraînement est constitué de cent documents étiquetés (textes sources et leurs résumés) en format HTML (au moyen de trois pages par document). Les résumés de référence sont faits par trois experts humains afin d'apprendre au système comment produire des résumés similaires à ceux des experts humains de manière automatique.

Les critères d'extraction sont utilisés pour annoter les phrases des documents constituant notre corpus d'entraînement.

Nous avons classé les critères dans deux classes : les critères positionnels et les critères lexicaux. Ces derniers associent un score normalisé à chaque phrase, par contre les critères positionnels classent les phrases selon leurs postions dans le texte, présentés dans la table 4.

Critères positionnels	Position_ ph_texte	Classe la phrase selon sa position dans le texte : 1 si la phrase est dans le premier tiers du texte, 2 si elle est dans le deuxième tiers et 3 autrement.
	Position_ ph_sec	Classe la phrase selon sa position dans la section : 1 si la phrase est dans le premier tiers du texte, 2 si elle est dans le deuxième tiers et 3 autrement.
Critères lexicaux	Nb_mot_titre	Calcule le nombre d'apparition des mots du titre dans la phrase.
	Nb_exp_bonus	Calcule le nombre d'expressions bonus dans la phrase.
	Tf*Idf	Calcule le score tf*idf de la phrase.

TABLE 4 -Critères d'extraction

- **Segmentation et annotation du corpus**

Cette étape aboutit à la construction d'un vecteur d'extraction pour chaque unité du texte. L'ensemble des vecteurs d'extraction forme un fichier d'entrée pour l'étape d'apprentissage. La sous étape segmentation a pour but de découper le texte en unités minimales. Nous avons adopté la même segmentation utilisée dans la méthode symbolique. Concernant la sous étape d'annotation, l'acte annotatif consiste à donner une valeur ou un jugement à un segment du texte en se référant aux critères d'extraction. Cette étape a pour but d'annoter chaque segment du texte selon les différents critères d'extraction présentés précédemment. Chaque phrase de la collection est décrite par vecteur d'extraction, où la valeur donnée d'un critère correspond à la valeur d'analyse de la phrase selon ce critère.

- **Étape d'apprentissage**

L'algorithme d'apprentissage utilisé est l'algorithme SVM (Machines à Vecteurs de Support). Le choix de cet algorithme se justifie par sa robustesse de classification binaire, sa vitesse d'exécution et son adaptation aux problèmes non linéairement séparables. Cet algorithme génère une seule règle d'extraction appelée équation de l'hyperplan qui sépare les phrases pertinentes des phrases non pertinentes. Ainsi, l'algorithme d'apprentissage élimine les critères qui sont inutiles pour la phase d'apprentissage.

3.2.2 Phase d'utilisation

Cette phase permet à l'utilisateur du système de bénéficier des résultats de la phase d'apprentissage pour résumer un nouveau document. Les étapes par lesquelles passe le

texte à résumer sont : l'étape de segmentation et d'annotation, et l'étape de classement. L'étape de classification prend comme entrées les vecteurs d'extraction générés par l'étape de segmentation et d'annotation et l'équation de l'hyperplan générée par la phase d'apprentissage. L'équation de l'hyperplan est utilisée pour calculer le score de chaque phrase en se basant sur les vecteurs d'extraction. Cette méthode a été implémentée dans le système Résumeur Intelligent Arabe (R.I.A) (Boudabous et al., 2010).

4 Méthode hybride proposée

Dans cette section, nous proposons une méthode hybride pour le résumé automatique. Elle consiste à coupler la méthode linguistique et la méthode numérique.

4.1 Présentation

La méthode hybride, pour le résumé automatique des documents arabes, consiste à combiner la méthode symbolique basée sur la RST et la méthode numérique à base d'apprentissage. La figure 4 illustre le principe de cette méthode.

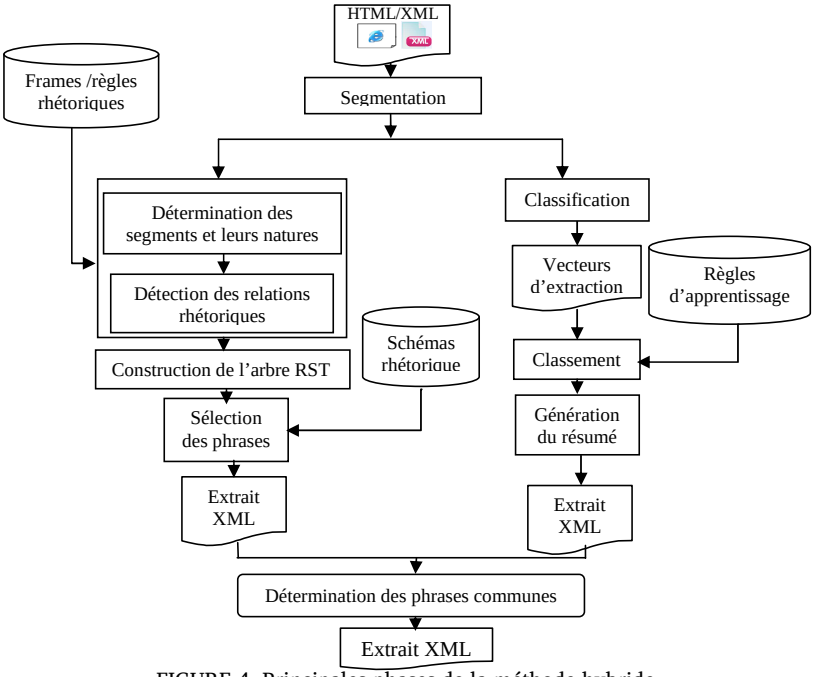


FIGURE 4 -Principales phases de la méthode hybride

4.2 Description détaillée de la méthode hybride

La méthode hybride que nous proposons se base sur la méthode symbolique et la méthode numérique, qui ont en commun le corpus d'étude et l'étape de segmentation des textes. Ces deux méthodes sont exécutées simultanément (en parallèle) comme nous l'avons décrit ci-dessus (section 2.2 et 3.2), puis, nous avons ajouté une étape de combinaison des résultats des deux méthodes.

L'étape de combinaison consiste à sélectionner les phrases communes des deux résumés générés par la méthode symbolique et la méthode numérique. Cette combinaison permet d'avoir un seul résumé pour chaque texte qui contient les phrases sélectionnées à la fois par la méthode symbolique et par la méthode numérique.

L'implémentation de cette méthode est basée sur l'intégration des deux systèmes ARSTRésume et R.I.A., à laquelle nous avons ajouté l'étape de combinaison. Le système développé s'appelle HybridResume.

5 Évaluation

Le corpus d'évaluation est formé de cent articles de presse, en langue arabe, rapatriés du journal Dar El Hayet[1] sans restriction quant a leurs contenu, taille, domaine et auteur. Ainsi, nous avons procédé à l'évaluation de la performance et de la pertinence des résumés générés par les trois systèmes, à l'aide d'une étude comparative qui mettra en jeu les résultats générés par les systèmes avec ceux réalisés par trois experts humain.

Nous avons utilisé le même corpus d'évaluation pour évaluer les trois systèmes (ARSTResume, R.I.A et HybridResume). Notons que ces trois systèmes ont utilisé le même module de segmentation pour avoir le même ensemble de phrases à traiter.

Nous avons procédé à trois expérimentations pour évaluer les trois systèmes. Chaque expérimentation compare les résumés de nos systèmes avec un résumé de l'expert. Le tableau suivant présente la moyenne de rappel, de précision et de f-mesure pour chacun des trois systèmes par rapport aux trois experts.

	ARSTResume			R.I.A.			HybridResume		
	Rappel	Précision	F-mesure	Rappel	Précision	F-mesure	Rappel	Précision	F-mesure
Expert 1	0.52	0.58	0.52	0.59	0.62	0.6	0.52	0.66	0.63
Expert 2	0.39	0.62	0.46	0.53	0.7	0.6	0.58	0.74	0.7
Expert 3	0.5	0.59	0.51	0.63	0.7	0.66	0.6	0.79	0.71
Moyenne	0.47	0.6	0.5	0.58	0.67	0.62	0.57	0.73	0.68

TABLE 5 –Résultats d'évaluation des trois systèmes

[1] Source : http://www.daralhayat.com

Nous remarquons que l'approche numérique est plus performante que l'approche symbolique et qu'HybridResume surclasse l'approche numérique sur ce corpus, et ce pour les 3 types de mesures effectuées.

6 Discussion des résultats obtenus

Suite à l'évaluation des trois systèmes, nous avons obtenu comme valeurs moyennes de rappel, de précision et de F-Mesure respectivement : 47%, 60% et 50% pour le système ARSTRésume, 58%, 67% et 62% pour le système R.I.A et 57%, 73% et 68% pour le système HybridResume. Nous remarquons, que ces mesures différentes d'un système à un autre et d'un expert à l'autre. Cela se justifie par le fait que chaque système à sa propre méthode, et que le résumé avec lequel nous faisons la comparaison dépend du jugement vis-à-vis du domaine d'intérêt de l'expert.

En comparant les mesures des trois systèmes simultanément, nous avons remarqué que le système HybridResume présente toujours les mesures les plus élevées. Voyons d'où cela provient en comparant les deux systèmes ARSTRésume et R.I.A.

En examinant ses mesures calculées sur le corpus d'évaluation pour chacun des deux systèmes, ARSTRésueme et R.I.A, nous avons remarqué que plus le texte est long, plus le système ARSTRésume présente les mesures de rappel et de précision les plus élevées. En effet, cette déduction se justifie par le fait que plus le texte est long, plus il contient de marqueurs linguistiques et de relations rhétoriques. Par conséquent, le système ARSTRésume fait le maximum de couverture pour générer un extrait semblable à celui réalisé par l'expert humain.

A contrario, le système R.I.A., présente ses mesures de rappel et de précision, les plus élevées lorsque le texte est court, car, plus le texte est long, plus nous avons un calcul complexe qui diminue la performance du système.

HybridResume se comporte mieux en moyenne sur un corpus de texte bien distribué entre textes longs et courts, ce qui justifie ses meilleures performances.

7 Conclusion

L'étude, que nous avons présentée, s'inscrit dans le cadre des travaux de recherche effectués sur les résumés automatiques de documents arabes. Dans ce contexte, nous avons présenté trois méthodes différentes de résumé automatique (i.e. une méthode symbolique, une méthode numérique et une méthode hybride). Nous avons implémenté ces trois méthodes respectivement dans les trois systèmes ARSTResume, R.I.A et HybridResume.

Ces trois systèmes ont été évalués sur un même corpus d'évaluation composé de cent textes résumés par trois experts. L'évaluation, a montré que le système R.I.A produit des résultats meilleurs que ceux produits par le système ARSTResume. En effet, les mesures de précision sont respectivement de 60% et 67% pour les systèmes ARSTResume et R.I.A. La performance relative au système R.I.A par rapport au système ARSTResume s'explique

par la difficulté de l'analyse linguistique. En effet, l'absence de relations rhétoriques, la présence des mots ambigus et le manque d'informations morphologiques ont une influence négative sur les valeurs de rappel et de précision. Toutefois, le système HybridResume, qui implémente une méthode hybride, donne les meilleurs résultats (73% de précision).

Suite à cette étude comparative, Nous avons conclu que l'approche numérique est plus robuste que l'approche symbolique, lorsque le texte est court et que l'approche symbolique est plus robuste lorsque le texte est long. Par conséquent, nous trouvons que la combinaison de ces deux approches en une approche hybride donne de meilleurs résultats.

Comme perspective, nous envisageons d'introduire une analyse morphologique pour la méthode symbolique en vue de mieux repérer les relations rhétoriques et d'améliorer les performances des systèmes.

8 Bibliographie

AMINI M.R.(2001). Apprentissage Automatique et Recherche d'information: Application à l'extraction d'information de surface et au résumé de texte. Thèse de doctorat, université Paris-6 France.

ASHER N.(1993). Reference to Abstract Objects in Discourse. Kluwer Academic Publishers, Netherlands.

Azmi A.M. et Al-Thanyyan S.(2012). A Text Summarizer for Arabic. *Computer Speech & Language. ISSN :0885-2308.*

BELGUITH H.L., BACCOUR L. et MOURAD G.(2005). Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules. 12ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'2005), Dourdan, France, 6-10 juin 2005, pp 451–456.

BOUDABOUS M.M., MAALOUL, M.H. et BELGUITH H. L.(2010). Digital Learning for Summarizing ARABIC Documents . IceTAL, Islande.

IRAKY K., ZAKAREYA A. et FARAWILA A.(2011). Arabic Discourse Segmentation Based on Rhetorical Methods. International Journal of Electric & Computer Sciences IJECS-IJENS Vol: 11 No: 01.

IRIA C., SILVIA F., PATRICIA v., VIVALDI J., SANJUAN E. et TORRES-MORENO J. M.(2007). A new hybrid summarizer based on Vector Space Model, Statistical Physics and Linguistics. Lecture Notes in Computer Science 4827. 872-882. ISSN 0302-9743.

KAMP H. et REYEL U.(1993), From Discourse To Logic , Dordrecht Kluwer.

KAMP H.(1981). Evénements, représentations discursives et référence temporelle. Langages, p 34-64.

KESKES I.(2011). Résumé automatique de textes arabes basé sur une approche symbolique. Editeur : EUE. ISBN-13 : 978-3841780232

KESKES I. et MAALOUL M. H.(2010). Résumé automatique de documents arabes basé sur la technique RST . Conférence international de Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (TALN /RECITAL 2010), 12ème edition, Montréal – Canada.

KESKES I., MAALOUL M. H. et BELGUITH L. H.(2010) ,(a). التلخيص الآلي للنصوص العربية اعتمادا على نظرية البنية البلاغية . International Computing Conference in Arabic, 6ème édition, Hammamet – Tunisie, prix du Best Paper.

KESKES I., MAÂLOUL M. H., BELGUITH L. H. et BLACHE P.(2010), (b). Automatic summarization of Arabic texts based on RST technique. International Conference on Enterprise Information Systems, 12ème edition, Madeira – Portugal.

LASCARIDES A. et ASHER N.(1993), Temporal Interpretation, Discourse Relations, and Commonsense Entailment , Linguistics and Philosophy, 16(5).

MAÂLOUL M. H.(2007). Al Lakas El'eli / الآلي اللخاص : Un système de résumé automatique de documents arabes . IBIMA.

MANN W. C. et THOMPSON S. A.(1988). Rhetorical structure theory: Toward a functional theory of text organization . Text, 8(3), p 243 – 281.

MARCU D.(1999). Discourse trees are good indicator of importance in text, Advances in Automatic Text Summarization. p123 – 136.

MINEL J.L.(2002). Filtrage sémantique : du résumé automatique à la fouille de textes. Hermès Science Publications, Paris.

MOURAD G.(1999). La segmentation de textes par l'étude de la ponctuation. CIDE'99, Document Electronique Dynamique, p 155 – 171, Damas, Syrie.

NICOLAS U., AMINI M.R. et GALLINARI P.(2005). Résumé automatique de texte avec un algorithme d'ordonnancement . CORIA.

TOFILOSKI M., BROOKE J. et TABOADA M.(2009). A Syntactic and Lexical-Based Discourse Segmenter. In Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics.

Etude sémantique des mots-clés et des marqueurs lexicaux stables dans un corpus technique

Ann Bertels[1,2] *Dirk De Hertog*[2] *Kris Heylen*[2]
(1) ILT, KU Leuven, Dekenstraat 6, B-3000 Leuven (Belgique)
(2) QLVL, KU Leuven, Faculty of Arts, Blijde-Inkomststraat 21, B-3000 Leuven (Belgique)
ann.bertels@ilt.kuleuven.be, dirk.dehertog@arts.kuleuven.be,
kris.heylen@arts.kuleuven.be

RESUME

Cet article présente les résultats d'une analyse sémantique quantitative des unités lexicales spécifiques dans un corpus technique, relevant du domaine des machines-outils pour l'usinage des métaux. L'étude vise à vérifier si et dans quelle mesure les mots-clés du corpus technique sont monosémiques. A cet effet, nous procédons à une analyse statistique de régression simple, qui permet d'étudier la corrélation entre le rang de spécificité des mots-clés et leur rang de monosémie, mais qui soulève des problèmes statistiques et méthodologiques, notamment un biais de fréquence. Pour y remédier, nous adoptons une approche alternative pour le repérage des unités lexicales spécifiques, à savoir l'analyse des marqueurs lexicaux stables ou *Stable Lexical Marker Analysis* (SLMA). Nous discutons les résultats quantitatifs et statistiques de cette approche dans la perspective de la corrélation entre le rang de spécificité et le rang de monosémie.

ABSTRACT

Semantic analysis of keywords and stable lexical markers in a technical corpus

This article presents the results of a quantitative semantic analysis of typical lexical units in a specialised technical corpus of metalworking machinery in French. The study aims to find out whether and to what extent the keywords of the technical corpus are monosemous. A simple regression analysis, used to examine the correlation between typicality rank and monosemy rank of the keywords, points out some statistical and methodological problems, notably a frequency bias. In order to overcome these problems, we adopt an alternative approach for the identification of typical lexical units, called *Stable Lexical Marker Analysis* (SLMA). We discuss the quantitative and statistical results of this approach with respect to the correlation between typicality rank and monosemy rank.

MOTS-CLES : unités lexicales spécifiques, analyse des mots-clés, analyse des marqueurs lexicaux stables, sémantique quantitative, analyse de régression.
KEYWORDS : typical lexical units, Keyword Analysis, Stable Lexical Marker Analysis (SLMA), quantitative semantics, regression analysis.

1 Introduction

Cette communication s'inscrit dans le contexte d'une étude sémantique quantitative effectuée sur un corpus de textes techniques relevant du domaine technique spécialisé des machines-outils pour l'usinage des métaux. L'étude vise à vérifier si et dans quelle mesure les unités lexicales spécifiques du corpus technique sont monosémiques. Selon la

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 239–252,
Grenoble, 4 au 8 juin 2012.

Terminologie traditionnelle, qui adopte une approche onomasiologique et prescriptive, la langue spécialisée se caractérise par la monosémie et l'univocité (Wüster, 1991). Les termes de la langue spécialisée sont idéalement monosémiques, tandis que la polysémie est réservée aux mots de la langue générale. Cela aboutit à une double dichotomie, qui oppose les termes de la langue spécialisée aux mots de la langue générale et la monosémie à la polysémie. Récemment, l'idéal de monosémie dans la langue spécialisée ainsi que la double dichotomie ont été remis en question par les partisans de la terminologie descriptive et linguistique, sémasiologique, distributionnelle et (con)textuelle (Bourigault et Slodzian, 1999 ; Cabré, 2000 ; Temmerman, 2000 ; Gaudin, 2003). Par ailleurs, des expérimentations ponctuelles menées sur des corpus spécialisés ont abouti à l'observation de cas de polysémie dans la langue spécialisée, même à l'intérieur d'un domaine spécialisé (Condamines et Rebeyrolle, 1997 ; Eriksen, 2002 ; Ferrari, 2002). Ces études sémantiques ponctuelles et qualitatives ainsi que les remises en question théoriques nous ont incités à évaluer la thèse monosémiste traditionnelle à grande échelle, dans un corpus de textes techniques, et à adopter une approche alternative, quantitative et scalaire.

Afin de procéder à une étude sémantique à grande échelle, à partir d'une analyse de corpus, il est nécessaire de reformuler la thèse monosémiste traditionnelle qualitative en une question de recherche quantitative. S'il est vrai que les unités lexicales d'un corpus spécialisé sont monosémiques, ce sera d'autant plus vrai pour les unités lexicales les plus spécifiques et les plus représentatives de ce corpus. Nous examinons donc si les unités lexicales les plus spécifiques du corpus technique sont effectivement les plus monosémiques, comme le prétendent les partisans de la terminologie traditionnelle, ou s'il existe des unités lexicales spécifiques qui sont polysémiques, comme le suggèrent les partisans de la terminologie descriptive. A cet effet, nous procédons à une double analyse quantitative. Pour l'extraction des unités lexicales spécifiques et pour le calcul de leur degré de spécificité, nous recourons à l'analyse des mots-clés (*Keyword Analysis*) (Scott, 2006). Pour quantifier l'analyse sémantique, nous calculons le degré de monosémie des mots-clés à partir du degré de recoupement de leurs cooccurrents de deuxième ordre, par le biais d'une mesure de recoupement (Bertels *et al.,* 2010). Ces données quantitatives de spécificité et de monosémie mènent ensuite à une analyse statistique de régression simple. Elle consiste à étudier la corrélation entre le rang de spécificité et le rang de monosémie, pour ainsi répondre à la question de recherche quantitative. Les résultats de l'analyse statistique confirment les observations des études sémantiques antérieures et permettent de réfuter la thèse monosémiste traditionnelle, comme nous l'avons décrit précédemment (Bertels *et al.,* 2010). Dans cet article, nous discutons la pertinence de l'analyse des mots-clés pour l'identification des unités spécifiques et nous proposons une approche méthodologique alternative.

Il est à noter que cet article ne se situe pas dans le domaine de l'extraction de termes. Cette discipline se caractérise par un classement catégoriel (termes vs non-termes), alors que nous visons à étudier la spécificité dans une perspective graduelle. Nous n'avons pas l'intention d'extraire la terminologie du domaine technique des machines-outils pour l'usinage des métaux. Nous cherchons avant tout à répondre à notre question de recherche, soulevée par le corpus de langue spécialisée, qui consiste à étudier la corrélation entre la spécificité et la monosémie.

Dans cet article, nous expliquons d'abord la méthodologie et les résultats de l'étude sémantique des mots-clés du corpus technique (section 2), ainsi que les problèmes statistiques et méthodologiques (section 3). Ensuite, nous présentons une autre approche pour l'identification des unités lexicales spécifiques et pour le calcul de leur degré de spécificité, à savoir l'analyse des marqueurs lexicaux stables ou *Stable Lexical Marker Analysis* (section 4). Nous discutons finalement les résultats quantitatifs et statistiques de cette approche dans la perspective de la corrélation entre le rang de spécificité et le rang de monosémie (section 5).

2 Etude sémantique des mots-clés du corpus technique

Le corpus technique constitué dans le cadre de cette étude relève du domaine spécialisé des machines-outils pour l'usinage des métaux et il comprend 1,7 million d'occurrences. Il a été étiqueté par Cordial 7 Analyseur et consiste en 4 sous-corpus, datant de 1996 à 2002, à savoir des revues électroniques (800.000 occurrences), des fiches techniques (300.000), des normes ISO et directives (300.000) et 4 manuels numérisés (360.000). Le corpus de référence de langue générale compte 15,3 millions d'occurrences lemmatisées et il est constitué d'articles du journal *Le Monde* de la même période (1998).

2.1 Identification des unités lexicales spécifiques

Pour le repérage des unités lexicales spécifiques, plusieurs approches méthodologiques sont envisageables. Elles permettent de générer une liste d'unités spécifiques, pourvues d'une indication de leur degré de spécificité. Les différences les plus importantes résident dans la méthodologie et les mesures statistiques sous-jacentes. La méthodologie du calcul des spécificités (Lafon, 1984 ; Labbé et Labbé, 2001) est basée sur la distribution hypergéométrique et sur le test statistique de Fisher Exact[1]. Elle est implémentée notamment dans Lexico3[2], Hyperbase[3] et TermoStat[4]. Du point de vue méthodologique, le calcul des spécificités procède par comparaison partie-tout. Une section d'un corpus est comparée au corpus entier dans le but d'identifier les mots spécifiques de la section par rapport au corpus entier. Le calcul des spécificités utilise seulement des informations appartenant au domaine en question. Le résultat est un coefficient de spécificité. Plus élevé ce coefficient, plus faible sera la probabilité de la fréquence observée (par rapport au corpus entier) et plus spécifique sera le mot. L'analyse des mots-clés (*Keyword Analysis*) (Scott, 2006) se caractérise par une approche contrastive, qui consiste à comparer les fréquences relatives des mots dans un corpus spécialisé à celles dans un corpus de référence de langue générale. Un mot est « clé » ou spécifique dans le corpus spécialisé si sa fréquence relative dans ce corpus est plus élevée que sa fréquence relative

[1] Le test statistique de Fisher Exact est généralement utilisé pour des données de taille modeste, des corpus peu volumineux et des fréquences plutôt faibles ($n < 20$).

[2] SYLED – CLA2T, Paris3 : http://www.tal.univ-paris3.fr/lexico/. [consulté le 20/01/2012].

[3] Hyperbase : http://ancilla.unice.fr/~brunet/pub/hyperbase.html. [consulté le 20/01/2012].

[4] TermoStat : http://olst.ling.umontreal.ca/~drouinp/termostat_web/doc_termostat/doc_termostat.html. [consulté le 20/01/2012]. Dans TermoStat le corpus de référence et le corpus d'analyse sont fusionné en un corpus virtuel, pour vérifier si le lexique du corpus d'analyse se comporte comme celui du corpus de référence.

dans le corpus de référence et si la différence de fréquence est statistiquement significative. Une mesure statistique, comme le rapport de vraisemblance (*Log-Likelihood Ratio* ou LLR ou G^2) (Dunning, 1993), permet de décider s'il s'agit d'un mot-clé du corpus spécialisé. L'analyse des mots-clés est implémentée notamment dans WordSmith[5], AntConc[6], TermoStat et AV Frequency List Tool[7].

Pour identifier le vocabulaire spécifique de notre corpus technique, nous adoptons l'analyse des mots-clés, parce qu'elle compare deux corpus différents. Nous recourons à la mesure statistique du log de vraisemblance (LLR), qui permet de conduire à des possibilités de classement précis et donc à des degrés de spécificité avec une granularité aussi fine que possible. Plusieurs études antérieures ont validé la mesure du LLR et démontré la pertinence de l'analyse des mots-clés pour relever les unités spécifiques d'un domaine particulier (Paquot *et al.,* 2009 ; Kwary, 2011). Nous nous limitons dans cette étude au niveau des unités simples, comme *fraisage, commande, machine*. Des recherches futures devront certainement porter sur l'étude des unités polylexicales, telles que *machine à fraiser*, *commande numérique*, puisque la plupart des unités terminologiques d'un domaine spécialisé se situent au niveau des unités complexes[8]. L'analyse des mots-clés est effectuée dans AV Frequency List Tool, à partir de deux listes de fréquence des lemmes des deux corpus, réalisées à l'aide de scripts en Python. Le logiciel AV génère une liste de lemmes spécifiques du corpus technique et indique leur degré de spécificité, à savoir la valeur du LLR (*keyness*), et une valeur p associée. Nous relevons 4717 mots-clés ($p < 0{,}05$), après suppression des mots grammaticaux, des noms propres et des hapax. Le degré de spécificité ou de *keyness* permet de situer ces 4717 mots-clés sur un continuum de spécificité et de leur accorder un rang de spécificité.

2.2 Quantification de l'analyse sémantique

Les 4717 mots-clés font ensuite l'objet d'une analyse sémantique quantitative et automatisée. A cet effet, nous recourons à l'analyse des cooccurrences (Grossmann *et al.,* 2003 ; Condamines 2005 ; Blumenthal et Hausmann, 2006 ; Mayaffre 2008), parce qu'elle permet de quantifier et d'objectiver la monosémie en l'implémentant en termes d'homogénéité sémantique (Habert *et al.,* 2005). Une unité lexicale monosémique

[5] WordSmith Tools : http://www.lexically.net/wordsmith/. [consulté le 20/01/2012].

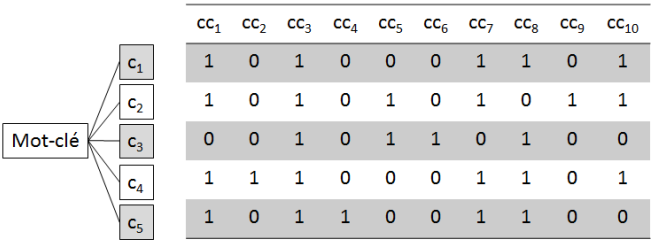
[6] AntConc : http://www.antlab.sci.waseda.ac.jp/software.html. [consulté le 20/01/2012].

[7] AV Frequency List Tool : http://wwwling.arts.kuleuven.be/av-tools/. [consulté le 20/01/2012].

[8] Plusieurs outils d'extraction terminologique, tels que LEXTER (Bourigault et al. 2001), permettent de repérer les unités polylexicales. Toutefois celles-ci posent problème lors du calcul des spécificités. Pour l'instant, il n'est pas possible de déterminer le degré de spécificité des unités complexes de façon fiable et statistiquement significative, notamment parce que la plupart d'entre elles sont absentes dans un corpus de référence de langue générale. Par ailleurs, les techniques d'extraction automatique de termes s'appuient souvent sur un algorithme hybride avec une composante syntaxique importante, c'est-à-dire des structures syntaxiques récurrentes (Lemay et al., 2005). Ainsi, plusieurs variables concourent au repérage des unités terminologiques complexes plutôt qu'une seule. Or, l'analyse de régression à laquelle nous procédons pour étudier la corrélation entre les données de spécificité et de monosémie, requiert une seule variable linguistique, c'est-à-dire un critère de spécificité clair et précis.

apparaît dans des contextes plutôt homogènes sémantiquement. Par contre, une unité lexicale polysémique se caractérise par des cooccurrents plus hétérogènes sémantiquement, qui appartiennent à des champs sémantiques différents (Véronis, 2003 ; Habert *et al.,* 2004). L'accès à la sémantique des cooccurrents de premier ordre (ou *c*) peut se faire à partir de leurs cooccurrents, c'est-à-dire les cooccurrents de deuxième ordre (ou *cc*). Ceux-ci se caractérisent principalement par des relations paradigmatiques avec le mot de base (hyponymes, hyperonymes, synonymes, antonymes) et dès lors ils sont intéressants pour caractériser sémantiquement le mot de base. Ils ont permis entre autres de mettre en évidence des relations de synonymie (Martinez, 2000).

Si les cooccurrents de premier ordre d'un mot de base, en l'occurrence un mot-clé, partagent beaucoup de cooccurrents de deuxième ordre, ces derniers se recoupent formellement, ce qui constitue une indication de l'homogénéité sémantique des cooccurrents de premier ordre et, dès lors, du mot de base. Le degré de monosémie d'un mot-clé pourra donc être déterminé en fonction du degré de recoupement formel de ses cooccurrents de deuxième ordre. Une représentation schématique (Cf. figure 1) fait intervenir un mot-clé, ses 5 c différents et tous leurs c (10 cc différents et 26 cc au total). Ce schéma permettra d'expliquer le poids de chaque cc pour le recoupement global. Un cc partagé par tous les c (p.ex. cc_3), figure 5 fois dans la liste des cc, constituée de 5 blocs de cc (un bloc par c). Le cc figurant 5 fois aura donc un poids maximal de 5/5. Par contre, un cc qui figure dans un seul bloc (p.ex. cc_2 ou cc_4) est un cc isolé avec un poids minimal de 1/5. De même, le poids d'un cc qui figure 2 fois dans la liste des cc ou dans 2 blocs (p.ex. cc_5) équivaut à 2/5. Ainsi, on pourra calculer facilement le poids de chaque cc dans la liste des 26 cc. Le poids de chaque cc correspond au rapport entre la fréquence du cc dans la liste des cc et le nombre de c. Pour connaître le recoupement global, calculé à partir du recoupement de tous les cc, on fera d'abord la somme des poids individuels (donc 26 réitérations du calcul précédent). Ce résultat sera divisé par 26 (nombre total de cc dans la liste), parce que chaque cc contribue pour 1/26 au recoupement global calculé pour le mot-clé. Le résultat final se situe toujours entre 0 et 1 et représente le degré de recoupement moyen pour un mot-clé, c'est-à-dire son degré d'homogénéité sémantique.



cc_1	cc_2	cc_3	cc_4	cc_5	cc_6	cc_7	cc_8	cc_9	cc_{10}
1	0	1	0	0	0	1	1	0	1
1	0	1	0	1	0	1	0	1	1
0	0	1	0	1	1	0	1	0	0
1	1	1	0	0	0	1	1	0	1
1	0	1	1	0	0	1	1	0	0

FIGURE 1 – Schéma : mot-clé + cooccurrents de premier et de deuxième ordre.

L'analyse des cooccurrences est effectuée, de façon récurrente, dans une fenêtre d'observation (ou *span*) de 5 mots à gauche et 5 mots à droite, sans informations de position ni d'orientation. Cette fenêtre apporte suffisamment d'informations sémantiques pertinentes, sans introduire trop de bruit, et permet un traitement informatique efficace.

Les cooccurrents de premier et de deuxième ordre sont considérés au niveau des formes graphiques, ce qui permet de faire la distinction entre, par exemple, *pièce usinée* (« résultat ») et *pièce à usiner* (« avant le processus d'usinage ») et dès lors de tenir compte de la différence sémantique. La mesure d'association utilisée pour déterminer les cooccurrents statistiquement pertinents est la mesure statistique du LLR (log du rapport de vraisemblance). Le seuil de significativité très sévère (valeur $p < 0{,}0001$) permet de relever uniquement les cooccurrents de premier et de deuxième ordre sémantiquement pertinents. La mesure de recoupement est concrétisée à l'aide de scripts en Python pour calculer le degré de recoupement des 4717 mots-clés à partir du recoupement formel de leurs cooccurrents de deuxième ordre. Ce degré de recoupement ou d'homogénéité sémantique permet de situer les 4717 mots-clés sur un continuum d'homogénéité sémantique ou de monosémie et permet de leur accorder un rang de monosémie.

Comme nous ne disposons pas de listes de sens préétablis, ni d'autres mesures sémantiques comparables, nous avons procédé à une validation manuelle de la mesure de recoupement à partir de l'analyse manuelle des cooccurrents les plus pertinents, ainsi qu'à une validation externe au moyen de dictionnaires. Les résultats de ces validations confirment les résultats de notre mesure de recoupement pour un échantillon de 50 mots-clés. Il est à noter que des recherches supplémentaires s'imposent pour examiner la relation précise entre, d'une part, notre mesure de monosémie, implémentant la monosémie en termes d'homogénéité sémantique, et, d'autre part, ce que l'on considère traditionnellement comme monosémie ou polysémie. Nous recourons à cette mesure, dans le but opérationnel de développer un critère mesurable. Sans recherches supplémentaires, il serait impossible d'affirmer que les degrés de monosémie calculés correspondent parfaitement à ce que les terminologues traditionnels considèrent comme monosémie ou polysémie. Notons toutefois que ces derniers omettent de fournir des critères opérationnels à ce sujet.

2.3 Corrélation entre la spécificité et la monosémie

Pour répondre à la question de recherche et pour examiner la corrélation entre le continuum de spécificité et le continuum de monosémie (ou d'homogénéité sémantique), les données quantitatives de spécificité et de monosémie sont soumises à une analyse statistique de régression linéaire simple. Celle-ci permet d'étudier l'impact d'une variable indépendante ou explicative (ici : le rang de spécificité) sur la variable dépendante ou expliquée (ici : le rang de monosémie). Le résultat de cette analyse est le coefficient de détermination ou le pourcentage de variation expliquée R^2. Il représente le pourcentage de la variation du rang de monosémie que l'on pourra expliquer ou prédire à partir de la variation du rang de spécificité des 4717 mots-clés.

Les résultats de l'analyse de régression simple permettent d'infirmer la thèse monosémiste traditionnelle, car ils montrent une corrélation négative (coefficient de corrélation Pearson de -0,72) et un pourcentage de variation expliquée R^2 de 51,57% (valeur $p < 2{,}2e^{-16}$). Il s'avère donc que les mots-clés les plus spécifiques du corpus technique ne sont pas les plus monosémiques, mais, au contraire, les plus hétérogènes sémantiquement (p.ex. *machine, pièce, tour*). En plus, les mots-clés les moins spécifiques du corpus technique sont les plus homogènes sémantiquement (par exemple *rationnellement, télédiagnostic*), à quelques exceptions près (*service* et *objet*). En effet, la

visualisation ci-dessous (Cf. figure 2) montre que la droite de régression s'incline vers le bas. Parmi les mots-clés spécifiques qui sont hétérogènes sémantiquement, nous retrouvons effectivement des unités polysémiques, telles que *découpe*, dont les sens « action de découper » et « résultat de la découpe » se caractérisent par une relation métonymique. Nous recensons également des homonymes (tels que *tour*), ainsi que des mots vagues (comme *usinage*) dont le sens sous-déterminé est précisé par le contexte linguistique.

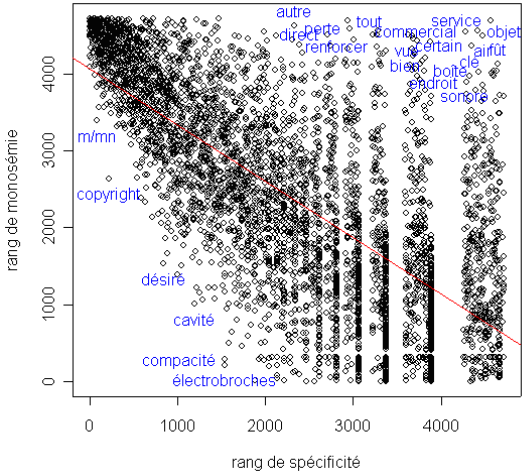


FIGURE 2 – Visualisation de l'analyse de régression.

3 Problèmes statistiques et méthodologiques

3.1 Hétéroscédasticité : mots généraux

La visualisation ci-dessus (Cf. figure 2) montre que la corrélation négative ne s'applique pas à tous les mots-clés et qu'elle n'est pas tout à fait linéaire. Le test statistique de Goldfeld-Quandt soulève effectivement un problème statistique d'hétéroscédasticité (statistique F du GQ-test 2,07), ce qui veut dire que les variances des erreurs ne sont pas constantes. En effet, certaines observations se caractérisent par un résidu important, c'est-à-dire par une grande différence entre leur valeur observée et la valeur estimée par le modèle de régression, par exemple *service, objet, commercial*. L'écart (ou le résidu) entre leur valeur observée (visualisée par la petite boule) et leur valeur estimée située sur la droite de régression est très important, ce qui donne lieu à une erreur importante lors de la prédiction de leur rang de monosémie à partir de leur rang de spécificité. Ces mots se situent principalement dans la partie supérieure droite, c'est-à-dire parmi les mots-clés les moins spécifiques. Ils sont plus polysémiques qu'on n'aurait cru en prenant en considération leur rang de spécificité.

Ce sont majoritairement des mots généraux, très fréquents dans le corpus de référence et dès lors peu spécifiques dans le corpus technique, en dépit de leur fréquence élevée dans le corpus technique. Pour ces mots, qui se trouvent dans la zone marginale de spécificité (valeur p légèrement inférieure à 0,05), le modèle de régression n'est pas une bonne prédiction de leur rang de monosémie à partir de leur rang de spécificité. Ces mots sont hétérogènes sémantiquement et se caractérisent par une polysémie à la fois générale et technique : leurs (divers) sens généraux se retrouvent aussi dans le corpus technique. Ils sont plutôt hétérogènes sémantiquement, quel que soit leur rang de spécificité.

3.2 Multicollinéarité : fréquence technique

Le deuxième problème est soulevé par une analyse statistique de régression multiple, qui évalue l'impact combiné et simultané de plusieurs variables indépendantes sur la variable dépendante. Parfois, deux ou plusieurs variables indépendantes sont corrélées les unes avec les autres. Elles expliquent en grande partie la même variation de la variable dépendante, ce que l'on qualifie de problème de multicollinéarité[9].

Pour les 4717 mots-clés, nous observons un problème de multicollinéarité pour trois variables, à savoir le log du LLR, le rang de spécificité et le rang de fréquence technique. En effet, il y a une corrélation (trop) importante (0,87) entre la valeur du LLR, utilisée pour identifier et classer les mots-clés, et la fréquence technique. Par ailleurs, la mesure statistique du LLR est trop sensible à la fréquence technique, car pour les fréquences techniques (très) élevées, elle gonfle la valeur du LLR et dès lors le degré de spécificité. Il s'ensuit que les mots très fréquents dans le corpus technique ont un degré de spécificité relativement plus élevé que les mots moyennement ou faiblement fréquents. Par conséquent, certains mots très fréquents se situent à tort parmi les mots les plus spécifiques. Bien entendu, dans l'analyse de régression simple, nous considérons le rang de spécificité, qui permet tout de même d'effacer les différences trop importantes en termes de degrés de spécificité. Notons également que la fréquence technique élevée de certains mots s'explique par leur fréquence très élevée dans une des parties du corpus, en dépit de leur fréquence plutôt normale dans les autres parties. Ce biais de fréquence local est souvent causé par un biais de sujet (*topical bias*). En effet, le calcul du degré de spécificité compare la fréquence relative dans le corpus technique entier (de 1,7 million de mots) à la fréquence dans le corpus général entier (de 15,3 millions de mots), sans tenir compte de la dispersion des mots à travers les corpus. Or, la prise en considération de la dispersion des mots s'avère importante lors de l'extraction des mots-clés (Paquot *et al.,* 2009). Comme notre corpus technique consiste en 4 sous-corpus (revues, fiches techniques, normes et manuels), cette hétérogénéité des sources aura probablement un impact sur la dispersion et la spécificité des unités lexicales spécifiques.

En conclusion, deux problèmes se posent. D'une part, il y a trop de mots généraux parmi les mots-clés et ils entraînent un effet perturbateur et de ce fait un problème statistique d'hétéroscédasticité. D'autre part, la mesure statistique du LLR est trop sensible à la fréquence technique élevée et elle souffre d'un biais de sujet.

[9] Valeurs VIF dans l'analyse de régression multiple : log du LLR (VIF 36,26), rang de spécificité (VIF 26,32) et rang de fréquence technique (VIF 14,72).

4 Solutions

Pour remédier à ces problèmes, nous proposons d'adopter une méthode alternative, qui permet de prendre en considération également la dispersion des mots à travers les corpus. Ainsi, on évite qu'un mot soit spécifique du corpus technique à cause de sa surreprésentation dans une seule partie. Or, la dispersion ne permet pas de résoudre tout le problème de la sensibilité à la fréquence. Par conséquent, nous adoptons également une autre mesure statistique, capable de refléter de façon plus fiable les unités lexicales spécifiques du corpus technique et leur degré de spécificité.

4.1 Stable Lexical Marker Analysis (SLMA)

La nouvelle méthode, appelée *Stable Lexical Marker Analysis* (SLMA) ou analyse des marqueurs lexicaux stables, a été développée dans le domaine de la linguistique variationnelle (Speelman *et al.,* 2006). Le but était d'identifier les variantes lexicales régionales typiques ou les « marqueurs lexicaux stables » des différences régionales entre le néerlandais utilisé aux Pays-Bas et en Flandre (Belgique) (Speelman *et al.,* 2008). La méthode s'applique aussi à l'extraction d'unités terminologiques, par exemple dans le domaine juridique de la législation financière (De Hertog *et al.,* 2010). La SLMA compare deux corpus à partir de leurs listes de fréquence et permet ainsi d'identifier les différences lexicales consistantes et stables entre les corpus. Elle s'inspire de la méthode des mots-clés de Scott (2006), en ce qu'elle consiste à comparer des listes de fréquence d'un corpus d'analyse à des listes de fréquence d'un corpus de référence. Toutefois, au lieu de comparer une liste de fréquence d'analyse à une liste de fréquence de référence, elle compare plusieurs fois de telles listes de fréquence. Elle fait donc intervenir de multiples tests d'hypothèse pour ainsi rendre compte de la dispersion.

En effet, le corpus spécialisé est subdivisé en plusieurs partitions (disons n partitions), tout comme le corpus de référence (m partitions). Pour chaque partition des deux corpus, on établit une liste de fréquence. Il y a donc $n*m$ listes de fréquence. Ensuite, chaque partition du corpus spécialisé A (p.ex. A_1, A_2, …, A_n) est comparée à chaque partition du corpus de référence B (p.ex. B_1, B_2, …, B_m), par le biais de leur liste de fréquence, ce qui revient à $n*m$ comparaisons de partitions. Chaque comparaison par paire de partitions permet de générer une liste de mots-clés spécifiques de la partition spécialisée (LLR et valeur $p < 0{,}05$). Les mots qui sont spécifiques dans la plupart de ces comparaisons (au maximum $n*m$) sont qualifiés de « marqueurs lexicaux stables », parce qu'ils sont stables et consistants à travers le corpus spécialisé entier. Le nombre de comparaisons significatives par paire de partitions (qualifié de SLM) est une première indication du degré de spécificité. Ces unités lexicales sont spécifiques (globalement relativement plus fréquentes dans le corpus spécialisé) et stables (avec une dispersion uniforme à travers le corpus spécialisé). Le découpage des corpus en partitions peut se réaliser à l'aide de scripts en Python, tout comme les multiples comparaisons des listes de fréquence.

4.2 Odds Ratio

La mesure statistique du log Odds Ratio (log OR), permet d'obtenir une indication de spécificité à granularité plus fine que le nombre de comparaisons significatives par paire de partitions (SLM). Le log OR permet également de prendre en considération la réelle

importance de la différence de fréquence d'un mot dans les deux (partitions de) corpus, ce que l'on qualifie de *effect size*. Le log OR fait intervenir la fréquence relative du mot, ainsi que celle de tous les autres mots, ce qui évite de gonfler le résultat pour les fréquences élevées (Cf. LLR). Pour un mot w_k donné, on calcule ainsi le score SMEA (*Stable Marker Effect size Analysis*), c'est-à-dire *SMEA(w_k, A, B)*, en calculant le log OR pour chaque comparaison significative, dans un corpus spécialisé *A* (*n* partitions) et un corpus de référence *B* (*m* partitions). La somme est divisée par le nombre total de comparaisons de partitions.

$$SMEA(w_k, A, B) = \frac{1}{n*m}\sum_{i=1}^{n}\sum_{j=1}^{m}(\log(\frac{F_{w_k}^{A^i} / F_{\neg w_k}^{A^i}}{F_{w_k}^{B^j} / F_{\neg w_k}^{B^j}}) * S(F_{w_k}^{A^i}, F_{\neg w_k}^{A^i}, F_{w_k}^{B^j}, F_{\neg w_k}^{B^j}))$$

avec $F_{w_k}^{A^i}$ la fréquence du mot w_k dans la partition *i* du corpus A et $F_{\neg w_k}^{A^i}$ la fréquence de tous les mots autres que w_k (et de même pour les fréquences du corpus B). *S( )* est une fonction booléenne qui égale 1 si la distribution des mots est significativement différente dans les corpus A et B ; sinon elle égale 0. Si le nombre de comparaisons significatives est plus élevé, donc si le mot spécifique est mieux dispersé, le score SMEA sera plus élevé. Le score SMEA est une indication de la spécificité du mot ainsi que de sa dispersion, mais elle échappe au gonflement pour les mots très fréquents. Sa granularité très fine permet de classer les marqueurs lexicaux stables et de déterminer leur nouveau rang de spécificité, appelé rang de SMEA. De Hertog *et al.* (2010) ont démontré la fiabilité de cette approche par l'extraction de candidats-termes à partir d'un corpus de textes juridiques européens et par leur validation contre la base de données terminologique officielle des services européens.

5 Etude sémantique des marqueurs lexicaux stables du corpus technique

5.1 Identification des marqueurs lexicaux stables

Pour déterminer les marqueurs lexicaux stables du corpus technique, celui-ci est subdivisé en 5 partitions, c'est-à-dire une partition par sous-corpus, pour les normes, les fiches et les manuels (entre 300.000 et 360.000 occurrences) et 2 partitions de 400.000 occurrences pour le sous-corpus des revues. Ces 5 partitions sont de taille comparable et raisonnable et respectent l'ordre des mots et les frontières des sous-corpus thématiques et stylistiques. Le corpus de référence de langue générale est réparti en 36 partitions de taille similaire à celle des partitions techniques (environ 400.000 occurrences). L'analyse de la SLMA est effectuée sur les lemmes au lieu des formes fléchies, à l'instar de l'analyse des mots-clés dans AV Frequency List Tool (Cf. section 2.1). Après l'extraction et avant l'interprétation, la liste des marqueurs lexicaux stables repérés subit le même traitement que la liste des mots-clés, à savoir la suppression des mots grammaticaux, des noms propres et des hapax. Dans le corpus technique, nous recensons ainsi 3479 marqueurs lexicaux stables, statistiquement significatifs ($p < 0,05$), dont 3123 formes (ou presque 90%) figurent aussi dans la liste des 4717 mots-clés.

5.2 Marqueurs lexicaux stables versus mots-clés

Le tableau ci-dessous (Cf. table 1) montre que les mots-clés les plus fréquents et les plus spécifiques comme *machine, outil, pièce*, visualisés dans la colonne de droite aux rangs de spécificité (LLR) 1, 2 et 4 respectivement, ne figurent pas parmi les marqueurs lexicaux stables les plus spécifiques, visualisés à gauche du tableau. En effet, ils se retrouvent respectivement aux rangs de SMEA 33, 55 et 170. On observe également que la prise en considération de la dispersion relègue certaines unités lexicales, très fréquentes dans les revues (p.ex. *Fig*) à un rang moins spécifique (37 au lieu de 9). Les vrais termes, qui sont spécifiques du domaine, occupent des rangs plus spécifiques (*usinage, broche, copeau, fraisage, serrage,* …). Ensuite, les mots généraux, qui ont des emplois généraux et techniques, occupent à juste titre des rangs un peu moins spécifiques (*machine, outil, pièce,* …). Enfin, les mots généraux peu fréquents et à peine spécifiques (i.e. la queue de la liste des 4717 mots-clés) ne se retrouvent pas parmi les 3479 marqueurs lexicaux stables. Il s'avère que la fréquence technique moyenne des 4717 mots-clés est plus faible (140,77) que celle des 3479 marqueurs lexicaux stables (182,16). Par ailleurs, la corrélation entre la fréquence dans le corpus technique et la valeur de SMEA, qui indique le degré de spécificité, est moins problématique dans la liste des marqueurs lexicaux stables (0,32) que dans la liste des 4717 mots-clés, où la corrélation entre la fréquence technique et la valeur du LLR était trop importante (0,87).

	lemme	**SMEA**	**SLM**	**fréq.tech.**			**mots-clés**
1	usinage	85,699726	180	6720		1	machine
2	broche	74,8200697	180	2893		2	outil
3	copeau	73,7392965	180	2557		3	usinage
4	fraisage	68,364653	180	1873		4	pièce
5	usiner	68,3239216	180	1577		5	mm
6	machine-outil	67,6419261	180	1005		6	vitesse
7	serrage	66,3394778	180	939		7	coupe
8	perçage	62,8188634	180	846		8	broche
9	fraise	62,0265842	180	1571		9	Fig
10	meule	61,8557297	180	776		10	axe

TABLE 1 – Top 10 des marqueurs lexicaux stables du corpus technique (à gauche), par rapport au top 10 des 4717 mots-clés (à droite).

5.3 Corrélation entre la spécificité et la monosémie

Pour étudier la corrélation entre le rang de spécificité (rang de SMEA) et le rang de monosémie des 3479 marqueurs lexicaux stables, nous procédons à une analyse statistique de régression simple. Elle montre une corrélation négative entre le rang de

spécificité et le rang de monosémie (-0,49). Il s'avère donc que les marqueurs lexicaux les plus stables et les plus spécifiques ne sont pas les plus monosémiques. Toutefois, la corrélation (-0,49) est moins convaincante que celle pour les 4717 mots-clés (-0,72). Le pourcentage de variation expliquée R^2 de 23,87% (valeur $p < 2{,}2e^{-16}$) est également moins convaincant que celui pour les 4717 mots-clés (51,57% et valeur $p < 2{,}2e^{-16}$). Ces résultats moins concluants pour les marqueurs lexicaux stables s'expliquent principalement par l'absence des mots généraux peu fréquents et très peu spécifiques, qui sont très monosémiques, et par le fait que les mots les plus fréquents occupent, à juste titre, des rangs moins spécifiques. En raison de leur fréquence plus élevée dans le corpus technique, ces derniers ont plus de chances d'être polysémiques et/ou de constituer la tête d'unités polylexicales, où ils sont désambiguïsés par les autres composants (par exemple *machine à fraiser, machine à rainurer*).

Notons que le test de Goldfeld-Quandt soulève aussi un problème d'hétéroscédasticité (statistique F du GQ-test 1,37), mais moins important que pour les 4717 mots-clés (2,07). Le problème de l'hétéroscédasticité est donc résolu en partie, mais suggère la présence d'une variable supplémentaire, cachée jusqu'à présent, qui prédit peut-être une partie de la variation du rang de monosémie. Cette variable pourrait être liée au fait que les mots spécifiques constituent la tête d'unités polylexicales dans le corpus technique. Des recherches futures permettront de vérifier si elle permet d'expliquer l'hétéroscédasticité et dans quelle mesure.

6 Conclusion

Dans cet article, nous avons étudié les unités lexicales spécifiques d'un corpus technique relevant du domaine spécialisé restreint des machines-outils pour l'usinage des métaux. Nous nous sommes tout particulièrement intéressés à la corrélation entre le rang de spécificité et le rang de monosémie de ces unités spécifiques.

Une double analyse quantitative a permis de générer une liste de 4717 mots-clés, avec un degré de spécificité et un degré de monosémie ou d'homogénéité sémantique. Ces données quantitatives ont permis de classer les 4717 mots-clés dans un continuum de spécificité et dans un continuum de monosémie afin d'examiner la corrélation entre le rang de spécificité et le rang de monosémie par le biais d'une analyse statistique de régression simple. Nous avons observé une corrélation négative, qui indique que les unités lexicales les plus spécifiques du corpus technique, relevées avec la méthodologie de l'analyse des mots-clés, ne sont pas les plus homogènes sémantiquement, au contraire. Cette observation a permis de remettre en cause la thèse monosémiste traditionnelle. La méthode alternative de l'analyse des marqueurs lexicaux stables (*Stable Lexical Marker Analysis* ou SLMA) a permis de remédier aux problèmes statistiques et méthodologiques d'hétéroscédasticité et de multicollinéarité, en prenant en considération la dispersion et en utilisant une autre mesure statistique. Elle a généré une liste de 3479 marqueurs lexicaux stables avec un nouveau rang de spécificité (rang de SMEA). Les résultats de l'analyse de régression simple confirment la corrélation négative entre le nouveau rang de spécificité (rang de SMEA) et le rang de monosémie des marqueurs lexicaux stables, bien qu'elle soit moins forte. Ces premières expérimentations montrent donc que l'analyse des marqueurs lexicaux stables constitue une alternative valable pour l'analyse des mots-clés.

Références

BERTELS, A., SPEELMAN, D. et GEERAERTS, D. (2010). La corrélation entre la spécificité et la sémantique dans un corpus spécialisé. *In Revue de Sémantique et de Pragmatique* n°27, pages 79–102.

BHREATNACH, U. et DE BARRA CUSACK, F., éditeurs (2010). *TKE 2010 : Presenting Terminology and Knowledge Engineering Resources Online: Models and Challenges*, Fiontar. Dublin City University.

BLUMENTHAL, P. et HAUSMANN, F.J., éditeurs (2006). *Collocations, corpus, dictionnaires. Langue française*, n° 150.

BOURIGAULT D., JACQUEMIN, C. et L'HOMME, M.-C., éditeurs (2001). *Recent advances in computational terminology*, Amsterdam/Philadelphia. John Benjamins Publishing Company.

BOURIGAULT, D. et SLODZIAN, M. (1999). Pour une terminologie textuelle. *In Terminologies Nouvelles* n°19, pages 29–32.

CABRE, M.T. (2000). Terminologie et linguistique : la théorie des portes. *In Terminologies Nouvelles* n°21, pages 10–15.

CONDAMINES, A. et REBEYROLLE, J. (1997). Point de vue en langue spécialisée. *In Meta*, n°42(1), pages 174–184.

CONDAMINES, A., éditeur (2005). *Sémantique et corpus*, Paris. Hermès-Science.

DE HERTOG, D., HEYLEN, K., SPEELMAN, D. et KOCKAERT, H. (2010). A variational linguistics approach to term extraction. *In* (Bhreatnach et de Barra Cusack, 2010), pages 229–248.

DUNNING, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* n°19(1), pages 61–74.

ERIKSEN, L. 2002. Die Polysemie in der Allgemeinsprache und in der juristischen Fachsprache. Oder : Zur Terminologie der ‚Sache' im Deutschen. *In Hermes – Journal of Linguistics* n°28, pages 211–222.

FERRARI, L. (2002). Un caso de polisemia en el discurso jurídico? *In Terminology* n°8(2), pages 221–244.

GAUDIN, F. (2003). *Socioterminologie* : *une approche sociolinguistique de la terminologie*. Bruxelles. Duculot.

GROSSMANN, F. et TUTIN, A., éditeurs (2003). *Les collocations, analyse et traitement, Travaux et Recherches en linguistique appliquée*, Série E, vol. 1.

HABERT, B., ILLOUZ, G., FOLCH, H. (2004). Dégrouper les sens : pourquoi ? comment ? *In Actes des JADT 2004 (Journées internationales d'analyse statistique des données textuelles),* Louvain-la-Neuve, pages 565–576.

HABERT, B., ILLOUZ, G., FOLCH, H. (2005). Des décalages de distribution aux divergences d'acception, *In* (Condamines, 2005), pages 277–318.

JUCKER, A., SCHREIER, D. et HUNDT, M. éditeurs (2009). *Corpora: Pragmatics and Discourse*, Amsterdam. Rodopi.

KRISTIANSEN, G. et DIRVEN, R., éditeurs (2008). *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems,* Berlin/New York. Mouton de Gruyter. KWARY, D.A. (2011). A hybrid method for determining technical vocabulary. *In System* n°39(2), pages 175-185.

KWARY, D.A. (2011). A hybrid method for determining technical vocabulary. *In System*, n°39(2), pages 175-185.LABBE, C. et LABBE, D. (2001). Que mesure la spécificité du vocabulaire ? *In Lexicometrica* n°3.

LAFON, P. (1984). *Dépouillements et statistiques en lexicométrie*, Genève-Paris. Slatkine-Champion.

LEMAY, C., L'HOMME, M.C. et DROUIN, P. (2005). Two methods for extracting specific single-word terms from specialized corpora. Experimentation and evaluation. *In International Journal of Corpus Linguistics*, n°10(2), pages 227–255.

MARTINEZ, W. (2000). Mise en évidence de rapports synonymiques par la méthode des cooccurrences. *In Actes des JADT 2000 (Journées internationales d'analyse statistique des données textuelles),* Lausanne, pages 78–84.

MAYAFFRE, D. (2008), Quand 'travail', 'famille', 'patrie' co-occurrent dans le discours de Nicolas Sarkozy. Etude de cas et réflexion théorique sur la co-occurrence, *In Actes des JADT 2008 (Journées internationales d'analyse statistique des données textuelles),* Lyon, pages 811–822.

PAQUOT, M. et BESTGEN, Y. (2009). Distinctive words in academic writing: a comparison of three statistical tests for keyword extraction », *In* (Jucker *et al.*, 2009), pages 247–269

SCOTT, M. et TRIBBLE, C. (2006). *Textual Patterns. Key words and corpus analysis in language education.* Studies in Corpus Linguistics, vol. 22. Amsterdam. Benjamins.

SPEELMAN, D., GRONDELAERS, S. et GEERAERTS, D. (2006). A profile-based calculation of region and register variation: the synchronic and diachronic status of the two main national varieties of Dutch. *In* (Wilson *et al.*, 2006), pages 195–202.

SPEELMAN, D., GRONDELAERS, S. et GEERAERTS, D. (2008). Variation in the choice of adjectives in the two main national varieties of Dutch. *In* (Kristiansen et Dirven, 2008), pages 205–233.

TEMMERMAN, R. (2000). *Towards new ways of terminology description. The sociocognitive approach,* Amsterdam/Philadelphia. John Benjamins Publishing Company.

VERONIS, J. (2003). Cartographie lexicale pour la recherche d'informations. *Actes de TALN 2003(Traitement automatique des langues naturelles)*, Batz-sur-Mer, pages 265–274.

WILSON, A., ARCHER, D. et RAYSON, P., éditeurs (2006). *Corpus Linguistics around the World*, Amsterdam. Rodopi.

WÜSTER, E. (1991). *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*, (3. Aufl.), Bonn. Romanistischer Verlag.

Fouille de graphes sous contraintes linguistiques pour l'exploration de grands textes

Solen Quiniou[1, 2] Peggy Cellier[3] Thierry Charnois[1] Dominique Legallois[2]
(1) GREYC Université de Caen Basse-Normandie, Campus 2, 14000 Caen
(2) CRISCO Université de Caen Basse-Normandie, Campus 1, 14000 Caen
(3) IRISA-INSA de Rennes, Campus de Beaulieu, 35042 Rennes Cedex
solen.quiniou@unicaen.fr, peggy.cellier@irisa.fr,
thierry.charnois@unicaen.fr, dominique.legallois@unicaen.fr

RÉSUMÉ

Dans cet article, nous proposons une approche pour explorer des textes de taille importante en mettant en évidence des sous-parties cohérentes. Cette méthode d'exploration s'appuie sur une représentation en graphe du texte, en utilisant le modèle linguistique de Hoey pour sélectionner et apparier les phrases dans le graphe. Notre contribution porte sur l'utilisation de techniques de fouille de graphes sous contraintes pour extraire des sous-parties pertinentes du texte (c'est-à-dire des collections de sous-réseaux phrastiques homogènes). Nous avons réalisé des expérimentations sur deux textes anglais de taille conséquente pour montrer l'intérêt de l'approche que nous proposons.

ABSTRACT

Graph Mining Under Linguistic Constraints to Explore Large Texts

In this paper, we propose an approach to explore large texts by highlighting coherent sub-parts. The exploration method relies on a graph representation of the text according to the Hoey linguistic model which allows the selection and the binding of sentences in the graph. Our contribution relates to using graph mining techniques under constraints to extract relevant sub-parts of the text (*i.e.*, collections of homogeneous sentence sub-networks). We have conducted some experiments on two large English texts to show the interest of the proposed approach.

MOTS-CLÉS : Fouille de graphes, réseaux phrastiques, analyse textuelle, navigation textuelle.

KEYWORDS: Graph Mining, sentence networks, textual analysis, textual navigation.

1 Introduction

L'interprétation critique des textes et l'analyse textuelle et discursive ont été renouvelées ces dernières années grâce à la numérisation de nombreux textes. Cependant, ce renouvellement s'accompagne de difficultés, notamment techniques : la numérisation ne suffit pas en elle-même, l'investigation des textes doit s'appuyer sur des méthodes et outils offrant à la fois une visualisation et une navigation pertinentes dans les textes. Les chercheurs peuvent ainsi par exemple focaliser leur analyse sur des thématiques particulières. La nécessité de tels méthodes et outils est d'autant plus forte que les textes sont généralement de taille conséquente. Deux types d'approches peuvent aider les linguistes dans des tâches d'exploration ou d'analyse de textes : les

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 253–266,
Grenoble, 4 au 8 juin 2012.

méthodes de résumé automatique et les techniques de visualisation de collections de textes.

D'un côté, les méthodes de résumé automatique visent à produire une vue contiguë du texte sous la forme d'un texte réduit formé de phrases saillantes. Il existe deux principales catégories d'approches pour le résumé automatique. Le premier type d'approche s'appuie sur l'extraction de phrases du texte original (Lin et Hovy, 2002). Un sous-ensemble de phrases saillantes du texte original est ainsi sélectionné. Dans le second type d'approche, appelé compression de phrases(Knight et Marcu, 2000), l'objectif est de réduire les phrases tout en préservant leur sens. Cependant, les méthodes de résumé automatique ne permettent pas de produire une vue relationnelle de la structure ou de l'organisation des différentes parties du texte.

D'un autre côté, les techniques de visualisation de collections de textes ont connu un intérêt grandissant ces dernières années (Newman *et al.*, 2010; Don *et al.*, 2007; Fekete et Dufournaud, 2000; Plaisant *et al.*, 2006). Par exemple, Newman *et al.* (2010) utilisent un modèle probabiliste pour produire un ensemble de thématiques décrivant une collection afin que l'utilisateur puisse effectuer une recherche de documents liés à une thématique particulière. Don *et al.* (2007) ont proposé un outil pour visualiser une collection de textes et permettre à l'utilisateur de l'explorer en affichant des motifs textuels fréquents (par exemple, des mots fréquents ou des ensembles de trigrammes). Les occurrences des motifs sont alors mises en relief dans le texte. Cependant, les approches de visualisation présentent un inconvénient commun aux approches de résumé automatique : le texte est visualisé de manière globale, sans mettre en évidence les relations entre les phrases.

Parmi les travaux intéressants en linguistique concernant l'exploration de textes, Hoey a présenté un modèle linguistique pour analyser des textes non-narratifs en s'appuyant sur les répétitions lexicales (Hoey, 1991). L'approche met en évidence l'organisation du texte (par exemple, le développement du texte ou son contenu conceptuel) en révélant les appariements entre phrases contiguës ou non contiguës, ce qui permet de construire une représentation du texte sous forme de *réseaux phrastiques*. Cette approche est intéressante pour plusieurs tâches tel que le raisonnement logique sur un sujet particulier du texte, l'étude de la cohésion lexicale du texte (Legallois *et al.*, 2011), le résumé de texte (Renouf et Kehoe, 2004) ou encore la segmentation de texte (Sardinha, 1999). Plusieurs études ont montré l'intérêt de la méthode sur des textes en anglais (Hoey, 1997; Károly et Francis, 2000) mais aussi sur des textes en français (Legallois, 2006). Alors qu'il est difficile de l'appliquer à la main sur des textes conséquents, peu de travaux utilisent une implémentation du modèle de Hoey. Dans Renouf et Kehoe (2004), les auteurs ont développé un outil de résumé basé sur le modèle de Hoey : SEAGULL (*Summary Extraction Algorithm Generated Using Lexical Links*). Ils ont réalisé des expérimentations pour montrer que leur outil obtient de meilleures performances que d'autres outils de résumé automatique mais ils n'ont utilisé pour cela qu'un petit texte en anglais (730 mots). Dans Legallois *et al.* (2011) les auteurs ont proposé un processus automatique appliquant le modèle de Hoey sur des textes de grande taille, afin d'étudier la cohésion lexicale de ces textes. Les expérimentations menées sur différents types de textes en français (narratif, expositif) ont permis de montrer l'intérêt de ce modèle pour cette tâche. Cependant, la grande taille des réseaux phrastiques obtenus par l'application de ce modèle demeure un inconvénient. En effet, cette représentation ne permet pas de visualiser de longs textes en entier et l'extraction de sous-parties potentiellement intéressantes n'est pas prévue par le modèle.

Dans cet article, nous proposons une approche pour extraire automatiquement, à partir d'un texte, des sous-ensembles de phrases cohérents d'un point de vue lexical. De plus, les sous-

ensembles sont représentés par des graphes, ce qui offre une vue des relations entre les phrases de ces sous-ensembles. Enfn, la taille des sous-ensembles de phrases étant raisonnable, cela permet aux linguistes de les analyser. Notre approche s'appuie sur une représentation du texte sous forme de graphe par application du modèle linguistique de Hoey. Pour pouvoir analyser de grands textes, nous proposons une implémentation du modèle de Hoey permettant de traiter des textes de grande taille. La principale contribution est l'utilisation d'une technique particulière de fouille de graphes, appelée *fouille de CoHoP*, pour extraire des sous-parties cohérentes du texte représenté sous forme de graphes. À notre connaissance, cette technique de fouille de graphes n'a jamais été utilisée dans le domaine du traitement automatique des langues. Dans notre approche, le processus de fouille est dit « sous contraintes linguistiques » car le graphe initial représentant le texte est construit par application du modèle de Hoey. D'autres contraintes linguistiques sur les sommets du graphe guident également le processus de fouille.

La fouille de graphes a connu un intérêt grandissant dans le domaine de la fouille de données pour la découverte de nouvelles connaissances (Washio et Motoda, 2003), et plus particulièrement la fouille de graphes *enrichis* (des attributs sont alors associés aux sommets). De telles méthodes de fouille ont été utilisées avec succès pour des tâches comme le clustering (Ge *et al.*, 2008; Zhou *et al.*, 2010) ou l'extraction de sous-graphes approximatifs (Tong *et al.*, 2007). Dans cet article, nous nous focalisons sur la fouille d'un certain type de motifs à partir de graphes enrichis : des *collections de k-PC homogènes* (CoHoP) (Mougel *et al.*, 2012). Nous les utilisons pour extraire des sous-parties homogènes du texte.

Dans la suite de l'article, nous présentons le modèle linguistique de Hoey, dans la section 2, puis les techniques de fouille de graphes pour extraire des motifs de type CoHoP, dans la section 3. Nous décrivons ensuite notre approche permettant d'extraire des sous-parties cohérentes des réseaux phrastiques en s'appuyant sur des méthodes de fouille de graphes, dans la section 4. Nous discutons enfin des expérimentations menées sur deux textes anglais, dans la section 5.

2 Modèle linguistique de Hoey

Le modèle linguistique introduit dans Hoey (1991) repose sur la notion de répétition lexicale au sein d'un texte. Il consiste alors à identifier les phrases du texte qui partagent au moins trois unités lexicales. Une *répétition lexicale* peut correspondre à la stricte répétition de l'unité lexicale (*cerveau/cerveau*) mais aussi à la répétition d'unités lexicales partageant le même lemme ou une autre forme dérivée (*produire/production*). La répétition lexicale peut également correspondre à une reprise anaphorique, une relation de synonymie (*acheter/acquérir*), une relation d'hypo/hyperonymie (*chien/animal*), une relation « implicative » (*conduire/voiture*) ou encore une suite ordonnée (*lundi/mardi/mercredi/...*).

Lorsque deux phrases partagent au moins trois unités lexicales, la paire de phrases est appariée. Un *appariement* entre deux phrases correspond ainsi à un chemin entre ces phrases. On appelle alors *réseau phrastique* un ensemble d'au moins trois phrases tel que, quelles que soient deux phrases de cet ensemble, ces phrases sont soit directement appariées, soit indirectement reliées par une succession de chemins dans l'ensemble de phrases [1]. L'ensemble des réseaux phrastiques d'un texte est appelé *hypotexte*. L'hypotexte peut être vu en quelque sorte comme un résumé du

1. Notons qu'en théorie des graphes, un réseau phrastique correspond à une composante connexe du graphe constitué des appariements.

texte original. Il est à noter que les phrases non appariées n'apparaissent pas dans l'hypotexte.

La figure 1 montre un extrait d'un réseau phrastique de "*Love Online: Emotions on the Internet*" (Ben-Ze'ev, 2004). Dans cet exemple, la répétition lexicale repose uniquement sur les lemmes des lexèmes communs des phrases (les *lexèmes* correspondent aux noms, adjectifs, adverbes et verbes). Ainsi, seules les répétitions strictes sont considérées. Le réseau phrastique est interprétable avec quelques « aménagements » mineurs. Dans l'exemple de la figure 1, certains mots ont été ajoutés au texte original et mis entre crochets afin de faciliter la lecture des enchaînements phrastiques du réseau (par exemple, [*However*]). La numérotation au début de chaque phrase (entre crochets) correspond à la position de la phrase dans le texte original ; il est alors intéressant de constater que l'empan de cet extrait de réseau est relativement conséquent.

Le modèle de Hoey permet de représenter un texte afin d'analyser sa cohésion lexicale. Cependant, la représentation sous forme d'hypotexte construite par ce modèle est de taille trop importante pour être visualisée en entier, ce qui rend fastidieuse l'exploration et l'analyse du texte. Ainsi, il est intéressant de disposer d'une méthode permettant d'extraire des sous-parties homogènes de réseaux phrastiques afin de faciliter l'analyse de ces réseaux. Dans ce but, nous introduisons, dans la section suivante, une approche de fouille de graphes permettant d'extraire des motifs appelés CoHoP.

3 Fouille de graphes : motifs de type CoHoP

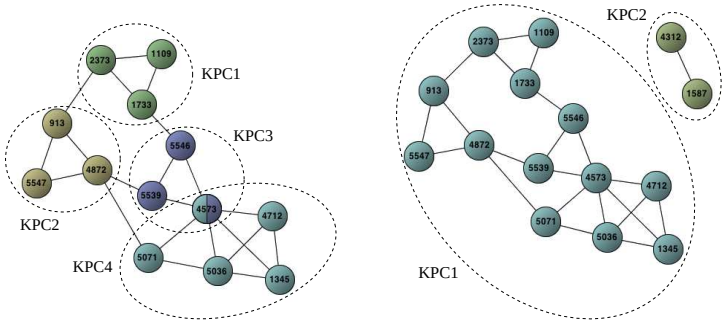
Les CoHoP (*collection de k-cliques percolées*) sont des motifs extraits à partir de graphes attribués booléens (Mougel *et al.*, 2012). Elle peuvent être vues comme des ensembles de communautés dont les éléments partagent des propriétés communes : chaque communauté correspond à une *k-clique percolée* (*k*-PC). Les CoHoP rendent ainsi compte d'une structure cachée, sous-jacente au graphe initial. Dans cette section, nous présentons les deux principales notions sur lesquelles s'appuie cette technique particulière de fouille de graphe : les *k*-PC et les CoHoP.

3.1 *k*-cliques percolées (*k*-PC)

Dans un graphe, une *k-clique* est un ensemble de k sommets dans lequel chaque paire de sommets distincts est connectée par une arête. La notion de *k-clique percolée* (*k*-PC) peut être vue comme une version relâchée du concept de clique. Une *k*-PC a été définie par Derenyi *et al.* (2005) comme étant l'union de toutes les *k*-cliques connectées par des chevauchements de $k-1$ sommets. Ainsi, dans une *k*-PC, chaque *k*-clique peut être atteinte de n'importe quelle autre *k*-clique par un chemin de *k*-cliques adjacentes et chaque sommet d'une *k*-PC peut être atteint

[1109] **Online**$_1$ **emotional**$_2$ **experiences**$_3$ may be compared to receiving a salary without earning it by hard work.
[1733] [However] **Online**$_1$ **relationships**$_4$ have a profound impact upon our **emotional**$_2$ **experiences**$_3$.
[2373] Since **emotional**$_2$ self-disclosure is more important to the **experience**$_3$ of intimacy than factual self-disclosure, 13 **online**$_1$ **relationships**$_4$ often have a higher degree of intimacy than offline **relationships**$_4$.

FIGURE 1: Extrait d'un réseau phrastique de "*Love Online: Emotions on the Internet*"



(a) $k = 3$ (b) $k = 2$

FIGURE 2: Exemple de CoHoP extraite à partir des attributs $\{a_1, a_2\}$, pour deux valeurs de k

par n'importe quel autre sommet de cette k-PC par un chemin de sous-ensembles de sommets bien connectés (les k-cliques).

Dans la figure 2a, il y a quatre k-PC ($k = 3$) : {1109, 1733, 2373}, {913, 4872, 5547}, {4573, 5539, 5546} et {1345, 4573, 4712, 5036, 5077}. Les trois premières k-PC contiennent une seule 3-clique alors que la dernière k-PC contient cinq 3-cliques (*e.g.*, {1345, 4712, 5036} et {1345, 4712, 4573}). Revenons sur la création de cette dernière k-PC. Nous pouvons tout d'abord constater que les sommets 1345, 4573, 4712 et 5036 sont directement connectés les uns aux autres : ils appartiennent ainsi à la même k-PC. Le sommet 5071 appartient également à cette k-PC puisqu'il est accessible à partir de chacun des quatre sommets précédents, par une série de k-cliques se chevauchant (le paramètre k a un impact sur le nombre de sommets à considérer dans les k-cliques ; dans cet exemple, sa valeur est fixée à 3) : par exemple, pour aller du sommet 5071 au sommet 4712, un chemin de 3-cliques se chevauchant peut être {4712, 4573, 5036} suivi de {4573, 5036, 5071} (avec $k = 3$, les chevauchements de 3-cliques contiennent deux sommets). En revanche, le sommet 4872 n'appartient pas à cette k-PC. En effet, pour cela il faudrait qu'il y ait une 3-clique entre les sommets 4573, 5071 et 4872, ce qui n'est pas le cas.

Il est à noter que le calcul des k-PC est indépendant des ensembles d'attributs associés aux sommets (les graphes utilisés étant attribués). De plus, une k-clique ne peut appartenir qu'à au plus une k-PC alors qu'un sommet peut se trouver dans plusieurs k-PC, puisqu'il peut appartenir à plusieurs k-cliques. Un sommet appartenant à plusieurs k-PC est appelé nœud relais ou *bridging node* (Musial et Juszczyszyn, 2009). Ainsi, dans la figure 2a, le sommet 4573 est un nœud relais car il appartient à deux k-PC : KPC_3 et KPC_4.

3.2 Collections de k-PC homogènes (CoHoP)

Une *collection de k-PC homogènes* (CoHoP) a été définie par (Mougel *et al.*, 2012) comme étant un ensemble de sommets tels que, étant donnés k, α et γ des entiers positifs définis par des utilisateurs :

– tous les sommets sont *homogènes*, c'est-à-dire qu'ils partagent au moins α attributs ;

– la collection contient au moins γ k-PC ;
– toutes les k-PC ayant les mêmes attributs sont présentes dans la collection (contrainte de *maximalité*).

La figure 2a représente ainsi une CoHoP extraite à partir de l'ensemble d'attributs $\{a_1, a_2\}$; comme vu dans la section 3.1, elle contient quatre k-PC ($\alpha = 2, k = 3, \gamma = 4$). Il est à noter que, contrairement au calcul des k-PC, l'extraction des CoHoP dépend fortement de l'ensemble d'attributs associés aux sommets du graphe. Sur la figure 2a, les ensembles d'attributs des sommets ne sont pas illustrés (pour ne pas surcharger la figure) mais chaque sommet est en fait étiqueté par un ensemble d'attributs qui contient au moins les attributs a_1 et a_2. En effet, cette CoHoP a été extraite à partir de ces deux attributs.

Les trois paramètres - k, α et γ - ont un impact inportant sur la structure des CoHoP extraites. Comme précisé précédemment, le paramètre α fixe le nombre minimal d'attributs communs associés aux sommets des CoHoP extraites et le paramètre γ fixe le nombre minimal de k-PC présentes dans les CoHoP. Le paramètre k a également un impact important sur la structure des CoHoP extraites. En effet, augmenter sa valeur a pour conséquence d'augmenter le degré de cohésion entre les sommets appartenant à une même k-PC. La figure 2b représente la CoHoP extraite à partir du même ensemble d'attributs que celle illustrée par la figure 2a mais en fixant cette fois $k = 2$. Cette CoHoP comporte maintenant 15 sommets répartis en seulement deux k-PC, la plus grosse k-PC (KPC_1) correspondant en fait aux quatre k-PC de la figure 2a. Ainsi, le choix de la valeur de k permet de choisir le degré de cohésion souhaité entre les sommets de chaque k-PC. En effet, un plus grand nombre de sommets doit être directement relié les uns aux autres lorsque la valeur de k augmente (la valeur de k représente ce nombre minimal de sommets).

4 Fouille de réseaux phrastiques

Dans cette section, nous proposons une nouvelle approche qui s'appuie à la fois sur le modèle de Hoey et sur la fouille de motifs de type CoHoP. Notre approche permet d'extraire des sous-parties homogènes de réseaux phrastiques afin de faciliter leur analyse.

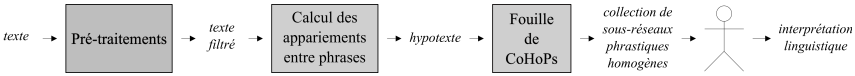


FIGURE 3: Vue d'ensemble de l'extraction de collections de sous-réseaux phrastiques homogènes

La figure 3 illustre les différentes étapes de notre approche, allant du pré-traitement du texte à l'extraction des collections de sous-réseaux phrastiques homogènes, en passant par la construction de l'hypotexte représentant le texte étudié. Ces différentes étapes sont décrites plus en détail dans les sous-sections suivantes.

4.1 Construction de l'hypotexte

Le texte est tout d'abord étiqueté à l'aide de TreeTagger (Schmid, 1994). Le texte étiqueté est ensuite découpé en phrases à chaque signe de ponctuation de l'ensemble suivant : { « . », « ? »,

« ! », « : »}. Les phrases sont finalement filtrées afin de ne conserver que leurs unités lexicales pertinentes. Dans notre cas, cela consiste à ne garder que les *lexèmes* (noms, adjectifs, adverbes et verbes hormis les auxiliaires). En fait, nous considérons les lemmes de ces lexèmes. À l'issue de cette étape, chaque phrase du texte filtré est alors représentée par les lemmes de ses lexèmes. Par exemple, la phrase « *Online emotional experiences may be compared to receiving a salary without earning it by hard work.* » est représentée par « *online emotional experience compare receive salary earn hard work* ». Tous les mots non pertinents sont ainsi filtrés et les mots restants (*e.g.*, les lexèmes) sont remplacés par leur lemme.

À partir du texte filtré, nous construisons ensuite la représentation du texte sous forme de graphe en appliquant le modèle linguistique de Hoey, comme présenté à la section 2. L'hypotexte est ainsi créé en appariant toutes les paires de phrases qui partagent au moins trois unités lexicales communes, correspondant dans notre cas aux lemmes des lexèmes. Il est également à noter que les phrases non appariées ne sont pas conservées dans l'hypotexte.

4.2 Fouille de collections de sous-réseaux phrastiques homogènes

L'objectif de cette dernière étape est d'extraire des sous-parties homogènes de l'hypotexte. L'hypotexte créé comme décrit précédemment peut être vu comme un graphe attribué. En effet, un hypotexte est un graphe dont chaque sommet représente une phrase appariée et dont chaque arête représente un appariement entre deux phrases qui partagent au moins trois unités lexicales communes. De plus, l'ensemble des unités lexicales d'une phrase peut étiqueter le sommet correspondant, en tant qu'ensemble d'attributs.

Avec cette représentation de l'hypotexte comme un graphe attribué, nous pouvons utiliser des algorithmes de fouille de CoHoP, comme présenté à la section 3. Dans notre approche, le processus de fouille est effectué « sous contraintes linguistiques » car, d'une part, la structure initiale du graphe est construite en utilisant le modèle linguistique de Hoey et, d'autre part, les ensembles d'attributs qui étiquettent les sommets sont les unités lexicales des phrases correspondantes.

Dans notre approche, chaque CoHoP extraite correspond alors à ce que nous appelons une *collection de sous-réseaux phrastiques homogènes* (CoHoSS). En effet, de la même façon qu'une CoHoP est constituée de k-PC homogènes (*i.e.*, des ensembles de sommets partageant un sous-ensemble de α attributs communs), une CoHoSS est constituée de sous-réseaux phrastiques homogènes (*i.e.*, des ensembles de phrases partageant un sous-ensemble de α unités lexicales communes). Chaque sous-réseau phrastique correspond alors à la définition d'une k-PC. Ainsi, dans un sous-réseau phrastique, chaque phrase est soit directement appariée aux autres phrases du sous-réseau par une arête (si elle partage au moins trois unités lexicales avec chacune de ces phrases), soit indirectement accessible depuis n'importe quelle autre phrase par une série de sous-ensembles de phrases bien connectés (chaque sous-ensemble correspond à une k-clique, comme défini à la section 3.1). Ainsi, la CoHoP représentée par la figure 2a correspond à une collection de sous-réseaux phrastiques homogènes (*i.e.*, une CoHoSS) dont toutes les phrases partagent les mots *emotional* et *experience* (correspondant aux attributs a_1 et a_2, respectivement). En effet, chaque sommet de la CoHoP correspond à une phrase, le numéro du sommet indiquant la position de la phrase dans le texte (la figure 1 donne les phrases associées aux sommets 1109, 1733 et 2373 de la CoHoP). De plus, les ensembles d'attributs associés aux sommets correspondent aux ensembles d'unités lexicales représentant les phrases. Par exemple, l'ensemble d'attributs A_{4712} de la phrase 4712 correspond à l'ensemble d'unités lexicales suivant :

{*parallel, world, help, preserve, actual, not, give, exciting, emotional, experience*}.

Les CoHoSS représentent ainsi des collections de sous-réseaux phrastiques de l'hypotexte initial qui ont une certaine cohésion lexicale par rapport à l'ensemble des unités lexicales à partir desquelles elles ont été extraites. Les CoHoSS ainsi que leur structure peuvent ensuite être analysées par des linguistes pour interpréter, par exemple, chacun des sous-réseaux ainsi que la façon dont ils sont connectés les uns aux autres, notamment par les *phrases relais* (correspondant aux nœuds relais présentés dans la section 3.1).

5 Résultats expérimentaux

Dans cette section, nous présentons les résultats expérimentaux sur deux textes anglais. Les textes ainsi que les outils utilisés pour extraire et visualiser les CoHoP sont tout d'abord présentés dans la section 5.1. Nous discutons ensuite les résultats quantitatifs sur l'utilisation du modèle de Hoey et sur les CoHoSS extraites, dans les sections 5.2.1 et 5.2.2 respectivement. Enfin, dans la section 5.2.3, nous présentons un exemple de CoHoSS extraite et son interprétation linguistique afin de montrer l'intérêt de notre approche pour l'exploration de textes.

5.1 Paramètres : données et outils

5.1.1 Textes étudiés

Pour évaluer l'approche que nous proposons, nous avons choisi deux textes de grande taille, chacun correspondant à un texte expositif en anglais : "*The Origin of Speech*" (MacNeilage, 2008) et "*Love Online: Emotions on the Internet*" (Ben-Ze'ev, 2004). Les caractéristiques de ces textes sont données dans la table 1.

Texte	Titre	Auteur	Année	Nb. pages
Speech	*The Origin of Speech*	Peter F. MacNeilage	2008	416
Love	*Love Online : Emotions on the Internet*	Aaron Ben-Ze'ev	2004	302

TABLE 1: Caractéristiques des textes étudiés

5.1.2 Extraction de motifs de type CoHoP

Afin d'extraire les CoHoP, comme présenté dans la section 4.2, nous utilisons *CoHoP Miner* (Mougel *et al.*, 2012). Cet outil permet d'extraire des CoHoP en fixant les valeurs des différents paramètres utilisés lors du processus de fouille (k, α et γ). Il inclut également une interface graphique pour visualiser les CoHoP extraites, comme l'illustre la figure 4. Les k-PC de la CoHoP affichée peuvent ainsi être visualisées, chaque k-PC étant représentée par une couleur différente. Les sommets appartenant à plusieurs k-PC peuvent également être visualisés afin de se focaliser sur eux lors de l'analyse des CoHoP, par exemple. Il est également possible d'afficher

l'ensemble d'attributs de chaque sommet. En fait, cet outil offre une visualisation globale des motifs extraits puisqu'il permet de sélectionner les motifs à analyser par la suite selon l'ensemble des attributs à partir desquels ils ont été extraits ou encore selon leur structure (par exemple, le nombre de *k*-PC qu'ils contiennent ou la présence de sommets appartenant à plusieurs *k*-PC). Cependant, afin d'analyser les CoHoP, les linguistes ont besoin d'afficher les phrases correspondant aux sommets ainsi que les mots à partir desquels les phrases ont été appariées (et conduisant à des arêtes entre les sommets). C'est pourquoi nous utilisons également un outil de visualisation de graphes.

5.1.3 Visualisation des réseaux phrastiques

Cet outil de visualisation de graphes a été développé en Java. Il offre une visualisation des sous-réseaux phrastiques correspondant aux CoHoP extraites, comme illustré par la figure 5. La CoHoSS affichée correspond à la CoHoP de la figure 4 : les sommets correspondent aux phrases du texte et les arêtes sont étiquetées par les unités lexicales partagées par chaque paire de phrases. Cet outil permet aux linguistes d'analyser plus facilement les réseaux phrastiques en offrant une visualisation focalisée sur une collection particulière de sous-réseaux phrastiques homogènes. Cependant, il n'offre pas d'affichage des différentes *k*-PC de la CoHoP extraite, ce qui rend plus difficile l'analyse des phrases appartenant à plusieurs sous-réseaux (*i.e.*, les phrases relais), par exemple. Il est donc intéressant d'utiliser cet outil en complément de *CoHoP Miner*.

5.2 Expérimentations sur la fouille de réseaux phrastiques

5.2.1 Résultats quantitatifs sur les réseaux phrastiques

Nous présentons tout d'abord des résultats quantitatifs sur l'hypotexte créé pour représenter chaque texte. Ces résultats sont donnés dans la table 2. Nous pouvons tout d'abord remarquer que chaque phrase contient en moyenne 10 lexèmes pour *Speech* et 11 lexèmes pour *Love* alors que les phrases comportent respectivement 24 et 20 mots, en moyenne. Représenter les phrases par leurs lexèmes permet ainsi de réduire le nombre d'attributs qui leur est associé. Nous pouvons également constater que peu de réseaux phrastiques ont été créés : deux pour chaque texte.

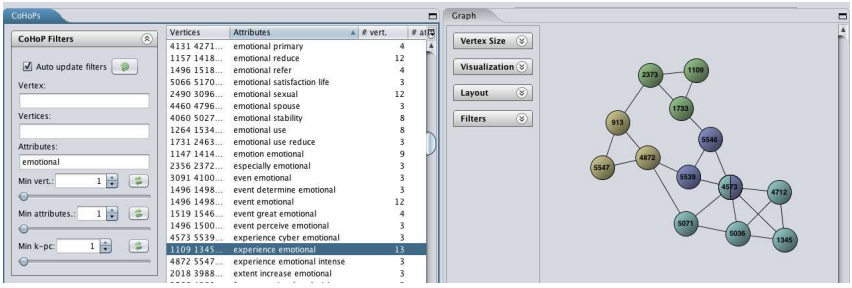


FIGURE 4: Visualisation d'une CoHoP du texte *Love*, avec *CoHoP Miner*

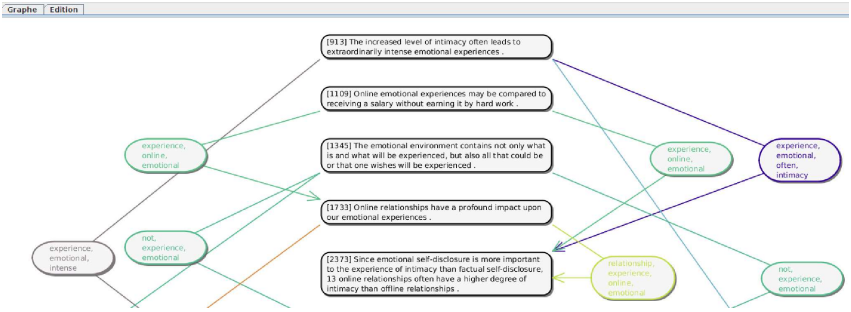


FIGURE 5: Extrait de la visualisation d'une CoHoSS du texte *Love*

De plus, l'hypotexte de chacun des textes (constitué ainsi des deux réseaux phrastiques) est de taille importante puisqu'il contient plus de 75 % des phrases. Cela suggère une forte cohésion lexicale à l'intérieur des textes (chaque phrase de l'hypotexte est en moyenne appariée avec 13 phrases pour *Speech* et 30 phrases pour *Love*).

5.2.2 Résultats quantitatifs sur les CoHoSS extraites

La table 3 donne le nombre de CoHoSS extraites pour chaque texte, pour différentes valeurs de k, α et γ. Par exemple, avec $(k = 3, \alpha = 2, \gamma = 2)$, 523 CoHoSS sont extraites pour le texte *Love* (voir la ligne 2 de la table 3a).

Une première série d'expériences a été menée en fixant $\gamma = 2$: cela signifie que l'on fixe le nombre de sous-réseaux phrastiques contenus dans une CoHoSS à au minimum deux. La table 3a donne le nombre de CoHoSS extraites avec ce paramétrage de γ et en faisant varier α et k. Nous observons que quelle que soit la valeur de k, le fait d'augmenter la valeur de α (c'est-à-dire le nombre d'attributs communs aux phrases) implique logiquement une diminution significative du nombre de CoHoSS (environ 50 % en faisant varier α de 1 à 2). Plus particulièrement, nous remarquons qu'aucune CoHoSS n'est extraite pour une valeur $\alpha \geq 3$. Cela signifie que les CoHoSS sont extraites à partir d'un ou deux attributs (ici, les lemmes des lexèmes des phrases).

Une autre série d'expériences a été menée en fixant cette fois $\alpha = 2$: cela signifie que l'on fixe le nombre d'attributs à au minimum deux. La table 3b donne alors le nombre de CoHoSS extraites avec ce paramétragede α et en faisant varier γ et k. Nous constatons qu'augmenter la valeur de γ

Texte	Nb. mots	Nb. total lexèmes	Nb. lexèmes différents	Nb. phrases	Nb. appariements	Nb. réseaux phrastiques	% phrases dans hypotexte
Speech	127 563	59 657	4 728	5 308	50 277	2	75,6%
Love	112 325	53 035	6 919	5 571	131 497	2	79,0%

TABLE 2: Résultats quantitatifs sur les hypotextes

Texte	k	α		
		1	2	3
Love	2	1 010	555	0
	3	924	523	0
	4	729	403	0
Speech	2	1 420	793	0
	3	973	523	0
	4	678	384	0

(a) $\gamma = 2$

Texte	k	γ				
		1	2	3	4	5
Love	2	38 061	555	15	0	0
	3	16 437	523	41	9	1
	4	8 425	403	51	13	4
Speech	2	39 442	793	25	2	0
	3	13 673	523	84	17	2
	4	5 552	384	75	24	2

(b) $\alpha = 2$

TABLE 3: Nombre de CoHoSS extraites en fixant la valeur (a) de γ et (b) de α

(c'est-à-dire le nombre de sous-réseaux phrastiques contenus dans une CoHoSS) diminue moins rapidement le nombre de CoHoSS extraites pour des valeurs élevées de k. En effet, comme présenté dans la section 3, augmenter la valeur de k permet d'obtenir des sous-réseaux avec un degré de cohésion plus élevé. Cela s'accompagne d'un nombre plus élevé de sous-réseaux dans les CoHoSS et donc d'une valeur plus élevée pour γ. Ainsi, en fixant les valeurs de α et γ, le nombre de sous-réseaux dans les CoHoSS extraites sera généralement plus élevé lorsque la valeur de k est également élevée.

D'une manière générale, nous pouvons observer que le nombre de CoHoSS extraites diminue lorsque les valeurs de k, α et γ augmentent. En conclusion, la valeur de k doit être choisie de manière judicieuse en fonction du degré de cohésion lexicale souhaitée pour les sous-réseaux phrastiques contenus dans les CoHoSS extraites. La valeur de γ, quant à elle, permet de limiter le nombre total de CoHoSS extraites, en ne choisissant que les plus grosses en termes de nombre de sous-réseaux phrastiques qu'elles contiennent. Enfin, la valeur de α joue un rôle important lorsque l'on souhaite focaliser l'analyse linguistique sur les relations entre les phrases partageant un certain nombre d'unités lexicales.

5.2.3 Exemple de CoHoSS extraite et interprétation linguistique

Nous présentons maintenant un exemple de CoHoSS. La figure 6 illustre la CoHoP extraite à l'aide de *CoHoP Miner*, sur le texte *Speech*, et la figure 7 donne les phrases correspondantes.

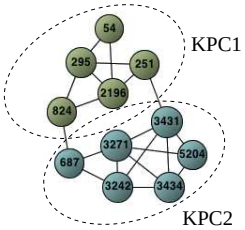


FIGURE 6: CoHoP correspondant à la CoHoSS extraite à partir de {*adaptation*} ($k = 3$)

La CoHoSS a été extraite à partir de l'attribut *adaptation*, en utilisant les valeurs suivantes pour

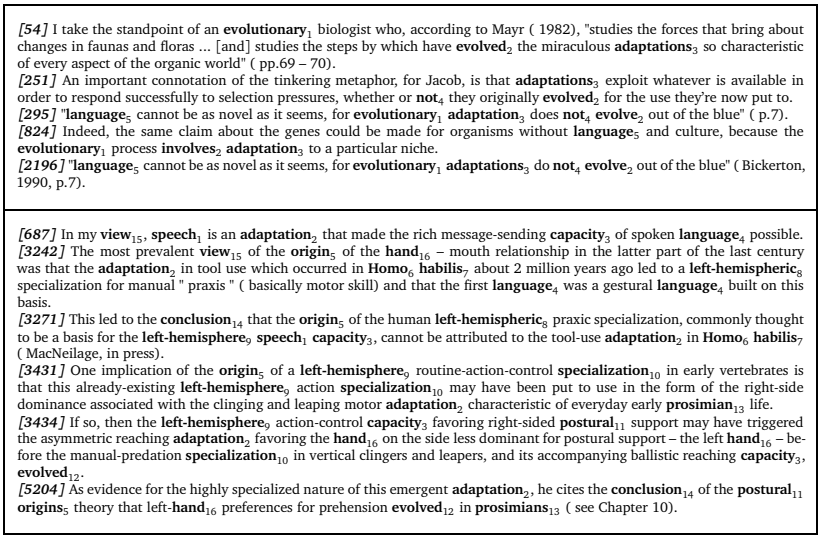


FIGURE 7: Phrases correspondant à la CoHoP de la figure 6

les paramètres de fouille : $k = 3, \alpha = 1, \gamma = 2$. Cette CoHoSS est constituée de deux sous-réseaux phrastiques. Le premier réseau (contenant les phrases 54, 251, 295, 824 et 2196) traite du thème général de la CoHoSS : le phénomène d'adaptation. Nous pouvons remarquer que ce réseau est relativement cohérent alors qu'il parcourt un ensemble considérable du texte (correspondant à un empan de plus de 2 000 phrases). Le second réseau, quant à lui, développe une thématique plus spécifique de l'adaptation : la spécialisation de l'hémisphère gauche. Ce sous-réseau commence avec la phrase 687 qui est connectée au sous-réseau précédent par la phrase 824. Nous pouvons également constater que les deux sous-réseaux se « chevauchent » au niveau du texte puisque la phrase 687 appartient au second sous-réseau alors que les phrases 824 et 2196 appartiennent au premier. De plus, nous pouvons remarquer que l'empan de la CoHoSS (mais aussi de chacun des sous-réseaux) est relativement important puisqu'elle commence à la phrase 54 et qu'elle se termine à la phrase 5204. Ainsi, cette propriété intéressante de non-contiguïté des phrases des réseaux phrastiques se retrouve également au niveau des sous-réseaux phrastiques constituant les CoHoSS extraites.

6 Conclusion

Dans cet article, nous avons proposé une approche pour explorer des textes de taille conséquente en se focalisant sur des sous-parties cohérentes. Cette méthode d'exploration s'appuie sur une représentation du texte à l'aide d'un graphe, en utilisant le modèle linguistique de Hoey

pour sélectionner et apparier les phrases conservées dans le graphe. Notre contribution porte sur l'utilisation de techniques issues de la fouille de graphes pour extraire des sous-parties du texte cohérentes d'un point de vue lexical (c'est-à-dire des collections de sous-réseaux phrastiques homogènes) dont la taille permet à un linguistique de les analyser. Nous avons réalisé des expérimentations sur deux textes anglais de la taille d'un livre pour valider cette approche. Cela nous a permis de montrer que le graphe généré à l'aide du modèle de Hoey était difficilement exploitable par un humain à cause du trop grand nombre de sommets et d'arêtes. En utilisant notre approche pour sélectionner des sous-parties pertinentes du graphe, il est alors possible d'appliquer le modèle de Hoey sur de grands textes. De plus, les différents paramètres utilisés lors du processus de fouille du graphe offrent la possibilité de définir le niveau de granularité des collections de sous-réseaux phrastiques homogènes extraites. D'un point de vue linguistique, cela signifie que le degré de cohésion lexicale entre les phrases des sous-réseaux phrastiques extraits est mis en évidence.

Remerciements

Les auteurs tiennent à remercier chaleureusement Pierre-Nicolas Mougel et Christophe Rigotti (LIRIS, Lyon) pour la mise à disposition de *CoHoP Miner*.

Ce travail bénéficie du soutien de la région Basse-Normandie et de l'ANR (projet Hybride ANR-11-BS02-002).

Références

Ben-Ze'ev, A. (2004). *Love Online : Emotions on the Internet*. Cambridge University Press.

Derenyi, I., Palla, G. et Vicsek, T. (2005). Clique percolation in random networks. *Physical Review Letters*, 94:160–202.

Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B. et Plaisant, C. (2007). Discovering interesting usage patterns in text collections : integrating text mining with visualization. *In Proc. of the Conference on Information and Knowledge Management*, pages 213–222.

Fekete, J. et Dufournaud, N. (2000). Compus : visualization and analysis of structured documents for understanding social life in the 16th century. *In Proc. of the ACM Conference on Digital Libraries*, pages 47–55.

Ge, R., Ester, M., Gao, B., Hu, Z., Bhattacharya, B. et Ben-Moshe, B. (2008). Joint cluster analysis of attribute data and relationship data. *ACM Transactions on Knowledge Discovering Data*, 2(2):1–35.

Hoey, M. (1991). *Patterns of Lexis in Text*. Describing English Language. Oxford University Press.

Hoey, M. (1997). Language and the subject. *In* Simms, K., éditeur : *Critical Studies*, chapitre The discourse's disappearing (and reappaering) subject : an exploration of the extent of Intertextual interference in the production of texts, pages 245–264. Rodopi.

KÁROLY, S. et FRANCIS, G. (2000). *Pattern Grammar, a corpus-driven approach to the lexical grammar of English*. John Benjamins.

KNIGHT, K. et MARCU, D. (2000). Statistics-Based Summarization — Step One : Sentence Compression. *In Proc. of the National Conference of the American Association for Artificial Intelligence*, pages 703–710.

LEGALLOIS, D. (2006). Des phrases entre elles à l'unité réticulaire du texte. *Langages*, 164:56–70.

LEGALLOIS, D., CELLIER, P. et CHARNOIS, T. (2011). Calcul de réseaux phrastiques pour l'analyse et la navigation textuelle. *In Actes de la Conférence sur le Traitement Automatique des Langues Naturelles*.

LIN, C.-Y. et HOVY, E. (2002). From single to multi-document summarization. *In Proc. of the Annual Meeting of the Association for Computational Linguistics*, pages 457–464.

MACNEILAGE, P. (2008). *The Origin of Speech*. UOP Oxford.

MOUGEL, P.-N., RIGOTTI, C. et GANDRILLON, O. (2012). Finding collections of k-clique percolated components in attributed graphs. *In Proc. of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. À paraître.

MUSIAL, K. et JUSZCZYSZYN, K. (2009). Properties of bridge nodes in social networks. *In Proc. of the International Conference on Computational Collective Intelligence*, pages 357–364.

NEWMAN, D., BALDWIN, T., CAVEDON, L., HUANG, E., KARIMI, S., MARTÍNEZ, D., SCHOLER, F. et ZOBEL, J. (2010). Visualizing search results and document collections using topic maps. *Web Semantics Science Services and Agents on the World Wide Web*, 8(2-3):169–175.

PLAISANT, C., ROSE, J., YU, B., AUVIL, L., KIRSCHENBAUM, M., SMITH, M., CLEMENT, T. et LORD, G. (2006). Exploring erotics in emily dickinson"s correspondence with text mining and visual interfaces. *In Proc. of the ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 141–150.

RENOUF, A. et KEHOE, A. (2004). Textual distraction as a basis for evaluating automatic summarisers. *In Proc. of the International Conference on Language Resources and Evaluation*.

SARDINHA, T. B. (1999). Looking at discourse in a corpus : The role of lexical cohesion. *In Proc. of the World Congress of Applied Linguistics*.

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proc. of the International Conference on Knowledge Discovery and Data Mining*.

TONG, H., GALLAGHER, B., FALOUTSOS, C. et ELIASSI-RAD, T. (2007). Fast best-effort pattern matching in large attributed graphs. *In Proc. of the International Conference on Knowledge Discovery and Data Mining*.

WASHIO, T. et MOTODA, H. (2003). State of the art of graph-based data mining. *SIGKDD Explorations*, 5(1):59–68.

ZHOU, Y., CHENG, H. et YU, J. (2010). Clustering large attributed graphs : An efficient incremental approach. *In Proc. of the International Conference on Data Mining*, pages 689–698.

Une étude en 3D de la paraphrase : types de corpus, langues et techniques

Houda Bouamor Aurélien Max Anne Vilnat
LIMSI-CNRS
Univ. Paris-Sud
Orsay, France
prenom.nom@limsi.fr

RÉSUMÉ

Cet article présente une étude détaillée de l'impact du type du corpus sur la tâche d'acquisition de paraphrases sous-phrastiques. Nos expériences sont menées sur deux langues et quatre types de corpus, et incluent une combinaison efficace de quatre systèmes d'acquisition de paraphrases. Nous obtenons une amélioration relative de plus de 27% en F-mesure par rapport au meilleur système, en anglais et en français, ainsi qu'une amélioration relative à notre combinaison de systèmes de 22% pour l'anglais et de 5% pour le français quand tous les types de corpus sont utilisés pour l'acquisition depuis le type de corpus le plus couramment disponible.

ABSTRACT

A study of paraphrase along 3 dimensions : corpus types, languages and techniques

In this paper, we report a detailed study of the impact of corpus type on the task of sub-sentential paraphrase acquisition. Our experiments are for 2 languages and 4 corpus types, and involve an efficient machine learning-based combination of 4 paraphrase acquisition systems. We obtain relative improvements of more than 27% in F-measure over the best individual system on English and French, and obtain a relative improvement over the combination system of 22% for English and 5% for French when using all other corpus types as additional training data for our most readily available corpus type.

MOTS-CLÉS : acquisition de paraphrases, constitution de corpus.

KEYWORDS: paraphrase acquisition, corpus collection.

1 Introduction

La variation paraphrastique est probablement l'une des caractéristiques les plus fascinantes de la langue naturelle : différentes expressions peuvent être utilisés pour véhiculer des sens très proches. Par exemple, les segments soulignés dans les phrases *elle semblait heureuse*$_1$ *de retrouver sa famille*$_2$ et *elle avait l'air contente*$_1$ *d'être à nouveau parmi les siens*$_2$ constituent des paires de paraphrases acceptables pouvant être exploitées dans divers contextes.

Il s'agit cependant de l'une des principales sources de complexité pour les processus de traitement automatique des langues. Les thésaurus encodés manuellement sont par nature incomplets, et ne sont pas disponibles pour toutes les langues. De plus, ils ne comprennent souvent pas d'ex-

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 267–280,
Grenoble, 4 au 8 juin 2012.

pressions de plusieurs mots qui sont nécessaires pour produire ou reconnaître automatiquement des paraphrases plus complexes. Le besoin d'acquérir automatiquement des paraphrases à partir de corpus de textes a ainsi été à l'origine de nombreux travaux. L'acquisition de paraphrases sous-phrastiques, que nous appellerons simplement *paraphrases* dans la suite, repose la plupart du temps sur l'appariement préalable d'unités de plus grande taille (des paires de phrases ou des documents comparables). Ces unités peuvent être obtenues directement par un processus supervisé, tel que la traduction humaine multiple, ou l'appariement automatique fondé sur la similarité entre textes (Mihalcea *et al.*, 2006). On observe que les techniques pour l'acquisition de paraphrases sont généralement très dépendantes des types de corpus sur lesquels elles ont été développées (Madnani et Dorr, 2010). Dans l'ordre inverse de leur disponibilité, ces types de corpus peuvent être grossièrement définis comme :

1. **corpus monolingues parallèles** : des paires d'énoncés de sens équivalents alignées de façon supervisée (comme les traductions multiples de livres (Barzilay et McKeown, 2001) ou les groupes de questions ayant la même réponse (Bernhard et Gurevych, 2008)).
2. **corpus multilingues parallèles** : des paires d'énoncés disponibles dans deux langues ou plus (Bannard et Callison-Burch, 2005) (comme les transcriptions des débats parlementaires européens)
3. **corpus monolingues comparables** : des paires de textes associés en fonction de similarité textuelle (comme des extraits de documents du Web (Pasça et Dienes, 2005)) en suivant éventuellement certaines heuristiques (tels que les articles de journaux publiés dans le même intervalle de temps (Dolan *et al.*, 2004))

Les ressources dans lesquelles les paraphrases abondent, ce qui facilite généralement une extraction *précise*, sont peu fréquentes à l'état naturel, alors que les unités de textes *comparables* sont potentiellement très nombreuses à l'échelle du Web (Pasça et Dienes, 2005; Bhagat et Ravichandran, 2008). Ces considérations nous conduisent à envisager d'améliorer les performances des techniques d'acquisition de paraphrases sur un type de corpus en utilisant du matériel d'apprentissage (i.e. des exemples annotés) à partir d'autres types de corpus.

Dans cet article, nous présentons une analyse détaillée de la tâche d'acquisition de paraphrases sur quatre types de corpus monolingues représentatifs, que nous avons nommés en fonction du *type de signal du contenu sémantique d'origine* :

– TEXTE : des paires de phrases résultant de traductions multiples d'un même texte.
– PAROLE : des paires d'énoncés résultant de traductions multiples de mêmes extraits de parole.
– SCÈNE : des paires d'énoncés résultant de descriptions multiples d'une même scène visuelle.
– ÉVÉNEMENT : des paires d'énoncés résultant de descriptions multiples d'un même événement ou de deux événements proches.

Notre étude sera menée sur des collections constituées d'un nombre identique de paires de phrases pour chacun des types de corpus, ceci pour deux langues, l'anglais et le français. Nous utiliserons quatre systèmes d'acquisition de paraphrases (Bouamor *et al.*, 2011) et décrirons une architecture efficace pour valider la combinaison de leurs hypothèses par apprentissage automatique. Nous détaillerons les quantités de paraphrases par type accessibles à partir de chacun des types de corpus étudiés, et nous donnerons les performances de chaque système individuel ainsi que notre système de combinaison sur chaque type de corpus pris indépendamment, et sur chaque type de corpus en ajoutant d'autres types de corpus comme données d'entraînement pour la validation.

L'un de nos principaux résultats est que, pour les deux langues étudiées, l'acquisition de paraphrases peut être significativement améliorée à l'aide des données d'entraînement de types de corpus différents. Ceci est notablement le cas pour le corpus ÉVÉNEMENT, source la plus facile à acquérir pour l'acquisition de paraphrases, ce qui ouvre d'intéressantes perspectives pour des études ultérieures sur l'acquisition de paraphrases à grande échelle et leur utilisation pour l'amélioration de la performance d'applications de TAL. Nous avons également élaboré une typologie, sur les deux langues, quantifiée des types de paraphrases sur chacun des types de corpus, à la fois pour les paraphrases de référence et pour celles que notre système de combinaison, le plus performant, parvient à acquérir sur chaque type de corpus, ce qui fournira des informations précieuses pour guider la suite de ces travaux.

Dans la suite de cet article, nous commencerons par un rapide état de l'art sur l'acquisition de paraphrases (section 2), puis nous décrirons la méthodologie de construction de nos corpus et leurs caractéristiques (section 3). Nous détaillerons ensuite nos résultats de l'évaluation de l'acquisition de paraphrases (section 4). À la section 5.1, nous présenterons tout d'abord la performance d'un système de combinaison sur chacun des types de corpus, puis la performance de ce système lorsque sont utilisées des données d'entraînement additionnelles provenant des autres types de corpus (5.2). Enfin nous conclurons en évoquant différentes pistes de recherche ouvertes par nos travaux (section 6).

2 État de l'art

Au cours du temps, l'acquisition et la génération de paraphrases ont attiré un grand nombre de travaux de recherche, qui sont trop nombreux pour être correctement résumés ici : Madnani et Dorr (2010) présentent une revue relativement complète des principales approches. L'acquisition de paraphrases au niveau de phrases entières a été abordée à partir de ressources spécifiques augmentant la probabilité de trouver des phrases en relation de paraphrase (Dolan *et al.*, 2004; Bernhard et Gurevych, 2008; Wubben *et al.*, 2009), à partir de corpus monolingues comparables (Barzilay et Elhadad, 2003; Fung et Cheung, 2004; Nelken et Shieber, 2006) ainsi qu'à partir du Web (Pasça et Dienes, 2005; Bhagat et Ravichandran, 2008).

Diverses techniques ont été proposées pour l'acquisition de paraphrases à partir de paires de phrases en relation (Barzilay et McKeown, 2001; Pang *et al.*, 2003) et à partir de corpus parallèles bilingues (Bannard et Callison-Burch, 2005; Kok et Brockett, 2010). Le lien entre la construction du corpus et le développement et l'évaluation des techniques d'acquisition est l'objet de (Cohn *et al.*, 2008; Callison-Burch *et al.*, 2008). À notre connaissance, il n'existe pas d'autres travaux portant sur l'acquisition de paraphrases qui soient menés sur plusieurs types de corpus et sur plusieurs langues de façon comparable. Pour sa part, le travail de Chan *et al.* (2011) explore la complémentarité de corpus bilingues et monolingues en acquisition de paraphrases. Faruqui et Padó (2011) s'intéressent à l'acquisition de *paires en relation d'implication* (prémisse et hypothèse), en menant des expériences dans trois langues sur des corpus journalistiques de différents domaines pour une langue. Bien que leur travail ne soit pas directement comparable au nôtre, ces auteurs montrent que la robustesse entre domaines est difficile à obtenir.

Enfin, l'évaluation de paraphrases générées automatiquement a été l'objet de quelques travaux récents (Liu *et al.*, 2010; Chen et Dolan, 2011; Metzler *et al.*, 2011), bien que ce problème reste difficile et globalement peu résolu. La génération de paraphrases motivée par une application

particulière offre un moyen indirect pour l'évaluation de la performance de la génération de paraphrases (Zhao *et al.*, 2009). Par exemple, le domaine de la Traduction Automatique Statistique est à l'origine de travaux montrant l'utilité à la fois de paraphrases produites par des humains (Schroeder *et al.*, 2009; Resnik *et al.*, 2010) et produites automatiquement (Madnani *et al.*, 2008; Marton *et al.*, 2009; Max, 2010) pour améliorer la performance en traduction.

3 Collecte de corpus de paires de phrases

Dans cet article, nous nous intéressons à l'acquisition de paraphrases à partir de paires de phrases en relation, caractéristiques de quatre types de corpus. Un corpus pour chaque type a été construit en deux langues, l'anglais et le français, et comporte 625 paires de phrases par langue. Nous détaillons maintenant la méthode de constitution de ces corpus.

TEXTE Pour l'anglais, nous avons utilisé le corpus MTC [1] (décrit dans (Cohn *et al.*, 2008)), qui regroupe des ensembles d'articles d'actualité traduits plusieurs fois depuis le chinois, et pour le français le corpus CESTA [2] regroupant des ensembles d'articles d'actualité traduits depuis l'anglais. Pour chaque groupe de phrases, nous retenons les paires de phrases ayant la plus petite distance d'édition au-dessus d'un seuil fixé empiriquement, en les extrayant d'abord de chacun des groupes et en reconsidérant par la suite des groupes déjà utilisés pour atteindre le nombre visé de paires de phrases.
Exemple : « *Dans l' autre cas , le gel des terres est destiné à maîtriser l' offre .* ⟷ *Le deuxième type de gel de terres doit servir à la gestion de l' offre .* »

PAROLE Pour l'anglais, nous avons utilisé des fichiers [3] librement disponibles de sous-titres de films tournés en français, *Le Fabuleux Destin d'Amélie Poulain* et *Les Choristes*, et pour le français nous avons pris les fichiers de la série télévisée tournée en anglais américain *Desperate Housewives*. Nous avons d'abord aligné chaque corpus parallèle en utilisant l'algorithme décrit dans (Tiedemann, 2007), basé sur des indices de durée et développé pour des sous-titres multilingues, puis nous avons extrait des paires de phrases en dessous d'un seuil minimal de distance d'édition, et filtré manuellement les erreurs apparentes de l'algorithme précédent.
Exemple : « *Vous pourriez passer ce soir et regarder ma tuyauterie ?* ⟷ *Pourriez-vous venir inspecter ma tuyauterie ce soir ?* »

SCÈNE Nous avons utilisé le *Multiple Video Description Corpus* (Chen et Dolan, 2011) obtenu à partir de descriptions multiples de courtes vidéos. De façon analogue à ce qui a été fait pour TEXTE, nous avons choisi des paires de phrases au sein de ces groupes en fonction d'une distance minimale d'édition au-dessus d'un certain seuil. Un point important est que pour l'anglais nous avons pu utiliser des descriptions qualifiées de "vérifiées". Cependant, les descriptions en français dans cette ressource sont disponibles dans des quantités bien moins importantes, et en outre aucune n'a le statut de "vérifiée". Nous avons tout de même décidé d'utiliser ce corpus, mais en gardant à l'esprit que cette source est de nettement moins bonne qualité. [4]
Exemple : « *une personne met du lait sur du riz.* ⟷ *un homme fait du riz au lait.* »

1. `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002T01`
2. `http://www.elda.org/article125.html`
3. `http://www.opensubtitles.org`
4. Ce type de corpus sera désigné entre parenthèse pour le français ("(SCÈNE)") dans tous les tableaux dans la suite pour rappeler son caractère particulier.

ÉVÉNEMENT Nous avons utilisé des titres de groupes d'articles d'actualité provenant du service d'agrégation Google News [5]. Nous avons ensuite affiné l'algorithme de regroupement en retenant les paires de titres dont les dates de publication des articles n'étaient pas séparées de plus d'un jour. Nous avons reproduit la même procédure de sélection que pour TEXTE et SCÈNE pour obtenir une couverture maximale sur l'ensemble des groupes.

Exemple : « *700 000 décès liés au Sida ont pu être évités en 2010* ⟷ *Forte baisse des décès et des infections liés au sida en 2010* »

	Statistiques du corpus 500 paires de phrases		**Accords inter-annotateur** 50 paires de phrases		**Stat. sur les formes dans les paraphrases** sans les paraphrases identiques			
					para. sûres		para. possibles	
	# formes	# formes par phrases	para. sûres	para. possibles	% formes	# formes	% formes	# formes
ANGLAIS								
TEXTE	21 473	21,0	66,1	20,4	18,6	4 004	12,3	2 651
PAROLE	11 049	10,5	79,1	10,9	17,5	1 942	31,6	3 500
SCÈNE	7 783	7,5	80,5	35,2	10,9	851	14,0	1 094
ÉVÉNEMENT	8 609	8,0	65,3	20,5	17,5	1 506	14,5	1 251
FRANÇAIS								
TEXTE	24 641	24,0	64,6	16,6	29,2	7 218	6,2	1 527
PAROLE	11 850	11,5	82,7	20,8	22,5	2 667	16,7	1 981
(SCÈNE)	7 012	6,5	42,8	9,3	3,9	275	9,4	664
ÉVÉNEMENT	9 121	9,1	67,8	3,8	19,6	1 793	9,6	876

TABLE 1 – Description de l'ensemble des corpus collectés et des annotations de référence pour les paraphrases en anglais et en français. Pour rappel, SCÈNE pour le français apparaît entre parenthèses car nous ne le considérons pas de la même qualité que les autres corpus.

Nous avons ensuite réalisé une annotation des paraphrases dans ces corpus, en suivant l'essentiel des consignes décrites dans (Cohn *et al.*, 2008) [6] à l'aide de l'outil YAWAT (Germann, 2008), mis à part le fait que nous ne sommes pas partis d'alignements initiaux obtenus automatiquement afin de ne pas biaiser le travail des annotateurs. Les principales consignes étaient que les paraphrases *sûres* et *possibles* devaient être distinguées, que les alignements les plus petits devaient être privilégiés sans décourager néanmoins les alignements groupe-à-groupe (i.e. *n-m*), et que les phrases devaient être alignées autant que possible. Nous ne considérerons dans la suite, pour toutes les statistiques et les expériences, que les paraphrases qui ne sont pas des paires identiques (telles que « *petit pont de bois* ⟷ *petit pont de bois* »), car on peut les considérer comme triviales au regard de la tâche d'acquisition.

La table 1 indique différentes statistiques pour les corpus collectés. La première observation est que TEXTE contient des phrases significativement plus longues que les autres types, plus de deux fois plus longues que celles de PAROLE par exemple. La table contient également les valeurs d'accords inter-annotateurs [7] calculées sur des sous-ensembles de 50 paires de phrases annotées indépendamment par deux annotateurs. Nous considérons comme acceptables les valeurs obtenues pour les paraphrases sûres, mais les valeurs obtenues pour les paraphrases possibles sont faibles. Ce dernier résultat était relativement prévisible, étant donné le nombre d'in-

5. `http://news.google.com`

6. Voir `http://staffwww.dcs.shef.ac.uk/people/T.Cohn/paraphrase_guidelines.pdf`

7. Pour chaque type de paraphrase, nous calculons la moyenne des valeurs de rappel obtenues par chaque annotateur comme référence et nous effectuons la moyenne.

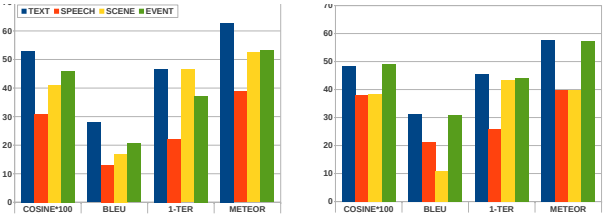


FIGURE 1 – Moyenne des similarités des paires de phrases pour tous les corpus pour l'anglais (à gauche) et le français (à droite) en utilisant le cosinus des vecteurs de formes, BLEU (Papineni *et al.*, 2002), TER (Snover *et al.*, 2006) et METEOR (Lavie et Agarwal, 2007) (à noter que les synonymes de WordNet ne pourraient être utilisés que pour l'anglais).

terprétations pour les paraphrases entrant possiblement dans cette catégorie. Cela ne constituera toutefois pas un problème dans la suite dans nos expériences : comme nous le verrons dans la section 4, notre métrique d'évaluation ne les considèrera pas comme des solutions attendues, et se limitera à ne pas les considérer comme fausses lorsqu'elles apparaitront parmi les hypothèses d'un système.

La table 1 montre enfin les pourcentages et les nombres de paraphrases pour chaque niveau de certitude pour chacun des corpus. Nous obtenons approximativement le même nombre total de paraphrases pour l'anglais (16 799) et le français (17 001). Les corpus anglais ont à peu près le même nombre de paraphrases sûres et possibles (8 303 par rapport à 8 496) alors qu'en français on trouve davantage de paraphrases sûres (11 953 contre 5 048). Ceci peut s'expliquer par le fait que les annotateurs ont travaillé indépendemment, avec des interprétations possiblement différentes de la tâche, et que les corpus, aussi *comparables* soient-ils entre langue, sont différents par nature. Les autres faits remarquables sont que TEXTE contient beaucoup plus de paraphrases que les autres corpus et que PAROLE comporte proportionnellement plus de paraphrases possibles que les autres corpus, et que SCÈNE contient nettement moins de paraphrases, en pourcentage et en nombre.

Dans la figure 1 différentes mesures typiquement utilisées, notamment en Traduction Automatique, de similarité entre paires de phrases sont données. TEXTE contient les paires de phrases les plus similaires selon toutes les métriques, suivi de près par ÉVÉNEMENT (dont les phrases sont beaucoup plus courtes). SCÈNE contient des paires de phrases qui sont plus similaires que celles de PAROLE pour l'anglais, ce qui n'est pas le cas pour le français.

4 Évaluation de l'acquisition de paraphrases

Nous adoptons la méthodologie PARAMETRIC de Callison-Burch *et al.* (2008) pour évaluer la performance des systèmes sur la tâche d'acquisition de paraphrases sur les corpus décrits dans la section précédente. Dans PARAMETRIC, un ensemble de paraphrases candidates extraites d'une paire de phrases en relation est comparé à un ensemble de paraphrases de référence, obtenues

par annotation manuelle, en calculant les mesures habituelles de *précision* (P) et *rappel* (R). La première valeur correspond à la proportion de paires d'hypothèses de paraphrases, ensemble noté H, produites par un système qui sont correctes par rapport à l'ensemble de référence contenant les paraphrases *sûres* et *possibles*, noté R_{tout}. Le rappel est obtenu en calculant la proportion de l'ensemble de référence de paraphrases *sûres*, noté $R_{\text{sûr}}$,qui sont trouvées par un système. Nous calculons également une valeur de F-mesure (F_1), qui considère le rappel et la précision comme également importants. Ces valeurs sont donc données par les formules suivantes :

$$P = \frac{|H \cap R_{\text{tout}}|}{|H|} \quad R = \frac{|H \cap R_{\text{sûr}}|}{|R_{\text{sûr}}|} \quad F_1 = \frac{2PR}{P+R}$$

Il est à noter que la façon dont les ensembles R_{tout} et $R_{\text{sûr}}$ de paires de paraphrases de référence sont définis garantit que les hypothèses de paraphrases incluant les paraphrases de référence annotées comme *possibles* ne pénaliseront pas la précision sans toutefois augmenter le rappel.

Toutes les valeurs de performance fournies dans les sections suivantes sont obtenues en effectuant une validation croisée 10 fois [8] au lieu d'utiliser le découpage des corpus en corpus de test/ corpus d'apprentissage. Nous moyennons, par la suite, les résultats sur chaque ensemble d'évaluation individuel pour obtenir des valeurs stables. Tous nos ensembles de données pour la validation croisée contiennent 500 paires de phrases. [9]

5 Expériences bilingues

5.1 Une architecture pour l'acquisition de paraphrases sous-phrastiques

Nous allons maintenant décrire les systèmes qui seront testés sur nos divers corpus décrits dans la section 3 utilisant la méthodologie décrite dans la section 4. Ces systèmes individuels sont décrits plus en détails dans (Bouamor *et al.*, 2011). Un système de combinaison est en outre utilisé pour valider automatiquement les hypothèses de paraphrases produites par les systèmes individuels en utilisant un ensemble de traits visant à reconnaître des paraphrases. Quatre systèmes individuels ont été utilisés et sont décrits ci-dessous : les raisons pour avoir retenu ces systèmes incluent leur libre disponibilité et/ou le coût raisonnable de leur développement, la possibilité d'utiliser des ressources comparables là où pertinent pour les deux langues étudiées, ainsi que les caractéristiques spécifiques à chaque technique.

Apprentissage statistique d'alignements de mots (GIZA) L'outil GIZA++ (Och et Ney, 2004) calcule des modèles statistiques d'alignement de mots de complexité croissante à partir de corpus parallèles. Il a été lancé sur chacun des corpus monolingues de paires de phrases dans les deux directions, les alignements ont été *symétrisés* puis les heuristiques classiques d'extraction de bi-segments *cohérents* ont été appliquées (Koehn *et al.*, 2003), sans toutefois agrandir les bi-segments par ajout de mots non alignés aux frontières.

Connaissances linguistiques sur la variation de termes (FASTR) L'outil FASTR (Jacquemin, 1999) permet de repérer des variantes de termes dans de grands corpus, les variations étant décrites à l'aide de métarègles spécifiant les dérivations morpho-syntaxiques possibles à partir

8. La validation croisée nous permet d'utiliser la totalité des données disponibles.

9. Il faut noter que, sur les 625 paires de phrases de départ pour chaque corpus, 125 paires de phrases sont extraites pour optimiser les paramètres d'un système basé sur la métrique TER_p (voir section 5.1).

d'un terme donné au moyen d'expressions régulières sur les catégories morpho-syntaxiques. La variation paradigmatique peut aussi s'exprimer au moyen de contraintes entre mots, imposant qu'ils appartiennent à la même famille morphologique ou sémantique en utilisant des ressources existantes disponibles pour nos deux langues. Les variantes pour tous les groupes de mots d'une des phrases d'une paire sont extraites dans l'autre phrase, et l'on conserve l'intersection des ensembles obtenus dans les deux directions.

Transformations optimales entre séquences de mots (TER$_p$) L'outil TER$_p$ (Snover *et al.*, 2010) peut être utilisé pour calculer un ensemble optimal (modulo quelques approximations) d'éditions au niveau des mots et des segments qui permettent de transformer une phrase en une autre. [10] Les types d'éditions sont paramétrés par un ou plusieurs poids qui sont optimisés par *hill climbing* pour maximiser la F-mesure, avec 100 redémarrages aléatoires, en utilisant les 125 paires de phrases réservées à cette fin dans chaque type de corpus.

Équivalence de traduction (PIVOT) Nous avons exploité la probabilité de paraphrase définie par Bannard et Callison-Burch (2005) pour des paraphrases extraites de corpus parallèles multilingues. Nous avons utilisé le corpus Europarl [11] de débats parlementaires en anglais et en français, comprenant environ 1,7 millions de phrases parallèles, en prenant chaque langue comme source et pivot à tour de rôle. GIZA++ a été utilisé pour aligner les mots et les probabilités de traduction de segments ont été estimées à partir de ces alignements par les méthodes standards du système de traduction statistique MOSES (Koehn *et al.*, 2007). Pour chaque segment d'une paire de phrases, nous avons construit son ensemble de paraphrases, et extrait sa paraphrase de l'autre phrase ayant la plus grande probabilité. Nous avons réitéré ce processus dans les deux directions, et finalement conservé pour chaque segment la paire de paraphrases issue d'une des deux directions avec la probabilité la plus forte.

Combinaison de systèmes par validation En calculant l'union de toutes les hypothèses de paraphrases issues de tous les systèmes précédents pour chaque paire de phrases, nous avons procédé à une classification en deux classes (soit, "paraphrase" ou "non paraphrase") en utilisant un classifieur à maximum d'entropie MAXENT [12]. Ceci permet d'inclure des traits qui n'étaient pas nécessairement pris en compte ou possibles à considérer dans les systèmes individuels. Plus généralement, ceci permet de tenter d'apprendre une caractérisation plus générique des paraphrases, qui pourrait s'adapter trivialement à un nombre quelconque de systèmes en entrée. Les exemples positifs pour le classifieur sont ceux provenant de l'union des hypothèses qui sont également présentes dans l'ensemble de référence $R_{\text{sûr}}$. Les exemples négatifs sont constitués du complément de cet ensemble dans l'union. Les traits que nous utilisons sont résumés dans la table 2.

Résultats expérimentaux Les résultats pour les systèmes individuels, leur union et nos systèmes de combinaison entraînés sur chaque type de corpus (colonne "appr.=C") sont donnés dans la Figure 3. Nous constatons tout d'abord que tous les systèmes obtiennent de meilleurs résultats sur TEXTE, pour lequel il y avait plus de données d'apprentissage disponibles et dans lequel les équivalences sémantiques entre les paires de phrases étaient plus probables. ÉVÉNEMENT apparaît comme le type de corpus le plus difficile, ce qui pourrait être considéré comme un

10. Il est à noter que contrairement à ce que TER$_p$ permet, nous n'utilisons pas les équivalents de mots ou de segments proposés par défaut car ceux-ci ne sont disponibles que pour l'anglais. Ce type de connaissance sera néanmoins apporté par les systèmes FASTR et PIVOT.

11. `http://statmt.org/europarl`

12. Nous avons utilisé l'implémentation disponible à : `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`

Traits dérivés des paires de segments – distance d'édition entre les paraphrases, racines identiques, mêmes formes, longueur de la phrase
Traits dérivés des paires de phrases – similarité entre paires de phrases (cosinus, BLEU, TER, METEOR), position relative des paraphrases, présence de formes communes aux frontières des paraphrases, présence d'une autre paire de paraphrases de chaque système aux frontières de la paraphrase, présence d'une paraphrase à un autre endroit dans l'autre phrase
Traits distributionnels – similarité (cosinus) des vecteurs de formes du contexte pour chaque segment d'une paraphrase (dérivée de fréquences obtenues dans le grand corpus parallèle anglais-français fourni pour la campagne d'évaluation WMT'11 (`http://www.statmt.org/wmt11/translation-task.html`, soit environ 30 millions de phrases parallèles)
Traits dérivés des systèmes – combinaison des systèmes individuels qui proposent la paire de paraphrase

TABLE 2 – Principaux traits utilisés par nos classifieurs. Des intervalles discrétisés basés sur les valeurs médianes sont utilisés pour les valeurs réelles, et des valeurs binarisées sont utilisées pour indiquer les configurations présentes pour les combinaisons.

Corpus type (C)	Systèmes individuels												Combinaison de systèmes								
	GIZA			FASTR			$TER_{\rightarrow F}$			PIVOT			union			appr.=C			appr.=tout		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
ANGLAIS																					
TEXTE	48,2	58,9	53,0	63,1	5,9	10,7	41,2	66,4	50,9	73,4	25,8	38,2	20,8	**80,8**	33,1	68,4	62,8	**65,5**	**77,5**	56,1	65,1
PAROLE	39,7	44,2	41,8	27,1	3,5	6,3	25,0	50,3	33,4	**79,2**	15,3	25,7	25,5	**71,4**	37,6	51,0	56,3	53,5	67,7	48,7	**56,6**
SCÈNE	44,8	57,7	50,5	47,4	5,2	9,5	40,1	67,9	50,4	**84,6**	14,6	25,0	36,2	**83,4**	50,5	44,9	66,8	**53,7**	33,2	59,7	42,7
ÉVÉN.	19,0	33,9	24,3	62,9	3,1	6,0	28,8	68,7	40,6	**97,4**	11,2	20,1	20,8	**75,5**	32,7	35,0	67,1	46,0	56,4	56,1	**56,2**
FRANÇAIS																					
TEXTE	52,5	58,9	55,5	56,9	4,9	9,1	46,4	61,4	52,8	64,5	30,3	41,2	41,5	**77,9**	54,1	**74,7**	61,0	**67,1**	**74,5**	60,2	66,6
PAROLE	44,0	54,9	48,9	30,7	4,3	7,6	34,8	60,2	44,1	**75,5**	19,0	30,4	31,4	**76,2**	44,5	60,2	59,7	**60,0**	55,1	61,0	57,9
(SCÈNE)	14,4	43,6	21,7	53,0	4,0	7,4	13,8	75,3	23,4	**94,6**	5,21	9,8	12,7	**86,4**	22,2	19,9	59,8	**29,8**	12,5	69,4	21,1
ÉVÉN.	28,7	44,2	34,8	34,4	2,3	4,3	29,9	58,9	39,7	**79,5**	15,0	25,2	25,2	**72,5**	37,4	40,0	56,3	46,8	62,4	40,7	**49,3**

TABLE 3 – Résultats de l'évaluation pour chaque système individuel (à gauche) et les systèmes combinés (à droite) sur tous les types de corpus, pour l'anglais (en haut) et le français (en bas). Les valeurs en gras indiquent les meilleurs résultats pour une métrique donnée pour chaque type de corpus et chaque langue.

résultat décevant dans la mesure où il s'agit du type pour lequel il existe le plus de données prêtes à être utilisées : nous reviendrons sur ce point dans la section 5.2.

En termes de performance en F-mesure par type de corpus, GIZA obtient de meilleurs résultats sur TEXTE et PAROLE, qui contiennent des phrases longues, avec d'éventuelles répétitions, alors que $TER_{\rightarrow F}$ a de meilleurs résultats sur SCÈNE et ÉVÉNEMENT, où les équivalences qui sont rares au niveau corpus sont plus fréquentes. Pour des raisons de place, nous ne détaillerons pas plus dans cet article les performances des systèmes individuels, pour nous concentrer sur nos combinaisons de systèmes.

Dans toutes les configurations, la combinaison de systèmes améliore de façon importante la F-mesure relativement au meilleur des systèmes individuels pour chaque type de corpus, ainsi que relativement à l'union des résultats de l'ensemble des systèmes. Les améliorations sont importantes sur TEXTE (respectivement +12,5 et +11,6 sur l'anglais et le français) et sur PAROLE (+11,7 et +11,1) et assez bonnes sur SCÈNE (+3,2 et +6,4) et sur ÉVÉNEMENT (+5,4 et +7,1).

	+Texte	+Parole	+Scène	+Événement	+Tous
ANGLAIS					
# ex+	7 342	2 296	1 784	1 171	12 593
Texte	65,5	**66,2**	65,1	**66,2**	65,1
Parole	56,0	53,5	52,8	54,8	**56,6**
Scène	49,7	**54,3**	53,7	53,8	42,7
Événement	51,1	45,3	42,5	46,0	**56,2**
FRANÇAIS					
# ex+	12 961	3 340	966	2 160	19 427
Texte	**67,1**	**67,2**	66,7	67,0	66,6
Parole	57,6	**60,0**	56,4	59,6	57,9
(Scène)	23,7	22,0	**29,8**	23,9	21,1
Événement	45,2	45,6	44,3	46,8	**49,3**

Table 4 – Résultats de l'évaluation (scores F_1) pour tous les types de corpus pour l'anglais (en haut) et le français (en bas) quand sont ajoutées les données d'entraînement des autres types de corpus (les valeurs sur fond grisé de la diagonale correspondent aux cas où aucune donnée n'est ajoutée). Les rangées "*#ex+*" indiquent le nombre d'exemples positifs de paraphrases apporté par chaque type de corpus supplémentaire sur le même nombre de paires de phrases.

Nous avons constaté (voir la table 1) que Texte et Parole sont les deux types de corpus ayant le plus grand nombre d'exemples de paraphrases sûres pour les deux langues : les résultats montrent que notre classifieur a été capable de les utiliser efficacement.

Les valeurs de rappel pour l'union sont assez grandes pour tous les types de corpus, allant de 71,4 (pour Parole en anglais) à 83,4 (pour Scène en anglais). Il y a, cependant, une nette baisse entre les valeurs de rappel pour les unions et pour les résultats de nos classifieurs, bien que ces dernières soient toutes autour de 6/10, ce qui peut être considéré comme une valeur acceptable pour une tâche de cette complexité. Une étude plus approfondie des faux négatifs pourrait nous aider à déterminer de nouveaux traits pour reconnaître des paraphrases plus difficiles à identifier. Enfin, nous pouvons noter que la précision est en général meilleure pour un des systèmes (Pivot), et atteint des valeurs intéressantes en particulier sur Texte, où nous disposons du plus grand nombre d'exemples (F-mesure de respectivement 68,4 et 74,6 pour l'anglais et le français).

5.2 Expériences sur l'apport des autres types de corpus

Nous considérons à présent la possibilité d'améliorer la performance de notre système de combinaison par l'utilisation de données d'apprentissage provenant d'autres types de corpus. Pour cela, nous construisons des systèmes en utilisant tout d'abord les données additionnelles provenant d'un autre type de corpus, puis de l'ensemble des types de corpus disponibles. Les résultats obtenus sont donnés dans la table 4 [13].

Nous observons qu'il existe deux cas de figure. Dans le premier, la performance en F-mesure est améliorée pour l'anglais sur Texte (+0,7), Parole (+3,1) et Scène (+0,6) en utilisant soit un seul type de corpus supplémentaire, soit l'ensemble des corpus disponibles, alors que pour le français

13. Nos résultats sont toujours donnés en procédant à une validation croisée qui réalise une moyenne des résultats obtenus sur 10 ensembles de test pour chaque type de corpus testé.

aucun ajout de données d'apprentissage n'améliore la performance pour ces types de corpus. Dans le second cas, ÉVÉNEMENT est amélioré à la fois pour l'anglais (+10,2) et pour le français (+2,5) en utilisant toutes les données d'apprentissage supplémentaires disponibles. Hormis la condition où les données provenant de TEXTE sont ajoutées pour l'anglais, tous les ajouts d'autres types de corpus diminuent la performance quand ils sont ajoutés individuellement : on observe donc ici nettement une contribution collective attribuable à l'ajout d'au moins deux sources. La nature des exemples pertinents ainsi ajoutés retiendra notre attention pour de futurs travaux : la sélection plus fine d'exemples pourrait effectivement repousser davantage la performance atteinte.

On peut encore noter que TEXTE n'est pratiquement pas touché par l'ajout de données supplémentaires, ce qui peut s'expliquer en partie par le fait que ce type de corpus contient à lui seul la moitié du nombre total d'exemples dans les deux langues. À l'opposé, SCÈNE, qui a le plus petit nombre d'exemples d'entraînement, voit sa performance baisser sensiblement, assez fortement par exemple avec l'ajout des données provenant de TEXTE (respectivement -4,0 et -6,1 pour l'anglais et le français) et par tous les corpus ensemble (respectivement -11,0 et -8,7). Ceci souligne à nouveau la nature spécifique de ce type de corpus : des descriptions indépendantes de la même scène vidéo peuvent être verbalisées de façons très diverses, à différents niveaux. Finalement, il y a nettement plus d'exemples positifs en français (19 427) qu'en anglais (12 593) : ceci peut s'expliquer par le fait que les phrases en français dans nos corpus contiennent plus de formes (voir Table 1) et que les paraphrases en français contiennent plus de variantes morphologiques telles que différentes formes conjuguées des verbes.

6 Discussion et perspectives

Dans cet article, nous nous sommes intéressés au problème de l'acquisition de paraphrases sur des types de corpus et entre ces types de corpus, en définissant les types de corpus à partir de l'origine du signal du contenu sémantique des paires de phrases utilisées : un texte dans différentes langues (TEXTE), de la parole transcrite dans une autre langue (PAROLE), une scène visualisée (SCÈNE), et une courte description (un titre d'article) d'un événement donné (ÉVÉNEMENT). Nous avons décrit un grand corpus annoté, contenant 2 500 paires de phrases pour l'anglais et pour le français, et nous avons réutilisé les principes généraux d'une méthodologie existante pour évaluer l'acquisition automatique de paraphrases (Callison-Burch *et al.*, 2008). Nous avons évalué un système efficace de combinaison exploitant les hypothèses de quatre systèmes, ainsi que l'impact produit par l'utilisation des données d'entraînement des autres types de corpus.

Notre résultat le plus prometteur est certainement l'amélioration obtenue sur le type de corpus ÉVÉNEMENT en utilisant les données d'entraînement de tous les corpus disponibles. Étant donné que les autres types de corpus sont beaucoup plus rares par nature, il semble que la disponibilité de tels corpus permet néanmoins d'apporter des connaissances utiles pour améliorer la reconnaissance des paraphrases sur ce qui s'est avéré être le type de corpus le plus difficile dans notre étude. Un résultat de cette nature incite à appliquer et améliorer nos techniques pour l'acquisition de paraphrases à l'échelle du Web (Pasça et Dienes, 2005; Bhagat et Ravichandran, 2008), où les paires de phrases en relation peuvent être très nombreuses.

Une piste intéressante porte sur l'amélioration des traits de détermination du statut de paraphrases, en particulier si l'on travaille sur les résultats d'un moteur de recherche sur le Web,

incluant des mesures de similarités entre paires de texte plus informées, par exemple en exploitant les structures thématiques des documents (Barzilay et Elhadad, 2003), des mesures de similarité lexicale en contexte (Dinu et Lapata, 2010; Erk et Pado, 2010), ou des résultats de systèmes d'implication textuelle (Kouylekov et Negri, 2010) [14]

Une analyse fine des différents types de paraphrases serait nécessaire pour servir de guide pour des travaux futurs afin de repousser les limites des systèmes actuels : de premiers résultats d'une telle analyse quantitative, pour tous les types de corpus et les deux langues de notre étude, sont donnés dans la Table 5. La principale observation est que la synonymie (comme dans *dans l'affirmative ⟷ le cas échéant*) est le phénomène le plus courant, qui de plus représente le principal type d'hypothèses correctes proposées par nos systèmes. En revanche, il n'est pas surprenant de voir que nos systèmes ne sont pas compétents pour reconnaître des paraphrases dans la catégorie "pragmatique", ce qui requiert de nombreuses et coûteuses informations sur le monde et sur le contexte des paires de phrases. Enfin, il est intéressant de noter que le type de corpus ÉVÉNEMENT contient des paraphrases de référence de tous les types.

	synonymie		**typographie**		**inclusion**		**pragmatique**		**syntaxe**		**dérivation**		**flexion**	
	%réf	%sys	%réf	%sys	%réf	%sys	%réf	%sys	%réf	%sys	%réf	%sys	%réf	%sys
ANGLAIS														
TEXTE	51,2	43,5	7,6	7,0	12,1	16,4	0,6	0,0	4,4	4,7	12,1	10,5	11,5	17,6
PAROLE	39,8	34,0	25,6	38,2	12,3	6,3	1,7	0,0	3,5	0,0	3,5	2,1	13,2	19
SCÈNE	50,0	46,8	1,3	2,1	21,6	23,4	0,0	0,0	1,3	0,0	5,4	8,5	20,2	19,1
ÉVÉNEMENT	36,9	41,6	15,0	22,2	19,1	16,6	1,3	0,0	6,8	2,7	6,8	2,7	13,6	13,8
FRANÇAIS														
TEXTE	46,9	26,0	9,0	20,6	2,1	1,0	3,6	1,0	6,6	0,0	3,0	3,2	28,5	47,7
PAROLE	45,5	43,9	14,2	19,5	8,0	7,3	2,6	0,0	11,6	2,4	3,5	2,4	14,2	24,3
(SCÈNE)	46,4	51,3	5,3	2,7	8,9	5,4	0,0	0,0	5,3	0,0	0,0	0,0	33,8	40,5
ÉVÉNEMENT	28,3	16,6	19,7	27,7	16,0	11,1	7,4	0,0	8,6	5,5	7,4	0,0	12,2	38,8

TABLE 5 – Distribution des types de paraphrases mesurée dans 50 paires de phrases annotées (%réf) choisies aléatoirement et des hypothèses de paraphrases sur ces phrases pour notre meilleur système (%sys) pour l'anglais (en haut) et le français (en bas). À noter que %sys doit être examiné en relation avec le rappel du système donné dans la table 3. Les types sont illustrées par les exemples suivants : (*dans l'affirmative ⟷ le cas échéant*) (**synonymie**), (*Classement ⟷ Class.*) (**typographie**)„ (*BNP ⟷ BNP Paribas*) (**inclusion**), (*de plus en plus sales ⟷ ne se brossent plus les dents*) (**pragmatique**), (*il y a 6 mois ⟷ six mois avant*) (**syntaxe**), (*refroidie ⟷ froide*) (**dérivation**), (*crevette ⟷ crevettes*, *moque ⟷ moquait*) (**flexion**)

Références

BANNARD, C. et CALLISON-BURCH, C. (2005). Paraphrasing with bilingual parallel corpora. *In Actes de ACL*, Ann Arbor, USA.

BARZILAY, R. et ELHADAD, N. (2003). Sentence alignment for monolingual comparable corpora. *In Proceedings of EMNLP*, Sapporo, Japan.

14. Concernant les systèmes d'implication textuelle, nous nous trouvons face à un problème de dépendance circulaire, car ces systèmes se fondent typiquement sur des connaissances préalables, notamment des paraphrases. Nous pensons quand même que l'utilisation de telles connaissances doit être faite quand celles-ci sont disponibles, comme nous l'avons fait nous-mêmes par l'utilisation du système FASTR et de ses ressources lexico-sémantiques associées.

BARZILAY, R. et MCKEOWN, K. (2001). Extracting paraphrases from a parallel corpus. *In Actes de ACL*, Toulouse, France.

BERNHARD, D. et GUREVYCH, I. (2008). Answering Learners' Questions by Retrieving Question Paraphrases from Social Q&A Sites. *In Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications*, Columbus, USA.

BHAGAT, R. et RAVICHANDRAN, D. (2008). Large scale acquisition of paraphrases for learning surface patterns. *In Actes de ACL-HLT*, Columbus, USA.

BOUAMOR, H., MAX, A. et VILNAT, A. (2011). Combinaison d'informations pour l'alignement monolingue. *In Actes de TALN*, Montpellier, France.

CALLISON-BURCH, C., COHN, T. et LAPATA, M. (2008). Parametric : An automatic evaluation metric for paraphrasing. *In Actes de COLING*, Manchester, UK.

CHAN, T. P., CALLISON-BURCH, C. et VAN DURME, B. (2011). Reranking bilingually extracted paraphrases using monolingual distributional similarity. *In Proceedings of the EMNLP Workshop on Geometrical Models of Natural language Semantics*, Edinburgh, UK.

CHEN, D. et DOLAN, W. (2011). Collecting highly parallel data for paraphrase evaluation. *In Proceedings of ACL-HLT*, Portland, Oregon, USA.

COHN, T., CALLISON-BURCH, C. et LAPATA, M. (2008). Constructing corpora for the development and evaluation of paraphrase systems. *Comput. Linguist.*, 34(4).

DINU, G. et LAPATA, M. (2010). Measuring distributional similarity in context. *In Proceedings of EMNLP*, Cambridge, USA.

DOLAN, B., QUIRK, C. et BROCKETT, C. (2004). Unsupervised construction of large paraphrase corpora : Exploiting massively parallel news sources. *In Proceedings of Coling*, Switzerland.

ERK, K. et PADO, S. (2010). Exemplar-based models for word meaning in context. *In Proceedings of ACL*, Uppsala, Sweden.

FARUQUI, M. et PADÓ, S. (2011). Acquiring entailment pairs across languages and domains : A data analysis. *In Proceedings of IWCS*, Oxford, UK.

FUNG, P. et CHEUNG, P. (2004). Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. *In Proceedings of COLING*, Geneva, Switzerland.

GERMANN, U. (2008). Yawat :Yet Another Word Alignment Tool. *In Proceedings of the ACL-HLT*, Columbus, Ohio.

JACQUEMIN, C. (1999). Syntagmatic and paradigmatic representations of term variation. *In Actes de ACL*, College Park, USA.

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A. et HERBST, E. (2007). Moses : Open source toolkit for statistical machine translation. *In Proceedings of ACL*, Czech Republic.

KOEHN, P., OCH, F. J. et MARCU, D. (2003). Statistical Phrase-Based Translation. *In Proceedings of NAACL HLT*, Edmonton, Canada.

KOK, S. et BROCKETT, C. (2010). Hitting the Right Paraphrases in Good Time. *In Proceedings of NAACL*, Los Angeles, USA.

KOUYLEKOV, M. et NEGRI, M. (2010). An open-source package for recognizing textual entailment. *In Proceedings of the ACL*, Uppsala, Sweden.

LAVIE, A. et AGARWAL, A. (2007). METEOR : An automatic metric for MT evaluation with high levels of correlation with human judgments. *In Proceedings of the ACL Workshop on Statistical Machine Translation*, Prague, Czech Republic.

LIU, C., DAHLMEIER, D. et NG, H. T. (2010). PEM : A paraphrase evaluation metric exploiting parallel texts. *In Proceedings of EMNLP*, Cambridge, MA.

MADNANI, N. et DORR, B. J. (2010). Generating Phrasal and Sentential Paraphrases : A Survey of Data-Driven Methods . *Computational Linguistics*, 36(3).

MADNANI, N., RESNIK, P., DORR, B. et SCHWARTZ, R. (2008). Are multiple reference translations necessary ? investigating the value of paraphrased reference translations in parameter optimization. *In Proceedings of AMTA*, Waikiki, Hawaii.

MARTON, Y., CALLISON-BURCH, C. et RESNIK, P. (2009). Improved Statistical Machine Translation Using Monolingually-derived Paraphrases. *In Proceedings of EMNLP*, Singapore.

MAX, A. (2010). Example-based paraphrasing for improved phrase-based statistical machine translation. *In Proceedings of EMNLP*, Cambridge, MA.

METZLER, D., HOVY, E. et ZHANG, C. (2011). An empirical evaluation of data-driven paraphrase generation techniques. *In Proceedings of ACL-HLT*, Portland, USA.

MIHALCEA, R., CORLEY, C. et STRAPPARAVA, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *In Proceedings of AAAI*, Boston, USA.

NELKEN, R. et SHIEBER, S. M. (2006). Towards robust context-sensitive sentence alignment for monolingual corpora. *In Proceedings of EACL*, Trento, Italy.

OCH, F. J. et NEY, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4).

PANG, B., KNIGHT, K. et MARCU, D. (2003). Syntax-based alignement of multiple translations : Extracting paraphrases and generating new sentences. *In Actes de NAACL-HLT*, Canada.

PAPINENI, K., ROUKOS, S., WARD, T. et ZHU, W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. *In Proceedings of ACL*, Philadelphia, USA.

PASÇA, M. et DIENES, P. (2005). Aligning Needles in a Haystack : Paraphrase Acquisition Across the Web. *In Proceedings of IJCNLP*, Jeju Island, South Korea.

RESNIK, P., BUZEK, O., HU, C., KRONROD, Y., QUINN, A. et BEDERSON, B. B. (2010). Improving translation via targeted paraphrasing. *In Proceedings of EMNLP*, Cambridge, MA.

SCHROEDER, J., COHN, T. et KOEHN, P. (2009). Word Lattices for Multi-Source Translation. *In Proceedings of EACL*, Athens, Greece.

SNOVER, M., DORR, B. J., SCHWARTZ, R., MICCIULLA, L. et MAKHOUL, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *In Proceedings of AMTA*, Boston, USA.

SNOVER, M., MADNANI, N., DORR, B. J. et SCHWARTZ, R. (2010). TER-Plus : paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, 23(2-3).

TIEDEMANN, J. (2007). Building a multilingual parallel subtitle corpus. *In CLIN17*, Belgium.

WUBBEN, S., van den BOSCH, A., KRAHMER, E. et MARSI, E. (2009). Clustering and matching headlines for automatic paraphrase acquisition. *In EWNLG*, Athens, Greece.

ZHAO, S., LAN, X., LIU, T. et LI, S. (2009). Application-driven Statistical Paraphrase Generation. *In Proceedings of ACL-AFNLP*, Suntec, Singapore.

Détection et correction automatique d'erreurs d'annotation morpho-syntaxique du French TreeBank

Florian Boudin Nicolas Hernandez
Université de Nantes
prénom.nom@univ-nantes.fr

RÉSUMÉ

La qualité de l'annotation morpho-syntaxique d'un corpus est déterminante pour l'entraînement et l'évaluation de méthodes d'étiquetage. Cet article présente une série d'expériences que nous avons menée sur la détection et la correction automatique des erreurs du *French Treebank*. Deux méthodes sont utilisées. La première consiste à identifier les mots sans étiquette et leur attribuer celle d'une forme correspondante observée dans le corpus. La seconde méthode utilise les variations de *n*-gramme pour détecter et corriger les anomalies d'annotation. L'évaluation des corrections apportées au corpus est réalisée de manière extrinsèque en comparant les scores de performance de différentes méthodes d'étiquetage morpho-syntaxique en fonction du niveau de correction. Les résultats montrent une amélioration significative de la précision et indiquent que la qualité du corpus peut être sensiblement améliorée par l'application de méthodes de correction automatique des erreurs d'annotation.

ABSTRACT

Detecting and correcting POS annotation in the French TreeBank

The quality of the Part-Of-Speech (POS) annotation in a corpus has a large impact on training and evaluating POS taggers. In this paper, we present a series of experiments that we have conducted on automatically detecting and correcting annotation errors in the French TreeBank. Two methods are used. The first simply relies on identifying tokens with missing tags and correct them by assigning the tag the same token observed in the corpus. The second method uses n-gram variations to detect and correct conflicting annotations. The evaluation of the automatic correction is performed extrinsically by comparing the performance of different POS taggers in relation to the level of correction. Results show a statistically significant improvement in precision and indicate that the POS annotation quality can be noticeably enhanced by using automatic correction methods.

MOTS-CLÉS : Étiquetage morpho-syntaxique, correction automatique, qualité d'annotation.

KEYWORDS: Part-Of-Speech tagging, automatic correction, annotation quality.

1 Introduction

Le corpus arboré de Paris 7, également appelé *French Treebank* (FTB), est la plus grande ressource disponible de textes annotés syntaxiquement et morpho-syntaxiquement pour le français (Abeillé

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 281–291,
Grenoble, 4 au 8 juin 2012.

et al., 2003). Il est le résultat d'un projet d'annotation supervisée d'articles du journal Le Monde mené depuis plus d'une dizaine d'années. La quasi-totalité des méthodes d'étiquetage morpho-syntaxique du français utilisent cet ensemble de données que ce soit pour leur phase d'entraînement ou d'évaluation, e.g. (Crabbé et Candito, 2008; Denis et Sagot, 2010). La qualité de l'annotation du corpus est donc déterminante.

De la même manière que la plupart des corpus annotés morpho-syntaxiquement, e.g. le *Penn TreeBank* (Marcus *et al.*, 1993) pour l'anglais, le FTB a été construit de manière semi-automatique. Un étiqueteur automatique est d'abord appliqué sur l'ensemble des textes. Les sorties sont ensuite corrigées manuellement des éventuelles erreurs commises par l'outil. Malgré cette dernière étape, il est presque certain que des erreurs existantes ne sont pas corrigées et que de nouvelles erreurs sont introduites (les humains n'étant pas infaillibles). Plusieurs études illustrent d'ailleurs cette problématique en décrivant quelques unes des erreurs d'annotation récurrentes du FTB telles que l'absence d'étiquette ou la présence d'éléments XML vides [1] (Arun et Keller, 2005; Green *et al.*, 2011).

Dans cette étude, nous présentons une série d'expériences que nous avons menée sur la correction automatique du FTB. Nous détaillons les différentes erreurs que nous avons rencontrées ainsi que les solutions que nous appliquons. Deux méthodes sont utilisées. La première consiste à identifier les mots sans étiquette et leur attribuer celle d'une forme correspondante observée dans le corpus. La seconde méthode utilise les variations de *n*-gramme pour détecter et corriger les anomalies d'annotation. L'évaluation de la correction du corpus est réalisée de manière extrinsèque en étudiant l'impact du niveau de correction sur les performances de plusieurs méthodes d'étiquetage morpho-syntaxique.

Le reste de cet article est organisé comme suit. La section 2 présente le corpus *French Treebank* que nous utilisons dans cette étude. La section 3 est consacrée à la description de la méthode que nous proposons. Nous décrivons ensuite en section 4 nos résultats expérimentaux avant de présenter les travaux connexes aux nôtres. La section 6 conclut cette étude et donne quelques perspectives de travaux futurs.

2 Description du corpus *French Treebank*

Notre intérêt pour le FTB est motivé par deux objectifs : d'une part réaliser des traitements automatiques sur le corpus, et d'autre part, construire des modélisations permettant de prédire l'étiquette grammaticale d'un mot à l'aide d'approches statistiques ; ce deuxième objectif est un cas particulier du premier. Nos observations concernent donc à la fois la structure du corpus qui porte les annotations et la qualité de ses annotations grammaticales.

Le corpus est toujours en développement. La version que nous utilisons dans cette étude est datée de juillet 2010, elle est composée de 21 562 phrases pour 628 767 mots (*tokens*). Les fichiers qui composent le corpus sont au format XML (voir la Figure 1). Les mots sont répartis en 13 catégories principales (attribut `cat`), elles mêmes réparties en 34 sous-catégories (attribut `subcat`). De plus, les traits flexionnels (attribut `mph`), les lemmes (attribut `lemma`) et les mots composés

1. Certaines des erreurs recencées ne sont que des choix de représentation qui ne sont pas forcément des choix les plus adaptés dans une perspective de traitement automatique du corpus. C.f. `http://www.llf.cnrs.fr/Gens/Abeille/guide-morpho-synt.02.pdf` pour la représentation de « *du*» en deux balises tokens « *de*» et « *le*», la seconde ayant un contenu textuel vide.

(e.g. « *aujourd'hui* », « *mettre en garde* ») sont explicités. Ces derniers sont très nombreux dans le corpus : ≈14% des occurrences de tokens entrent dans un mot composé. On peut noter que la structure originale de la phrase (avec les caractères espaces) ainsi que l'identifiant du document source ne figurent pas dans le corpus.

```
<SENT nb="226">
 <NP>
  <w cat="D" [...] lemma="son" mph="1fss" subcat="poss">Ma</w>
  <w cat="N" [...] lemma="position" mph="fs" subcat="C">position</w>
 </NP>
 <VN>
  <w cat="V" [...] lemma="être" mph="P3s" subcat="">est</w>
 </VN>
 <NP>
  <w cat="D" [...] lemma="le" mph="fs" subcat="def">la</w>
  <w cat="N" [...] lemma="suivant" mph="fs" subcat="C">suivante</w>
 </NP>
 <w cat="PONCT" [...] lemma="." subcat="S">.</w>
</SENT>
```

FIGURE 1 – Exemple de phrase annotée extraite du fichier `lmf300_13000ep.cat.xml`, certains attributs ont été supprimées pour faciliter la lecture (`[...]`).

L'encodage natif du corpus est *iso-8859-1*. Le premier traitement que nous avons opéré est sa conversion en *utf-8* via l'outil GNU iconv. L'encodage *utf-8* est utilisé par défaut par l'ensemble des outils et des applications que nous utilisons. Seul le fichier `lmf7ad1co.aa.xml` fut récalcitrant et nous avons été amené à corriger les caractères accentués à l'aide de quelques règles de conversion ad hoc. Le FTB nécessite ensuite de nombreux pré-traitements avant de pouvoir être utilisé pour l'entraînement et l'évaluation d'étiqueteurs morpho-syntaxiques. Le format XML d'origine doit tout d'abord être converti au format d'entrée standard [2]. Cette première conversion du corpus nous a permis d'identifier et de corriger quelques problèmes liés à sa structure : absence d'attribut, étiquette de catégorie morpho-syntaxique non valide, etc.

2.1 Choix du jeu d'étiquettes et du découpage en unités lexicales

Plusieurs possibilités s'offrent à nous quant au choix du jeu d'étiquettes morpho-syntaxiques. (Crabbé et Candito, 2008) ont proposé un jeu d'étiquettes optimisé en 29 catégories (utilisant l'information supplémentaire du mode des verbes et de certaines sous-catégories). Les résultats obtenus avec leur méthode indiquent une amélioration de la précision par rapport à l'utilisation du jeu de 13 étiquettes du FTB. Cependant, les tokens présents dans les mots composés ne contiennent que l'information de la catégorie principale (attribut `catint`). Il n'est donc pas toujours possible de leur attribuer une étiquette optimisée automatiquement. La solution retenue par (Arun et Keller, 2005) et les travaux suivants consiste à fusionner les tokens et de leur affecter l'étiquette du mot composé. Par exemple, les tokens du mot composé illustré dans la Figure 2 seront fusionnés en « `levée_de_boucliers` » avec l'étiquette `NC (cat="N"+subcat="C")`. Cette méthodologie simplifie artificiellement la tâche d'étiquetage mais facilite la comparaison avec les approches précédentes.

2. Une phrase par ligne dans laquelle chaque mot est suivi d'un séparateur et de son étiquette. Par exemple, la phrase de la Figure 1 doit être converti en : `Ma/D position/N est/V la/D suivante/N ./PONCT`

```
<w cat="N" [...] lemma="levée de boucliers" mph="fs" subcat="C">
  <w catint="N">levée</w>
  <w catint="P">de</w>
  <w catint="N">boucliers</w>
</w>
```

FIGURE 2 – Exemple de mot composé extrait du fichier `lmf300_13000ep.cat.xml`.

Néanmoins, entraîner des méthodes d'étiquetage avec un ensemble de données dans lequel les mots composés sont fusionnés suppose par la suite l'utilisation d'un *tokenizer* capable de détecter les mots composés. Or, les méthodes existantes ne sont pas encore arrivées à un niveau de maturité satisfaisant. De plus, la notion de mot composé reste encore ambiguë et spécifique aux choix faits par les annotateurs du FTB. En effet, la définition du mot composé dans le FTB est assez large, avec par exemple « *à tout prix* » ou « *seconde guerre mondiale* ». Nous avons fait ici un choix restrictif en ne fusionnant que les mots composés lexicaux dont le lemme ne contient pas le caractère espace (e.g. « *aujourd'* » + « *hui* » → « *aujourd'hui* », « *celles* » + « - » + « *ci* » → « *celles-ci* ») ainsi que les nombres décimaux (e.g. « *16* » + « *,* » + « *7* » → « *16,7* ») et découpés (e.g. « *500* » + « *000* » → « *500000* »). Un total de 8 967 mots-composés sont fusionnés de cette manière.

Pour l'ensemble des raisons que nous avons évoquées précédemment, nous utilisons dans cette étude le jeu de 13 étiquettes dérivées des catégories principales du FTB. Ce choix est également appuyé par le fait que nous souhaitons proposer un ensemble de règles de corrections automatiques ne nécessitant pas de ressources externes ou d'intervention manuelle. Le corpus généré à partir de cette conversion directe du FTB contient 7 747 tokens pour lesquels aucune étiquette n'a pu être affectée, i.e. soit l'étiquette est manquante, soit l'étiquette présente n'est pas valide. Au total, le corpus généré contient 2 090 phrases dans lesquelles au moins un token sans étiquette est présent.

D'un point de vue pratique, la plupart des systèmes d'étiquetage nécessitent des données d'entraînement complètement étiquetées (i.e. sans étiquette manquante). Nous attribuons donc l'étiquette U (pour *Unknown*) aux tokens pour lesquels aucune étiquette n'a pu être affectée.

3 Correction automatique des erreurs d'annotation

La méthodologie de correction automatique des erreurs d'annotation peut être décomposée en deux étapes : i. identifier les occurrences des mots incorrectement étiquetés (ou ayant une étiquette manquante) dans le corpus ; ii. assigner les bonnes étiquettes correspondant à ces occurrences. Concernant la seconde étape, nous avons avant tout cherché à privilégier la précision des corrections. Ainsi, notre choix s'est porté sur des méthodes cherchant à assigner une étiquette corrective avec la plus grande confiance possible au détriment du nombre d'erreurs candidates corrigées.

Nous proposons deux méthodes pour corriger les erreurs : la première vise la correction des étiquettes manquantes de certains mots à l'aide de la fréquence d'occurrence des étiquettes associées à d'autres occurrences du mot (Section 3.1), que nous désignerons par FTB+corr. 1. La seconde vise la correction des erreurs d'annotation par la détection des variations d'étiquetage pour des *n*-grammes de mots (Section 3.2), que nous désignerons par FTB+corr. 2.

3.1 Correction des étiquettes manquantes

Une solution simple au problème des étiquettes manquantes consiste à attribuer l'étiquette de la forme correspondante dans le corpus. Dans le cas ou plusieurs étiquettes ont été attribuées à un même token, la plus fréquente sera choisie. Cette stratégie peut s'avérer être problématique dans le cas où les fréquences des différentes étiquettes d'un token sont égales ou très proches. Par exemple, « *quelque* » apparaît 47 fois en tant qu'adverbe, 46 fois en tant que déterminant et 34 fois en tant qu'adjectif. Le choix de l'étiquette serait dans ce cas ambigu. Pour minimiser le risque d'introduire des erreurs d'annotations, nous n'attribuons l'étiquette la plus fréquente que si sa fréquence dans le corpus est supérieure à la somme des fréquences des autres étiquettes candidates. Seules les étiquettes avec une fréquence supérieure à 1 sont utilisées.

Le nombre de tokens sans étiquette est ainsi ramené à 901, tandis que le nombre de phrases contenant au moins une étiquette manquante est réduit de 2 090 à 582. Malgré les contraintes que nous avons mises en place, il est probable que cette méthode de correction introduit des étiquettes erronées. La seconde méthode que nous décrivons dans la section suivante permet de détecter et de corriger les éventuelles séquences d'étiquettes erronées.

3.2 Détection et correction des variations d'annotation

Afin de détecter les erreurs d'annotation nous mettons en oeuvre l'approche proposée par (Dickinson et Meurers, 2003) puis reprise par (Loftsson, 2009) pour évaluer les corpus *Wall Street Journal* (WSJ) et *Icelandic Frequency Dictionary* (IFD). L'approche repose sur la détection de variations d'étiquetage pour un même *n*-gramme de mots. On utilisera le terme de *variation de n-gramme* (*variation n-gram*) pour désigner un *n*-gramme de mots dans un corpus qui contient un mot annoté différemment dans une autre occurrence du même *n*-gramme dans le corpus. Le(s) mot(s) sujet(s) à la variation (qui ont une étiquette différente dans les différentes occurrences) est(sont) appelé(s) *noyau de variation* (*variation nucleus*).

La présence au sein d'un corpus d'une variation d'annotations pour un même *n*-gramme de mots peut s'expliquer soit par l'ambiguïté des mots noyaux de la variation (une même forme peut admettre des étiquettes distinctes dans un contexte d'occurrence différent) soit par une erreur. L'hypothèse que l'on pose est : plus des contextes d'une variation sont similaires, plus grande est la probabilité qu'il s'agisse d'une erreur. La notion de contexte se traduit ici par le nombre de mots, *n*, que l'on considère dans les *n*-grammes observés. La table 1 rapporte une comparaison en chiffres des observations menées sur les différents corpus traités par cette méthode.

Comme l'explique (Loftsson, 2009), une même erreur candidate peut être détectée plusieurs fois du fait du fait qu'un *n*-gramme de mots peut se retrouver contenu dans un autre pour une valeur de *n* supérieure. De plus, une variation de *n*-gramme contient à minima deux annotations possibles pour le même *n*-gramme de mots. Il n'est ainsi pas facile de calculer la précision de cette méthode (i.e. le ratio d'erreurs correctement détectées sur toutes les erreurs candidates).

(Dickinson et Meurers, 2003; Loftsson, 2009) avaient pour objectif d'évaluer manuellement le nombre de variations distinctes correctes. Pour cette raison, ils ont choisi un *n* minimal suffisamment grand pour que le contexte soit discriminant. Ils ont par ailleurs considéré la plus longue variation de *n*-grammes pour chaque occurrence de mot présentant une variation afin d'avoir le plus de matériel sous les yeux pour permettre la levée de l'ambiguïté. Notre objectif

Corpus	WSJ	IDP	FTB+corr. 2	FTB+corr. 1&2
# tokens	1,3 M	590 297	628 767	
# étiquettes	36	700	13**	
+ longue variation	224	20	87	
Valeur de n observé	$n \geq 6$	$n \geq 5$	$n \geq 5$	
Variations distinctes	2495	752	293	147
Vraies erreurs	2436 (97,6%)	254	–	–
# tokens corrigés	4417 (0,34%)	236* (0,04%)	741	169

TABLE 1 – Comparatifs des corpus en chiffres sur lesquels des *variations de n-gramme* ont été calculées. * Nous constatons que le nombre réel de tokens corrigés, calculé à partir du pourcentage fourni par (Loftsson, 2009), est inférieur au nombre de variations étant de vraies erreurs ; nous supposons que cela est peut être dû à une erreur dans le recensement des variations distinctes observées. ** A ce nombre il faut rajouter une étiquette supplémentaire que l'on utilise pour tous les mots qui n'ont pas nativement une des 13 étiquettes retenues.

diffère puisque nous souhaitons détecter et corriger des erreurs automatiquement. Nous sommes néanmoins sensibles au fait de poser un n suffisamment grand pour discriminer mais aussi suffisamment petit pour que la différence du nombre d'occurrences entre les variations puisse être utilisée pour filtrer les variations les moins probables. Du fait que l'IDP et le FTB ont une taille proche, nous suivons le choix de (Loftsson, 2009) et optons pour $n \geq 5$.

Nous proposons une heuristique pour corriger certaines variations. Celle-ci est la suivante : nous considérons les n-grammes par taille décroissante, puis par nombre d'occurrences décroissant. Nous sélectionnons les candidats pour une correction selon deux contraintes : i. la présence d'au moins deux unités lexicales et ii. la présence d'une variation, sans étiquette manquante, dont le nombre d'occurrence est strictement supérieur à la somme des occurrences des autres variations. De fait seuls les n-grammes apparaissant au moins trois fois sont considérés. Cette dernière contrainte nous sert aussi de base pour proposer une correction. En effet, la variation qui valide la contrainte est considérée comme la séquence d'étiquettes correcte.

Les exemples 1, 2 et 3 illustrent des corrections opérées avec cette heuristique. Les mots corrigés sont soulignés.

(1) ,/PONCT l'/D une/N des/P plus/ADV → ,/PONCT l'/D une/PRO des/P plus/ADV

(2) produit/N intérieur/N brut/A (/PONCT PIB/N)/PONCT → produit/N intérieur/A brut/A (/PONCT PIB/N)/PONCT

(3) d'/P état/N chargé/N de/P la/D → d'/P état/N chargé/V de/P la/D

Pour $n \geq 5$, lorsque l'on applique cette méthode directement sur le FTB, nous comptons 293 variations distinctes (vérifiant les contraintes citées ci-dessus) et le nombre de tokens corrigé est 741 (dont 593 étaient sans étiquette). Le nombre de tokens corrigés augmente lorsque l'on diminue la taille minimale des n-grammes traités. Nous avons néanmoins préféré garder un n suffisamment haut pour maintenir une certaine confiance dans le choix de considérer certaines des variations détectées comme erreurs.

Intrinsèquement la méthode par détection de variations de n-gramme repose sur le nombre d'occurrences des n-grammes. La méthode est donc sensible à la taille du corpus et l'on peut

s'attendre à ce qu'elle fournisse de meilleurs résultats sur des corpus homogènes (i.e. d'un genre spécifique) utilisant un jeu d'étiquette à gros grain ; caractéristiques que nous retrouvons dans le FTB.

4 Résultats

L'évaluation des corrections apportées au corpus est réalisée de manière extrinsèque. L'idée derrière cette méthodologie est simple, il s'agit de comparer les scores de performance de différentes méthodes d'étiquetage morpho-syntaxique en fonction du niveau de correction du FTB. Une amélioration de la précision de l'étiquetage est une indication indirecte de la bonne correction du corpus.

Les méthodes d'étiquetage morpho-syntaxique fondées sur des modèles probabilistes discriminants atteignent des niveaux de performance très élevés. Dans cette étude, nous avons choisi deux systèmes utilisant des modèles par maximum d'entropie (*MaxEnt*) : la version 3.0.4 du *Stanford POS Tagger* (Toutanova *et al.*, 2003) et l'étiqueteur morpho-syntaxique de la suite *Apache OpenNLP* [3]. Le *Stanford POS Tagger* a été entraîné avec un ensemble standard [4] de traits bidirectionnels sur les mots et les étiquettes. L'étiqueteur d'*Apache OpenNLP* a, quant à lui, été entraîné avec l'implémentation par défaut qui caractérise chaque mot à l'aide de traits caractéristiques des trois mots précédents et suivants. Ces traits sont les préfixes et les suffixes de quatre caractères, la classe rudimentaire d'information de ces caractères (e.g. débute avec une majuscule, est un nombre, est un symbole), l'étiquette grammaticale et la forme de surface des mots. On note que les ensembles de traits que nous utilisons n'ont pas été optimisés pour le français, cette tâche sortant du cadre de notre étude. Les résultats que nous présentons ici ne correspondent donc pas à la performance maximale des systèmes. De plus, nous souhaitons préciser que nous n'entendons pas comparer ici les deux systèmes. Il faudrait utiliser les mêmes ensembles de traits pour discuter a minima de leur implémentation de l'algorithme *MaxEnt*. Les résultats sont donc donnés à titre informatif principalement parce qu'ils sont tous deux utilisés dans la communauté.

Les deux systèmes sont entraînées et évaluées à partir des différents niveaux de correction du FTB. L'ensemble de données FTB+corr. 1 correspond à la correction des étiquettes manquantes par la fréquence (Section 3.1), FTB+corr. 2 correspond à la correction des erreurs d'annotation par la méthode des variations de *n*-grammes (Section 3.2). FTB+corr. 1&2 et FTB+corr. 2&1 correspondent à l'utilisation successive des deux méthodes de correction : corr. 1 puis +corr. 2 et inversement.

Dans la littérature, les méthodes d'étiquetage morpho-syntaxique pour le français ont presque toujours été évaluées à partir d'un découpage du FTB en trois sous-ensembles : 80% pour l'entraînement, 10% pour le développement et 10% pour le test, e.g. (Denis et Sagot, 2010; Constant *et al.*, 2011). Intuitivement, une évaluation fondée uniquement sur 10% des données ne peut pas être représentative du niveau de performance réel d'une méthode. Une première série d'expériences nous a conforté dans cette idée puisque nous avons observé une variation de plus de 2% (en absolu) de la précision en fonction du découpage effectué. La construction incrémentale du FTB ainsi que la nature des documents annotés joue un rôle prépondérant dans ce phénomène. Pour palier ce problème, les résultats que nous présentons dans cette étude ont

3. `http://incubator.apache.org/opennlp/`
4. Nous avons utilisé la macro `naacl2003unknowns` décrite dans (Toutanova *et al.*, 2003).

tous été obtenus en validation croisée en 10 strates, ils ne sont donc pas directement comparables à ceux présentés dans les travaux précédents.

Trois mesures d'évaluation sont considérées comme pertinentes pour nos expériences : la précision sur les tokens, la précision sur les phrases (nombre de phrases dans lesquelles tous les tokens ont été correctement étiquetés par rapport au nombre de phrases total) et la précision sur les mots inconnus. L'écart type (σ) des scores sur les 10 strates est également calculé.

Les résultats sont présentés dans les tables 2 et 3. Les corrections apportées au corpus permettent d'améliorer les scores de précision des méthodes d'étiquetage de manière significative. Ainsi, la précision sur les tokens passe de 96,39 à 97,53 pour le *Stanford POS tagger* et de 95,70 à 97,05 pour *Apache OpenNLP*. De plus, on peut observer que l'écart type des scores calculé sur les 10 strates diminue fortement. Cette mesure est un indicateur de l'amélioration de la stabilité du niveau de performance des systèmes. Une amélioration encore plus importante est observée sur la précision au niveau des phrases, elle passe de 53,05% à 57,05% pour le *Stanford POS tagger* et de 47,56% à 51,67% pour *Apache OpenNLP*. Cette augmentation s'explique par la réduction du nombre de phrases contenant au moins un token auquel aucune étiquette n'a pu être affectée. Concernant la précision au niveau des mots inconnus, la stabilité des scores est normale puisque nous n'introduisons pas de nouveaux tokens dans les données.

Correction	**Stanford POS Tagger**		
	Prec. tokens	Prec. phrases	Prec. inconnus
FTB non corrigé	$96,39\ (\sigma = 0,96)$	$53,05\ (\sigma = 3,71)$	$83,36\ (\sigma = 3,43)$
FTB + corr. 1	$97,52^{\dagger}(\sigma = 0,26)$	$56,76^{\dagger}(\sigma = 2,15)$	$83,52^{\dagger}(\sigma = 3,40)$
FTB + corr. 2	$96,51^{\dagger}(\sigma = 0,89)$	$53,57^{\dagger}(\sigma = 3,61)$	$83,37\ (\sigma = 3,42)$
FTB + corr. 1&2	$97,53^{\dagger}(\sigma = 0,26)$	$57,05^{\dagger}(\sigma = 1,15)$	$83,50\ (\sigma = 3,38)$
FTB + corr. 2&1	$97,53^{\dagger}(\sigma = 0,27)$	$57,02^{\dagger}(\sigma = 2,25)$	$83,51\ (\sigma = 3,40)$

TABLE 2 – Scores de précision obtenus avec le Stanford POS tagger en fonction du niveau de correction du FTB. σ correspond à l'écart type des scores calculé sur les 10 strates. Les scores indiqués par les caractères † sont statistiquement significatifs par rapport au FTB non corrigé ($\rho < 0,01$ avec un t-test de Student).

Correction	**Apache OpenNLP**		
	Prec. tokens	Prec. phrases	Prec. inconnus
FTB non corrigé	$95,82\ (\sigma = 0,95)$	$47,56\ (\sigma = 3,25)$	$85,50\ (\sigma = 1,57)$
FTB + corr. 1	$97,03^{\dagger}(\sigma = 0,26)$	$51,40^{\dagger}(\sigma = 2,04)$	$85,68\ (\sigma = 1,57)$
FTB + corr. 2	$95,94^{\dagger}(\sigma = 0,88)$	$48,08^{\dagger}(\sigma = 3,21)$	$85,50\ (\sigma = 1,60)$
FTB + corr. 1&2	$97,05^{\dagger}(\sigma = 0,26)$	$51,67^{\dagger}(\sigma = 2,13)$	$85,70\ (\sigma = 1,55)$
FTB + corr. 2&1	$97,04^{\dagger}(\sigma = 0,26)$	$51,67^{\dagger}(\sigma = 2,11)$	$85,68\ (\sigma = 1,57)$

TABLE 3 – Scores de précision obtenus avec Apache OpenNLP en fonction du niveau de correction du FTB. σ correspond à l'écart type des scores calculé sur les 10 strates. Les scores indiqués par les caractères † sont statistiquement significatifs par rapport au FTB non corrigé ($\rho < 0.01$ avec un t-test de Student).

La méthode de correction par détection de variations de *n*-grammes permet de ramener davantage d'erreurs candidates lorsque l'on traite des *n*-grammes de taille inférieure à 5. De plus nous avons constaté que pour une taille strictement supérieure à 1, les résultats des étiqueteurs morpho-syntaxiques étaient améliorés. Nous n'avons pas gardé ces résultats car un premier retour au corpus nous conduisait à nous interroger sur la qualité des corrections opérées et par conséquent sur l'amélioration qui pourrait bien être due à un phénomène de lissage des annotations du corpus. Ce dernier point nécessitera une évaluation plus approfondie dans le futur.

5 Travaux connexes

Les méthodes d'étiquetage morpho-syntaxique actuelles, basées sur des modèles probabilistes, offrent un niveau de performance élevé. L'analyse des erreurs restantes suggèrent néanmoins que le gain de précision potentiel venant de meilleurs traits ou d'une méthode d'apprentissage plus performante reste très limité. Les problèmes relevés montrent que les inconsistances et les erreurs d'annotations présentes dans les données d'entraînement et de test sont en partie responsables du palier auquel les méthodes sont confrontées. Partant de ce constat, (Manning, 2011) propose un ensemble de règles manuelles visant à corriger les erreurs d'annotation présentes dans le *Penn Treebank*. Une évaluation comparative de la précision d'un système d'étiquetage morpho-syntaxique sur les données ainsi corrigées a permis de montrer l'efficacité des règles de correction proposées.

Concernant la problématique de détection et de correction automatique d'erreurs d'annotation, la majorité des travaux s'est penchée sur le premier problème avec une attention particulière sur l'étiquetage grammatical (Loftsson, 2009). Outre la question d'annotation d'unité lexicale, le projet DECCA [5] aborde aussi les problèmes de détection d'erreurs d'annotations [6] continues (concernant une séquence de mots), discontinues et de type dépendance.

Quelles que soient les approches et le type d'annotations observé, le principe de détection d'une erreur repose sur la recherche d'annotations inconsistantes au sein du corpus ; c'est-à-dire d'étiquetages différents pour des occurrences comparables du phénomène observé.

Concernant la détection d'erreur d'étiquetage grammatical, cinq approches ont été proposées. (Loftsson, 2009) compare trois méthodes de détection : la première fondée sur la détection de variation de *n*-gramme, la seconde fondée sur l'utilisation de plusieurs étiqueteurs automatiques et la troisième fondée sur de l'analyse syntaxique en constituants. La seconde méthode consiste à utiliser plusieurs étiqueteurs (l'auteur en a utilisé cinq) et à les combiner en utilisant un simple mécanisme de vote (chaque étiqueteur vote pour une étiquette et l'étiquette avec le plus grand nombre de votes est sélectionné). La troisième méthode consiste à exploiter l'étiquette syntaxique produite par l'analyse en constituants pour corriger certaines erreurs éventuelles d'étiquetage grammatical interne. Cette méthode requiert d'identifier dans un premier temps les types d'erreurs d'étiquetage grammatical possibles sous chaque constituant puis d'écrire les règles de correction correspondante. Les résultats de (Loftsson, 2009) montrent que ces méthodes permettent toutes de détecter effectivement des erreurs et qu'elles agissent en complémentarité.

(Loftsson, 2009) rapporte aussi les travaux de (Nakagawa et Matsumoto, 2002) et de (Kveton

5. `http://decca.osu.edu`
6. La nature syntaxique ou sémantique de l'annotation est discutée mais le problème est secondaire.

et Oliva, 2002). (Nakagawa et Matsumoto, 2002) ont utilisé le poids de confiance que leur algorithme de classification (machines à vecteurs de support) utilise pour décider de la classe d'un mot afin de déterminer si la classe assignée était une erreur candidate. Cette méthode est intéressante même si l'entraînement des modèles peut être coûteuse pour de larges jeux d'étiquettes.

(Kveton et Oliva, 2002) décrivent, quant à eux, une méthode semi-automatique pour détecter des *n*-grammes d'étiquettes grammaticales invalides en partant d'un ensemble construit à la main de paires d'étiquettes adjacentes invalides (e.g. un déterminant suivi d'un verbe). La méthode consiste pour chaque bigramme invalide à construire par collecte successive dans les phrases du corpus l'ensemble d'étiquettes pouvant apparaître entre deux étiquettes du bigramme invalide. Tout *n*-gramme d'étiquettes débutant et finissant par les étiquettes d'un bigramme invalide et ayant une étiquette n'appartenant pas à l'ensemble d'étiquettes avérées est considéré comme une erreur potentielle dans un nouveau corpus. Cette méthode requiert d'une part une construction manuelle (par un linguiste) de bigrammes invalides et d'autre part ne permet pas de détecter des *n*-grammes valides d'étiquettes utilisés incorrectement dans certains contextes.

6 Conclusion et perspectives

Nous avons présenté une étude menée sur la détection et la correction automatique des erreurs d'annotation morpho-syntaxique du *French TreeBank*. Deux méthodes ont été utilisées. La première consiste à identifier les mots sans étiquette et leur attribuer celle d'une forme correspondante observée dans le corpus. La seconde méthode utilise les variations de *n*-gramme pour détecter et corriger les anomalies d'annotation. Les résultats que nous avons obtenus montrent que les corrections apportées au corpus permettent d'améliorer de manière significative les scores de précision de deux différentes méthodes d'étiquetage morpho-syntaxique.

Les perspectives de cette étude sont nombreuses. Dans un premier temps, nous souhaitons poursuivre nos travaux en utilisant un jeu d'étiquettes plus étendu. Nous envisageons également d'améliorer la détection des erreurs en utilisant conjointement les variations de *n*-gramme et la combinaison des sorties de plusieurs étiqueteurs (Loftsson, 2009). A plus long terme, nous voulons étudier la possibilité d'utiliser la correction automatique des erreurs d'annotation soit comme une étape préliminaire à la vérification manuelle des annotations, soit comme une alternative. Pour ce dernier point, l'objectif serait d'étendre le FTB par l'ajout de données annotées et corrigées automatiquement.

Les modèles construits à partir des données corrigées pour les étiqueteurs *Stanford POS tagger* et *Apache OpenNLP* sont disponibles à l'adresse : `http://www.lina.univ-nantes.fr/?-TALN-.html`

Références

ABEILLÉ, A., CLÉMENT, L. et TOUSSENEL, F. (2003). Building a treebank for French. *Treebanks : building and using parsed corpora*, pages 165–188.

ARUN, A. et KELLER, F. (2005). Lexicalization in crosslinguistic probabilistic parsing : The case of French. *In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 306–313, Ann Arbor, Michigan. Association for Computational Linguistics.

CONSTANT, M., TELLIER, I., DUCHIER, D., DUPONT, Y., SIGOGNE, A. et BILLOT, S. (2011). Intégrer des connaissances linguistiques dans un CRF : application à l'apprentissage d'un segmenteur-étiqueteur du français. *In Actes de Traitement automatique des langues naturelles (2011)*.

CRABBÉ, B. et CANDITO, M. (2008). Expériences d'analyse syntaxique statistique du français. *In Actes de Traitement automatique des langues naturelles (2008)*.

DENIS, P. et SAGOT, B. (2010). Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français. *In Actes de Traitement automatique des langues naturelles (2010)*.

DICKINSON, M. et MEURERS, W. D. (2003). Detecting errors in part-of-speech annotation. *In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*, pages 107–114, Budapest, Hungary.

GREEN, S., de MARNEFFE, M.-C., BAUER, J. et MANNING, C. D. (2011). Multiword expression identification with tree substitution grammars : A parsing tour de force with french. *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 725–735, Edinburgh, Scotland, UK. Association for Computational Linguistics.

KVETON, P. et OLIVA, K. (2002). (semi-)automatic detection of errors in pos-tagged corpora. *In COLING*.

LOFTSSON, H. (2009). Correcting a POS-tagged corpus using three complementary methods. *In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 523–531, Athens, Greece. Association for Computational Linguistics.

MANNING, C. (2011). Part-of-speech tagging from 97linguistics ? *In* GELBUKH, A., éditeur : *Computational Linguistics and Intelligent Text Processing*, volume 6608 de *Lecture Notes in Computer Science*, pages 171–189. Springer Berlin / Heidelberg.

MARCUS, M., MARCINKIEWICZ, M. et SANTORINI, B. (1993). Building a large annotated corpus of English : The Penn Treebank. *Computational linguistics*, 19(2):313–330.

NAKAGAWA, T. et MATSUMOTO, Y. (2002). Detecting errors in corpora using support vector machines. *In COLING*.

TOUTANOVA, K., KLEIN, D., MANNING, C. et SINGER, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *In Proceedings of the 3rd Conference of the North American Chapter of the ACL (NAACL 2003)*, pages 173–180. Association for Computational Linguistics.

Annotation sémantique du French Treebank à l'aide de la réécriture modulaire de graphes

Bruno Guillaume[1, 2] Guy Perrier[1, 3]
(1) LORIA - Campus Scientifique - BP 239 - 54506 Vandœuvre-lès-Nancy cedex
(2) INRIA Grand Est - 615, rue du Jardin Botanique - 54600 Villers-lès-Nancy
(3) Université de Lorraine - 34, cours Léopold - CS 25233 - 54502 Nancy cedex
bruno.guillaume@loria.fr, guy.perrier@loria.fr

RÉSUMÉ

Nous proposons d'annoter le French Treebank à l'aide de dépendances sémantiques dans le cadre de la DMRS en partant d'une annotation en dépendances syntaxiques de surface et en utilisant la réécriture modulaire de graphes. L'article présente un certain nombre d'avancées concernant le calcul de réécriture utilisé : l'utilisation de règles pour faire le lien avec des lexiques, en particulier le lexique des verbes de Dicovalence, et l'introduction de filtres pour écarter à certaines étapes les annotations incohérentes. Il présente aussi des avancées dans le système de réécriture lui-même, qui a une plus large couverture (constructions causatives, verbes à montée, ...) et dont l'ordre des modules a été étudié de façon plus systématique. Ce système a été expérimenté sur l'ensemble du French Treebank à l'aide du prototype GREW, qui implémente le calcul de réécriture utilisé.

ABSTRACT

Semantic Annotation of the French Treebank using Modular Graph Rewriting

We propose to annotate the French Treebank with semantic dependencies in the framework of DMRS starting from an annotation with surface syntactic dependencies and using modular graph rewriting. The article presents some new results related to the rewriting calculus: the use of rules to make a link with lexicons, especially with the lexicon of verbs Dicovalence, and the introduction of filters to discard inconsistent annotations at some computation steps. It also presents new results related to the rewriting system itself: the system has a larger coverage (causative constructions, rising verbs, ...) and the order between modules has been studied in a more systematic way. This system has been experimented on the whole French Treebank with the prototype GREW, which implements the used rewriting calculus.

MOTS-CLÉS : réécriture de graphes, interface syntaxe-sémantique, dépendances, DMRS.

KEYWORDS: graph rewriting, syntax-semantics interface, dependencies, DMRS.

Introduction

Les résultats présentés dans cet article se situent dans la continuité de (Bonfante *et al.*, 2010, 2011; Morey, 2011). Le but de ces différents travaux est de produire une annotation sémantique de gros corpus à partir d'une annotation syntaxique en dépendances. L'annotation sémantique se situe au niveau de la phrase et elle est réalisée, comme pour la syntaxe, sous forme de dépendances.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 293–306,
Grenoble, 4 au 8 juin 2012.

L'idée est d'avoir une annotation lisible et minimale, c'est-à-dire évitant tout engagement dans des choix linguistiques trop pointus, qui susciteraient la controverse. Nous avons choisi comme cadre formel la DMRS (Dependency Minimal Recursion Semantics) (Copestake, 2009), car elle répond à ces exigences.

Pour le français, il existe très peu de corpus annotés syntaxiquement ; le plus grand actuellement est le *French Treebank*. Le French Treebank est un corpus de phrases extraites du journal « Le Monde » qui ont été annotées en constituants (Abeillé *et al.*, 2003). Ces annotations ont ensuite été converties en dépendances syntaxiques (Candito *et al.*, 2009) en suivant le format décrit dans le guide mis au point à cet effet[1]. C'est ce dernier corpus qui est utilisé ici sous l'abréviation FTB.

Même si les ambiguïtés syntaxiques sont levées par l'annotation du FTB, la production des structures sémantiques est encore ambiguë et plusieurs structures DMRS peuvent correspondre à un seule phrase du FTB.

Les structures DMRS produites font intervenir la notion d'actant sémantique mais les actants ne sont pas classés selon leur rôle. La DMRS n'a pas voulu s'engager par rapport à un choix particulier de rôles thématiques. Les actants d'un prédicat donné sont simplement distingués par un numéro et notés $arg1$, $arg2$, ... en suivant un ordre d'oblicité syntaxique fixé. Quand le prédicat est un verbe dans notre application, ces arguments font référence à un lexique externe qui est Dicovalence (Van den Eynde et Mertens, 2003). Bien que celui-ci soit avant tout un lexique syntaxique, il dispose d'informations fines et fait notamment des distinctions de lemmes en fonction de leur traduction en néerlandais et en anglais ; il peut donc être utilisé comme un lexique sémantique.

Pour le calcul, le cadre formel que nous utilisons est celui de la *réécriture de graphes* qui est motivée par la forme des objets que nous manipulons. En effet, les structures DMRS produites sont des graphes de dépendances sémantiques. Même si, en entrée, les structures sont le plus souvent des arbres de dépendances syntaxiques, pour le calcul de la sémantique, nous avons besoin de compléter ces arbres avec, par exemple, certains actants syntaxiques des infinitifs et certains antécédents des pronoms déterminés par la syntaxe. Nous obtenons alors des graphes dans toute leur généralité, avec des nœuds qui ont plusieurs antécédents et avec des cycles.

Contrairement au cas de la réécriture de termes, il n'y a pas de définition canonique de la réécriture de graphes. Nous avons choisi une définition opérationnelle (cf. Section 1) où la partie droite d'une règle est décrite par des commandes. Nous l'avons implanté dans un prototype baptisé GREW[2] qui a été utilisé pour l'expérimentation sur le FTB.

Les résultats présentés dans cet article constituent des avancées par rapport aux travaux précédents sur deux plans. D'un part, le calcul de réécriture de graphes a été enrichi par l'introduction de règles lexicales qui permettent de faire le lien avec des lexiques (Dicovalence dans notre application) et par l'introduction de filtres qui opèrent à la fin des modules pour ne conserver que les annotations cohérentes selon certains critères linguistiques ; ces nouveautés sont présentées en Section 2. D'autre part, le système de règles a été significativement étendu et l'ordonnancement des modules a été revu de façon systématique ; ces nouveautés sont présentées en Section 3. La Section 4 décrit alors l'ensemble des règles de notre système. Nous sommes ainsi en mesure de fournir des résultats expérimentaux plus riches et plus complets qui sont exposés à la Section 5.

[1] `http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html`
[2] `http://grew.loria.fr`

1 Présentation du calcul

1.1 La forme des règles de réécriture

Les règles de réécriture sont décrites en trois parties ; considérons une règle R qu'on cherche à appliquer à un graphe G ; R se compose de :

- un *patron positif* qui est un graphe qu'on va chercher à apparier avec une partie de G, l'appariement se faisant de façon injective ;
- un ensemble éventuellement vide de *patrons négatifs* qui sont des extensions[3] du patron positif et qui imposent des conditions négatives sur G ; s'il est possible d'étendre un appariement du patron positif de R avec l'un des patrons négatifs de R, la règle R ne s'applique pas ;
- une liste de *commandes* qui vont être exécutées dans l'ordre ; une commande va effectuer une transformation élémentaire sur G (ajout ou suppression d'un nœud, d'un arc ou d'un trait, fusion de deux nœuds, ...).

Si une règle s'applique à un graphe G et produit un graphe G', nous écrirons : $G \rightarrow G'$.

On peut maintenant s'interroger sur le coût de l'application d'une règle R à un graphe G. On doit procéder à l'appariement d'un graphe (un patron de la règle R) avec un sous-graphe du graphe G et on sait que ce problème, dans toute sa généralité, est NP-complet. Ici, néanmoins, cela se fait dans des conditions particulières :

- les patrons utilisés dans les règles sont très petits ;
- les patrons sont toujours connexes et ont généralement une seule racine, exe exceptionnellement deux (une racine est un nœud qui est un ancêtre de chacun des autres nœuds du graphe) ;
- le nombre d'arcs sortant d'un nœud d'un patron est borné (pour le système utilisé dans ce travail, la borne est de 3).

Dans ces conditions, pour un patron fixé, l'appariement est linéaire en temps par rapport à la taille du graphe auquel s'applique le patron. Il est important également de noter que dans le cas où le patron et le graphe sont tous les deux des arbres, l'appariement qui est utilisé suit le même algorithme que les outils standards de réécriture d'arbres.

1.2 L'organisation des règles en modules

Dans notre utilisation de la réécriture de graphes pour l'interface syntaxe-sémantique, chaque principe linguistique est traduit en quelques règles lisibles et simples. Leur intérêt est que leur effet est local mais le revers de la médaille est qu'il est difficile de contrôler l'interaction de centaines de règles pouvant agir en parallèle.

Pour faciliter la lisibilité et la maintenance de la cohérence globale, les règles sont regroupées en *modules*, chaque module comprenant un ensemble de règles qui ont une unité linguistique propre. L'ordre d'application des règles au sein d'un module est libre alors que l'ordre entre

[3] Un patron négatif respecte la même syntaxe qu'un patron positif et peut faire référence à des nœuds définis dans le patron positif.

modules peut être contraint.

Les modules jouent en plus un rôle décisif dans la mise en œuvre du calcul, sous l'angle de la terminaison et de la confluence. Un module M a la *propriété de terminaison* si pour tout graphe G, il n'existe pas de réécriture infinie $G \rightarrow G_1 \rightarrow G_2 \rightarrow \cdots$ par application de règles de M. Dans notre système, tous les modules ont la propriété de terminaison. Le problème aurait pu venir des règles créant de nouveaux nœuds dans un graphe. Or, tous les nœuds créés sont rattachés à un nœud initial et il n'est possible d'en créer qu'un nombre fini. Dans ces conditions, la quasi-totalité des modules peuvent être munis d'une mesure proportionnelle à la taille du graphe et qui décroit à chaque application de règles. Pour les autres modules, une mesure quadratique dans la taille du graphe existe.

Etant donné un module M, un graphe G est une M-forme normale si aucune règle de M ne s'applique à G. Compte tenu de la propriété de terminaison, on sait que tous les graphes se réécrivent en un nombre fini de M-formes normales. L'application d'un module M à un graphe consiste donc à calculer ses M-formes normales.

Un module M est *confluent* si tout graphe a une unique M-forme normale. Cette propriété est intéressante car elle rend le calcul indifférent à l'ordre d'application des règles. On ne peut pas espérer avoir tous les modules confluents car la transformation d'une annotation syntaxique en annotation sémantique est par essence non confluente, dans la mesure où une même annotation syntaxique peut donner lieu à plusieurs lectures sémantiques. Avec l'organisation en modules, on peut néanmoins contrôler les effets négatifs de la non confluence en la restreignant à certains modules particuliers.

En résumé, ce n'est ni l'appariement des patrons de règles, ni la longueur des chemins de réécriture qui est source de complexité des calculs. C'est avant tout la non confluence de certains modules mais ce n'est pas la méthode, la réécriture de graphes, qui est en cause. C'est l'essence même du problème auquel elle est confrontée, la production d'une représentation sémantique à partir d'une représentation syntaxique. On peut même ajouter que la source en est essentiellement l'ambiguité lexicale.

2 L'enrichissement du calcul

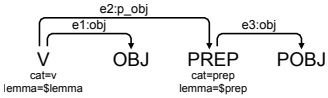
2.1 Branchement de lexiques

Dans les précédents travaux était déjà mis en évidence le fait que certaines règles font appel à des informations lexicales. C'est typiquement le cas pour les règles qui transforment les actants syntaxiques des verbes en arguments sémantiques. Elles utilisent les cadres de sous-catégorisation attachés aux différents sens des verbes. Des informations lexicales sont nécessaires pour traiter les expressions figées, les diverses classes sémantiques d'adverbes (Bonami *et al.*, 2004), d'adjectifs (Dixon et Aikhenvald, 2006; Partee, 2010) et les cadres de sous-catégorisation des noms prédicatifs(Giry-Schneider, 1987).

Dans les premières versions de notre système, les informations lexicales étaient gérées manuellement dans les règles elles-mêmes, ce qui était particulièrement lourd pour la maintenance et nuisait à la lisibilité du système. De plus, il n'était possible de paramétrer l'utilisation d'une règle que par un élément : la valeur du lemme. Les nouvelles règles lexicales permettent :

- d'utiliser plusieurs paramètres (repérés par le symbole $) dans les patrons de la règle (dans l'exemple ci-dessous, le lemme et une préposition) ;
- d'utiliser des paramètres (repérés par le symbole @) dans la partie commandes des règles (dans l'exemple, le numéro de l'entrée dicovalence).

Pour l'exemple des verbes transitifs avec un objet indirect introduit par une préposition (différente de *« à »* et *« de »*), la règle `suj_V_obj_pobj` qui transforme les actants syntaxiques en arguments sémantiques utilise le motif ci-contre et est décrite par le code suivant :




```
1 lex_rule suj_V_obj_pobj (feature $lemma, $prep, @dicoval_id; file "suj_V_obj_pobj.lp")  {
2   match{
3     V [cat=v, lemma=$lemma];
4     OBJ [];
5     e1: V -[obj]-> OBJ;
6     PREP [cat=prep, lemma=$prep];
7     e2:V -[p_obj]-> PREP;
8     POBJ [];
9     e3:PREP -[obj]-> POBJ;
0   }
1   without { V [frame=*] }
2   without { V -[a_obj|de_obj|ato|ats]-> * }
3   commands {
4     del_edge e1; del_edge e2; del_edge e3;
5     del_node PREP;
6     add_edge V -[arg2]-> OBJ; add_edge V -[arg3]-> POBJ;
7     V=@dicoval_id;
8   }
9 }
```

Cette règle fait référence au fichier `suj_V_obj_pobj.lp` qui contient la liste des 149 entrées de Dicovalence correspondant à ce cadre de sous-catégorisation. Chaque ligne décrit, dans l'ordre, le lemme, la préposition régie et le numéro de l'entrée Dicovalence. Les deux premiers éléments constituent les valeurs possibles des paramètres d'entrée de la règle alors que le dernier représente la valeur correspondante du paramètre de sortie.

```
accommoder#avec##900
accorder#avec##1090
accoupler#avec##1305
     ...
troquer#contre##84610
voir#en##86390
```

Lorsque l'on applique la règle à un graphe de dépendance, on cherche à apparier le patron positif de la règle (lignes 2 à 10) avec une partie du graphe et on vérifie que l'appariement ne peut être étendu en un appariement pour aucun des patterns négatifs. Si l'application de la règle réussit, on vérifie que le couple de valeurs dans le graphe qui coïncide avec `$lemma` et avec `$prep` correspond à l'une des lignes du fichier `suj_V_obj_pobj.lp`. Si c'est le cas, on instancie `@dicoval_id` avec la valeur correspondante de sortie du lexique. Les informations lexicales extraites de Dicovalence sont transformées automatiquement en une suite de règles lexicales (301 actuellement), accompagnées chacune d'un fichier d'instanciation des paramètres.

Lorsqu'il n'existe pas de ressources directement utilisables, des embryons de lexiques ont été construits manuellement. C'est le cas pour les cadres de sous-catégorisation des adjectifs et

des noms prédicatifs. C'est aussi le cas pour les adverbes qui ont une portée flottante, pour les adjectifs qui ne sont pas intersectifs.

2.2 Les filtres

Au sein de chaque module, on a défini précédemment la notion de graphe en forme normale. Or, parmi les formes normales, on peut distinguer celles qui sont cohérentes selon certains critères linguistiques de celles qui ne le sont pas.

Pour effectuer le tri entre les unes et les autres, nous avons introduit des *filtres*. Un filtre se présente comme une règle de réécriture standard mais sans la partie *commandes* ; il comporte seulement un patron positif et éventuellement des patrons négatifs. Exécuté en fin du module dans lequel il se situe, il ne fait que supprimer les graphes pour lesquels l'appariement entre le filtre et une partie du graphe réussit.

Les filtres peuvent être utiliser de deux façons différentes. Ils peuvent supprimer des graphes mal annotés initialement et dans lesquels l'information linguistique n'est plus cohérente. Mais ils peuvent également être utiliser pour simplifier l'écriture d'un jeu de règles : il est parfois plus simple d'écrire un ensemble de règles qui surgénèrent et d'utiliser ensuite un filtre pour ne garder que les solutions pertinentes. Par exemple, le filtre qui a été introduit dans le module qui transforme les actants syntaxiques des verbes en actants sémantiques joue ce double rôle. Il supprime les annotations dans lesquelles il subsiste des arguments syntaxiques. C'est le cas notamment des verbes intransitifs qui auraient un objet direct en syntaxe profonde.

3 L'enrichissement du système de règles

3.1 Extension de la couverture grammaticale

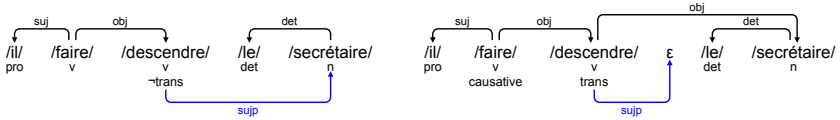
Les systèmes de règles présentés dans (Bonfante *et al.*, 2010, 2011; Morey, 2011) avaient une couverture grammaticale relativement limitée. Celle-ci a été étendue significativement. Nous en donnons ici les exemples les plus marquants.

Parmi les constructions qui ont été intégrées, on trouve les constructions causatives. Dans le FTB, un verbe causatif est traité comme auxiliaire de l'infinitif complément. Cet infinitif est considéré comme tête du noyau verbal et comme gouverneur de tous les arguments. L'annotation en dépendances suivant le guide du FTB de la phrase (1) est présentée ci-dessous.

(1) *Il fait descendre le secrétaire*

Cette approche est un peu trop simplificatrice et ne permet pas de prendre en compte certains aspects. Dans l'exemple ci-dessus, *« secrétaire »* est objet de *« descendre »* mais c'est ambigu : si par exemple le secrétaire est une personne, *« secrétaire »* peut être le sujet profond de *« descendre »* ; s'il s'agit d'un meuble, ou si *« descendre »* est

employé avec le sens de *« tuer »*, *« secrétaire »* est alors objet profond du verbe transitif *« descendre »*. Afin de distinguer les deux lectures, notre système de réécriture transforme l'auxiliaire du causatif en verbe plein, ce qui permet d'exprimer les deux lectures à travers les deux annotations ci-dessous.



Dans les deux schémas ci-dessus, a été introduite une dépendance représentant le sujet profond de *« descendre »*. Dans le premier cas, il s'agit de *« secrétaire »* et dans le second, ce sujet n'est pas exprimé dans la phrase et il est représenté par un mot vide noté ϵ. D'une façon générale, les dépendances syntaxiques de surface sont représentées au-dessus du texte qu'elles annotent, alors que les dépendances profondes, qu'elles soient syntaxiques (notées avec un 'p' terminal : **sujp**, **objp**, ...) ou sémantiques, sont placées au-dessous du texte.

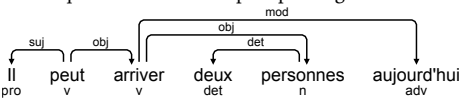
Les verbes à montée du sujet sont traités dans le FTB de la même façon que les verbes à contrôle comme des verbes pleins. Mais il y a de bonnes raisons (Rooryck, 1989) pour les traiter aussi comme des éléments hybrides qui se comportent parfois comme des auxiliaires. Considérons les exemples suivants où les verbes à montée sont indiqués en gras et où les infinitifs sont soulignés.

(2) *Il* **peut** *arriver deux personnes aujourd'hui*
(3) *La maison* **peut** *être vendue aujourd'hui*
(4) *La maison* **peut** *se vendre aujourd'hui*

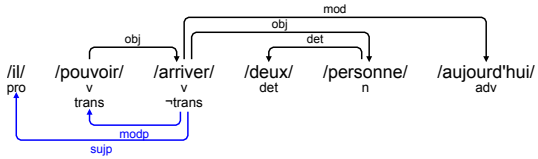
Nous avons successivement une construction impersonnelle, une construction à la voix passive et une construction à la voix moyenne. Si on veut éviter la multiplication du nombre des règles pour le calcul des arguments sémantiques, il est nécessaire de transformer chaque construction en une forme canonique. Classiquement, c'est la construction personnelle à la voix active. Pour nos trois exemples, on obtient :

(2') *Deux personnes* **peuvent** *arriver aujourd'hui.*
(3'),(4') *On* **peut** *vendre la maison aujourd'hui.*

Les règles de réécriture qui effectuent cette transformation sont grandement facilitées si les verbes à montée sont traités comme des auxiliaires. On peut alors appliquer les mêmes règles que pour les verbes non gouvernés par des verbes à montée. Détaillons la méthode sur l'exemple (2). Voici l'étiquetage de surface qui est donné au départ par le guide du FTB.

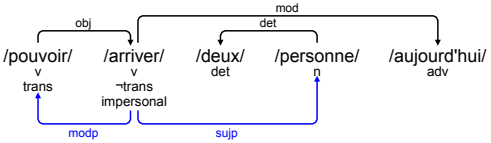


Une première règle transforme le verbe à montée *« pouvoir »* en auxiliaire de *« arriver »* qui devient la tête de la phrase.



Le fait que nous ne limitions pas la structure de dépendance à un arbre permet un plus grand pouvoir d'expression. Ainsi, le verbe *« pouvoir »* garde *« arriver »* comme objet tout en étant un modificateur de ce dernier, ce qui donne lieu à un cycle dans le graphe. Bien entendu, la tête syntaxique de la phrase n'est plus alors la racine d'un arbre de dépendance mais c'est elle qui reste destinée à recevoir toute dépendance externe, notamment quand la phrase est insérée comme proposition subordonnée dans une phrase complexe.

Maintenant, nous sommes en mesure de traiter la construction impersonnelle, qui est dorénavant ancrée à *« arriver »* comme n'importe quelle construction impersonnelle, et de lui appliquer une règle pour retrouver la construction canonique active correspondante. Cette règle consiste à supprimer le sujet impersonnel et l'objet de surface devient le sujet profond du verbe.



3.2 Ordonnancement des modules

L'organisation des règles de réécriture en modules est indissociable d'une définition d'un ordre partiel d'exécution de ces modules. Cet ordre partiel est déterminé par des choix linguistiques et des choix de représentation. Ainsi par exemple, le module de traitement des constructions impersonnelles précède celui du traitement de la voix passive, de façon à ce que soient correctement transformées les constructions passives impersonnelles.

Dans les premières version du système, cet ordre n'avait pas été étudié de façon systématique et certains choix simplificateurs posaient problème. En effet, le choix avait été fait d'enrichir en premier l'annotation du FTB en ajoutant les actants syntaxiques des infinitifs qui étaient présents dans la phrase. L'idée était d'enchaîner ensuite par une redistribution des actants pour effectuer le calcul des arguments sémantiques à partir d'une construction canonique. Or, en réalité, les choses peuvent être plus complexes. Dans chacune de ces phrases (5) et (6), on a un verbe à contrôle en gras qui gouverne le sujet d'un infinitif qui est souligné.

(5) *Jean est* **autorisé** *à* *arriver* *plus tard*

(6) *Jean* **permet** *à Marie d'* *être accompagnée* *par sa fille*

Pour la phrase (5), il est nécessaire d'appliquer en premier le module de redistribution du passif, noté PASSIVE_ARG dans notre système, pour retrouver la construction canonique (5'). Ensuite seulement, on est en mesure d'appliquer le module de détermination des sujets des infinitifs

dépendant de verbes à contrôle, noté INF_SUBJ. Dans ce module, une règle lexicale va indiquer que le sujet de *« arriver »* est l'objet direct de *« autorise »*.

(5') *On* **autorise** *Jean à arriver plus tard.*
(6') *Jean* **permet** *à Marie que sa fille l'accompagne*

Pour la phrase (6), c'est le contraire. Il faut d'abord appliquer le module INF_SUBJ pour trouver que le sujet de *« être accompagnée »* est l'objet indirect de *« permet »*. Ensuite, l'application du module PASSIVE_ARG permet de transformer la voix passive en voix active (6').

Ceci explique que l'on trouve dans l'ordre d'application des modules la séquence PASSIVE_ARG, INF_SUBJ, PASSIVE_ARG. On peut imaginer que cette suite de modules soit insuffisante et qu'il faille itérer l'aller-retour entre ces deux modules mais en pratique l'ordre ci-dessus est suffisant.

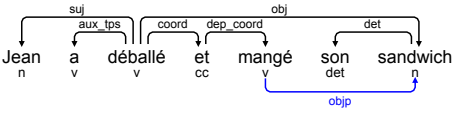
La coordination, de par sa spécificité, ne peut pas être traitée comme la plupart des constructions syntaxiques à l'aide d'un module particulier qui viendrait s'insérer quelque part dans le graphe d'ordonnancement des modules. Comme elle permet un partage de structures, elle interagit avec les autres constructions syntaxiques et cette interaction doit être modélisée par des règles qui sont distribuées dans les différents modules de calcul de la syntaxe profonde.

Bien entendu, chacune des règles est dépendante du choix qui a été fait pour annoter la coordination dans le FTB ; en particulier, le fait d'imposer que la structure de dépendance soit un arbre a de lourdes conséquences. Ainsi, on ne peut pas exprimer le partage de structures introduit fréquemment par la coordination, qui nécessite le partage de dépendants entre gouverneurs. Les règles de notre calcul vont donc consister essentiellement à retrouver le partage. Selon que la structure partagée est le sujet d'un verbe, un auxiliaire de temps ou du passif, ou encore l'antécédent d'un pronom relatif, le phénomène est traité respectivement dans le module d'introduction de sujets SUBJ_INTRO, celui de suppression des auxiliaires VERB_AUX, ou enfin dans celui de détermination de l'antécédent des pronoms relatifs ANT_REL_PRO.

Le partage de compléments est plus délicat à traiter dans la mesure où un complément n'est pas toujours obligatoire, ce qui rend l'annotation du FTB ambigüe. Considérons par exemple la phrase suivante annotée selon le guide du FTB.

(7) *Jean a déballé et mangé son sandwich*

Si on veut exprimer que *« sandwich »* est un objet partagé par *« déballé »* et *« mangé »*, on devrait ajouter la dépendance **objp** en dessous du dessin. Malheureusement ce n'est pas possible dans le FTB car le mot *« sandwich »* aurait alors deux gouverneurs. En conséquence, uniquement avec la structure de dépendance, on ne peut pas distinguer la phrase où *« son sandwich »* est un objet partagé de la phrase *« Jean a déballé son sandwich et mangé »*. Seul l'ordre linéaire des mots permettra de faire la différence. Dans notre système, nous avons choisi provisoirement de laisser ce problème de côté.



4 Description du système de règles de réécriture

Le système actuel comprend 458 règles, dont 325 règles lexicales, organisées en 35 modules. La Figure 1 présente le système de modules sous forme d'un graphe où les nœuds représentent les modules et les arcs les contraintes sur l'ordre d'application des modules. Seuls 9 des 35 modules sont non confluents ; sur le diagramme, ils se distinguent des autres parce qu'ils sont représentés par des ovales opaques.

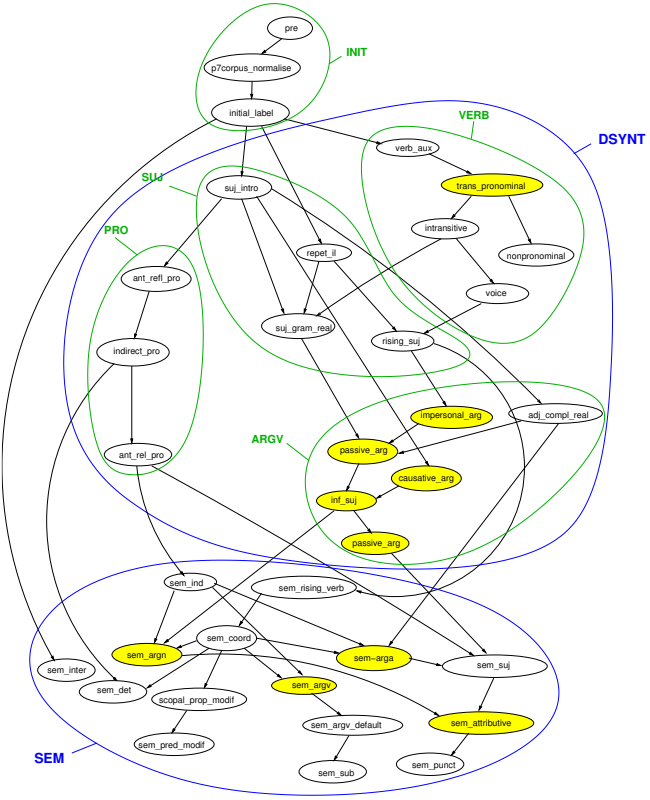


FIGURE 1 – Diagramme du système de modules

Ces modules sont eux-mêmes rangés dans 6 paquets :

- INIT qui transforme l'annotation CONLL initiale pour la mettre à un format compatible avec les règles de calcul de la syntaxe profonde,
- VERB dédié à l'uniformisation du noyau verbal par suppression des auxiliaires et à la

détermination du caractère transitif et pronominal des verbes et à la reconnaissance de leur voix (active, passive ou moyenne),
- SUBJ dédié au traitement du sujet des verbes et des adjectifs considérés comme prédicatifs,
- ARGV dédié au calcul des actants syntaxiques profonds des verbes, soit par redistribution des actants de surface (transformation des constructions impersonnelles et causatives et des voix passive et moyenne), soit par détermination lexicale de certains actants des infinitifs,
- PRO qui uniformise la syntaxe des différents pronoms et détermine l'antécédent des pronoms relatifs et des pronoms personnels réfléchis,
- SEM qui transforme l'annotation syntaxique profonde en annotation sémantique.

Les quatre paquets VERB, SUBJ, ARGV et PRO contribuent à la production de l'annotation des phrases en dépendances syntaxiques profondes. C'est pourquoi elles sont rassemblées dans un même super-paquet DSYNT. Ensuite, à partir de celle-ci, les règles du paquet SEM calculent une représentation sémantique de la phrase dans le format DMRS.

Le FTB ne suit pas complètement le guide d'annotation et les mêmes phénomènes ne sont pas annotés dans tout le corpus de façon cohérente. Notre système de modules tient compte de ces aspects et tente des récupérer certains de ces problèmes d'annotation qui sont les plus systématiques (comme par exemple des relations **p-obj** étiquetées **dep** ou des incohérences dans les coordinations de complétives).

(8) *Jean a pu être autorisé à partir et à regagner son domicile*

La Figure 2 illustre, pour la phrase (8), les deux étapes de calcul avec les trois structures :

- l'annotation initiale en dépendances syntaxiques de surface effectuée conformément au guide d'annotation du FTB ;
- l'annotation en dépendances profondes obtenue après exécution des règles des paquets INIT, VERB, SUBJ, ARGV et PRO ;
- la représentation sémantique en DMRS après exécution des règles du paquet SEM.

5 Résultats expérimentaux

Le système de modules présenté à la section précédente a été appliqué aux 12 351 phrases du FTB. Les statistiques complètes d'utilisation des règles par module sur l'ensemble du corpus et les résultats détaillés (avec les structures produites) sur 1% sur corpus sont accessible en ligne[4].

Le calcul étant non confluent, on s'intéresse en premier lieu au nombre de formes normales produites par notre système : on peut en effet avoir pour certaines longues phrases une forte ambiguïté. De plus, à la fin de certains modules, on filtre des structures que l'on considère comme incohérentes. Il peut donc arriver que certaines phrases ne conduisent à aucune forme normale : dans l'ensemble du corpus, c'est le cas pour 12,8% des phrases. Pour 31,3% des phrases, le système retourne une forme normale, pour 21,7%, il retourne 2 formes normales. 90,3% des phrases conduisent à un nombre de formes normales inférieur ou égal à 8.

Pour estimer la couverture de notre système nous avons observé le nombre de relations de dépendances présentes dans le FTB qui ne sont pas traitées lors de la réécriture. Dans le tableau

[4] http://wikilligramme.loria.fr/doku.php?id=taln_2012

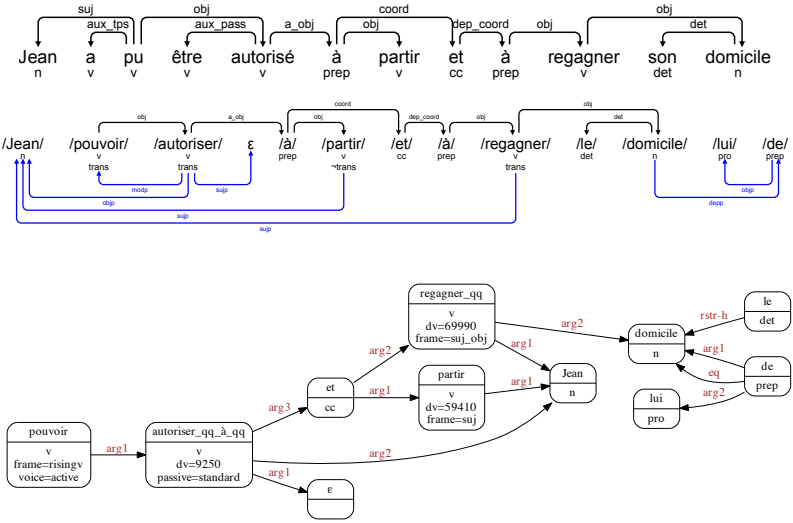


FIGURE 2 – Syntaxe de surface, syntaxe profonde et sémantique pour la phrase (8).

suivant, pour les phrases qui ont au moins une forme normale et pour chaque type de lien de dépendance du FTB, nous donnons le nombre total de liens au départ et le nombre qui restent après réécriture (en valeur absolue et relative) :

aff	**arg**	**dep**	aux_caus	aux_tps	aux_pass	comp	coord	**dep_coord**	det
1 581	465	32 821	128	3 806	1 672	37	6 358	7 299	40 355
91	375	2 331	5	27	15	37	203	1 251	608
5,8%	80,6%	6,8%	3,9%	0,7%	0,9%	100,0%	3,2%	17,1%	1,5%

mod	**mod_rel**	ponct	suj	**obj**	a_obj	de_obj	p_obj	ato	ats
58 353	2 352	35 810	15 100	58 132	2 192	1 668	1 663	160	2 521
10 452	427	1 561	162	3 222	16	23	4	15	15
17,9%	17,7%	4.3%	1,0%	5,5%	0,7 %	1,4%	0,2%	9,4%	0,6%

Il est difficile de séparer les cas qui proviennent d'une erreur d'annotation des cas qui se sont pas gérés par notre système. Néanmoins, pour les relations pour lesquelles il reste un nombre important de liens, et pour les cas où l'annotation est correcte, nous pouvons donner quelques commentaires. Les comparatifs ne sont pas traités par notre système, la relation **comp** n'est donc pas réécrite. La relation **arg** n'a pas été traitée. Les liens **mod** restant concernent les noms qui modifient un nom ou un verbe et les liens **mod_rel** peuvent correspondre à des cas d'ellipse dans la relative. Pour la coordination, il n'est pas rare d'avoir plusieurs liens **dep_coord** qui ont le même gouverneur. Le lien **dep** est sous-spécifié, il est utilisé dans des contextes assez différents et pas tous prévus par le guide d'annotation ou par notre système. Pour finir, la relation **obj** est utilisé pour les objets directs de verbe mais aussi pour le lien entre une préposition et le terme

qu'elle introduit ; c'est ces dernières relations qui peuvent rester.

Par ailleurs, pour estimer la couverture de notre usage de Dicovalence, nous donnons des statistiques sur les verbes repérés. Le FTB complet contient 39 108 verbes et les phrases pour lesquelles on fournit au moins une forme normale contiennent 32 255 verbes. Chaque fois que la réécriture réussit, on s'intéresse au nombre de verbes qui sont associés à une entrée de Dicovalence dans au moins une des structure finales. C'est le cas pour 17 987 verbes ; cela représente 55,8% des 32 255 verbes réécrits et 46,0% de l'ensemble des 39 108 verbes.

Pour 120 phrases du corpus, les résultats complets sont disponibles. Parmi elles, 11 phrases ne conduisent à aucune forme normale. Pour 5 de ces 11 phrases, Dicovalence ne décrit pas le cadre nécessaire : *hispaniser* transitif, *acheter* avec à-objet et sans objet, *maintenir* avec un complément locatif, *souscrire* transitif et *vendre* intransitif. Les autres cas correspondent à des problèmes d'annotation.

Plus précisément, nous avons évalué manuellement les résultats pour les 30 premières phrases des 120. Lorsque notre système fournit pour l'une d'elles plusieurs annotations sémantiques en sortie, nous avons considéré uniquement celle qui se rapproche le plus de l'annotation souhaitée. Pour 3 phrases, le système ne fournit aucune annotation en sortie et pour 3 autres, il fournit une annotation sans erreurs. Pour le reste, nous avons noté 67 erreurs qui se répartissent ainsi : 21 sont dues à des annotations incorrectes ou insuffisantes dans le FTB au départ ; dans 23 autres cas, le système échoue à produire une annotation sémantique parce qu'il ne couvre pas certains phénomènes (les 8 cas les plus fréquents concernent des expressions entre parenthèses insérées au milieu d'une phrase) ; pour les 23 cas restants, le système produit une annotation sémantique mais celle-ci n'est pas correcte ; les deux erreurs les plus fréquentes concernent les expressions figées (7 cas) qui ne sont pas considérées comme telles au niveau sémantique et les négations (6 cas) qui sont exprimées généralement par un couple de mots grammaticaux mais pour lesquels le système ignore le lien entre les deux éléments du couple.

Conclusion

A travers l'exemple du FTB, nous avons montré qu'il est pertinent d'utiliser la réécriture de graphes pour annoter de façon automatique un gros corpus en dépendances sémantiques à partir d'une annotation en dépendances syntaxiques de surface.

Les modules dans le système de règles de réécriture jouent un rôle majeur tant pour contrôler les calculs que pour construire et maintenir le système. Celui qui a été présenté dans cet article est conçu en fonction des formats d'entrée et de sortie choisis. Toutefois le caractère modulaire du système fait qu'il est possible de l'adapter à peu de frais à d'autres formats.

Pour en rester au FTB, l'annotation obtenue présente deux limites principales : l'annotation de départ présente de nombreuses erreurs et incohérences et le cadre formel choisi pour l'annotation sémantique, celui de la DMRS, n'offre aucune réponse quant à la modélisation de certaines propriétés sémantiques (l'intentionnalité ou les constructions comparatives par exemple). Pour ce qui est de la première limite, on peut espérer utiliser la réécriture de graphes pour corriger les erreurs qui sont les plus systématiques. Pour ce qui est de la seconde, nous ne pouvons que nous en remettre aux linguistes travaillant sur ces questions.

Références

Abeillé, A., Clément, L. et Toussenel, F. (2003). *Building a Treebank for French*, chapitre 10. Kluwer Academic Publishers.

Bonami, O., Godard, D. et Kampers-Manhe, B. (2004). Adverb classification. *In Handbook of French Semantics*, chapitre 11, pages 143–184. CLSI Publications.

Bonfante, G., Guillaume, B., Morey, M. et Perrier, G. (2010). Réécriture de graphes de dépendances pour l'interface syntaxe-sémantique. *In Actes de TALN 2010 (Traitement automatique des langues naturelles)*, Montréal. ATALA.

Bonfante, G., Guillaume, B., Morey, M. et Perrier, G. (2011). Modular graph rewriting to compute semantics. *In IWCS 2011*, pages 65–74, Oxford, UK.

Candito, M.-H., Crabbé, B., Denis, P. et Guérin, F. (2009). Analyse syntaxique statistique du français : des constituants aux dépendances. *In Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis. ATALA, LIPN.

Copestake, A. (2009). *Invited Talk : S*lacker semantics : Why superficiality, dependency and avoidance of commitment can be the right way to go. *In Proceedings of EACL 2009*, pages 1–9, Athens, Greece.

Dixon, R. W. et Aikhenvald, A. Y., éditeurs (2006). *Adjective Classes - a Cross-Linguistic Typology*. Oxford University Press.

Giry-Schneider, J. (1987). *Les prédicats nominaux en français*. Librairie DROZ Genève-Paris.

Morey, M. (2011). *Étiquetage grammatical symbolique et interface syntaxe-sémantique des formalismes grammaticaux lexicalisés polarisés*. Thèse, Université de Lorraine.

Partee, B. H. (2010). Privative adjectives : subsective plus coercion. *In* Bäuerle, R. et Zimmermann, T. E., éditeurs : *Presuppositions and Discourse : Essays Offered to Hans Kamp*, pages 273–285. Bingley, UK : Emerald Group Publishing.

Rooryck, J. (1989). Les verbes à montée et à contrôle "ambigus". *Revue québécoise de linguistique*, 18(1):189–206.

Van den Eynde, K. et Mertens, P. (2003). La valence : l'approche pronominale et son application au lexique verbal. *French Language Studies*, 13:63–104.

Enrichissement du FTB : un treebank hybride constituants/propriétés

Philippe Blache & Stéphane Rauzy
LPL, 5 Avenue Pasteur, 13100 Aix-en-Provence
{blache;rauzy}@lpl-aix.fr

RÉSUMÉ

Cet article présente les mécanismes de création d'un treebank hybride enrichissant le *FTB* à l'aide d'annotations dans le formalisme des *Grammaires de Propriétés*. Ce processus consiste à acquérir une grammaire *GP* à partir du treebank source et générer automatiquement les structures syntaxiques dans le formalisme cible en s'appuyant sur la spécification d'un schéma d'encodage adapté. Le résultat produit, en partant d'une version du *FTB* corrigée et modifiée en fonction de nos besoins, constitue une ressource ouvrant de nouvelles perspectives pour le traitement et la description du français.

ABSTRACT

Enriching the French Treebank with Properties

We present in this paper the hybridation of the French Treebank with Property Grammars annotations. This process consists in acquiring a PG grammar from the source treebank and generating the new syntactic encoding on top of the original one. The result is a new resource for French, opening the way to new tools and descriptions.

MOTS-CLÉS : Treebank hybride, French Treebank, Grammaires de Propriétés.

KEYWORDS: Hybrid treebank, French Treebank, Property Grammars.

1 Introduction

La constitution d'un treebank pour le français reste une priorité pour la description de notre langue. Il n'existe à ce jour quasiment qu'une seule ressource de ce type : le *French Treebank* (Abeillé *et al.*, 2003) à partir duquel quelques travaux ont été proposés (voir par exemple (Candito *et al.*, 2010), (Pynte *et al.*, 2001)). D'autres projets sont actuellement en cours dans différents laboratoires (voir (Cerisara *et al.*, 2010)), mais aucun ne propose à ce jour de véritable distribution. L'enrichissement du *FTB* est donc nécessaire pour disposer d'une ressource aussi complète que possible à partir de laquelle différents outils peuvent être développés. Parmi ces enrichissements, la constitution d'un treebank hybride comportant des analyses dans différents formalismes (voir par exemple (Candito *et al.*, 2010)) s'avère extrêmement utile non seulement d'un point de vue théorique (comparaison entre les différentes analyses), mais également pratique (notamment en termes d'apprentissage et d'évaluation). Par ailleurs, l'hybridation d'un treebank constitue un procédé économique pour la constitution d'une nouvelle ressource dans le formalisme cible : le fait de disposer d'une structure syntaxique (élaborée et vérifiée

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 307–320,
Grenoble, 4 au 8 juin 2012.

manuellement dans le cas du *FTB*) permet de générer automatiquement et de façon très contrôlée les structures dans le formalisme cible, héritant au passage des qualités du treebank d'origine.

Nous décrivons dans cet article l'enrichissement du *FTB* permettant la construction d'un treebank hybride *Grammaire Syntagmatique / Grammaire de Propriétés* . Dans une première partie, sans revenir sur une présentation du *FTB* disponible ailleurs, nous décrirons les corrections que nous avons effectuées (essentiellement des erreurs d'étiquetage) ainsi que les modifications apportées au format d'origine de façon à faciliter l'hybridation. Dans une seconde partie, nous proposons un schéma abstrait d'annotation pour les *Grammaires de Propriétés* (Blache, 2001) et son encodage en XML. La troisième partie sera consacrée à la présentation du procédé d'acquisition à partir du *FTB* de la grammaire dans le formalisme des Grammaires de Propriétés avant de décrire sa mise en oeuvre pour l'enrichissement du treebank. La dernière partie sera consacrée à une présentation et une discussion des résultats.

2 Le treebank FTB_{LPL}

Nous avons pour cette étude travaillé sur un sous-ensemble du *FTB*, que nous noterons désormais FTB_{LPL}(pour *LPL French Treebank*). Il a été constitué à partir du corpus *MFT* (*Modified French Treebank*, voir (Schluter et van Genabith, 2007)), lui-même sous-ensemble du corpus *FTB*.

Nous avons choisi d'apporter quelques modifications au format d'origine du *FTB* de façon à assurer une meilleure homogénéité avec les ressources existantes dans d'autres langues ou pour d'autres domaines (par exemple description de la multimodalité, études psycholinguistique, etc.). Ces modifications portent d'une part sur le niveau morphosyntaxique, pour lequel nous avons adopté le jeu de traits Multext (Ide et Véronis, 1994).

Dans un premier temps, l'étiqueteur du LPL (un étiqueteur stochastique basé sur le modèle des patrons (Blache et Rauzy, 2008; Rauzy et Blache, 2009) qui dans sa version actuelle atteint un score de F-Mesure de 0.975) a été appliqué à l'ensemble du corpus. Une fouille d'erreurs a ensuite été effectuée par correction ou validation manuelles des passages pour lesquels notre étiquetage présentait une différence avec celui proposé dans le *MFT*. D'autre part nous avons opéré une régularisation des positions des marqueurs de ponctuation, en déplaçant ces marqueurs à l'extérieur des syntagmes composant l'arbre.

Sur le plan syntaxique, nous avons systématisé certaines représentations (par exemple la projection des têtes) ou encore introduit de nouvelles catégories (plus conformes aux ressources existantes). Ces choix ne remettent pas en question la structure syntaxique d'origine, ils concernent essentiellement la forme.

Ces modifications ont été appliquées sur les 4.741 phrases du MFT (soit 134.445 mots). La moitié d'entre elles ont ensuite été validées manuellement. Environ 30% des phrases de ce sous-échantillon ont été temporairement rejetées faute d'un consensus clair dans la description de leur structure syntaxique ou si l'arbre proposé présentait des erreurs de catégorisation ou de rattachement. Le sous-ensemble du FTB_{LPL}utilisé dans cette étude compte finalement 1.471 phrases validées manuellement (soit environ 26.000 mots). La validation de la totalité du *MFT* est en cours.

A titre d'information, le schéma de la figure 1 récapitule le nombre d'occurrences des catégories

Catégories	*Modifications dans le FTB_{LPL}*
AP, PP, AdP	- Projection des têtes unaires
NP	- Projection des clitiques
	- Le N épithète fait partie du NP ("une tarte *maison*")
	- Plusieurs AP possibles dans le NP ("un *très bon premier* ministre")
Srel	- ProRel est directement rattaché à Srel (pas au NP ou PP de la relative)
VN	- Tous les participes se projettent en VN (sauf les participes passés des temps composés)
Coord :	- Projection d'un noeud COORD indiquant le type et la fonction des conjoints
VP	- Projection systématique, avec pour tête VN, incluant compléments et adjoints

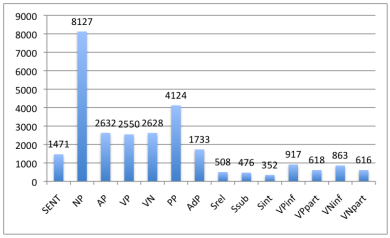


FIGURE 1 – Répartition des catégories dans le corpus

réalisées dans le FTB_{LPL}, indiquant une fréquence nettement supérieure du NP et dans une moindre mesure du PP par rapport aux autres catégories.

3 Un schéma d'annotation pour les *Grammaires de Propriétés*

L'hybridation du *FTB* consiste à ajouter aux données d'origine les annotations syntaxiques issues de l'analyse en *Grammaires de Propriétés* (voir (Blache, 2001)). Il s'agit d'une approche reposant sur la notion de construction (à la manière des *Grammaires de Construction* (voir (Kay et Fillmore, 1999))), dans laquelle les unités de base sont des objets (également appelés *signes*) comportant certaines caractéristiques décrites par des traits. En *Grammaires de Propriétés* (notées dorénavant *GP*), les signes ne sont porteurs que d'informations statiques (ou traits endogènes). Pour les signes lexicaux, il s'agit par exemple des traits morpho-syntaxiques, de la forme, sa position dans la phrase, etc. Ces informations sont récapitulées dans la structure de traits de la figure 3.

Les signes entretiennent entre eux des relations appelées *propriétés*. Les *GP* reposant sur une représentation explicite de toutes les informations syntaxiques, chaque type d'information correspond ainsi à un type de propriété qui sont, dans la grammaire élaborée pour ce treebank, au nombre de six :

– *Linéarité* : relations de précédence linéaire
– *Obligation* : identification de la tête du syntagme
– *Dépendance* : relation syntactico-sémantique entre les catégories
– *Unicité* : catégories qui dans un syntagme ne peuvent être répétées

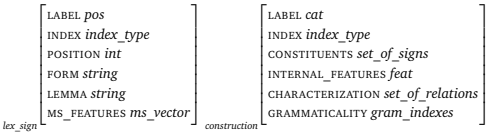


FIGURE 2 – Structures des signes lexicaux et des constructions

– *Exigence* : cooccurrence obligatoire de catégories
– *Exclusion* : impossibilité de cooccurrence de catégories

A ces types de propriétés s'ajoute la spécification de l'ensemble des constituants pouvant intervenir dans la réalisation d'un syntagme. Notons que l'ensemble de propriétés décrit ici n'est pas limitatif, d'autres types d'informations peuvent également être ajoutés, comme l'adjacence ou la représentation explicite de l'accord. Au total, un syntagme sera décrit par un couple *<constituants, propriétés>* comportant d'une part la liste de ses constituants possibles et d'autre part l'ensemble des propriétés qui forment des relations entre les constituants (ou dans le cas des propriétés unaires comme l'obligation ou l'unicité, entre le constituant et la racine).

Dans le vocabulaire des *GP*, les relations unissant les constituants d'un syntagme forment l'ensemble des propriétés évaluées, appelé *caractérisation*. Un ensemble de signes reliés par des propriétés correspond dans notre approche à la description d'une *construction*. D'une façon générale, en *GP*, toutes les unités de niveau non lexical, et notamment les unités syntaxiques, sont considérées comme des constructions. De plus, de la même façon que pour les signes lexicaux, une construction peut posséder des informations intrinsèques, décrites de façon statique par des traits spécifiques (par exemple, des informations morpho-syntaxiques ou grammaticales). Une construction correspond donc à la seconde structure de la figure 3.

Notons qu'un des traits spécifiques de la construction concerne la grammaticalité. Il s'agit d'un ensemble d'indices permettant de décrire le niveau de grammaticalité (Blache *et al.*, 2006). Ces indices s'appuient notamment sur la mesure de la densité des relations (nombre de propriétés satisfaites), leur importance (poids des propriétés) etc. à partir desquels un indice global de grammaticalité peut être proposé. L'indice de grammaticalité constitue une information intéressante, y compris dans le cas de constructions "bien formées" comme dans le cas du FTB. Cet indice permet en effet de mesurer la densité d'information syntaxique construite et constitue un élément d'évaluation de la complexité de la structure, utile notamment dans la perspective d'expériences en psycholinguistique.

La description d'une construction forme un *graphe de description* dans lequel les signes et les propriétés sont respectivement représentés par des nœuds et des arcs. Dans cette représentation, toute construction correspond à un graphe dont la racine correspond à l'identification de cette construction : une construction peut à son tour être utilisée comme constituant d'une autre construction. Les signes, correspondant aux nœuds du graphe, sont donc aussi bien des objets lexicaux que des constructions.

L'exemple de la figure 3 propose le graphe de description pour une phrase extraite du FTB_{LPL}. les nœuds du graphe représentent des signes (pour des raisons de simplicité, seuls quelques traits sont représentés), tandis que les arcs portent les différentes propriétés : précédence, dépendance,

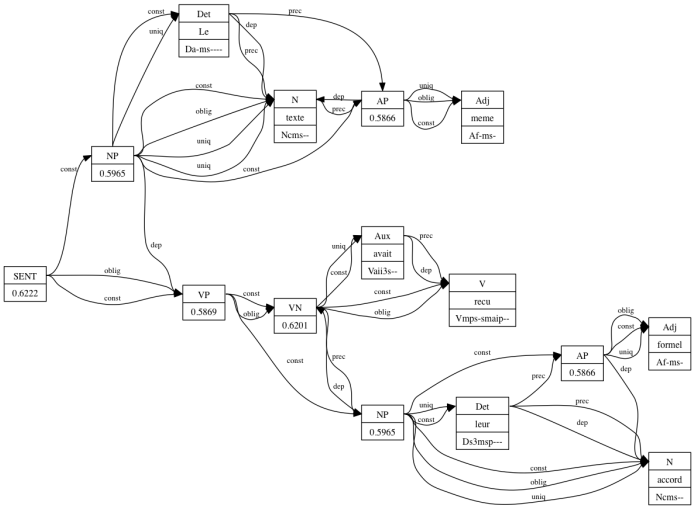


FIGURE 3 – Graphe de description de la phrase "*Le texte même avait reçu leur accord formel*"

obligation, constituance, unicité. Notons que les relations d'exclusion (qui jouent surtout un rôle en cas de violation de contrainte) ne sont pas indiquées là encore pour des raisons de lisibilité.

En *GP*, toutes les relations sont indiquées au même niveau. En isolant les sous-graphes de certaines de ces relations, il est possible d'extraire du graphe un arbre syntagmatique (sous-graphe formé des arcs étiquetés *const*) ou de dépendance (arcs étiquetés *dep*).

Dans le cadre du FTB_{LPL}, s'agissant de données écrites, les constructions concernent seulement le domaine syntaxique : chaque syntagme correspond à une construction définie par l'ensemble des nœuds participants (les constituants) et l'ensemble des relations qui les relient.

4 Schéma d'encodage XML

Le schéma abstrait défini dans la section précédente permet de définir un schéma d'encodage XML. Les signes sont tout d'abord caractérisés par leur type (signe lexical ou construction) qui induira des structures différentes pour les structures de traits qui les décrivent. Tous les types ont les attributs suivants :

< `sign` >

type	type du signe (lexical ou construction)
label	identification de la catégorie
index	index du noeud permettant sa référence
features	vecteur de traits morpho-syntaxiques, traits syntaxiques

Les signes lexicaux sont décrits par les attributs suivants :

Type *lex*

position	identification de la position dans la chaîne
form	forme du mot
lemma	lemme

De leur côté, les signes de type *construction* dans l'encodage *GP* sont décrits par un ensemble d'attributs portant les informations intrinsèques ainsi que l'indice du syntagme correspondant dans l'encodage *FTB* :

Type *const*

label	identification de la catégorie
index	index du nœud permettant sa référence
ftb_index	index du nœud correspondant dans l'encodage *FTB*

Par ailleurs, un signe de type *construction* contient un certain nombre d'éléments décrivant l'ensemble de constituants, leurs propriétés et l'évaluation de leur grammaticalité :

Type *const* : :=

< `constituents` >	ensemble des signes participant à la construction
< `characterization` >	ensemble des relations (propriétés) entre les signes
< `gram_indexes` >	ensemble des indices d'évaluation de la grammaticalité

La caractérisation d'une construction est formée quant à elle par l'ensemble des propriétés. Chaque propriété est décrite par un ensemble d'attributs décrivant le type de la relation, les nœuds sources et cibles ainsi que la satisfaction (ou violation) de la propriété :

< `characterization` >

type	type de la propriété (linéarité, dépendance, etc.)
source	index du nœud source de la relation (nœud racine en cas de relation unaire)
target	index du nœud cible
sat	satisfaction de la propriété (valeur '*plus*' ou '*moins*')

L'exemple suivant illustre l'encodage du dernier syntagme nominal de la phrase "*Le texte même avait reçu leur accord formel*" dont le graphe de description est donné plus haut. La première partie de l'exemple fournit l'encodage des annotations sous format FTB modifié comme indiqué en première section. Il s'agit d'une représentation arborescente classique, dans laquelle toutes les catégories, lexicales et syntagmatiques, sont représentées par des éléments de type `sign`.

Encodage FTB

```
<sign type="const" label="NP" features="NP_OBJ" index="f-10">
     <sign type="lex" label="Det" features="Ds3msp--" index="f-11" form="leur" lemma="leur"/>
     <sign type="lex" label="N" features="Ncms-" index="f-12" form="accord" lemma="accord"/>
     <sign type="const" label="AP" features="AP" index="f-13">
          <sign type="const" label="Adj" features="Af-ms-" index="f-14" form="formel" lemma="formel"/>
     </sign>
</sign>
```

L'encodage en *GP* suit le schéma abstrait décrit plus haut. Un graphe de description est un ensemble de constructions, chacune formant un graphe composé d'un ensemble de nœuds (les constituants) et d'un ensemble de relations indiquées dans la caractérisation et constituant le cœur de la description syntaxique. Dans cet exemple, pour des raisons de lisibilité, elles sont représentées sous forme compacte pour la représentation de la caractérisation du *NP* (les propriétés sont des attributs dont les valeurs sont des couples d'index). Par ailleurs, le treebank étant hybride, les deux encodages *GP* et *FTB* sont représentés simultanément. Deux options sont

possibles pour l'encodage XML : la première consistant à représenter dans un fichier unique (donc à l'intérieur d'un arbre XML) les informations d'origine et les propriétés. La seconde option, plus fidèle à l'approche des GP repose sur une représentation parallèle : l'arbre XML correspondant au FTB d'un côté et les informations GP de l'autre. L'identification des signes lexicaux dans la partie d'encodage *GP* se fait donc par référence à l'index des éléments correspondants décrits dans l'encodage *FTB*.

Encodage GP

```
<sign type="const" label="AP" index="g-15" ftb_index="f-13">
     <constituents>
          <constituent sign_index="f-14"/>
     </constituents>
     <characterization>
          <property type="oblig" source="f-13" target="f-14" sat="p"/>
          <property type="unic" source="f-13" target="f-14" sat="p"/>
     </characterization>
     <gram_indexes gram="0.585" sat="1.0" complete="0.117" quality="1.0" precision="0.584"/>
</sign>
<sign type="const" label="NP" index="g-16" ftb_index="f-10">
     <constituents>
          <constituent sign_index="f-11"/>
          <constituent sign_index="f-12"/>
          <constituent sign_index="g-15"/>
     </constituents>
     <characterization prec_p="11:12;11:13" oblig_p="12" dep_p="11:12;13:12"
          exig_p="12:11;13:12" unic_p="11;12"/>
     <gram_indexes gram="0.595" sat="1.0" complete="0.145" quality="1.0" precision="0.597"/>
</sign>
```

Il est important de préciser que dans l'encodage *GP*, les constructions sont toutes représentées au même niveau : les relations entre les différents signes (les arcs du graphe) sont données par les propriétés et ne suivent pas nécessairement une hiérarchie stricte. En d'autres termes, les informations linguistiques sont représentées sous forme de graphe, et non d'arbre. Ce type de représentation est indispensable dans la perspective d'intégration d'informations issues de domaines autres que la syntaxe (discours, pragmatique ou encore, pour les corpus oraux, prosodie, phonétique, etc.). Son intérêt réside en particulier dans sa capacité à décrire des phénomènes discontinus, fréquents à l'oral, mais également à l'écrit.

Concrètement, l'encodage proposé suit ainsi les recommandations du format *GrAF* (voir (Ide et Suderman, 2007), (ISO24612, 2008)) : les différents éléments du graphe sont représentés sous la forme d'éléments non hiérarchisés, leur structuration étant fournie par les relations spécifiées entre les index. Cependant, à la différence de *GrAF*, nous avons choisi de représenter les informations statiques sous la forme d'attributs (lorsqu'il s'agit de vecteurs de traits) ou d'éléments (pour les structures de traits) à l'intérieur des nœuds. Dans notre représentation, les éléments `<sign>` correspondent aux nœuds. Ils peuvent faire référence au signe correspondant dans l'encodage *FTB* (attribut *ftb_index*). Ils comportent par ailleurs la spécification de l'ensemble des nœuds qui vont former le graphe (éléments `<constituents>`). La liste des relations (éléments `<property>`) est indiquée dans l'élément `<characterization>`. Elle est complétée par un élément portant les indices de grammaticalité.

Cette représentation permet d'identifier les constructions comme des unités à part entière et qui peuvent à leur tour devenir des nœuds du graphe de description. Ce dernier correspond donc à un *hypergraphe* dans lequel chaque nœud correspond soit à un élément atomique, soit à un

graphe. Dans notre exemple, la construction *AP* (index *g-15*) est ainsi décrite par un graphe qui est utilisé comme nœud dans la construction *NP* (index *g-16*).

5 Acquisition d'une GP à partir du FTB_{LPL}

La construction de la grammaire *GP* acquise à partir du FTB_{LPL}a été faite manuellement. Un outil de navigation et d'édition du FTB_{LPL}a pour cela été développé. Sa première fonctionnalité permet de lister pour chaque syntagme toutes ses réalisations possibles. Il s'agit en d'autres termes d'identifier, dans le cadre d'une grammaire syntagmatique, toutes les parties droites de règles. Le résultat se présente sous la forme d'un tableau HTML (voir l'exemple de la figure 4) permettant d'éditer pour chaque syntagme ses réalisations dans le corpus. La colonne de droite est constituée de liens vers les positions correspondantes dans le FTB_{LPL}, permettant ainsi de visualiser les exemples comme représenté dans la figure 5.

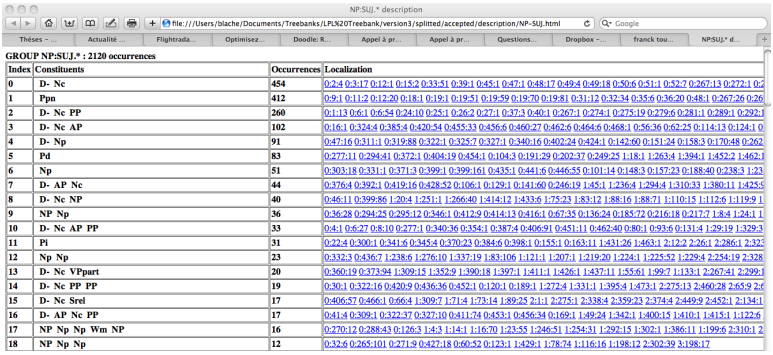


FIGURE 4 – Edition des constructions du NP sujet

Nous avons répertorié, grâce à l'éditeur ci-dessus, l'ensemble des constructions possibles de toutes les unités syntaxiques. Cette base d'information fournit directement la liste des constituants et leurs propriétés. Celles-ci sont construites de la façon suivante :

- *Linéarité* : pour chaque constituant, toutes les catégories pouvant la précéder, mais pas la suivre
- *Obligation* : liste des catégories, mutuellement exclusives, apparaissant obligatoirement dans une construction du syntagme. Il s'agit des têtes, il peut y en avoir plusieurs pour un même syntagme (mais jamais réalisées simultanément)
- *Unicité* : liste des catégories n'apparaissant qu'une fois dans chacune des constructions
- *Exigence* : pour chaque constituant, liste des catégories du syntagme apparaissant toujours avec lui
- *Exclusion* : pour chaque constituant, liste des catégories du syntagme n'apparaissant jamais avec lui

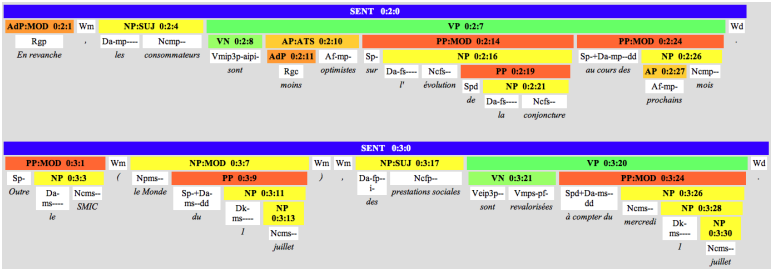


FIGURE 5 – Visualisation d'arbres

Remarquons que ces propriétés peuvent être acquises automatiquement à partir de la liste des constructions possibles du syntagme sur la base de laquelle les informations sont extraites. A ces propriétés s'ajoute la relation de *dépendance*. Celle-ci correspond, dans le cas de cette grammaire, aux relations de complémentation, adjonction et spécification. Il est toujours possible d'enrichir ou de modifier cet ensemble de relations. Le tableau suivant récapitule pour chaque catégorie le nombre des propriétés acquise à partir du *FTB* :

	const	lin	dep	unic	oblig	exig	excl
SENT	8	5	3	3	1	0	2
NP	12	36	18	10	4	6	44
AP	5	7	4	3	1	0	2
PP	7	6	5	3	1	0	0
AdP	10	18	5	4	1	0	1
VP	10	24	8	3	1	0	0
VPinf	9	13	0	7	1	0	0
VPpart	8	7	0	7	1	0	0
VN	6	11	5	2	1	0	0
VNinf	7	6	5	4	3	4	0
VNpart	7	7	6	5	2	4	0
Srel	7	7	0	5	1	0	3
Ssub	10	14	1	1	1	0	3
Sint	8	4	0	5	1	0	6

Ces propriétés sont inégalement réparties, le *NP* étant la catégorie la plus riche. On constate par ailleurs que le nombre de propriétés et leur répartition entre les types ne sont pas totalement dépendants du nombre de catégories, mais plutôt de la variété des constructions possibles. La figure 6 propose l'exemple de l'ensemble des propriétés du *NP* observées dans le FTB_{LPL}.

6 Enrichissement automatique du FTB_{LPL}

L'enrichissement du FTB_{LPL}par une description en *GP* se fait automatiquement. Nous utilisons pour cela un ensemble de solveurs de contraintes appliqué à l'entrée sous format XML du FTB_{LPL}. Le processus d'évaluation des contraintes permettant de construire l'analyse en *GP* est donc très fortement contraint : il prend en effet en entrée un syntagme déjà identifié, ainsi que la liste de ses constituants. Ceci permet donc de sélectionner dans la grammaire le sous-ensemble

const	Det, N, N_p, Pro, Clit, AdP, AP, NP, VPpart, VPinf, Srel, PP, Ssub
lin	Det $\prec$ {N, N_p, AdP, AP, VPpart, VPinf, Ssub, Srel, PP, NP} N $\prec$ {AdP, VPpart, VPinf, Ssub, Srel, PP, NP} N_p $\prec$ {AP, VPpart, Srel, PP} AP $\prec$ {VPpart, VPinf, Ssub, Srel, NP} AdP $\prec$ {Ssub, Srel, PP} NP $\prec$ {PP} Pro $\prec$ {Srel, PP} PP $\prec$ {VPpart, VPinf, Srel}
dep	{Det, AP, AdP, NP, VPpart, VPinf, PP, Srel, Ssub} $\rightsquigarrow$ N {AP, NP, VPpart, PP, Srel} $\rightsquigarrow$ N_p {PP, Srel} $\rightsquigarrow$ Pro
unic	unic = {Det, N, Pro, Clit, AdP, NP, VPpart, VPinf, Srel, Ssub}
oblig	oblig = {N, N_p, Pro, Clit}
exig	{Det, PP, AP, VPpart, VPinf, Ssub} $\Rightarrow$ N
excl	Pro $\otimes$ {Det, N, N_p, AP, AdP, NP, VPpart, VPinf, Ssub} AdP $\otimes$ {VPpart, VPinf} Ssub $\otimes$ {AP, NP, VPpart, VPinf, Srel, PP} NP $\otimes$ {Pro, VPpart, VPinf, Srel, Ssub} VPpart $\otimes$ {VPinf, Srel, Ssub} VPinf $\otimes$ {AP, Srel, Ssub} Srel $\otimes$ {Ssub} N_p $\otimes$ {N, Pro, AdP, VPinf, Ssub} Clit $\otimes$ {Det, N, Pro, AP, AdP, NP, VPpart, VPinf, Ssub}

FIGURE 6 – Propriétés du NP

de contraintes décrivant le syntagme en question là où un processus normal d'analyse en *GP* consisterait à parcourir l'ensemble des contraintes de la grammaire. Au total, le principe de construction de la description en *GP* consiste donc à parcourir le treebank initial et calculer la caractérisation de chaque syntagme du corpus. Celle-ci, comme indiqué plus haut, est formée d'une part par la liste des constituants (qui formera la liste de nœuds du graphe de description) et la liste des relations qui les relient. Cette dernière est obtenue en appliquant les solveurs de contraintes. La description de leur implantation peut être décrite de façon simplifiée par une présentation ensembliste.

On note : $|E|$ la cardinalité de l'ensemble E ; $\mathscr{C}$ la suite ordonnée des constituants du syntagme analysé, $\mathscr{C}_{i..j}$ le sous-ensemble de $\mathscr{C}$ entre les positions i et j ; c_i un constituant de $\mathscr{C}$ à la position i ; n le nombre de constituants de $\mathscr{C}$.

Par ailleurs, les propriétés dans la grammaire sont représentées soit par des ensembles (obligation, unicité) soit par des relations binaires. Dans le premier cas, on note $\mathscr{P}$ les catégories spécifiées dans l'ensemble et p_i une propriété de $\mathscr{P}$. Dans le second cas, on note *left* la catégorie de la partie gauche de la relation et *right* celle de la partie droite.

Obligation	$\lvert\mathscr{C} \cap \mathscr{P}\rvert > 0$
Linéarité	$left \in \mathscr{C}_{1..i} \Rightarrow right \in \mathscr{C}_{i+1..n}$
Unicité	$\forall c_i \in p_i, \lvert\{c_i\} \cap \mathscr{C}\rvert = 1$
Exigence	$left \in \mathscr{C} \Rightarrow right \in \mathscr{C}$
Exclusion	$left \in \mathscr{C} \Rightarrow right \notin \mathscr{C}$

Les descriptions opérationnelles ci-dessus sont décrites pour une propriété identifiée. La construction de la caractérisation totale consiste à appliquer cette évaluation à l'ensemble des propriétés décrivant le syntagme. Concrètement, nous avons développé un système couplant à un analyseur syntaxique probabiliste (entraîné sur le *FTB*) l'évaluateur *GP* développé comme indiqué

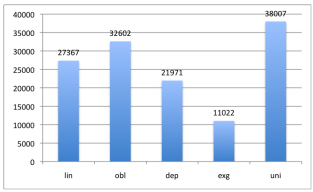


FIGURE 7 – Répartition des propriétés dans le corpus

ci-dessus. Nous pouvons donc désormais générer automatiquement un treebank hybride constituants/propriétés au format FTB_{LPL}.

7 Analyse de la répartition des propriétés

La génération des analyses en propriétés du *FTB* permet de préciser la distribution des propriétés de façon à en mesurer l'impact. L'ensemble des caractérisations permet en effet d'identifier les propriétés effectivement utilisées pour la description du corpus (on parle alors de propriétés *pertinentes*). L'étude de la distribution de ces propriétés fournit des indications précises au niveau général sur le rôle des différents types de propriétés dans la grammaire, mais également sur l'impact spécifique d'une propriété particulière dans la description d'un syntagme. Il devient donc possible d'envisager l'identification sur la base de corpus d'une répartition entre propriétés fortes et faibles, comme cela est proposé dans les approches en psycholinguistique [1].

La figure 7 indique la répartition du nombre total des propriétés pertinentes pour la description des unités du FTB_{LPL}. Nous avons isolé des indicateurs la propriété d'exclusion, celle-ci étant en effet inégalement répartie entre les syntagmes (beaucoup n'en contiennent pas). De plus, ce type de propriété consiste à vérifier l'absence de certaines catégories (restriction de cooccurrence), elle est donc presque toujours pertinente (ou évaluée), ce qui n'est pas le cas des autres propriétés. Enfin, cette propriété est directement dépendante du nombre de catégories pouvant être réalisées comme constituant d'un syntagme. D'une façon générale, les résultats montrent une fréquence importante pour les contraintes d'*unicité*, d'*obligation* et de *linéarité*. Les propriétés de *dépendance* et d'*exigence* sont quant à elles moins fréquentes dans les caractérisations. Cette information souligne l'importance de trois types d'informations non présentes explicitement dans une représentation syntagmatique classique : la présence de la tête, l'impossibilité de répétition d'un élément et l'ordre linéaire. En étudiant plus précisément la répartition de ces types de propriétés dans les syntagmes (voir figure 8), on constate que les catégories *SENT, NP* et dans une moindre mesure *PP* mobilisent l'essentiel des propriétés évaluées. Ce phénomène révèle en fait un certain degré de figement des constructions : les syntagmes ayant une grande variabilité notamment en termes de nombre de constituants et de variété des constructions ont recours à un plus grand nombre de propriétés pour leur description que les autres. Cette tendance se confirme en étudiant la répartition des propriétés évaluées pour les syntagmes montrant des description

1. Une telle répartition en deux types est plus opérationnelle pour le type d'analyse que nous proposons qu'une véritable probabilisation de l'espace des propriétés.

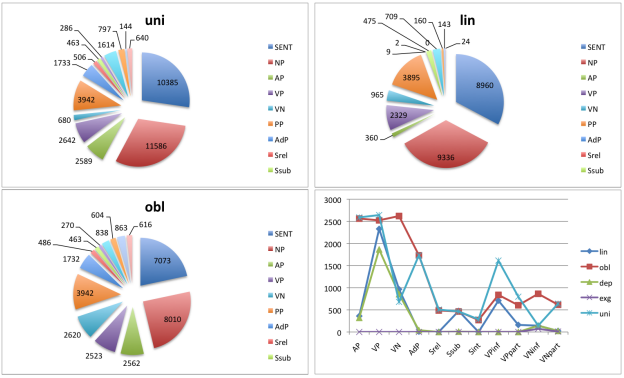


FIGURE 8 – Répartition des propriétés par type

moins denses.

L'étude des caractérisations des principaux syntagmes les plus variables (voir figure 9) fait apparaître trois grands types de répartition : *SENT* et *NP* qui présentent un certain équilibre entre les différentes propriétés ; *PP* et *VP* dont les description n'utilisent pas de propriété d'exigence (cooccurrence obligatoire de catégories) ; *AP* et *AdP* enfin qui n'utilisent quasiment que des propriétés d'unicité et d'obligation.

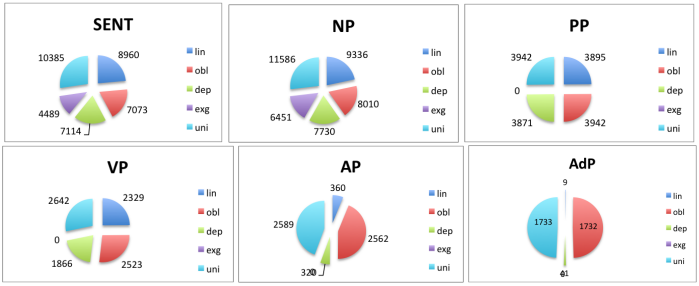


FIGURE 9 – Répartition des propriétés par catégorie

Le niveau le plus fin de description consiste à analyser pour chaque syntagme la répartition des propriétés prises individuellement. Celle-ci n'est en effet pas homogène, ce qui nous permet d'identifier avec précision quels types de propriété jouent un rôle plus important que les autres. Une estimation sur la base de la fréquence constitue un élément indicatif sur la base de laquelle une telle estimation peut être faite. Les tableaux de la figure 10 fournissent ces résultats. Ces schémas présentent en abscisses les indices des propriétés dans la grammaire et en ordonnées

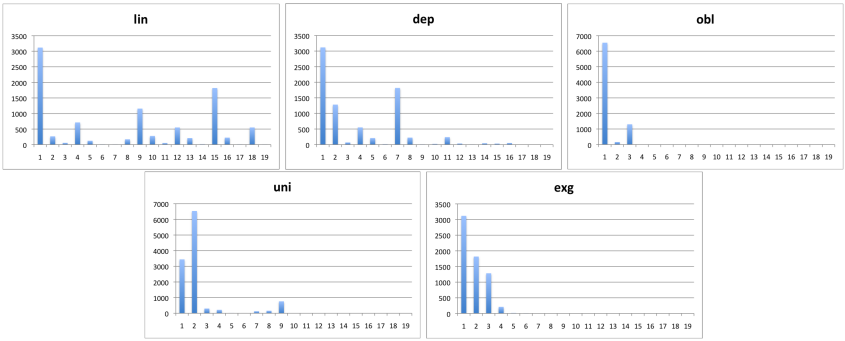


FIGURE 10 – Répartition des propriétés pour le NP

leurs occurrences.

Nous choisissons de retenir comme propriétés à poids fort celles présentant plus de 500 occurrences. Le tableau de la figure 11 récapitule les propriétés fortes du *NP*.

Ce résultat nous offre la possibilité d'acquérir automatiquement une méthode de calcul des poids des propriétés pour chacune des constructions. Cette information, intégrée dans la grammaire, est extrêmement utile tout d'abord en termes de contrôle du processus d'analyse : les propriétés fortes seront traitées comme impérativement satisfaites tandis que les faibles pourront être relâchées. Par ailleurs, du point de vue de la modélisation des processus cognitifs, il s'agit également d'une information très importante, contribuant notamment à l'évaluation de la difficulté de traitement (la violation de contraintes fortes entraînera une difficulté de traitement plus importante).

8 Conclusion

Nous avons décrit dans cet article un processus d'enrichissement du *FTB* permettant de produire un treebank hybride incluant les analyses en *Grammaires de Propriétés*. Le résultat constitue

Type	Indice	Propriété
Linéarité	1	Det $\prec$ N
	4	Det $\prec$ AP
	9	Det $\prec$ PP
	15	N $\prec$ Srel
	18	N $\prec$ NP
Exigence	1	Det $\Rightarrow$ N
	2	PP $\Rightarrow$ N
	3	AP $\Rightarrow$N

Type	Indice	Propriété
Dépendance	1	Det $\rightsquigarrow$ N
	2	AP $\rightsquigarrow$ N
	4	NP $\rightsquigarrow$ N
	7	PP $\rightsquigarrow$ N
Obligation	1	N
	3	Pro
Unicité	1	Det
	2	N

FIGURE 11 – Liste des propriétés *fortes* du NP

une nouvelle ressource pour la description du français à partir de laquelle le développement de nouveaux outils sera possible. Le processus d'enrichissement est totalement automatique, il est donc possible d'envisager la constitution de nouveaux treebanks hybrides. Par ailleurs, ce type treebank inclut une évaluation quantifiée de la grammaticalité grâce à la description en propriétés, ce qui en fait une ressource précieuse notamment pour les études en psycholinguistiques sur corpus.

Références

Abeillé, A., Clément, L. et F., T. (2003). Building a treebank for french. *In* Abeillé, A., éditeur : *Treebanks*, Kluwer, Dordrecht.

Blache, P. (2001). *Les Grammaires de Propriétés : Des contraintes pour le traitement automatique des langues naturelles*. Hermès.

Blache, P., Hemforth, B. et Rauzy, S. (2006). Acceptability prediction by means of grammaticality quantification. *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 57–64, Sydney, Australia. Association for Computational Linguistics.

Blache, P. et Rauzy, S. (2008). Influence de la qualité de l'étiquetage sur le chunking : une corrélation dépendant de la taille des chunks. *In Actes de Traitement Automatique des Langues Naturelles*, pages 290–299, Avignon, France.

Candito, M.-H., Crabbé, B. et Denis, P. (2010). Statistical french dependency parsing : treebank conversion and first results. *In LREC'2010*.

Cerisara, C., Gardent, C. et Anderson, C. (2010). Building and Exploiting a Dependency Treebank for French Radio Broadcast. *In TLT9 – the ninth international workshop on Treebanks and Linguistic Theories*, Tartu, Estonie.

Ide, N. et Suderman, K. (2007). Graf : A graph- based format for linguistic annotations. *In First Linguistic Annotation Workshop*.

Ide, N. et Véronis, J. (1994). MULTEXT : Multilingual text tools and corpora. *In Proceedings of the 15th. International Conference on Computational Linguistics (*Coling 94*)*, volume I, pages 588–592, Kyoto, Japan.

ISO24612 (2008). Language resource management - linguistic annotation framework. *In ISO/TC 37/SC 4 N522/WG 1/CD 24612*.

Kay, P. et Fillmore, C. (1999). Grammatical Constructions and Linguistic Generalizations : the *what's x doing y ?* Construction. *Language*, 75(1):1–33.

Pynte, J., Abeillé, A. et Toussenel, F. (2001). Constituent length and attachment preferences in french. *In AMLAP'2001*.

Rauzy, S. et Blache, P. (2009). Un point sur les outils du lpl pour l'analyse syntaxique du français. *In Actes du workshop ATALA 'Quels analyseurs syntaxiques pour le français ?'*, pages 1–6, Paris, France.

Schluter, N. et van Genabith, J. (2007). Preparing, restructuring and augmenting a french treebank : Lexicalized parsers or coherent treebanks ? *In Proceedings of PACLING 07*, Melbourne, Australia.

Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical

Marie Candito[1] Djamé Seddah[1, 2]
(1) Alpage (Univ. Paris Diderot & INRIA), 175 rue du Chevaleret, 75013 Paris, France
(2) Univ. Paris Sorbonne, 28, rue Serpente, 75006 Paris, France
marie.candito@linguist.jussieu.fr, djame.seddah@paris-sorbonne.fr

RÉSUMÉ

Nous présentons dans cet article la méthodologie de constitution et les caractéristiques du corpus Sequoia, un corpus en français, syntaxiquement annoté d'après un schéma d'annotation très proche de celui du French Treebank (Abeillé et Barrier, 2004), et librement disponible, en constituants et en dépendances. Le corpus comporte des phrases de quatre origines : Europarl français, le journal *l'Est Républicain*, Wikipédia Fr et des documents de l'Agence Européenne du Médicament, pour un total de 3204 phrases et 69246 tokens. En outre, nous présentons une application de ce corpus : l'évaluation d'une technique d'adaptation d'analyseurs syntaxiques probabilistes à des domaines et/ou genres autres que ceux du corpus sur lequel ces analyseurs sont entraînés. Cette technique utilise des clusters de mots obtenus d'abord par regroupement morphologique à l'aide d'un lexique, puis par regroupement non supervisé, et permet une nette amélioration de l'analyse des domaines cibles (le corpus Sequoia), tout en préservant le même niveau de performance sur le domaine source (le FTB), ce qui fournit un analyseur multi-domaines, à la différence d'autres techniques d'adaptation comme le self-training.

ABSTRACT

The Sequoia corpus : syntactic annotation and use for a parser lexical domain adaptation method

We present the building methodology and the properties of the Sequoia treebank, a freely available French corpus annotated following the French Treebank guidelines (Abeillé et Barrier, 2004). The Sequoia treebank comprises 3204 sentences (69246 tokens), from the French Europarl, the regional newspaper *l'Est Républicain*, the French Wikipedia and documents from the European Medicines Agency. We then provide a method for parser domain adaptation, that makes use of unsupervised word clusters. The method improves parsing performance on target domains (the domains of the Sequoia corpus), without degrading performance on source domain (the French treenbank test set), contrary to other domain adaptation techniques such as self-training.

MOTS-CLÉS : Corpus arboré, analyse syntaxique statistique, adaptation de domaine.

KEYWORDS: Treebank, statistical parsing, parser domain adaptation.

1 Introduction

L'analyse syntaxique statistique a fait de grands progrès ces quinze dernières années, avec de très nombreux travaux, majoritairement sur l'anglais, fondés sur un apprentissage sur les sections du

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 321–334,
Grenoble, 4 au 8 juin 2012.

Wall Street Journal du Penn Treebank (Marcus *et al.*, 1993). D'autres langues ont bénéficié de ces avancées, à la condition, de taille, que soit disponible pour ces langues un corpus arboré, en constituants ou en dépendances. Cependant, les analyseurs ainsi obtenus, appris sur un corpus bien précis, ont leur performance maximale sur des textes similaires à ce corpus, mais sont peu robustes : ils montrent une qualité nettement dégradée lorqu'ils sont évalués sur des textes de domaine ou genre différents. C'est particulièrement vrai pour l'anglais, car le WSJ montre peu de variété de thèmes : (McClosky *et al.*, 2006) rapporte que l'analyseur de Charniak (Charniak, 2000) obtient une F-mesure en constituants labelés de 89.7% sur la section de test du WSJ, mais chute à 82.9% sur le corpus de test du Brown corpus (Francis et Kucera, 1964), corpus anglais de genres variés.

Pour le français, le French Treebank (ci-après FTB) (Abeillé et Barrier, 2004) a servi de corpus d'entraînement pour des analyseurs initialement développés pour l'anglais (voir (Seddah *et al.*, 2009) pour une comparaison de plusieurs analyseurs en constituants, et (Candito *et al.*, 2010b) pour une comparaison d'analyseurs en dépendances, pour le français). Le FTB est un corpus de phrases du journal *Le Monde*, annotées en morphologie et en constituants. Les évaluations disponibles des analyseurs appris sur le FTB sont dites *intra-domaine* : elles sont classiquement faites sur une partie du FTB, non vue à l'apprentissage. Les évaluations dites *hors-domaine*, c'est-à-dire simplement sur des phrases d'origine différente de celles du corpus d'apprentissage se heurtent à l'absence de corpus annotés dans le même schéma que le FTB. Le corpus EASy (Paroubek *et al.*, 2005) comprend des phrases de domaines et genres textuels divers, mais son format mixte entre constituants (chunks) et dépendances (dépendances entre chunks) rend difficile l'évaluation des performances d'un analyseur en constituants sur ces textes.

Pour cette raison, nous avons entrepris l'annotation syntaxique de quatre corpus en suivant, à quelques exceptions près, le schéma d'annotation du FTB, regroupés sous le nom de *corpus Sequoia* [1]. Nous présentons ici la méthodologie d'annotation et les caractéristiques du corpus arboré obtenu, ainsi que l'application sur ces corpus d'une méthode d'adaptation à de nouveaux domaines d'un analyseur statistique appris sur le FTB [2]. Si l'objectif premier est de pouvoir tester et améliorer la robustesse d'analyseurs statistiques, ces corpus, librement disponibles [3], sont utilisables à d'autres fins, en particulier pour des études linguistiques.

Nous décrivons section 2 la méthodologie d'annotation et les caractéristiques du corpus, puis section 3 la méthode d'adaptation d'analyseur et les travaux antérieurs dans ce domaine, et en section 4 les expériences réalisées et les résultats obtenus. Enfin nous concluons en section 5.

2 Les corpus Sequoia

2.1 Origine et méthode de sélection

Le corpus Sequoia comporte des phrases ou textes de quatre origines différentes : l'agence européenne du médicament, Europarl, le journal régional *l'Est Républicain* et Wikipedia Fr. Le choix de ces quatre origines est en partie conjoncturel, car lié à la disponibilité des corpus : nous avons en effet eu le souci que les corpus soient librement disponibles, et qu'ils offrent une

1. Du nom du projet (SEQUOIA ANR-08-EMER-013) ayant financé l'annotation manuelle.
2. Cet article étend un article court publié à IWPT 2011 (Candito *et al.*, 2011), relatant des expériences d'adaptation d'analyseurs sur un des quatre sous-corpus aujourd'hui disponibles.
3. https ://www.rocq.inria.fr/alpage-wiki/tiki-index.php ?page=CorpusSequoia

diversité variable par rapport au genre journalistique du FTB (diversité évaluée a priori, non précisément). D'autres critères ont guidé notre choix, comme l'existence d'autres annotations pour les phrases sélectionnées et la disponibilité de gros volume de corpus brut de même origine, en vue d'expériences d'apprentissage semi-supervisé.

2.1.1 Domaine médical

Nous avons sélectionné le domaine médical comme domaine potentiellement très éloigné de celui du FTB. Plus précisément, nous avons retenu deux documents provenant de la partie en français du corpus EMEA, lui-même inclus dans le corpus OPUS (Tiedemann, 2009) [4].

Le corpus EMEA contient des documents concernant des médicaments, essentiellement des rapports public d'évaluation (EPAR), chaque rapport étant dédié à la justification de l'autorisation ou l'interdiction de la mise sur le marché d'un médicament. La partie française que nous utilisons contient environ 1000 documents convertis d'un format pdf, et concaténés. Il s'agit pour la majorité de traductions de versions originales anglaises. D'après les procédures standards de l'Agence Européenne du médicament pour les EPARs [5] les documents sont d'abord écrits en anglais, dans des termes "compréhensibles par quelqu'un qui n'est pas expert du domaine". La traduction dans les différentes langues officielles de l'Union Européenne est gérée par le Centre de Traduction de l'UE (CdT), avec une terminologie standardisée pour la biomédecine. D'après ce que nous avons pu juger, la traduction est de très bonne qualité.

Pour l'annotation manuelle, nous avons sélectionné deux EPARs, pour constituer un corpus de développement et un corpus de test (ci-après **EMEA-dev** et **EMEA-test**). Ces deux sous-corpus sont particulièrement éloignés des phrases journalistiques, pour ce qui est du domaine (ici médical) et du genre textuel (rapport scientifique). Lexicalement, ils contiennent de la terminologie spécialisée (protocoles de test et administration de médicaments, descriptions de maladies, symptômes et contre-indications). Syntaxiquement on peut noter de nombreux impératifs (pour les instructions d'utilisation), la description de dosages, et un usage fréquent de précisions apparaissant entre parenthèses (gloses de termes savants, abréviations, information de fréquence).

2.1.2 FrWiki

La deuxième source retenue est la Wikipédia en français. Nous avons pioché dans le corpus Wikipedia Fr faisant partie du corpus PASSAGE (Villemonte De La Clergerie *et al.*, 2008) [6], le texte correspondant à 19 entrées Wikipedia, concernant des "affaires" sociales ou politiques célèbres, pour la plupart récentes. Chaque entrée correspond à une description, en général chronologique, de l'affaire en question. Ainsi nous obtenons un sous-corpus d'un genre textuel narratif, pour lequel d'autres annotations existent (PASSAGE).

2.1.3 EstRépublicain

Le corpus *L'Est Républicain* est un corpus librement disponible au CNRTL [7], rassemblant les articles de deux années de ce quotidien régional (pour un total de 150 millions de tokens ponctuation

4. opus.lingfil.uu.se/EMEA.php
5. Document 3131, sur : www.ema.europa.eu
6. Les 19 premières entrées du fichier frwiki_50.txt
7. *http ://www.cnrtl.fr/corpus/estrepublicain*

comprise). Nous avons retenu 39 articles, qui sont ceux sélectionnés dans le cadre du projet ANR ANNODIS [8] dédié à l'annotation discursive pour le français, avec comme critère d'obtenir des textes intéressants du point de vue discursif.

Avec ce choix, d'une part nous espérons qu'il sera profitable de disposer pour ce corpus à la fois des annotations syntaxiques et des annotations discursives. D'autre part, nous obtenons un sous-corpus dont le domaine est éloigné du FTB. En effet les articles retenus relatent des informations essentiellement locales (faits divers, inaugurations, ...), ce qui n'est pas le cas du FTB.

2.1.4 Europarl

Enfin, nous avons sélectionné des phrases manuellement annotées dans le cadre du projet PASSAGE (Villemonte De La Clergerie *et al.*, 2008), en choisissant une sous-partie des phrases d'Europarl sélectionnées dans le cadre de ce projet.

Trois raisons principales expliquent ce choix : (i) Europarl constitue un corpus très utilisé en TAL, un corpus arboré peut en permettre une étude fine ; (ii) le type textuel d'Europarl, débat parlementaire, montre a priori des caractéristiques syntaxiques qui peuvent différer du type journalistique, ne serait-ce que par exemple le recours fréquent à la première personne et au vocatif, et enfin (iii) les phrases choisies ont également été annotées dans le schéma d'annotation Easy (pour le projet PASSAGE), ce qui peut aider à la conversion de schémas des corpus PASSAGE, Easy vers FTB et vice-versa.

2.2 Annotation morpho-syntaxique

2.2.1 Schéma d'annotation

Choix linguistiques

Notre objectif est d'obtenir des corpus compatibles avec le FTB, et donc en suivant les choix linguistiques du FTB, caractérisé comme un schéma syntagmatique surfacique, avec des annotations fonctionnelles pour les dépendants des verbes. Ainsi avons-nous suivi autant que possible les guides d'annotation du FTB (Abeillé et Clément, 2006; Abeillé *et al.*, 2004; Abeillé, 2004).

Une exception notable concerne le traitement des mots composés. Pour les composés ni nominaux ni verbaux, nous nous sommes appuyés sur les composés existants dans le FTB. Pour les composés verbaux à syntaxe régulière, nous avons préféré n'en annoter aucun, et privilégier une analyse syntagmatique. En effet ils sont potentiellement discontinus, et leur notation est alors variable dans le FTB (par exemple, annotation *il est_en_train de...* versus *il est justement en train de ...*). Concernant les composés nominaux, le FTB contient de nombreuses incohérences (séquences de même sémantique parfois codées comme composés, parfois codées par un syntagme), en particulier pour les composés syntaxiquement réguliers à sémantique tout ou partiellement compositionnelle [9]. Nous avons donc choisi de systématiquement coder syntagmatiquement des séquences syntaxiquement régulières (comme *N prep N* ou *N A* par exemple), y compris celles pouvant être considérées comme des noms composés. Cela a le mérite de l'uniformité, mais

8. http ://w3.erss.univ-tlse2.fr/annodis

9. Par exemple *pays industrialisés* apparaît deux fois comme composé, et 41 fois comme deux mots ; *taux d'intérêt apparaît 80 comme composé, et 25 fois comme trois mots.*

appelle des traitements ultérieurs pour repérer en particulier les cas de composés sémantiquement non compositionnels.

Format

En ce qui concerne le format, au lieu de reproduire le format XML du FTB, nous avons opté pour un format certes moins riche mais beaucoup plus souple : un format parenthésé avec une ligne par phrase syntagmatiquement annotée, qui fournit la catégorie morpho-syntaxique des tokens, et leur structure syntagmatique. Ce format est celui du PennTreebank, qui s'est imposé comme format d'apprentissage des analyseurs syntagmatiques probabilistes pour diverses langues et c'est sous cette forme que nous utilisons le FTB dans nos expériences d'analyse syntaxique probabiliste.

Voici un exemple dans le format en constituants parenthésé, provenant du corpus médical :

((SENT (PP-MOD (P Afin_de) (VPinf (VN (VINF diminuer)) (NP-OBJ (DET le) (NC risque) (PP (P de) (NP (ADJ faibles) (NC valeurs) (PP (P d') (NP (NC ACT)))))))) (PONCT ,) (NP-SUJ (DET le) (NC produit) (VPpart (VPP reconstitué) (COORD (CC et) (VPpart (VPP dilué))))) (VN (V doit)) (VPinf-OBJ (VN (VINF être) (ADV bien) (VPP mélangé))) (COORD (CC puis) (VN (V doit)) (VPinf (VN (VINF être) (VPP administré)) (PP-MOD (P en) (NP (NC bolus))) (PP-MOD (P par) (NP (NC poussée) (AP (ADJ intraveineuse)) (AP (ADJ rapide)))))) (PONCT .)))

Le jeu de catégories morpho-syntaxiques que nous utilisons est celui mis ou point par (Crabbé et Candito, 2008), contenant 28 catégories, qui correspondent aux combinaisons entre une des 13 catégories grossières du FTB et des informations codées dans le FTB sous forme de traits (essentiellement distinction nom commun, nom propre, mode du verbe). Il y a appauvrissement des annotations par rapport au FTB, pour ce qui est des informations morphologiques. En effet, si une partie de celles disponibles dans le FTB est encodée dans l'étiquette morpho-syntaxique, d'autres comme le lemme, le genre et le nombre ne sont pas représentés. En outre, les catégories des composants de composés n'ont pas été explicitées (un composé est directement codé comme un seul token, avec ses composants séparés par '_'). Cet appauvrissement relatif est compensé par la souplesse d'utilisation de ce format, et la disponibilité d'outils de visualisation et validation, ce qui favorise clairement la qualité des annotations, par rapport à une validation faite directement sur format XML. D'autre part, comme indiqué supra, c'est ce format parenthésé qui est utilisé pour l'analyse syntaxique probabiliste.

Conversion en dépendances

Le corpus annoté en constituants a été automatiquement converti en dépendances en utilisant le convertisseur développé pour la conversion automatique du FTB (Candito *et al.*, 2010a). Au final, le corpus Sequoia est donc disponible sous deux formes : un format parenthésé annoté en constituants [10] décoré de fonctions syntaxiques, et un format tabulé CoNLL [11] pour la version en dépendances labelées.

2.2.2 Méthodologie d'annotation

Pour obtenir le corpus Sequoia, nous avons procédé en alternant traitements automatiques et validation de ces traitements pour passer à l'étape suivante. A toutes les étapes (segmentation, tagging, parsing, annotations des fonctions), les annotations précédentes pouvaient être remises

10. Plus précisément, deux formats en constituants sont disponibles : le format standard FTB, et un format avec une représentation modifiée des infinitives introduites par des prépositions, et un syntagme supplémentaire dans les complétives, tel que décrit dans (Candito et Crabbé, 2009). Ces modifications facilitent la conversion en dépendances. La conversion de l'un vers l'autre format est automatique.

11. `http://ilk.uvt.nl/conll/#dataformat`

en cause. La séquence a été la suivante :

- Prétraitements automatiques : Segmentation en phrases, reconnaissance hors contexte de composés et tokenisation via l'outil Bonsai [12]
- Etiquetage morpho-syntaxique en utilisant le tagger MElt (Denis et Sagot, 2009)
- Validation manuelle en éditeur simple, par un seul annotateur expert, du tagging, de la segmentation en phrases, et de la reconnaissance de composés
- Pour tous les sous-corpus sauf EMEA : Analyse syntagmatique automatique au moyen de deux parsers statistiques différents, en guidant les analyseurs avec les tags manuellement validés : les analyses doivent se conformer aux catégories fournies en entrée. Les analyseurs sont le parser de Berkeley (Petrov et Klein, 2007) et l'analyseur de Charniak (Charniak, 2000), tous deux adaptés et entraînés sur le FTB. Pour EMEA : la validation syntaxique a été faite par un annotateur expert.
- Validation manuelle indépendante des deux sorties d'analyseurs, via l'outil graphique WordFreak (Morton et LaCivita, 2003) adapté pour le tagset et le jeu de fonctions du FTB, puis adjudication.
- Annotation automatique des fonctions des dépendants des verbes finis, en utilisant l'annotateur en fonctions intégré à Bonsai
- Validation manuelle des annotations fonctionnelles par deux annotateurs indépendamment, via WordFreak, puis adjudiction.
- Vérifications systématiques par un expert de points repérés comme difficiles [13] ; vérification systématique de la cohérence du traitement des composés.

2.2.3 Evaluation de l'annotation

Pour évaluer l'accord inter-annotateurs, et la distance au corpus de référence après adjudication et vérifications systématiques, nous utilisons l'outil Evalb servant habituellement à l'évaluation des sorties d'un analyseur par rapport à des analyses de référence. Pour les sous-corpus Europarl, EstRépublicain et FrWiki, nous fournissons table 1 l'accord deux à deux entre trois résultats d'annotations : l'annotation A, l'annotation B et le résultat de l'adjudication de A et B plus vérification. La mesure utilisée est la moyenne harmonique (F-mesure) entre la précision et le rappel en constituants labelés. Nous avons dû contourner le problème de tokenisations divergentes, où une séquence de tokens analysée comme un mot composé dans un des fichiers et pas dans l'autre. Par exemple la séquence *en fait* peut avoir été codée (ADV en_fait) d'un côté et (CLO en) (V fait) de l'autre. Pour résoudre ce problème, nous transformons les annotations avant l'évaluation de l'accord : tous les composés sont transformés en structure contenant les composants, avec une catégorie unique pour les composants. Pour notre exemple '(ADV en_fait)' est transformé en (ADV (Z en) (Z fait)). Ainsi les divergences de tokenisation non seulement ne bloquent pas evalb, mais sont en outre prises en compte dans l'évaluation.

L'évaluation montre des résultats assez satisfaisants pour Europarl et EstRépu, avec une nette amélioration lors de l'évaluation avec la référence. Pour FrWiki, l'accord entre les deux annotations simples est bas : il est comparable avec les résultats obtenus par l'analyseur sur le domaine neutre (section 4). C'est en effet par ce corpus que l'annotation a commencé. On voit ici que la phase de formation est longue. Sachant cela, la vérification pour ce corpus a été plus poussée.

12. http ://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html

13. Entre autres : les clivées versus relatives, le causatif, les complétives en *de* objet direct versus oblique (de-obj), le repérage d'incohérences comme par exemple des verbes finis sans sujet.

	Annotations A vs. B	Annotation A vs. référence	Annotation B vs. référence
FrWiki	83.96	91.59	88.64
Europarl	90.14	94.20	92.26
EstRépu	90.45	94.22	93.72

TABLE 1 – Evaluation deux à deux (moyenne des F-mesures) des annotations simple A, simple B et de la référence (après adjudication de A et B et vérifications systématiques).

2.3 Caractéristiques

La table 2 fournit les caractéristiques des différents corpus annotés, en regard de celle des corpus de développement et d'entraînement du FTB (FTB-dev et FTB-train) utilisés pour les expériences (section 4) [14].

	Corpus Sequoia					**FTB**	
	Médical		Neutre				
	EMEA dev	**EMEA test**	**Est Rép.**	**Euro Parl**	**Fr Wiki**	**dev**	**train**
Nb de phrases	574	544	529	561	996	1235	9881
Longueur moyenne	16,3	22,0	21,0	26,3	22,2	29,6	28,1
Ecart type sur la longueur	14,7	15,0	12,9	15,0	18,0	16,0	16,5
Données pour tout type de formes fléchies (ponctuation y compris)							
Taille du vocabulaire	1916	1737	3337	3300	4687	7222	24110
% d'inconnus	41.4	35.8	29,2	20,6	34,2	22,5	-
Nb d'occ.	9343	11964	11114	14745	22080	36508	278083
% d'occ. d'inconnus	23.0	19.7	11,2	6,6	12,9	5,2	-
% d'occ. de Noms propres	1,7	2.7	5,1	2,9	9,7	4,1	4,0
Données pour les formes alphanumériques minusculisés							
Taille du vocabulaire	1695	1599	3173	3165	4410	6904	22526
% d'inconnus	36.6	34.0	28,0	20,1	32,6	21,6	-
Nb d'occ.	8107	10451	9552	13073	18619	30940	235105
% d'occ. d'inconnus	23.2	20.9	12,1	7,0	13,8	5,7	-

TABLE 2 – Caractéristiques chiffrées des corpus manuellement annotés. Les *inconnus* sont les formes absentes du FTB-train.

Les différents nouveaux corpus ont chacun environ 500 phrases, sauf FrWiki (961 phrases). Si la longueur moyenne des phrases varie nettement entre les différents corpus, on peut noter une grande variance. Ce sont les phrases du FTB qui sont les plus longues en moyenne (29,6 pour

14. Pour comparabilité avec nos résultats antérieurs, nous utilisons la partie du FTB annotée en fonctions grammaticales, telle que distribuée en 2007, qui contient 12351 phrases. La version actuellement disponible du FTB contient environ 4000 phrases supplémentaires. Nous utilisons le découpage initialement proposé par (Crabbé et Candito, 2008) en corpus de test (1235 premières phrases), corpus de développement (1235 phrases suivantes) et 9881 phrases restantes comme corpus d'apprentissage. Le corpus original XML est prétraité tel que décrit dans (Candito et Crabbé, 2009). En particulier les composés nominaux et verbaux syntaxiquement réguliers sont défaits et représentés syntagmatiquement, et chaque occurrence de composé restante traitée comme un seul token (par exemple *(N (P à) (N cause) (P de))* est remplacé par *(N à_cause_de)*).

FTB-dev et 28,1 pour FTB-train), devant même Europarl (26,3).
La table fournit également la taille des vocabulaires (de formes fléchies), et en leur sein la proportion de formes qui sont absentes du FTB-train. Nous fournissons les chiffres calculés en utilisant tous les tokens (y compris la ponctuation) ainsi que ceux calculés sur les tokens alphanumériques minusculisés [15], pour mieux évaluer la diversité lexicale. On peut constater que le corpus médical comporte de loin le vocabulaire le plus éloigné de celui du FTB (plus d'une forme sur trois est absente du FTB-train). Pour le corpus EMEA-dev, la proportion d'inconnus en comptant tous les types de formes fléchies est très haute, du fait d'un grand nombre de mots entièrement capitalisés (la proportion passe de 41,4 à 36,6 en ignorant la ponctuation et en minusculisant). Pour le corpus FrWiki, la forte proportion d'inconnus (34,2%) peut s'expliquer par une grande fréquence des noms propres (cf. la ligne *% d'occurrences de noms propres* : environ une occurrence sur 10 est un nom propre dans FrWiki).

Les lignes sur les nombres d'occurrences et le pourcentage d'inconnus parmi ces occurrences donnent une vision plus précise de la diversité lexicale des corpus. Dans les corpus médicaux, une occurrence sur 5 (et presque une sur 4 pour EMEA-dev) correspond à un inconnu du FTB-train, ce qui, avec la faible proportion d'occurrences de noms propres (1,7 et 2,7) indique que les mots inconnus sont plutôt des mots fréquemment utilisés dans ces corpus. Au contraire, pour FrWiki on voit que, calculée sur les occurrences, la proportion d'inconnus tombe à 12,9 (la majorité des inconnus du vocabulaire sont des noms propres, apparaissant rarement). Le corpus le plus proche lexicalement du FTB semble être Europarl : seulement 6,6% des occurrences sont des inconnus, formant un cinquième du vocabulaire, ce qui constitue moins d'occurrences d'inconnus que dans le FTB-dev.

3 Adaptation de domaine par pont lexical

Notre objectif est d'explorer une méthode d'amélioration des performances d'un analyseur statistique sur des textes d'origine différente de celle du corpus d'entraînement de l'analyseur, les différences pouvant relever du domaine et/ou du genre des textes. Pour simplifier, nous utilisons par la suite les termes *domaine source* pour les caractéristiques (domaine, genre, registre) du corpus d'entraînement, *domaines cibles* pour celles des textes d'origine différente et *analyse hors-domaine* pour l'analyse de textes des domaines cibles.

Pour améliorer l'analyse hors-domaine, nous proposons d'adapter une technique testée au départ pour le parsing intra-domaine. S'inspirant de l'utilisation par (Koo *et al.*, 2008) de clusters de mots comme traits d'un analyseur discriminatif en dépendances, (Candito et Crabbé, 2009) ont proposé une technique qui, en réduisant la dispersion des données lexicales, améliore les performances de parsing intra-domaine. Ils entraînent un analyseur statistique sur un corpus où les mots sont remplacés par des identifiants de clusters de mots, obtenus de manière non supervisée sur un corpus brut de grande taille. Le parsing se fait ensuite de la même manière, en remplaçant chaque mot par leur cluster correspondant, de manière déterministe et non contextuelle, puis en réinsérant les tokens originaux pour obtenir les sorties d'analyse.

Plus précisément, le regroupement de formes fléchies en clusters se fait en deux étapes :
– Les formes fléchies sont d'abord groupées en clusters morphologiques via un lexique morphologique. Il s'agit de ramener un ensemble de formes fléchies à une forme canonique, dite forme

15. Plus précisément les tokens comportant au moins une lettre ou un chiffre, et ramenés à une forme minusculisée.

défléchie, avec comme principe de conserver exactement la même ambiguïté de catégories morpho-syntaxiques (contrairement par exemple à une lemmatisation). On veut en effet déléguer la désambiguisation de catégories à l'analyseur, et ne pas trancher par pré-traitement. Pour cela, pour une forme donnée, on récupère la liste de ses catégories recensées dans le dictionnaire, puis, tant que cette liste de catégories ne varie pas, le pluriel est ramené au singulier, le féminin au masculin, et pour les formes verbales conjuguées non ambiguës, les personnes, mode et temps verbaux sont ramenées à la deuxième personne présent pluriel (moyen rapide de trouver une forme n'introduisant pas de nouvelles ambiguïtés). Par exemple, *analysées* est ramené à *analysé*, mais *entrées* est ramené à *entrée*, de manière à conserver l'ambiguïté nom/participe. Toutes les formes finies de *augmenter* sont ramenées à *augmentez*, mais par exemple *joue* est inchangé pour préserver son ambiguïté catégorielle.
– Ensuite un algorithme de clustering non supervisé (Brown *et al.*, 1992) est appliqué sur gros corpus préalablement segmenté en phrases, tokenisé et défléchi (i.e. où les formes fléchies sont remplacées par leur forme défléchie correspondante). On obtient ainsi des clusters de formes défléchies. Il s'agit d'un algorithme hiérarchique et agglomératif, où le critère de fusion de deux clusters est la perte minimale de vraisemblance dans un modèle bigramme de séquences de clusters.

Dans cet article, nous adaptons cette technique au problème spécifique de la non robustesse des analyseurs statistiques, en utilisant des clusters de mots appris sur la concaténation de corpus du domaine source (ou proche du domaine source) et des domaines cibles. L'objectif est d'obtenir que soient groupés sous le même cluster des mots appartenant au domaine source et des mots appartenant aux domaines cibles, de façon à réaliser un pont entre les vocabulaires respectifs de ces domaines (d'où le nom d'adaptation à de nouveaux domaines par "pont lexical").

3.1 Travaux antérieurs reliés

Différentes techniques ont été proposées pour adapter des modèles d'analyse existants à de nouveaux genres :
– Adaptation au domaine via de l'auto-entraînement (*self-training*) (Bacchiani *et al.*, 2006; McClosky *et al.*, 2006; Sagae, 2010) : un analyseur entraîné sur le domaine source est utilisé pour analyser du domaine cible, et on réentraîne un analyseur sur les données validées source et les données prédites cibles. Le corpus d'entraînement ainsi obtenu, bien que bruité, capture suffisamment de régularités du domaine cible pour améliorer les performances d'analyse sur ce domaine (tout en dégradant les performances sur le domaine source) ;
– co-entraînement avec sélection d'exemples (Steedman *et al.*, 2003) : deux analyseurs sont itérativement re-entraînés sur leurs sorties respectives, les phrases du domaine cible à utiliser étant choisies de manière à minimiser les erreurs d'analyse tout en maximisant l'utilité à l'entraînement ;
– transformation de treebank et adaptation du domaine cible (Foster, 2010) ;
– adaptation méticuleuse du domaine cible à la source d'entraînement (Foster *et al.*, 2007) ;

Bien que différentes, les techniques ici évoquées sont toutes conçues pour combler la variation syntaxique et lexicale entre le domaine source et les domaines cibles. La variation lexicale est en particulier problématique dans le cas d'une langue à la morphologie plus riche que l'anglais, la flexion augmentant la dispersion des données lexicales.

4 Expériences et résultats

4.1 Clusters de mots

Pour calculer les clusters de mots nous utilisons diverses concaténations de quatre corpus, avec d'une part le corpus *L'Est Républicain* déjà cité section 2, de 150 millions de tokens, qui va jouer le rôle de corpus proche du domaine source malgré des différences manifestes concernant les sujets traités [16]. Et d'autre part, nous utilisons des tronçons de corpus de même origine que les sous-corpus Sequoia annotés : Europarl, Wikipedia Fr et domaine médical. Cela donne quatre corpus :

– **ER** : 150 millions de tokens *L'Est Républicain*
– **MED** : 12 millions de tokens du domaine médical, dont 5 millions du corpus EMEA français [17] cité section 2 et 7 millions de tokens provenant du site *doctissimo* [18].
– **EP** : la même taille, soit 12 millions de tokens, d'Europarl français,
– **FW** : et 12 millions de tokens de Wikipedia Fr

Pour le calcul des clusters, les phrases contenues dans le corpus arboré Sequoia ont été retirées.

Le corpus ER, en tant que corpus journalistique régional, est choisi comme corpus proche du FTB, malgré des différences manifestes concernant les sujets traités. La concaténation du corpus ER et du corpus MED+EP+FW va jouer le rôle de pont lexical entre le domaine source (journalistique) et les domaines cibles.

Les corpus bruts ER, MED, EP et FW sont d'abord prétraités par l'outil Bonsai (segmentés en phrases, tokenisés, et des mots composés sont reconnus hors-contexte). Puis nous appliquons le processus de défléchissement décrit section 3, pour remplacer chaque forme fléchie par sa forme défléchie équivalente. Le lexique morphologique utilisé est le L*efff* (Sagot, 2010).

Enfin nous calculons des clusters de formes défléchies [19] en utilisant l'implémentation par (Liang, 2005) de l'algorithme de (Brown *et al.*, 1992) :

– les *clusters source* sont obtenus en appliquant l'outil sur le corpus ER,
– les *clusters pont mixtes* sont obtenus sur la concaténation de ER + MED + EP + FW (soit environ 186 millions de tokens).
– les *clusters pont er-med* sont obtenus sur la concaténation de ER + MED uniquement (soit environ 162 millions de tokens), pour tester la méthode avec des clusters plus ciblés sur le vocabulaire médical.

Dans les trois cas, le nombre de clusters générés est de 1000, et les formes défléchies considérées sont celles apparaissant au moins 100 fois dans le corpus d'apprentissage [20].

16. D'après les indicateurs de la table 2, c'est plutôt Europarl qui est le plus proche en termes de vocabulaire.
17. Le corpus fait initialement environ 14 millions de tokens, mais contient énormément de formules répétitives. La suppression des phrases doublons réduit sa taille à 5 millions de tokens.
18. Il s'agit des pages médicaments et des pages du glossaire. Le texte est bien formé et proche du corpus EMEA dans les thématiques. Les phrases doublons ont été retirées.
19. Nous avons réalisé des tests en utilisant des clusters calculés sur formes fléchies (sans le processus de défléchissement), ce qui donne systématiquement des résultats moins bons qu'en utilisant les clusters sur formes défléchies.
20. Nous avons constaté lors de tests qu'un seuil plus bas, a peu d'impact sur les résultats. Un seuil de 100 réduit le vocabulaire considéré ce qui limite le temps de calcul des clusters.

4.2 Protocole et expériences

Nous réalisons ces premiers tests en analyse en constituants sans annotations fonctionnelles. Tous les traitements (entraînement d'analyseur et tests) se font donc sur des versions des corpus où les annotations fonctionnelles sont supprimées [21]. Nous utilisons l'algorithme d'apprentissage et d'analyse de PCFG avec annotations latentes (ci-après PCFG-LA) de (Petrov et Klein, 2007), et son implémentation [22], avec modèle de lissage pour les mots rares et inconnus adapté au français.

Pour cet algorithme, (Petrov, 2010) montre une variabilité des résultats selon les valeurs aléatoires choisies à l'initialisation de l'algorithme EM d'apprentissage des probabilités de règles avec annotations latentes. Aussi, nous réalisons pour chaque expérience quatre exécutions de l'apprentissage, avec quatre graines aléatoires différentes. Tous les apprentissages se font en utilisant 5 cycles de fission-fusion.

Pour l'évaluation des performances, nous utilisons l'outil Evalb, et fournissons la moyenne des F-mesures de constituants labelés (moyenne sur les quatre graines aléatoires) pour les phrases de moins de 40 mots ainsi que pour toutes les phrases.

Nous utilisons PCFG-LA pour apprendre quatre analyseurs, sur quatre versions du FTB-train (cf. section 2) différant par les symboles terminaux utilisés (les feuilles lexicales) :
– **forme fléchie** : les formes fléchies sont laissées telles quelles
– **forme défléchie** : chaque forme fléchie est remplacé par sa forme défléchie équivalente
– **cluster source** : chaque forme défléchie est ensuite remplacée par son cluster source équivalent (clusters appris sur le corpus ER)
– **cluster pont mixte** : idem mais en utilisant les clusters appris sur ER + MED + EP + FW
– **cluster pont er-med** : idem mais en utilisant les clusters appris sur ER + MED

4.3 Résultats et discussion

Nous avons réalisé des tests en comparant les résultats sur le FTB et sur le corpus Sequoia. Plus précisément, d'une part avons considéré trois "domaines" : le domaine source (FTB), un domaine très éloigné (domaine médical, corpus Emea), et un domaine que nous appelons *neutre*, regroupant les autres parties du corpus Sequoia (phrases de Wikipédia Fr, Europarl et Est Républicain). D'autre part, pour chaque domaine (source, médical et neutre) nous avons séparé corpus de test pour les tests finaux, et corpus de développement pour la phase exploratoire, de la manière suivante :
– **domaine source** : FTB-dev et FTB-train tels que décrits note 14
– **domaine médical** : EMEA-dev et EMEA-train, cf. les colonnes 2 et 3 de la table 2
– **domaine neutre** : SequoiaN-dev et SequoiaN-test obtenus en découpant en deux chacun des sous-corpus annotés FrWiki, EstRep et Europarl (colonnes 4,5 et 6 table 2). Cela donne 1043 phrases pour SequoiaN-dev et autant pour SequoiaN-test.

La table 3 fournit les résultats obtenus. Dans le cas standard, où les symboles terminaux sont simplement les formes fléchies, on observe sans surprise une nette dégradation des performances entre le domaine source (F=83.6) et le domaine médical (F=78.5). La dégradation est nettement

21. En outre, nous utilisons une instantiation des corpus où des modifications automatiques de structure ont été faites, comme décrit dans (Candito et Crabbé, 2009), ceci pour faciliter la conversion en dépendances de tous les résultats d'analyse. Les modifications introduisent des syntagmes supplémentaires pour les prépositions introduisant une infinitive et les complétives.
22. `http://code.google.com/p/berkeleyparser`

moindre pour le domaine "neutre" avec F=82.2 pour le corpus SequoiaN-test. L'apprentissage sur les phrases du journal *Le Monde* se généralise donc assez bien sur ces trois autres types de corpus (FrWiki, Europarl et Est Républicain).

Les résultats obtenus avec défléchissement (ligne 2) sont meilleurs dans toutes les configurations. On note cependant que l'incrément est moindre pour les domaines cibles que pour le domaine source (les différences restent statistiquement significatives, $p < 0.05$) [23]

Enfin, les trois dernières lignes donnent les résultats lorsque les formes sont remplacées par des clusters. La technique améliore les résultats pour le parsing du domaine source, ce qui confirme des résultats précédents. Ici nous montrons qu'elle est valable également pour les deux domaines cibles. Cela constitue donc une technique qui rend plus robuste l'analyseur, en améliorant les performances sur les domaines cibles tout en améliorant également sur le domaine source, au contraire par exemple de la technique d'auto-entraînement.

En revanche, les trois configurations qui varient selon le corpus utilisé pour le calcul des clusters offrent peu de variation dans les résultats (la plupart des différences entre ces 3 lignes ne sont pas significatives (p-value > 0.05). Ceci invalide l'hypothèse selon laquelle il serait bénéfique d'utiliser un corpus permettant de faire un pont entre le vocabulaire du domaine source et celui du domaine cible. [24]

	Toutes les phrases			Phrases de moins de 40 mots		
	Médical **EMEA-test**	Neutre **SequoiaN-test**	Source **FTB-test**	Médical **EMEA-test**	Neutre **SequoiaN-test**	Source **FTB-test**
Nombre de phrases	544	1043	1235	486	919	969
Terminaux						
formes fléchies	78.5	82.2	83.6	80.5	84.4	85.7
formes défléchies	79.0	83.1	85.0	81.0	85.0	87.4
clusters source	80.8	84.1	86.0	82.6	86.0	88.3
clusters pont mixtes	80.2	84.4	86.0	82.2	86.3	88.2
clusters pont er-med	80.7	84.1	85.9	82.8	86.1	88.2

TABLE 3 – F-mesures calculées via evalb, en ignorant la ponctuation, chacune étant moyennée sur quatre graines aléatoires différentes.

5 Conclusion

Nous avons présenté le corpus arboré Sequoia, comportant quatre sous-corpus annotés syntaxiquement en suivant le schéma du French Treebank, à quelques exceptions près. Les corpus sont librement disponibles sous forme de constituants et de dépendances.

Nous avons exploité ces corpus pour évaluer une méthode d'adaptation d'un analyseur statistique à des domaines autres que celui de son corpus d'entraînement, méthode fondée sur l'utilisation de clusters de mots, proposée dans une version préliminaire de ce travail (Candito *et al.*, 2011). Nous montrons que cette technique améliore les performances sur les domaines cibles, tout en ne dégradant pas les résultats sur le domaine source, contrairement à toutes les techniques d'adaptation de parsers statistiques à notre connaissance. En revanche, les tests réalisés en faisant

23. En utilisant l'outil `http://www.cis.upenn.edu/~dbikel/software.html#comparator`.

24. Nous contredisons ici les résultats publiés à IWPT (Candito *et al.*, 2011) où pour le corpus médical, les résultats étaient légèrement meilleurs avec les clusters pont er-med. D'une part le corpus médical a légèrement été modifié lors de la phase de vérification systématique d'erreurs d'annotation, d'autre part, il semble que cette amélioration n'était pas stable lors des tests avec différentes graines aléatoires.

varier le corpus brut sur lequel calculer les clusters ne montrent pas d'avantage clair à utiliser du texte brut du domaine cible.

Remerciements

Nous remercions chaleureusement les trois annotatrices Vanessa Combet, Catherine Moreau-Mocquay et Virginie Mouilleron pour leur travail très consciencieux. L'annotation a été financée par l'ANR (projet SEQUOIA ANR-08-EMER-013).

Références

ABEILLÉ, A. (2004). Annotation fonctionnelle, version du 1er mars 2004. `http://www.llf.cnrs.fr/Gens/Abeille`.

ABEILLÉ, A. et BARRIER, N. (2004). Enriching a french treebank. *In Proc. of LREC'04*, Lisbon, Portugal.

ABEILLÉ, A. et CLÉMENT, L. (2006). Annotation morpho-syntaxique, version du 10 nov. 2006. `http://www.llf.cnrs.fr/Gens/Abeille`.

ABEILLÉ, A., TOUSSENEL, F. et CHÉRADAME, M. (2004). Corpus le monde, annotation en constituants, guide pour les correcteurs, version du 31 mars 2004. `http://www.llf.cnrs.fr/Gens/Abeille`.

BACCHIANI, M., RILEY, M., ROARK, B. et SPROAT, R. (2006). Map adaptation of stochastic grammars. *Computer speech & language*, 20(1):41–68.

BROWN, P. F., DELLA, V. J., DESOUZA, P. V., LAI, J. C. et MERCER, R. L. (1992). Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

CANDITO, M. et CRABBÉ, B. (2009). Improving generative statistical parsing with semi-supervised word clustering. *In Proceedings of IWPT 2009*, pages 138–141, Paris, France.

CANDITO, M., CRABBÉ, B. et DENIS, P. (2010a). Statistical french dependency parsing : Treebank conversion and first results. *In Proceedings of LREC'2010*, Valletta, Malta.

CANDITO, M., HENESTROZA ANGUIANO, E. et SEDDAH, D. (2011). A word clustering approach to domain adaptation : Effective parsing of biomedical texts. *In Proceedings of IWPT 2011*, pages 37–42, Dublin, Ireland.

CANDITO, M., NIVRE, J., DENIS, P. et ANGUIANO, E. H. (2010b). Benchmarking of statistical dependency parsers for french. *In Proceedings of COLING 2010*, Beijing, China.

CHARNIAK, E. (2000). A maximum entropy inspired parser. *In Proceedings of NAACL 2000*, pages 132–139, Seattle, WA.

CRABBÉ, B. et CANDITO, M. (2008). Expériences d'analyse syntaxique statistique du français. *In Actes de TALN 2008*, pages 45–54, Avignon, France.

DENIS, P. et SAGOT, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. *In Proc. of PACLIC*, Hong Kong, China.

FOSTER, J. (2010). "cba to check the spelling" : Investigating parser performance on discussion forum posts. *In Proceedings of HLT-NAACL 2010*, pages 381–384, Los Angeles, California.

FOSTER, J., WAGNER, J., SEDDAH, D. et VAN GENABITH, J. (2007). Adapting wsj-trained parsers to the british national corpus using in-domain self-training. *In Proceedings of the Tenth IWPT*, pages 33–35.

FRANCIS, W. N. et KUCERA, H. (1964). *Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Brown University, Providence, Rhode Island.

KOO, T., CARRERAS, X. et COLLINS, M. (2008). Simple semi-supervised dependency parsing. *In Proceedings of ACL-08*, pages 595–603, Columbus, USA.

LIANG, P. (2005). Semi-supervised learning for natural language. *In MIT Master's thesis*, Cambridge, USA.

MARCUS, M., MARCINKIEWICZ, M. et SANTORINI, B. (1993). Building a large annotated corpus of english : The penn treebank. *Computational linguistics*, 19(2):313–330.

MCCLOSKY, D., CHARNIAK, E. et JOHNSON, M. (2006). Reranking and self-training for parser adaptation. *In Proceedings of COLING-ACL 2006*, pages 337–344, Sydney, Australia.

MORTON, T. et LACIVITA, J. (2003). Wordfreak : an open tool for linguistic annotation. *In Proceedings of NAACL 2003, Demonstrations*, pages 17–18.

PAROUBEK, P., POUILLOT, L.-G., ROBBA, I. et VILNAT, A. (2005). Easy : Campagne d'évaluation des analyseurs syntaxiques. *In Proceedings of TALN'05, EASy workshop : campagne d'évaluation des analyseurs syntaxiques*, Dourdan.

PETROV, S. (2010). Products of random latent variable grammars. *In Proceedings of HLT-NAACL 2010*, pages 19–27, Los Angeles, California.

PETROV, S. et KLEIN, D. (2007). Improved inference for unlexicalized parsing. *In Proceedings of HLT-NAACL 2007*, pages 404–411, Rochester, New York.

SAGAE, K. (2010). Self-training without reranking for parser domain adaptation and its impact on semantic role labeling. *In Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 37–44, Uppsala, Sweden.

SAGOT, B. (2010). The L*efff*, a freely available and large-coverage morphological and syntactic lexicon for french. *In Proceedings of LREC'10*, Valetta, Malta.

SEDDAH, D., CANDITO, M. et CRABBÉ, B. (2009). Cross parser evaluation and tagset variation : a french treebank study. *In Proceedings of IWPT 2009*, IWPT '09, pages 150–161, Stroudsburg, PA, USA.

STEEDMAN, M., HWA, R., CLARK, S., OSBORNE, M., SARKAR, A., HOCKENMAIER, J., RUHLEN, P., BAKER, S. et CRIM, J. (2003). Example selection for bootstrapping statistical parsers. *In Proceedings of the NAACL 2003*, pages 157–164.

TIEDEMANN, J. (2009). News from opus - a collection of multilingual parallel corpora with tools and interfaces. *Recent advances in natural language processing V : selected papers from RANLP 2007*, 309:237.

VILLEMONTE DE LA CLERGERIE, E., HAMON, O., MOSTEFA, D., AYACHE, C., PAROUBEK, P. et VILNAT, A. (2008). Passage : from french parser evaluation to large sized treebank. *In Proceedings of LREC'2008*.

ACOLAD
Plateforme pour l'édition collaborative dépendancielle

Francis Brunet-Manquat et Jérôme Goulian
LIG-GETALP, Université Pierre Mendès France Grenoble 2
Francis.Brunet-Manquat@imag.fr et Jerome.Goulian@imag.fr

RESUME

Cet article présente une plateforme open-source pour l'édition collaborative de corpus de dépendances. Cette plateforme, nommée ACOLAD (Annotation de COrpus Linguistique pour l'Analyse de Dépendances), propose des services manuels de segmentation et d'annotation multi-niveaux (segmentation en mots et en syntagmes minimaux (chunks), annotation morphosyntaxique des mots, annotation syntaxique des chunks et annotation syntaxique des dépendances entre mots ou entre chunks). Dans cet article, nous présentons la plateforme ACOLAD, puis nous détaillons la représentation pivot utilisée pour gérer les annotations concurrentes, enfin décrivons le mécanisme d'importation de ressources linguistiques externes.

ABSTRACT

ACOLAD: platform for collaborative dependency annotation

This paper presents an open-source platform for collaborative editing dependency corpora. ACOLAD platform (Annotation of corpus linguistics for the analysis of dependencies) offers manual annotation services such as segmentation and multi-level annotation (segmentation into words and phrases minimum (chunks), morphosyntactic annotation of words, syntactic annotation chunks and annotating syntactic dependencies between words or chunks). In this paper, we present ACOLAD platform, then we detail the representation used to manage concurrent annotations, then we describe the mechanism for importing external linguistic resources.

MOTS-CLES : annotation collaborative de corpus, annotations concurrentes, dépendances
KEYWORDS : corpus collaborative annotation, concurrent annotations, dependencies

1 Introduction

De nombreuses applications de TAL nécessitent de grandes quantités de données annotées manuellement. La production de ces données est coûteuse. D'autre part, la nature et la qualité des annotations à produire dépendent très largement des besoins en terme d'exploitations futures du corpus (Valli et Véronis, 1999). Pour faciliter la production de tels corpus, plusieurs outils récents ont été développés parmi lesquels on peut citer : l'application Web 2.0 *System EasyRef* développé dans le cadre de l'ANR action Passage pour annoter des corpus syntaxiques dans les formats Easy et Passage (Paroubek et al., 2009), l'extension firefox *WebAnnotator* qui permet d'annoter des pages Web selon une DTD définie par l'utilisateur (Xavier, 2012). Ces outils tentent de résoudre de nombreux problèmes liés à la création de corpus, en particulier, l'aspect collaboratif pour *EasyRef* et l'aspect générique pour *WebAnnotator*. Dans cet article, nous nous intéressons à 2 problèmes cruciaux liés à l'annotation de corpus : Comment

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 335–342,
Grenoble, 4 au 8 juin 2012.

représenter les annotations concurrentes ? Comment importer et utiliser de manière générique des ressources linguistiques *externes* comme des dictionnaires ou des analyses morphosyntaxiques ?

Nous tentons avec la plateforme ACOLAD (Annotation de COrpus Linguistique pour l'Analyse de Dépendances) de répondre à ces questions. Cette plateforme open-source a été développée avec pour objectif de faciliter la tâche d'édition collaborative lors de la création d'un corpus de dépendance. Elle propose des services manuels de segmentation et d'annotation multi-niveaux (segmentation en mots et en syntagmes minimaux (chunks), annotation morphosyntaxique des mots, annotation syntaxique des chunks et annotation syntaxique des dépendances entre mots ou entre chunks).

Après une vue d'ensemble de la plateforme ACOLAD, nous nous focalisons dans cet article sur la représentation pivot utilisée pour gérer les annotations concurrentes et le mécanisme d'importation de ressources linguistiques externes.

2 ACOLAD, une plateforme pour l'édition de corpus de dépendances

2.1 Présentation

L'idée est de proposer sur la même plateforme une palette de services permettant la création de corpus annotés pour les besoins classiques de développement d'outils linguistiques. Les intérêts de notre environnement sont les suivants :

– visualisation et édition graphique simple des annotations (voir figure 1) ;

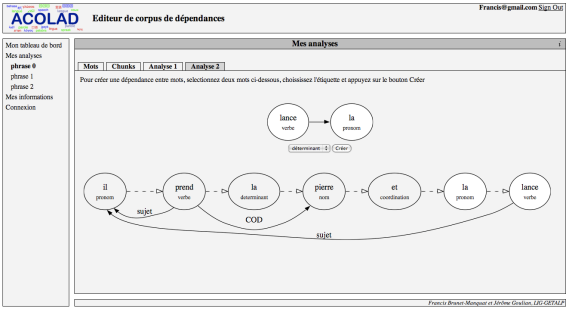


Figure 1 : création d'une analyse de dépendances

– édition manuelle configurable :
  - choix entre différentes granularités d'unités d'annotations (dépendances entre mots et/ou entre chunks) ;
  - possibilité d'utiliser différents jeux d'étiquettes (grammaticales, syntaxiques au niveau des chunks et des dépendances) pour tenir compte des spécificités de chaque corpus (écrit, oral transcrit ou oral issu de la reconnaissance de parole par exemple) et des besoins en terme d'exploitation future du corpus ;

- choix des contraintes structurelles de dépendances à associer à l'édition en cours[1] (projectivité ou non projectivité, analyse totale ou partielle) ;

– possibilité d'éditer simultanément les ambiguïtés d'analyse tant au niveau mots, chunks, qu'au niveau des dépendances syntaxiques.

2.2 Architecture de la plateforme ACOLAD

La plateforme ACOLAD est une application web consacrée au développement collaboratif de corpus de dépendances. La plateforme est organisée en une architecture 3-tiers classique (voir figure 2) : une couche de *présentation* (responsable de l'interface avec les utilisateurs), une couche de traitement (qui fournit les services) et une couche *données* (responsable du stockage des données persistantes).

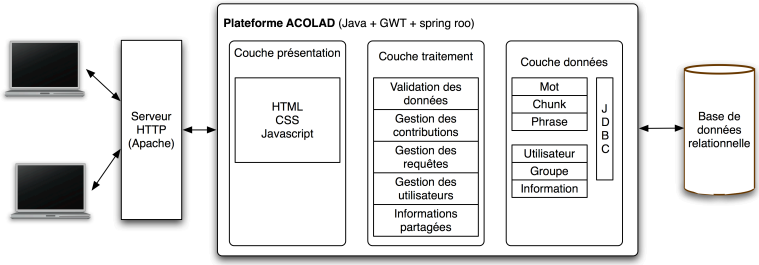


Figure 2 : architecture générale de la plateforme ACOLAD

Pour utiliser la plateforme ACOLAD, le responsable de la tâche d'annotation du corpus n'a pas à écrire de code spécifique Java ni de pages Web dynamiques spécifiques. Il doit simplement fournir, outre la liste des annotateurs et des modérateurs, une description du jeu d'étiquettes (sous forme XML) et éventuellement les dictionnaires nécessaires à la tâche (quelques dictionnaires sont déjà intégrés, par exemple LEFFF/DELAF pour le français).

3 Représentation pivot pour l'annotation concurrente

La plateforme ACOLAD se base sur une représentation par pivot pour distinguer les différences de segmentation, d'étiquetage et d'annotation dépendancielle.

3.1 Matrice de dépendances (MD)

Toute annotation est décrite, dans notre plate-forme, par une représentation matricielle. Notre représentation, nommée matrice de dépendance (MD), est un couple <L, M> composé de :

– Une liste de nœuds (L), un nœud étant composé d'informations linguistiques

[1] Pour le moment il s'agit d'une liste pré-établie de contraintes.

relatives aux mots et/ou aux chunks[2] qu'il décrit ;
- Une matrice carrée (M) permettant de décrire les dépendances entre nœuds (implémentée sous forme de matrice creuse). La case (i, j) contient l'ensemble des dépendances entre le nœud i et le nœud j de la liste de nœuds.

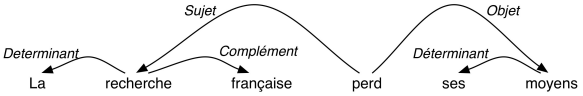


Figure 3 : exemple d'analyse de dépendances

La MD correspondant à la structure de dépendance syntaxique de la figure 3 est :

L =

la	::	cat=déterminant
recherche	::	cat=nom
française	::	cat= adjectif
perd	::	cat=verbe
ses	::	cat=déterminant
moyens	::	cat=nom

M =

relation →	la	recherche	française	perd	ses	moyens
la						
recherche	Déterminant		Complément			
française						
perd		Sujet				Objet
ses						
moyens					Déterminant	

Cette représentation matricielle des données présente trois avantages pour le traitement informatique :

- Maniabilité : de nombreux outils mathématiques sont associés aux matrices : ajout, suppression, comparaison, etc. Tous ces outils permettent un traitement simple de l'information contenue dans une matrice.
- Efficacité : les méthodes utilisant les matrices comme structures de données, telle que la reconnaissance de motifs ou la fusion de matrice, se montrent très efficaces et très simples à mettre en place.
- Simplicité de la description d'analyse multi-niveaux (toutes les informations peuvent être présente dans la même matrice de dépendances)

3.2 Annotations concurrentes

Pour illustrer le mécanisme de gestion d'annotations concurrentes d'ACOLAD, nous prenons comme exemple dans la suite le cas de segmentations multiples. Le principe consiste à regrouper les nœuds représentant la même segmentation dans la phrase (information commune minimale). Mais elle consiste également à représenter les discordances issues des différentes segmentations des structures et dues, par exemple, aux mots composés, aux entrées des dictionnaires (États Unis ou États-Unis), etc.

Pour ce faire, nous créons une structure, appelée réseau de segmentation (RS), représentant les différentes segmentations de la phrase et permettant de lier les nœuds

[2] Dans la suite de l'article, on ne traitera pour l'exemple que les relations de dépendance entre mots.

des structures. Ce réseau peut être vu comme un « pivot de liaison » entre ces structures. Ce réseau est un treillis, chaque nœud du réseau représentant une segmentation possible d'un mot et servant de liaison entre les nœuds des structures de dépendance. Concrètement, un nœud Nrs d'un RS contient deux informations :

- SNODE(Nrs) : intervalles représentant la sous-chaîne dans la phrase correspondant au nœud Nrs, Par exemple, les mots de la phrase « On avait dénombré cent vingt-neuf candidats » auront pour intervalles : On[1-2], avait[3-7], dénombré[8-15], etc. Cette information est basée sur les SSTC (Structured String-Tree Correspondences) proposée par (Boitet et Zaharin, 1988) ;
- L : un ensemble contenant les nœuds des structures liés au nœud Nrs.

Le réseau de segmentation final obtenu représente les segmentations possibles et lie les nœuds des structures entre eux. Une fois que la correspondance entre les nœuds des structures est établie, la plateforme peut fusionner ces structures pour fournir une unique représentation de dépendance combinant toutes les informations linguistiques relatives aux structures (illustré dans la figure 4).

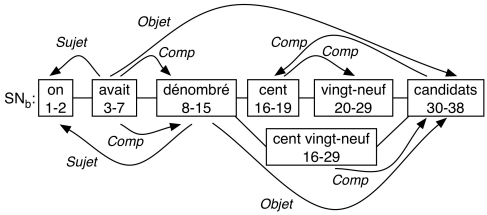


Figure 4 : exemple d'annotations concurrentes

4 Importer des informations linguistiques *externes*

La plate-forme ACOLAD doit être capable de récupérer des informations linguistiques provenant d'autres outils (par exemple d'analyseurs lexicaux ou morphosyntaxiques), c'est-à-dire d'extraire les informations des résultats qu'ils produisent puis de les interpréter. Nous pensons en effet que cette fonctionnalité doit faire partie intégrante de l'outil afin de fournir un cadre simple de transformation des données aux utilisateurs diffusable en tant que « plugin » avec l'outil. Le processus se compose de deux phases :

- La phase d'extraction proprement dite dépendante des analyseurs utilisés ;
- La phase de projection des informations extraites.

4.1 Extraction des informations

Un module d'extraction peut être généré facilement pour chaque outil externe. Ce module permet de lire les fichiers résultats produits, de vérifier leurs bonnes structurations et d'extraire les données linguistiques contenues dans ces fichiers.

4.1.1 Production des modules d'extraction

La production de modules d'extraction s'organise en deux étapes. Dans un premier temps, il faut définir un fichier de spécifications qui décrit le format du fichier résultat produit par l'outil (une grammaire de description). Il faut ensuite enrichir cette grammaire avec des méthodes[3] qui permettent d'extraire les données linguistiques contenues dans les résultats d'analyse. Un module d'extraction peut alors être produit à partir de la grammaire et des méthodes.

Chaque module d'extraction généré est constitué de deux analyseurs :

- Un analyseur lexical qui convertit une séquence de caractères provenant du fichier résultat à analyser en une séquence d'objets, nommés tokens ;
- Un analyseur syntaxique qui consomme la séquence de tokens calculée précédemment pour vérifier la syntaxe du fichier résultat. Il exécute également les actions associées à la grammaire.

4.1.2 Grammaire de description

Une grammaire de description représente le format des fichiers résultats fournis par un analyseur. Cette grammaire est de type hors contexte et elle est composée de deux ensembles :

- Un ensemble de tokens, représentant les objets (nombre, chaîne, etc.) contenus dans un fichier résultat ;
- Un ensemble de règles syntaxiques, représentant la syntaxe d'un fichier résultat en fonction des tokens.

4.1.3 Grammaire de description enrichie

Pour extraire les données linguistiques présentes dans les fichiers résultats, des méthodes d'extraction sont introduites dans les grammaires de description. Ces méthodes se présentent sous la forme de procédures insérées dans les règles syntaxiques. Elles permettent d'extraire les données linguistiques. Quand une règle syntaxique de la grammaire est appliquée, les méthodes correspondantes sont exécutées. La grammaire ainsi créée est nommée grammaire de description enrichie. Nous donnons ci-après un extrait d'une grammaire de description augmentée pour une analyse de dépendance.

```
// Règle syntaxique
void RelationDeDependance() {
   // Déclaration des variables
   String etiquette ; Mot mot1, mot2 ;

   etiquette=<Chaine> "(" mot1=Element() "," mot2=Element() ")"
      // où   - Element() est une règle syntaxique retournant un objet Mot
      //      - <Chaine> est un token représentant une chaîne
      {
         // Extraction d'une relation syntaxique avec la
         // méthode d'extraction AjoutRelationSyntaxique
         phraseCourante.AjoutRelationSyntaxique(etiquette, mot1, mot2)
```

[3] Un panel de méthodes est proposé pour réaliser l'extraction des informations linguistiques.

```
   }
}
void Element() {
   <Chaine> "^" <Chaine> "^" ( <Chaine> )+ ( "_" <Chaine> ) )? ":" <Nombre>
      // <Nombre> est un token représentant un nombre positif
      // le + signifie que la chaîne peut apparaître plusieurs fois
      // le ? signifie que la chaîne peut apparaître au moins une fois
}
```

La règle syntaxique précédente s'applique sur chacune des relations de dépendance extraites de l'analyse faite par l'analyseur XIP de la phrase « La recherche française perd ses moyens. » :

```
SUBJ(<perd^perdre^+VERB_P3SG:3>,<recherche^recherche^+NOUN_SG:1>)
VARG(<perd^perdre^+VERB_P3SG:3>,<moyens^moyen^+NOUN_PL:5>)
NN(<recherche^recherche^+NOUN_SG:1>,<française^française^+NOUN_INV:2>)
DETERM(<La^le^+DET_SG:0>,<recherche^recherche^+NOUN_SG:1>)
```

4.2 Projection des données extraites

Après avoir extrait toutes les données linguistiques des résultats d'un analyseur, il faut faire correspondre ces données à une norme commune établie, spécifiée en amont en fonction des jeux d'étiquettes donnés pour la tâche d'annotation.

Cette phase peut être reproduite facilement autant de fois que nécessaire et ainsi s'adapter aux différentes granularités des jeux d'étiquettes envisagés en fonction des besoins de la tâche.

Pour normaliser les données extraites, un ensemble de règles de projection est défini. Une règle de projection est constituée d'une partie gauche représentant les données linguistiques à reconnaître et d'une partie droite représentant les mêmes informations normalisées. Par exemple, pour les données brutes extraites dans l'exemple précédent, les règles instanciables sont[4] :

```
SUBJ($var1, $var2) ::= Sujet($var1, $var2)
VARG($var1, $var2) ::= Objet($var1, $var2)
NN($var1, $var2) ::= Complement($var1, $var2)
DETERM($var1, $var2) ::= Determinant($var2, $var1)
             // La relation DETERM est projetée en une relation
             // Determinant en modifiant l'ordre relatif des mots
```

L'extraction d'information d'ACOLAD est basée sur les outils développés pour la plateforme de combinaison d'analyseurs syntaxiques DepAn (Brunet-Manquat, 2005). L'approche a été expérimentée et validée dans (Brunet-Manquat, 2004).

5 Perspectives

ACOLAD propose des services manuels permettant de segmenter une phrase ou de produire des analyses de dépendances. Cet environnement est actuellement utilisé pour produire des analyses de dépendances pour le français. Mais cet outil, par son

[4] une variable $vari représente soit un mot soit un chunk

formalisme de dépendances et son aspect configurable (choix des jeux d'étiquettes, choix des contraintes structurelles de dépendances -- en particulier pour tenir compte de la variabilité de l'ordre des mots selon la langue (Holan, 2000)), a un potentiel multilingue.

Des expérimentations sont actuellement menées pour intégrer les premiers résultats d'ACOLAD dans nos travaux sur l'analyse syntaxique automatique et la production de documents auto-explicatifs (Blanchon et al., 2006).

Les services proposés par ACOLAD pourront également être intégrés à des environnements tel que Sectra_w, un système collaboratif permettant d'évaluer, de présenter, d'exploiter et de réviser des corpus de traduction automatique (Huynh et al., 2008), par exemple pour proposer l'ajout de dépendances syntaxiques.

Enfin, ACOLAD pourra être proposé dans le cadre de campagne d'évaluation d'analyseurs syntaxiques de dépendances pour aider à fabriquer les analyses références.

Notre plateforme est proposée à la communauté sous licence publique générale limitée GNU (GNU Lesser General Public License). Elle sera prochainement disponible à l'url suivante : https://forge.imag.fr/projects/acolad/.

Références

BLANCHON, H., BOITET, C. AND CHOUMANE, A. (2006). Traduction automatisée fondée sur le dialogue et documents auto-explicatifs: bilan du projet LIDIA. *in TAL*. vol. 47(3):30 p.

BRUNET-MANQUAT F. (2005). Improving dependency analysis by Syntactic parser combination. *Proceedings of IEEE NLP-KE 2005*, Wuhan, China, Oct 30- Nov 1, 2005.

BRUNET-MANQUAT, F. (2004). CRÉATION D'ANALYSEURS DE DÉPENDANCE PAR COMBINAISON D'ANALYSEURS SYNTAXIQUES. THÈSE EN INFORMATIQUE. UNIVERSITÉ JOSEPH FOURIER - GRENOBLE 1. 21 DÉCEMBRE 2004. 169 P.

HOLAN T., KUBON, OLIVA K., PLATEK M. (2000). On complexity of word order, *in TAL*., 41(1), pp. 273-300, Hermès, Paris, France.

HUYNH C.-P., BOITET C. & BLANCHON H. (2008). SECTra_w : an Online Collaborative System for Evaluating, Post-editing and Presenting MT Translation Corpora. *Proc. LREC-08,* Marrakech, 27-31/5/08, ELRA/ELDA, ed., 8 p.

PAROUBEK P., VILLEMONTE DE LA CLERGERIE E., LOISEAU S., VILNAT A., ET FRANCOPOULO G. (2009) The PASSAGE Syntactic Representation, *7th International Workshop on Treebanks and Linguistic Theories* (TLT7, Groningen, January 23-24, 2009)

VALLI, A. VERONIS, J. (1999). Etiquetage grammatical des corpus de parole : problèmes et perspectives. *Revue Française de Linguistique Appliquée, IV(2),* 113-133. (dossier : l'oral spontané)

XAVIER T. (2012). WebAnnotator, an Annotation Tool for Web Pages. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012, to appear)*. Istanbul, Turkey, 2012.

Extraction de préférences à partir de dialogues de négociation

Anaïs Cadilhac Farah Benamara Vladimir Popescu Nicholas Asher
Mohamadou Seck
IRIT, 118, Route de Narbonne, 31062 Toulouse Cedex 9
{cadilhac, benamara, popescu, asher, seck}@irit.fr

RÉSUMÉ
Cet article présente une approche linguistique pour l'extraction d'expressions de préférence à partir de dialogues de négociation. Nous proposons un nouveau schéma d'annotation pour encoder les préférences et les dépendances exprimées linguistiquement dans deux genres de corpus différents. Ensuite, nous proposons une méthode d'apprentissage qui extrait les expressions de préférence en utilisant une combinaison de traits locaux et discursifs. Finalement, nous évaluons la fiabilité de notre approche sur chaque genre de corpus.

ABSTRACT
Towards Preference Extraction From Negotiation Dialogues

This paper presents an NLP based approach for preference expression extraction from negotiation dialogues. We propose a new annotation schema for preferences and dependencies among them and illustrate on two different corpus genres. We then suggest a learning approach that efficiently extracts preference expressions using a combination of local and discursive features and assess the reliability of our approach on each corpus genre.

MOTS-CLÉS : Préférence, dialogue, apprentissage automatique.

KEYWORDS: Preference, dialogue, machine learning.

1 Introduction

Modéliser les préférences est incontournable dans de nombreux problèmes de la vie courante, que ce soit pour la prise de décision individuelle ou collective (Arora et Allenby, 1999), les interactions stratégiques entre agents (Brainov, 2000) ou la théorie des jeux (Hausman, 2000). Un aperçu des travaux sur les préférences en Intelligence Artificielle est donné par Kaci (2011).

Une préférence est généralement définie comme un ordre donné par un agent sur différentes options. Les options dépendent du domaine : elles peuvent être des vols d'avions, des voitures, des horaires et lieux de rendez-vous, etc. L'ordonnancement des préférences peut être total (strict ou non), rendant chaque paire d'options comparable, ou partiel, quand certaines options ne peuvent pas être comparées par un agent donné. Parmi ces options, certaines sont acceptables pour l'agent, c'est-à-dire qu'il est prêt à agir pour les réaliser, et d'autres ne le sont pas. Parmi les options acceptables, l'agent en préfère généralement certaines par rapport aux autres.

Il est important de différencier les notions de préférence et d'opinion. Alors que les opinions sont un point de vue, un sentiment ou un jugement qu'un agent peut avoir sur un objet ou une personne, les préférences, comme nous les avons définies, impliquent un ordre de la part de

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 343–350,
Grenoble, 4 au 8 juin 2012.

l'agent et sont ainsi relationnelles et comparatives. Les opinions concernent donc un jugement absolu sur des objets ou des personnes (positif, négatif ou neutre), tandis que les préférences concernent un jugement relatif sur des options, les préférant, ou non, aux autres. Par exemple, *Ce film n'est pas mauvais* exprime une opinion directe positive sur un film mais nous ne savons pas si ce film est le « plus » préféré. *J'aimerais aller au cinéma. Allons voir Madagascar 2* exprime deux préférences, l'une dépendant de l'autre. La première est que l'auteur préfère aller au cinéma par rapport aux autres actions alternatives ; la seconde est qu'étant donnée cette préférence, il préfère aller voir *Madagascar 2* plutôt que les autres films possibles.

Traiter les préférences n'est pas aisé. Tout d'abord, il est nécessaire de connaître au moins partiellement l'ensemble des options sur lesquelles portent les préférences. Ensuite, il faut pouvoir définir un ordre a priori sur les options acceptables mais ce n'est pas toujours trivial. De plus, donner un ordre entre deux options (appareils photos) peut être difficile à cause de la nécessité de tenir compte des compromis et des interdépendances entre les différents critères (durée de vie de la batterie, poids, etc.). Ensuite, les utilisateurs manquent souvent d'informations complètes sur leurs préférences initiales qui tendent à changer au cours du temps. En effet, les utilisateurs peuvent apprendre du domaine, des préférences des autres et même de leurs propres préférences au cours du processus de prise de décision. Plusieurs méthodes ont été proposées en Intelligence Artificielle pour éliciter les préférences (Chen et Pu, 2004). Cependant, à notre connaissance, aucun travail ne montre comment les préférences pourraient être déterminées à partir de dialogues en utilisant une approche linguistique.

L'approche que nous proposons a pour but d'étudier le rôle du discours pour extraire et raisonner sur les préférences. Notre approche comporte trois étapes :

1. *Extraire les options.* L'objectif est de repérer, au sein de chaque segment de discours, les expressions linguistiques sur lesquelles portent les préférences d'un agent.
2. *Identifier les éventuelles dépendances entre les options extraites à l'étape 1 en utilisant un ensemble d'opérateurs spécifiques*. Ces dépendances nous permettent d'inférer les préférences de l'agent et d'identifier, étant données deux options, leur ordonnancement.
3. *Proposer une description formelle des préférences de chaque agent.* Nous étudions comment les relations de discours permettent de suivre l'évolution des préférences au cours du dialogue. Cette description se fait en utilisant une représentation compacte des préférences, les CP-nets (Conditional Preference Networks) (Boutilier *et al.*, 2004).

Une description détaillée de ces étapes est donnée dans (Cadilhac *et al.*, 2011). Le travail présenté ici est un premier pas vers l'automatisation de ce processus en se focalisant sur la première étape. Nous analysons comment les préférences sont exprimées linguistiquement, c'est-à-dire comment les options et les dépendances sont lexicalisées. Nous montrons comment les options peuvent être extraites automatiquement grâce à un algorithme d'apprentissage supervisé utilisant des traits locaux et discursifs et nous évaluons la fiabilité de notre approche.

2 Données

Nos données viennent de deux corpus. Le premier corpus, `Verbmobil`, est composé de 35 dialogues choisis au hasard dans le corpus Verbmobil déjà existant (Wahlster, 2000), dans lequel deux agents discutent pour fixer la date et le lieu d'un rendez-vous. Il a été utilisé pour créer la liste des traits d'apprentissage. Le second corpus, `Réservations`, a été utilisé pour évaluer à quel point notre méthode est dépendante du domaine. Il a été construit à partir de plusieurs ressources d'apprentissage de l'anglais disponibles sur Internet (par exemple, www.bbc.co.uk/worldservice/learningenglish). Il contient 21 dialogues choisis au hasard, dans

lesquels un agent, le client, appelle un service pour réserver une chambre, un vol d'avion, un taxi, etc. En voici un exemple typique :

π_1 *A* : Northwind Airways, good morning. May I help you ?
π_2 *B* : Yes, do you have any flights to Sydney next Tuesday afternoon ?
π_3 *A* : Yes, there's a flight at 16:45 and one at 18:00.
π_4 *A* : Economy, business class or first class ticket ?
π_5 *B* : Economy, please.

Afin d'analyser comment les options et les dépendances sont exprimées linguistiquement dans les dialogues de négociation, nous avons réalisé une annotation à deux niveaux : d'abord, au niveau du discours, séparant le texte en segments (les π_i ci-dessus) liés entre eux par des relations rhétoriques ; puis, au niveau des préférences exprimées dans chaque segment.

2.1 Annotation du discours

Les dialogues sont structurés par des tours de parole qui permettent à chaque agent de participer au dialogue. Un agent peut, par exemple, répondre aux questions d'autres agents, poser ses propres questions, etc. Dans chacun de ces tours, les agents s'engagent sur leurs croyances et préférences. En général, les modèles formels de dialogue n'explicitent pas le lien entre les énoncés et les préférences (voir, par exemple, (Ginzburg, 2012)). Il est alors nécessaire d'avoir, d'une part, une méthode qui permet une extraction partielle des préférences et de leurs dépendances et, d'autre part, une méthode qui permet d'exploiter cette description partielle afin d'identifier l'ensemble des options préférées.

Notre approche exploite la structure du discours selon la Théorie des Représentations Segmentées du Discours, SDRT (Asher et Lascarides, 2003) où des unités discursives (UD) sont liées entre elles par des relations rhétoriques telles que *Paire Question-Réponse (QAP)*, *Plan-Correction (P-Corr)*, etc. Bien que le problème d'extraction de la structure du discours reste redoutable, on peut approximer ces relations relativement bien en utilisant des caractéristiques qui peuvent facilement être obtenues automatiquement (par exemple, Baldridge et Lascarides (2005b) réalisent un F-score d'environ 69,2%). Notre étude montre ici l'importance des caractéristiques du discours pour l'extraction de préférences, en supposant que celles-ci sont données par un oracle. Pour `Verbmobil`, nous avons utilisé l'annotation de Baldridge et Lascarides (2005a). Pour `Réservations`, l'annotation a été faite par consensus en utilisant le même ensemble de relations rhétoriques qui a été utilisé pour annoter `Verbmobil`.

2.2 Annotation des préférences

Notre objectif est d'analyser comment les préférences sont linguistiquement exprimées dans des segments de dialogues. Deux étapes sont nécessaires : (i) identifier l'ensemble O des options (des termes) sur lesquelles portent les préférences d'un agent, (ii) identifier les éventuelles dépendances entre les éléments de O en utilisant un ensemble d'opérateurs spécifiques, c'est-à-dire identifier les préférences de l'agent parmi les options énoncées. Par exemple, dans *Rencontrons nous lundi ou mardi*, nous avons $O = \{lundi,\ mardi\}$ où les options sont linguistiquement reliées par la conjonction *ou* qui signifie que l'agent est prêt à réaliser une de ces options, les préférant de manière égale.

Dans une UD, les préférences peuvent être exprimées de différentes manières. Elles peuvent être atomiques, par exemple, « Je veux X » ou « Je préfère X » où « X » est une option acceptable. Cette option peut être un groupe nominal (*lundi*), un groupe prépositionnel (*à mon bureau*) ou un groupe verbal (*se rencontrer*). Les préférences peuvent aussi être exprimées dans des constructions comparatives et/ou superlatives (*un vol moins cher*). Elles sont aussi exprimées

d'une manière indirecte en utilisant des questions. Bien que toutes les questions n'impliquent pas que l'auteur s'engage sur une préférence, dans beaucoup de cas elles le font. C'est-à-dire si un agent demande *Pouvons-nous nous rencontrer la semaine prochaine ?*, il implique une préférence pour se rencontrer. Des expressions de sentiment ou de politesse peuvent aussi être utilisées pour introduire indirectement des préférences. Dans `Réservations`, le segment *Economique, s'il vous plaît* indique que l'agent préfère être dans la classe économique.

Les expressions de préférences peuvent aussi être complexes, exprimant des négations, conjonctions, disjonctions, ou dépendances. Nous associons à chacune de ces expressions des opérateurs spécifiques (non-booléens), que nous désignons respectivement par not, &, or et $\mapsto$. Les réalisations linguistiques de ces opérateurs nous seront utiles dans la phase d'extractions des options (voir section 3). Les négations indiquent ce que l'agent ne préfère pas, c'est-à-dire que l'option exprimée est non-préférée. La négation peut être explicite, comme dans *Je ne veux pas qu'on se rencontre vendredi*, ou inférée à partir du contexte, comme dans *Je suis occupé mardi*. Un exemple de conjonction entre préférences est *Pourrais-je avoir un petit déjeuner et un repas végétarien ?* où l'agent exprime deux préférences qu'il souhaite satisfaire et il aimerait en avoir au moins une des deux s'il ne peut pas les avoir toutes. La sémantique des disjonctions est une modalité de choix libre. Par exemple, *Je suis libre lundi ou mardi* signifie que lundi ou mardi est un jour possible pour se rencontrer et que l'agent est indifférent entre les deux. Finalement, certaines UD expriment des engagements sur des préférences dépendantes. Par exemple, dans la phrase *Pourquoi pas lundi, dans l'après-midi ?*, il y a deux préférences : une pour le jour lundi et, étant donné la préférence pour lundi, une pour la période de l'après-midi (au moins pour une des interprétations syntaxiques du segment).

Pour chaque UD, nous avons demandé à deux annotateurs d'identifier comment les options sont exprimées et ensuite d'indiquer comment les préférences sur ces options sont liées entre elles en utilisant les opérateurs spécifiques not, &, or et $\mapsto$. Nous donnons ci-dessous un exemple de comment certains segments sont annotés. $< o >$_i indique que o est l'option numéro i dans le segment, et le symbole // est utilisé pour séparer les deux niveaux d'annotation. Une description détaillée de ce schéma d'annotation est donnée dans (Cadilhac *et al.*, 2012).

π_1 : Je suis libre <à quatre>_1 ou <cinq heures>_2 <ces jours-là>_3. // $3 \mapsto (1\ or\ 2)$

π_2 : <Mardi 16>_1, j'ai séminaire <de 9h à midi>_2. // $1 \mapsto not\ 2$

En utilisant le coefficient Kappa de Cohen, nous avons calculé deux taux d'accord inter-annotateurs sur l'identification des options. L'un est basé sur un accord *exact* où deux annotations (c'est-à-dire, les unités de texte correspondant à une option) doivent correspondre exactement pour être considérées comme correctes. L'autre est basé sur un accord souple où deux annotations correspondent s'il y a un chevauchement entre leurs unités de texte (comme pour *2 heures* et *environ 2 heures*). Nous avons obtenu un accord exact de 0,66 et un accord souple de 0,85. L'accord souple étant bon pour `Verbmobil`, nous avons décidé d'annoter `Réservations` par consensus.

Le gold standard pour les deux corpus a été construit après discussion des cas de désaccord. Nous avons observé quatre cas. (1) Le premier concerne la *redondance des préférences* et nous avons décidé de ne pas garder les préférences redondantes dans le gold standard. En effet, les agents répètent souvent des préférences qui ont déjà été établies, comme dans l'exemple suivant, π_1 *A : jeudi, vendredi et samedi je ne suis pas là.* π_2 *A : Ces 3 jours ne sont pas possibles pour moi*, où nous avons *Resultat*(π_1, π_2). (2) Le second cas de désaccord vient des préférences qui sont exprimées par des *anaphores*. Nous avons décidé de les annoter dans le gold standard car elles sont souvent utilisées dans les corpus pour introduire ou préciser des préférences. Comme dans

l'exemple suivant, π_1 *A : A 2 heures, le 17 ?* π_2 *B : C'est parfait*, où nous avons *Q-Elab*(π_1, π_2). (3) Le troisième cas de désaccord concerne *l'explication de préférence*. Dans le gold standard, nous avons choisi de ne pas annoter les expressions qui sont utilisées pour expliquer des préférences déjà établies. Comme dans l'exemple suivant, π_1 *A : pas lundi,* π_2 *A : j'ai un cours de 9 à 12 heures*, où nous avons *Explication*(π_1, π_2). (4) Finalement, le dernier cas de désaccord provient des préférences qui ne sont pas directement liées à l'action de fixer une date pour le rendez-vous mais plutôt à d'autres actions comme déjeuner ensemble. Même si ces préférences ont souvent été omises par les annotateurs, nous avons décidé de les garder.

3 Extraction des objets de préférences

Le problème de l'extraction est de décider si un terme est une option ou non. L'objectif est donc de classer les termes en deux catégories : *Option* et *Non-option* indiquant respectivement que le terme exprime une option faisant l'objet des préférences, ou non. Nous rappelons que les options peuvent être des groupes nominaux, groupes prépositionnels ou groupes verbaux. Nous devons donc choisir quels groupes de mots doivent être classés. Dans les données, les agents négocient pour se mettre d'accord sur une action : se rencontrer un jour donné, réserver un certain vol, etc. Nous sommes généralement informés de ces actions dans les groupes verbaux. Cependant, les termes correspondants aux options de préférences sont plutôt contenus dans les groupes nominaux (GN). Par exemple, pour fixer un rendez-vous, la négociation porte sur les jours et les heures. Pour réserver un hôtel, la négociation porte sur des options plus spécifiques comme *une chambre double*. Il semble donc approprié d'extraire les GN. Pour les classer dans une des deux classes, nous utilisons deux genres de traits (tous binaires) : les traits locaux et les traits discursifs. Le classifieur est basé sur les Machines à Vecteurs de Support.

La portée des traits locaux est soit l'unité qui doit être classée, c'est-à-dire le GN, soit le segment qui contient le GN. Certains de ces traits reposent sur une ontologie qui modélise un calendrier (date, temps, etc.) inspirée de deux ontologies de haut niveau, SUMO et COSMO. Nous avons cinq traits au niveau du GN qui testent si le GN contient : le label d'un concept appartenant à l'ontologie, un comparatif, un superlatif, une disjonction ou une conjonction. Nous avons dix traits au niveau du segment : (1) le voisin gauche du GN correspond à un label d'un concept de l'ontologie. Puisque la liste des termes associés à chaque concept de notre ontologie est courte, ce trait aide à retrouver des lexicalisations supplémentaires ; (2) le segment contient une disjonction ou une conjonction ; (3) le GN est dans la portée d'une négation, d'un modal ou d'un verbe d'action du domaine (*se rencontrer*, *réserver*). La portée des négations et des modaux est résolue de manière simplifiée en utilisant l'arbre syntaxique de l'UD ; (4) le segment contient un mot d'opinion (*bon*, *mauvais*, *OK*, etc.), un mot de politesse ou un mot qui introduit des préférences (*préférer*, *favori*, *choix*, *trop*, etc.) ; (5) le segment contient une référence à l'autre agent. Ce trait est un indice pour la classe *Non-option*. Dans des segments comme *Tu as dit que tu n'es pas libre mardi matin ou mercredi après-midi ?*, l'agent n'apporte pas de nouvelle information sur les préférences mais répète seulement ce qui a déjà été établi par l'autre agent.

Nous avons neuf traits au niveau du discours : (1) les relations rhétoriques qui lient l'UD courante à l'UD précédente et à l'UD suivante impliquent des préférences. Nous avons remarqué que certaines relations de discours peuvent aider à repérer des segments qui contiennent, ou non, des préférences. Nous dissocions les relations de discours en trois catégories : (a) celles qui impliquent « généralement » une *Non-option* comme *Explication*, *Commentaire*, *Résumé*, (b) celles qui impliquent « peut-être » une *Option* comme *Elaboration*, *Continuation*, *Correction* et

		C_V			C_R			$C_V + C_R$		
		P	R	F	P	R	F	P	R	F
Baselines	Tous les GN	40.9	100.0	58.1	28.0	100.0	43.8	28.3	100.0	44.1
	Ontologie seule	95.6	61.3	74.7	55.6	16.7	25.7	49.2	13.5	21.2
	Classifieur simplifié	65.2	71.1	68.0	68.4	43.3	53.1	43.9	55.7	49.1
Traits Locaux	Tous les traits (GN)	95.7	62.0	75.2	100.0	3.3	6.5	50.7	16.0	24.4
	+ Tous les traits (Segment)	94.1	78.9	85.8	68.4	43.3	53.1	60.2	26.2	36.5
Traits Discursifs	+ Relation Précédente	94.9	78.9	86.2	67.6	41.7	51.6	60.2	26.2	36.5
	+ Relation Suivante	94.0	77.5	84.9	66.7	40.0	50.0	59.4	25.3	35.5
	+ Questions	95.6	75.4	84.3	79.0	50.0	61.2	59.4	25.3	35.5
	+ ≥ 2 occurrences du GN	90.8	83.1	**86.8**	75.6	56.7	**64.8**	62.9	32.9	**43.2**

TABLE 1 – Résultats (pourcentages) pour les trois évaluations.

(c) celles qui impliquent « généralement » une *Option*. Dans `Verbmobil`, 86 % des relations de discours sont de la catégorie (a) alors que 14 % des relations annotées appartiennent à la catégorie (b). Nous observons la même tendance pour `Réservations`. Il n'y a pas d'instances de la catégorie (c) dans les relations de discours utilisées lors de l'annotation des deux corpus. Ainsi, nous avons six traits : trois pour tester si la relation entre l'UD courante et l'UD précédente appartient, ou non, à une des trois catégories, et trois autre pour la relation entre l'UD courante et l'UD suivante ; (2) l'UD courante ou l'UD précédente est une question. Dans nos corpus, les formes interrogatives ne sont pas toujours suivies par une marque de question. Pour détecter les questions, nous utilisons donc les relations de discours spécifiques, comme *QAP*, *Q-Elab* ; (3) le GN apparaît au moins deux fois dans le dialogue.

4 Evaluation et Résultats

Plusieurs évaluations sont réalisées pour évaluer la validité de notre méthode d'extraction. La première est effectuée sur les 35 dialogues de `Verbmobil` (C_V) pour un total de 1272 UD. Nous le séparons au hasard en un corpus d'entraînement constitué de 25 dialogues, soit 2374 GN, et un corpus de test de 10 dialogues, soit 700 GN. Dans la seconde (C_R), le classifieur est entraîné sur 15 dialogues de `Réservations`, soit 837 GN et testé sur 6 dialogues pris au hasard, soit 312 GN. Les 21 dialogues de ce deuxième corpus comportent au total 348 UD. Pour la troisième, le classifieur est évalué en utilisant `Verbmobil` pour l'entraînement (les 35 dialogues) et `Réservations` pour le test (les 21 dialogues) ($C_V + C_R$). Cette dernière évaluation, plutôt inhabituelle, est supposée aider à déterminer si notre méthode permet l'entrainement sur un corpus plus grand et disponible et le test sur un corpus plus petit et parfois d'un domaine différent. Pour toutes ces évaluations, nous utilisons le logiciel SVM-light (http ://svmlight.joachims.org).

Nous comparons les résultats du classifieur avec ceux de trois baselines : la première classe tous les GN dans la catégorie *Option*, la seconde classe dans la catégorie *Option* tous les GN qui contiennent un concept appartenant à l'ontologie, et la troisième baseline est une version simplifiée de notre classifieur qui utilise seulement un sous-ensemble de nos traits (nous enlevons les traits basés sur l'ontologie ainsi que tous les traits basés sur les relations de discours). La table 1 présente les résultats, sous forme de précision (P), rappel (R) et F-mesure (F). Elle montre d'abord les résultats des baselines. Nous développons ensuite notre modèle en considérant les traits locaux au niveau du GN, puis nous ajoutons les traits locaux au niveau du segment et ajoutons progressivement les traits au niveau du discours (l'ajout des traits est symbolisé par le signe +). La dernière ligne présente le résultat final, obtenu en utilisant tous les traits.

Les résultats dans la table 1 montrent que, parmi les trois baselines, la seconde donne les meilleurs résultats pour `Verbmobil`. Ceci était attendu puisque l'ontologie a été construite pour ces données. Cependant, cette baseline ne permet pas de retrouver toutes les options car certains GN qui contiennent des concepts de l'ontologie ne sont pas des options (ce sont des répétitions, des commentaires, etc.) et bien sûr toutes les options exprimées par les agents ne sont pas couvertes par les concepts de l'ontologie. Pour `Réservations`, l'ontologie dégrade le rappel par rapport à la première baseline, puisqu'il y a un faible recouvrement entre les concepts dans l'ontologie et ceux dans le corpus. Il en va de même pour la troisième évaluation ($C_V + C_R$). Cependant, ce n'est pas un problème critique puisque des ontologies adaptées sont également disponibles pour le domaine touristique. Dans tous les cas, la troisième baseline donne des résultats assez stables, toujours meilleurs que ceux de la première baseline et, dans les deuxième et troisième évaluations (pour lesquelles nous n'avons pas utilisé d'ontologie adaptée), ces résultats sont également meilleurs que ceux de la deuxième baseline. Le classifieur donne un meilleur rappel pour la troisième évaluation que pour la deuxième. Cela peut montrer un problème de rareté des données lors de l'entrainement uniquement sur `Réservations` (configuration (C_R)).

Les évaluations montrent que notre méthode a une tendance similaire sur `Verbmobil` et `Réservations`. Nous voyons que les traits locaux au niveau du GN sont pertinents pour obtenir une bonne précision. Les traits au niveau du segment et les traits discursifs améliorent le rappel et la F-mesure dans les trois configurations. L'amélioration est mieux marquée dans les deuxième et troisième évaluations. Peut-être parce que l'ontologie, moins bien adaptée pour ces évaluations, a moins d'impact sur les performances. Finalement, pour `Verbmobil`, nous obtenons une F-mesure de 86,8 %, i.e. presque 20 % au-dessus de la troisième baseline (classifieur simplifié) et plus de 10 % au-dessus de la deuxième baseline (basée sur l'ontologie). Pour `Réservations`, nous obtenons une F-mesure de 64,8 %, i.e. plus de 10 % au dessus du classifieur simplifié. Pour la troisième évaluation, les résultats ne montrent pas d'amélioration par rapport aux baselines. C'est probablement dû à l'influence de l'ontologie qui adapte mieux les vecteurs de support au corpus d'entrainement (`Verbmobil`), les rendant moins pertinents pour le corpus de test. En désactivant les deux traits basés sur l'ontologie, nous obtenons 50,2 % de précision, 62,9 % de rappel et 55,8 % de F-mesure, soit une amélioration par rapport aux baselines.

Pour les traits discursifs, nous remarquons que, pour `Verbmobil`, les relations rhétoriques entre l'UD courante et l'UD précédente apportent plus d'amélioration que les autres informations discursives. Cela peut s'expliquer par la nature du corpus, où le contexte (exprimé dans les tours de dialogues précédents) est important. Pour `Réservations`, le trait qui teste si l'UD courante ou l'UD précédente sont des questions apporte la meilleure amélioration des performances car ce corpus contient principalement des paires question-réponse. Pour la troisième évaluation, les traits discursifs n'apportent pas d'amélioration importante par rapport aux baselines. C'est peut-être causé par l'incapacité des informations discursives à compenser les différences entre les données d'entrainement et de test : en effet, en principe, il y a plus d'instances des traits locaux (au niveau du GN et du segment) associées à des cas positifs, que d'instances des traits discursifs associées à des cas positifs. Et quand le classifieur est entrainé sur des traits extraits d'un domaine de corpus et testé sur un autre domaine, le poids des traits discursifs peut ne pas suffire à compenser les autres traits, locaux.

Dans ces trois configurations, le trait testant la présence d'un GN au moins deux fois dans le dialogue apporte une amélioration conséquente par rapport aux autres. C'était plutôt attendu puisqu'en principe la fréquence d'un GN apporte de l'information sur le sujet principal, et cela a du sens, puisque les agents ont tendance à exprimer des préférences sur le sujet de la discussion.

5 Conclusion et futurs travaux

Nous avons présenté une méthode linguistique pour l'extraction des expressions de préférence dans des dialogues de négociation. Nous avons d'abord proposé un schéma d'annotation pour étudier comment les préférences sont exprimées dans des dialogues dans deux domaines différents. Nous avons ensuite proposé une méthode d'apprentissage qui extrait les expressions de préférence des dialogues en utilisant une combinaison de traits locaux et discursifs. Les résultats montrent que la structure discursive couplée avec une ontologie est utile pour extraire les expressions de préférence de manière efficace. Dans nos futurs travaux, nous voulons évaluer la méthode sur des corpus plus grands et variés, pour vérifier sa pertinence et sa robustesse sur différents domaines de conversation et registres de discours. Pour le moment, la méthode de classification traite uniquement des GN. Ceci est justifié pour les corpus sur lesquels nous avons travaillé mais nous devons étudier s'il est toujours pertinent d'utiliser uniquement des GN pour d'autres corpus et, si nécessaire, étendre la méthode à d'autres types de syntagmes. Ce travail d'extraction des expressions de préférence est, nous le rappelons, la première étape d'un processus plus complexe d'élicitation des préférences (Cadilhac *et al.*, 2011) qui sera complètement automatisé afin de l'appliquer à des cas pratiques de négociation et marchandage.

Les auteurs remercient le projet STAC ERC Grant n°269427.

Références

ARORA, N. et ALLENBY, G. M. (1999). Measuring the influence of individual preference structures in group decision making. *Journal of Marketing Research*, 36:476–487.

ASHER, N. et LASCARIDES, A. (2003). *Logics of Conversation*. Cambridge University Press.

BALDRIDGE, J. et LASCARIDES, A. (2005a). Annotating discourse structures for robust semantic interpretation. *In Proceedings of the 6th IWCS*.

BALDRIDGE, J. et LASCARIDES, A. (2005b). Probabilistic head-driven parsing for discourse structure. *In Proceedings of CoNLL*.

BOUTILIER, C., BRAFMAN, C., DOMSHLAK, C., HOOS, H. H. et POOLE, D. (2004). Cp-nets : A tool for representing and reasoning with conditional *ceteris paribus* preference statements. *Journal of Artificial Intelligence Research*, 21:135–191.

BRAINOV, S. (2000). The role and the impact of preferences on multiagent interaction. *In Proceedings of ATAL*, pages 349–363. Springer-Verlag.

CADILHAC, A., ASHER, N., BENAMARA, F. et LASCARIDES, A. (2011). Commitments to preferences in dialogue. *In Proceedings of SIGDIAL*, pages 204–215. ACL.

CADILHAC, A., BENAMARA, F. et ASHER, N. (2012). Annotating preferences in negotiation dialogues. À paraître.

CHEN, L. et PU, P. (2004). Survey of preference elicitation methods. Rapport technique.

GINZBURG, J. (2012). *The Interactive Stance : Meaning for Conversation*. Oxford University Press.

HAUSMAN, D. M. (2000). Revealed preference, belief, and game theory. *Economics and Philosophy*, 16(01):99–115.

KACI, S. (2011). *Working with Preferences : Less Is More*. Cognitive Technologies. Springer.

WAHLSTER, W., éditeur (2000). *Verbmobil : Foundations of Speech-to-Speech Translation*. Springer.

Détection de conflits dans les communautés épistémiques en ligne

Alexandre Denis[1] Matthieu Quignard[2] Dominique Fréard[3]
Françoise Détienne[3] Michael Baker[3] Flore Barcellini[4]
(1) UMR 7503 LORIA, CNRS Campus scientifique 54 506 Vandoeuvre-lès-Nancy
(2) UMR 5191 ICAR, CNRS 5 parvis René Descartes 69342 Lyon Cedex 07
(3) UMR 5141 LTCI, CNRS 46 rue Barrault 75 634 Paris Cedex 13
(4) CNAM-CRTD, 41 rue Gay-Lussac 75 005 Paris
denis@loria.fr, matthieu.quignard@univ-lyon2.fr, {dominique.freard,
francois.detienne, michael.baker}{@telecom-paristech.fr},
flore.barcellini@cnam.fr

RÉSUMÉ
La présence de conflits dans les communautés épistémiques en ligne peut s'avérer bloquante pour l'activité de conception. Nous présentons une étude sur la détection automatique de conflit dans les discussions entre contributeurs Wikipedia qui s'appuie sur des traits de surface tels que la subjectivité ou la connotation des énoncés et évaluons deux règles de décision : l'une découle d'un modèle dialectique en exploitant localement la structure linéaire de la discussion, la subjectivité et la connotation ; l'autre, plus globale, ne s'appuie que sur la taille des fils et les marques de subjectivité au détriment des marques de connotation. Nous montrons que ces deux règles produisent des résultats similaires mais que la simplicité de la règle globale en fait une approche préférée dans la détection des conflits.

ABSTRACT
Conflicts detection in online epistemic communities

Conflicts in online epistemic communities can be a blocking factor when producing knowledge. We present a way to automatically detect conflict in Wikipedia discussions, based on subjectivity and connotation marks. Two rules are evaluated : a local rule that uses the structure of the discussion threads, connotation and subjectivity marks and a global rule that takes the whole thread into account and only subjectivity. We show that the two rules produce similar results but that the simplicity of the global rule makes it a preferred approach to detect conflicts.

MOTS-CLÉS : wikipedia, conflit, syntaxe, sémantique, interaction.

KEYWORDS: wikipedia, conflict, syntax, semantics, interaction.

1 Problématique

Les communautés épistémiques en ligne sont des communautés qui rassemblent des individus dans le but de concevoir collectivement des ressources : ontologies, articles, spécifications, etc. (Barcellini *et al.*, 2008). Nous nous plaçons dans la perspective de l'étude de ces communautés et proposons un outil automatique de détection de fils de discussion conflictuels. L'objectif est

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 351–358,
Grenoble, 4 au 8 juin 2012.

double. Tout d'abord un outil de détection de conflit permet de repérer automatiquement les fils de discussion qui peuvent être bloquants ou au contraire productifs pour l'activité de conception. Ensuite, l'outil permet de détecter automatiquement les *individus* les plus conflictuels. Dans les deux cas, cette détection automatique constitue un outil utile pour les gestionnaires de communautés.

La construction d'un outil de détection de conflits s'inscrit dans le cadre du projet CCCP-Prosodie [1], au sein duquel la communauté Astronomie de Wikipedia fut étudiée. En particulier, l'article dédié à l'astre céleste Pluton, qui a perdu son statut de planète en 2006 fit l'objet d'intenses discussions à propos de son renommage. Dans (Fréard *et al.*, 2010), une annotation manuelle des contributions des participants de la page de discussion autour de Pluton est effectuée. La communauté est alors étudiée sous l'angle de l'évolution du conflit autour du renommage grâce à cette annotation manuelle. Toutefois les catégories d'annotation proposées par (Fréard *et al.*, 2010), elles-mêmes inspirées de (Baker *et al.*, 2009) sont difficiles à reproduire automatiquement (acte de dialogue, catégorisation du contenu propositionnel, niveau d'expertise du contributeur) et particulièrement sur le texte libre des pages de discussion Wikipedia.

Nous proposons dans cet article d'explorer les traits accessibles par une méthode automatique qui permettent de caractériser le conflit et proposons alors une méthode pour la détection de ces conflits. Nous discutons d'abord la définition du conflit dans la partie 2, proposons plusieurs méthodes pour détecter le conflit dans la partie 3, et concluons par une évaluation des différentes méthodes dans la partie 4.

2 Qu'est-ce qu'un conflit ?

2.1 Le conflit dans l'argumentation

La notion de *conflit* (ou *conflit d'opinions avouées*) figure à la base des modèles dialectiques de l'argumentation dans le dialogue (Barth et Krabbe, 1982; Mackenzie, 1985), qui analysent le processus argumentatif comme un jeu d'attaques et de défenses visant à déterminer si une proposition (ou thèse) est tenable sous le feu de la critique.

Selon le modèle de Barth & Krabbe (*op.cit*), le conflit est défini par la distribution de rôles vis-à-vis d'une thèse : le *proposant* qui doit défendre la thèse sans se dédire, l'*opposant* qui a le droit d'attaquer ou mettre en doute les allégations du proposant sans toutefois revenir sur ses propres concessions. En pratique, on observe qu'un conflit est déclaré lorsque 3 actes de dialogue (ou *argumentation moves*) ont été produits : (1) l'affirmation d'une proposition p, (2) l'attaque ou la mise en doute de p, (3) la défense de p par une justification ou une contre-attaque.

Dans le cadre d'une discussion critique (ou argumentation rationnelle), les arguments ne doivent porter que sur les énoncés et les objets de discours. Or, en réalité, et dans Wikipedia en particulier, les participants recourent fréquemment à des raisonnements fallacieux (van Eemeren et Grootendorst, 1992) – arguments d'autorité, attaques à la personne, *etc.* – dans le but de décrédibiliser l'interlocuteur, relativiser la portée de ses dires, voire l'exclure du débat (cas des *trolls*). Il est donc nécessaire d'étendre la notion de conflit de sorte à inclure les conflits *personnels*, qui portent sur la légitimité d'une personne à participer au débat, et les conflits *méta-argumentatifs*, portant

1. Projet ANR n° ANR-08-CORD-004

sur le fait qu'un participant a respecté ou non les règles du débat instituées par la communauté.

2.2 Approche discursive du conflit

La mise en œuvre d'un modèle dialectique ci-dessus, qui plus est dans sa version étendue, est largement problématique tant elle nécessite de mobiliser des connaissances du domaine (étendues aux personnes et aux règles communautaires) et de techniques d'analyse de discours (van Eemeren *et al.*, 1993; Asher et Lascarides, 2003).

Sans toutefois abandonner le modèle dialectique, nous proposons d'exploiter des indices de surface : la *polarité* (connotation positive ou négative des énoncés) et *subjectivité* (mentions directe ou indirecte des locuteurs dans le discours). Si la connotation est un indice relativement faible dans le cadre général et ne suffit pas à marquer fiablement une opinion et encore moins un argument, cet indice prend tout son sens dans un contexte argumentatif. La mobilisation d'une connotation négative peut marquer les attaques tandis que les connotations positives peuvent marquent les concessions et les défenses. La subjectivité, quant à elle, est d'une part une marque d'implication du locuteur dans le discours et sert d'autre part au démarquage des thèses dans un conflit mixte (ce que tu dis vs. ce que je dis). Elle est en outre un indice essentiel au repérage de conflits personnels ou méta-argumentatifs.

L'extrait de la figure 1, tiré de la page de discussion de Wikipedia sur Pluton, illustre un conflit entre deux participants, *M* et *R*. Il est remarquable que chaque message porte une connotation négative (*neg*) et que les deux participants s'engagent personnellement en utilisant la première ou la seconde personne : *M* attaque *R* avec la seconde personne, et *R* se défend avec la première personne.

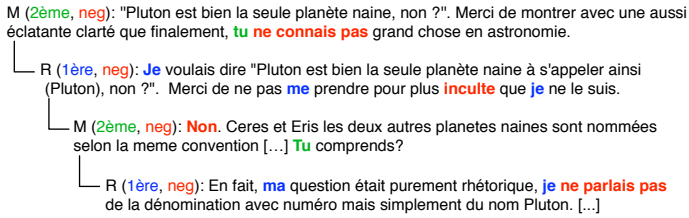


FIGURE 1 – Extrait d'un conflit alternant les marqueurs subjectifs dans un contexte négatif

Cet exemple nous incite à considérer une règle naïve de détection de conflit où le conflit est défini comme une situation négative dans lequelle les participants alternent les marques subjectives. Cette règle est implémentée et évaluée dans la section 4.

3 Mise en œuvre

Nous proposons d'explorer la pertinence des dimensions de subjectivité et de polarité pour la détection du conflit en deux temps : d'abord nous proposons une annotation de bas niveau des

énoncés des participants en termes de subjectivité et de polarité, ensuite une annotation de haut niveau afin de déterminer si un fil de discussion porte ou non un conflit.

3.1 Annotation de la subjectivité et de la polarité

L'annotation des traits de subjectivité et de polarité consiste à associer à chaque énoncé et à chaque message une liste de traits.

L'acquisition de la dimension de subjectivité peut être compliquée si on considère la subjectivité inhérente des adjectifs (comparer par exemple "absurde" et "incorrect"). Nous nous limitons alors à la présence de marques pronominales. Chaque pronom personnel (je, nous), posssessif (le mien), et déterminant possessif (mon, nôtre) de première personne (respectivement de deuxième personne) est annoté *subjective-1st* (respectivement *subjective-2nd*), un énoncé pouvant comporter plusieurs de ces marques.

La reconnaissance de la polarité d'un énoncé est un problème complexe car la polarité d'un mot change en fonction du co-texte (négation syntaxique, rôle syntaxique d'un mot, etc.) ou du contexte d'énonciation (point de vue de celui qui énonce). De nombreuses approches se sont attaquées au problème général dans le cadre de la détection d'opinion ou de sentiments (Pang *et al.*, 2002; Wilson *et al.*, 2009; Pak et Paroubek, 2011). Nous pouvons cependant supposer que dans le cadre du débat argumentatif, le problème est moins difficile. Par exemple l'énoncé "l'école élémentaire Gavin a été *condamnée* en Avril 2004" tiré de (Wilson *et al.*, 2009) illustre l'emploi d'un mot négatif qui exprime une opinion neutre dans le cadre général mais qui peut être considérée comme une *attaque* dans le débat. Le problème est alors plus simple mais non trivial, puisque la négation syntaxique intervient tout de même. Y compris dans le débat, "ce n'est pas absurde" ne peut être considéré comme négatif. Nous proposons alors deux méthodes d'annotation de la polarité, une méthode purement lexicale, et une méthode syntaxique qui permet de filtrer des cas comme "ce n'est pas absurde". Ces deux méthodes sont évaluées et comparées dans la section 4.

Analyse lexicale L'analyse lexicale ne repose que sur la polarité des mots sans considérer l'influence du co-texte, le problème se résumant à obtenir cette polarité. Il existe en Anglais des lexiques de polarités comme Wordnet-affect (Valitutti *et al.*, 2004), ou SentiWordnet (Baccianella *et al.*, 2010). En Français, on peut citer le lexique de (Mathieu, 2004) mais celui-ci est limité à 950 mots. Nous avons alors procédé à l'annotation manuelle en termes de polarités (positive, négative ou neutre) du lexique du Français fondamental qui comporte les 8000 lemmes les plus fréquents du Français (Gougenheim *et al.*, 1964). Ce lexique étant relativement limité en taille, nous l'avons étendu automatiquement en nous appuyant sur EuroWordnet-FR (Vossen, 1998), considérant que les hyponymes d'un lemme connoté l'étaient également. Etant donné le lexique, l'annotation d'un énoncé se limite à collecter les polarités des mots qu'il contient. L'analyse purement lexicale considère alors des énoncés du type "ce n'est pas absurde" comme des énoncés négatifs.

Analyse syntaxico-sémantique profonde L'analyse profonde s'appuie sur l'analyseur syntaxique LLP2 (Lopez, 2000) que nous avons utilisé dans plusieurs projets et dont nous souhaitions déterminer la capacité à s'adapter à du texte libre. La grammaire est une grammaire LTAG (Joshi

et Schabes, 1997), comportant 1500 arbres, ancrée avec le LEFFF (Sagot, 2010), un lexique d'environ 530 000 formes fléchies. Les dérivations partielles de l'analyseur LLP2 font l'objet de réécritures successives (Bedaride et Gardent, 2009) afin de produire une forme sémantique plus facile à manipuler (120 règles). La forme sémantique est ensuite annotée selon la polarité grâce à des règles qui s'appuient sur le lexique de polarités précédemment construit. Par exemple, un énoncé est annoté négativement s'il comporte un verbe négatif, un modifieur négatif (adverbe ou adjectif) ou un verbe positif nié syntaxiquement : "je n'aime pas" sera négatif car *aimer* est positif. Les règles les plus complexes s'appuient également sur la taxonomie EuroWordnet, par exemple si un énoncé contient un verbe hyponyme de "penser", nié syntaxiquement, dont la subordonnée est polarisée, l'énoncé est annoté selon l'inverse de la polarité de la subordonnée, typiquement "je ne pense pas que cela soit absurde" sera positif, car la subordonnée est négative et la principale niée syntaxiquement. Au total, l'annotation de la polarité contient 55 règles de ce type.

Qu'il s'agisse de la méthode lexicale ou syntaxique, un énoncé est annoté comme une liste de marques de subjectivité et de polarité. Un énoncé peut alors contenir *plusieurs* marques de même type, mais également des marques à la fois négatives ou positives. Un message est annoté en faisant l'union des marques des énoncés qui le compose. Par exemple avec l'annotation lexicale, le message "*Merci* [...] Maintenant que *vous* avez fait *votre critique* et répandu *vos insultes*, qu'attendez-vous pour nous *aider* à *améliorer* les articles en utilisant votre immense culture orthographique, historique et philosophique ?" sera annoté par : [positive, subjective-2nd, subjective-2nd, negative, subjective-2nd, negative, positive, positive].

3.2 Annotation des conflits

L'annotation des conflits consiste à déterminer, pour un fil de discussion, s'il contient ou non un conflit. Afin d'effectuer automatiquement cette décision, nous proposons de comparer deux types de règles qui s'appuient toutes deux sur les traits de subjectivité et de polarité acquis automatiquement par l'annotation lexicale ou syntaxique.

Le premier type de règle que nous désirons tester est une règle dialectique inspirée de l'exemple de la figure 1 qui s'appuie sur la structure hiérarchique des messages sur Wikipedia, similaire à la structure qu'on peut trouver dans une liste de diffusion. Un conflit est alors défini comme la présence de deux messages dans la hiérarchie, tous deux négatifs, et qui alternent les marques subjectives, c'est-à-dire un message s'appuyant sur la première personne, suivi plus loin d'un message s'appuyant sur la seconde personne ou vice versa. Un fil de discussion est annoté comme *conflict* s'il contient une telle séquence et *no_conflict* sinon. En vertu de sa dépendance à la structure locale d'un fil, nous référerons à cette règle comme la règle "locale".

Le second type de règle est une règle apprise à partir d'un corpus annoté et d'un classifieur de type arbre de décision (C4.5). Deux annotateurs ont annoté manuellement les 153 fils de discussion de la page de discussion associée à l'article Astrologie, un des articles appartenant à la catégorie Wikipedia *article sujet à controverses*. La page de discussion Astrologie contient 982 messages et rassemble 88 auteurs. Chaque fil est annoté selon la présence ou l'absence de conflit. L'accord inter-annotateur obtenu par la méthode du kappa est plutôt bon $\kappa = 0.644$ ($p < 0.0001$). Après adjudication ($\kappa = 1.0$), la page de discussion Astrologie comporte environ 20% de fils conflictuels (30 fils conflictuels pour 123 non conflictuels). Les traits d'apprentissage sont issus de l'annotation automatique de la polarité et de la subjectivité : un fil de discussion est

représenté comme un vecteur à cinq dimensions $\langle n, p, s_1, s_2, t\rangle$ où n est le taux de négativité du fil, p le taux de positivité, s_1 et s_2 respectivement les taux de marques subjectives de première et seconde personne et t la taille du fil en termes de nombre de messages. Les taux sont calculés comme le rapport du nombre d'occurrence d'un marqueur dans le fil sur le nombre de mots du fil, et sont en conséquence très inférieurs à 1. La méthode statistique construit alors un arbre de décision pour les classes *conflict* ou *no_conflict* à partir de ces traits. L'arbre de décision ainsi obtenu s'appuie seulement sur la taille du fil et le taux de subjectivité seconde personne : si un fil contient strictement plus de 4 messages et que son taux s_2 est supérieur à 0.00274, alors le fil est considéré comme conflictuel. En vertu de son application sur l'ensemble d'un fil, nous référerons à cette règle comme la règle "globale".

Nous évaluons dans la section suivante les types de règles d'annotation des conflits ainsi que la pertinence de l'analyse syntaxique par rapport à une simple analyse lexicale.

4 Evaluation et résultats

Nous avons choisi de tester les différentes méthodes et règles sur les pages de discussion associées aux articles Communisme et Jésus de Nazareth, deux articles controversés. Les deux pages réunies comportent 320 fils de discussion, 2659 messages et 256 auteurs. Seule l'annotation de la page Communisme a été faite en double, et elle fournit un bon accord inter-annotateur de $\kappa = 0.79$ ($p < 0.0001$). Les deux pages réunies comportent davantage de conflits que la page Astrologie puisqu'environ 38% des fils sont conflictuels (122 fils conflictuels pour 198 fils non conflictuels). Le corpus de test est annoté automatiquement en termes de polarité et de subjectivité selon la méthode lexicale et la méthode syntaxique, et on évalue la capacité des différentes règles d'annotation des conflits à reproduire correctement l'annotation manuelle.

Nous résumons l'ensemble des résultats dans la table 1 et donnons pour chaque combinaison de techniques (locale ou globale pour l'annotation des conflits, lexicale ou syntaxique pour l'annotation des polarités), le kappa, la précision et le rappel pour chaque catégorie (la présence de conflit C ou son absence $\overline{C}$). La combinaison globale/syntaxique n'est pas présentée étant donné qu'elle fournit des résultats similaires à la combinaison globale/lexical en vertu du fait que la règle globale ne repose pas sur la polarité.

Méthode	Kappa	Catégories	Rappel	Précision	F-Mesure
locale/lexicale	0.70 ± 0.08	C	77.8	84.0	80.6
		$\overline{C}$	90.9	87.0	88.9
locale/syntaxique	0.66 ± 0.08	C	79.5	78.9	79.2
		$\overline{C}$	86.9	87.3	87.1
globale/lexicale	0.72 ± 0.08	C	77.9	87.2	82.3
		$\overline{C}$	92.9	87.2	90.0

TABLE 1 – Résultats des différentes techniques sur Communisme et Jésus

Le premier axe de comparaison, entre la règle locale et la règle globale montre que la règle globale produit des résultats légèrement meilleurs. Le second axe de comparaison entre l'approche lexicale ou syntaxique suggère que la méthode syntaxique n'apporte rien : si elle permet de gagner environ 2 points en rappel pour la présence de conflits, elle en perd 7 en précision. Notre

étude ne permet pas de conclure sur l'importance des polarités dans l'établissement d'un conflit étant donné que les trois méthodes produisent des résultats qui ne sont pas significativement différents et que nous ignorons la qualité de l'annotation *a priori* des polarités. En revanche, cette étude met en avant l'intérêt de la subjectivité. Il s'agit en effet d'un problème beaucoup moins complexe que celui des polarités mais qui offre pourtant de très bons résultats. L'extrême simplicité de la règle globale (elle peut se limiter au repérage des pronoms de seconde personne) suggère alors que la subjectivité est une piste de recherche qui doit être privilégiée pour la détection automatique des conflits.

5 Discussion et perspectives

Nous avons évalué deux règles de détection de conflit dans des fils de discussion, s'appuyant sur les mêmes traits de surface, subjectivité et polarité, l'une fonctionnant au niveau local (deux contributions dans le même fil, liées dans un rapport proposition/réplique), l'autre au niveau global (l'ensemble des traits des contributions est rapporté au niveau du fil). L'évaluation a montré que malgré leur différence de nature ces règles obtiennent des résultats comparables et de niveau tout à fait correct. De plus, la règle globale obtient un bon score malgré le fait qu'elle *n'exploite pas la polarité* et que d'autre part la longueur du fil et le taux de subjectivité de seconde personne sont des indices *suffisants*. Selon cette règle, il *suffit* de prendre en compte le nombre de contributions (un conflit se règle rarement en moins de 5 interventions, car il en faut au minimum 3 pour se déclarer) et une proportion minimale de marques subjectives de seconde personne, c'est-à-dire la présence durable de 2 voix, d'un « *dialogisme diffus* » (car la structure des contributions n'est pas prise en compte). Nous estimons que la raison principale pour laquelle la méthode globale est si efficace tient à l'activité épistémique de la communauté Wikipedia : la page de discussion est un espace de parole finalisé dédié à la tâche de conception. Or dans ce cas, l'usage de la seconde personne est nettement moins fréquent que dans des discussions non finalisées.

Cette expérience a montré que la piste de la subjectivité est une piste intéressante à exploiter pour la détection des conflits. Nous pouvons tirer parti de la simplicité de la règle globale – indépendante du domaine et facilement transposable à d'autres langues – pour évaluer si elle reste efficace sur des discussions du Wikipedia anglophone. Par ailleurs, il est possible d'étudier une autre facette de la subjectivité grâce aux termes marqués subjectivement ("absurde" versus "incorrect") et de vérifier alors si leur présence ou absence contribue à améliorer la détection des conflits. Enfin, il peut être intéressant de tester la place de la subjectivité dans d'autres types d'interactions moins finalisées, par exemple dans des forums généralistes.

Références

ASHER, N. et LASCARIDES, A. (2003). *Logics of Conversation*. Cambridge University Press, Cambridge.

BACCIANELLA, S., ESULI, A. et SEBASTIANI, F. (2010). Sentiwordnet 3.0 : An enhanced lexical resource for sentiment analysis and opinion mining. *In* CALZOLARI, N., CHOUKRI, K., MAEGAARD, B., MARIANI, J., ODIJK, J., PIPERIDIS, S., ROSNER, M. et TAPIAS, D., éditeurs : *Proceedings of the*

Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

BAKER, M., DÉTIENNE, F., LUND, K. et SÉJOURNÉ, A. (2009). *Méthodologies d'analyse de situations coopératives de conception : le corpus MOSAIC*, chapitre Etude des profils interactifs dans la conception collective en architecture. Presses Universitaires de Nancy.

BARCELLINI, F., DÉTIENNE, F. et BURKHARDT, J.-M. (2008). User and developer mediation in an open source software community : boundary spanning through cross participation in online discussions. *International Journal of Human Computer Studies*, 66(7):558–570.

BARTH, E. M. et KRABBE, E. C. (1982). *From Axiom to Dialogue*. de Gruyter, Berlin.

BEDARIDE, P. et GARDENT, C. (2009). Semantic normalisation : a framework and an experiment. *In 8th International Workshop on Computational Semantics*, Tilburg, Netherland.

FRÉARD, D., DENIS, A., DÉTIENNE, F., BAKER, M., QUIGNARD, M. et BARCELLINI, F. (2010). The role of argumentation in online epistemic communities : the anatomy of a conflict in wikipedia. *In Proceedings of European Conference of Cognitive Ergonomics*.

GOUGENHEIM, G., MICHEA, R., RIVENC, P. et SAUVAGEOT, A. (1964). *L'élaboration du français fondamental : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Didier, Paris.

JOSHI, A. et SCHABES, Y. (1997). Tree-adjoining grammars. *In* ROZENBERG, G. et SALOMAA, A., éditeurs : *Handbook of Formal Languages*, volume 3, pages 69 – 124. Springer, Berlin, New York.

LOPEZ, P. (2000). Extended partial parsing for lexicalized tree grammars. *In Proc. of the Sixth International Workshop on Parsing Technologies (IWPT 2000)*, pages 159–170.

MACKENZIE, J. D. (1985). No Logic before Friday. *Synthese*, 63:329–341.

MATHIEU, Y. Y. (2004). A semantic lexicon for emotions and feelings. *In* YAN QU, James G. Shanahan, J. W., éditeur : *American Association for Artificial Intelligence (AAAI) Spring symposium on Exploring Attitude and Affect in Text*, AAAI Technical Report Series, AAAI Press, Menlo Park, USA, pages 89–93.

PAK, A. et PAROUBEK, P. (2011). Classification en polarité de sentiments avec une représentation textuelle à base de sous-graphes d'arbres de dépendance. *In Proceedings of TALN 2011*, Montpellier.

PANG, B., LEE, L. et VAITHYANATHAN, S. (2002). Thumbs up ? Sentiment classification using machine learning techniques. *In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86.

SAGOT, B. (2010). The LEFFF, a freely available and large-coverage morphological and syntactic lexicon for french. *In Proceedings of LREC 2010*, La Valette, Malte.

VALITUTTI, A., STRAPPARAVA, C. et STOCK, O. (2004). Developing affective lexical resources. *PsychNology Journal*, pages 61–83.

van EEMEREN, F. H. et GROOTENDORST, R. (1992). *Communication, Argumentation, Fallacies*. Erlbaum, Mahwah, N. J.

van EEMEREN, F. H., GROOTENDORST, R., JACKSON, S. et JACOBS, S. (1993). *Reconstructing Argumentative Discourse*. Studies in Rhetoric and Communication. University of Alabama Press, London.

VOSSEN, P. (1998). *EuroWordNet : a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.

WILSON, T., WIEBE, J. et HOFFMANN, P. (2009). Recognizing contextual polarity : An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.

Quel est l'apport de la détection d'entités nommées pour l'extraction d'information en domaine restreint ?

Camille Dutrey[1, 2, 3], Chloé Clavel[1], Sophie Rosset[2], Ioana Vasilescu[2], Martine Adda-Decker[2, 3]

(1) EDF R&D, 1 avenue du Général de Gaulle 92141 Clamart

(2) LIMSI-CNRS, rue John von Neumann 91403 Orsay

(3) LPP, 19 rue des Bernardins 75005 Paris

camille.dutrey@{edf,limsi}.fr, chloe.clavel@edf.fr, sophie.rosset@limsi.fr, vasilescu@limsi.fr, madda@{univ-paris3,limsi}.fr

RÉSUMÉ

Les travaux liés à la définition et à la reconnaissance des entités nommées sont généralement envisagés en domaine ouvert, à travers la conception de catégories génériques (noms de personnes, de lieux, etc.) et leur application à des données textuelles issues de la presse (orale comme écrite). Par ailleurs, la fouille des données issues de centres d'appel est stratégique pour une entreprise comme EDF, compte tenu du rôle crucial joué par l'opinion pour les applications marketing, ce qui passe par la définition d'entités d'intérêt propres au domaine. Nous comparons les deux types de modèles d'entités – génériques et spécifiques à un domaine précis – afin d'observer leurs points de recouvrement, via l'annotation manuelle d'un corpus de conversations en centres d'appel. Nous souhaitons ainsi étudier l'apport d'une détection en entités nommées génériques pour l'extraction d'information métier en domaine restreint.

ABSTRACT

What is the contribution of named entities detection for information extraction in restricted domain ?

In the framework of general domain dialog corpora a particular focus is dedicated to Named Entities definition and recognition, which are mostly very generic (personal names, locations, etc.). Moreover, call-centre data mining is strategic for a company like EDF, the public opinion analysis playing a significant role in EDF services quality evaluation and for marketing applications. In this purpose a domain dependant definition of entities of interest is essential. In this primary work we compare two types of entities models (generic and specific to the domain) in order to observe their respective coverage. We annotated manually a sub-corpus extracted from a large corpus of oral dialogs recorded in an EDF call-centre. The respective proportion of generic vs domain-specific Named Entities is then estimated. Impact for future work on building EDF domain-specific entities models is discussed.

MOTS-CLÉS : entités nommées, concepts métier, extraction d'information, données conversationnelles, annotation.

KEYWORDS: named entities, business concept, information extraction, conversational data, annotation.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 359–366,
Grenoble, 4 au 8 juin 2012.

1 Introduction

La fouille des données issues de centres d'appel est stratégique pour EDF, compte tenu du rôle crucial joué par l'opinion pour les applications marketing ; elle participe d'une amélioration de la connaissance du client et de ce fait du développement de sa fidélité vis à vis de l'entreprise. Il est de plus nécessaire de mettre en place des méthodes adaptées pour un traitement automatisé afin d'extraire et d'organiser efficacement le contenu informationnel de ces interactions téléphoniques.

L'extraction d'information sur des transcriptions orales est généralement abordée du point de vue de la détection d'entités nommées (EN) et de la recherche d'information, avec des campagnes d'évaluation comme ESTER2 (ESTER2, 2008; Galliano *et al.*, 2009). Ces campagnes sont toutefois principalement basées sur des données de type bulletins d'information et abordent peu la question de l'extraction d'information d'une part sur des conversations téléphoniques, au sein desquelles les spécificités de la parole spontanée sont plus fréquentes, et d'autre part sur des données en domaine restreint. On peut tout de même citer le projet Luna [1], axé sur la compréhension temps-réel de la parole spontanée, et notamment sur la détection de thèmes dans un contexte de dialogue homme-machine (corpus MEDIA, (Bonneau-Maynard *et al.*, 2006)).

Notre objectif est d'explorer la possible utilisation sur des données en domaine restreint de typologies d'EN élaborées pour des données en domaine ouvert, en se basant sur une comparaison entre entités génériques et entités spécialisées. Nous nous focalisons dans cette étude sur des transcriptions manuelles de conversations en centres d'appel. Nous avons pour cela annoté manuellement un corpus en entités génériques structurées définies dans (Grouin *et al.*, 2011) pour le projet Quaero ainsi qu'en entités métier définies par EDF. [2]

Après un aperçu des travaux sur la définition des EN ainsi qu'une description des modèles retenus pour cette étude (section 2), nous décrivons les expériences d'annotation que nous avons menées autour de la comparaison d'entités (section 3). Nous présentons ensuite les résultats de cette étude (section 4) puis nos perspectives de recherche (section 5).

2 Entités nommées génériques et entités d'intérêt spécifiques

2.1 Positionnement

Les EN sont traditionnellement définies en référence aux trois classes développées pour la tâche de reconnaissance d'entités de la conférence MUC-6 [3] : *ENAMEX* (noms de personnes, de lieux et d'organisations), *TIMEX* (expressions temporelles) et *NUMEX* (expressions numériques de pourcentages et de monnaies). Cette première définition a depuis été élargie : ainsi, les EN sont présentées comme suit dans (ESTER2, 2008) : « les EN sont des types particuliers d'unités lexicales (groupes de mots) qui font référence à une entité du monde concret dans certains domaines spécifiques [...] et qui ont un nom (typiquement un nom propre ou un acronyme) ». Cette définition a servi de base à l'élaboration des entités structurées du projet Quaero (Grouin *et al.*,

1. `http://www.ist-luna.eu/project_description.htm`
2. (Danesi et Clavel, 2010) ont abordé la problématique de la détection d'entités métier EDF à partir de transcriptions automatiques de conversations en centres d'appel, ce travail s'appuyant sur les transcriptions manuelles du même corpus. La question de l'impact des erreurs de reconnaissance automatique de la parole n'est pas abordée dans notre étude.
3. `http://www.cs.nyu.edu/cs/faculty/grishman/NEtask20.book_1.html`

2011; Rosset *et al.*, 2011), qui propose d'étendre leur portée à des expressions ne comportant pas de noms propres. (Sekine *et al.*, 2002) proposent une typologie hiérarchique et étendue avec près de 200 catégories. Les trois classes décrites supra ont été enrichies d'autres types d'entités, comme les *produits* (Sekine et Nobata, 2004) ou les *fonctions* (Galliano *et al.*, 2009). La notion recouvrant une réalité sémantico-lexicale de plus en plus large, l'appellation même d'« entité nommée » paraît parfois restrictive. La campagne ESTER2, du fait de l'intégration des *temps* et *montants*, parle d'« entités spécifiques ». Dans cette étude, nous préfèrerons le terme « entités d'intérêt » à « entités nommées ».

(Ehrmann, 2008) analyse la problématique des EN du point de vue théorique des difficultés définitoires et catégorielles, proposant la définition suivante : « étant donnés un modèle applicatif et un corpus, on appelle EN toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus ». Cette définition dépendante d'un corpus fait écho aux travaux de (Boufaden, 2004), celle-ci étudiant l'adéquation d'EN traditionnelles à un corpus spécialisé dans un cadre d'extraction d'information sur des transcriptions de conversations téléphoniques. Cette étude met en avant la différence d'actualisation lexicale entre les EN classiques et les entités d'intérêt en domaine restreint. Nous nous sommes penchés sur l'étude d'une telle approche contrastive des EN dites génériques – dans la mesure où leur définition tend vers une extension sémantique indépendante d'un corpus spécifique et d'un domaine précis – et des entités propres à un domaine en confrontant le modèle Quaero à un modèle issu du domaine du corpus d'étude, la relation entre EDF et ses clients.

2.2 Entités génériques : le modèle Quaero

Le modèle d'entités du projet Quaero, défini dans (Grouin *et al.*, 2011), se différencie des définitions classiques par deux aspects : *i)* les entités sont étendues à des expressions ne contenant pas de noms propres et *ii)* les entités sont structurées grâce à des sous-types et composants [4]. Les différents types et sous-types du modèle, décrits dans (Rosset *et al.*, 2011), sont présentés dans la table 1. Les composants décrits dans le modèle n'ont pas été utilisés pour notre étude.

Types	**Sous-types et/ou Exemples**
Personne	individuelle, collective (*ex. Bertrand Delanoë*)
Fonction	individuelle, collective (*ex. le maire de Paris*)
Organisation	entreprise, administration (*ex. le parti socialiste*)
Localisation	administrative, physique, voie, bâtiment, adresse (*ex. Paris*)
Production humaine	marque, œuvre, production médiatique, produit financier, logiciel, prix, ligne de transport, doctrine, loi/accord (*ex. socialisme*)
Quantité	(*ex. un car de CRS*)
Point dans le temps	date absolue/relative, heure absolue/relative (*ex. le second Empire*)
Événement	(*ex. la coupe du monde*)

TABLE 1 – Entités nommées structurées du projet Quaero

4. Par exemple, l'entité « Bertrand Delanoë » sera typée *personne individuelle* et chacun de ses composants « Bertrand » et « Delanoë » sera respectivement sous-typé *prénom* et *nom*.

2.3 Entités d'intérêt métier : le modèle EDF

Les entités définies à l'EDF correspondent à des concepts métier identifiés par des experts du domaine. Elles ne relèvent pas d'une terminologie globale à l'échelle de l'entreprise, comme le serait une ontologie, mais constituent une multitude de modèles élaborés en fonction *i)* d'un besoin applicatif (formation des conseillers clientèle, analyse marketing...) et *ii)* d'un type de données (champs de commentaires rédigés par le conseiller suite à un appel, enquêtes de satisfaction...). De plus, tout comme les entités Quaero élargissent le champ des EN traditionnelles, les entités d'intérêt EDF recouvrent une réalité lexicale de taille et de composition variables au travers de segments correspondant aux expressions caractérisant les concepts métier. Nous avons sélectionné un modèle d'entités d'intérêt très spécifique élaboré afin d'identifier les motifs des réclamations, des problèmes techniques aux problèmes de contact (présenté en table 2).

Concepts	**Sous-Concepts et/ou Exemples**
Facturation	Index (*ex. facture estimée*) ; Relèvement/Redressement (*ex. facture rectificative*) ; Recouvrement/Relance (*ex. relance pour impayés*) ; Paiement (*ex. mensualisation*)
Incident technique	Assurance (*ex. direct assurance*) ; Panne (*ex. l'ordinateur a grillé*) ; Sinistre (*ex. tempête*) ; Coupure/Surtension (*ex. coupure de courant*)
Intervention technique	Relève (*ex. l'agent de relève*) ; Autre hors relève (*ex. dépannage*) ; Frais d'intervention (*ex. frais de dédit*)
Contact	Attente (*ex. veuillez patienter*) ; Accueil
Traitement de la demande	Délai ; Échanges client-agent
Prix	(*ex. moins cher*)
Contrat/Tarif	(*ex. Jours tempo*)
Vente directe	Offres EDF (*ex. Suivi conso*)
Agence en ligne	(*ex. site Internet EDF bleu ciel*)
Concurrence	(*ex. Powéo*)
Autre	Entités d'intérêt EDF ne relevant pas de la thématique réclamation

TABLE 2 – Entités d'intérêt EDF orientées « réclamation »

3 Expériences

Afin de confronter les deux modèles d'entités décrits en section 2 sur un corpus lié au modèle spécifique EDF, nous avons sélectionné un corpus composé de conversations issues de centres d'appel (section 3.1) que nous avons annoté manuellement en entités Quaero et EDF (section 3.2).

3.1 Le corpus CallSurf

Dans le cadre du projet Infom@gic – CallSurf (Garnier-Rizet *et al.*, 2008), EDF a effectué une campagne d'enregistrement au sein d'un de ses centres d'appel : dix agents volontaires ont été enregistrés durant quatre mois lors de leurs appels avec des clients professionnels. Cette campagne a permis de constituer un large corpus de parole spontanée (5 755 appels, 620 heures de conversations) dont les caractéristiques sont détaillées dans (Danesi et Clavel, 2010). Les auteures ont relevé les spécificités propres à l'oral de ce corpus (par exemple les phénomènes liés au travail de mise en mots et les effets disfluents).

Notre corpus de développement, le corpus Cal10, est composé de transcriptions manuelles produites à l'aide de `Transcriber` (Barras *et al.*, 1998). Le corpus Cal10 est composé de 90 appels téléphoniques soit environ 10h de signal en langue française ; ses principales caractéristiques sont présentées en table 3.

Nous en avons extrait – via un tirage aléatoire – un sous-corpus Cal10a de 10 conversations, soit environ 10 000 mots répartis sur 2 250 tours de parole.

Nombre d'appels	90
Nombre de tours de parole	12 825 (moy. 142,5 / appel)
Nombre de locuteurs	moy. 2,3 / appel
Nombre de mots	100 677 (moy. 1 118,6 / appel – 3,6 / sec.)
Durée des tours de parole	08 : 22 : 16 (moy. 00 : 00 : 02)
Durée totale retranscrite	08 : 51 : 20

TABLE 3 – Caractéristiques générales du corpus Cal10

3.2 Annotation manuelle en entités d'intérêt

Deux annotateurs ont respectivement annoté la totalité du corpus Cal10a en entités Quaero et en entités EDF, avec la plateforme d'annotation `Glozz` (Widlöcher et Mathet, 2009) [5]. L'objectif de cette étude consistant en une comparaison de la couverture d'un modèle d'entités génériques et d'un modèle d'entités spécifiques sur un corpus en domaine restreint, nous avons créé deux modèles d'annotation : l'un pour les entités Quaero et l'autre pour les entités d'intérêt EDF. La figure 1 présente un aperçu de l'interface `Glozz`. Le texte annoté comporte l'entité EDF « il y a pas les heures creuses » (catégorie *contrat*) recouvrant l'entité Quaero « heures creuses » (catégorie *production humaine*).

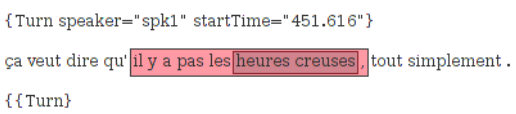


FIGURE 1 – Aperçu de Glozz cadrant un tour de parole annoté

4 Résultats

Suite à l'annotation manuelle du corpus, 368 entités Quaero et 273 entités EDF ont été identifiées dans notre corpus, soit en moyenne 0,3 entités Quaero et 0,2 entités EDF par tour de parole équivalent à une entité Quaero tous les 31 mots environ et une entité EDF tous les 42 mots environ. La figure 2 présente la répartition des entités EDF et Quaero (les entités absentes du corpus ne sont pas apparentes), répartition inégale et ce au sein des deux modèles.

5. `Glozz` permet l'annotation d'unités de granularité variable (caractère, mot, phrase, paragraphe, etc.), de relations entre ces unités et de schémas, constructions complexes de relations.

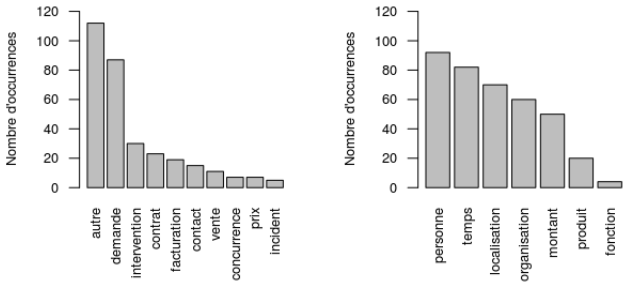


FIGURE 2 – Répartitions des entités EDF et Quaero par catégorie

Concernant les entités d'intérêt EDF, la catégorie *autre* est la plus présente (41% des entités sont de ce type) : cette catégorie concerne des entités d'intérêt pour EDF ne relèvant pas du modèle choisi ; l'on observe ainsi que la thématique de la réclamation clients ne représente qu'une partie des termes métier présents dans le corpus. Au sein du modèle, 32% des entités concernent un *traitement de la demande* ; les autres types d'entités sont relativement peu présents dans notre corpus. Concernant les entités Quaero, la répartition est davantage homogène : environ 25% des entités sont de type *personne*, ce qui est logique compte tenu de la teneur informationnelle du corpus (l'agent est tenu de se présenter ainsi que l'entreprise, et les références client – y compris son nom et le nom de son entreprise – sont fréquemment rappelées) [6]. La forte présence d'entités de type *localisation* (19%) participe également de cette teneur informationnelle. La table 4 présente des informations concernant le chevauchement des deux modèles : le nombre d'entités relevant des deux modèles est relativement faible (respectivement 23,8% des entités EDF et 17,7% des entités Quaero). Ce résultat relève en fait d'un double aspect : certaines entités Quaero sont peu représentées dans notre corpus (*fonction*, *événement*), en raison d'une différence thématique justifiée, alors que d'autres parmi les plus fréquentes (*personne*) ne sont pas prises en compte par le modèle EDF.

Entités EDF ayant au moins un mot en commun avec une entité Quaero	23,8 %
Entités Quaero ayant au moins un mot en commun avec une entité EDF	17,7 %
Entités EDF n'ayant aucun recouvrement avec le modèle Quaero	76,2 %
Entités Quaero n'ayant aucun recouvrement avec le modèle EDF	82,3 %

TABLE 4 – Recouvrement des modèles d'entités Quaero et EDF sur le corpus EDF

Le corpus ayant servi pour l'annotation comporte 11 429 mots, dont 2 413 ont été annotés (soit environ 21% de la totalité du corpus).

Les tables 5 et 6 présentent la couverture des deux modèles d'entités ainsi que leur intersection eu égard à l'ensemble des mots du corpus (table 5) et à l'ensemble des mots annotés dans le corpus (table 6). Bien que davantage d'entités Quaero aient été identifiées, elles concernent seulement 6% des mots annotés ; de plus, elles sont en moyenne composées de 1,8 mot (contre

6. Les données ayant été anonymisées, en accord avec les exigences de la CNIL, l'annotation en noms de personnes a été effectuée grâce à la présence de labels de remplacement tels *NOMPERS*, *NOMSOCIETE*, etc.

6,9 pour les entités EDF). Ces chiffres sont cohérents avec la définition des deux modèles, les entités d'intérêt EDF identifiant des expressions allant au delà du syntagme nominal.

Modèle	Taux de mots annotés
EDF	15 %
Quaero	4,5 %
Intersection	1,5 %
Union	*21 %*

TABLE 5 – Couverture des entités relatives à l'ensemble des mots du corpus

Modèle	Taux de mots annotés
EDF	71,86 %
Quaero	21,59 %
Intersection	6,55 %
Union	*100 %*

TABLE 6 – Couverture des entités relatives à l'ensemble des mots annotés du corpus

Une comparaison des définitions des modèles permet d'établir un rapprochement sémantique entre certaines classes ; nous avons cherché à savoir si ces classes s'actualisaient de la même manière dans notre corpus. En prenant le modèle Quaero comme référence, nous observons notamment que 15,3% des mots annotés *point dans le temps* étaient également annotés *demande/délai* dans le modèle EDF. De même, 23,6% des mots annotés *quantité* relèvent des types EDF *vente* ou *prix*. Enfin, 48,6% des mots annotés *production humaine* relèvent également du type EDF *contrat*. En revanche, seul 1,6% des mots relevant du modèle Quaero (tous types confondus) fait partie de la classe *autre* du modèle EDF.

5 Conclusions & Perspectives

Nous avons mis en avant la présence partielle d'un modèle d'entités génériques comme Quaero dès lors qu'il est appliqué à un corpus de textes issus d'un domaine précis, eu égard aux entités d'intérêt métier pour une entreprise.

Notre étude laisse toutefois apparaître un lien non négligeable entre les deux modèles considérés, grâce au fort recouvrement entre certains types Quaero et EDF : nous souhaitons approfondir cet aspect en étudiant les segments textuels identifiés dans chaque modèle, par types et sous-types. Ces équivalences sont encourageantes pour la suite de nos travaux : nous souhaitons en effet nous pencher sur la détection d'EN en soutient à l'extraction d'entités métier.

Nous continuons cette démarche comparative doublement enrichissante pour affiner les thématiques propres au domaine mais aussi pour élaborer des modèles facilement adaptables aux différents volets du corpus analysé ou à d'autres corpus conversationnels, en mesurant plus finement l'écart observé entre modèle générique et spécifique.

Par ailleurs, un volet essentiel que nous souhaitons étudier est l'impact des erreurs de reconnaissance automatique de la parole sur la détection des entités d'intérêt EDF, dans la continuité des travaux initiés par (Danesi et Clavel, 2010) : un travail de mise en rapport des résultats de la détection d'entités sur les transcriptions manuelles et automatiques représente un objectif crucial du travail futur. Par exemple, de nombreuses erreurs d'analyse concernent les noms propres (personnes, lieux...), comme en témoigne l'exemple suivant : « j'habite à *Beauvoisin* » analysé « j'habite à *vos voisins* ». Nous souhaitons étudier les méthodes existantes de reconnaissance des EN afin de pallier ce type d'erreurs.

Références

BARRAS, C., GEOFFROIS, E., WU, Z. et LIBERMAN, M. (1998). Transcriber : a Free Tool for Segmenting, Labeling and Transcribing Speech. *In Proceedings of LREC'98*.

BONNEAU-MAYNARD, H., AYACHE, C., BECHET, F., DENIS, A., KUHN, A., LEFEVRE, F., MOSTEFA, D., QUIGNARD, M., ROSSET, S., SERVAN, C. et VILLANEAU, J. (2006). Results of the French Evalda-Media evaluation campaign for literal understanding. *In Proceedings of LREC'06*.

BOUFADEN, N. (2004). *Extraction d'information à partir de transcriptions de conversations téléphoniques spécialisées*. Thèse de doctorat, Université de Montréal.

DANESI, C. et CLAVEL, C. (2010). Impact of Spontaneous Speech Features on Business Concept Detection : a Study of Call-Centre Data. *In Proceedings of the ACM Multimedia SSCS Workshop*.

EHRMANN, M. (2008). *Les Entités Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*. Thèse de doctorat, Université Paris 7 – Denis Diderot.

ESTER2 (2008). *ESTER2, Convention d'annotation en Entités Nommées, Dates, heures et montants*.

GALLIANO, S., GRAVIER, G. et CHAUBARD, L. (2009). The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts. *In Proceedings of Interspeech'09*.

GARNIER-RIZET, M., ADDA, G., CAILLIAU, F., GAUVAIN, J.-L., GUILLEMIN-LANNE, S. et LAMEL, L. (2008). Callsurf : Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content. *In Proceedings of LREC'08*.

GROUIN, C., ROSSET, S., ZWEIGENBAUM, P., FORT, K. et QUINTARD, L. (2011). Proposal for an Extension of Traditional Named Entities : From Guidelines to Evaluation, an Overview. *In Proceedings of Linguistic Annotation Workshop*.

ROSSET, S., GROUIN, C. et ZWEIGENBAUM, P. (2011). *Entités nommées structurées : guide d'annotation Quaero*. LIMSI-CNRS.

SEKINE, S. et NOBATA, C. (2004). Definition, dictionaries and tagger for Extended Named Entity Hierarchy. *In Proceedings of LREC'04*.

SEKINE, S., SUDO, K. et NOBATA, C. (2002). Extended Named Entity Hierarchy. *In Proceedings of LREC'02*.

WIDLÖCHER, A. et MATHET, Y. (2009). La plateforme Glozz : environnement d'annotation et d'exploration de corpus. *In Actes de TALN'09*.

Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des sentiments

Egle Eensoo Mathieu Valette
INALCO, ERTIM, 2 rue de Lille, 75343 Paris Cedex 07
{prenom.nom}@inalco.fr

RÉSUMÉ

Depuis une dizaine d'années, le TAL s'intéresse à la subjectivité, notamment dans la perspective d'applications telles que la fouille d'opinion et l'analyse des sentiments. Or, la linguistique de corpus outillée par des méthodes textométriques a souvent abordé la question de la subjectivité dans les textes. Notre objectif est de montrer d'une part, ce que pourrait apporter à l'analyse des sentiments l'analyse textométrique et d'autre part, comment mutualiser les avantages d'une association entre celle-ci et une méthode de classification automatique basée sur l'apprentissage supervisé. En nous appuyant sur un corpus de témoignages issus de forums de discussion, nous montrerons que la prise en compte de critères sélectionnés suivant une analyse textométrique permet d'obtenir des résultats de classification satisfaisants par rapport à une vision purement lexicale.

ABSTRACT

About the application of textometric methods for developing classification criteria in Sentiment analysis

Over the last ten years, NLP has contributed to applied research on subjectivity, especially in applications such as Opinion mining and Sentiment analysis. However, corpus linguistics and textometry have often addressed the issue of subjectivity in text. Our purpose is to show, first, what textometric analysis could bring to sentiment analysis, and second, the benefits of pooling linguistic/textometric analysis and automatic classification methods based on supervised learning. By processing a corpus of posts from fora, we will show that the building of criteria from a textometric analysis could improve classification results, compared to a purely lexical approach.

MOTS-CLÉS : linguistique de corpus, textométrie, analyse de sentiments, classification automatique supervisée.

KEYWORDS : corpus linguistics, textometry, sentiment analysis, supervised learning.

1 Introduction

L'extraction d'information subjective (Pang et Lee, 2008) est depuis une dizaine d'années un vaste domaine d'applications en croissance régulière. Malgré quelques travaux (par exemple Vernier, 2009 ; Béchet *et al.*, 2008) le savoir-faire linguistique y est peu sollicité. La subjectivité a pourtant fait l'objet de nombreux travaux linguistiques, dans différents courants théoriques – linguistique de l'énonciation, analyse de discours, sémantique des textes. La textométrie, aux confins de la linguistique générale et du TAL, propose par ailleurs une archive intéressante de travaux sur corpus susceptibles d'intéresser les

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 367–374,
Grenoble, 4 au 8 juin 2012.

applications d'analyse des sentiments (AS). On songe, dans le discours politique, aux travaux de (Salem, 1993), dans les sondages d'opinion, à (Lebart et Salem, 1988) ou dans la littérature, à (Brunet, 2009).

On souhaiterait ici susciter une rencontre entre d'une part le TAL ingénierique et ses applications et, d'autre part, la textométrie, à partir des constats suivants : (1) l'évaluation des méthodes en TAL repose sur un ensemble restreint de mesures (telles que précision, rappel, f-mesure) qui ont pour but de vérifier la qualité des méthodes plus que de valider des hypothèses et des méthodologies linguistiques. Leurs résultats ne nécessitent pas d'interprétation pour être valides ; (2) la textométrie relève, au contraire, d'une tradition descriptive. Elle se focalise sur l'interprétation des résultats de traitements statistiques, davantage que sur l'amélioration desdits traitements. À la différence du TAL, l'évaluation n'est pas un enjeu en textométrie[1].

Notre projet repose sur l'hypothèse que la textométrie, discipline descriptive, est à même d'apporter des solutions méthodologiques pour les applications généralement dévolues au TAL. Nous tenterons d'évaluer l'apport potentiel de la conjonction d'une analyse textométrique et de méthodes d'apprentissage pour une application d'AS.

2 État de l'art

La catégorisation des textes, qu'elle soit bipolaire (positif/négatif) ou multiclasse (mauvais/bon/excellent), est l'application principale en extraction d'information subjective. Elle peut être réalisée au moyen d'algorithmes *ad hoc* (Turney, 2002 ; Snyder et Barzilay, 2007) ou des méthodes d'apprentissage comme Naive Bayes, Support Vector Machines, etc. (Pang *et al.*, 2002 ; Mihalcea et Liu, 2006), en utilisant des attributs différents pour caractériser les documents. Même si perdurent d'autres méthodes – ayant principalement recours à l'utilisation de ressources lexicales, construites (Kim et Hovy, 2004) ou automatiquement acquises (Turney, 2002 ; Riloff *et al.*, 2003), avec la banalisation des corpus annotés, les méthodes de catégorisation basées sur l'apprentissage supervisé sont de plus en plus utilisées. Elles utilisent diverses caractéristiques textuelles : (i) tous les mots du texte (sac de mots, unigrammes ou n-grammes) (Pang *et al.*, 2002 ; Dave *et al.* 2003) ; (ii) la présence ou l'absence d'un ensemble de mots déterminés ; (iii) l'emplacement de certains mots (Kim et Hovy, 2006) ; (iv) certaines parties du discours seules : adjectifs (Kamps et Marx, 2002), collocations adverbe-adjectif (Turney, 2002) ; substantifs ; enfin (v) les dépendances syntaxiques (Nakagawa *et al.*, 2010 ; Wi *et al.*, 2009 ; Wiegand et Klakow, 2010). Nous nous inscrivons donc pleinement dans cette démarche en proposant des critères de classification issus d'analyses textométriques pour servir de base aux divers algorithmes d'apprentissage supervisé.

[1] Autrement dit, les études textométriques ne sont validées que par l'assentiment d'une communauté qui, dans le meilleur des cas, est distante (par exemple, critique littéraire, sociologie), mais, dans le pire des cas, n'est peut-être qu'un avatar du *jugement d'acceptabilité* pourtant honni de ladite communauté.

3 Présentation du corpus

3.1 Contexte applicatif de l'étude

Le corpus est constitué de 300 textes courts réunis par SAMESTORY (http://www.same-story.com), un service d'agrégation d'ego-documents. Il s'agit, en l'occurrence, de témoignages et récits d'histoires vécues postés par les internautes sur différents forums de discussion (aufeminin.com, doctissimo.fr, etc.). Les catégorisations sont multicritères : thématiques, tonalité, conseil *vs* demande, sexe de l'émetteur, situation familiale, etc. Nous traitons, dans des textes portant sur la santé, la tonalité « gai/triste ». De prime abord, elle s'apparente à une analyse thymique, mais il s'agit de catégories complexes où les phénomènes discursifs (ex : structure du récit) interviennent dans la classification autant que l'expression linguistique des sentiments. Ainsi, notre tâche est de modéliser l'art de témoigner d'une histoire vécue.

3.2 Annotation tonale du corpus

L'annotation tonale du corpus a été effectuée par SAMESTORY. Nous en avons analysé un échantillon pour en déduire la stratégie d'annotation de façon à caractériser plus finement l'opposition binaire gai/triste. Un témoignage « triste » est (i) une histoire qui finit mal, (ii) un témoignage exprimant des doutes, des interrogations, ou sollicitant de l'aide. Un témoignage « gai » est (i) une histoire triste qui finit bien, (ii) un témoignage modulant la gravité de la situation et soulignant les points positifs (iii) un conseil.

4 Description de l'expérience

4.1 Étape 1 : Choix des caractéristiques textuelles au moyen des méthodes textométriques

Nous tentons de mettre en évidence les phénomènes textuels qui différencient les témoignages de nos deux catégories. Nous avons une double ambition : trouver des critères de classification linguistiquement explicables et suffisamment robustes pour servir de critères aux méthodes d'apprentissage supervisé. Nous faisons l'hypothèse que les critères de classification *interprétables* sont plus performants que les critères trouvés par des méthodes d'apprentissage, souvent non signifiants d'un point de vue textuel et incidents au corpus d'apprentissage (ex : présence de fautes d'orthographe non pertinentes par rapport aux catégories de classification). Ainsi, lors de l'étape de sélection de critères, l'analyste écarte les critères liés à l'échantillon du corpus et choisit les critères textuels cohérents avec les composantes textuelles (thématique, dialogique, etc. *cf.* § 5) actualisées dans le corpus. Pour l'expérience, nous avons utilisé trois types de critères : (i) critères unitaires : un choix de formes, lemmes ou catégories morphosyntaxiques ; (ii) critères composites adjacents (n-grammes) ; (iii) cooccurrences multiniveaux (combinant les éléments de différents niveaux de description linguistique : formes, lemmes ou catégories morphosyntaxiques). Tous les critères sont sélectionnés selon 4 principes : leur caractère spécifique à un sous-corpus, leur répartition uniforme dans le sous-corpus, leur fréquence et leur pertinence linguistique.

L'analyse du corpus et l'extraction des critères ont été effectuées avec deux logiciels

textométriques – Lexico3 (Salem *et al.* 2003) et TXM (Heiden *et al.* 2010) – qui implémentent les algorithmes de spécificités (Lafon, 1980) et de cooccurrences (Lafon, 1981). Nous avons choisi les deux premiers types de critères selon le procédé suivant :

1. calcul des spécificités des items isolés (formes, lemmes et catégories morphosyntaxiques) et de leur n-grammes (fonction « Segments Répétés » de Lexico 3) pour chaque sous-corpus (gai/triste) ;
2. analyse des contextes d'apparition des items spécifiques (au moyen de concordances textuelles) afin de s'assurer de leur pertinence textuelle et de l'unicité de leur fonction (les critères ayant une seule fonction et signification ont été privilégiés) ;
3. vérification de la répartition uniforme des items dans le sous-corpus (fonctionnalité « Carte de Sections » du Lexico 3) ;

La sélection des cooccurrences s'est fait comme suit :

1. calcul des cooccurrences (fonction « Cooccurrences » de TXM) des items spécifiques fréquents et uniformément repartis sur la totalité du corpus ;
2. analyse des contextes d'apparition de ces cooccurrences ;
3. sélection des cooccurrences spécifiques à un sous-corpus ;

Dans les deux cas, les critères de classification pour chaque texte sont des fréquences ou des valeurs booléennes (présence/absence) des items sélectionnés.

4.2 Étape 2 : Classification

La deuxième étape consiste à utiliser des algorithmes d'apprentissage supervisé pour classer les textes. En utilisant Weka[2], nous en avons expérimenté trois, chacun d'une famille différente : les arbres de décision (J48 ; Quinlan, 1993), Naive Bayes (John et Langley, 1995) et les Machines à Vecteurs de Support (SMO ; Platt, 1998). L'objectif est d'observer les différences et similitudes au niveau des performances en changeant la nature et la quantité des critères.

Le corpus contient 300 textes équitablement répartis entres les deux catégories (147 « gaies » et 153 « tristes »). L'évaluation a été effectuée avec la méthode de validation croisée sur cinq parties.

- *Expérimentation 1.1* : première expérimentation avec des mots simples sans aucune modification (avec pour valeur leur fréquence dans un texte) ; on considère ces résultats comme la base de comparaison (*baseline*) pour d'autres expérimentations. La base de comparaison est donc l'expérimentation qui nécessite l'effort computationnel minimal sur les textes en considérant ces derniers comme un matériau brut, directement accessible (au moyen d'une segmentation en mots). Toutes les autres expérimentations effectuent des traitements supplémentaires sur les textes visant à améliorer les résultats. L'évaluation a été effectuée avec la validation croisée sur 5 parties du corpus.

[2] http://www.cs.waikato.ac.nz/ml/weka/

- *Expérimentation 1.2* : A la place des mots, nous avons utilisés leurs lemmes (casse normalisée).
- *Expérimentation 1.3* : Utilisation des n-grammes de mots (longueur maximale 3).

Dans la série des expérimentations 2, nous avons utilisé les critères élaborés selon la méthodologie décrite dans la partie précédente.

- *Expérimentation 2.1* : Utilisation de critères unitaires et de critères composites adjacents pour un total de 30 critères.
- *Expérimentation 2.2* : Ajout de critères cooccurrentiels et augmentation du nombre (total : 46 critères).
- *Expérimentation 2.3* : Augmentation du nombre de critères (total : 61 critères).

5 Résultat et discussion

Type d'attributs	**Algorithme de classification**	**% des textes bien catégorisés**
1.1. Mots simples (1200 critères)	J48	53
	Naive Bayes	63
	SMO	70
1.2. Lemmes (370 critères)	J48	55
	Naive Bayes	63
	SMO	64
1.3 N-grammes de mots (1357 critères)	J48	56
	Naive Bayes	64
	SMO	74
2.1. Critères textométriques (30 critères)	J48	67
	Naive Bayes	64
	SMO	65
2.2. Critères textométriques (43 critères)	J48	62
	Naive Bayes	72
	SMO	72
2.3. Critères textométriques (61 critères)	**J48**	**70**
	Naive Bayes	**74**
	SMO	**77**

TABLE 1 – Résultat des expérimentations

Comme dans des expériences similaires (Pang *et al.*, 2002), on constate que la

classification sur les mots simples et les n-grammes permet d'obtenir des résultats convenables compte tenu de la difficulté de la tâche. Néanmoins, cela constitue un plafond que l'on ne peut dépasser. La généralisation des critères apportée par la lemmatisation ne permet pas d'améliorer les résultats. Ce phénomène a fait l'objet de nombreux débats dans la communauté textométrique (par exemple Mellet, 2003).

A la différence de la première série d'expérimentations, nos critères textométriques sont peu nombreux mais ils constituent une base facilement extensible. L'ajout des critères augmente systématiquement les performances de Naive Bayes et SMO. Ainsi, nous observons une progression sensible sur l'ensemble des algorithmes. Notre meilleur résultat (avec SMO) dépasse de 7 points celui obtenu avec des mots simples et de 3 points celui des n-grammes. Par ailleurs, l'amélioration des résultats pour J48 et Naive Bayes est systématique.

L'interprétation des résultats chiffrés et des critères obtenus participe selon nous de la validation de l'expérimentation et en constitue une valeur ajoutée. Ainsi, nous avons organisé nos critères selon une typologie inspirée de travaux sémiotiques. Les critères thymiques (Courtès, 1991), qui relèvent d'une lecture axiologique des textes, sont essentiellement dysphoriques et concernent donc les textes tristes : *« avoir peur »*, *« je souffre »*, *« douleur »*, *« stress »*. Le seul critère thymique retenu pour la classification des textes gai est *« heureux »* (euphorique). Au-delà des critères thymiques courants, nous nous sommes intéressés à ceux relatifs à des *composantes textuelles* (Rastier, 2001) parce que, ne relevant pas de typologies axiologiques classiques (positif/négatif) (Charaudeau, 1992), ils sont rarement pris en compte en AS. La composante *dialectique* concerne l'organisation linéaire et temporelle du récit. Ces critères, dans les textes « gais », sont différents marqueurs de structuration argumentative (*« par contre »*, *« car »*) et temporelle (*« après »*, *« puis »*) absents des textes « tristes ». Dans ceux-ci, la structuration est cumulative (*« en plus »*, *« de nouveau »*) ou indice d'incertitude (*« ne pas arriver à »*, *« avoir l'impression de »*). La composante *dialogique* est relative au positionnement des acteurs. Elle met en œuvre un fort contraste entre les textes « gais », où le destinateur-énonciateur s'adresse explicitement à un *« tu »* destinataire actualisé par des pronoms de 2ème personne (pronoms personnel, possessifs, etc.), relate une expérience édifiante (*« mon expérience »*, *« pour ma part »*) et prodigue des conseils (présence d'hyperliens *« www »*) et des encouragements (*« bon courage »*) sans pour autant mettre en avant un *je*. Les témoignages « tristes » mettent en texte un *« je »* massif. Enfin, la composante *thématique* n'a pas été négligée mais nous nous sommes efforcés de ne sélectionner que des critères d'un grand niveau de généralité relatifs au domaine de la santé. Ainsi, aux noms de symptômes, maladies, traitements ou médicaments, nous avons préféré, pour les textes « tristes » : *« urgences »*, *« hôpital »*, *« rendez-vous »*, *« analyses »*, *« médecins »*, ou la locution *« être atteint de »*. Pour les textes « gais » : *« rémission »*, *« produit naturel »*, *« homéopathie »* permettent d'obtenir des résultats convaincants.

6 Conclusion

Il est admis que les méthodes efficaces en classification thématique (par exemple, l'apprentissage supervisé sur mots simples) sont peu performantes pour les tâches d'analyse de la subjectivité. La difficulté réside dans le fait que la subjectivité ne relève pas seulement du lexique, mais d'autres niveaux de description : organisation temporelle

du récit, structure argumentative, etc. Nous avons proposé ici quelques éléments d'analyse pour la prise en compte de ces niveaux de description et leur implémentation pour la classification. Le coût en temps de notre méthode d'élaboration de critères n'a pas été quantifié mais nous estimons qu'il est comparable à d'autres méthodes semi-automatiques. Le domaine manquant de méthodes éprouvées, notre expérience nous a permis de mieux comprendre la tâche et sa complexité et d'esquisser une proposition méthodologique tenant compte d'une caractérisation textuelle de la subjectivité.

7 Références

BRUNET, É. (2009). *Écrits choisis,* Volume 1, *Comptes d'auteurs. Études statistiques. De Rabelais à Gracq*. Textes édités par D. Mayaffre, Champion, Paris

BÉCHET, F., EL-BÈZE, M. et TORRES-MORENO, J.-M. (2008). En finir avec la confusion des genres pour mieux séparer les thèmes *Actes de l'atelier de clôture de la 4ème édition du Défi Fouille de Texte*.

CHARAUDEAU P. (1992). *Grammaire du sens et de l'expression*. Hachette Education.

COURTÈS, J. (1991). *Analyse sémiotique du discours. De l'énoncé à l'énonciation*, Paris, Hachette.

DAVE, K., LAWRENCE, S., et PENNOCK, D.M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *In Proceedings of the 12th international WWW conference*, May 20-24, 2003, Budapest, Hungary, pages 519-528.

HEIDEN, S., MAGUÉ, J-P. et PINCEMIN, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In I. C. Sergio Bolasco (Ed.),*JADT 2010*, Vol. 2, pages 1021-1032. [logiciel disponible sur http://textometrie.ens-lyon.fr/]

JOHN, G. H. et LANGLEY, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. *Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo, pages 338-345.

KAMPS, J. et MARX, M. (2002). Words with Attitude. *1st International WordNet Conference*, pages 332-341.

KIM, S-M. et HOVY, E. (2004). Determining the sentiment of opinions.*Proceedings of the 20th international conference on Computational Linguistics (COLING '04). Association for Computational Linguistics, Stroudsburg*, PA, USA.

KIM, S.-M. et HOVY, E. (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. *SST '06: Proceedings of the Workshop on Sentiment and Subjectivity in Text, Association for Computational Linguistics*, pages 1-8.

LEBART L. et SALEM A., (1988). *Analyse statistique des données textuelles. Questions ouvertes et lexicométrie*, Paris, Dunod.

LAFON, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus, *Mots,* 1, pages 127-165.

LAFON, P. (1981). Analyse lexicométrique et recherche des cooccurrences, *Mots, 3*, pages 95-148.

MELLET, S. (2003). Lemmatisation et encodage grammatical : un luxe inutile ? *Lexicometrica : Autour de la lemmatisation*, Dominique Labbé, éd.

MIHALCEA, R. et LIU, H. A (2006). Corpus-Based Approach to Finding Happiness AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW).

NAKAGAWA, T., INUI, K. et KUROHASHI, S. (2010). Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables. *Proceedings of Human Language Technologies.*

PANG, B., LEE, L. et VAITHYANATHAN, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79-86.

PANG, B. et LEE, L. (2008). Opinion Mining and Sentiment Analysis, Now Publishers Inc.

PLATT, J. (1998). Machines using Sequential Minimal Optimization. B. Schoelkopf, C. Burges and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning.*

RASTIER, F. (2001). *Arts et sciences du texte*, Paris, PUF.

RILOFF, E., WIEBE, J. et WILSON (2003). T. Learning subjective nouns using extraction pattern bootstrapping. *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pages 25-32.

QUINLAN, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

SALEM, R. (1993). Méthodes de la statistique textuelle, Thèse pour le doctorat d'État ès lettres et sciences humaines, Université de la Sorbonne Nouvelle – Paris 3, 998 pages.

SALEM A., LAMALLE C., MARTINEZ W., FLEURY S., FRACCHIOLLA B., *et al.* (2003). Lexico3 – Outils de statistique textuelle. Manuel d'utilisation. http://www.tal.univ-paris3.fr/lexico/]

SNYDER, B. et BARZILAY, R. (2007). Multiple aspect ranking using the Good Grief algorithm. *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL)*, pages 300-307.

TURNEY, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the Association for Computational Linguistics (ACL)*, pages 417-424.

VERNIER, M., MONCEAUX, l. et DAILLE, B. (2009). DEFT'09 : détection de la subjectivité et catégorisation de textes subjectifs par une approche mixte symbolique et statistique *Actes de l'atelier de clôture de la 5ème édition du Défi Fouille de Textes.*

WI, Y., ZHANG, Q., Huang, X., et Wu, L. (2009). Phrase Dependency Parsing for Opinion Mining. *Proceedings of EMNLP-2009*, Singapore.

WIEBE, J.M., WILSON, T., BRUCE, R., BELL, M. et MARTIN, M. (2004). Learning subjective language. *Computational Linguistics*, 30(3), pages 277-308.

WIEGAND, M. et KLAKOW, D. (2010). Convolution Kernels for Opinion Holder Extraction. *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, L.A., CA*.

Méthodologie d'exploration de corpus et de formalisation de règles grammaticales pour les langues des signes

Michael Filhol *Annelies Braffort*
LIMSI-CNRS, Campus d'Orsay bat 508, BP133, 91403 Orsay cx
michael.filhol@limsi.fr, annelies.braffort@limsi.fr

RÉSUMÉ

Cet article présente une méthodologie visant, à partir d'une observation de corpus vidéo de langue des signes, à repérer puis formaliser les régularités de structure dans les constructions linguistiques. Cette méthodologie est applicable à tous les niveaux du langage, du sub-lexical à l'énoncé complet. En s'appuyant sur deux exemples, il présente une application de cette méthodologie ainsi que le modèle AZee qui, intégrant la souplesse nécessaire en termes de synchronisation des articulateurs, permet une formalisation des règles repérées.

ABSTRACT

Methodology for corpus exploration and grammatical rule building in Sign Language

This paper presents a methodology for Sign Language video observation to extract and then formalise observed linguistic structure. This methodology is relevant to all linguistic layers from sub-lexical to discourse as a whole. Relying on two examples, we apply this methodology and describe the AZee model, which integrates the required flexibility for synchronising articulators, hence enables a specification of any new systematic rule observed.

MOTS-CLÉS : Langue des signes, analyse de corpus, modèle grammatical, synchronisation.

KEYWORDS : .Sign Language, corpus analysis, grammatical models, synchronisation.

1 Introduction

Notre objectif général est de représenter de manière formelle le fonctionnement des Langues des Signes (LS), ses éléments et ses règles. Ces représentations doivent nous permettre de générer des énoncés et produire automatiquement des animations en LS via un signeur virtuel (personnage virtuel en 3d s'exprimant en LS), et à terme d'envisager la traduction d'une langue écrite vers la LS (Filhol, 2011).

Les LS sont des langues peu dotées, dont les ressources (dictionnaires, livres de grammaire, méthodes pédagogiques, corpus...) sont très limitées. En France, la LSF n'est reconnue comme langue à part entière que depuis 2005[1]. Les quelques descriptions existantes[2] sont sommaires et nous avons pu observer pour certaines d'entre-elles qu'elles ne résistent pas à la vérification sur corpus. C'est pourquoi notre démarche comporte

[1]Loi 2005-102

[2]La langue des signes, Histoire et grammaire, IVT (ed) 1998 ; La LSF mode d'emploi, M. Companys (ed) 2003.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 375–382,
Grenoble, 4 au 8 juin 2012.

l'annotation et l'analyse d'un corpus de LS pour identifier des régularités de forme qui conduiront à la description de règles grammaticales. Ensuite, nous élaborons des modèles formels permettant de représenter ces règles.

Cet article décrit tout d'abord la méthodologie employée pour mettre en évidence les phénomènes systématiques d'une LS puis décrit le nouveau formalisme « AZee » proposé pour les représenter, en s'appuyant sur deux exemples.

2 La conception des règles

Cette section présente la méthodologie et le résultat d'une étude multilingue sur les LS française (LSF), grecque (GSL), anglaise (BSL) et allemande (DGS), réalisée pendant le projet européen Dicta-Sign[3]. Une structure linguistique identifiée est celle que nous nommons la structure de «qualification/désignation». C'est un exemple représentatif du type de règle à représenter.

2.1 Méthodologie : de la forme à la fonction, et réciproquement

Deux approches sont possibles pour déterminer une règle systématique entre une structure ou une relation sémantique d'une part, et une production de surface (phonétique) d'autre part : soit à partir de la fonction sémantique, soit à partir de la forme de la surface. La structure présentée ici a été découverte au moyen de la deuxième approche, qui a comporté trois étapes :

(1) Choix des occurrences à collecter dans le corpus : Nous avons tout d'abord repéré un grand nombre d'occurrences où la posture de la main dominée était maintenue pendant que les gestes de la main dominante continuait, sans que les deux mains ne soient en relation pour des raisons géométrique ou topologique (comme lorsque la main dominante pointe vers la main dominée). Ceci nous a conduit à définir le critère de repérage de ces structures de forme, nommé «persistance indépendante de la main dominée », comme suit :

> *Un signe bimanuel s0, suivi par un ou plusieurs signes monomanuels de la main dominante, pendant que la posture finale de s0 est maintenue par la main dominée.*
>
> Main dominante : |_ *s0* _| |_ signes monomanuels *s'* _ _ _
> Main dominée : |_ *s0* _____ maintenue _____ _ _

(2) De la forme à la fonction en LSF : Dans la partie LSF du corpus, nous avons recueilli un minimum de 150 occurrences claires de la forme de la surface décrite en (1), et nous avons constaté que toutes correspondaient à l'une des deux catégories ci-dessous:

a) *Qualification/dénomination* : La suite *s'* réalisée par la main dominante qualifie le signe *s0* tel un adjectif, le nomme avec un « nom-signe » (nom propre en LS), ou encore épelle un mot (avec la dactylologie) pour l'identifier. Cela peut être une combinaison de ces réalisations.

b) *Conservation de l'activation* : *s0* est tenu par la main dominée parce qu'il est à

[3]http://www.dictasign.eu/

nouveau nécessaire, après la séquence monomanuelle de la main dominante (s0 est souvent répété ensuite). Cela peut être considéré comme une parenthèse dans un discours, au cours de laquelle s0 doit être conservé « actif »

(3) De la fonction à la forme en différentes LSs : L'étape suivante a été de commencer un processus de vérification multilingue sur les parties LSF, DGS et GSL (respectivement LS française, allemande et grecque) du corpus Dicta-Sign. Toutes les langues ont été fouillées pour trouver des occurrences de la fonction sémantique de qualification/dénomination (2a) et les formes correspondantes observées. Les LS ont été analysées indépendamment par des experts de chaque LS et les résultats nous ont permis de confirmer nos observations sur la LSF et de proposer la règle suivante, commune aux trois LSs :

> *Lorsque s0 est un signe bimanuel suivi par un ou plusieurs signes monomanuels de qualification ou de dénomination, la main dominée a tendance à garder de manière ferme la dernière posture de s0, tandis que les autres signes sont effectués avec la main dominée.*

Dans l'exemple LSF montré figure 1, il s'agit d'une qualification suivie d'une dénomination d'une ligne de métro sur un plan : le locuteur identifie la ligne jaune U3. La main dominée garde très clairement la posture finale du premier signe et est maintenue fermement tout au long des trois signes suivants, jusqu'à ce que les deux mains soient relâchées.

FIGURE 1 – Combinaison des quatre signes LIGNE JAUNE U 3.

2.2 Discussion

Ces premiers résultats doivent être affinés, que ce soit sur les formes associées à la fonction ou sur les fonctions associées à la forme. Mais dès à présent, ils nous permettent de mettre en lumière un certain type de contraintes pouvant s'exercer sur les événements manuels et qui sont probablement accompagnés d'autres contraintes à découvrir, sur des éléments non-manuels par exemple.

Nous avons remarqué la présence de ces structures (2a) dans la base de données de lexique de LSF construite pendant le projet. Dans cette base de données, chaque entrée est une unité lexicale lemmatisée, associée à un ou plusieurs concepts. Une entrée possédant une telle structure doit-elle être considérée comme une unité lexicale, ou s'agit-il d'une construction « syntaxique » à laquelle on peut associer un concept ? Le signe est-il une étape de la lexicalisation d'une structure, et comment trancher ? De plus, il est possible que des articulateurs non manuels soient porteurs d'une structure, comme

chez certaines observées dans le projet DictaSign, et leur synchronisation peut devenir d'autant plus complexe.

Pour en revenir à notre motivation initiale de concevoir des modèles informatiques, ces considérations plaident en faveur de représentations qui ne sont ni organisées autour de l'activité manuelle a priori, ni limités à des niveaux linguistiques spécifiques (lexique, syntaxe, etc.) mais proposent un point de vue global.

3 La représentation des règles

Cette section présente un état des lieux des modèles existants, le cahier des charges auquel selon nous doit répondre un modèle de description de la LS, puis le nouveau formalisme que nous proposons nommé AZee.

3.1 Modèles à composante temporelle

Le projet le plus abouti reste celui élaboré durant le projet européen ViSiCAST, basé sur HPSG (Marshall, 2004). Il est intégré dans un système de génération automatique d'énoncés en LS qui définit des séquences de mouvements ou de signes élémentaires séparés par des transitions de même nature (Elliott, 2004). Ce type de représentation n'intègre pas de système de synchronisation suffisant pour représenter les phénomènes liés à la multi-linéarité de la LSF.

Deux modèles font tout de même apparaître la multi-linéarité dans les descriptions. Liddell & Johnson (1989) ont montré que les signes étaient divisibles en unités temporelles où les articulateurs du corps se synchronisaient en *postures*, séparées par des unités de *transition*, alternant sur une ligne temporelle de description. Ce modèle reste en revanche comme ses prédécesseurs porté sur l'activité manuelle et la description lexicale dans sa forme de citation (dictionnaire), or les LS permettent la création spontanée et sémantiquement productive d'unités non répertoriées qui contrastent avec le vocabulaire figé (« standard ») en cela qu'elles mettent souvent en jeu de nombreux articulateurs non manuels (épaules, buste, muscles faciaux, etc.) qu'il faut synchroniser.

Le modèle P/C de Huenerfauth (2006) permet de diviser par endroits une ligne de temps en deux lignes parallèles pour spécifier deux activités simultanées. L'énoncé peut se représenter sous la forme d'un arbre où les feuilles sont des signes lexicaux et les nœuds intermédiaires sont chacun :

- soit de type C (constituant), dont les enfants sont des sous-parties de l'énoncé à concaténer ;
- soit de type P (partition), dont les enfants sont des sous-parties de l'énoncé à paralléliser.

Le problème est alors que les nœuds P et C partagent systématiquement les mêmes bornes temporelles et ne peuvent se chevaucher librement à moins d'utiliser des nœuds spéciaux « Ø » qui ne représentent rien linguistiquement et rendent les descriptions fastidieuses.

3.2 Clés pour un nouveau modèle

Nous proposons un nouveau formalisme de description nommé Azee. En utilisant deux méthodes de synchronisation combinées, AZee peut décrire n'importe quel motif de synchronisation des articulateurs du corps en LS.

Dans le cas général, un groupe d'articulateurs dans une production signée a une période d'activité pendant laquelle ils concourent à l'énoncé et hors de laquelle ils sont ou retournent dans une position de repos. Cette période est appelée « intervalle » et notée « TI » (time interval). Par exemple, le schéma suivant montre 5 TI synchronisés sur un axe temporel qui correspondant à l'exemple de la figure 1.

Main dominante : |_LIGNE_| |_JAUNE_| |__U__| |__3__|
Main dominée : |_LIGNE_____ _ _ _

Chaque production linguistique met en jeu un certain nombre de TI qu'il faut synchroniser, et chaque TI contient une partie de la signation qu'il faut spécifier.

Notons en outre que l'observation d'un corpus de vidéos montre que pour une construction linguistique donnée, tous les locuteurs ne synchronisent pas nécessairement les TI de manière rigoureusement identique. On remarque que le maintien de la configuration finale du signe LIGNE par la main dominée peut varier dans sa durée, mais que la synchronisation initiale des deux mains au début du signe LIGNE reste identique pour tous les signeurs.

À propos de cette variabilité et en vue de spécifier la structure linguistique, nous posons les trois objectifs suivants :

- toute variabilité dans la production n'entraînant pas de modification du sens doit rester possible (pas de sur-spécification) ;
- tout changement entraînant une modification du sens de l'énoncé fait l'objet d'un paramètre de la règle ;
- toute spécification valable quelle que soit le contexte et le signeur doit être fixé par la règle (on appelle ces éléments les invariants de la structure).

Pour traiter ce problème, nous proposons :

- le recours à des ensembles minimaux de contraintes (gestuelles et temporelles) suffisantes pour énoncer une règle sans contraindre trop la signation ;
- la possibilité pour les éléments de spécification de dépendre de variables contextuelles non fixées par la règle mais qui prendront une valeur selon leur utilisation.

3.3 Le modèle AZee

Le modèle Azee permet de représenter les contraintes nécessaires et suffisantes (CNS) de synchronisation et de réalisation d'un énoncé en LS. Il est composé de deux modèles, Zebedee et Azalee.

Zebedee est un langage de description qui implémente des CNS ainsi que des dépendances contextuelles pour donner aux séquences posture-transition ces mêmes propriétés. Il a été initialement conçu pour décrire les unités lexicales de la LS. Nous ne détaillons pas ce formalisme ici mais une page web lui est dédiée[4].

Azalee est un formalisme capable de décrire tous types de synchronisation entre TI. En Azalee, les TI, généralement superposés, doivent être agencés sur la ligne de temps selon des contraintes temporelles à déterminer, puis chaque TI doit être spécifié, séparément, spécifiant ainsi la totalité de la structure.

Soit un ensemble de TI numérotés TI1, TI2, etc. concourant à une structure linguistique. Azalee décrit cette structure en un « azalisting », en les encapsulant comme suit :

```
[[
      règle de synchro,   %% Liste des contraintes temporelles
      règle de synchro,   %% nécessaires et suffisantes agençant
      règle de synchro... %% les TI sur l'axe temporel
|| TI1 :
      bloc de spécification
|| TI2 :
      bloc de spécification
|| ... :
       %% etc. (un bloc pour chaque TI apparaissant dans le bloc de synchro)
]]
```

À l'instar des CNS de Zebedee, les TI sont agencés sur l'axe temporel avec un ensemble minimal de contraintes temporelles nécessaires sur les bornes des intervalles (relations $<$, $=$, $\geq$...) ou sur les intervalles tout entiers (Allen, 1983). Ces contraintes apparaissent dans la première section de l'azalisting, encadrées par « [[» et le premier séparateur « || ». Nous en donnons quelques exemples ci-dessous :

- |gaze = <|pt → le début du TI nommé « gaze » précède immédiatement le début du TI « pt » (pointage manuel) ;
- gaze| < pt| → la fin de « gaze » précède celle de « pt » ;
- pt |d| md → le TI « pt » est inclus dans « md » (p. ex. la tenue de la main dominée) – le « d » est pour « during » ;
- |A = B| ~ C| → « A » débute entre la fin de « B » et la fin de « C ».

Chaque TI doit ensuite être spécifié dans son propre bloc de spécification. Cela peut représenter un simple regard sur une cible, une suite de signes manuels, un geste des épaules, du buste, des sourcils ou une combinaison de ceux-ci. Formellement, cela peut représenter :

- une simple liste de contraintes qui sera à maintenir durant toute sa durée, en utilisant des contraintes articulatoires élémentaires – ces contraintes ciblent les articulateurs du corps comme un os du squelette ou un muscle du visage, éventuellement paramétrés par un numéro d'ordre (p.ex. la « n-ième » phalange

[4]http://perso.limsi.fr/filhol/zebedee

du doigt) et/ou par un côté du corps (gauche/droit ou dominant/dominé) ;

- une synchronisation de postures séparées par des transitions, en utilisant le langage Zebedee prévu à cet effet ;
- une structure temporellement plus complexe, à savoir un azalisting imbriqué, répartissant ainsi les TI contenus sur le morceau de l'axe temporel dédié au TI englobant.

Les deux formalismes Azalee et Zebedee, et ce faisant les deux stratégies de synchronisation, se combinent et permettent l'imbrication libre de structures réutilisables.

Voici un exemple complet d'azalisting pour une structure activant une zone de l'espace de signation par un pointage dont le schéma général est décrit ci-dessous :

- le regard précède toujours d'un temps très court le signe du pointage ;
- le regard et le pointage ciblent tous deux le même point de l'espace qui dépend de l'énoncé ;
- si le regard est maintenu, il ne dépasse jamais la rétractation du pointage manuel.

```
AZOP "activation + pointage de l'espace"
 %% Ci-dessous, déclaration d'une dépendance au contexte
 DEP spaceloc : Point  %% représente l'emplacement activé
[[
      |gaze = <|pt, %% regard débute juste avant pointage
      gaze| < pt|   %% regard termine avant la fin du pointage
 || gaze:
       LOOK at [spaceloc]    %% regarder l'emplacement à activer
 || pt:
        SEQ "pointage" WITH  %% ce même emplacement comme cible
          target = [spaceloc]
        END
]]
END
```

où “pointage” est une « zebedescription » définie par ailleurs avec une dépendance contextuelle nommée « target » qui représente la cible du pointage.

4 Conclusion et perspectives

Cet article a présenté une méthodologie visant, à partir d'une observation de corpus vidéo, à repérer puis formaliser les régularités de structure apparaissant entre le sens et la forme de constructions linguistiquement motivées. Cette méthodologie se veut le moins possible empreinte de courant linguistique a priori, et applicable à tous les niveaux du langage, du sub-lexical à l'énoncé complet. Le repérage vidéo s'organise autour de deux démarches inverses et complémentaires : rechercher une forme et généraliser sur le sens ou rechercher une valeur sémantique et en extraire les invariants

surfaciques. La formalisation de règles de production à partir de ces régularités observées est rendue possible grâce au modèle AZee, dont nous avons présenté les bases. Celui-ci permet une combinaison des deux stratégies de synchronisation qu'offrent les langages Zebedee et Azalee, respectivement la synchronisation par postures et celle par agencement contraint d'intervalles temporels sur une ligne de temps.

Si la partie Zebedee de ce modèle est déjà bien évaluée (2000 signes LSF décrits), nous n'avons pour l'instant exploré qu'une dizaine de structures. Les cinq heures de corpus DictaSign annotées[5] devraient nous permettre de recueillir plus de ces structures, et plus d'occurrences pour chacune, et ainsi augmenter notre échantillon d'étude.

Nous comptons poursuivre ce travail pour multiplier le nombre de structures ainsi décrites et à terme pouvoir conclure sur la capacité d'AZee à couvrir l'ensemble des structures linguistiques mises en évidence. Notre hypothèse est qu'une fois combinées, ces règles pourront permettre de décrire des énoncés complets en LS. Ceci sera le point de départ d'une évaluation approfondie du modèle AZee et de sa mise en œuvre au sein d'applications dédiées à la génération et LS et à plus long terme à la traduction du texte vers la LS, en utilisant les entrées de ces règles comme sortie d'un système d'analyse textuelle. Ces règles pourront aussi faire suite au travail déjà entamé avec Zebedee (Gonzalez, 2012), pour étendre les travaux de reconnaissance des signes lexicaux aux structures plus larges des énoncés.

5 Références

Allen, J. F. (1983). Maintaining Knowledge about Temporal Intervals. *In Communications of the ACM 26:11, pp. 832–843*, New-York, USA.

Elliott, R., Glauert, J., Jennings V., Kennaway, R. (2004). An overview of SiGML notation and SiGMLSigning software system. *In Proceedings of LREC 2004 (Language Resources and Evaluation Conference) workshop RPSL (Representation and Processing of Sign Languages) : From SignWriting to Image Processing.Information techniques and their implications for teaching, documentation and communication*, Lisbon, Portugal.

Filhol, M. (2011). Text-sign parallel corpus study to start designing an automatic translation system. *In proceedings of SLTAT workshop 2011 (Sign Language Translation and Avatar Technology)*, Dundee, Scotland.

Gonzalez, M., Filhol, M., Collet, C. (2012). *Semi-automatic Sign Language corpora annotation using lexical representations of sig*ns, Language Resource and Evaluation Conference, Istanbul.

Huenerfauth, M. (2006). Generating American Sign Language classifier predicates for English-to-ASL machine translation. *PhD thesis*, University of Pennsylvania, USA.

Liddell, S. K., Johnson, R. E. (1989). American Sign Language, the phonological base. Sign Language studies 64, Cambridge University press.

Marshall, I., Sáfár, É. (2004). Sign Language Generation in an ALE HPSG . *In proceedings of HPSG-11*, Leuven, Belgique.

[5]http://www.sign-lang.uni-hamburg.de/dicta-sign/portal/

Annotation manuelle de matchs de foot : Oh la la la ! l'accord inter-annotateurs ! et c'est le but !

Karën Fort[1, 2] Vincent Claveau[3]
(1) INIST-CNRS, 2 allée de Brabois, 54500 Vandoeuvre-lès-Nancy
(2) LIPN, Université Paris 13 & CNRS, 99 av. J.B. Clément, 93430 Villetaneuse
(3) IRISA - CNRS, Campus de Beaulieu, 35200 Rennes
karen.fort@inist.fr, vincent.claveau@irisa.fr

RÉSUMÉ

Cet article présente une campagne d'annotation de commentaires de matchs de football en français. L'annotation a été réalisée à partir d'un corpus très hétérogène, contenant à la fois des comptes-rendus minute par minute et des transcriptions des commentaires vidéo. Nous montrons ici comment les accords intra- et inter-annotateurs peuvent être utilisés efficacement, en en proposant une définition adaptée à notre type de tâche et en mettant en exergue l'importance de certaines bonnes pratiques concernant leur utilisation. Nous montrons également comment certains indices collectés à l'aide d'outils statistiques simples peuvent être utilisés pour indiquer des pistes de corrections des annotations. Ces différentes propositions nous permettent par ailleurs d'évaluer l'impact des modalités sources de nos textes (oral ou écrit) sur le coût et la qualité des annotations.

ABSTRACT

Manual Annotation of Football Matches : Inter-annotator Agreement ! Goooooal !

We present here an annotation campaign of commentaries of football matches in French. The annotation was done from a very heterogeneous text corpus of both match minutes and video commentary transcripts. We show how the intra- and inter-annotator agreement can be used efficiently during the whole campaign by proposing a definition of the markables suited to our type of task, as well as emphasizing the importance of using it appropriately. We also show how some clues, collected through statistical analyses, could be used to help correcting the annotations. These statistical analyses are then used to assess the impact of the source modality (written or spoken) on the cost and quality of the annotation process.

MOTS-CLÉS : annotation manuelle, accords inter-annotateurs.

KEYWORDS: manual annotation, inter-annotator agreement.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 383–390,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Nous étudions dans cet article la création d'un corpus textuel annoté construit à partir de transcriptions de commentaires vidéos et de sites Web spécialisés. Ce corpus annoté est développé dans le but de mettre au point des techniques automatiques d'analyse, tels que le résumé vidéo, le *repurposing* (transformation du contenu et du format pour un autre support de diffusion) ou l'extraction d'information pour les vidéos d'événements sportifs. Cette application, développée dans le cadre d'un partenariat industriel, n'est pas détaillée plus avant dans cet article, mais il est important de noter qu'elle guide la définition des éléments à annoter (cf. section 2).

Outre la présentation d'une nouvelle ressource annotée, cet article a pour objectif de montrer l'intérêt d'analyses fines pour évaluer la qualité d'une telle ressource hétérogène. En particulier, nous proposons une définition des mesures d'accord inter- et intra-annotateur adaptée à ce type d'annotation où seuls certains éléments des corpus sont annotés. Nous montrons également comment certains indices collectés à l'aide d'outils statistiques simples peuvent être utilisés pour souligner les difficultés de la tâche d'annotation et indiquer des pistes de corrections des annotations. Ces différentes propositions nous permettent par ailleurs d'évaluer l'impact des modalités sources de nos textes (oral ou écrit) sur le coût et la qualité des annotations.

D'un point de vue applicatif, quelques travaux (Nemrava *et al.*, 2007, par exemple) font référence à un corpus annoté du domaine du football, mais à notre connaissance, aucun ne détaille l'annotation du corpus utilisé. D'autres études ont fait usage de corpus de football pour créer des lexiques monolingues (Gasiglia, 2003) or multilingues (Schmidt, 2008) plus ou moins détaillés. Dans ces cas, si les publications associées détaillent l'annotation du corpus utilisé, les annotations elles-même sont de nature linguistique plutôt que du domaine et soulèvent des questions différentes. D'un point de vue méthodologique, l'analyse statistique des annotations repose principalement sur les calculs d'accord inter-annotateurs (Artstein et Poesio, 2008, pour une revue détaillée). Ces derniers sont généralement fournis sur les corpus annotés comme mesure d'évaluation de la qualité de la ressource produite (Dandapat *et al.*, 2009, *inter alia*). Les méthodes d'annotation agiles (Voormann et Gut, 2008) proposent d'utiliser ces mesures pendant toute l'annotation du corpus, pour assurer la cohérence des annotations et limiter les divergences dans les cas, majoritaires, où l'on ne peut pas tout annoter en double avec adjudication. Notre travail se situe dans ce cadre mais aborde plusieurs problèmes posés par les particularités de nos annotations. Après une présentation des données et des annotations en section 2, nous détaillons les différentes analyses menées en section 3 et nous concluons en donnant quelques perspectives à ce travail.

2 Campagne d'annotation

2.1 Données, annotations et méthodologie

Le corpus annoté couvre 16 matchs de football. Il est composé de 24 transcriptions de commentaires tirés de vidéos (1 par mi-temps, 12 matchs) et de 16 fichiers contenant une description minute-par-minute du match (dont les 12 de la transcription et 4 matchs additionnels) tirés de sites Web spécialisés. La parole contenue dans les vidéos a été transcrite manuellement en utilisant TRANSCRIBER (Barras *et al.*, 1998) et son guide de transcription par défaut. L'ensemble du corpus a une taille d'environ 250 000 mots. Sa principale caractéristique est d'être très hétérogène

(Fort *et al.*, 2011), que ce soit d'un point de vue des types de match (ligues, championnats...), de la taille des fichiers (de 1 116 tokens par match pour les minutes à 21 000 tokens pour les transcriptions), ou de la source (chaînes de diffusion des vidéos, commentateurs, sites Web...).

Le jeu d'étiquettes a été construit en définissant les éléments intéressants pour l'application finale et ensuite affiné durant les phases d'entraînement et de pré-campagne. L'ensemble des étiquettes retenues a été divisé en trois couches, *Unités*, *Actions* et *Relations* (cf. tableau 1 [1]), chacune correspondant à un niveau d'analyse de complexité croissante à aborder successivement par les annotateurs. Par cohérence avec les besoins applicatifs et pour prendre en compte le style elliptique de l'oral (« `Makoun. Et c'est récupéré. Clerc, avec Cris. Boumsong, Makoun.` »), nous avons décidé de ne pas faire porter les annotations sur les prédicats dénotant les actions ou les relations, souvent absents, mais sur les acteurs impliqués.

Unités	acteurs	*Joueur, Equipe, Arbitre, Entraineur, ArbitreAssistant, Président*
	circonstants	*EspaceSurTerrain, LieuDuMatch, TempsDansMatch*
Actions	arbitrales	*TirerCoupFrancDirect, TirerCoupFrancIndirect, TirerCorner, TirerPenalty, FaireFauteDeJeu, HorsJeu, MarquerBut, PrendreCartonJaune, PrendreCartonRouge, PrendreRappelALOrdre*
	autres	*Centrer, FaireTentative2Centre, Dribbler, RaterBut, ArreterBut, IntercepterBallon, PossederBallon, ActionDuPublic*
Relations	arbitrales	*FaireFauteSurJoueur, TaclerFaute, RemplacerJoueur*
	autres	*FaireCombinaison, FairePasse, FaireTentative2Passe*

TABLE 1 – Couches d'annotations retenues et étiquettes correspondantes

La méthodologie employée pour l'annotation de ce corpus suit les recommandations de Bonneau-Maynard *et al.* (2005) et Gut et Bayerl (2004) ; elle est décrite en détail dans (Fort et Claveau, 2012). L'annotation a été réalisée par deux annotateurs experts du domaine avec l'outil d'annotation GLOZZ (Widlöcher et Mathet, 2009), choisi en raison de sa facilité d'utilisation et de la possibilité qu'il offre d'annoter des relations. Les temps d'annotation par couche ont été mesurées à l'aide de l'outil TIMETRACKER [2]. Nous avons également invité les annotateurs à ajouter des commentaires sur leurs annotations, et un attribut *Incertitude* a été mis à leur disposition dans GLOZZ.

2.2 Données générales sur le processus d'annotation

Le nombre total d'annotations produites s'élève à 37 784 dont 27 736 (soit plus de 73 %) pour les transcriptions. Toutes les catégories ont été utilisées, mais avec une grande disparité : par exemple, *TirerCoupFrancIndirect* et *TirerPenalty* n'ont servi que 2 fois (et uniquement dans les minutes), *PrendreCartonRouge* 6 fois et *Président* 9 fois.

Le tableau 2 présente le temps d'annotation moyen (pour 1 000 tokens) par annotateur et par source. Un t-test de Welsh à deux échantillons (avec $p = 0,05$) montre que les différences entre annotateurs ne sont pas significatives, que ce soit pour les transcriptions ou pour les

1. Le regroupement des étiquettes à l'intérieur de ces couches (circonstants, acteurs, etc) est proposé ici pour faciliter la lecture et l'analyse, mais n'existait pas dans le modèle de données utilisé pour l'annotation.
2. `http://www.formassembly.com/time-tracker/#`

	Minutes	Transcriptions
Annotateur 1	36,92	20,03
Annotateur 2	41,30	16,06

TABLE 2 – Temps moyen d'annotation par source et par annotateur, en minute/1 000 tokens

minutes. En revanche, les différences entre modalités sont jugées statistiquement significatives, pour les deux annotateurs. Cela s'explique par la différence (statistiquement significative) de densité d'annotations (nombre d'annotations par token) : 0,16 pour les minutes et 0,08 pour les transcriptions. En effet, les commentateurs sportifs ne parlent pas uniquement des événements du matchs et ont tendance à digresser. En revanche, si l'on rapporte le temps d'annotation au nombre d'annotations produites, aucune différence n'est constatée entre minutes et transcriptions. Les différences de temps entre les deux modalités s'expliquent donc uniquement par le nombre plus important d'annotations à produire à volume de texte constant.

3 Analyse statistique des annotations

3.1 Mesures d'accord et estimation des "annotables"

Les calculs d'accords inter- et intra-annotateur servent à quantifier la fiabilité, et donc la qualité, des annotations produites, mais aussi à fixer une limite supérieure aux performances que l'on peut attendre d'un système automatique, et enfin, dans notre cas, à mesurer la difficulté de la tâche selon la modalité d'origine. Pour ce faire, les Kappa (κ) de Cohen (Cohen, 1960) et de Carletta (Carletta, 1996) sont préférés aux mesures plus simples telles que la F-mesure car ils normalisent l'accord observé en fonction de l'accord attendu (ou dû au hasard). Carletta considère que l'annotation par hasard se traduit par une unique distribution valable pour les deux annotateurs, alors que Cohen considère que ces distributions dépendent de chaque annotateur (Artstein et Poesio, 2008, pour une description complète et des comparaisons).

Cependant, ces définitions posent problème dès lors que ce ne sont pas seulement les étiquettes qui peuvent varier, mais aussi les éléments à annoter (les *marquables* ou *annotables*), puisqu'elles ne précisent en rien comment le désaccord sur les annotables doit être traité. Nous proposons donc d'étendre les κ en décomposant l'accord en un accord sur l'annotable et un accord sur l'étiquette. De telles mesures nécessitent donc de connaître le nombre d'*annotables* $\mathscr{M}$. Ce nombre d'annotables est évident ou connu *a priori* pour certaines tâches (comme l'étiquetage morphosyntaxique : tous les tokens sont annotables), mais ne peut être qu'estimé *a posteriori* pour des tâches comme la nôtre (Grouin *et al.*, 2011). Nous proposons pour ce faire une estimation originale basée sur une procédure EM (*Expectation-Maximization*) décrite dans l'algorithme 1. Celui-ci énumère itérativement le nombre d'annotables δ (étape de *Maximization*) en utilisant la probabilité γ (estimée itérativement) que tous les annotateurs aient manqué le même annotable, elle-même calculée grâce à l'estimation du nombre d'annotables δ de l'itération précédente (*expectation*).

Avoir une estimation la plus exacte possible du nombre d'annotables est un enjeu d'importance pour obtenir des accords inter-annotateurs réalistes. Par exemple, si l'on considère que tous

Algorithme 1 *Estimation EM des annotables*

Entrées : $\{\mathscr{M}_j\}$ (ensembles des éléments annotés par les annotateurs A_j) ; $\delta_0 = \left|\bigcup_j \mathscr{M}_j\right|$

for (i=1 ; $\delta_i \neq \delta_{i-1}$; i++) **do**

expectation : $\gamma_i = \prod_j P(A_j \text{ manque un marquable}) = \prod_j \frac{\delta_{i-1} - |\mathscr{M}_j|}{\delta_{i-1}}$

maximization : $\delta_i = \frac{\delta_0}{1-\gamma_i}$

end for

return δ

les mots (tokens) des textes sont des annotables (et donc ceux non annotés sont considérés annotés par défaut par une étiquette *sans-annotation*), le Kappa de Cohen pour les accords intra- et inter-annotateurs atteindrait respectivement 0,9456 et 0,9404, principalement par l'abondance des accords sur les très nombreux mots *sans-annotation*. De telles valeurs masquent des différences qui sont révélées avec l'estimation plus réaliste des annotables que nous proposons (voir sous-section 3.2).

Les deux κ, tels que nous les avons implémentés, sont aussi très stricts, puisque la moindre différence dans les annotations (étiquette bien sûr, mais aussi délimitation des entités) est considérée comme un désaccord. Quand cela est possible, nous fournissons donc également la mesure d'accord entropique implémentée dans GLOZZ (Mathet et Widlöcher, 2011) ; celle-ci autorise en effet les correspondances partielles d'annotation et fournit donc des valeurs d'accord prenant en compte ces accords partiels. Elle ne s'applique cependant pas encore aux relations.

3.2 Accords inter-annotateurs

Le tableau 3 présente l'accord inter- et intra-annotateur, selon la modalité, calculés avec le κ de Cohen, et, à des fins de comparaison, la mesure d'entropie de GLOZZ. Le κ de Carletta a également été calculé et est très proche dans la quasi-totalité des cas au κ de Cohen ; nous ne reportons donc pas ses valeurs par manque de place. Cette proximité signifie qu'il n'y a pas de biais d'annotateur : les distributions des annotations produites par chacun des annotateurs sont très similaires (Artstein et Poesio, 2008). On constate sans surprise que l'accord (aussi bien inter- qu'intra-annotateur) a tendance à être plus faible dans les transcriptions que dans les minutes, à l'exception d'une transcription pour laquelle les unités/actions ont produit un accord bien supérieur (près de 0,65). Cette tendance générale se manifeste spécialement dans les cas d'annotations complexes comme les relations. Les spécificités de l'oral mentionnées précédemment, et en particulier le style elliptique propre aux commentaires, expliquent facilement cette différence.

Si le calcul d'accord inter-annotateurs est devenu une bonne pratique standard du développement de ressources annotées, nous souhaitons promouvoir dans cet article l'intérêt d'une analyse plus détaillée. Cela est d'autant plus important quand les éléments annotés relèvent de catégories différentes et que ces catégories elles-mêmes ont des populations très différentes, comme c'est le cas ici. En effet, les valeurs présentées précédemment masquent des disparités importantes entre catégories d'annotation. Dans le tableau 4, colonnes 2 et 5, nous développons les résultats d'accord inter-annotateurs par regroupements de catégories. Les difficultés accrues sur les transcriptions se vérifient à cette échelle, mais l'on constate en outre de très faibles accords pour

	inter-annotateurs		intra-annotateur A1		intra-annotateur A2	
	κ de Cohen	Glozz	κ de Cohen	Glozz	κ de Cohen	Glozz
Minutes unités/actions	0,5992	0,7627	0,7531	0,8753	0,7109	0,8519
Minutes relations	0,5707	-	0,6377	-	0,5983	-
Transcriptions unités/actions	0,6234	0,7498	0,7558	0,8327	0,6812	0,8179
Transcriptions relations	0,4345	-	0,4010	-	0,4701	-

TABLE 3 – Accords inter-annotateurs et intra-annotateur par modalité

	Minutes			Transcriptions		
	κ	Incertitude	Gain d'entropie	κ	Incertitude	Gain d'entropie
Acteurs	0,9228	0,5	-	0,8974	1,0	-1
Circonstants	0,4827	1,9	49	0,4441	10,0	15
Actions arbitrales	0,5999	4,3	-	0,5082	19,7	7
Actions autres	0,3240	1,3	92	0,1407	9,8	26
Relations arbitrales	0,6355	10,7	-	0,4520	18,4	8
Relations autres	0,5540	10,2	8	0,3793	69,9	23

TABLE 4 – Accords inter-annotateurs par modalité et par famille d'annotations

certaines catégories. Les accords sur les entités offrent un grand contraste entre les annotations des acteurs et des circonstants, davantage sujets à interprétations. De la même manière, les événements (actions ou relations) sanctionnés par une action de l'arbitre obtiennent des accords bien supérieurs aux autres événements. Un examen détaillé des résultats montre que les annotateurs sont rarement en désaccord sur les types des éléments annotés, mais qu'ils annotent des éléments différents. Ce dernier point justifie d'autant plus l'emploi de notre technique d'estimation des annotables et explique pourquoi la définition standard des κ sur-estime tant l'accord.

3.3 Incertitudes

Les annotateurs avaient la possibilité d'indiquer les annotations leur posant problème, pour quelque raison que ce soit, à l'aide d'un champ *Incertitude*. Ces incertitudes permettent, lors de la campagne, de préciser les instructions d'annotations, de comprendre certaines annotations lors de l'utilisation du corpus, mais aussi d'aider à l'analyse automatique des résultats, comme indicateur de la difficulté d'annotation. Il est à noter qu'un seul des annotateurs de la campagne a véritablement utilisé les incertitudes, mais de manière systématique.

Dans les colonnes 3 et 6 du tableau 4, nous présentons les taux d'incertitude par catégorie d'annotations et par modalité. On y constate encore une fois que proportionnellement plus d'incertitude concerne l'oral retranscrit (différence statistiquement significative, test de Student pour deux ensembles, avec $p = 0,05$).

Nous nous sommes intéressés au lien éventuel entre incertitude et désaccord. Nous avons cherché à savoir si la présence d'une incertitude est liée au désaccord. Par contre, nous considérons non interprétable l'absence d'incertitude. Pour ce faire, nous avons calculé la différence entre l'entropie de l'accord $H(Acc)$ (eqn 2) de la variable aléatoire Acc indiquant s'il y a accord ou non ($\mathscr{D}_{Acc} = \{vrai\,;\,faux\}$) et l'entropie conditionnelle de l'accord sachant qu'une incertitude est

présente ($H(Acc|Inc = présent)$, eqn 2). Un gain positif signifie que l'incertitude aide à discerner les accords des désaccords. Autrement dit, pour une catégorie donnée, un gain positif indique que l'incertitude peut aider à prédire les catégories susceptibles de désaccord.

$$H(Acc) = -\sum_{v \in \mathcal{D}_{Acc}} P(Acc = v) \log P(Acc = v) \quad (1)$$

$$H(Acc|Inc = vrai) = -\sum_{v \in \mathcal{D}_{Acc}} P(Acc = v|Inc = vrai) \log P(Acc = v|Inc = vrai) \quad (2)$$

Ces gains sont indiqués en colonnes 4 et 7 du tableau 4 pour les familles d'annotation (dans trois cas, il n'y a pas assez d'incertitudes pour les calculer). À une exception près, ils sont tous positifs, ce qui signifie que ces incertitudes sont des bons indicateurs d'erreurs, même si elles n'ont été posées que par un seul annotateur. Que ce soit pour les minutes ou les transcriptions, il faut remarquer que le gain est d'autant plus fort que le taux de désaccord est important. L'étude des causes de ces incertitudes est donc une piste privilégiée pour la correction systématisée des désaccords et donc des éventuelles erreurs d'annotation.

4 Conclusion et perspectives

L'analyse de la campagne d'annotation présentée dans cet article [3] a mis en exergue différents éléments. D'un point de vue méthodologique, notre technique d'estimation des annotables doit permettre un calcul d'accord inter-annotateurs plus réaliste dans les cas où leur nombre peut varier selon l'annotateur. Nous avons aussi montré que les bonnes pratiques ne sauraient se limiter à un calcul d'accord inter-annotateurs unique pour l'ensemble des annotations quand celles-ci relèvent de catégories différentes et d'effectifs non équilibrés. Enfin, nous avons montré que l'étude statistique des incertitudes met au jour une possibilité de détecter systématiquement les désaccords ou erreurs potentiels. Ces différentes analyses nous ont aussi permis de montrer que le coût d'annotation des textes issus de l'oral est moindre que pour ceux issus de l'écrit, du fait de la différence de densité des annotations. En revanche, les indicateurs de qualité (désaccord, incertitudes) indiquent sans ambiguïté la difficulté accrue de traiter de l'oral. Les annotations seront librement disponibles sous licence LGPL-LR à `http://www.irisa.fr/texmex/people/claveau/corpora/FootQuaero/` dès que les corrections identifiées auront été effectuées. Le guide d'annotation mis à jour sera lui-aussi fourni.

En suite de ce travail, et aussi bien d'un point de vue théorique que pratique, nous souhaitons développer des approches permettant de propager automatiquement des corrections d'annotations à partir de quelques corrections apportées à une petite quantité de données. Ces approches s'appuieraient d'une part sur les analyses précédentes pour détecter les catégories les plus problématiques, et éventuellement sur des approches d'apprentissage artificiel pour proposer des corrections.

3. Nous remercions chaleureusement les annotateurs de la campagne, C. Ris et A. Zérouki, de l'INIST-CNRS, pour leur travail minutieux et leurs précieux retours. Nous remercions également V. Lux et A.-R. Ebadat pour leur participation à la préparation de la campagne, et Technicolor pour la mise à disposition d'une partie des données. Ce travail a été réalisé dans le cadre du programme Quæro (`http://www.quaero.org`), financé par OSEO, agence nationale de valorisation de la recherche.

Références

ARTSTEIN, R. et POESIO, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

BARRAS, C., GEOFFROIS, E., WU, Z. et LIBERMAN, M. (1998). Transcriber: a free tool for segmenting, labeling and transcribing speech. *In Actes de First International Conference on Language Resources and Evaluation (LREC 1998)*, Grenade, Espagne.

BONNEAU-MAYNARD, H., ROSSET, S., AYACHE, C., KUHN, A. et MOSTEFA, D. (2005). Semantic annotation of the french media dialog corpus. *In Actes de InterSpeech*, Lisbonne, Portugal.

CARLETTA, J. (1996). Assessing Agreement on Classification Tasks: the Kappa Statistic. *Computational Linguistics*, 22:249–254.

COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

DANDAPAT, S., BISWAS, P., CHOUDHURY, M. et BALI, K. (2009). Complex Linguistic Annotation - No Easy Way Out ! A Case from Bangla and Hindi POS Labeling Tasks. *In Proceedings of the third ACL Linguistic Annotation Workshop*, Singapour.

FORT, K. et CLAVEAU, V. (2012). Annotating football matches: : Influence of the source medium on manual annotation. *In Actes de Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turquie.

FORT, K., NAZARENKO, A. et RIS, C. (2011). Corpus linguistics for the annotation manager. *In Actes de Corpus Linguistics*, Birmingham, Angleterre.

GASIGLIA, N. (2003). Pistes méthodologiques pour l'exploration d'un corpus à haut rendement relatif au parler du football, une langue de spécialité de grande diffusion. *In 3es journées de linguistique de corpus*. Centre de Recherche en Littérature, Linguistique et Civilisation (CRELLIC), Université de Bretagne-Sud, Lorient.

GROUIN, C., ROSSET, S., ZWEIGENBAUM, P., FORT, K., GALIBERT, O. et QUINTARD, L. (2011). Proposal for an extension of traditional named entities: from guidelines to evaluation, an overview. *In Actes de 5th Linguistic Annotation Workshop*, pages 92–100, Portland, Oregon, USA. Association for Computational Linguistics.

GUT, U. et BAYERL, P. S. (2004). Measuring the reliability of manual annotations of speech corpora. *In Actes de Speech Prosody*, pages 565–568, Nara, Japon.

MATHET, Y. et WIDLÖCHER, A. (2011). Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs. *In Actes de Traitement Automatique des Langues Naturelles 2011 (TALN 2011)*, Montpellier, France.

NEMRAVA, J., SVATEK, V., SIMUNEK, M. et BUITELAAR, P. (2007). Mining over: football match data: seeking associations among explicit and implicit events. *In Proc. of Znalosti 2007*.

SCHMIDT, T. (2008). *The Linguistics of Football (Language in Performance 38)*, volume 38, chapitre The Kicktionary: Combining corpus linguistics and lexical semantics for a multilingual football dictionary, pages 11–23. Gunter Narr, Tübingen, Allemagne.

VOORMANN, H. et GUT, U. (2008). Agile corpus creation. *Corpus Linguistics and Linguistic Theory*, 4(2):235–251.

WIDLÖCHER, A. et MATHET, Y. (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. *In Actes de Traitement Automatique des Langues 2009 (TALN 2009)*, Senlis, France.

Étude de différentes stratégies d'adaptation à un nouveau domaine en fouille d'opinion

Anne Garcia-Fernandez Olivier Ferret
CEA, LIST, Laboratoire Vision et Ingénierie des Contenus
Gif-sur-Yvette, F-91191 France.
prénom.nom@cea.fr

RÉSUMÉ

Le travail présenté dans cet article se situe dans le contexte de la fouille d'opinion et se focalise sur la détermination de la polarité d'un texte en adoptant une approche par apprentissage. Dans ce cadre, son objet est d'étudier différentes stratégies d'adaptation à un nouveau domaine dans le cas de figure fréquent où des données d'entraînement n'existent que pour un ou plusieurs domaines différents du domaine cible. Cette étude montre en particulier que l'utilisation d'une forme d'auto-apprentissage par laquelle un classifieur annote un corpus du domaine cible et modifie son corpus d'entraînement en y incorporant les textes classés avec la plus grande confiance se révèle comme la stratégie la plus performante et la plus stable pour les différents domaines testés. Cette stratégie s'avère même supérieure dans un nombre significatif de cas à la méthode proposée par (Blitzer *et al.*, 2007) sur les mêmes jeux de test tout en étant plus simple.

ABSTRACT

Study of various strategies for adapting an opinion classifier to a new domain

The work presented in this article takes place in the field of opinion mining and aims more particularly at finding the polarity of a text by relying on machine learning methods. In this context, it focuses on studying various strategies for adapting a statistical classifier to a new domain when training data only exist for one or several other domains. This study shows more precisely that a self-training procedure consisting in enlarging the initial training corpus with texts from the target domain that were reliably classified by the classifier is the most successful and stable strategy for the tested domains. Moreover, this strategy gets better results in most cases than (Blitzer *et al.*, 2007)'s method on the same evaluation corpus while it is more simple.

MOTS-CLÉS : fouille d'opinion, adaptation à un nouveau domaine, auto-apprentissage.

KEYWORDS: opinion mining, domain adaptation, self-training.

1 Introduction

Le travail présenté dans cet article part de deux constats bien connus en Traitement Automatique des Langues, et notamment en fouille d'opinion, lorsqu'une approche par apprentissage supervisé est mise en place. D'une part, la constitution de ressources annotées est un processus très coûteux. D'autre part, un système ayant de bonnes performances dans un domaine donné n'est pas nécessairement adapté à un autre domaine. De nombreux travaux ont été menés en

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 391–398,
Grenoble, 4 au 8 juin 2012.

fouille d'opinion, en particulier sur la problématique de l'adaptation à un nouveau domaine. Les approches ainsi développées vont de l'identification des correspondances terminologiques entre domaines à l'utilisation de ressources externes telles que des lexiques d'opinion, lexiques au sein desquels chaque terme se voit typiquement associer une *polarité* positive, négative voire même neutre. Mais que se passe-t-il si l'on ne dispose pas de telles ressources ? Comment procéder en n'ayant accès qu'à quelques données annotées ne relevant pas du domaine ciblé ?

Nous nous attachons dans cet article à étudier des stratégies d'apprentissage automatique pour la classification de textes n'utilisant pas de ressources externes. Différentes configurations de données annotées pour la tâche de classification de textes en polarité positive ou négative sont testées afin de déterminer celles permettant d'obtenir les meilleures performances lorsqu'il s'agit de classer des textes relevant d'un nouveau domaine. En particulier, nous exposons une approche d'apprentissage dit *itératif* par laquelle un corpus annoté manuellement est enrichi au cours de boucles successives par des données annotées automatiquement.

2 Fouille d'opinion et classification de textes multi-domaine

La fouille d'opinion regroupe un grand nombre de travaux centrés sur l'exploration de textes afin d'en déterminer le caractère objectif ou subjectif ou encore d'extraire les avis qui y sont exprimés par leurs auteurs. Une des caractéristiques de ce type d'analyses est leur caractère transversal : leur nature ne dépend pas le plus souvent du domaine des textes considérés. Ainsi, dans le travail considéré ici, qui se focalise sur une tâche de classification de textes subjectifs, en l'occurrence des critiques issues du site AMAZON, les deux classes considérées, polarité positive ou négative, sont générales et non dépendantes des différents domaines qu'abordent ces critiques. En revanche, les moyens pour effectuer cette classification peuvent être plus ou moins liés à un domaine, ce qui rend la problématique de l'adaptation à un nouveau domaine de ce type d'analyses particulièrement importante. Dans ce qui suit, nous illustrerons d'abord le caractère contextuel de la détermination de la polarité d'un énoncé avant de passer en revue les principales approches pour l'adaptation d'une telle classification à un nouveau domaine.

2.1 Classer des textes selon leur polarité

Une des principales approches pour déterminer la polarité d'un texte consiste à se focaliser sur des termes porteurs d'opinion. De tels termes peuvent être trouvés dans des ressources de référence, telles que SENTIWORDNET (Esuli et Sebastiani, 2006). Néanmoins, la disponibilité de ces ressources n'est pas suffisante pour déterminer la polarité d'un texte. En effet, la polarité d'un terme peut dépendre de son contexte. Cette dépendance existe d'abord à un niveau local. Ainsi, la présence d'une négation dans une phrase peut inverser la polarité de celle-ci alors même qu'elle contient un ou plusieurs termes négatifs ("Cela dit le réalisateur sait y faire et c'est bien pour ça que le film n'est pas mauvais du tout, ni ennuyeux, ni lent."). Cette dimension est prise en compte par de nombreux travaux. Taboada *et al.* (2011) utilisent ainsi un lexique de termes porteurs d'opinion mais pondèrent l'importance de ces termes en fonction du caractère subjectif ou objectif des paragraphes dans lesquels ils apparaissent. Ils intègrent par ailleurs les négations, les modifieurs (notamment les modifieurs d'intensité) et la modalité (en particulier les suppositions). Choi et Cardie (2009) proposent au travers de l'algorithme Vote & Flip de

prendre en compte la présence d'une ou plusieurs négations dans le contexte d'un terme porteur d'opinion modifiant ainsi sa polarité. L'objet auquel un terme porteur d'opinion se rapporte peut également influer sur la valeur de ce terme. Ainsi *mortel* est porteur d'opinion négative dans *un ennui mortel* mais d'opinion positive dans *cette fête était vraiment mortelle*. Enfin, la polysémie des termes peut aussi jouer un rôle puisque certains termes n'ont pas la même valeur d'opinion selon leur sens. C'est le cas par exemple de *navet* qui, dans la phrase "C'est un navet.", n'est pas porteur d'opinion si l'on fait référence au légume mais renvoie à une opinion négative s'il qualifie un film. La polarité d'un terme dans un texte est alors influencée plus globalement par le domaine auquel ce texte se rattache. Dans cette perspective, Harb *et al.* (2008) proposent d'acquérir un lexique d'opinion lié à une thématique en sélectionnant dans des corpus liés à cette thématique des termes cooccurrents avec des termes dont la polarité est déjà connue et en utilisant une mesure de similarité entre les termes candidats et les termes connus.

2.2 D'un domaine à l'autre

Dans le prolongement des travaux de la section précédente fondés sur des lexiques d'opinion, une première voie pour aborder le problème de la dépendance par rapport au domaine en matière de fouille d'opinion consiste à définir des capacités d'adaptation automatique de ces lexiques à un domaine donné. Dans cette optique, Jijkoun *et al.* (2010) proposent d'adapter un lexique d'opinion général à un domaine spécifique en caractérisant les termes de ce lexique par des profils de contextes syntaxiques obtenus à partir d'un corpus général. Ces profils sont ensuite utilisés pour identifier les termes porteurs d'opinion pertinents pour ce domaine à partir de corpus représentatifs de celui-ci. Gindl *et al.* (2010) adaptent pour leur part un lexique d'opinion en fonction du domaine considéré en supprimant les termes de ce lexique dont la polarité varie selon le contexte.

Outre l'utilisation de lexiques d'opinion, la détermination de la polarité d'un texte peut bénéficier de l'utilisation de corpus annotés. La dépendance de ceux-ci par rapport à un domaine donné rend néanmoins leur usage délicat et conduit à définir différentes approches pour compenser cette dépendance. Denecke (2009) associent ainsi le lexique général SentiWordNet et un corpus d'entraînement relevant de différents domaines autres que le domaine cible pour classer des textes comme subjectifs ou objectifs. Pour la même tâche, Aue et Gamon (2005) s'affranchissent de lexiques d'opinion et comparent différentes approches utilisant des corpus d'apprentissage d'un domaine autre que le domaine cible. La disponibilité de corpus annotés relevant de différents domaines est ainsi un élément clef dans la tâche de classification de textes d'un nouveau domaine. Blitzer *et al.* (2007) construisent un tel corpus, le *Multi-domain Sentiment Dataset* (MDSD), en s'appuyant pour minimiser les coûts d'annotation sur les critiques rédigées sur le site amazon portant sur des objets variés appartenant à 25 grands domaines. Par ailleurs, Blitzer *et al.* (2007) proposent également de chercher des correspondances entre domaines par la technique du *Structural Correspondence Learning* (SCL) en s'appuyant sur des traits et des prédicteurs pivots. Li et Zong (2008) réutilisent ce corpus et proposent deux approches. L'approche par fusion de traits (ou *feature fusion*) se fonde sur le regroupement des corpus d'apprentissage de différents domaines en un seul. L'approche par fusion de classifieurs (ou *classifier fusion*) consiste à construire autant de classifieurs que de domaines sources disponibles et à entraîner un classifieur (un méta-classifieur) sur les sorties de ces modèles.

Quelle que soit la méthode, les performances obtenues pour des textes relevant d'un nouveau domaine cible sont moins bonnes que celles obtenues en disposant de données annotées dans le

domaine cible. Pour réduire le coût de cette annotation, des approches dites d'*Active learning* ont été proposées. L'idée est de détecter les exemples classés avec un faible score de confiance par le modèle, d'annoter ces exemples manuellement et d'intégrer ces nouvelles données aux données d'apprentissage. Si ces méthodes permettent de limiter la quantité d'annotation manuelle à effectuer, elles restent tout de même coûteuses. De ce fait, des méthodes dites d'auto-apprentissage (*self-training*) proposent, à l'image de (Drury *et al.*, 2011), de s'appuyer sur des données non annotées manuellement mais classées avec un fort score de confiance par un classifieur pour enrichir le corpus d'apprentissage de ce dernier et élargir ainsi sa couverture. C'est l'approche que nous privilégierons ici en la transposant au cas de textes appartenant à d'autres domaines.

3 Stratégies d'adaptation

L'étude que nous présentons dans cet article aborde la classification de textes en termes de polarité positive ou négative selon une approche supervisée fondée sur un corpus d'entraînement annoté manuellement, sans s'appuyer sur un lexique d'opinion constitué *a priori*. Dans ce cadre, qui reprend celui de (Blitzer *et al.*, 2007) et de (Li et Zong, 2008), son objectif est de déterminer, partant de corpus d'entraînement dans un ou plusieurs domaines sources et d'un corpus non annoté dans un domaine cible (corpus dit de développement), la stratégie la plus adaptée de constitution d'un nouveau corpus d'entraînement à partir de ces corpus disponibles afin d'obtenir les meilleures performances possibles sur le domaine cible. Les différentes stratégies considérées se différencient selon deux facteurs principaux : l'utilisation de corpus appartenant à un seul ou à plusieurs domaines sources ; l'utilisation ou non d'un corpus du domaine cible non annoté. Nous avons plus précisément testé les stratégies suivantes, chacune reposant sur le même volume de textes annotés manuellement pour constituer leur corpus d'entraînement :

Un corpus source [BASELINE] Cette stratégie baseline utilise un corpus annoté d'un unique domaine source autre que le domaine cible.
Apprentissage itératif à partir d'un corpus source [ITE-FIXE, ITE-SEUIL] Dans cette approche, nous entraînons un modèle sur un seul corpus source, comme précédemment, classifions les textes du corpus du domaine cible, sélectionnons les exemples ayant le meilleur score de confiance et intégrons ces exemples au corpus d'entraînement. Cette boucle est réitérée jusqu'à épuisement du corpus du domaine cible. La sélection des meilleurs exemples peut se faire en fonction d'un seuil sur le score de confiance (approche dite ITE-SEUIL) ou bien en fonction d'un nombre fixe d'exemples intégrés à chaque itération (ITE-FIXE).
Plusieurs corpus sources [MULTI-DOMAINE] Dans cette configuration, les corpus de plusieurs domaines sources sont associés pour construire le corpus d'entraînement. Nous prenons ici tous les domaines sources, ce qui représente une autre forme de baseline.
Méthode par vote [MULTI-VOTE] Dans cette stratégie, elle aussi classique, nous entraînons un modèle par corpus source et procédons à une classification finale par vote : un exemple donné est ainsi classé en fonction de la décision majoritaire observée parmi les classifieurs associés à chaque domaine source.
Apprentissage itératif à partir de plusieurs corpus sources [ITE-MULTI-VOTE] Cette stratégie est une hybridation de la méthode par vote et de l'apprentissage itératif. À partir d'un corpus source, nous entraînons plusieurs modèles (un par domaine source) ; puis nous sélectionnons les meilleurs exemples qui sont ensuite intégrés dans chacun des corpus sources de départ. Nous sélectionnons alors les exemples classés unanimement par tous les modèles.

4 Mise en œuvre et résultats

Nos expériences ont été menées sur le corpus MDSD évoqué précédemment, corpus composé de critiques portant sur des produits variés et triées par domaine (livre, cuisine et articles ménagers, vêtements ...). Une critique est composée d'un texte, d'un titre et d'une note. Les textes sont courts, quelques phrases seulement, et rédigés en anglais. Les notes varient de 1 à 5, 1 indiquant l'avis le plus négatif sur le produit et 5 l'avis le plus positif. L'exemple présenté à la figure 1 est une critique issue du sous-corpus "Book". Nous avons utilisé la même configuration de données

```
<review>
  (...)
  <rating>5.0</rating>
  <review_text>
    I read Les Misérables after I saw the opera, and it has inspired
    in me more than any book I've ever read. I don't believe one
    could ever find a better novel anywhere. For everyone (...)
  </review_text>
</review>
```

FIG. 1 – Exemple de critique issue du MDSD relevant du domaine BOO

que celle de (Blitzer *et al.*, 2007) afin de pouvoir comparer nos résultats aux leurs à la différence près que les corpus de test ont été scindés en corpus de développement et corpus de test. Le tableau 1 présente une description générale du corpus utilisé, organisé en quatre domaines. Les données d'apprentissage sont équilibrées entre critiques positives et négatives[1] et sont présentes en même quantité dans les quatre domaines.

	Corpus d'entraînement		Corpus de développement	Corpus de test
Domaine	#critiques	#formes/critiques	# critiques	# critiques
Cuisine & articles ménagers (KIT)	2 000	96	2 000	3 945
Livres (BOO)	2 000	174	2 000	2 465
DVD (DVD)	2 000	189	2 000	3 945
Matériel électronique (ELE)	2 000	113	2 000	3 945

TAB. 1 – Description générale du corpus

À l'instar de (Torres-Moreno *et al.*, 2007), nous avons utilisé un modèle de classification à base de boosting pour effectuer nos tests, modèle mise en œuvre grâce à l'outil BoosTexter (Shapire et Singer, 2000). Le modèle produit est composé d'un ensemble de règles binaires (ou *weak learners*) portant chacune sur la présence d'un n-gramme et se voyant associer une probabilité par rapport à chaque classe considérée. Lors de la phase de classification, un score est calculé pour chaque classe en fonction des règles déclenchées par le texte traité et la classe de plus haut score est attribuée au texte. La configuration utilisée a été sélectionnée empiriquement en optimisant les paramètres pour notre approche `baseline`. Le nombre de tours a ainsi été fixé à 50. Les règles utilisent des n-grammes de taille 1 et les textes ne sont pas lemmatisés.

Les résultats de classification sont donnés en termes d'exactitude (*accuracy*) afin de pouvoir comparer nos résultats avec ceux de (Blitzer *et al.*, 2007). Le tableau 2 présente les résultats pour

[1] Comme Blitzer, nous considérons une critique comme positive si sa note est > 3 et négative si elle est < 3.

l'ensemble des approches. Si l'on se concentre en premier lieu sur les stratégies ne mettant en jeu qu'un seul domaine source, on peut observer que l'approche `baseline` donne des exactitudes supérieures à 70% quels que soient les domaines d'entraînement (TRN) et les domaines de test (TEST). L'approche `ite-fixe` est simple mais ne prend pas en compte le score de confiance donné par BoosTexter. Elle ne dispose donc pas de critère naturel d'arrêt de prise en compte de nouveaux exemples, ce qui la conduit à « consommer » tout le corpus de développement. Elle obtient en pratique des performances très inférieures à celles de `baseline` et se révèle être la moins bonne de nos stratégies. La prise en compte du score de confiance de BoosTexter (approche `ite-seuil`) donne en revanche la meilleure exactitude (notée par $*$) pour la plupart des couples (domaine source, domaine cible). Il est à noter que les cas où cette approche ne donne pas de meilleurs résultats que `baseline` font tous intervenir les données du domaine ELE (matériel électronique).

TRN	TEST	baseline	ite-fixe	ite-seuil	multi-domaine	multi-vote	ite-multi-vote
DVD	BOO	79,7	48,8	84,4$*$	68,1	69,6	75,6†
ELE	BOO	75,4	41,6	79,3$*$			
KIT	BOO	70,9	38,1	81,8$*$			
BOO	DVD	77,2	69,5	82,0$*$	72,2	70,3	81,4†
ELE	DVD	76,2	54,3	73,4			
KIT	DVD	76,9	54,0	78,2$*$			
BOO	ELE	77,5$*$	64,5	65,5	69,1	64,7	77,5†
DVD	ELE	74,1$*$	69,7	62,2			
KIT	ELE	86,8$*$	60,4	75,7			
BOO	KIT	78,9	68,4	82,7$*$	75,4	71,0	81,4†
DVD	KIT	81,4	61,1	82,3$*$			
ELE	KIT	85,9	65,6	78,7			

TAB. 2 – Résultats en termes d'exactitude pour l'ensemble des approches

Les résultats des stratégies utilisant des données annotées relevant de plusieurs domaines sont présentés dans les 3 dernières colonnes du tableau 2. L'approche `multi-domaine` constitue dans ce cas notre baseline. La comparaison de ses résultats avec ceux de l'approche `baseline` (un unique domaine source) est clairement en défaveur de l'approche `multi-domaine` dans tous les cas de figure. Ce constat tend à montrer qu'utiliser un corpus d'entraînement composé de données sources hétérogènes du point de vue thématique (approche `multi-domaine`) est une moins bonne option pour classer des textes dans un domaine cible qu'utiliser un corpus d'entraînement source thématiquement homogène. Il est néanmoins possible qu'une telle observation soit à nuancer en fonction de la taille des corpus et du nombre de domaines. L'approche par vote (`multi-vote`) ne permet pas quant à elle d'obtenir une performance de classification plus élevée que la baseline. Cette approche, tout comme l'approche `ite-fixe`, ne prend pas en compte le score de confiance accordé par le modèle lors de la classification et le fait que la majorité des modèles catégorisent un exemple dans une classe donnée n'est apparemment pas un indice suffisant pour compenser cette insuffisance. Cependant, là encore, le nombre de domaines considérés peut avoir une influence. L'approche `ite-multi-vote` est celle des trois approches utilisant plusieurs corpus annotés donnant les meilleurs résultats (notés par †) et ce, quel que soit le domaine cible. La méthode d'apprentissage itérative se révèle donc particulièrement intéressante dans ce cas de figure comme elle l'est dans le cas d'un domaine source unique. L'exactitude moyenne de `ite-multi-vote`, égale à 79,0, est même légèrement supérieure à l'exactitude moyenne de `ite-seuil`, égale à 77,2, en particulier du fait d'un

meilleur comportement pour le domaine cible ELE.

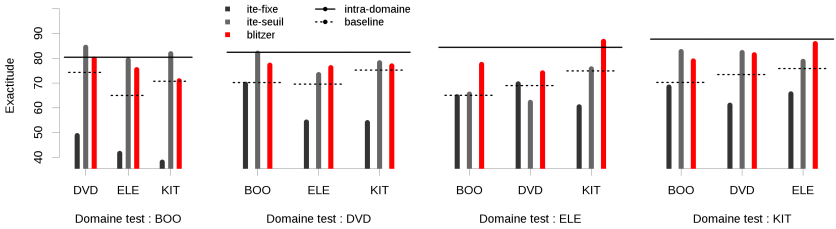


FIG. 2 – Exactitude obtenue par les approches utilisant un seul domaine d'apprentissage[2]

La figure 2 propose une comparaison de nos résultats avec ceux de (Blitzer *et al.*, 2007). Les barres horizontales pleines indiquent l'exactitude obtenue par une approche intra-domaine (les corpus source et cible relèvent du même domaine). Les barres horizontales pointillées indiquent l'exactitude obtenue par notre approche `baseline`. Pour chaque couple de domaines sources et cibles sont indiqués nos résultats ainsi que ceux de Blitzer. On peut observer que notre meilleure approche (`ite-seuil`) obtient de meilleurs résultats que ceux de Blitzer en dehors du domaine ELE (qu'il soit source ou cible). L'approche proposée dans (Blitzer *et al.*, 2007) permet de mettre en correspondance des termes supposés équivalents d'un domaine à un autre, termes prenant la forme de n-grammes tels que *must read* (BOO) ou *excellent product* (KIT). Il semble néanmoins que ces correspondances concernent essentiellement des termes constitués de mots pleins. Nous expliquons la performance de notre approche par le fait que nos modèles ne favorisent pas un type d'unités plutôt qu'un autre, ce qui leur permet d'utiliser aussi bien des mots outils, qui se retrouvent dans tous les domaines, que des mots pleins, plus spécifiques à un domaine. Or les mots outils jouent un rôle dans l'expression des opinions puisqu'ils permettent notamment d'exprimer la négation et l'intensité (Taboada *et al.*, 2011).

5 Conclusion et perspectives

Dans cet article, nous avons présenté une étude sur différentes stratégies possibles pour construire un classifieur statistique pour un domaine cible en ne disposant de données annotées pour son entraînement que pour un ou plusieurs autres domaines. Nous avons en particulier montré l'efficacité pour cette tâche d'une stratégie d'apprentissage itératif assimilable à une forme d'auto-apprentissage et consistant à incorporer progressivement dans le corpus d'entraînement du classifieur les textes d'un corpus du domaine cible que ce classifieur annote avec la plus grande confiance. Cette stratégie se révèle même dans un nombre significatif de cas plus efficace que la méthode présentée dans (Blitzer *et al.*, 2007) tout en étant plus simple. Une des prolongations les plus immédiates de ce travail est sa généralisation à d'autres types de classifieurs que le

[2]L'approche *intra-domaine* correspond à la même configuration que notre approche BASELINE à la différence que les corpus d'entraînement et de test relèvent du même domaine.

boosting utilisé ici. Au-delà, nous envisageons la transposition à notre contexte inter-domaine de la démarche d'auto-apprentissage présentée dans (Wiebe et Riloff, 2005), démarche fondée sur l'utilisation d'un classifieur supplémentaire, de nature différente du classifieur initial, pour la constitution non supervisée du corpus d'entraînement.

Remerciements

Ce travail a été financé par la Fondation Jean-Luc Lagardère. Nous tenons également à remercier Morgane Marchand et Romaric Besançon pour leur contribution aux prémices de ce travail.

Références

Aue, A. et Gamon, M. (2005). Customizing sentiment classifiers to new domains : a case study. *In RANLP 2005*.

Blitzer, J., Dredze, M. et Pereira, F. (2007). Biographies, Bollywood, Boom-boxes and Blenders : Domain Adaptation for Sentiment Classification. *In ACL 2007*, Prague, Czech Republic.

Choi, Y. et Cardie, C. (2009). Adapting a Polarity Lexicon using Integer Linear Programming for Domain-Specific Sentiment Classification. *In EMNLP 2009*, pages 590–598, Singapore.

Denecke, K. (2009). Are SentiWordNet scores suited for multi-domain sentiment classification ? *In 4th International Conference on Digital Information Management (ICDIM 2009)*, pages 1–6.

Drury, B., Torgo, L. et Almeida, J. J. (2011). Guided self training for sentiment classification. *In Workshop on Robust Unsupervised and Semisupervised Methods in Natural Language Processing*.

Esuli, A. et Sebastiani, F. (2006). SentiWordNet : A Publicly Available Lexical Resource for Opinion Mining. *In 5th Conference on Language Resources and Evaluation (LREC 2006)*.

Gindl, S., Weichselbraun, A. et Scharl, A. (2010). Cross-Domain Contextualization of Sentiment Lexicons. *In 19th European Conference on Artificial Intelligence*, pages 771–776.

Harb, A., Dray, G., Plantié, M., Poncelet, P., Roche, M. et Trousset, F. (2008). Détection d'opinion : Apprenons les bons adjectifs ! *In INFORSID'08 - Atelier FODOP'08*, pages 59–66.

Jijkoun, V., de Rijke, M. et Weerkamp, W. (2010). Generating focused topic-specific sentiment lexicons. *In 48th Annual Meeting of the Association for Computational Linguistics*, pages 585–594.

Li, S. et Zong, C. (2008). Multi-domain sentiment classification. *In 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, pages 257–260.

Shapire, R. E. et Singer, Y. (2000). BoosTexter : A boosting-based system for text categorization. *Machine Learning*, 39(1):135–168.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. et Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Torres-Moreno, J.-M., El-Bèze, M., Béchet, F. et Camelin, N. (2007). Comment faire pour que l'opinion forgée à la sortie des urnes soit la bonne ? Application au défi DEFT 2007. *In Atelier DEFT'07 - Plate-forme AFIA 2007*, Grenoble, France.

Wiebe, J. et Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. *In CICLing-2005*.

Le *Lexicoscope* : un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques

Olivier Kraif [1], *Sascha Diwersy* [2]

(1) LIDILEM, Université Stendhal Grenoble 3, BP 25, 38040 Grenoble Cedex

(2) Université de Cologne

olivier.kraif@u-grenoble3.fr, sascha.diwersy@uni-koeln.de

RÉSUMÉ

Dans le cadre du projet franco-allemand Emolex, dédié à l'étude contrastive de la combinatoire du lexique des émotions en 5 langues, nous avons développé des outils et des méthodes permettant l'extraction, la visualisation et la comparaison de profils combinatoires pour des expressions simples et complexes. Nous présentons ici l'architecture d'ensemble de la plate-forme, conçue pour effectuer des extractions sur des corpus de grandes dimensions (de l'ordre de la centaine de millions de mots) avec des temps de réponse réduits (le corpus étant interrogeable en ligne[1]). Nous décrivons comment nous avons introduit la notion de pivots complexes, afin de permettre aux utilisateurs de raffiner progressivement leurs requêtes pour caractériser des constructions lexico-syntaxiques élaborées. Enfin, nous donnons les premiers résultats d'un module d'extraction automatique d'expressions polylexicales récurrentes.

ABSTRACT

The Lexicoscope : an integrated tool for combinatoric profiles observation and lexico-syntactic constructs extraction.

The German-French research project Emolex whose aim is the contrastive study of the combinatorial behaviour of emotion lexemes in 5 languages has led to the development of methods and tools to extract, display and compare the combinatorial profiles of simple and complex expressions. In this paper, we present the overall architecture of the query platform which has been conceived to ensure efficient processing of huge annotated text corpora (consisting of several hundred millions of word tokens) accessible through a web-based interface. We put forward the concept of “complex query nodes” introduced to enable users to carry out progressively elaborated extractions of lexical-syntactic patterns. We finally give primary results of an automated method for the retrieval of recurrent multi-word expressions, which takes advantage of the complex query nodes implementation.

MOTS-CLÉS : collocations, cooccurrences, profil combinatoire, expressions polylexicales, lexique des émotions.

KEYWORDS : collocations, combinatorial profiles, multi-word expressions.

[1] L'accès au corpus sera rendu public, moyennant authentification, d'ici quelques mois.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 399–406,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Cette communication présente des travaux réalisés dans le cadre du projet Emolex, projet franco-allemand cofinancé par l'ANR et la DFG. Dans le cadre de cette recherche, nous avons rassemblé des corpus massifs comportant plusieurs centaines de millions de mots pour 5 langues différentes (l'allemand, le français, l'anglais, l'espagnol et le russe). L'objectif du projet est d'analyser, dans une perspective formulée par Sinclair (2004) ou encore Hoey (2005) et d'un point de vue contrastif, les valeurs sémantiques et les rôles discursifs à partir de la combinatoire du lexique des émotions, afin d'élaborer une cartographie permettant de mieux structurer ce champ lexical, avec des applications en lexicographie mais aussi en didactique des langues et traductologie. Cette étude porte plus précisément sur le développement d'une approche automatisée permettant de guider l'observation linguistique par l'extraction de cooccurrences autour d'un pivot.

2 Un modèle de cooccurrence flexible

Pour caractériser le profil combinatoire d'une entrée, nous reprenons le concept de *lexicogramme*, introduit par Maurice Tournier et repris dans le logiciel WebLex (Heiden, Tournier 1998) : il s'agit d'établir, pour un pivot donné, la liste de ses cooccurrents les plus fréquents, à gauche et à droite, en faisant l'extraction des fréquences de cooccurrence et en calculant des mesures d'association statistiques (telles que rapport de vraisemblance ou t-score). Pour construire ces lexicogrammes, nous proposons un modèle de cooccurrence flexible permettant à l'utilisateur de définir lui-même les *unités de cooccurrences* : formes, lemmes, catégories morphosyntaxiques, traits additionnels (p.ex. sémantiques), relations syntaxiques (dans le cas des *colligations*) ou des combinaisons de ces informations. La possibilité de faire intervenir des combinaisons de ses traits nous semble importante pour permettre à l'utilisateur d'ajuster la focale de ses observations en allant du général au particulier (ou vice-versa), de préciser des contraintes pour désambiguïser certains contextes, et de combiner les aspects lexicaux et syntaxiques dans ses observations. Par ailleurs nous proposons également une caractérisation flexible de *l'espace de cooccurrence,* qui conditionne les points de rencontre entre pivot et collocatifs, ainsi que la manière de les dénombrer. On peut par exemple définir la cooccurrence à l'intérieur d'un empan de largeur fixe, éventuellement différente à droite et à gauche du pivot. Mais on peut aussi rechercher la *cooccurrence syntaxique*, à l'instar de Kilgariff et Tugwell (2001) ou Charest et al. (2010), mise en jeu lorsqu'une relation fonctionnelle (du type sujet, complément d'objet, modifieur, etc.) a été identifiée entre deux unités. Evert (2007), signale l'intérêt de ce type de cooccurrence en terme de bruit et de silence : "(...) unlike surface cooccurrence, it does not set an arbitrary distance limit, but at the same time introduces less “noise” than textual cooccurrence". Pour la cooccurrence syntaxique, nous exploitons les relations de dépendances obtenues grâce à différents analyseurs : XIP pour l'anglais (Aït-Mokhtar et al. 2001), Connexor pour l'allemand, le français et l'espagnol (Tapanainen & Järvinen 1997), DeSR pour le russe (Attardi et al. 2007), basé sur un modèle stochastique créé à partir du corpus arboré SyntagRus (Nivre *et al.,* 2008). Un post-traitement a permis d'harmoniser et de standardiser l'annotation des relations de dépendance entre les

langues (l'annotation de Connexor ayant servi de référence). Nous avons par la suite complété ces relations pour obtenir des dépendances plus pertinentes sur le plan sémantique (p. ex. sujet profond dans les constructions passives, etc.).

Avec le modèle de cooccurrence ainsi défini, on peut viser des aspects très génériques de la combinatoire (par exemple : quels sont les principaux collocatifs de la forme *surprise* toutes relations confondues) ou beaucoup plus spécifiques et circonscrits (par exemple : quels sont les principaux collocatifs verbaux à l'imparfait du nom lemmatisé *surprise* en tant qu'objet direct). Le tableau 1 montre un tel lexicogramme :

l1	l2	f	f1	f2	loglike
surprise_N	créer_V	614	2098	21658	4548,43
surprise_N	réserver_V	230	2098	2869	2143,50
surprise_N	avoir_V	484	2098	423602	627,50
surprise_N	constituer_V	94	2098	13778	406,80
surprise_N	éviter_V	43	2098	16296	109,30
surprise_N	manifester_V	22	2098	2424	106,62
surprise_N	causer_V	19	2098	2210	90,06
surprise_N	ménager_V	15	2098	1495	75,58
surprise_N	exprimer_V	23	2098	6186	72,54
surprise_N	provoquer_V	23	2098	10551	50,61
surprise_N	feindre_V	9	2098	676	50,31

TABLEAU 1 : extrait du lexicogramme pour le nom lemmatisé *surprise* pris en tant qu'objet direct (f = fréquence de cooccurrence, f1 = fréquence de l1, f2 = fréquence de l2)

3 Visualisations comparatives

A partir de ces lexicogrammes, nous offrons différentes modalités d'exploration :

– pour l'analyse linguistique, le "retour au texte" est indispensable : un simple clic sur un collocatif permet de retrouver, sous forme de concordance, tous les contextes de cooccurrence avec le pivot.

– pour comparer de manière synthétique divers profils combinatoires, nous proposons d'identifier les lexicogrammes à des points dans un espace vectoriel, en ne retenant que la mesure jugée la plus pertinente (fréquence, loglike, t-score, etc.). Il est dès lors possible d'utiliser des méthodes d'analyse de données pour visualiser les similarités entre pivots : analyse factorielle des correspondances (AFC), échelonnement multidimensionnel (MDS) ou classification hiérarchique ascendante (hClust). La figure 1 montre ces sorties pour des unités du domaine sémantique de la 'colère' (obtenues grâce aux modules du projet 'GNU R'). La classification, réalisée pour la relation "objet", indique une hiérarchisation assez bien corrélée à l'intensité du sentiment. Quant à la 'factor map', réalisée pour des relations quelconques

concernant des collocatifs adjectivaux, elle permet de distinguer trois groupes : *révolte*, *indignation* - souvent lié à la sphère publique et politique ; *fureur*, *rage*, *colère* - lié à l'expression ponctuelle et plus ou moins intense de l'affect ; enfin *énervement*, *irritation*, *exaspération* - qui concernent plutôt des états émotionnels précurseurs de cette manifestation. Ces cas montrent de façon assez éclairante le lien entre les valeurs sémantiques et la combinatoire lexico-syntaxique.

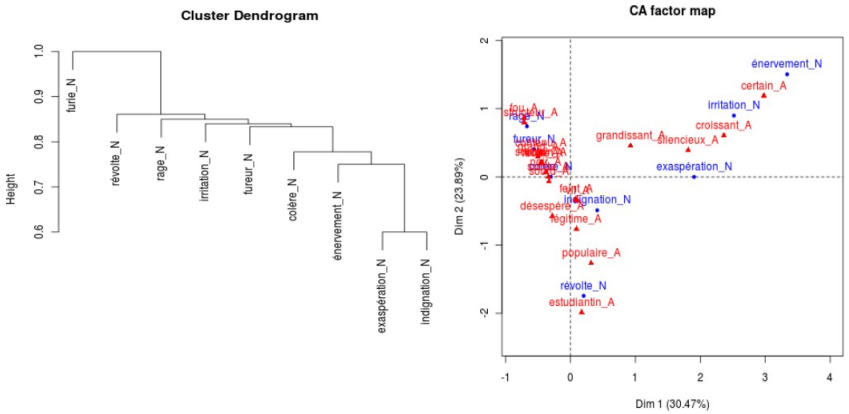


FIGURE 1 : Classification hiérarchique et AFC (domaine sémantique de la 'colère')

4 Architecture logicielle

Comment répondre rapidement à une requête d'utilisateur lorsqu'on interroge des corpus contenant des centaines de millions d'occurrences ? La réponse est simple, a priori : grâce à une indexation préalable des unités et des cooccurrences. Mais la difficulté de notre système tient au fait que ni les unités, ni l'espace de cooccurrence ne sont définis à l'avance : on peut interroger des lemmes, des formes, des combinaisons lemmes-catégories, et toute combinaison de forme, lemme, catégorie et traits (ces derniers pouvant être caractérisés par des expressions régulières). En outre, l'espace de cooccurrence est établi dynamiquement, au moment de la requête, par des expressions régulières définissant l'ensemble des relations à prendre en compte.

Pour répondre à la double exigence de flexibilité et d'efficacité, nous avons élaboré une indexation multi-niveaux, sous la forme de hachages de hachages sérialisés : chaque forme pointe vers l'ensemble des lemmes correspondants ; chaque lemme pointe vers l'ensemble de ses catégories possibles (dans le corpus) ; chaque lemme-catégorie pointe vers l'ensemble des traits associés (dans le corpus) ; chaque lemme-catégorie-traits pointe vers l'ensemble des relations associées ; chaque lemme-catégorie-traits-relation pointe vers un ensemble de paires (collocatif,fréquence). Les expressions régulières liées au contraintes portant sur les catégories, traits et relations sont appliquées lors du parcours de l'index. Les ensembles de catégories, traits et relations étant réduits (et fermés) et le temps de recherche dans le hachage étant en O(1), la succession de ces recherches n'est

pas très couteuse.

En ce qui concerne l'implémentation, nous avons opté pour le langage Perl, pour son traitement très efficace des expressions régulières. Pour les index, nous avons testé deux systèmes de bases de données réputés pour leur efficacité : BerkeleyDB 5.1.25[2] et KyotoCabinet 1.2.48[3]. Les résultats du tableau 3 montrent que le système qui est apparu le plus efficace pour nos requêtes était celui de KyotoCabinet::BTree.

Corpus presse (fr) 2007-2008 (87 807 463 tokens)	Taille des index	**test 2 1 pivot**	**test 2 5 pivots**	**test 3 24 pivots**
BerkeleyDB:Hash	1800 Mo	125 s. / 1,3 s.	275 s. / 206 s.	892 s. / 766 s.
KyotoCabinet::Hash	1200 Mo	50 s. / 1 s.	376 s. / 180 s.	749 s. / 702 s.
KyotoCabinet::Btree	955 Mo	76 s. / 1.5 s.	247 s./ 231 s.	416 s. / 315 s.

TABLEAU 2 : comparaison des tailles et des temps de réponse pour différents types de DBM (le 2ème temps est obtenu lorsqu'une requête est immédiatement réitérée).

Ces temps sont donnés à titre de comparaison : ils ont été obtenus sur un PC ancien et assez lent. Sur notre matériel actuel (Intel Core2 Quad CPU Q9550 2.83GHz, avec 4Go de RAM) nous obtenons des temps environ 4 fois supérieurs. La différence importante entre le 1er et le 2ème temps indique que ce sont les accès disques qui pénalisent les traitements, car lorsque la DBM est en cache, la réponse est presque instantanée. En utilisant un disque SSD ultra-rapide, nous prévoyons d'améliorer les temps de réponse de manière drastique.

5 Prise en compte des pivots multimots

L'aspect exclusivement binaire des relations de dépendance directe peut aboutir à un rétrécissement du contexte des observations et faire manquer des phénomènes intéressants sur le plan phraséologique. Ces limitations empêchent notamment l'extraction automatique de séquences polylexicales à valeur d'unité minimale de sens (les « meaning units » selon Sinclair 2004), qui peuvent présenter une variabilité considérable sur le plan de l'expression.

Cependant, en ce qui concerne les « collocations lexicales », Tutin (2008) affirme que la plupart d'entre elles ont une structure binaire, même pour celles qui s'étendent à plus de deux éléments, car elles correspondent sémantiquement à une structure prédicat-argument : "*Collocations can be considered as predicate-argument structures, and as such, are prototypically binary associations, where the predicate is the collocate and the argument is the base. Most ternary (and over) collocations are merged collocations (collocational clusters) or recursive collocations.*"

Et en effet, de nombreux travaux dédiés à l'extraction de collocations étendues à plus de deux mots se basent en fait sur des modèles binaires, appliqués à deux éléments composés : collocation d'arbres syntaxiques (Charest et al., 2010), construction itérative

[2]http://www.oracle.com/technetwork/database/berkeleydb/overview/index.html

[3]http://fallabs.com/kyotocabinet/

de cooccurrence multimots à partir de cooccurrences binaires (Seretan et al., 2003), ou encore calcul de mesure d'association multimots en combinant des mesures à deux termes.

De la même manière, il est possible d'étendre notre architecture pour le calcul des lexicogrammes d'un pivot donné, en la généralisant à des configurations plus complexes : la solution consiste à définir le pivot non plus seulement à partir d'une forme prise isolément, mais comme *une forme associée à un certain contexte lexico-syntaxique*. Une fois déterminé ce contexte, il est possible de calculer le tableau de contingence comme précédemment, le pivot et son contexte formant en quelque sorte une nouvelle unité pour laquelle il est possible de calculer à la fois les fréquences de cooccurrence (en se basant sur les relations du pivot) et la fréquence marginale dans le corpus.

Pour l'écriture des contextes, nous utilisons le formalisme de méta-expressions régulières proposé par Kraif (2008). Par exemple, pour rechercher le pattern V+ DET(poss.) + admiration_N + POUR, nous définissons le contexte suivant :

pivot : #1= *admiration#N*
contexte : <l=son> <#1> && <#2> && <pour,#3>::(.*,#1,#2)(.*,#2,#3)

Le calcul est seulement un peu plus long à mettre en œuvre, car les pivots multimots n'étant pas connus a priori, il n'est pas possible de les indexer tels quels. Seuls les tokens (formes ou lemmes) composant le contexte, ainsi que les relations de dépendances entre deux tokens définis, sont indexés, ce qui permet de réduire significativement l'ensemble des phrases à analyser. Pour des expressions comportant plusieurs relations, comme c'est l'intersection des phrases indexées pour chaque relation qui est retenue, la recherche est plus rapide : en d'autres termes, plus un pivot complexe est long, plus sa recherche est rapide. Dans le tableau 3 ci-dessous, on constate que pour le contexte donné en exemple, la mesure du log-likelihood fait clairement ressortir les verbes *cacher* et *dissimuler*, qui correspondent tous deux à la même construction stéréotypée : *X ne pas cacher/dissimuler son admiration pour Y*.

l1	l2	f	f1	f2	N	loglike
admiration_N	cacher_V	4	14	527	544994	38,83
admiration_N	dissimuler_V	2	14	107	544994	22,70
admiration_N	proclamer_V	2	14	176	544994	20,70
admiration_N	exprimer_V	2	14	642	544994	15,53
admiration_N	redire_V	1	14	76	544994	10,57
admiration_N	manifester_V	1	14	193	544994	8,70
admiration_N	confier_V	1	14	1319	544994	4,91

TABLEAU 3 - extrait de lexicogramme pour le pivot multimot *son admiration pour* pris en tant qu'objet direct

Ainsi conçue, l'extraction des lexicogrammes pour les pivots multimots se veut surtout être un outil d'observation permettant aux utilisateurs, par complexification progressive, de mieux préciser le contexte des phénomènes qui les intéressent (comme ici en précisant la détermination ou la structure prépositionnelle).

Cette approche qui va du simple vers le complexe peut néanmoins, d'une certaine

manière, s'automatiser. Partant d'un pivot simple, on peut retenir ses collocatifs les plus saillants pour former de nouveaux pivots multimots. Et l'on peut réitérer l'opération de manière récursive sur les nouveaux pivots, jusqu'à une taille limite fixée arbitrairement. Nous avons implémenté ce processus jusqu'à une taille maximale de 5 mots, en ne retenant, à chaque itération, que les candidats à l'extension qui cooccurrent au moins 3 fois et pour lesquelles la valeur de loglike sont supérieure à 10. Ne sont retenus que les pivots multimots maximaux (de 5 mots) ou qui ne peuvent être étendus par un pivot multimot plus long.

Dans l'exemple ci-dessous, pour mieux cibler l'extraction autour du nom *admiration*, nous avons imposé que le premier collocatif soit issu de la relation d'objet direct (on trouve donc, pour commencer, un verbe). Voici les résultats obtenus, sans filtrage, pour les 3 verbes les plus saillants.

1 : précision_N qui_PRON forcer_V la_DET admiration_N
2 : précision_N forcer_V la_DET admiration_N
3 : vouer_V une_DET admiration_N sans_PREP borne_N
4 : vouer_V une_DET profond_A admiration_N
5 : vouer_V une_DET grand_A admiration_N
6 : il_PRON vouer_V une_DET grand_A admiration_N
7 : qui_PRON vouer_V une_DET admiration_N
8 : qui_PRON pas_ADV cacher_V son_PRON admiration_N
9 : ne_ADV pas_ADV cacher_V son_PRON admiration_N
10 : avoir_V cacher_V son_PRON admiration_N
11 : il_PRON pas_ADV cacher_V son_PRON admiration_N
12 : qui_PRON ne_ADV cacher_V pas_ADV admiration_N

Comme souvent dans les extractions d'expressions multimots, on trouve un ensemble d'expressions de natures diverses (collocations simples, collocations récursives, locutions, etc.), avec notamment des fragments incomplets d'expressions plus larges (cf. exemple. 2) ou des expressions qui agrègent des éléments de contexte non pertinent (cf. exemple 10, avec *avoir*). On obtient cependant, et ceci de façon assez précise, des constructions récurrentes et stéréotypées caractéristiques de la combinatoire du nom *admiration* pris en tant qu'objet.

6 Conclusion

Nous avons présenté un nouvel outil d'exploration de la combinatoire lexico-syntaxique, que nous avons baptisé le *lexicoscope*. Cet outil s'appuie sur un modèle de cooccurrence flexible permettant à l'utilisateur de définir lui même les unités qui l'intéressent (en combinant forme, lemme, catégorie et traits) ainsi que l'espace de cooccurrence visé (en précisant les relations de dépendance concernées). Le lexicoscope permet en outre d'effectuer des comparaisons des profils combinatoires, synthétisés sous la forme de lexicogrammes, et propose en sortie des visualisations du type AFC, MDS ou hClust.

Enfin, pour permettre à l'utilisateur de ne pas se limiter aux seules dépendances directes autour d'un pivot, nous avons ajouté la possibilité de définir des pivots multimots avec leurs contextes syntaxiques. Ce nouvel outil est actuellement à l'essai, dans le cadre des

observations contrastives effectuées pour le projet Emolex. L'interface sera accessible pour le grand public d'ici quelque mois (mais les corpus, qui sont soumis à des restrictions de droits d'auteur, ne pourront être diffusés dans leur intégralité). D'ici là, nous pourrons effectuer une analyse plus précise des possibilités offertes par le lexicoscope pour la comparaison des profils combinatoires de différents pivots, et en dégager une méthodologie d'observation adaptée.

7 Références

Aït-Mokhtar, S., Chanod, J.-P., Roux C. (2002) "Robustness beyond Shallowness: Incremental Deep Parsing", *Natural Language Engineering*, 8 :121-144.

Attardi, G., Dell'Orletta, F., Simi, M., Chanev, A., Ciaramita, M. (2007) "Multilingual Dependency Parsing and Domain Adaptation using DeSR", In *Proc. of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, Prague.

Charest,S. Brunelle E., Fontaine J. (2010) Au-delà de la paire de mots : extraction de cooccurrences syntaxiques multilexémiques, *Actes de TALN 2010,* Montréal, juillet 2010

Evert, Stefan (2007). Corpora and collocations. in A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, article 58. Mouton de Gruyter, Berlin.

Heiden S., Tournier M. (1998) Lexicométrie textuelle, sens et stratégie discursive, actes *I Simposio Internacional de Análisis del Discurso*, Madrid.

Hoey, M. (2005) : *Lexical Priming: A New Theory of Words and Language*, London, Routledge.

Kilgariff A.,Tugwell D. (2001) WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography, *Proc ACL workshop on COLLOCATION Computational Extraction Analysis and Exploitation*, Toulouse July 2001.

Kraif, O. (2008) Comment allier la puissance du TAL et la simplicité d'utilisation ? l'exemple du concordancier bilingue ConcQuest, *JADT 2008*, PUL, 625-634, vol. 2.

Nivre, J., Boguslavsky, I. M., Iomdin, L. L. (2008) "Parsing the SYNTAGRUS Treebank of Russian", *Proceedings of the 22nd International Conference on Computational Linguistics* (Coling 2008), Manchester, August 2008, p. 641–648.

Seretan V., Nerima L., Wehrli E. (2003). Extraction of Multi-Word Collocations Using Syntactic Bigram Composition. *Proceedings of the Fourth International Conference on Recent Advances in NLP*, (RANLP-2003), 424–431.

Sinclair, John McH. (2004) *Trust the text : language, corpus and discourse*, London, Routledge.

Tapanainen, P., Järvinen, T. (1997) "A non-projective dependency parser", In *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, DC, p. 64-74.

Tutin A. (2008), For an extended definition of lexical collocations, *Proceedings of Euralex*, Barcelone 15-19 juillet 2008, Université Pompeu Fabra.

Analyse des contextes et des candidats dans l'identification des équivalents terminologiques en corpus comparables

Audrey Laroche[1]
(1) RALI–DIRO, Université de Montréal, C.P. 6128, Succ. Centre-Ville Montréal (Québec) H3C 3J7, Canada
audrey.laroche@umontreal.ca

RÉSUMÉ

L'approche standard d'identification d'équivalents terminologiques à partir de corpus comparables repose sur la comparaison de mots contextuels en langues source et cible et sur l'utilisation d'un lexique bilingue. Nous analysons manuellement, selon des critères linguistiques (parties du discours, spécificité et relations sémantiques), les propriétés des mots contextuels et des erreurs commises par l'approche standard appliquée à la terminologie médicale pour suggérer des améliorations basées sur la sélection de mots contextuels.

ABSTRACT

Analysis of contexts and candidates in term-translation spotting in comparable corpora

The standard approach for identifying terminological equivalents from comparable corpora is based on the comparison of source and target language context words using a bilingual lexicon. We cary a manual analysis of the linguistic properties (parts of speech, specificity and semantic relations) of the context words and the inacurrate equivalents given by the standard approach applied to medical terminology, in order to suggest improvements based on the selection of context words.

MOTS-CLÉS : équivalents terminologiques, vecteurs contextuels, corpus comparables, terminologie médicale, étude qualitative.

KEYWORDS: terminological equivalents, contextual vectors, comparable corpora, medical terminology, qualitative study.

1 Introduction

L'identification automatique d'équivalents de termes (comme *accident vasculaire cérébral* et *stroke*) est un sujet qui intéresse de nombreux chercheurs. Les applications potentielles en sont multiples : aide à la rédaction de dictionnaires bilingues spécialisés ou à la traduction, recherche d'information multilingue, traduction automatique, etc. (Rapp, 1999; Li et Gaussier, 2010). La plupart des techniques proposées pour identifier des équivalents de termes reposent sur l'hypothèse selon laquelle le contexte d'un mot en langue source est similaire au contexte de son équivalent en langue cible (Rapp, 1999).

Les équivalents terminologiques sont, dans l'approche dite standard (Rapp, 1999), repérés en

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 407–414,
Grenoble, 4 au 8 juin 2012.

comparant, à l'aide d'un lexique bilingue de projection, leurs vecteurs contextuels extraits de corpus comparables [1]. Parmi les travaux récents s'inspirant de cette approche, mentionnons ceux de (Rubino, 2011; Rubino et Linarès, 2011), qui combinent par vote trois niveaux de représentation pour chaque terme (contexte, thème et graphie). (Morin et Prochasson, 2011) utilisent un lexique de projection spécialisé formé à partir de phrases parallèles extraites automatiquement de corpus comparables. (Prochasson et Fung, 2011) combinent des vecteurs contextuels et des modèles de cooccurrence entre mots pour trouver les équivalents de termes rares. (Li et Gaussier, 2010) et (Li *et al.*, 2011) proposent deux méthodes pour améliorer le degré de comparabilité des corpus, paramètre également étudié par (Prochasson, 2009). Toutes ces stratégies contribuent à améliorer la performance de l'identification d'équivalents terminologiques. Notons que la majorité des travaux mentionnés portent sur le domaine médical, qui fait aussi l'objet du présent article.

L'analyse de la performance de ces divers systèmes repose sur des mesures classiques comme la précision et le rappel. Dans le présent article, nous analysons qualitativement les mots contextuels pris en compte dans l'approche standard, notamment du point de vue de leur partie du discours, de leur degré de spécificité et de leur relation sémantique avec le terme en langue source. Cette analyse servira éventuellement à déterminer s'il est possible de sélectionner les mots du contexte pour augmenter la performance de l'identification d'équivalents. De plus, nous examinons les candidats équivalents erronés obtenus avec l'approche standard, de façon à catégoriser les erreurs typiques et ainsi proposer des heuristiques pouvant améliorer la performance.

2 Approche par projection de contextes

L'approche standard d'extraction d'équivalents terminologiques à partir de corpus comparables est basée sur la comparaison de vecteurs de mots contextuels tirés de corpus de langues source et cible. Chaque terme dont on cherche l'équivalent est caractérisé par ses mots voisins (la définition de *voisinage* variant d'une étude a l'autre) qui sont pondérés (à l'aide d'une mesure d'association) en fonction de leur fréquence de cooccurrence avec ce terme source. Les mots voisins sont projetés dans la langue cible à l'aide d'un lexique bilingue de projection (constitué de mots spécialisés ou généraux), formant ainsi un vecteur projection. Ce dernier est comparé aux vecteurs de mots contextuels des termes extraits du corpus en langue cible : les candidats équivalents sont ceux dont le vecteur contextuel ressemble le plus (selon une certaine mesure de similarité) au vecteur projection.

Cette approche par projection de contextes comporte plusieurs paramètres de base. (Laroche et Langlais, 2010) en ont fait une étude détaillée en utilisant des corpus comparables anglais et français tirés de Wikipédia avec NlGbAse (Charton et Torres-Moreno, 2010), ainsi que des lexiques de projection contenant des mots généraux (provenant de *Freelang*) et spécialisés (provenant du *Multilingual glossary of technical and popular medical terms* du Heymans Institute of Pharmacology). Les expériences portaient sur 5 000 termes nominaux simples et complexes [2]

1. Des corpus sont comparables s'ils ne sont pas des traductions l'un de l'autre, mais portent sur le même domaine.
2. Une approche « en amont » (Morin, 2007) est utilisée pour extraire les contextes (mots simples seulement) des termes complexes sources ; les mots des vecteurs projections sont simples ou complexes, selon leur traduction dans le lexique de projection. Des vecteurs contextuels sont extraits du corpus cible pour tous les mots simples et tous les bigrammes composés de deux mots lexicaux (99,5 % des équivalents de référence comptant au plus deux composantes).

du domaine médical tirés du MeSH [3]. Les meilleures valeurs de paramètres, selon les expériences de (Laroche et Langlais, 2010) sur 70 configurations différentes, sont :
– Longueur du contexte : la phrase [4].
– Mesure d'association : le ratio log-odds.
– Mesure de similarité entre vecteurs contextuels : le cosinus.
– Taille du lexique bilingue de projection : 9 000 termes (pour des corpus de 90 328 mots (source) et 38 929 mots (cible) en moyenne).
– Contenu du lexique de projection : mots généraux et termes spécialisés.

Tout comme (Prochasson, 2009), (Laroche et Langlais, 2010) ont observé que les mesures d'association et de similarité sont les paramètres ayant la plus grande influence sur la performance et que les différents paramètres s'influencent les uns les autres. D'autres paramètres importants ont fait l'objet d'articles, comme le degré de comparabilité des corpus (Li et Gaussier, 2010; Li *et al.*, 2011) et la fréquence d'occurrence des termes (Prochasson et Fung, 2011; Li *et al.*, 2011). Tel que mentionné en introduction, de nombreuses techniques ont récemment été proposées pour améliorer l'approche standard.

3 Analyse des mots contextuels

Nous avons analysé manuellement le contenu des vecteurs projections de 30 termes nominaux simples ou complexes choisis arbitrairement (mais commençant par la lettre *a* [5]), et ce, pour quatre tailles de lexiques de projection différentes (5 000, 7 000, 9 000 et 11 000 entrées, dont 2 000 du domaine médical et le reste de langue générale) ; l'analyse porte donc sur 120 vecteurs projections. Ces vecteurs ont été obtenus en utilisant les ressources, l'implémentation et la configuration paramétrique optimale de (Laroche et Langlais, 2010). Rappelons que les vecteurs projections correspondent aux mots voisins des termes en langue source qui sont projetés vers la langue cible à l'aide du lexique bilingue de projection. Le Tableau 1 présente le contenu de quelques-uns des vecteurs étudiés (obtenus avec le lexique de projection de 5 000 entrées).

Nous examinons la répartition des parties du discours de même que le degré de spécificité (langue spécialisée ou générale) et la sémantique des 20 plus forts mots contextuels des 30 termes français [6]. Ceci nous permet d'identifier des critères pouvant améliorer la qualité de l'identification d'équivalents. Cette analyse est complémentaire à l'inspection manuelle des mots contextuels dans l'approche standard menée par (Morin, 2007) pour 100 termes simples du domaine de la foresterie, qui est centrée sur l'apport des termes complexes dans les vecteurs contextuels en langue source et les vecteurs projections. Ses résultats ne peuvent pas être directement comparés aux nôtres, puisque notre implémentation n'extrait du corpus source que des termes simples pour peupler les vecteurs contextuels.

3. Les ressources utilisées dans l'étude en question sont disponibles sur `http://olst.ling.umontreal.ca/~audrey/coling2010`

4. Selon (Prochasson, 2009; Rubino et Linarès, 2011), la longueur optimale dépend de la fréquence du terme source.

5. *Abcès* (*abscess*), *acétylène* (*acetylene*), *acétylcholine* (*acetylcholine*), *accident vasculaire cérébral* (*stroke*), *acide lactique* (*lactic acid*), *acides* (*acids*), *adhésifs* (*adhesives*), *aine* (*groin*), *albinisme* (*albinism*), *allèles* (*alleles*), *alliages* (*alloys*), *aloès* (*aloe*), *amiante* (*asbestos*), *amidon* (*starch*), *amnésie* (*amnesia*), *amphétamines* (*amphetamines*), *analyse harmonique* (*Fourier analysis*), *anatomie* (*anatomy*), *anesthésie* (*anesthesia*), *anorexie mentale* (*anorexia nervosa*), *antigenes* (*antigens*), *antioxydants* (*antioxidants*), *anxiété* (*anxiety*), *apnée* (*apnea*), *appendicite* (*appendicitis*), *artère pulmonaire* (*pulmonary artery*), *artères* (*arteries*), *articulations* (*joints*), *atrophie* (*atrophy*).

6. (Laroche et Langlais, 2010) étudient en détail la pondération statistique des mots contextuels.

Terme source	Vecteur projection
albinisme	*syndrome*, *Parkinsonism*, *deficit*, *hypoplasia*, *anomaly*, *ocular*, *corpus luteum*, *pigments*, *absence*, *mutation*, *origin*, *retina*, *nystagmus*, *nobody*, *pigmentation*, *abatement*, *synthesis*, *lyophilisate*
artères	*aorta*, *pulmonary*, *Parkinsonism*, *cardiac*, *coronary*, *circulation*, *infarction*, *fact*, *risk patient*, *vascular*, *fat*, *afterload*, *members*, *network*, *hypertension*, *myelosuppression*, *myocardium*, *arterial*, *lyophilisate*, *fabrics*
artère pulmonaire	*pulmonary*, *cardiac*, *aorta*, *afterload*, *arterial*, *duct*, *ventricular*, *coronary*, *venous*, *hypertension*, *Parkinsonism*, *systolic*, *function*, *stenosis*, *absence*, *fat*, *fact*, *circulation*, *diameter*, *anomaly*

TABLE 1 – Contenu des vecteurs projections

3.1 Parties du discours

Les noms sont fortement majoritaires dans les vecteurs projections. Par exemple, avec le lexique de projection de 5 000 entrées, 80,9 % des mots contextuels dans les vecteurs sont des noms. Par comparaison, ce même lexique de projection compte 68,7 % de noms. Rappelons que dans nos expériences, les termes dont on cherche l'équivalent sont des noms ou des syntagmes nominaux (de deux composants) ; il serait intéressant de voir, avec de nouveaux équivalents de référence, si la majorité des mots contextuels seraient aussi des noms si les termes dont on cherche l'équivalent étaient d'une autre partie du discours.

D'autre part, il n'y a pas de différence significative quant aux proportions des parties du discours des mots contextuels selon que les termes dont on cherche l'équivalent sont simples (par ex. *artères*) ou complexes (par ex. *artère pulmonaire*), ces deux types de termes étant traités de la même façon (c'est-à-dire d'un seul bloc) dans notre implémentation.

3.2 Spécificité

Pour analyser le degré de spécificité des mots contextuels dans les vecteurs projections, nous avons classé chacun d'entre eux dans l'une de trois catégories : « domaine médical », « langue générale » ou « ambigu ». Certains mots sont ambigus parce qu'ils appartiennent à la fois à la langue générale et à la langue médicale. Par exemple, dans le vecteur projection d'*anatomie* se trouve le mot *fingers*, que nous avons considéré comme ambigu. Cette tâche est relativement difficile étant donné que nous ne sommes ni spécialiste du domaine médical, ni terminologue, ni anglophone ; certaines tendances se dessinent tout de même (Figure 1).

En moyenne, 57,5 % des mots contextuels sont du domaine médical lorsque le lexique de projection compte 5 000 entrées, bien que celui-ci contienne 36 % d'entrées du domaine médical : les mots contextuels projetés et qui ont un score d'association fort avec le terme source ont donc tendance à être des mots spécialisés. La quantité de termes spécialisés dans les vecteurs projections dépend tout de même de la proportion de termes spécialisés dans le lexique de projection, comme le montre la Figure 1 (36 % des entrées sont du domaine médical dans le lexique de taille 5 000, contre 16 % dans le lexique de 11 000 entrées).

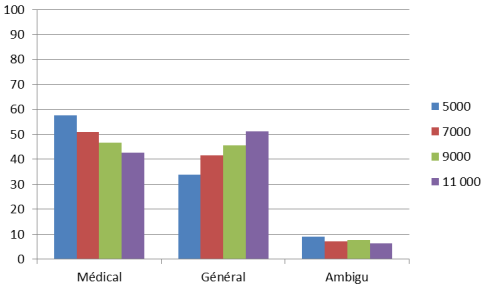


FIGURE 1 – Répartition des domaines dans les vecteurs projections (en %) selon la taille du lexique de projection

Avec notre échantillon de 30 termes en langue source, il n'y a pas de corrélation entre la proportion de mots contextuels qui appartiennent au domaine médical et le rang du bon équivalent. Toutefois, en utilisant l'approche standard pour identifier les équivalents de 122 termes simples du domaine médical, (Morin et Prochasson, 2011) ont montré que les résultats sont meilleurs lorsque le lexique de projection contient des termes spécialisés en plus de mots de la langue générale. Puisque le contenu du lexique de projection a une influence directe sur le contenu des vecteurs projections, il est possible qu'avec un échantillon de taille comparable nous observions un effet analogue ; de plus amples analyses seraient nécessaires pour le vérifier.

3.3 Relations sémantiques

Les mots faisant partie du vecteur projection sont souvent sémantiquement liés au terme source. Selon nos observations sur un sous-ensemble de 10 termes, les mots contextuels peuvent être en relation de collocation (*amnésie* et *infantile*), de synonymie (*anorexie mentale* et *anorexia*) ou d'hyponymie (*artères* et *aorta*) avec le terme source, ils peuvent lui être liés morphologiquement (*anatomie* et *anatomical*), avoir un lien sémantique spécifique au domaine médical (par exemple *est un symptôme de* ou autres relations plus difficiles à caractériser comme celle entre *albinisme* et *pigmentation*), n'être pas liés sémantiquement au terme source ou encore être trop génériques. Les relations spécifiques au domaine médical sont les plus fréquentes (moyenne de 7,3/20 mots contextuels en utilisant le lexique bilingue de projection de 5 000 entrées), suivies des relations de collocation (moyenne de 4,7/20). Les mêmes mots contextuels de sémantisme faible (comme *nobody*) ou qui ne sont pas liés sémantiquement au terme source reviennent dans plusieurs vecteurs ; les premiers pourraient être éliminés en les incluant dans un antidictionnaire, et la présence des seconds peut s'expliquer par certaines faiblesses de nos ressources, tel qu'expliqué dans la prochaine section.

3.4 Cas problématiques

Dans environ 10 % des vecteurs projections examinés, les domaines d'où proviennent les mots contextuels sont très disparates. Par exemple, le vecteur projection obtenu pour *aine* (avec le lexique de projection de 5 000 entrées) est formé des mots : *grand mal*, *meadow*, *radio*, *music*, *region*, *quarter*, *network*, *yearly*, *policy*, *family*, *country*, *origin*, *population*, *venous*, *dance*, *diffusion*, *orange*, *contest*, *prince*, *hockey*. Les candidats équivalents pour *aine* sont à leur tour très diversifiés sémantiquement. Plusieurs éléments peuvent expliquer cette disparité : la façon dont le corpus a été construit, son contenu, la polysémie.

Nous avons remarqué, au cours de l'analyse, certaines lacunes concernant le lexique bilingue de projection (dans lequel les mots contextuels retenus figurent nécessairement). Dans ce lexique, l'équivalent de *maladie* est *Parkinsonism*, celui de *double* est *doubleblind*, celui de *mal* est *grand mal*, celui de *corps* est *corpus luteum*, celui de *produit* est *lyophilisate*, celui de *risque* est *risk patient*. Ces équivalents du lexique bilingue de projection sont beaucoup plus spécifiques que les termes français. Le mot *Parkinsonism* se retrouve ainsi dans plusieurs vecteurs projections (comme ceux des trois exemples du Tableau 1), mais, étant donné qu'il est rare dans les corpus anglais, il ne figure virtuellement pas dans les vecteurs contextuels des candidats équivalents. La performance de l'identification d'équivalents pourrait sans doute être améliorée si le lexique bilingue de projection était d'une meilleure qualité (Morin et Prochasson, 2011).

Enfin, les mots contextuels sont parfois redondants, étant donné que, dans notre implémentation, ils ne sont pas lemmatisés (ceci afin de ne pas faire dépendre l'approche standard d'outils externes qui ne sont pas entraînés sur des corpus médicaux). Pratiquement tous les chercheurs lemmatisent leur corpus avant de former les vecteurs contextuels, et notre examen manuel des mots projetés leur donne raison.

4 Analyse des candidats équivalents

Nous avons analysé manuellement les 20 premiers candidats équivalents de 30 termes français (les mêmes que ceux de la section 3) obtenus avec différents paramètres dans les expériences de (Laroche et Langlais, 2010), pour un total de 360 listes de 20 candidats. Avec la configuration paramétrique optimale, pour 18 des 30 termes examinés, le bon équivalent n'est pas au premier rang. On y trouve plutôt 3 candidats correspondant à une composante du terme complexe attendu (*acid* pour *acide lactique*), 1 qui ne diffère de l'équivalent de référence que par la morphologie (*joint* pour *articulations*), 11 qui sont sémantiquement liés à l'équivalent de référence (*eating* pour *anorexie mentale*), 2 qui sont des mots génériques (*causes* pour *albinisme* et *species* pour *antioxydants*) et 1 qui n'a aucun lien avec l'équivalent de référence (*combo* pour *aine*).

Pour toutes les valeurs de paramètres testées dans (Laroche et Langlais, 2010), il y a systématiquement environ deux candidats sur 20 qui sont des termes complexes, et ce, peu importe si l'équivalent de référence est simple ou complexe. Les termes équivalents d'une langue à l'autre n'ont pas toujours la même complexité (*accident vasculaire cérébral* et *stroke*) (Morin et Daille, 2004). Mais le fait que, peu importe les valeurs des paramètres, le système récupère le même nombre de candidats équivalents complexes suggère qu'il pourrait être amélioré, par exemple en extrayant préalablement les termes (simples et complexes) dans les corpus sources et cibles (Daille et Morin, 2005). Par ailleurs, dans plusieurs cas, les composantes des termes complexes

(ex. *fatty* et *acids*) sont situées à de meilleurs rangs que l'équivalent de référence (*fatty acids*) dans la liste des candidats équivalents. Ceci peut être attribué à notre implémentation, dans laquelle des vecteurs contextuels en langue cible sont construits à la fois pour les termes complexes (bigrammes) et pour chacune de leurs composantes. Une heuristique pourrait donner plus de poids au terme complexe dans ces cas.

Dans presque toutes les listes de candidats équivalents, de un à cinq (environ) candidats parmi les 20 sont morphologiquement liés à l'équivalent de référence. L'étiquetage des parties du discours et la lemmatisation (que font plusieurs chercheurs) permettraient de regrouper les candidats dont seule la flexion varie (par ex., *acids* et *acid*) pour ensuite proposer comme équivalent le candidat qui a le même nombre que le terme source. Le fait que l'approche par projection permette de trouver des candidats liés morphologiquement (comme *oxygen*, *oxidative*, *antioxidant* et *reactive oxygen* pour *antioxydants*) montre que les indices contextuels sont pertinents pour trouver automatiquement les flexions et les dérivations d'un mot donné.

Tous nos équivalents de référence sont des noms ou des syntagmes nominaux ; or, environ 25 % des candidats équivalents ont une autre partie du discours. De façon générale, les parties du discours ne sont pas toujours identiques entre les termes équivalents dans les langues distinctes (Névéol et Ozdowska, 2006), mais, étant donnés notre paire de langues source et cible (français et anglais) et notre domaine (le lexique médical), il serait justifié de réordonner les candidats équivalents pour favoriser les termes nominaux.

Enfin, dans pratiquement toutes les listes de candidats équivalents observées, au moins 10 (et souvent au-delà de 15) des 20 candidats sont sémantiquement liés à l'équivalent de référence. Ces candidats seraient pertinents pour construire des thésaurus, des dictionnaires de synonymes ou d'analogies, etc. Parmi les autres candidats, ceux qui sont des mots génériques comme *process* et *levels* pourraient être inclus dans un antidictionnaire. Si les candidats équivalents ne sont pas génériques, mais appartiennent à des domaines très différents, cela indique que le terme source est polysémique et devrait être désambiguïsé.

5 Conclusion

L'approche standard par projection de contextes pour l'identification d'équivalents terminologiques en corpus comparables est une technique sur laquelle se basent plusieurs travaux récents qui proposent des stratégies pour améliorer la précision. Nous avons analysé manuellement, selon des critères linguistiques, le contenu des vecteurs projections et les listes de candidats équivalents obtenus avec l'implémentation de (Laroche et Langlais, 2010) appliquée à 30 termes du domaine médical. Les mots contextuels ont souvent la même partie du discours que le terme source, ils ont tendance à être spécialisés et à être liés au terme source par des relations spécifiques au domaine et par la collocation ; la qualité des ressources a une influence directe sur celle des vecteurs projections. Parmi les candidats équivalents, ceux qui ont le même nombre, le même degré de complexité et la même partie du discours que le terme source sont à privilégier (du moins pour le domaine médical). Le fait que la très grande majorité des candidats soient sémantiquement liés à l'équivalent de référence confirme l'intérêt de l'approche basée sur la projection de contextes. Les travaux futurs vérifieront l'influence sur la performance des heuristiques proposées ici et de la sélection de mots contextuels en fonction des caractéristiques que nous avons fait ressortir.

Remerciements

Nous remercions Raphaël Rubino pour les ressources ainsi que Philippe Langlais, Patrick Drouin et les relecteurs pour leurs commentaires pertinents. Nous reconnaissons le soutien du FQRSC.

Références

CHARTON, E. et TORRES-MORENO, J.-M. (2010). Nlgbase : A free linguistic resource for natural language processing systems. *In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*.

DAILLE, B. et MORIN, E. (2005). French-English terminology extraction from comparable corpora. *In 2nd International Joint Conference on Natural Language Processing*, pages 707–718.

LAROCHE, A. et LANGLAIS, P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. *In Proceedings of the 23rd International Conference on Computational Linguistics*, pages 617–625.

LI, B. et GAUSSIER, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. *In Proceedings of the 23rd International Conference on Computational Linguistics*, pages 644–652.

LI, B., GAUSSIER, E., MORIN, E. et HAZEM, A. (2011). Degré de comparabilité, extraction lexicale bilingue et recherche d'information interlingue. *In Actes de TALN 2011*.

MORIN, E. (2007). Apport des termes complexes à l'acquisition lexicale multilingue à partir de corpus comparables spécialisés : entre intuition et réalité. *In Actes, 7ème Rencontres Terminologie et Intelligence Artificielle*, pages 11–20.

MORIN, E. et DAILLE, B. (2004). Extraction de terminologies bilingues rtir de corpus comparables. *Traitement automatique des langues*, 45(3):103–122.

MORIN, E. et PROCHASSON, E. (2011). Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. *In Proceedings of the 4th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web*, pages 27–34.

NÉVÉOL, A. et OZDOWSKA, S. (2006). Terminologie médicale bilingue anglais/français : usages cliniques et bilingues. *Glottopol*, 8.

PROCHASSON, E. (2009). *Alignement multilingue en corpus comparables spalisés : Caractérisation terminologique multilingue*. Thèse de doctorat, Université de Nantes.

PROCHASSON, E. et FUNG, P. (2011). Rare word translation extraction from aligned comparable documents. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 1327–1335.

RAPP, R. (1999). Automatic identification of word translations from unrelated English and German corpora. *In 37th Annual Meeting of the Association for Computational Linguistics*, pages 66–70.

RUBINO, R. (2011). *Traduction automatique statistique et adaptation à un domaine spécialisé*. Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse.

RUBINO, R. et LINARÈS, G. (2011). Une approche multi-vue pour l'extraction terminologique bilingue. *In Conférence en Recherche d'Infomations et Applications*, pages 97–111.

BiTermEx : un prototype d'extraction de mots composés à partir de documents comparables via la méthode compositionnelle

Emmanuel Planas[1,2]
(1) UNAM, LINA, 2 rue de la Houssinière, BP 92208, 44322 Nantes
(2) UNAM, UCO, ST, 3, place André Leroy, 49008 Angers
emmanuel.planas@univ-nantes.fr

RÉSUMÉ

Nous décrivons BiTermEx, un prototype d'expérimentation de l'extraction de terminologie bilingue de mots composés, à partir de documents comparables, via la méthode compositionnelle. Nous expliquons la variation morphologique et la combinaison des constituants lexicaux des termes composés. Cette permet une précision TOP1 de 92% et 97,5% en français anglais, et de 94% en français japonais pour l'alignement de termes composés (textes scientifiques et de vulgarisation scientifique).

ABSTRACT

BiTermEx , A prototype for the extraction of multiword terms from comparable documents through the compositional approach.

We describe BiTermEx, a prototype for extracting multiword terms from comparable corpora, using the compositional method. We focus on morphology-based variations of multiword constituents and their recombinaison. We experimented our approach on scientific and popular science corpora. We record TOP1 precisions of 92% and 97,5% on French to English alignments and 94% on French to Japanese.

Mots-clés : extraction terminologique, prototype, terminologie bilingue, documents comparables, méthode compositionnelle, mots composés, corpus.

Keywords : term extraction, prototype, bilingual terminology, comparable documents, compositional method, multiword terms, corpus.

1 Introduction

Les documents comparables sont caractérisés par le partage d'un ensemble significatif de termes traduits en commun, tels les articles de Wikipédia relatifs à un sujet. (Déjean & Gaussier, 2011) en donnent cette définition : « Deux corpus de deux langues l1 et l2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue l1, respectivement l2, dont la traduction se trouve dans le corpus de langue l2, respectivement l1. » Ils sont plus difficiles à utiliser pour un alignement de termes que les documents parallèles qui sont, eux, des traductions l'un de l'autre (ex : les manuels de téléphones portables en plusieurs langues). Ceci provient du fait que les premiers ne présentent pas de repères positionnels et distributionnels présents dans les derniers. Ils ont cependant un avantage important : leur nombre nettement plus élevé (Fung, 1998).

En outre, Internet en est un réservoir important, qui s'incrémente quotidiennement.

2 Principes d'extraction de terminologie de corpus comparables

Les méthodes d'alignement de termes d'une langue à l'autre sont habituellement classées en deux grandes catégories : la **méthode contextuelle** (Fung, 1998) et la **méthode compositionnelle** (Robitaille et al., 2006). Elles reposent sur quatre phases.

La **Phase 1** de collecte les corpus sources et cibles. Elle peut être automatique, semi-automatique ou manuelle, éventuellement guidée par une liste fermée de termes sources et de leur traduction (graines), comme par exemple dans (Robitaille et al., 2006). Des outils ont été précédemment développés pour cette tache d'extraction, comme BootCat (Baroni & Bernardini, 2004.), ou Babouk (De Groc, 2011).

La **Phase 2** a pour but d'extraire une liste de termes sources et cibles candidats, souvent après une phase de pré-traitement du texte extrait des corpus ; léger : un effacement de mots vides (Fung, 1995), ou plus profond : analyse syntaxique chez (Yu & Tsujii, 2009). L'identification de termes candidats peut se faire : de façon monolingue, par exemple à l'aide de patrons de catégories grammaticales comme dans (Takeuchi et al., 2009) ; qui utilisent Acabit (Daille, 2003) ; ou encore de façon bilingue, comme dans (Fung, 1998), via la recherche d'une corrélation entre les termes source et cible.

La différentiation entre les deux grandes méthodes (**contextuelle et compositionnelle**) s'effectue en **Phase 3**. Celle-ci a pour mission de construire une liste de candidats cibles associés à un candidat source dont on cherche la traduction.

Dans la **méthode contextuelle**, cette construction se fait par le rapprochement statistique de **contextes** construits autour des mots sources d'une part et des mots cibles d'autre part. Les « contextes » peuvent être des vecteurs lexicaux (Fung, 1998) ou syntaxiques (Yu & Tsujii, 2009). La méthode contextuelle s'applique mieux aux termes simples car ceux-ci ont en général une fréquence plus importante que les termes complexes, qui peut être exploitée par les méthodes statistiques.

Dans la **méthode compositionnelle**, les termes sources dont on cherche la traduction sont «transposés» en des candidats cibles par construction morphologique, lexicale, syntaxique, ou sémantique à partir de leurs composants lexicaux (Morin & Daille, 2009). L'ensemble des combinaisons des composants est généré et traduit pour obtenir des candidats cibles.

C'est à la **Phase 4** que sont sélectionnées la où les meilleures traductions. En contextuel, cela s'effectue souvent à l'aide d'une mesure de similarité entre les contextes source et cible ; dans la méthode compositionnelle, les candidats résultent d'une « construction théorique ». La sélection des meilleurs candidats peut alors se faire par simple « projection » : on ne retient les candidats qui apparaissent dans la liste des termes cibles.

Pour compléter ce tour d'horizon, on pourra consulter (Laroche & Langlais, 2010) qui passent en revue l'ensemble des facteurs de la méthode contextuelle, et (Morin & Daille, 2009) qui présentent une large vue de la méthode compositionnelle.

3 Description de BiTermEx

BiTermEx est un prototype permettant de tester la méthode d'alignement de terminologie compositionnelle (Morin & Daille, 2009). Il est écrit en Java. Il a été testé sur Linux Ubuntu, Windows XP et 7. Nous décrivons ici les choix théoriques.

3.1 Phase 1 : Identification de corpus comparable

La collecte des corpus se fait actuellement manuellement : l'automatisation sera traitée dans une version ultérieure de l'outil.

3.2 Phase 2 : Extraction de termes candidats sources et cibles

BiTermEx extrait des listes de mots composés candidats via l'application de patrons de catégories grammaticales et de lemmes. Ceci est réalisé après une catégorisation et identification des lemmes relatifs aux mots du corpus. Cette analyse est obtenue par l'intégration du TreeTagger d'Helmut Schmid (Schmid, 1994). En post traitement de l'analyse de Tree Tagger, nous effectuons une standardisation des étiquettes des catégories grammaticales, de telle façon à pouvoir exprimer les patrons d'extraction de terminologie dans un langage commun à l'ensemble des langues traitées (Ex : français : DET:ART → DET@art ; anglais: RB → DET).

Chaque phrase du texte est exprimée comme une chaîne de caractères résultant de la concaténation de l'analyse de chacun de ses mots (la catégorie grammaticale et le lemme), en voici un exemple ; les mots sont séparés par des tirets bas :

…._cat=VERB@be3sp:lem=be_cat=DET:lem=a_cat=NOUN@sing:lem=design_cat=NOUN@sing:lem=concept_cat=PREP@in:lem=for_cat=DET:lem=a_…

Les règles d'extraction de terminologie sont du type suivant :

[cat1=DET_cat2=NOUN_cat3=NOUN_lem4=for#lem2_lem3]

Dans cet exemple, la correspondance entre le texte analysé et la règle d'extraction se fait successivement sur le *DET/a*, *NOUN/design*, *NOUN/concept*, et *lem/for*. Le terme extrait est *design concept*.

Cette méthode permet non seulement d'identifier des patrons, mais aussi de les contextualiser (*design concept* est extrait entre *DET/a* et *lem/for*). Nous avons ici une contribution à la question de la « termicité » (« termhood » en anglais : la séquence extraite est-elle vraiment un « vrai » terme ?) des termes candidats extraits (Robitaille et al., 2006). De plus, cette méthode est facilement adaptable : les règles peuvent être modifiées ou ajoutées par simple édition d'un fichier externe.

3.3 Phase 3 : Génération d'une liste de candidats cibles pour chacun des termes composés sources

3.3.1 Modifications de l'approche classique

i) Modification des unités lexicales par variation morphologique

Supposons que l'un des termes sources extraits en Phase 2 soit *production annuelle*, lemmatisé en *production annuel,* et que dans le dictionnaire bilingue, *production* soit traduit par *production* et *output*, et *annuel* soit traduit par *yearly* et *annual*. Alors l'approche compositionnelle consiste en phase 3 à combiner les traductions : production *yearly, yearly production, production annual, annual production, output yearly, yearly output, output annual, annual output*. Et la phase 4 à sélectionner les termes qui apparaissent dans la liste de termes composés cibles extraite en Phase 2 : *yearly production* et *annual production*. Le seul fait qu'ils apparaissent dans le texte source étant discriminant.

Mais pour le terme *biologie tumorale* qui est traduit en anglais par *tumor biology*, le substantif *tumor* n'est pas la traduction de l'adjectif *tumorale*, mais du substantif dérivé de la même famille : *tumeur*.

Cette difficulté peut être résolue par variation graphique, morphologique, lexicale, ou syntaxique des éléments constituant les termes composés (Morin & Daille, 2009). Nous traitons les dérivations lexicales par l'application de règles de dérivation. Dans le cas de notre exemple, la lemmatisation ayant transformé *biologie tumorale* en *biologie tumoral*, la règle : *oral* → *eur* permet de transformer l'adjectif *tumoral* en le nom *tumeur* qui sera bien traduit par *tumor*. Dans le prototype, les règles sont consignées dans un fichier texte pour chaque langue. L'utilisateur peut ajouter ou modifier ces règles par l'édition de ces fichiers.

ii) Combinaison

Ordre de combinaison

Les unités lexicales des mots composés sources sont combinées pour chacune des variations générées précédemment. Pour une variation fixée, un mot composé ABC de longueur N, présente alors N! combinaisons de longueur N. Voici une illustration :

ABC → **C**AB, A**C**B, AB**C** ; **C**BA, B**C**A, BA**C** (N = 3, 3! = 6)

L'ensemble de ces permutations sont construites par récurrence sur N, illustrée ici par le positionnement de 'C' sur les différentes permutations de AB.

Profondeur de recherche

Dans la récurrence, nous gardons les sous-permutations de dimension N-p, et nommons p : « profondeur de recherche » de combinaison d'unités lexicales. Notons que pour une profondeur p, le nombre de combinaisons est $A^{p}_{N} = (N)\,!/(N\text{-}p)\,!$

Cela permet, après traduction, de générer des termes candidats de longueur différente. C'est une réponse possible au problème de fertilité (Morin & Daille, 2009).

3.3.2 Réduction du nombre de permutations

Chacune des combinaisons sources est traduite. Nous utilisons simplement un lexique bilingue pour effectuer le transfert de la langue source à la langue cible.

Pour réduire le nombre de combinaisons traduite, nous cherchons **avant traduction** l'ensemble des traductions de variantes d'unités lexicales trouvées en (ii) dans le dictionnaire. Les variantes qui ne possèdent pas de traduction, ou dont les traductions n'apparaissent pas parmi les composants des termes cibles extraits en Phase 2 sont éliminées avant combinaison, réduisant de façon importante cette complexité.

3.4 Phase 4 : Projection des candidats cibles sur la liste des Termes Composés Cibles extraits.

La sélection des meilleurs candidats cibles construits en phase 3 se fait par « projection ».

4 Expériences et Résultats

4.1 Expérience 1 : Wikipédia, français anglais, 935 termes candidats

Corpus

Nous avons téléchargé manuellement 13 articles Wikipédia français (source) et 15 anglais (cible) liés au thème de l'énergie éolienne. Après l'application d'un filtre retirant les métadonnées, les balises, et les parties textuelles liés à la navigation de Wikipédia, nous avons obtenu 36077 tokens français et 39761 tokens anglais. L'analyse de TreeTagger nous a permis d'identifier les lemmes et catégories grammaticales.

Extraction de termes monolingues candidats

L'extraction monolingue de candidats sources français donne 846 termes composés français et 935 termes composés anglais de longueur comprise entre 2 et 5 unités lexicales, comme *autonomie énergétique* en français, ou encore *vertical axis wind turbine* en anglais. En voici quelques caractéristiques de description statistique :

	Termes initiaux	**Freq min.**	**Freq Max.**	**Freq. Moy.**	**Freq. = 1**	**Freq. = 2**	**Freq 3-5**	**Freq 6-10**	**Freq > 10**
FR	846	1	62	1,5	703 (83%)	94 (11%)	34 (4%)	9 (1%)	6 (0,7%)
EN	935	1	56	1,4	772 (83%)	121 (13%)	35 (4%)	0 (0%)	7 (0,7%)

TABLE 1 – Répartition statistique des mots composés monolingues candidats – Wikipedia - Éoliennes – FR - EN

Alignement de Termes

Nous utilisons le dictionnaire ELRA français anglais contenant 103.190 entrées françaises correspondant à 238.742 traductions. L'alignement compositionnel entre les 846 termes sources français et les 935 termes cibles anglais est réalisé aux profondeurs 0 (longueur N), 1 (longueur >=N-1, et 2. (>=N-2). L'alignement des termes a été évalué

manuellement. Une erreur est de type A si la traduction est partielle, de type B si la traduction est complètement fausse. Les résultats sont rassemblés ici :

Profondeur	Nb alignés	Nb Erreurs A	Nb Erreurs B	Rappel	Précision	Score F
0	79	1 (1,2%)	5 (6,3%)	9,3%	92,4% (74)	5,9
1	94	2 (2,5%)	5 (6,3%)	11%	91,1% (73)	5,1
2	144	49 (34%)	12 (8,3%)	17%	57,6% (83)	3,8

TABLE 2 – Statistiques d'alignement bilingue suivant la profondeur de recombinaison – Wikipedia - Éoliennes – FR - EN

On enregistre un taux de précision TOP1 de 92,4% en profondeur 0 et de 91,1 en profondeur 1, pour une légère amélioration du rappel. Le traitement de la fertilité est amélioré par la prise en compte de traductions anglaises possédant les mêmes lexèmes que le français, sans préposition. Ex : *production de électricité | electricity production*. Ce taux chute pour la profondeur 2. Voici la répartition des termes alignés suivant leur nombre d'occurrence dans le corpus, pour la profondeur 1 :

	Termes alignés	Freq min.	Freq Max.	Freq. Moy.	Freq. = 1	Freq. = 2	Freq 3-5	Freq 6-10	Freq > 10
FR	94	1	62	4,4	53 (56%)	12 (13%)	12 (13%)	7 (7%)	10 (11%)
EN	94	1	23	2,5	63 (67%)	13 (13%)	11 (12%)	0 (0%)	7 (7%)

TABLE 3 – Répartition des mots composés alignés – Wikipedia - Éoliennes – FR - EN

Plus de la moitié de l'effectif sont des hapax, et deux tiers sont de fréquence inférieure ou égale à 2. Cela montre que la méthode compositionnelle est très adaptée à l'alignement de termes de très basse fréquence. On note cependant, en comparant ce tableau au tableau de répartition de l'ensemble de l'effectif, que les mots composés de plus hautes fréquences ont proportionnellement une tendance à mieux s'aligner que ceux de basse fréquence : le quartile des mots composés alignés de fréquence supérieure à 10 est de 11% pour le français, alors que seulement 0,7% de l'ensemble des mots composés français extraits en Phase 2 ont une telle fréquence. Il est intéressant de noter que pour ce corpus, contrairement au corpus suivant, l'application de règles morphologiques ne fourni que peu de résultats : 3 termes 94.

4.2 Expérience 2 : Corpus médical, français anglais

Le module d'alignement de termes étant indépendant du module d'extraction de termes du corpus, nous avons pu tester l'alignement en profondeur 0 d'une liste de termes déjà extraite, issue d'un corpus médical spécialisé sur le cancer du sein de 3483 termes composés français avec une liste de 6642 termes composés anglais. Aussi, nous ne nous intéressons ici qu'à l'analyse de l'alignement de termes composés.

Alignés	classique	morpho	Erreurs A	Erreurs B	Rappel	Précision
808	702 (87%)	106 (13%)	2%	0,5%	24 %	97,5 %

TABLE 4 – Statistique d'alignement de mots composés – Cancer du sein – FR – EN

La précision est très forte, pour un rappel significatif. On note ici l'importance des variations morphologiques puisqu'elles engendrent 13% des solutions.

4.3 Expérience 3 : Corpus médical, français japonais

Nous avons testé l'alignement entre 23487 mots composés français et 26188 mots composés japonais extraits d'un corpus médical sur le diabète et l'obésité. Ces listes ne sont pas bien nettoyées : des dates et des extractions ne correspondant pas à des termes complets subsistent. Le petit dictionnaire généraliste utilisé est celui de Jean-Marc Desperrier[1] (18039 entrées françaises, 32444 traductions japonaises). Nous n'avons pas de données sur le degré de comparabilité de ce corpus. Les résultats confirment la bonne précision de la méthode, et l'importance des variations morphologiques.

Alignés	classique	morpho	Erreurs A	Erreurs B	Rappel	Précision
140	85 (61%)	55 (39%)	5%	3%	0,6%	92%

TABLE 5 – Statistique d'alignement de mots composés – Diabète - Obésité – FR – JP

5 Conclusions et perspectives

Cette étude montre la très bonne précision de la méthode compositionnelle, en particulier pour les hapax et les termes très peu fréquents. Nos résultats confirment ceux de (Morin & Daille, 2009) qui obtiennent des taux de précision de 88 % pour des termes composés de basse fréquence. Ainsi que ceux de (Robitaille et al., 2006) qui obtiennent des précisions comprises entre 49% et 92% sur une extraction français – japonais. Ces derniers mesurent un rappel élevé, calculé sur une population restreinte des termes français et japonais vérifiant dans chaque langue une cohérence forte avec une liste de graines bilingues (test de Jacquard $>=$ 0,01), alors que nous n'imposons pas de telle restriction. Les raisons générales du faible rappel ont été identifiées dans des travaux précédents : non-compositionnalité des termes composés et couverture des dictionnaires bilingues (Morin & Daille 2009).

Remerciements

Ce travail, qui s'inscrit dans le cadre du projet METRICC (www.metricc.com), a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-08-CORD-009. Merci à Koichi Takeuchi pour sa coopération et à Emmanuel Morin et Béatrice

[1] http://dico.fj.free.fr/dico.php [15 janvier 2012]

Daille pour leurs conseils.

Références

BARONI, M., & BERNARDINI, S. BOOTCAT (2004). Bootstrapping corpora and terms from the web. Dans E. L. R. A. Elra (Éd.), *Proceedings of LREC* (Vol. 2004, p. 1313–1316). ELRA.

DAILLE, B. (2003). Conceptual Structuring through Term Variations. *Proceeding MWE '03 Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment - Volume 18*.

DE GROC, C. (2011). Babouk : Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. *Proceedings of the IEEEWICACM International Conferences on Web Intelligence*, 497–498.

DÉJEAN, H., & GAUSSIER, E. (2011). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables hal. *Lexicometra*.

FUNG, P. (1995). Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus. *Proceedings of the Third Workshop on Very Large Corpora* (p. 173–183).

FUNG, P. (1998). A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora. *Parallel Text Processing* (p. 1–17). Springer.

LAROCHE, A., & LANGLAIS, P. (2010). Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. *COLING* (p. 617-625).

MORIN, E., & DAILLE, B. (2009). Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation*.

ROBITAILLE, X., SASAKI, Y., TONOIKE, M., SATO, S., & UTSURO, T. (2006). Compiling French-Japanese Terminologies from the Web. *EACL*.

SCHMID, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, United Kingdom.

TAKEUCHI, K., KAGEURA, K., KOYAMA, T., DAILLE, B., & ROMARY, L. (2009). Pattern Based Term Extraction Using ACABIT System. *CoRR*.

YU, K., & TSUJII, J. (2009). Extracting Bilingual Dictionary from Comparable Corpora with Dependency Heterogeneity. *HLT-NAACL (Short Papers)* (p. 121-124).

Combinaison d'approches pour l'extraction automatique d'événements

Laurie Serrano[1, 2] Thierry Charnois[1] Stephan Brunessaux[2]
Bruno Grilheres[2] Maroua Bouzid[1]
(1) Laboratoire GREYC, Université de Caen Basse-Normandie
Campus Côte de Nacre, Boulevard du Maréchal Juin, BP 5186 - 14032 Caen
(2) Département IPCC, Cassidian
Parc d'Affaires des Portes - 27600 Val de Reuil
prenom.nom@unicaen.fr, prenom.nom@cassidian.com

RÉSUMÉ

Dans cet article, nous présentons un système d'extraction automatique d'événements fondé sur deux approches actuelles en extraction d'information : la première s'appuie sur des règles linguistiques construites manuellement et la seconde se fonde sur un apprentissage automatique de patrons linguistiques. Les expérimentations réalisées montrent que combiner ces deux méthodes d'extraction permet d'améliorer significativement la qualité des événements extraits (amélioration de près de 10 points de F-mesure).

ABSTRACT

Automatic events extraction by combining multiple approaches

In this paper, we present an automatic system for extracting events based on the combination of two existing information extraction approaches : the first one is made of hand-crafted linguistic rules and the second one is based on an automatic learning of linguistic patterns. We have shown that this mixed approach leads to a significant improvement of extraction performances.

MOTS-CLÉS : Extraction d'information, événements, approche symbolique, apprentissage de patrons linguistiques.

KEYWORDS: Text mining, events, symbolic extraction, linguistic pattern learning.

1 Introduction

Face à l'augmentation vertigineuse des informations disponibles librement (en particulier sur le Web), repérer efficacement celles qui peuvent nous intéresser s'avère une tâche longue et complexe. En réponse à cela, l'équipe IPCC [1] développe le WebLab [2], une plateforme d'intégration de services de "media mining" [3] pour la découverte de connaissances et l'aide à la décision. Nous présentons dans cet article une étude comparative de trois approches d'extraction d'événements : une approche symbolique basée sur des règles linguistiques construites manuellement, une méthode d'apprentissage de patrons linguistiques et une approche mixte. Les expériences menées

1. Information Processing, Control and Cognition, Cassidian
2. http://weblab-project.org/, consulté le 21/03/2012
3. Fouille de documents multimédia

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 423–430,
Grenoble, 4 au 8 juin 2012.

montrent que la combinaison des deux premières méthodes permet d'améliorer significativement la qualité des événements extraits. Les travaux présentés ici sont réalisés dans le cadre de l'élaboration d'un système plus global de capitalisation des connaissances exploitant les technologies du Web sémantique (Serrano *et al.*, 2012). Précisons également que les événements que nous souhaitons extraire ont été au préalable définis dans une ontologie de domaine (nommée WOOKIE [4]). Nous proposons, dans un premier temps, un rapide tour d'horizon des travaux existants en extraction d'événements. Puis, nous décrivons les trois approches proposées, les expérimentations mises en place, nos premiers résultats et les perspectives envisagées.

2 Extraction des événements : tour d'horizon

L'extraction d'information est une discipline récente qui consiste à analyser un texte de manière automatique afin d'en extraire un ensemble d'informations jugées pertinentes (Poibeau, 2003). Deux approches dites "classiques" émergent : l'extraction basée sur des techniques linguistiques et les systèmes statistiques. Les tâches les plus communes en extraction d'information sont l'extraction d'entités nommées (Nadeau et Sekine, 2007), de relations entre entités et d'événements. L'extraction d'événements est particulièrement utilisée dans les activités de veille économique et stratégique ((Capet *et al.*, 2011) pour la détection de crise). Celle-ci peut-être conçue comme une forme particulière d'extraction de relations où une "action" est liée à d'autres entités telles qu'une date, un lieu, des participants, *etc*. Plusieurs campagnes MUC [5] s'y sont intéressé avec notamment des tâches de remplissage automatique de formulaires ("template filling"). Comme en extraction d'information de façon générale, la littérature du domaine offre à la fois des travaux basés sur des approches symboliques et des techniques purement statistiques. (Aone et Ramos-Santacruz, 2000) développe REES, un extracteur d'événements basé sur des règles linguistiques construites manuellement couplées à une analyse syntagmatique. Dans la lignée, (Grishman *et al.*, 2002) s'intéresse à la détection d'événements épidémiques au moyen d'un transducteur à états finis. Toutefois, ces méthodes purement linguistiques, bien que généralement très précises, ont pour principales faiblesses d'être spécifiques à un domaine donné, d'avoir un taux de rappel plutôt faible et un coût de développement manuel élevé. Du côté des approches statistiques, (Ahn, 2006) propose de combiner plusieurs classifieurs pour l'extraction des événements dans la campagne ACE. L'apprentissage statistique permet de prévoir de nombreux contextes d'apparition mais nécessite une grande quantité de données annotées pour être performant et construit un modèle de type "boite noire" non-accessible et non-modifiable. Face à cela, les méthodes d'apprentissage de patrons ou les approches semi-supervisées apparaissent intéressantes comme par exemple le système de (Xu *et al.*, 2006). Observant que toutes ces approches prises séparément restent imparfaites, nous proposons d'élaborer une approche hybride permettant d'exploiter les points forts des méthodes "classiques". Pour cela, nous avons choisi, de compléter les performances d'un extracteur d'événements symbolique par un système d'apprentissage de patrons linguistiques.

3 Modélisation des événements

L'événement étant l'objet central de nos travaux, il est nécessaire de définir plus précisément ce concept. Considéré comme une entité aux propriétés spécifiques, l'événement a particulièrement

4. Weblab Ontology for Open sources Knowledge and Intelligence Exploitation
5. Message Understanding Conference

été étudié en philosophie (Davidson, 2001) et en linguistique (Van De Velde, 2006). Après avoir considéré différents travaux, nous prenons pour point de départ la définition de (Krieg-Planque, 2009) qui nous paraît adaptée : "un événement est une occurrence perçue comme signifiante dans un certain cadre". Afin de proposer une représentation plus formelle d'un événement, nous nous appuyons sur les travaux de (Saval *et al.*, 2009) qui propose une extension sémantique pour la modélisation des événements de type "catastrophes naturelles". Celui-ci définit un événement E comme la combinaison de 3 composantes : une propriété sémantique S, un intervalle temporel I, et une entité spatiale SP. Un événement est donc représenté sous la forme $E\langle I, SP, S\rangle$. Dans notre cas, la propriété sémantique est définie par les différents types d'événement de notre ontologie (que nous décrivons plus loin), la composante temporelle constitue la date ou période d'occurrence d'un événement et l'entité spatiale correspond à son lieu d'occurrence. Nous proposons d'adapter cette représentation à notre domaine d'application en l'enrichissant d'une composante supplémentaire A correspondant aux différents participants impliqués dans l'événement. Nous avons donc maintenant $E\langle I, SP, S, A\rangle$ où A est un ensemble de participants jouant un ou plusieurs rôle(s). Un participant est noté P_i où $0 \leq i < n$ et un rôle est noté r_j où $0 \leq j < k$. La composante A est donc définie de la façon suivante : $A = \{(P_\alpha, r_\beta)\}$ tel que le participant P_α joue le rôle r_β dans l'événement en question. Cette modélisation a été implémentée au sein de notre ontologie de domaine WOOKIE, centrée sur 5 classes supérieures : "Event", "Person", "Unit", "Place" et "Equipment". Les différentes approches comparées dans cet article visent à extraire la vingtaine d'événements suivants : "AttackEvent", "BombingEvent", "ShootingEvent", "CrashEvent", "DamageEvent", "DeathEvent", "FightingEvent", "InjureEvent", "KidnappingEvent", "MilitaryOperation", "ArrestOperation", "HelpOperation", "PeaceKeepingOperation", "SearchOperation", "SurveillanceOperation", "TrainingOperation", "TroopMovementOperation", "NuclearEvent", "TrafficEvent".

4 Extraction des événements : approches proposées

4.1 Approche à base de règles linguistiques

L'approche que nous présentons ici a été implémentée grâce à la plateforme d'ingénierie textuelle GATE [6] et vise à extraire un ensemble d'événements tels que définis ci-dessus ainsi que les participants et circonstants suivants : la date de l'événement, son lieu d'occurrence et les entités de type "personne" et "organisation" impliquées. La figure 1 résume les différentes étapes de notre approche (le repérage des rôles ne sera pas traité dans cet article, pour une description détaillée se référer à (Serrano *et al.*, 2011)). Notre système repose sur une chaine de traitement composée de différents modules d'analyse linguistique ("tokenisation", découpage en phrases, repérage lexical, étiquetage grammatical, analyse syntaxique, transducteur à états finis, *etc.*). Nous définissons tout d'abord un ensemble de termes considérés comme possibles déclencheurs d'événement (dits "noms d'événement"). Nous choisissons de nous limiter, pour l'instant, aux déclencheurs verbaux et nominaux et de constituer des listes de lemmes, plus courtes et permettant d'étendre le repérage à toutes les formes fléchies. Ces déclencheurs (139 lemmes actuellement) sont répartis en différentes listes, chacune étant associée à un type d'événement (c'est-à-dire à une classe d'événement de notre ontologie) afin d'être repérés et annotés dans le corpus à analyser. Nous associons ensuite à ces "noms d'événement" les différentes entités qu'ils impliquent. Pour

6. General Architecture for Text Engineering

cela, nous effectuons, dans un premier temps, une extraction automatique d'entités nommées [7] ainsi qu'une analyse en constituants syntaxiques (syntagmes verbaux, nominaux, *etc.*). Enfin, ces différents éléments sont rattachés au "nom d'événement" grâce à une analyse syntaxique en dépendance (réalisée par le Stanford Parser [8]) ainsi que des règles de grammaire élaborées manuellement (dans le langage JAPE [9]). A l'heure actuelle, nous obtenons une annotation positionnée sur le "nom d'événement" résumant ses différents participants et circonstants et indiquant pour chacun d'eux s'il correspond à une entité nommée détectée précédemment.

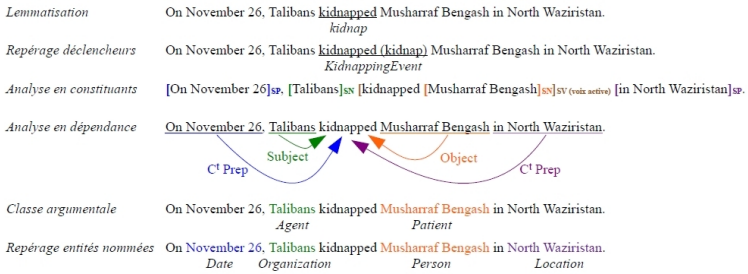


FIGURE 1 – Chaine d'extraction d'événements pour l'anglais

4.2 Apprentissage de patrons linguistiques

Par ailleurs, nous nous intéressons à l'extraction d'événements par une technique d'extraction de motifs séquentiels fréquents. Ce type d'approche permet d'apprendre automatiquement des patrons linguistiques compréhensibles et modifiables par un expert linguiste. La découverte de motifs séquentiels a été introduite par (Agrawal *et al.*, 1993) dans le domaine du "data mining" et adaptée par (Cellier et Charnois, 2010) à l'extraction d'information dans les textes. Ceux-ci s'intéressent en particulier à l'extraction de motifs séquentiels d'itemsets. Il s'agit de repérer, dans un ensemble de séquences, des enchainements d'items ayant une fréquence d'apparition supérieure à un seuil donné (dit "support"). La recherche de ces motifs s'effectue dans une base de séquences ordonnées d'itemsets où chaque séquence correspond à une unité de texte (ici la phrase). Un itemset est un ensemble d'items décrivant un mot de cette séquence. Un item correspond à une caractéristique particulière de ce mot telle que la catégorie grammaticale, le lemme, la forme fléchie, *etc*. Un certain nombre de paramètres peuvent être adaptés selon l'application visée : nature de la séquence et des items, nombre d'items, support, *etc*. La fouille sur un ensemble de séquences d'itemsets permet l'extraction de motifs combinant plusieurs types d'item et d'obtenir ainsi des patrons génériques, spécifiques ou mixant les informations (ce qui n'est pas permis par les motifs d'items simples), comme par exemple les patrons suivants : <homme de culture> <homme de N> <N PRP N> [10], *etc*. De plus, contrairement aux différentes approches que nous venons de mentionner, l'apprentissage de patrons ne nécessite ni corpus annoté avec les entités-cibles, ni analyse syntaxique. Cela constitue un réel avantage car, tout

7. Extraction des dates, lieux, personnes et organisations réalisée par une chaine GATE (Serrano *et al.*, 2011)
8. `http://nlp.stanford.edu/software/lex-parser.shtml`
9. Java Annotation Patterns Engine
10. N pour la catégorie nom, PRP pour préposition

d'abord, l'annotation manuelle de corpus reste un effort important et l'analyse syntaxique est encore une technologie aux performances inégales et peu disponible librement selon les langues. Le point faible partagé par toutes ces méthodes d'apprentissage symbolique reste le nombre important de motifs extraits. Pour pallier ce problème, (Cellier et Charnois, 2010) propose l'ajout de contraintes pour diminuer la quantité de motifs retournés. Dans la lignée de ces travaux, nous utilisons l'outil d'extraction de motifs séquentiels développé au GREYC (Béchet *et al.*, 2012). Celui-ci présente plusieurs points forts : il extrait uniquement des motifs dits "clos" (c'est-à-dire non redondants) et génère ainsi moins de motifs que d'autres systèmes. De plus, ce logiciel s'avère robuste et permet la fouille de séquences d'itemsets, fonctionnalité qui est rarement proposée par les outils existants. Nous avons adapté la fouille de motifs à notre domaine d'application et au traitement de dépêches de presse dans le but d'obtenir des patrons linguistiques permettant la détection d'événements. Ainsi, notre approche propose tout d'abord de pré-traiter un corpus grâce à l'outil TreeTagger [11] afin d'obtenir un découpage en séquences (ici en phrases) ainsi que différents types d'items : forme fléchie, lemme, catégorie grammaticale. Elle nécessite également une annotation sémantique en entités nommées. Enfin, nous effectuons un repérage lexical des "noms d'événement" et de leur type. Comme prévu, le nombre de motifs retournés par l'outil s'avère élevé, nous introduisons donc un ensemble de contraintes spécifiques à notre application : des contraintes linguistiques d'appartenance (nous pouvons par exemple choisir de ne retourner que des motifs contenant au moins un "nom d'événement" et une date) mais aussi une contrainte dite de "gap" (Dong et Pei, 2007), autorisant l'extraction de motifs ne contenant pas nécessairement des itemsets consécutifs (contrairement aux n-grammes dont les éléments sont strictement contigus). Ainsi un "gap" d'une valeur maximale n signifie qu'au maximum n itemsets (mots) sont présents entre chaque itemset du motif dans les séquences correspondantes. Cette approche non-supervisée nécessite une sélection manuelle des motifs pertinents. Pour cela, nous utilisons l'outil Camelis (Ferré, 2009) permettant d'ordonner et visualiser les motifs des plus généraux aux plus spécifiques puis de filtrer les plus pertinents. Les motifs ainsi sélectionnés sont ensuite appliqués sur un nouveau corpus afin d'en extraire les relations visées.

4.3 Vers une approche mixte

Nous travaillons actuellement à définir une méthode d'hybridation qui permette d'exploiter les forces de chacune des approches présentées. En effet, comme nous l'avons déjà souligné, l'extraction d'information à base de règles écrites manuellement s'avère généralement très précise mais peu couvrante alors que les techniques d'apprentissage montrent habituellement un meilleur rappel. Nous proposons donc, dans un premier temps, d'effectuer une simple union des résultats des deux extracteurs afin de maximiser le rappel de notre approche mixte (les expérimentations reportées dans cet article sont basées sur cette approche).

5 Expérimentations

Les expérimentations présentées consistent à extraire un ensemble d'événements d'un corpus journalistique en anglais et dont le thème est d'intérêt pour le renseignement. Nous nous focalisons sur une vingtaine de types d'événement (*cf* section 3) et sur les relations suivantes : la date de l'événement, son lieu d'occurrence et les actants impliqués (personnes et organisations).

11. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

5.1 Corpus et paramètres d'apprentissage

Cette évaluation a nécessité les deux corpus suivants :

– Le premier corpus est un corpus d'apprentissage nécessaire à la mise en place de l'outil d'extraction de motifs séquentiels fréquents : il s'agit d'un corpus de textes anglais abordant une thématique militaire et annoté avec des entités nommées et des "noms d'événement". Nous avons constitué ce corpus de manière semi-automatique à partir de 400 dépêches de presse sur l'engagement du Canada en Afghanistan [12] et de 700 dépêches parues entre 2003 et 2009 sur le site de l'ISAF [13]. Ce corpus a, dans un premier temps, été annoté automatiquement en entités nommées et "noms d'événement" grâce à notre outil d'extraction basé sur GATE (Serrano *et al.*, 2011), puis nous avons revu manuellement ces annotations pour corriger les éventuels erreurs/oublis et ainsi garantir la qualité des données d'apprentissage.
– Le second corpus est un corpus de test permettant de comparer notre extraction d'événements par rapport à une vérité-terrain. Pour cela nous avons choisi d'utiliser un corpus fourni dans la campagne d'évaluation MUC-4 et constitué de 100 dépêches de presse relatant des faits terroristes en Amérique du Sud. Notre évaluation porte sur une partie de ce corpus annotée manuellement [14], soit environ 210 événements et près de 240 relations (55 relations de type "date", 65 relations de type "lieu" et 120 relations de type "participant").

Pour mettre en place notre approche d'apprentissage symbolique, nous avons, tout d'abord, opéré un apprentissage de motifs séquentiels fréquents sur le premier corpus en considérant quatre caractéristiques (quatre types d'item) : la forme fléchie du mot, sa catégorie grammaticale, son lemme et sa classe sémantique ("nom d'événement", "date", "lieu", "personne" ou "organisation"). Nous avons choisi de réaliser une tâche d'apprentissage par type d'entité impliquée en utilisant le système des contraintes d'appartenance proposé par l'outil de (Béchet *et al.*, 2012). Nous obtenons donc quatre séries de motifs de type "nom d'événement"-"date", "nom d'événement"-"lieu", "nom d'événement"-"personne" et "nom d'événement"-"organisation". Nous avons également procédé à plusieurs essais de paramétrage et, au regard de ces tests, nous avons choisi de fixer un "gap" maximal de 3 itemsets (correspondant à 3 mots possibles entre chaque élément du motif) et un support absolu relativement bas (10 en valeur absolue, soit 6% des séquences pour tous les types de relation) afin d'obtenir des motifs intéressants mais en nombre raisonnable pour une exploration et une validation manuelles (environ 12000 motifs au total).

5.2 Résultats et discussion

Nous avons réalisé manuellement une comparaison des extractions obtenues par chacune des deux approches (appliquée séparément) et celles résultant de l'approche mixte. Le tableau 1 présente les scores de précision, rappel et F-mesure de chaque approche, globalement et par type de relation. Précisons que ces résultats proviennent d'une extraction de relations fondée sur l'annotation manuelle des entités nommées que nous avons réalisée sur le corpus de test (et non pas sur une extraction automatique) afin d'éviter que des erreurs dans l'extraction des entités viennent perturber l'extraction de relations. Nous pouvons constater que l'approche à base de règles et l'apprentissage de motifs obtiennent tous deux une très bonne précision globale et que,

12. `http://www.afghanistan.gc.ca/canada-afghanistan`, consulté le 21/03/2012
13. `http://www.nato.int/isaf/docu/pressreleases`, consulté le 21/03/2012
14. Nous avons choisi de ne pas réutiliser les "templates" de référence fournis avec le corpus car le nombre et le type des événements ne correspondaient pas à notre modélisation et ne permettaient pas d'évaluer la totalité de nos extractions.

	Approche à base de règles manuelles			Apprentissage de motifs			Approche mixte		
	Précision	*Rappel*	*F-mesure*	*Précision*	*Rappel*	*F-mesure*	*Précision*	*Rappel*	*F-mesure*
Date	0,93	0,25	0,39	0,90	0,64	0,75	0,90	0,68	0,78
Lieu	0,92	0,37	0,53	0,86	0,49	0,63	0,81	0,60	0,69
Participants	0,97	0,49	0,42	0,93	0,32	0,47	0,92	0,51	0,66
Toutes relations	0,94	0,37	**0,45**	0,90	0,48	**0,62**	0,88	0,60	**0,71**

TABLE 1 – Extraction d'événements : évaluation des trois approches

comme attendu, le rappel est meilleur pour cette dernière approche. Par ailleurs, nous avons été assez surpris par la bonne précision de la méthode par apprentissage, que nous expliquons par une sélection manuelle restrictive et précise des motifs. Les taux de rappel obtenus sont peu élevés : ce résultat est conforme à l'état de l'art pour l'approche à base de règles. Pour l'approche à base d'apprentissage, l'utilisation d'un "gap" maximal trop restreint ne permet pas d'extraire les relations distantes. Ce qu'il faut retenir de ces expérimentations est que l'approche mixte obtient une F-mesure nettement supérieure (près de 10 points par rapport à la meilleure des deux approches), ce qui dénote une amélioration globale de la qualité d'extraction pour tout type de relation. De plus, nous remarquons que l'apprentissage de patrons complète avec succès notre approche symbolique en augmentant sensiblement le taux global de rappel. Nous constatons cependant une légère perte de précision qui résulte du nombre plus élevé de règles et patrons linguistiques au sein de l'approche mixte entraînant une augmentation des faux positifs.

Apport de l'analyse syntaxique Parallèlement à ces résultats, nous nous sommes intéressés à l'apport de l'analyse syntaxique au sein de notre approche mixte : les résultats du tableau sont issus de notre système avec analyse en dépendance, sans cela nous aurions eu une perte de performances considérable (11 points de F-mesure, 19 points de précision et 1 point de rappel). Bien que les outils d'analyse syntaxique soient inégalement disponibles selon les langues, cette observation confirme l'intérêt de cette technique pour l'extraction d'événements.

Résultats avec repérage automatique des entités nommées Pour compléter les résultats précédents basés sur une annotation manuelle des entités nommées, nous avons évalué notre approche mixte avec une annotation automatique des entités. Nous observons une baisse globale des performances des trois approches et plus particulièrement du taux de rappel bien que les performances restent acceptables (64% de F-mesure pour l'approche mixte). Ce point est important car dans une application réelle d'extraction d'événements, les entités nommées sont toujours repérées par des outils d'extraction automatique.

Améliorations Pour améliorer la combinaison, nous envisageons de tester plusieurs techniques d'hybridation. Tout d'abord, afin d'obtenir de meilleurs résultats de façon plus globale (c'est-à-dire maximiser la F-mesure), nous expérimenterons l'ajout d'un système d'estimation de confiance au sein de nos extracteurs. Cela peut être réalisé par différents moyens : (1) faire évaluer à la main par un expert linguiste chaque règle/motif composant les deux approches précédentes et reporter cette confiance sur les événements/relations extraits ; (2) estimer la confiance de chaque règle/motif automatiquement en réalisant une évaluation préalable. Une dernière piste à explorer est l'apport des approches statistiques : nous souhaitons apprendre automatiquement un modèle de performances qui permettrait une sélection contextuelle d'approche en suggérant, lors du traitement d'un corpus, la meilleure approche à utiliser.

6 Conclusion et perspectives

Dans cet article nous avons proposé une étude comparative de deux approches et de leur combinaison pour l'extraction automatique d'événements. Nos résultats montrent que la méthode

mixte améliore significativement la qualité des événements extraits. Malgré une combinaison plutôt simple, ces résultats sont encourageants et nous invitent à explorer de nouveaux modes d'hybridation afin de tirer le meilleur parti des deux premières approches (améliorer le taux de rappel sans perdre trop en précision).

Références

AGRAWAL, R., IMIELIŃSKI, T. et SWAMI, A. (1993). Mining association rules between sets of items in large databases. SIGMOD '93, New York. ACM.

AHN, D. (2006). The stages of event extraction. ARTE '06, pages 1–8, Stroudsburg, USA. ACL.

AONE, C. et RAMOS-SANTACRUZ, M. (2000). Rees : A large-scale relation and event extraction system. *In ANLP*, pages 76–83.

BÉCHET, N., CELLIER, P., CHARNOIS, T. et CRÉMILLEUX, B. (2012). Discovering linguistic patterns using sequence mining. *In CICLing (1)*, pages 154–165.

CAPET, P., DELAVALLADE, T., GÉNÉREUX, M., POIBEAU, T., SÁNDOR, Á. et VOYATZI, S. (2011). Un système de détection de crise basé sur l'extraction automatique d'événements. *In* et P. HOOGSTOEL, M. C., éditeur : *Sémantique et multimodalité en analyse de l'information*, pages 293–313. Lavoisier.

CELLIER, P. et CHARNOIS, T. (2010). Fouille de données séquentielle d'itemsets pour l'apprentissage de patrons linguistiques. *In TALN (short paper)*.

DAVIDSON, D. (2001). *Essays on Actions and Events*. Oxford University Press.

DONG, G. et PEI, J. (2007). *Sequence Data Mining*. Advances in Database Systems. Kluwer.

FERRÉ, S. (2009). Camelis : a logical information system to organize and browse a collection of documents. *In Int. J. General Systems*, volume 38.

GRISHMAN, R., HUTTUNEN, S. et YANGARBER, R. (2002). Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, 35(4):236–246.

KRIEG-PLANQUE, A. (2009). *A propos des noms propres d'événement*, volume 11, pages 77–90. Les carnets du Cediscor.

NADEAU, D. et SEKINE, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26. Publisher : John Benjamins Publishing Company.

POIBEAU, T. (2003). *Extraction automatique d'information : Du texte brut au web sémantique*. Lavoisier.

SAVAL, A., BOUZID, M. et BRUNESSAUX, S. (2009). A semantic extension for event modelisation. *21st IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2009)*.

SERRANO, L., BOUZID, M., CHARNOIS, T. et GRILHERES, B. (2012). Vers un système de capitalisation des connaissances : extraction d'événements par combinaison de plusieurs approches. *In SOS-DLWD'2012 at EGC'2012*.

SERRANO, L., GRILHERES, B., BOUZID, M. et CHARNOIS, T. (2011). Extraction de connaissances pour le renseignement en sources ouvertes. *In SOS'2011 at EGC'2011*.

VAN DE VELDE, D. (2006). *Grammaire des événements*. Presses Universitaires du Septentrion.

XU, F., USZKOREIT, H. et LI, H. (2006). Automatic event and relation detection with seeds of varying complexity. *In AAAI Workshop Event Extraction and Synthesis*, Boston.

Apprentissage automatique d'un chunker pour le français

Isabelle Tellier[1,2], Denys Duchier[2], Iris Eshkol[3],
Arnaud Courmet[2], Mathieu Martinet[3]
(1) LaTTiCe, université Paris 3 - Sorbonne Nouvelle
(2) LIFO, université d'Orléans
(3) LLL, université d'Orléans
isabelle.tellier@univ-paris3.fr, denys.duchier@univ-orleans.fr,
iris.eshkol@univ-orleans.fr, arnaud.coumet@gmail.com,
mathieu_martinet@hotmail.fr

RÉSUMÉ
Nous décrivons dans cet article comment nous avons procédé pour apprendre automatiquement un chunker à partir du French Tree Bank, en utilisant les CRF (Conditional Random Fields). Nous avons réalisé diverses expériences, pour reconnaître soit l'ensemble de tous les chunks possibles, soit les seuls groupes nominaux simples. Nous évaluons le chunker obtenu aussi bien de manière interne (sur le French Tree Bank lui-même) qu'externe (sur un corpus distinct transcrit de l'oral), afin de mesurer sa robustesse.

ABSTRACT
Machine Learning of a chunker for French

We describe in this paper how to automatically learn a chunker for French, from the French Tree Bank and CRFs (Conditional Random Fields). We did several experiments, either to recognize every possible kind of chunks, or to focus on simple nominal phrases only. We evaluate the obtained chunker on internal data (i.e. also extracted from the French Tree Bank) as well as on external (i.e from a distinct corpus) ones, to measure its robustness.

MOTS-CLÉS : chunking, apprentissage automatique, French Tree Bank, CRF.

KEYWORDS: chunking, Machine Learning, French Tree Bank, CRF.

1 Introduction

Nous présentons dans cet article la démarche ayant permis d'apprendre automatiquement à partir du French Tree Bank (Abeillé *et al.*, 2003) un "chunker" ou analyseur syntaxique superficiel (Abney, 1991) du français. Alors que cette tâche a fait l'objet du challenge CoNLL 2000 [1] pour l'anglais, aucun chunker du français n'avait encore, semble-t-il, été appris automatiquement à partir de données annotées. Ce travail est une suite naturelle à l'acquisition d'un étiqueteur morpho-syntaxique (ou POS) du français réalisée précédemment à partir du même corpus

1. http ://www.cnts.ua.ac.be/conll2000/chunking/

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 431–438,
Grenoble, 4 au 8 juin 2012.

(Constant *et al.*, 2011), et sur lequel il s'appuie. Comme l'étiqueteur POS, notre chunker a été appris à l'aide des CRF (Conditional Random Fields) (Lafferty *et al.*, 2001; Tellier et Tommasi, 2011). Comme lui, il est librement disponible en téléchargement (Tellier *et al.*, 2012).

La notion de chunk peut recouvrir plusieurs niveaux de détails possibles, suivant que l'on se concentre sur les seuls groupes nominaux simples, à la façon de (Sha et Pereira, 2003) ou sur l'ensemble de tous les constituants non récursifs possibles. Ces deux variantes ont été testées et évaluées par validation croisée sur le corpus, en s'appuyant sur un étiquetage POS parfait, et montrent qu'identifier les différents types de chunks est de difficulté variable suivant leur nature. De plus, la variante qui se concentre sur les groupes nominaux simples a aussi été testée sur un autre corpus totalement différent, constitué de transcriptions de l'oral et annoté avec notre étiqueteur, donc imparfaitement. Ces évaluations permettent de mesurer la sensibilité du modèle à des conditions d'utilisation dégradées.

L'article suit la structure suivante. Tout d'abord, nous évoquons la tâche de chunking et ses différentes variantes. Nous décrivons ensuite les différentes instances du French Tree Bank de Paris 7 qui ont permis la constitution du corpus d'apprentissage ainsi que les CRF qui ont été utilisés pour cet apprentissage. Nous fournissons enfin les résultats de nos différentes expériences.

2 Le chunking

2.1 Le chunking du français

Les chunks sont des constituants continus et non-récursifs (Abney, 1991). Ils définissent la structure syntaxique superficielle des phrases et, à ce titre, sont moins coûteux et plus faciles à obtenir que leur structure en constituants complète. Pour certains textes non normés (transcriptions de l'oral par exemple), ils représentent le degré d'analyse le plus poussé qu'on puisse espérer.

A notre connaissance, peu de solutions spécifiques sont disponibles pour le chunking du français, et celles qui existent ont été écrites à la main :
– soit pour réaliser une analyse syntaxique superficielle de textes non normés, en particulier ceux transcrits de l'oral (Antoine *et al.*, 2008; Blanc *et al.*, 2010)
– soit en tant que composant d'un analyseur syntaxique complet, comme par exemple les systèmes ayant participé aux campagnes d'évaluation Easy et Passage (Paroubek *et al.*, 2006)
– soit encore en tant que composant d'une plateforme généraliste et multilingue comme Gate [2]
Nous proposons à la place de coder la tâche de chunking comme une annotation, et de l'apprendre automatiquement à l'aide d'un CRF, en nous inspirant des expériences de (Sha et Pereira, 2003).

2.2 Découpages en chunks

La notion de chunk n'est pas toujours très précisément définie. Deux niveaux de détails sont possibles pour caractériser les chunks :

2. http ://www.semanticsoftware.info/munpex

– soit on s'intéresse aux seuls groupes nominaux simples (i.e. non récursifs), qui sont chacun constitués d'un unique nom ou pronon, incluant ses éventuels groupes adjectivaux immédiats, déterminants et adjectifs numériques. Les compléments du nom sont dans des chunks distincts de celui du nom qu'ils qualifient.
– soit on s'intéresse à tous les groupes possibles, en cherchant à obtenir un parenthésage complet de la phrase. Dans ce cas, les différents types possibles de chunks, tels qu'ils apparaissent dans le French Tree Bank, sont :
  – les groupes nominaux ou NP, définis comme précédemment sauf quand ils sont inclus dans un des autres types suivants ;
  – les groupes verbaux ou VN, incluant les formes interrogatives, infinitives, modales.. ;
  – les groupes prépositionnels ou PP, incluant tous les groupes nominaux introduits par une préposition ainsi que tous ceux qui qualifient les VN ;
  – les groupes adjectivaux ou AP, incluant les éventuels adverbes modifieurs d'adjectifs ;
  – les groupes adverbiaux ou Adp, incluant les modifieurs de phrases ;
  – les groupes coordonnés ou COORD, introduits par une conjonction de coordination, et pouvant aussi inclure des groupes nominaux.

Ces différents chunks peuvent bien sûr être obtenus à partir de la structure en constituants de la phrase. Par exemple, l'arbre de la Figure 1 donne lieu aux deux découpages suivants :
– (La commercialisation efficace)$_{NP}$ est plus exigeante.
– (La commercialisation efficace)$_{NP}$ (est)$_{VN}$ (plus exigeante)$_{AP}$.

Dans le cas des compléments du nom ou des groupes nominaux coordonnés par exemple, le découpage de premier type n'est pas strictement inclu dans celui de deuxième type, comme l'illustre le cas suivant :
– (La commercialisation)$_{NP}$ de (la marchandise)$_{NP}$ et (des services)$_{NP}$ est plus exigeante.
– (La commercialisation)$_{NP}$ (de la marchandise)$_{PP}$ (et des services)$_{COOR}$ (est)$_{VN}$ (plus exigeante)$_{AP}$.

Pour aborder la tâche de chunking comme une tâche d'annotation, il suffit d'associer à chaque mot appartenant à un chunk une étiquette donnant son type (voit NP, soit un type parmi {NP, VN, PP, AP, AdP, VCOOR}) accompagnée du codage BIO (Begin/In/out) qui permet de délimiter ses frontières. Dans le cas d'un parenthésage total, le type O est inutile car la fin d'un chunk coincide toujours avec le début d'un autre :
– La/B-NP commercialisation/I-NP efficace/I-NP est/O plus/O exigeante/O.
– La/B-NP commercialisation/I-NP efficace/I-NP est/B-VN plus/B-AP exigeante/I-AP.

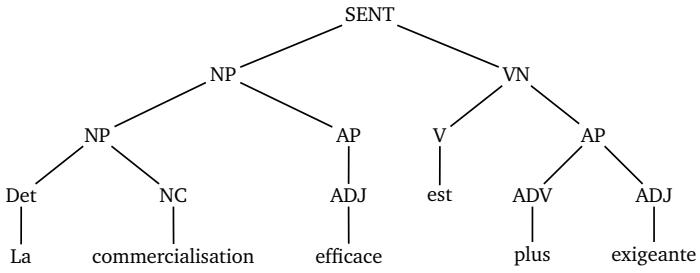


FIGURE 1 – Arbre d'analyse syntaxique extrait du French Tree Bank

3 Constitution de corpus de référence

3.1 Le French Tree Bank

Le French Tree Bank (Abeillé *et al.*, 2003) est la ressource à partir de laquelle nous avons pu constituer des ensembles d'exemples de référence pour l'apprentissage automatique. Ce corpus est composé d'articles du journal "Le Monde". Les phrases qui y figurent sont complètement analysées syntaxiquement en constituants, comme dans la Figure 1. Nous en avons extrait deux variantes de corpus annotés en chunks, correspondant aux deux notions possibles de chunks évoquées précédemment.

3.2 Homogénéisation des étiquettes

Pour ce travail, nous disposions en fait de deux versions complémentaires du corpus :
– la version arborée, composée d'environ dix mille fichiers XML (un par phrase). Ces fichiers décrivent donc la structure syntaxique complète des phrases ainsi que de leurs unités. Les mots sont associés à une liste d'attributs qui les caractérisent (lemme, catégorie, ...).
– une version où ne figurent plus que les mots et leur catégorie morpho-syntaxique, ayant servi à acquérir un étiqueteur (Constant *et al.*, 2011). Son jeu d'étiquettes comprend 29 catégories POS distinctes. Ces catégories ne correspondent pas exactement avec la valeur de l'attribut "cat" associé aux mots de la version arborée (des simplifications ont eu lieu), ce qui nous a containt à quelques prétraitements.

Il était indispenable d'harmoniser les catégories morpho-syntaxiques figurant dans ces deux versions du corpus, car notre chunker doit pouvoir s'appuyer sur l'étiqueteur POS appris précédemment à partir des catégories simplifiées. L'étiqueteur POS utilisé ne prend pas en compte pour l'instant la reconnaissance des unités multimots.

4 Le modèle d'apprentissage

4.1 Les CRF

Les champs markoviens conditionnels ou CRF (Tellier et Tommasi, 2011) sont des modèles probabilistes discriminants introduits par (Lafferty *et al.*, 2001) pour l'annotation de données séquentielles. Ils ont été utilisés dans de nombreuses tâches de traitement automatique des langues, pour lesquelles ils sont particulièrement bien adaptés (McCallum et Li, 2003; Sha et Pereira, 2003; Tsuruoka *et al.*, 2009; Tellier *et al.*, 2010).

Les CRF permettent d'associer à une observation x une annotation y en se basant sur un ensemble d'exemples annotés (x, y). La plupart du temps (et ce sera le cas ici), x est une *séquence d'unités* (ici, une suite d'unités lexicales associées à leur catégorie POS) et y la *séquence des annotations correspondante* (ici, la suite des étiquettes BIO couplées au type des chunks). Ils sont définis par X et Y, deux champs aléatoires décrivant respectivement chaque unité de l'observation x et de son annotation y, et par un graphe dont $V = X \cup Y$ est l'ensemble des nœuds (vertices) et $E \subseteq V \times V$ l'ensemble des arcs (edges). Deux variables sont reliées dans le graphe

si elles dépendent l'une de l'autre. Le graphe sur le champ Y des CRF linéaires est une simple chaîne qui traduit le fait que chaque étiquette est supposée dépendre de l'étiquette précédente et de la suivante et, implicitement, de la donnée x complète.

Dans un CRF linéaire, on a la relation suivante :

$$p(y|x) = \frac{1}{Z(x)} \prod_{c \in \mathscr{C}} \exp\Big(\sum_k \lambda_k f_k(y_i, x, c) \Big) \quad avec$$

- $Z(x)$ est un coefficient de normalisation, défini de telle sorte que la somme sur y de toutes les probabilités $p(y|x)$ pour une donnée x fixée soit égale à 1.
- $\mathscr{C}$ est l'ensemble des cliques (sous-graphes complètement connectés) de $\mathscr{G}$ sur Y : dans le cas des CRF linéaires, ces cliques sur Y sont constituées soit d'un nœud isolé, soit d'un couple de nœuds successifs.
- Les fonctions f_k sont appelées *fonctions caractéristiques* (features) : elles sont définies à l'intérieur de chaque clique c et sont à valeurs réelles, mais souvent choisies pour donner un résultat binaire (0 ou 1). Elles doivent être founies au système par l'utilisateur. Par définition, la valeur de ces fonctions peut dépendre des annotations y_c présentes dans une certaine clique c ainsi que de la valeur de x *n'importe où dans la donnée* (et pas uniquement aux indices correspondants à la clique c, ce qui donne beaucoup d'expressivité aux CRF).
- y_c est l'ensemble des valeurs prises par les variables de Y sur la clique c pour une annotation y donnée : ici, c'est donc soit la valeur y_i d'une seule étiquette soit celles d'un couple d'étiquettes successives (y_i, y_{i+1}).
- Les poids λ_k, à valeurs réelles, permettent d'accorder plus ou moins d'importance à chaque fonction f_k dont ils caractérisent le *pouvoir discriminant*. Ce sont les paramètres du modèle : l'enjeu de la phase d'apprentissage est de fixer leur valeur en cherchant à maximiser la log-vraisemblance sur l'ensemble des exemples annotés constituant le corpus d'apprentissage.

Le logiciel que nous avons utilisé est Wapiti avec pénalisation L1 (Lavergne *et al.*, 2010), reconnu comme actuellement le plus efficace pour les CRF linéaires, car il procède lors de sa phase d'apprentissage à une *sélection* des features les plus discriminantes.

4.2 Features utilisées

Pour nos expériences, nous nous sommes contenté de features simples. Le seul attribut associé aux mots M est leur catégorie POS, notée C. A chaque fois qu'un couple $x_i = (M_i, C_i)$ est associé à une étiquette $y_i = E_i$ à une position i dans le corpus d'apprentissage, on crée une feature "unigramme" (c'est-à-dire ne prenant en compte qu'une seule étiquette) caractérisant l'association du mot et de l'étiquette, ainsi qu'une autre caractérisant l'association de la catégorie et de l'étiquette. On fait de même avec chacun des mots situés dans une fenêtre de taille 5 (de deux places avant à deux places après) centrée sur le mot courant. Les features "bigrammes" (c'est-à-dire portant sur un couple d'étiquettes successives) sont construites de la même façon, en ne tenant compte que des catégories et pas des mots, parce qu'elles varient moins que ces derniers. Les features sont donc toutes les configurations attestées dans les exemples de la forme suivante :

- $f_{1,i,j}(y_i, x) = 1$ si $y_i = E_i$ et $mot_j = M_j$, $\forall j \in [i-2, i+2]$ (=0 sinon)
- $f'_{1,i,j}(y_i, x) = 1$ si $y_i = E_i$ et $POS_j = C_j$, $\forall j \in [i-2, i+2]$ (=0 sinon)
- $f_{2,i,j}(y_i, y_{i+1}, x) = 1$ si $y_i = E_i$ et $y_{i+1} = E_{i+1}$ et $POS_j = C_j$, $\forall j \in [i-2, i+2]$ (=0 sinon)

5 Les résultats

5.1 Validation interne

Les premières évaluations ont été réalisés par validation croisée en répartissant le corpus d'apprentissage dans 5 ensembles distincts, 4 servant pour l'apprentissage et 1 pour le test. Dans chacun de ces ensembles, on dispose d'un étiquetage POS parfait, puisqu'il est lui-même issu du French Tree Bank. Un chunk est considéré comme reconnu si à la fois ses frontières et son type sont corrects.

Les seuls "groupes nominaux simples" NP sont identifiés avec une précision de 97,49%, un rappel de 97,40% et une F-mesure de 97,45. Ces excellents résultats dépassent ceux obtenus pour la tâche CoNLL 2000 sur l'anglais (où les meilleurs dépassaient à peine 94 points de F-mesure), mais ces comparaisons sont à prendre avec précautions, car ni les données ni le jeu de catégories POS utilisées n'étaient les mêmes.

Il faut remarquer que, dans le cas du chunking complet, une erreur de frontière rend erronés les deux chunks que cette frontière devrait séparer. Les taux d'erreurs sont donc naturellement globalement plus bas :

type de chunk	proportion (%)	Précision (%)	Rappel (%)	F-mesure
AP	10	68,36	68,61	68,49
AdP	2	53,57	39,47	45,45
COORD	6	80,81	76,35	78,52
NP	26	84,99	86,10	85,54
PP	34	77,79	77,82	77,81
VN	22	83,3	85,52	84,39

Ce tableau montre que les NP sont les mieux reconnus, mais avec tout de même près de 12 points de F-mesure de moins que quand ils sont la seule cible, les goupes adverbiaux étant quant à eux à la fois les plus rares et les plus difficiles à identifier. La "micro-average" (moyenne des F-mesure pondérées par les effectifs des différents chunks) vaut 79,73, tandis que la "macro-average" (moyenne donnant autant d'importance à chaque type de chunk, indépendamment de sa fréquence d'apparition) vaut : 73,37. Il n'y a pourtant pas toujours corrélation entre la fréquence d'un chunk et sa propention à être reconnu. Ainsi, PP est le type de chunks le plus fréquent car il couvre à la fois les compléments du nom qui suivent un NP et les groupes prépositionnels associés à un VN. Cette variabilité de construction explique sans doute la relative difficulté à les retrouver. Inversement, les COORD sont assez rares, mais comme ils doivent être nécessairement introduits par une conjonction de coordination, ils ne sont pas si durs à repérer.

Les résultats de notre système de chunking complet sont moins bons que ceux habituellement obtenus par les analyseurs syntaxiques complets (qui peuvent atteindre une exactitude d'environ 85%) : la simple identification des chunks est apparemment plus difficile quand elle n'est pas couplée avec celle des relations qu'ils entretiennent les uns avec les autres.

Nous aurions pu mesurer l'importance de la catégorie POS sur cette identification en cherchant à retrouver les chunks à partir d'un corpus annoté par notre étiqueteur, c'est-à-dire imparfaitement. Cependant, cet étiqueteur POS a été appris sur ce même corpus, il y fait moins de 2 points d'erreur en exactitude (puisqu'il n'en faisait déjà pas beaucoup plus en validation croisée), et l'ef-

fet de ces très rares erreurs sur le chunking sera donc difficile à mesurer. A la place, nous avons testé le résultat final du traitement : étiquetage + chunking NP sur un corpus complètement différent.

5.2 Test sur un corpus oral

Afin de tester la robustesse du chunker qui se concentre sur les groupes nominaux simples dans un contexte différent de celui dans lequel il a été appris, nous avons évalué ses performances sur un extrait du corpus de transcriptions orales ESLO [3]. Le corpus a été annoté en POS avec SEM [4], l'étiqueteur POS appris sur le French Tree Bank, sans que les catégories fournies par ce programme ne soient corrigées. Seuls les résultats du chunking ont, eux, été vérifiés à la main. Sur un corpus comprenant 575 "phrases" (i.e. tours de parole ou "groupe de souffle") et environ 9 280 mots, la performance de notre chunker tombe à moins de 40 en F-mesure, très loin de ses 97,45 points obtenus par validation croisée. L'exactitude de l'étiquetage B_NP est d'environ 56%, celui des I_NP de 61%.

Il n'est pas facile d'analyser la raison de ces résultats. Certaines erreurs semblent provenir de la segmentation qui n'est pas traitée par notre étiqueteur POS : les mots composés, entités nommées ou expressions figées devraient rester dans le même chunk et ne pas être considérés comme des compléments du nom. Les irrégularités propres à l'oral (disfluences, hésitations, amorces) sont aussi courantes et rendent bien sûr l'étiquetage POS moins fiable (même si nous n'avons pas mesuré la qualité de l'étiquetage POS indépendamment de celle du chunking), donc la reconnaisance des chunks plus délicate. En fait, la notion même de chunk doit être amendée dans ce contexte. En effet, quand le nom principal d'un chunk est oralement répété, les deux formes transcrites sont incluses dans le même chunk qui comporte donc deux noms, ce qui est en principe interdit par notre définition des NP. Si la répétition d'un déterminant ne provoque pas un changement de chunk NP, en revanche celle d'un pronom en entraîne un : est-ce toujours souhaitable ? Et doit-on considérer que des interruptions comme "heu", "oui", "ah bon" doivent être incluses dans le chunk NP qui les englobe, le découper en deux NP distincts ou en constituer un nouveau à part ? Le statut syntaxique de ces formes propres à l'oral reste sujet à discussion.

6 conclusion

Dans cet article, nous avons présenté comment obtenir efficacement deux variantes de chunkers du français par apprentissage automatique à partir du French Tree Bank.

Nos expériences montrent que la tâche de chunking est de difficulté très variable en fonction du contexte dans lequel on l'applique. La reconnaissance des NP seuls dans des textes normés ne pose pas de problèmes, mais ils sont difficiles à distinguer des autres groupes qui peuvent aussi intégrer des noms dans le cas d'un chunking complet. Enfin, la robustesse d'un chunker acquis par apprentissage automatique est très limitée quand on l'applique à des types de textes présentant des propriétés très différentes. La notion même de chunking est peut-être à préciser dans le cas des corpus oraux.

3. http ://eslo.in2p3.fr
4. http ://www.lattice.cnrs.fr/sites/itellier/SEM.html

Il nous reste à étudier en quoi la reconnaissance des unités multimots dans la phase préliminaire d'étiquetage modifie ou non les propriétés du chunking, et à repérer les dépendances entre chunks, pour se rapprocher des performances des analyseurs syntaxiques profonds. Il est aussi envisageable d'apprendre directement un segmenteur-étiqueteur POS-chunker en une seule étape, afin d'éviter de cumuler les erreurs.

Références

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for french. *In* Abeillé, A., éditeur : *Treebanks*. Kluwer, Dordrecht.

Abney, S. (1991). Parsing by chunks. *In* Berwick, R., Abney, R. et Tenny, C., éditeurs : *Principle-based Parsing*. Kluwer Academic Publisher.

Antoine, J.-Y., Mokrane, A. et Friburger, N. (2008). Automatic rich annotation of large corpus of conversational transcribed speech : the chunking task of the epac project. *In Proceedings of LREC'2008*.

Blanc, O., Constant, M., Dister, A. et Watrin, P. (2010). Partial parsing of spontaneous spoken french. *In Proceedings of LREC'2010*.

Constant, M., Tellier, I., Duchier, D., Dupont, Y., Sigogne, A. et Billot, S. (2011). Intégrer des connaissances linguistiques dans un CRF : application à l'apprentissage d'un segmenteur-étiqueteur du français. *In Actes de TALN'11*.

Lafferty, J., McCallum, A. et Pereira, F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *In Proceedings of ICML 2001*, pages 282–289.

Lavergne, T., Cappé, O. et Yvon, F. (2010). Practical very large scale CRFs. *In Proceedings of ACL'2010*, pages 504–513. Association for Computational Linguistics.

McCallum, A. et Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. *In Proceedings of CoNLL'03*.

Paroubek, P., Robba, I., Vilnat, A. et C., A. (2006). Data annotations and measures in easy, the evaluation campain for parsers of french. *In Proceedings of LREC'2006*, pages 315–320.

Sha, F. et Pereira, F. (2003). Shallow parsing with conditional random fields. *In Proceedings of HLT-NAACL 2003*, pages 213 – 220.

Tellier, I., Dupont, Y. et Courmet, A. (2012). Un segmenteur-étiqueteur et un chunker pour le français. *In Actes de TALN'12, session démo*.

Tellier, I., Eshkol, I., Taalab, S. et Prost, J. P. (2010). Pos-tagging for oral texts with crf and category decomposition. *Research in Computing Science*, 46:79–90. Special issue "Natural Language Processing and its Applications".

Tellier, I. et Tommasi, M. (2011). Champs Markoviens Conditionnels pour l'extraction d'information. *In* Eric Gaussier et François Yvon, éditeurs : *Modèles probabilistes pour l'accès à l'information textuelle*. Hermès.

Tsuruoka, Y., Tsujii, J. et Ananiadou, S. (2009). Fast full parsing by linear-chain conditional random fields. *In Proceedings of EACL 2009*, pages 790–798.

Effacement de dimensions de similarité textuelle pour l'exploration de collections de rapports d'incidents aéronautiques

Tulechki Nikola[1,2] Tanguy Ludovic[1]
(1) CLLE-ERSS : CNRS et Université de Toulouse 2, 5 allées Antonio Machado, 31058 Toulouse CEDEX 9
(2) Conseil en Facteurs Humains, 4 impasse Montcabrier. 31500 Toulouse
{tanguy,tulechki}@univ-tlse2.fr

RÉSUMÉ

Cet article étudie le lien entre la similarité textuelle et une classification extrinsèque dans des collections de rapports d'incidents aéronautiques. Nous cherchons à compléter les stratégies d'analyse de ces collections en établissant automatiquement des liens de similarité entre les documents de façon à ce qu'ils ne reflètent pas l'organisation des schémas de codification utilisés pour leur classement. Afin de mettre en évidence les dimensions de variation transversales à la classification, nous calculons un score de dépendance entre les termes et les classes et excluons du calcul de similarité les termes les plus corrélés à une classe donnée. Nous montrons par une application sur 500 documents que cette méthode permet effectivement de dégager des thématiques qui seraient passées inaperçues au vu de la trop grande saillance des similarités de haut niveau.

ABSTRACT

Deletion of dimensions of textual similarity for the exploration of collections of accident reports in aviation

In this paper we study the relationship between external classification and textual similarity in collections of incident reports. Our goal is to complement the existing classification-based analysis strategies by automatically establishing similarity links between documents in such a way that they do not reflect the dominant organisation of the classification schemas. In order to discover such transversal dimensions of similarity, we compute association scores between terms and classes and exlude the most correlated terms from the similarity calculation. We demonstrate on a 500 document corpus that by using this method, we can isolate topics that would otherwise have been masked by the dominant dimensions of similarity in the collection.

MOTS-CLÉS : similarité textuelle, classification de documents, corpus spécialisé.

KEYWORDS: textual simliarity, document classification, specialised corpora.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 439–446,
Grenoble, 4 au 8 juin 2012.

1 Introduction et contexte applicatif

Dans toute industrie à risque, le retour d'expérience (REX) occupe une place capitale dans les mécanismes de gestion de la sûreté. Des politiques de recueil, d'analyse et de stockage sont mises en place afin de garder une trace de tout évènement qui s'écarte de la norme, de tout incident ou accident qui survient lors des opérations. Les informations ainsi recueillies servent ensuite de support aux experts de sûreté pour mettre à jour les règles et les procédures d'exploitation en les adaptant à un contexte en perpétuelle évolution.

1.1 Texte et codification des rapports d'incidents aéronautiques

L'objet de notre étude est un sous-ensemble particulier de REX, les rapports de type Aircraft Safety Report (ASR) recueillis dans le service de sécurité de la compagnie aérienne Air France. Les ASR sont des textes relativement courts (105 mots en moyenne) rédigés par les pilotes eux-mêmes immédiatement ou peu après qu'un incident s'est produit, et décrivant celui-ci en langage libre. Lorsqu'ils sont soumis, ces rapports sont saisis dans la base de données de la compagnie et enrichis d'un certain nombre d'informations factuelles, telles que le modèle de l'avion, les conditions météo, la localisation ou encore le poids de l'appareil le jour de l'incident.

Ensuite les rapports subissent une première analyse visant à "coder" l'évènement suivant un schéma préétabli. Un schéma de codification est une abstraction d'un scénario d'accident, composée de plusieurs taxonomies de codes en rapport avec différents aspects d'un accident. En pratique, l'expert en charge du codage doit décrire l'évènement en utilisant quelques centaines de codes, à partir de listes fermées. (Voir (Ponvert, 2009) pour les détails de l'élaboration et la mise en place du schéma de codification actuel d'Air France). Une fois codés, les rapports sont stockés dans la base de données et peuvent être interrogés *via* des requêtes portant sur les informations factuelles et la codification. Un expert peut ainsi, par exemple, extraire de la base l'ensemble d'incidents, où il y a eu une panne du radar météo, dans un Boeing 747 survenue lorsque l'avion était en phase de montée initiale.

1.2 Limites de la codification

Avec du recul, on peut voir le procédé de codification comme un effort visant à maîtriser la variation inhérente des rapports afin d'atteindre un niveau d'abstraction suffisamment stable pour une exploitation informatisée d'une base de REX. Sans rentrer dans les détails, nous dirons que cet effort est nécessairement accompagné d'un appauvrissement du contenu informationnel directement accessible aux experts. Le fait de réduire un texte à un squelette prédéterminé a pour effet de ne garder que les éléments les plus saillants de l'évènement au détriment de subtilités qui, tout en étant présentes dans le texte original, ne trouvent pas leur place dans la codification.

Une autre limite de ces stratégies est leur caractère intrinsèquement réactif. Un schéma de codification est une représentation de la réalité figée à un instant précis, alors que la réalité qu'elle reflète est en perpétuelle évolution. Toute changement majeur du contexte doit être reflété dans le schéma, ce qui correspond à un effort considérable et prends un temps précieux aux experts, pendant lequel un risque nouvellement apparu peut se trouver sans code associé.

1.3 Objectifs applicatifs

Conscients des limites des stratégies d'analyse des REX par codification, notre objectif est de concevoir des techniques et outils venant en complément de ces stratégies et permettant aux experts d'explorer les collections de rapports en fonction des particularités de leur contenu textuel et de leur distribution chronologique. S'affranchissant de la rigidité de la codification, dans l'idéal ces outils devront être capables d'alerter leurs usagers de configurations particulières d'évènements, de tendances émergentes ou encore d'évènements anormaux (Tulechki, 2011).

2 Similarité textuelle

Dans un premier temps nous avons cherché à rapprocher les textes en fonction de leur contenu en utilisant des méthodes classiques en recherche d'information (RI) : la similarité cosinus (Salton *et al.*, 1975), une mesure du recouvrement lexical qui attribue un score compris entre 0 et 1 à chaque paire de documents dans la collection. Un score de 0 signifie une absence de termes en commun et un score de 1 une identité complète du contenu lexical des deux textes. Ce score est obtenu en calculant le cosinus entre deux vecteurs dans un espace à n dimensions correspondant aux termes présents dans la collection.

En plus de son utilisation immédiate dans des applications de type moteur de recherche, ce calcul permet de superposer automatiquement une couche de structure sur une collection et transformer un matériau symbolique et qualitatif en des données numériques et ouvre la voie à d'autres traitements comme l'apprentissage non supervisé (Steinbach *et al.*, 2000), ou encore la détection d'anomalies (Chandola *et al.*, 2009) pour en citer quelques uns. Toutefois à l'heure actuelle, nous avons préféré tout d'abord évaluer l'apport *per se* de la similarité textuelle en développant un outil utilisant ce calcul et en le soumettant aux experts de sûreté.

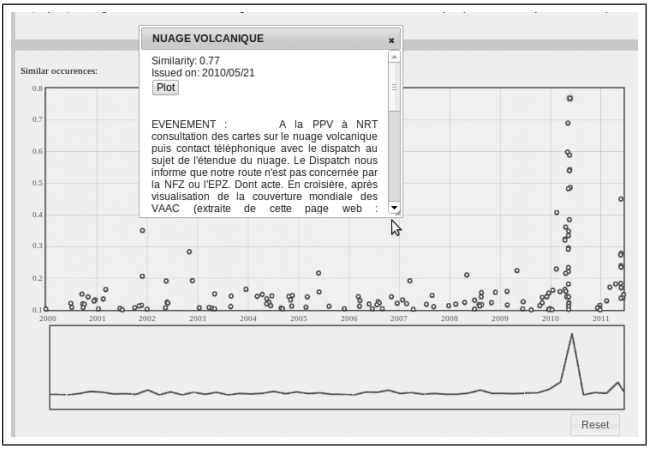


FIGURE 1 – Similarité de rapports d'incidents de sur un axe chronologique

L'outil *timePlot* présenté en figure 1 permet, à partir d'un rapport pivot, de visualiser les rapports similaires et leur distribution dans le temps. Les rapports sont présentés à l'utilisateur sur un graphique interactif qui permet un accès direct à leur contenu. Différentes configurations chronologiques peuvent apparaître, comme ici le pic de rapports associés à l'éruption volcanique du printemps 2010, ou encore des phénomènes saisonniers, comme par exemple des incidents liés à la neige qui, naturellement, apparaissent dans les périodes de grand froid.

2.1 Limites de la similarité

L'approche par similarité textuelle a connu un succès auprès des experts qui ont apprécié son coté intuitif et le potentiel de rapprochement de rapports dont le lien n'est nullement reflété par un codage commun. La logique d'utilisation simplifiée et l'intuitivité de l'interface, conçue d'emblée comme un support à une exploration de la collection sans a priori, ont aussi contribué à la validation de cette approche par ses usagers.

Néanmoins, nous nous sommes vite aperçus que le calcul de similarité, dans notre contexte précis et compte tenu des spécificités du matériau textuel auquel nous avons affaire, souffrait d'un manque de finesse évident. Étant donné que l'ensemble des textes sont issus d'un domaine très circonscrit, celui de l'aviation civile, ils sont tous plus ou moins similaires, la plupart parlant d'"avions", de "pilotes" et de "vols". Ce fond lexical commun est en partie géré par les techniques de pondération, comme le TF/IDF (Spärck-Jones, 2004), mais compte tenu de la variation lexicale inhérente du domaine, notamment la multitude de termes [1] désignant un même objet qui sont employés par les rédacteurs, des rapprochements sont faits sans pour autant designer des facteurs de similarité pertinents pour une analyse.

Une autre limite directement liée à la visée applicative de nos travaux apparaît aussi. On s'aperçoit de l'existence d'un fort recouvrement entre le codage des rapports et les regroupements mises à l'évidence par le calcul de similarité textuelle. Un lien entre le termes du texte et leur codage existe dans le corpus. Il est clair que dans les rapports parlant de chocs aviaires [2], on trouve des termes comme "oiseau", "mouette", et "aviaire". Un rapprochement basé sur ces termes retrouvera plus ou moins la catégorie *choc aviaire* qui est déjà mise en évidence dans le codage. Dit autrement, la similarité textuelle à tendance de retrouver les dimensions les plus saillantes dans le corpus, dimensions qui sont pour la plupart déjà bien identifiées et reflétées dans les schémas de codification.

Par ailleurs, un système de classification automatique de ces données, permet déjà, sur la base du contenu textuel, de proposer des codes aux experts (Hermann *et al.*, 2008). Or, un de nos objectifs est notamment de chercher des facteurs communs plus subtils, pouvant rapprocher des incidents sur des critères différents et transversaux au codage.

3 Dimensions de la similarité textuelle

La similarité textuelle, telle qu'elle est calculée, représente toute parenté qui peut exister entre deux textes sur une dimension unique, sans tenir compte des multiples facteurs qui peuvent contribuer à cette parenté.

1. Pour parler du pilote, on trouve dans les textes le terme "pilote", mais aussi un ensemble de termes et d'acronymes spécifiques au domaine comme "cdb" (commandant de bord), "opl" (officier pilote de ligne), "copilote", "copi", "pf" (pilot flying), "pnf" (pilote not flying) etc.
2. Il arrive très fréquemment que les avions percutent des oiseaux.

Une des thématiques actuelles en recherche d'information est de regrouper les résultats des moteurs de recherche par thème en utilisant des méthodes de *clustering*, afin de mettre en évidence les différentes thématiques qui sont présentes dans la liste des résultats. Une requête comme "japon" par exemple, peut ramener des documents traitant du tourisme au japon et de gastronomie japonaise (Navarro *et al.*, 2011). En se basant sur les similitudes entre ces documents un système d'apprentissage non supervisé regroupe ensuite ces résultats en deux paquets et permet à l'utilisateur de focaliser sa recherche sur le sous-ensemble qui l'intéresse. Ces méthodes, tout en raffinant et classant les résultats ne peuvent pas encore gérer des collections où les thématiques varient simultanément sur plusieurs dimensions. Les résultats sur le japon peuvent concerner des localisations différentes ("Tokyo" et "Osaka", par exemple) sans qu'une localisation soit particulièrement associée à un des thèmes. Un système de clustering peinera à isoler ces deux dimensions de variation et à proposer un découpage des résultats selon les deux critères (thème et localisation) simultanément. De travaux sont en cours visant à développer des méthodes efficaces de clustering avec recouvrement [3], notamment pour répondre à l'unidimensionalité des techniques actuelles.

Principalement orientées vers un usage dans un moteur de recherche "classique" et sur des collections larges de textes hétérogènes, ces méthodes n'assument pas une organisation à priori de la collection. Or dans un corpus spécialisé, comme les bases de rapports d'incidents, les schémas de codification visent justement à organiser la collection, de façon pertinente compte tenu de spécificités de son contexte d'utilisation, tout en intégrant l'hétérogénéité des facteurs de similarité et en représentant. Illustrons ceci par les trois exemples suivants que nous avons construit en nous inspirant de textes réels intitulés comme suit :

```
1) Choc aviaire au décollage.
2) Turbulences au décollage.
3) Choc aviaire à l'atterrissage.
```

Entre ces trois textes un score de similarité comparable sera calculé entre 1) et 2) et entre 1) et 3). Pourtant, les raisons de ce rapprochement sont différentes dans les deux cas. 1) et 3) traitent d'un même type d'incident, alors que 1) et 2) partagent les circonstances dans lesquels sont survenus des incidents différents. Ces deux aspects sont pris en compte dans le schéma de codification, grâce aux champs "type d'incident" et "phase de vol". Le "type d'incident" pour 1) et 3) sera *choc aviaire* et *turbulences* pour 2). La "phase de vol" sera *décollage* pour 1) et 2) et *atterrissage* pour 3). Nous allons donc regarder de près comment mettre en évidence le lien entre le codage des rapports et leur contenu textuel.

3.1 Lien entre codage et contenu

Nous avons déjà vu que certains termes des textes étaient fortement liés à certaines classes du schéma de codification et que ces mêmes termes font en sorte que la similarité textuelle retrouve souvent les classes du schéma.

Afin d'étudier ce lien, nous avons constitué un corpus de test en prenant des rapports traitant de chocs aviaires et de turbulences, survenus lors de l'atterrissage et lors du décollage, de manière à avoir une collection équilibrée que nous savons varier sur deux dimensions, la phase de vol et le type d'incident. Le corpus est constitué de 482 rapports que nous avons choisis en nous basant sur le codage de leur champs *type d'incident* et *phase de vol* :

3. Concrètement,une telle méthode doit être capable de classer un même document dans plusieurs classes, en fonction de critères de rapprochement différents.

	Turbulences	Choc aviaire	Total
Atterrissage	118	133	251
Décollage	107	124	231
Total	225	257	482

Le premier test a consisté à mesurer le degré de recouvrement entre similarité et catégorisation dans le corpus. Pour cela nous avons, pour chaque document, automatiquement sélectionné les 30 documents les plus similaires et, pour chacun de ces documents, testé s'il partage la même valeur pour les champs *type d'incident* et *phase de vol*. En moyenne, 89% des documents partagent la catégorie et 75% des documents partagent la phase de vol, alors que si aucun lien entre codage et similarité n'existait, nous nous attendrions à ce que ces valeurs avoisinent les 50%.

3.2 Effacement des dimensions principales

Afin d'isoler les dimensions de similarité, nous avons tout d'abord cherché les termes qui sont les plus liés à chacune d'entre elles en utilisant une mesure d'interdépendance statistique : l'information mutuelle (IM), en nous inspirant des techniques de sélection de traits utilisées en recherche d'information (voir [4] (Manning *et al.*, 2008, Section 13.5.1) pour l'algorithme utilisé). En RI cette technique permet, pour une collection catégorisée, de réduire l'espace des termes en ne sélectionnant que ceux qui sont statistiquement corrélés à une classe donnée. L'IM est aussi communément utilisée en classification automatique. Étant donné un terme t et une classe C, plus l'information mutuelle $IM(t,C)$ est élevée, plus t permet de correctement prédire C.

Pour la RI, l'hypothèse sous-jacente qui justifie ce procédé est que ce sont typiquement les termes décrivant au mieux la variation relative à une organisation particulière repérée par un humain *via* une classification donnée qui seront aussi les plus performants pour l'indexation de la même collection. Notre objectif est exactement inverse. Nous allons exclure ces termes du calcul de similarité, afin qu'il ne reflète pas l'organisation déjà présente dans le schéma de codification.

Nous avons calculé l'IM entre tous les termes d'un corpus de 4450 rapports, et les 4 classes que nous avons isolées. Voici les 5 termes les plus corrélés par catégorie.

	Turbulences	**Choc aviaire**	**Atterrissage**	**Décollage**
1	vent	aviaire	approche	décollage
2	turbulence	collision	finale	poussée
3	gaz	oiseau	atterrissage	rotation
4	arrière	impact	stabilisation	t/o [5]
5	windshear [6] [7]	bird [7]	arrondir	vr [8]

Nous avons de nouveau mesuré la moyenne de recouvrement (MR) entre similarité et codification, mais cette fois ci en excluant soit les 50 termes les plus associés aux deux phases de vol (phVol), soit les 50 termes les plus associés aux types d'évènement (typEve). Nous avons aussi calculé un taux de perturbation (TP) en regardant, pour chaque document le nombre de documents qui apparaissent dans les 30 documents les plus similaires lors de l'application d'un filtrage.

4. version disponible en ligne à `http://nlp.stanford.edu/IR-book/`
5. Take-off (Décollage)
6. Cisaillement du vent
7. Il est très courant que des termes anglais soient employés dans ces textes pourtant écrits en français.
8. Vitesse de rotation

	MR phVol	MR typEve	TP [9]
Sans filtre	75%	89%	-
Filtre sur phVol	64%	84%	9,8
Filtre sut typEve	73%	69%	13,6

On peut voir que le recouvrement entre la similarité textuelle et une dimension donnée varie en fonction du filtrage des termes associés à cette même dimension, alors que le recouvrement sur l'autre dimension est moins affecté. Après filtrage on trouve, en moyenne, respectivement 9,8 et 13,6 nouveaux documents dans la liste des 30 premiers, ce qui témoigne de l'effet du filtrage sur le classement des résultats.

Concrètement ceci signifie que, pour un rapport traitant de turbulences au décollage, un filtrage des termes associés avec les phases de vol privilégiera les rapports traitant de turbulences alors qu'un filtrage des termes associés avec le type d'évènement privilégiera les rapports traitant d'évènements survenus lors du décollage.

3.3 Dimensions transversales

En effaçant ces dimensions de similarité, le filtrage des termes associés possède la capacité de mettre en évidence des facteurs de similarité secondaires. Voici un exemple d'une telle dimension qui a émergé de notre corpus. Le rapport suivant traite de turbulences à l'atterrissage, mais mentionne en plus un double pilotage [10] :

> INCURSION VFE SUITE CISAILLEMENT EN FINALE. [REPORT]. Fort cisaillement en finale reporté par les avions précédents. La soudaineté du phénomène surprend l'OPL PF. Légère incursion dans la VFE (3 ou 4 kts). Réponse des commandes par CDB (double pilotage pendant 1 à 2 s.). Avion stabilisé, l'OPL reprend les commandes. Atterrissage sans problème. -FIN-

Lorsque nous regardons la liste des rapports similaires sans filtre, au premiers rangs nous trouvons ceux qui évoquent des turbulences à l'atterrissage, comme par exemple :

> FORT CISAILLEMENT DE VENT EN FINALE 26R CDG. [REPORT]. FORT CISAILLEMENT DE VENT EN FINALE. -FIN-

Pour le même document, lorsque l'on filtre les termes associés avec la phase de vol et le type d'évènement, les rapports parlant uniquement de turbulences à l'atterrissage apparaissent plus bas dans la liste des rapports similaires et on retrouvera leur place ceux qui partagent des termes non associés avec les phases et les types d'évènement, notamment des rapports traitant de double pilotage, information qui n'est pas reflétée dans le codage. Ce facteur commun permet d'établir un lien entre ces deux rapports, qui dans certains cas peut s'avérer pertinent pour un expert.

> BREF DOUBLE PILOTAGE AU DECOLLAGE. [REPORT]. OPL PF au décollage. Vent travers avec rafales. Brève action réflexe en latéral du CDB pour contrer rafale et début d'inclinaison à droite. Prise de priorité peu pertinente pour effet immédiat. -FIN-

9. Les valeurs ici sont des moyennes pour les 482 documents.
10. Double pilotage signifie que les deux pilotes agissent simultanément sur les commandes de l'avion.

4 Conclusion et perspectives

La technique que nous avons présentée s'inscrit dans un effort global dont l'objectif est de proposer des outils d'exploration de collections de documents et la mise en évidence de liens de similarité "faibles" entre les documents qui seraient autrement masqués par la dimension de similarité la plus saillante. D'emblée conçues pour un usage par un utilisateur averti, notre intention est d'en évaluer l'apport applicatif en les proposant à des experts en sûreté aérienne sous forme d'un outil de visualisation et d'exploration permettant dynamiquement à l'utilisateur de choisir les dimensions à ne pas prendre en compte lors du calcul à partir d'une liste des dimensions les plus saillantes pour le sous-ensemble en cours d'analyse. Le fait de se baser sur la codification assure que les choix de filtres qu'auront les experts reflètent des concepts qu'ils ont l'habitude de manipuler dans leur activité d'analyse.

Disposant à l'heure actuelle d'une preuve de concept, nous comptons, dans les mois qui viennent, passer à l'échelle en prenant en compte l'intégralité de la codification des collections. S'agissant de techniques exploratoires et fortement dépendantes du domaine et de leur objectif applicatif précis, nous ne sommes pas en mesure de proposer un protocole d'évaluation classique et comptons sur un évaluation par l'usage et un échange constant avec les usagers pour pouvoir juger de la pertinence de ces méthodes.

Références

Chandola, V., Banerjee, A. et Kumar, V. (2009). Anomaly detection : A survey. *ACM Computing Surveys (CSUR)*, 41(3):15.

Hermann, E., Leblois, S., Mazeau, M., Bourigault, D., Fabre, C., Travadel, S., Durgeat, P. et Nouvel, D. (2008). Outils de Traitement Automatique des Langues appliqués aux comptes rendus d'incidents et d'accidents. *In 16e Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement, Avignon*.

Manning, C. D., Raghavan, P. et Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press, New York.

Navarro, E., Chudy, Y., Gaume, B., Cabanac, G. et Pinel-Sauvagnat, K. (2011). Kodex ou comment organiser les résultats d'une recherche d'information par détection de communautés sur un graphe biparti ? *In Actes de Coria 2011 : Conférence en Recherche d'Information et Applications*.

Ponvert, M. (2009). Définition des besoins nécessaires à la mise en place d'un Data WareHouse dans le cadre du SGS Air France. Mémoire de D.E.A., École Nationale de l'Aviation Civile.

Salton, G., Wong, A. et Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Spärck-Jones, K. (2004). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 60(5):493–502.

Steinbach, M., Karypis, G., Kumar, V. *et al.* (2000). A comparison of document clustering techniques. *In KDD workshop on text mining*, volume 400, pages 525–526. Boston.

Tulechki, N. (2011). Des outils de TAL en support aux experts de sûreté industrielle pour l'exploitation de bases de données de retour d'expérience. *In Actes des 13èmes Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2011)*.

Traduction automatique à partir de corpus comparables: extraction de phrases parallèles à partir de données comparables multimodales

Haithem AFLI Loïc BARRAULT Holger SCHWENK
Laboratoire d'Informatique de l'Université du Maine
prénom.nom@lium.univ-lemans.fr

RÉSUMÉ

Les performances des systèmes de traduction automatique statistique dépendent de la disponibilité de textes parallèles bilingues, appelés aussi *bitextes*. Cependant, les corpus parallèles sont des ressources limitées et parfois indisponibles pour certains couples de langues ou domaines. Nous présentons une technique pour l'extraction de phrases parallèles à partir d'un corpus comparable multimodal (audio et texte). Ces enregistrements sont transcrits avec un système de reconnaissance automatique de la parole et traduits avec un système de traduction automatique. Ces traductions sont ensuite utilisées comme requêtes d'un système de recherche d'information pour sélectionner des phrases parallèles sans erreur et générer un bitexte. Plusieurs expériences ont été menées sur les données de la campagne IWSLT'11 (TED) qui montrent la faisabilité de notre approche.

ABSTRACT

Automatic Translation from Comparable corpora : extracting parallel sentences from multimodal comparable corpora

Statistical Machine Translation (SMT) systems depend on the availability of bilingual parallel text, also called bitext. However parallel corpora are a limited resource and are often not available for some domains or language pairs. We present an alternative method for extracting parallel sentences from multimodal comparable corpora. This work extends the use of comparable corpora, in using audio instead of text on the source side. The audio is transcribed by an automatic speech recognition system and translated with a base-line SMT system. We then use information retrieval in a large text corpus of the target language to extract parallel sentences. We have performed a series of experiments on data of the IWSLT'11 speech translation task (TED) that shows the feasibility of our approach.

MOTS-CLÉS : Reconnaissance de la parole, traduction automatique statistique, corpus comparables multimodaux, extraction de phrases parallèles.

KEYWORDS: Automatic speech recognition, statistical machine translation, multimodal comparable corpora, extraction of parallel sentences.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 447–454,
Grenoble, 4 au 8 juin 2012.

1 Introduction

La construction d'un système de traduction automatique statistique (TAS) nécessite un corpus dit parallèle pour l'apprentissage du modèle de traduction et des données monolingues pour construire le modèle de langue cible. Un corpus parallèle est une collection de textes bilingues alignés au niveau de la phrase, c'est-à-dire des textes en langue source avec leurs traductions.

Malheureusement, les textes parallèles librement disponibles sont aussi des ressources rares : la taille est souvent limitée, la couverture linguistique insuffisante ou le domaine n'est pas approprié. Il y a relativement peu de paires de langues pour lesquelles des corpus parallèles de taille raisonnable sont disponibles comme l'anglais, le français, l'espagnol, l'arabe, le chinois et quelques langues européennes (Hewavitharana et Vogel, 2011). De plus, ces corpus proviennent principalement de sources gouvernementales, comme le parlement canadien ou européen, ou de l'Organisation des Nations Unies. Ceci est problématique en TAS, parce que les systèmes de traduction appris sur des données provenant, par exemple, d'un domaine politique ne donnent pas de bons résultats lorsqu'ils sont utilisés pour traduire des articles scientifiques.

Une façon de pallier ce manque de données parallèles est d'exploiter les corpus comparables qui sont plus abondants. Un corpus comparable est un ensemble de textes dans deux langues différentes, qui ne sont pas parallèles au sens strict du terme, mais qui contiennent les mêmes informations. On peut par exemple citer les actualités multilingues produites par des organismes de presse tels que l'Agence France Presse (AFP), Xinhua, l'agence Reuters, CNN, BBC, etc. Ces textes sont largement disponibles sur le Web pour de nombreuses paires de langues (Resnik et Smith, 2003). Le degré de parallélisme peut varier considérablement, en allant de documents peu parallèles, aux documents quasi parallèles ou « parallèles bruités » qui contiennent de nombreuses phrases parallèles (Fung et Cheung, 2004). Ces corpus comparables peuvent couvrir différents sujets.

Ces travaux s'inscrivent dans le cadre du projet DEPART (Documents Écrits et PAroles – Reconnaissance et Traduction) dont l'un des objectifs est l'exploitation de données multimodales et multilingues pour la TAS. Nous considérons le cas, assez fréquent pour des domaines de spécialité, où un manque de données textuelles peut être pallié par l'exploitation de données audio. Un domaine de spécialité est un sous-domaine possédant un vocabulaire spécifique, tel que la chirurgie dans le domaine plus large de la médecine. Nous pouvons également considérer les conférences ou séminaires scientifiques et leurs articles associés pour un domaine de recherche spécifique.

La question que nous nous posons alors est la suivante : un corpus comparable multimodal permet-il d'apporter des solutions au problème du manque de données parallèles ? Dans ce travail nous proposons une méthode pour l'utilisation de corpus comparables multimodaux, en se limitant aux modalités texte et audio, pour l'extraction de données parallèles.

2 Recherches précédentes

Plusieurs travaux ont traité de l'extraction des données parallèles à partir d'un corpus comparable bilingue. Un critère de maximum de vraisemblance est proposé par Zhao et Vogel (2002) qui ont combiné des modèles de longueur de phrases avec un lexique extrait d'un corpus parallèle

aligné existant. Le lexique est itérativement adapté avec un processus de réapprentissage en utilisant les données extraites. Resnik et Smith (2003) ont montré qu'ils peuvent générer un grand nombre de documents parallèles à partir du WEB en utilisant leur système d'extraction de textes parallèles, « STRAND ». Do *et al.* (2010) ont utilisé une méthode non-supervisée pour extraire des paires de phrases parallèles à partir d'un corpus comparable et ont montré que cette approche est intéressante surtout pour les langues peu dotées. La détection des paires de phrases parallèles est faite en utilisant un système de traduction automatique de base qui est amélioré avec un processus itératif.

Afin de construire un corpus parallèle anglais/japonais, Utiyama et Isahara (2003) utilisent la recherche d'information cross-langue et la programmation dynamique pour l'extraction de phrases parallèles à partir d'un corpus comparable dans le domaine des actualités. Les paires d'articles similaires sont identifiées et traitées comme des textes parallèles afin d'aligner leurs phrases. La procédure d'alignement commence par la traduction mot à mot des textes japonais en utilisant un dictionnaire bilingue, qui sont ensuite pris comme requêtes de recherche d'information dans la partie anglaise des textes. L'approche de Fung et Cheung (2004) utilise la mesure « cosinus » pour calculer le degré de similarité des phrases. Toutes les paires de phrases possibles d'un corpus « non-parallèle »ont été considérées, et celles ayant un niveau de similarité supérieur à un certain seuil sont conservées pour construire un dictionnaire qui sera réappris itérativement.

Une méthode d'extraction des segments de phrases parallèles est présentée par Munteanu et Marcu (2005). Un dictionnaire bilingue existant est utilisé pour traduire chaque document en langue source vers la langue cible afin d'extraire le document cible qui correspond à cette traduction. Pour chaque paire de documents, des paires de phrases et de segments parallèles sont extraites en utilisant un lexique de traduction et un classifieur à maximum d'entropie pour le choix final des phrases parallèles. Rauf et Schwenk (2011) présentent une technique similaire à celle de Munteanu et Marcu (2005). Les différences majeures résident dans l'utilisation d'un système de TA statistique à la place du dictionnaire bilingue, et dans l'utilisation de mesures d'évaluation, comme le taux d'erreur mot (WER) ou le taux d'édition de la traduction (TER), pour évaluer le degré de parallélisme des phrases extraites.

Toutes ces méthodes sont présentées comme des techniques efficaces pour extraire des données parallèles à partir d'un corpus comparable. Certains travaux exploitent la modalité audio pour l'extraction de données parallèles. Paulik et Waibel (2009) ont montré que les modèles de traductions statistiques peuvent être appris automatiquement d'une manière non-supervisée à partir des données parallèles audio. Dans notre contexte de travail, nous nous intéressons à l'exploitation des corpus comparables multimodaux avec différents niveaux de similitude. La multimodalité concernera l'utilisation de documents textuels et audio.

3 Architecture générale

Notre but est d'exploiter les données comparables multimodales afin d'en extraire des données parallèles nécessaires pour construire, adapter et améliorer nos systèmes de traduction automatique statistique. L'architecture générale de notre approche, qui se résume en 3 étapes, est décrite dans la figure 1.
Notre corpus comparable multimodal est constitué de données audio en langue source L1 et de données textuelles en langue cible L2. Les données audio sont tout d'abord transcrites par un système de Reconnaissance Automatique de la Parole (RAP). Ce système produit une hypothèse

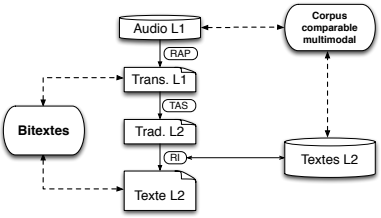


FIGURE 1 – Architecture générale du système

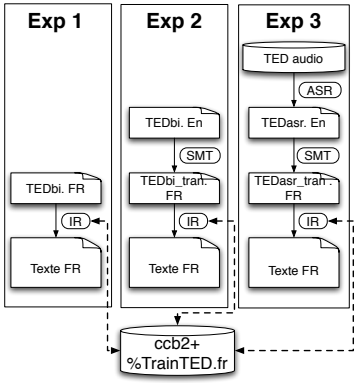


FIGURE 2 – Expériences permettant de mesurer l'impact des différents modules mis en jeu sur le corpus bilingue extrait.

de transcription qui est ensuite traduite par le système TAS. La meilleure hypothèse de traduction est utilisée comme requête dans le système de recherche d'information (RI), dont le corpus indexé correspond à la partie textuelle en langue cible du corpus comparable multimodal. Dans cette approche, qui se base sur les travaux de Rauf et Schwenk (2011), nous utilisons le logiciel libre Lemur (Ogilvie et Callan, 2001) pour effectuer la RI. Au final, nous obtenons un bitexte constitué d'une part de la transcription automatique et d'autre part du résultat de la RI, qui pourra être réinjecté dans le système de base.

Ce cadre de travail soulève toutefois plusieurs problèmes. Chaque module mis en jeu pour la traduction de la parole introduit un certain nombre d'erreurs. Il est donc important de mettre en évidence la faisabilité de l'approche ainsi que l'impact de chaque module sur les données générées. Pour cela, nous avons effectué 3 types d'expérience différents, décrits dans la figure 2. Le premier type d'expérience (*Exp 1*) consiste à utiliser la référence de traduction comme requête pour la RI. Ce cas est le plus favorable, cela simule le fait que les modules de RAP et de TAS ne commettent aucune erreur. Le second type d'expérience (*Exp 2*) utilise la référence de transcription pour alimenter le système de traduction automatique. Cela permet de mettre en évidence l'impact des erreurs de traduction. Enfin, le troisième type d'expérience (*Exp 3*) met en œuvre l'architecture complète décrite ci-dessus. Cela correspond au cas réel auquel nous sommes confrontés.

Une autre problématique concerne l'importance du degré de similitude (*comparabilité*) des corpus comparables utilisés. Nous avons donc artificiellement créé des corpus comparables plus ou moins ressemblants en intégrant une quantité plus ou moins grande (*25%, 50%, 75% et 100%*) de données du domaine dans le corpus indexé par la RI.

Les résultats de la RI ne sont pas toujours satisfaisants, il est donc nécessaire de filtrer ces résultats afin de ne pas ajouter de phrases non parallèles dans le bitexte final. Nous considérons le Taux d'Edition de la Traduction (*Translation Edit Rate* - TER) calculé entre les phrases retournées par la RI et la requête, comme mesure de filtrage des phrases trouvées. Les phrases ayant un TER

bitextes	# de mots	du domaine ?
nc7	3,7M	non
eparl7	56,4M	non
ccb2_px70	1,3M	non
TEDasr	1,8M	oui
TEDbi	1,9M	oui

TABLE 1 – Données utilisées pour l'apprentissage des systèmes de traduction automatique.

Dev	# de mots
dev.outASR	36k
dev.refSMT	38k

Test	# de mots
tst.outASR	8,7k
tst.refSMT	9,1 k

TABLE 2 – Données de développement (Dev) et de Test.

supérieur à un certain seuil (déterminé empiriquement) sont exclues.

Dans tous les cas, l'évaluation de l'approche est nécessaire. Ainsi, les données parallèles extraites sont réinjectées dans le système de base, qui est ensuite utilisé pour traduire les données de test à nouveau. L'évaluation peut ensuite se faire avec une mesure automatique comme BLEU (Papineni *et al.*, 2002).

4 Expériences et résultats

Pour nos expériences, nous exploitons les données de la campagne d'évaluation IWSLT'11 dans laquelle des données bilingues multimodales sont disponibles. Cette tâche, détaillée dans Rousseau *et al.* (2011), consiste à traduire des discours de TED [1] de l'anglais vers le français. Le système de RAP est appris sur 773 discours représentant 118 heures de parole. Les données de développement et de test officielles sont utilisées pour évaluer notre approche.

Le corpus de développement est composé de 19 discours représentant un peu plus de 4 heures de parole. Les corpus bilingues suivants sont utilisés pour l'apprentissage des modèles de traduction : News-Commentary version 7(*nc7*), le corpus des actes du parlement européen (*eparl7*) et le corpus Gigaword_EnFr (ccb2_px70). [2] De ce dernier, ne sont conservées que les paires de phrases dont la perplexité du côté cible (calculée avec le modèle de langue utilisé pour le système TAS) est inférieure à un seuil (ici 70). Le détail des données disponibles est présenté dans le tableau 1.

Le système de reconnaissance de la parole utilisé est basé sur le système libre CMU Sphinx (*version 3 et 4*), modifié et amélioré. Le système anglais qui été développé pour transcrire les données audio de TED utilise cinq passes similaires à celui du français décrit dans Deléglise *et al.* (2009). Les systèmes de traduction mis en œuvre sont fondés sur Moses (Koehn *et al.*, 2007), approche par segments (*phrase-based*). Le modèle de langue est un modèle 4-gramme construit avec l'outil SRILM (Stolke, 2002). Nous avons utilisé toutes les données monolingues disponibles et le côté cible des bitextes.

Comme mentionné précédemment, le score *TER* est utilisé comme métrique de filtrage des phrases résultantes de la RI, c'est-à-dire que les phrases ayant un TER supérieur à un certain seuil ne sont pas conservées. Ce seuil est déterminé expérimentalement. Pour cela, nous avons filtré les corpus extraits dans les différentes conditions d'expérimentation avec différents seuils *TER* (de 0 à 100). Pour chaque seuil TER nous obtenons un nombre de phrases parallèles. Le corpus obtenu est ajouté aux données d'entraînement du système de base (*eparl7* et *nc7*) pour

1. http ://www.ted.com/
2. Ces corpus sont librement disponibles sur le site de la campagne IWSLT'11 et WMT'11.

obtenir le système adapté. Les résultats en terme de score *BLEU* sur le corpus de développement obtenus avec les différents systèmes adaptés sont présentés dans la figure 3.

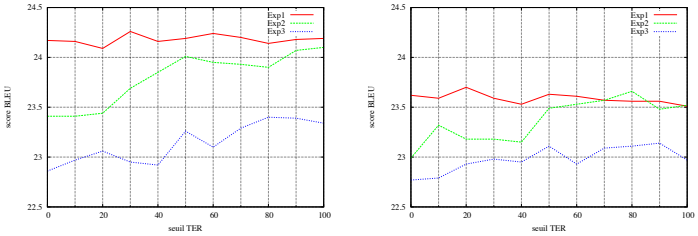


FIGURE 3 – Score BLEU de la traduction du Dev en utilisant les systèmes adaptés avec les bitextes correspondant à différents seuils TER, extraits d'un corpus d'index constitué par *ccb2 + 100% TEDbi* (à gauche) et *ccb2+ 25% TEDbi* (à droite).

Ces résultats montrent que le choix du seuil de *TER* adéquat dépend de la nature des données. En effet, pour la condition de l'*Exp1* où les requêtes de la RI sont sans erreur, nous remarquons que le meilleur résultat est obtenu pour un seuil proche de 0. Dans les deux autres conditions (*Exp2* et *Exp3*), le meilleur seuil est dans l'intervalle [80-90]. Dans nos expériences, nous retiendrons le seuil de 80 pour le filtrage des résultats de la RI.

Dans l'expérience *Exp2*, les traductions automatiques sont utilisées en tant que requêtes pour la RI. On peut espérer que la RI elle-même n'est pas trop affectée par les erreurs de traduction, mais ceci influence bien sûr le filtrage basé sur le score TER. Nous n'avons pas observé un maximum du score BLEU en fonction du seuil sur le score TER - dans nos expériences les performances semblent augmenter de façon continue. Néanmoins, afin de limiter l'impact des phrases bruitées, nous avons choisi un seuil de 70. On peut observer que le score BLEU du système adapté est très proche de celui de *Exp1*. Ainsi, nous pouvons conclure que les erreurs commises par la TAS n'ont pas une influence importante sur l'algorithme d'extraction des phrases parallèles. Ceci confirme l'analyse de (Rauf et Schwenk, 2011).

Notre système de base entraîné avec des données génériques obtient un score BLEU de 22,93. Dans *Exp1*, nous utilisons les traductions de référence en tant que requêtes et la RI devrait en principe trouver toutes les phrases avec un TER de zéro. Les figures montrent que la RI fonctionne comme attendu : l'amélioration du score BLEU ne dépend pas du seuil sur le score TER puisque la plupart des phrases ont effectivement un score TER de zéro. L'amélioration du score BLEU dépend bien sûr de la quantité de données extraites : le score BLEU augmente de 22,93 à 24,14 lorsque 100% des données ont été injectées, alors que nous n'obtenons que 23,62 avec 20% des données TED. Ces résultats nous donnent une borne supérieure des résultats envisageables avec l'utilisation d'un corpus multimodal.

Finalement, dans *Exp3*, la RAP est utilisée dont le taux d'erreur est d'environ 18%. Les phrases extraites du corpus multimodal permettent d'améliorer le système de traduction : le score BLEU n'est que 0,5 points en dessous de celui obtenu dans *Exp1* ou *Exp2*. Les résultats obtenus après adaptation du système de base sont présentés dans le tableau 4. Dans ce cas, le corpus indexé par la RI est constitué des corpus *ccb2_px70* et *TEDbi* (100%).

	Phrase extraite
Français	vous allez chez ibm et vous prenez un superordinateur ...
Anglais	you get a supercomputer because they know ...
	Test audio
Sortie ASR	a supercomputer has calculated that humans and only ...
Référence	a supercomputer has calculated that humans have only ...
	Traductions de la sortie ASR
Système de base	un supercomputer a calculé que les humains et seulement ...
Système adapté	un superordinateur a calculé que les humains et seulement ...
Référence	un superordinateur a calculé que les humains n' avaient plus que ...
	Traductions améliorées
Sys de base	j'ai écrit un article sur la nourriture génétiquement modifiée
Sys adapté	j'ai écrit un article sur les produits alimentaires génétiquement modifiés
Sys de base	yeah tu as raison de réparer
Sys adapté	euh oui tu as raison il faut réparer

TABLE 3 – Exemples d'amélioration du système de base : vocabulaire enrichi à partir des phrases parallèles extraites dans la condition *Exp3*.

Le tableau 5 présente les résultats des systèmes adaptés en fonction du degré de similitude du corpus comparable, dans les conditions d'expérimentation *Exp3*. Des exemples d'adaptation sont présentés dans le tableau 3. Nous pouvons remarquer que le degré de similitude est un facteur important. Un résultat attendu est que lorsque nous augmentons la proportion de corpus du

Expérience	Dev	Test
Système de base	22,93	23,96
Exp1	24,14	25,14
Exp2	23,90	25,15
Exp3	23,40	24,69

TABLE 4 – % BLEU obtenus sur le Dev et Test après l'ajout des bitextes extraits au système de base, dans les conditions *Exp1, Exp2 et Exp3*.

Expérience	Dev	Test	# mots
Système de base	22,93	23,96	-
25% TEDbi	23,11	24,40	∼110k
50% TEDbi	23,27	24,58	∼215k
75% TEDbi	23,43	24,42	∼293k
100% TEDbi	23,40	24,69	∼393k

TABLE 5 – Résultats (%BLEU) obtenus avec les systèmes adaptés lorsque le degré de similitude du corpus comparable varie.

domaine dans le corpus indexé, les performances sont meilleures. Il est important de noter que lorsque les corpus sont moins similaires, le nombre de phrases conservé est réduit drastiquement par le filtrage, et donc l'impact de l'adaptation est plus faible. Sans filtrage, les performances du système de base peuvent être dégradées.

5 Conclusion

Dans ce travail nous avons proposé une méthode permettant d'extraire des textes parallèles à partir de corpus comparables multimodaux (audio et texte) pour adapter et améliorer des systèmes de traduction automatique statistique. Plusieurs modules sont utilisés pour extraire du

texte parallèle : reconnaissance automatique de la parole, traduction automatique et recherche d'information. Nous validons notre méthode en injectant les données produites dans l'apprentissage de nouveaux systèmes de TAS. Des améliorations en termes de BLEU sont obtenues pour différents cadres expérimentaux. Il en ressort que l'enchaînement des modules ne dégrade que faiblement les résultats, mais le filtrage des résultats de la RI est nécessaire. Le degré de similitude du corpus comparable est un facteur important qu'il faudra prendre en compte lorsque cette architecture sera exploitée dans des conditions réelles.

Remerciements

Ces recherches ont été financées par la région des Pays de la Loire sous le projet DEPART[3].

Références

Deléglise, P., Estève, Y., Meignier, S. et Merlin, T. (2009). Improvements to the LIUM french ASR system based on CMU Sphinx : what helps to significantly reduce the word error rate ? *In Interspeech 2009*.

Do, T. N. D., Besacier, L. et Castelli, E. (2010). Apprentissage non supervisé pour la traduction automatique : application à un couple de langues peu doté. *TALN 2010*.

Fung, P. et Cheung, P. (2004). Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. *In Proceedings of COLING '04*.

Hewavitharana, S. et Vogel, S. (2011). Extracting parallel phrases from comparable data. *In Proceedings of the 4th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web*, BUCC '11, pages 61–68.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. et Herbst, E. (2007). Moses : open source toolkit for statistical machine translation. *In Proceedings of ACL'07*, pages 177–180.

Munteanu, D. S. et Marcu, D. (2005). Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504.

Ogilvie, P. et Callan, J. (2001). Experiments using the lemur toolkit. *Procedding of the Trenth Text Retrieval Conference (TREC-10)*.

Papineni, K., Roukos, S., Ward, T. et Zhu, W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. *In Proceedings of ACL '02*, pages 311–318.

Paulik, M. et Waibel, A. (2009). Automatic translation from parallel speech : Simultaneous interpretation as mt training data. *ASRU*.

Rauf, S. A. et Schwenk, H. (2011). Parallel sentence generation from comparable corpora for improved SMT. 25(4):341–375.

Resnik, P. et Smith, N. A. (2003). The web as a parallel corpus. *Comput. Linguist.*, 29:349–380.

Rousseau, A., Bougares, F., Deléglise, P., Schwenk, H. et Estève, Y. (2011). LIUM's systems for the IWSLT 2011 speech translation tasks. *In Proceedings of IWSLT'11*.

Stolke, A. (2002). Srilm - an extensible language modeling toolkit. *ICSLP*, pages 901–904.

Utiyama, M. et Isahara, H. (2003). Reliable measures for aligning japanese-english news articles and sentences. *In Proceedings of ACL'03*, volume 1, pages 72–79.

Zhao, B. et Vogel, S. (2002). Adaptive parallel sentences mining from web bilingual news collection. *In Proceedings of IEEE International Conference on Data Mining*, page 745.

3. http://www.projet-depart.org/

La reconnaissance automatique de la fonction des pronoms démonstratifs en langue arabe

Yacine Ben Yahia, Souha Mezghani Hammami, Lamia Hadrich Belguith
ANLP Research Group – Laboratoire MIRACL/ FSEGS Sfax, Tunisie.
anlp-reserach-group@googlegroups.com
benyacine.sint@gmail.com, souha.mezghani@fsegs.rnu.tn, l.belguith@fsegs.rnu.tn

RESUME

La résolution d'anaphores est l'une des tâches les plus difficiles du Traitement Automatique du Langage Naturel (TALN). La capacité de classifier les pronoms avant de tenter une tâche de résolution d'anaphores serait importante, puisque pour traiter un pronom cataphorique le système doit chercher l'antécédent dans le segment qui suit le pronom. Alors que, pour le pronom anaphorique, le système doit chercher l'antécédent dans le segment qui précède le pronom. En outre, le nombre des pronoms a été jugée non-trivial dans la langue arabe. C'est dans ce cadre que se situe notre travail qui consiste à proposer une méthode pour la classification automatique des pronoms démonstratifs arabes, basée sur l'apprentissage. Nous avons évalué notre approche sur un corpus composé de 365585 mots contenant 14318 pronoms démonstratifs et nous avons obtenu des résultats encourageants : 99.3% comme F-Mesure.

ABSTRACT

Automatic recognition of demonstrative pronouns function in Arabic

Anaphora resolution is one of the most difficult tasks in NLP. Classifying pronouns before attempting a task of anaphora resolution is important because to handle the cataphoric pronoun, the system should determine the antecedent into the segment following the pronoun. Although, for the anaphoric pronoun, the system should look for the antecedent into the segment before the pronoun. In addition, the number of demonstrative pronouns is very important in Arabic. In this paper, we describe a machine learning method for classifying demonstrative pronouns in Arabic. We have evaluated our approach on a corpus of 365585 words which contain 14318 demonstrative pronouns and we have obtained encouraging results: 99.3% as F-Measure.

MOTS-CLES : Pronoms démonstratifs, résolution des anaphores, traitement de la langue arabe.
KEYWORDS: Demonstrative pronouns, anaphora resolution, ANLP.

1 Introduction

La résolution des anaphores pronominales est l'une des branches les plus actives du domaine de Traitement Automatique du langage Naturel (TALN). Elle consiste à identifier l'antécédent pour chaque pronom. La première étape de la tâche de résolution des pronoms anaphoriques consiste à distinguer les occurrences référentielles (anaphoriques et cataphoriques) de celles non-référentielles. Au niveau de cette étape, le système détecte toutes les occurrences non-référentielles afin d'éviter la recherche d'un antécédent qui n'existe pas. En outre, la classification des occurrences référentielles en anaphoriques et cataphoriques est importante pour un système de résolution. En effet,

pour les pronoms anaphoriques, le système doit chercher l'antécédent dans le segment localisé avant le pronom, alors que, pour les pronoms cataphoriques, le système doit chercher l'antécédent dans le segment qui suit le pronom. Par conséquent, cette classification peut améliorer la performance du système.

Considérons les exemples suivants :

(1) انه من الصعب إيجاد حل لهذا المشكل

/ Ain~ahu mina AlS~aEobi Ii?jaAdu HalK liha*aA Almu$okili/

Il est difficile de trouver une solution à ce problème.

(2) عندما دخلت أختي المستشفى العام الماضي، كنّا نحضر لها الكسكسي والعجين بأنواعه

/EinodamaA daxalato Oxotiy Almusota$ofaY AlEaAma AlmaADiy, kuna~A nuHaD~iru lahaA Alkusokusiy waAlEaji?na biOanowaAEihi/

Quand ma sœur entra à l'hôpital l'année dernière, nous lui avons apporté le couscous et les divers types de pâtes

(3) هذا الاختراع مهم جدا

/ha*aA AlIixotiraAEu muhimN jid~FA/

Cette invention est très importante.

Le pronom (هـ/hu/*il*), dans l'exemple (1) ne se réfère à aucun syntagme nominal. Il est donc non-référentiel. Cependant, dans l'exemple (2), le pronom (ها/hA/*lui*) possède comme antécédent le syntagme (أختي/>xty/*ma sœur*) ; de ce faite il est anaphorique. Le pronom (هذا/h*A/cette), dans l'exemple (3) est cataphorique puisqu'il se réfère au nom (الاختراع/AlAxtrAE/invention) situé après le pronom. Ainsi, un système de résolution des anaphores doit chercher les antécédents seulement pour les pronoms des exemples (2) et (3).

Vu l'importance de la classification automatique des pronoms, plusieurs chercheurs se sont intéressés à ce sujet. Certains travaux ont visé la distinction des pronoms personnels de ceux impersonnels ((Lappin et Leass, 1994), (Boyd et a, 2005), (Weissenbacher et Nazarenko, 2007), (Hammami et al, 2010)). D'autres travaux se sont intéressés aux pronoms démonstratifs tels que les travaux de (Muller, 2007), (Byron, 2002) pour l'anglais, le travail de (Navaretta, 2009) pour le danois et le travail de (Dutta et al, 2010) pour l'Hindo. A notre connaissance, il n'existe pas de travaux similaires pour la classification des pronoms démonstratifs en langue arabe.

Dans cet article, nous proposons une méthode d'apprentissage pour la classification automatique des pronoms démonstratifs en langue arabe. Cette méthode permet de classifier les pronoms démonstratifs en pronoms démonstratifs cataphoriques et pronoms démonstratifs anaphoriques. Elle se base, d'une part, sur un ensemble de critères contextuels et d'autre part sur des techniques d'apprentissage. La section 2 présente la spécificité des pronoms démonstratifs arabes. La section 3 donne un aperçu sur l'état de l'art pour la classification des pronoms. Dans la quatrième section, nous décrivons la méthode proposée pour la classification des pronoms démonstratifs. Enfin, nous présentons nos expérimentations et nous discutons les résultats obtenus.

2 Les pronoms démonstratifs en langue Arabe

Selon la littérature, les pronoms démonstratifs sont fréquemment utilisés en langue arabe. Comme pour l'anglais (this, these…) et le français (ceci, celui-ci, celle-là, celui-là,…), il existe dans la langue arabe des pronoms qui désignent le singulier (هذا/h*A/, هذه/h*h/) et le pluriel (هؤلاء/hWlA'/, أولئك/Awl}k/). Ce pendant, il existe des pronoms démonstratifs qui désignent le duel tels que : هذان/h*An/, هاتان/htAn/, ذلكما/*lkmA/, تلكما/tlkmA/, ذانك/*Ank/. Ce qui n'est pas le cas pour le français ou l'anglais. En outre, il existe des pronoms, qui sont considérés comme démonstratifs, et qui désignent le temps (حينذاك/Hyn*Ak/, أنذاك/On*Ak/) et le lieu (هنا/hnA/, هناك/hnAkA/, هنالك/hnAlkA/, ههنا/hhnA/, ثَمَّ/vm~/).

D'après notre étude statistique, il y a des pronoms qui sont utilisés beaucoup plus que d'autres tels que les pronoms هذا/h*A/، تلك/tlk/, هذه/h*h/، ذلك/*lk/. L'utilisation des pronoms démonstratifs (*lkmA/ذلكما, *lkm/ذلكم, tlkmA/تلكما, *lkn/ذلكن) est presque négligeable (ils apparaissent dans le saint coran ou dans les anciens livres arabes).

3 Travaux antérieurs

L'anaphore pronominale est le type d'anaphore le plus fréquent (Mitkov, 2002), c'est pourquoi la résolution automatique des pronoms est un domaine de recherche qui a suscité énormément d'attention depuis plusieurs années. Un système de résolution des anaphores pronominales doit être capable de distinguer les occurrences des pronoms non-référentielles de celles référentielles avant de s'attaquer à leur résolution. De nombreux travaux se sont intéressés particulièrement à cette étape vue son importance et sa difficulté.

La plus part des chercheurs se sont intéressés aux pronoms personnels. Nous pouvons distinguer trois types d'approches : une approche à base de règles telle que les travaux de (Paice et Husk, 1987), (Lappin et Leass, 1994) et (Denber, 1998) pour l'anglais, et (Hammami, 2009) pour l'arabe. Afin de remédier aux inconvénients rencontrés au niveau de l'approche précédente, d'autres auteurs ont adopté une approche numérique basée sur des méthodes d'apprentissages telle que les travaux d' (Evans, 2001) et (Bergsma, 2008).

D'autres, ont fait recours à la combinaison des deux approches précédentes telles que les travaux de (Boyd et al, 2005), (Weissenbacher et Nazarenko, 2007), (Hammami et al., 2010) et (AbdulMajeed, 2011).

Cependant, la classification des pronoms démonstratifs n'a pas encore reçu beaucoup d'attention. (Byron, 2002) décrit un système pour la résolution des pronoms it, this et that dans les dialogues dans un domaine spécifique. Ce système, appelé PHORA, est implémenté et basé sur des connaissances sémantiques. Le résultat d'évaluation était 67% et 62% respectivement pour la précision et le rappel.

(Müller, 2007) a proposé une approche basée sur l'apprentissage automatique pour la résolution des pronoms *it, this* et *that*. Cet algorithme a utilisé cinq corpus différents composés de dialogues pour l'apprentissage et le test, et il repose exclusivement sur l'annotation du corpus. Les résultats de cet algorithme sont moins performants que ceux

des algorithmes reposant sur des connaissances linguistiques et des structures de discoures complexes.

(Navaretta, 2009) décrit des expérimentations d'apprentissage supervisé (classification) et non supervisé (clustering) dans le but de reconnaitre la fonction du pronom neutre singulier dans la langue danoise. Le corpus utilisé est très hétérogène. Il est composé de quatre parties : des textes écrits, des transcriptions de monologue, des dialogues et une interview de TV. La classification de la fonction du pronom neutre singulier comprend neuf classes telles que la classe explétive (non-référentiel), cataphorique, déictique, anaphore individuelle, etc. Le meilleur algorithme pour le clustering est Expectation Maximisation de l'outil Weka. Les résultats obtenus pour les textes écrits sont plus intéressants que pour les autres corpus. Pour la classification, plusieurs algorithmes ont été examinés. Pour estimer la performance de son système, Navaretta utilise la méthode d'échantillonnage validation croisée (10 cross-validation). Elle a fixé l'algorithme ZeroR de Weka comme baseline. Les meilleurs algorithmes sont NBTree, SMO, SMO et KStar respectivement pour les corpus de textes, monologues, dialogue et l'interview.

(Dutta et al., 2010) proposent une application pour la classification des pronoms démonstratifs « yeh », « veh », « iss » et « uss » en langue Hindo. Cette classification est basée sur le formalisme du réseau de neurone probabiliste (PNN). Comme première étape, ils ont extrait des patrons et des caractéristiques pour l'identification des pronoms démonstratifs indirects. Ensuite, ils ont appliqué l'algorithme basé sur le modèle PNN en utilisant la validation croisée. Enfin, des expérimentations sont effectuées pour l'ensemble des données contenant les pronoms démonstratifs et aussi les occurrences des pronoms démonstratifs non référentielles. Les meilleurs résultats sont 94.90% pour tous les pronoms démonstratifs et 84.16% pour les pronoms non-référentiels comme taux de réussite.

4 Méthode proposée

La méthode que nous proposons pour la classification automatique des pronoms démonstratifs arabes est composée de deux phases à savoir la phase d'apprentissage et la phase de test.

La phase d'apprentissage permet d'apprendre à classifier les pronoms. Elle accepte, en entrée, un corpus annoté et elle est composée de trois étapes à savoir la segmentation, l'analyse morphologique et l'extraction des règles. L'étape de segmentation consiste à segmenter les textes de notre corpus. Les textes sont segmentés en phrases dans le but de connaitre les frontières des phrases contenant un pronom démonstratif. Le texte segmenté sera par la suite analysé morphologiquement afin d'identifier les caractéristiques morphologiques des mots de chaque texte de notre corpus. Cette étape va servir à déterminer les valeurs des critères de classification que nous utilisons dans l'étape d'extraction des règles. Pour établir cette dernière, nous avons dégagé huit critères de classification à savoir :

- POS+1 : (Part Of Speech) ce critère prend la catégorie du mot qui suit le pronom. Les valeurs possibles pour ce critère sont : Nom-propre, Nom, Particule, Délimiteur ou Inconnu (dans le cas où la catégorie du mot n'a pas pu être identifiée).

- Type : Dans le cas où le critère *POS+1* prend la valeur Particule, alors le critère *Type* prend le type de cette particule qui peut être : particule de coordination, particule de conjonction, particule d'exception, conjonction d'appel, conjonction de négation, conjonction de condition, etc.
- Determine : Dans le cas où le critère *POS+1* prend la valeur Nom, le critère *Determine* prend la valeur « match » si ce Nom est défini (c'est-à-dire agglutiné à ال). Sinon il prend la valeur « nomatch ».
- Enclitique : Si le mot qui suit le pronom est un *Nom*, et ce nom est agglutiné à un enclitique, alors, ce critère prend la valeur « match ».
- Proclitique+1 : Si le mot qui suit le pronom est agglutiné à un proclitique, alors, ce critère prend la valeur « match ».
- Bimafidhalika : Ce critère prend la valeur « match » si le pronom (*lk / ذلك) est précédé par une succession des deux mots (bmA /بما) et (fyفي/).
- MotSpec : Si le pronom est suivi d'un mot spécifique ce critère prend la valeur « match ». La liste des mots spécifiques est composée des mots suivants: OyDA/أيضا, gyr/غير, mmA/ممّا, kl/كل.
- Pronom : ce critère reçoit le pronom à apprendre.

L'étape d'extraction des règles exploite ces critères de classification pour produire des règles appelées règles de classification, en utilisant un algorithme d'apprentissage.

La phase de test permet de classifier un nouveau pronom démonstratif en pronom anaphorique ou pronom cataphorique. Elle accepte en entrée un texte brut qui sera par la suite segmenté en phrases et analysé morphologiquement. L'étape d'identification des pronoms démonstratifs consiste à identifier les pronoms démonstratifs figurant dans le texte afin de les classifier automatiquement. La détection des pronoms se fait d'une manière automatique en examinant le texte analysé morphologiquement et en faisant ressortir les mots qui ont comme catégorie pronom démonstratif (Asm I$Arp/ اسم إشارة). Enfin, les pronoms identifiés seront classifiés en pronoms démonstratifs anaphoriques ou pronoms démonstratifs cataphoriques en se basant sur les règles d'extraction générées par la phase d'apprentissage.

5 Expérimentations

5.1 Corpus

Le processus de classification à base d'apprentissage nécessite généralement un corpus annoté afin d'assurer la phase d'entraînement. Il est à signaler que la constitution d'un corpus de référence (corpus d'apprentissage) est coûteuse. Ainsi, et vu le manque de corpus étiquetés pour la langue arabe, nous avons procédé à une étape d'annotation binaire des documents constituant notre corpus. Il s'agit d'attribuer à chaque pronom démonstratif la classe anaphorique ou cataphorique. Pour accélérer notre travail, nous avons développé un système d'annotation manuelle AnnotAr. Ce système accepte en entrée un texte segmenté en mot sous le format XML et permet à l'utilisateur d'annoter les pronoms démonstratifs dans le texte en pronoms anaphoriques ou cataphoriques. Le texte annoté est enregistré sous format XML où chaque pronom démonstratifs est étiqueté par la balise qui lui correspond (c'est-à-dire <ANA> pour le pronom anaphorique et <CATA> pour le pronom cataphorique).

Le corpus d'apprentissage est composé d'un ensemble d'articles de presse d'ELMASRY ALYOUM 2010, de textes de livres scolaires, de l'enseignement Tunisien de différents niveaux (un texte contient en moyenne vingt cinq phrases), des manuels d'utilisation (la taille moyenne d'un manuel est de trente pages) et un extrait du Penn Arabic TreeBank (ATB). Nous avons choisi un corpus de nature variée parce que nous estimons que plus le corpus est diversifié plus il sera représentatif. Ce corpus contient un total de 365585 mots et nous a permis d'analyser 14318 pronoms démonstratifs (où 32.15% sont anaphoriques et 67.85% sont cataphoriques).

Corpus	Anaphorique		Cataphorique		Total
	Nombre	Pourcentage(%)	Nombre	Pourcentage(%)	
Livres	1072	33.29	2148	66.71	3220
Journaux	2562	31.88	5472	68.12	8034
Manuels	155	29.41	372	70.59	527
ATB	815	47.32	1722	52.68	2537
Total	4604	32.15	9714	67.85	14318

TABLE 1:Statistiques du corpus

5.2 Résultats et discussion

Nous avons effectué nos expérimentations de classification en utilisant le système Weka (Frank, Witten, 2005) qui permet de tester et de comparer plusieurs algorithmes. Nous avons choisi de tester les algorithmes suivants: IBk, JRip, NBTree et NaiveBayes.

Nous avons procédé à une validation croisée pour valider les résultats de nos expériences. Nous avons sélectionné aléatoirement le neuf dixième du corpus pour l'apprentissage. Nous avons ensuite appliqué notre système sur le un dixième restant. Nous avons réitéré dix fois ces opérations en changeant à chaque fois la partie de test, pour obtenir la moyenne des performances de chaque itération. Les attributs pertinents d'après Weka sont : POS + 1, Type, Determine, procltique + 1, Pronom et motSpec.

Afin de bien évaluer notre système, nous avons implémenté un système Baseline à base des règles. Le système Baseline repose sur six règles contextuelles qui se basent principalement sur les caractéristiques morphologiques du mot qui suit le pronom démonstratif (ex. pr_dem + Nom-défini --> pr-cataph, pr_dem + pr-relatif --> pr-cataph). Les résultats obtenus sont présentés dans le Tableau 2.

Algorithme	Résultats avec annotation morphologique automatique (MORPH2)			Résultats avec annotation morphologique manuelle		
	Précision	Rappel	F-Mesure	Précision	Rappel	F-Mesure
IBk	94.7%	94.7%	94.7%	99.3%	99.3%	99.3%

JRip	93.9%	93.9%	93.9%	99%	99%	99%
NBTree	94.7%	94.7%	94.7%	99.2%	99.2%	99.2%
NaiveBayes	92.8%	92.8%	92.8%	96.2%	96.2%	96.2%
Baseline	78.09%	75.39%	76.72%	97.87%	98.89%	98.38%

TABLE 2 : Résultats obtenus avec la validation croisée

En examinant les mesures de rappel, précision et F-mesure calculées sur le corpus d'évaluation, nous remarquons que les résultats sont très encourageants.

D'une part, l'utilisation de l'apprentissage a amélioré les résultats d'une manière significative (16.61% et 15.81% respectivement pour IBk et JRip) par rapport au Baseline (méthode à base de règles). Cela justifie notre choix d'une méthode d'apprentissage qui réduit l'erreur d'estimation en déterminant le poids des attributs discriminants pour le domaine du corpus.

D'autre part, nous remarquons que l'algorithme IBk (K Plus Proche Voisin) donne les meilleurs résultats par rapport aux autres algorithmes. Ensuite, nous avons mené deux évaluations. Au niveau de la première, nous avons utilisé l'analyseur morphologique MORPH2 (Chaaben et al, 2010) pour l'étiquetage morphologique de notre corpus. Au niveau de la deuxième évaluation, nous avons corrigé manuellement les résultats de MORPH2. En effet, l'utilisation d'un étiquetage morphologique manuel a amélioré les résultats d'une manière significative (environ 5% pour l'apprentissage et 21.66% pour le Baseline).

Les principales erreurs sont dues à des constructions spécifiques de phrases contenant un pronom démonstratif ainsi le manque de ponctuations en langue arabe (Belguith et al., 2005). Citons l'exemple suivant :

(4) ...قد يستهلك ذلك الكثير من الماء.
*/qado yasotaholiku *alika Alkaviyra mina AlmaA'i/*
Ça peut consommer beaucoup d'eau.

Dans cet exemple, le pronom démonstratif (ذلك, /*lk/, ça) est suivi par le nom défini (الكثير, /Alkvyr/, beaucoup). En appliquant l'apprentissage ou la méthode de Baseline, ce pronom démonstratif est classé cataphorique alors qu'il est anaphorique. Cette fausse classification est due à l'absence de la virgule après le pronom démonstratif.

6 Conclusion

La classification des pronoms démonstratifs est une étape très importante dans le processus de la résolution de l'anaphore pronominale. Dans cet article, nous avons proposé une méthode d'apprentissage pour la classification binaire des pronoms démonstratifs en langue arabe en pronom anaphoriques et cataphoriques. Cette méthode d'apprentissage proposée atteint des résultats meilleurs que celle à base des règles. Ainsi l'algorithme K-PPV a donné un meilleur résultat. En se basant sur ces résultats, nous envisageons de chercher les antécédents des pronoms et de terminer les étapes de la résolution d'anaphores.

Références

C. PAICE ET G. HUSK (1987). Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun it. *Computer Speech and Language.*

D. WEISSENBACHER ET A. NAZARENKO (2007). A bayesian classifier for the recognition of the impersonal occurrences of the it pronoun. *In Proceedings of DAARC'07, 2007.*

S. LAPPIN ET H. LEASS (1994). An algorithm for pronominal anaphora resolution, *Computational Linguistics, 20(4), 1994, p. 535–561.*

S. MEZGHANI HAMMAMI., R.SELLAMI, L. HADRICH BELGUITH (2010). A Bayesian Classifier for the Identification of Non-referential Pronouns in Arabic. *The 7th INFOS 2010.*

S. HAMMAMI, L. BELGUITH, A. BEN HAMADOU (2009). A Rule-Based Method for Detecting Arabic Anaphoric Pronouns. *Proceedings of the 7th DAARC'2009, Goa-India, 2009.*

R. EVANS (2001). Applying machine learning toward an automatic classifi cation of it. *Literary and linguistic computing, 16, 2001, p. 45–57.*

M. DENBER (1998). Automatic resolution of anaphora in English. *Eastman Kodak Co, 1998.*

ABDUL-MAGEED, M. 2011. Automatic detection of Arabic non-anaphoric pronouns for improving anaphora resolution. *Asian Lang. Inform. Process. (March 2011).*

S. BERGSMA, D. LIN ET R. GOEBEL (2008). Distributional Identification of Non-Referential Pronouns. *ACL, Columbus Ohio, 2008, p. 10-18.*

A. BOYD, W. GEGG-HARRISON ET D. BYRON (2005). Identifying non-referential it: a machine learning approach incorporating linguistically motivated features. *Workshop on Feature Engineering for Machine Learning in Natural Language Processing, 2005.*

C. MÜLLER (2007).Resolving It, This, and That in Unrestricted Multi-Party Dialog. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007.*

D. BYRON (2002). Resolving Pronominal Reference to Abstrac Entities. *In Proceedings of th 40th Annual Meeting of the Association of Computation Linguistics (ACL), Philadelphia 2002.*

C. NAVARETTA (2009). Automatic Recognition of the Function of Singular Neuter Pronouns in Texts and Spoken Data. *In DAARC 2009.*

K. DUTTA , N. PRAKASH, S. KAUSHIK (2010). Probabilistic neural network approach to the classification of demonstrative pronouns for indirect anaphora in Hindi. *International Journal Information Technology and Intelligent Computing, 2010.*

N. CHAÂBEN KAMMOUN, L. HADRICH BELGUITH ET A. BEN HAMADOU (2010). The MORPH2 new version: A robust morphological analyzer for Arabic texts. *JADT'2010.*

E. FRANK, IAN H. WITTEN (2005). Practical Machine Learning Tools and Techniques, Second Edition. *(Morgan Kaufmann series in data management systems).*

BELGUITH, L., BACCOUR, L. AND MOURAD, G. (2005). Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules *(TALN'2005).*

MITKOV (2002): R. Mitkov, «Anaphora resolution». Longman. 2002.

Un annotateur automatique d'expressions temporelles du français et son évaluation sur le TimeBank du français

André Bittar Caroline Hagège
XRCE, 6 Chemin de Maupertuis, 38240 Meylan, FRANCE
Andre.Bittar@xrce.xerox.com, Caroline.Hagege@xrce.xerox.com

RÉSUMÉ

Dans cet article, nous présentons un outil d'extraction et de normalisation d'un sous-ensemble d'expressions temporelles développé pour le français. Cet outil est mis au point et utilisé dans le cadre du projet ANR Chronolines[1] et il est appliqué sur un corpus fourni par l'AFP. Notre but final dans le cadre du projet est de construire semi-automatiquement des chronologies événementielles à partir de la base de depêches de l'AFP. L'une des étapes du traitement est l'analyse de l'information temporelle véhiculée dans les textes. Nous avons donc développé un annotateur d'expressions temporelles pour le français que nous décrivons dans cet article. Nous présenterons également les résultats de son évaluation.

ABSTRACT

An Automatic Temporal Expression Annotator and its Evaluation on the French TimeBank

In this article, we present a tool that extracts and normalises a subset of temporal expressions in French. This tool is being developed and used in the ANR (French National Research Agency) project Chronolines, applied to a corpus of provided by the Agence France Presse. The aim of the project is to semi-automatically construct event chronologies from this corpus. To do this, a detailed analysis of the temporal information conveyed by texts, is required. The system we present here is the first version of a temporal annotator that we have developed for French. We describe it in this article and present the results of an evaluation.

MOTS-CLÉS : Analyse temporelle, évaluation.

KEYWORDS: Temporal processing, evaluation.

1 Introduction

Le travail présenté ici s'insère dans un cadre plus ambitieux qui est la constitution semi-automatique de chronologies événementielles à partir d'une requête effectuée sur un grand ensemble de depêches de l'AFP pour le français et pour l'anglais. Afin de pouvoir constituer des chronologies événementielles, il est capital de pouvoir analyser dans un premier temps le contenu textuel (afin de repérer les événements) mais aussi de reconnaître et normaliser les expressions temporelles associées à ces événements. L'outil présenté ici est un annotateur d'un sous-ensemble d'expressions temporelles qui :
– repère les expressions temporelles

[1] ANR-10-CORD-010, http://www.chronolines.fr

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 463–470,
Grenoble, 4 au 8 juin 2012.

– normalise ces expresssion temporelles
– rajoute des annotations concernant la modalité des événements auxquels ces expressions se rapportent

Dans une première partie, nous présentons un bref état de l'art concernant l'analyse automatique de la temporalité en général et pour le français en particulier. Puis nous décrivons l'outil que nous avons développé pour le français. Nous présenterons enfin les résultats obtenus par l'annotateur en les comparant avec le TimeBank du français (TBF), un corpus de textes en français annotés selon la norme ISO-TimeML (Pustejovsky *et al.*, 2010).

2 Etat de l'art

L'analyse de la temporalité est un élément important pour un grand nombre de tâches et de ressources relevant du traitement informatique des textes. (Li *et al.*, 2005) montre l'importance de l'analyse de la composante temporelle pour les systèmes de Questions/Réponses. Pour le résumé multi-document, l'ajout de la composante temporelle aide à repérer les éléments textuels véhiculant une information similaire (Barzilay et Elhadad, 2002). Les grandes bases de connaissances constituées automatiquement ou semi-automatiquement grâce à l'extraction d'information textuelles s'enrichissent actuellement d'une composante temporelle (Wang *et al.*, 2010). Par ailleurs, la norme ISO-TimeML est aujourd'hui largement adoptée, tant dans le cadre de la mise en place de ressources annotées avec des informations temporelles pour diverses langues dont le français (Bittar *et al.*, 2011), mais aussi lors des compétitions TempEval (Pustejovsky et Verhagen, 2010) au cours desquelles divers outils d'annotation automatique des informations temporelles sont évalués. Plus spécifiquement pour le français, outre le TimeBank mentionné ci-dessus, plusieurs travaux visant à l'analyse automatique de la temporalité ont vu le jour. Concernant l'annotation et le typage des expressions temporelles (ET) nous pouvons citer les travaux de (Battistelli *et al.*, 2008) qui présente une représentation algébrique des expressions de type date, de (Ehrmann et Hagège, 2009) qui propose un typage des ET accompagné de critères syntaxiques et sémantiques, mais aussi dans le domaine de l'annotation automatique, l'outil décrit dans (Parent *et al.*, 2008), ainsi que celui de (Teissèdre *et al.*, 2011).

3 Description de l'annotateur

L'annotateur que nous avons développé est intégré à un analyseur linguistique (Aït-Mokhtar *et al.*, 2002) qui produit une analyse syntaxique en dépendances à partir d'un texte d'entrée (texte brut ou XML). L'annotateur a été développé pour les besoins du projet Chronolines qui vise à traiter le corpus de dépêches (couvrant les années allant de 2004 à 2011) mis à disposition par l'Agence France-Presse. Ce corpus est constitué d'environ 1 million de documents (chaque document correspondant à une dépêche) comprenant environ 9,4 millions d'expressions temporelles de tout type. Ces dépêches sont disponibles dans un format XML (NewsML).

Le module spécifique pour la reconnaissance des expressions temporelles est constitué de plusieurs éléments qui seront détaillés plus bas :
– Ajout d'information lexicale permettant de typer les adverbes de temps.
– Règles locales permettant de délimiter les ET et si possible de les typer. Ces règles sont intégrées

à la grammaire générale de l'analyseur.
– Règles utilisant les dépendances syntaxiques permettant de procéder à un typage plus fin de ces ET. Ces règles sont également intégrées à la grammaire générale.
– Programme Java externe à la grammaire qui utilise les informations linguistiques pour désambiguiser certaines expressions et pour procéder à la normalisation des ET sélectionnées.

3.1 Information lexicale

L'information lexicale spécifique à l'analyse temporelle consiste essentiellement en l'ajout de traits sémantiques sur des éléments lexicaux qui vont rentrer dans la composition d'une ET. Par exemple, les noms de jours (e.g. *lundi*) se voient attribuer un trait spécifique *[day :+]*. De même, les noms de mois, de fêtes, des adverbes de temps seront marqués dans le lexique.

3.2 Règles locales

Les règles locales vont permettre d'assembler des éléments lexicaux quand ils peuvent potentiellement correspondre à une expression temporelle incluant plusieurs constituants de base. Par exemple, une expression comme *mi-mars 2012* est segmentée originellement par l'analyseur morphologique en 4 segments *mi* + - + *mars* + *2012*. Une règle locale regroupe ces quatre segments afin de constituer une seule expression temporelle. Lors de l'application de ces règles locales, un premier typage des expressions temporelles est effectué dans le cas où celles-ci ne sont pas ambiguës. Par exemple, une expression comme *mi-mars 2012* peut être typée comme une date absolue[2] sans avoir recours au contexte .

3.3 Règles de dépendances raffinant le typage

Certaines expressions temporelles reconnues par une analyse lexicale ou par l'application de règles locales ne peuvent cependant pas être typées *a priori*. En effet, l'analyse de la seule expression ne permet de déterminer de quel type d'expression il s'agit et seul un contexte plus large permet de désambiguïser. Parfois, une même expression mettant en jeu des unités linguistiques constitutives d'une ET peut s'avérer, en contexte, ne pas être une ET. Par exemple, une expression telle que *trois ans*, n'est pas une ET dans (1) mais correspond à une durée dans (2). L'expression *en avril* dans (3) correspond à une date relative alors que dans (4), elle correspond à une date récurrente (qui peut être paraphrasée par *tous les mois d'avril*).

1. Jean fêtera bientôt ses **trois ans**.
2. Il est resté **trois ans** sans la voir.
3. Il était malade **en avril**.
4. **En Avril**, il fait le grand nettoyage de printemps.

Grâce à l'analyse syntaxique sous-jacente, des restrictions utilisant à la fois des informations syntaxiques (fonction syntaxique de l'ET potentielle), et des informations sémantiques lexicales

[2] nous détaillons à la section 3.5 les différents types de dates que nous extrayons et la terminologie adoptée pour les distinguer.

(verbe *rester* est un verbe de permanence), permettent de filtrer et de mieux typer des ET extraites lors des étapes précédentes de traitement (analyse lexicale et règles locales).

3.4 Programme externe

Une API java de l'analyseur permet d'étendre les traitements linguistiques et d'utiliser les résultats des analyses dans du code extérieur à l'analyseur. La normalisation[3] des ET extraites est effectuée par ce biais. Une fois de plus, à ce stade, l'analyse des expressions temporelles est encore raffinée afin de pouvoir procéder correctement à la normalisation. Par exemple, une expression comme *lundi* est une expression de type date relative par rapport au moment de l'énonciation (ME) mais elle peut être antérieure ou postérieur au ME selon les contextes, ainsi qu'en témoignent les exemples suivants.

1. Elle est partie **lundi**.
2. Elle partira **lundi**.

3.5 Type de dates extraites reconnues

Bien que nous délimitions tout type d'ET, nous avons mis l'accent pour une première utilisation de l'annotateur sur la normalisation d'un sous-ensemble d'expressions temporelles qui est le suivant :

Nous considérons les dates absolues, et les dates relatives au moment de l'énonciation qui correspondent à des intervalles bornés ou à des points. Nous avons pour ce faire défini des critères de segmentation et de typage décrits dans (Bittar *et al.*, 2012). Toutes ces dates que nous extrayons doivent être normalisées.

Parmi ces expressions nous avons des ET comme *En 2003*, *En janvier 2003*, *le 24 juin 2010*, *le mois dernier*, *lundi*, *dans quatre mois*, etc. Les trois premiers exemples sont des dates absolues, les exemples suivants sont des dates relatives au ME dans la mesure où pour procéder à la normalisation de ces dates, il est nécessaire de connaître la date correspondant à l'assertion durant laquelle cette date est mentionnée.

3.6 Extraction d'information sur la modalité

Dans le cadre plus large du projet, nous avons souhaité distinguer les événements datés factuels (c'est à dire les événements considérés comme avérés par l'auteur de la dépêche) des événements datés hypothétiques ou relevant du discours rapporté. Dans la mesure où nous utilisons un analyseur linguistique, nous disposons des liens syntaxiques entre l'ET extraite et le prédicat nominal ou verbal que cette ET modifie. La prise en compte de la modalité dans le traitement de la temporalité est un vaste et riche domaine (voir (Battistelli, 2009) pour une présentation détaillée) que nous n'avons pas considéré dans son ensemble. Nous avons cependant distingué

[3]Nous entendons ici par normalisation le fait d'attribuer une valeur correspondante au calendrier (éventuellement sous-spécifiée) à une ET.

<DCT value="20040101"/>
L' <EN TYPE="LOCORG">Irlande</EN> s'apprête à prendre <EC TYPE="DATE" SUBTYPE="REL" REF="ST" value ="20040101">jeudi</EC> la présidence tournante de l' <EN TYPE="ORG">Union européenne</EN> qui doit vivre <EC TYPE="DATE" SUBTYPE="REL" REF="ST" value ="20040501" FACTUAL="MODAL">le 1er mai</EC> un élargissement historique...

FIG. 1 – Exemple de sortie de l'annotateur.

les cas suivants, qui se produisent de manière assez fréquente sur les corpus que nous traitons et qui nous semblent pertinents pour la finalité applicative que nous avons dans le cadre du projet.

– L'expression temporelle se trouve dans une proposition dont le verbe principal est à une forme modale ou future.
– L'expression temporelle se trouve dans une enchâssée qui relève du discours rapporté.
– L'expression temporelle est associée à un verbe *dicendi* introduisant un discours rapporté.

Dans le premier cas, l'ET est soit le modifieur d'un verbe utilisé à une forme modale ou au futur (information dont nous disposons grâce à l'analyse morpho-syntaxique), soit le modifieur d'un nom argument d'un verbe utilisé à une forme modale ou future. L'exemple ci-dessous marque l'ET *en 2006* par un attribut *FACTUAL="MODAL"* dans la mesure où le prédicat auquel elle se rapporte (*s'achever*) est employé à une forme modale.

Exemple :
Il a réaffirmé la primauté de son mandat de cinq ans , qui doit s'achever <EC TYPE="DATE" SUBTYPE="ABS" value ="2006XXXX" FACTUAL="MODAL">*en 2006*</EC>

Dans le deuxième cas, l'ET est modifieur d'un verbe d'une enchâssée dépendante d'un verbe *dicendi*. Dans ce cas, l'annotation de l'ET est enrichie par l'attribut *REPORTED="YES"*.

Exemple :
Il a annoncé que des élections se dérouleraient <EC TYPE="DATE" SUBTYPE="ABS" value ="2004XXXX" REPORTED="YES" FACTUAL="MODAL">*en 2004*</EC>

Enfin, dans le dernier cas, l'ET est modifieur d'un verbe marqué comme étant un verbe introducteur d'un discours rapporté. Dans ce cas, l'annotation de l'ET comporte l'attribut *DECLARATION="YES"*.

Exemple :
La Libye avait annoncé <EC TYPE="DATE" SUBTYPE="REL" REF="ST" DECLARATION="YES" value ="20031219">*le 19 décembre*</EC> *sa décision de renoncer aux armes de destruction massive.*

3.7 Exemple de sortie

L'annotateur prend en entrée du texte brut ou un texte au format XML. Il produit en sortie un texte au format XML qui correspond au texte initial enrichi par les annotations des ET que nous avons décrites. Les entités nommées sont également marquées. Un exemple de sortie est illustré dans la Figure 1.

4 Evaluation sur TimeBank

Afin de pouvoir évaluer notre outil, nous avons comparé les annotations produites avec le corpus TimeBank du français (TBF) annoté selon la norme ISO-TimeML. Cette évaluation a nécessité quelques adaptations afin de faire correspondre les sorties de l'outil d'analyse aux données de TimeBank. Ces adaptations sont décrites ci-dessous. Par ailleurs, l'information concernant la modalité et le discours rapporté n'a pas été évaluée.

4.1 Adaptations

Les adaptations ont été nécessaires pour trois raisons principales : la délimitation des ET selon notre approche n'est pas tout à fait semblable à celle adoptée par la norme ISO-TimeML, nous ne considérons qu'un sous-ensemble des ET envisagées dans TBF et enfin, le format de normalisation que nous adoptons est différent de celui de ISO-TimeML.

Dans TBF, les balises qui marquent les ET (<TIMEX3>) n'incluent pas les éventuels marqueurs de relation (ex. les prépositions temporelles telles que *avant, après, etc.*). En effet, la norme ISO-TimeML préconise de les annoter séparément avec une autre balise (<SIGNAL>). Selon notre schéma d'annotation (Bittar *et al.*, 2012) cependant, les marqueurs de relations sont annotés à l'intérieur d'une seule balise délimitant l'expression temporelle. Afin de résoudre cette différence, nous avons converti le TimeBank selon notre format par application d'un simple transducteur qui place le contenu textuel de la balise <SIGNAL> à l'intérieur de la balise <TIMEX3> qui la suit directement. Par ailleurs, notre annotateur a été adapté pour fournir une sortie contenant la balise (<SIGNAL>). Nous obtenons donc pour ce type d'expression la représentation finale(1).

1. <TIMEX3><SIGNAL>depuis< /SIGNAL>mars 2003< /TIMEX3>

Notre annotateur ne traite pour l'instant que des dates absolues et des dates relatives au moment de l'énonciation correspondant à des intervalles bornés et dont la granularité n'est pas inférieure au jour. Nous avons donc retiré du corpus TBF toutes les annotations des expressions ne correspondant à ces catégories. Concrètement, toute balise <TIMEX3> obéissant à l'un des quatre critères mentionnés ci-dessous a été supprimée du corpus de référence.

– l'attribut `temporalFunction="true"` (qui indique qu'il s'agit d'une date relative) est présent et l'attribut `anchorTimeID` est différent de `t1` (la DCT). Ce critère permet de retirer du corpus de référence toutes les dates relatives à un référent textuel.
– l'attribut `type` a une des valeurs `DURATION`, `SET` ou `TIME`. Ce critère permet de retirer du corpus de référence toutes les durées, les heures ou les agrégats temporels.
– l'attribut `value` est `PAST_REF`, `PRESENT_REF` ou `FUTURE_REF`. Ce critère permet de retirer du corpus de référence toutes les expressions de dates floues.
– la balise a l'attribut `MOD` indiquant que la date a un modifieur (de début, de fin, d'approximation, etc.). Ce critère permet également de retirer du corpus de référence des dates floues.

Enfin, nous avons fait converger les formats de la valeur normalisée des ET afin d'obtenir une représentation comparable entre TBF et les sorties de l'annotateur. Le corpus adapté pour notre évaluation contient 299 expressions temporelles annotées (sur les 608 du corpus original).

4.2 Resultats pour le français

Les performances, en termes de rappel, précision et F-mesure, ainsi que la mesure kappa (Cohen, 1960), figurent dans le Tableau 1. Pour la détection des expressions temporelles, les performances sont satisfaisantes, mais peuvent encore être améliorées. Les erreurs sont dues essentiellement à des manques de couverture dans la grammaire, qui est encore en cours de développement. Aucune erreur de typage des expressions n'a été commise. De très bons résultats ont été obtenus pour la normalisation des expressions. La principale source d'erreurs pour la normalisation provient des cas où une ET apparaît dans un contexte où elle n'est pas reliée par une dépendance à un prédicat verbal. Ceci est parfois le cas lorsque la clause où apparaît l'ET ne contient effectivement pas de verbe, mais cela peut également se produire suite à une erreur de l'analyseur syntaxique. Dans un de ces cas, le temps verbal est donc indisponible pour le calcul de la valeur correcte normalisant l'ET. Enfin, lors de l'évaluation, un certain nombre de désaccords entre la sortie du système et la référence ont révélé des erreurs du TBF. Ces erreurs n'ont pas été prises en compte pour l'évaluation et elles ont été transmises au gestionnaire du corpus.

	Précision	Rappel	F-mesure	Kappa
Étendue des balises	0.90	0.84	0.87	0.71
Attribut `type`	1.0	1.0	1.0	1.0
Attribut `value`	0.94	1.0	0.96	0.92

TAB. 1 – Performances du système sur l'ensemble du corpus d'évaluation.

5 Conclusion

Nous avons développé une première version d'un outil d'analyse des ET du français à partir de textes tout-venant. Nous avons utilisé cet annotateur pour effectuer des expériences visant à extraire les dates importantes mentionnées dans de grands volumes de texte. L'annotateur a été évalué grâce au TimeBank du français. L'élargissement de cet annotateur à d'autres types d'ET est en cours (prise en compte d'autres des dates référentielles par rapport à un référent introduit dans le discours, des dates répétitives (fréquences dans la terminologie ISO-TimeML), et des dates qui ne peuvent être assimiliée à un point ou à un intervalle temporel borné).

Remerciements

Remerciement à l'ANR qui a financé une partie de ce travail, ainsi qu'à X. Tannier, R. Kessler et V. Moriceau qui ont utilisé et commenté les sorties de l'annotateur. Nous remercions également D. Teyssou de l'AFP qui nous a donné accès au corpus de dépêches.

Références

Aït-Mokhtar, S., Chanod, J.-P. et Roux, C. (2002). Robustness beyond Shallowness : Incremental Deep Parsing. *Natural Language Engineering*, 8:121–144.

Barzilay, R. et Elhadad, N. (2002). Inferring Strategies for Sentence Ordering in Multidocument News Summarization. *Journal of Artificial Intelligence Research*, 17:35–55.

Battistelli, D. (2009). *La temporalité linguistique : circonscrire un objet d'analyse ainsi que des finalités à cette analyse*. Université Paris-Ouest Nanterre La Défense (Paris 10).

Battistelli, D., Couto, J., Minel, J.-L. et Schwer, S. (2008). Représentation algébrique des expressions calendaires et vue calendaire d'un texte. *In* (Bechet *et al.*, 2008).

Bechet, F., Bellot, P., Bonastre, J.-F. et Jimenez, T., éditeurs (2008). *Actes de TALN 2008 (Traitement automatique des langues naturelles)*, Avignon. ATALA, LIA.

Bittar, A., Amsili, P., Denis, P. et Danlos, L. (2011). French TimeBank : An ISO-TimeML Annotated Reference Corpus. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers*, volume 2, Portland. Association for Computational Linguistics.

Bittar, A., Hagège, C., Tannier, X., Moriceau, V. et Teissèdre, C. (2012). Temporal Annotation : A Proposal for Guidelines and an Experiment with Inter-annotator Agreement. *In Proceedings of LREC 2012 - to appear*, Istanbul. ELRA.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 43(6):551–558.

Ehrmann, M. et Hagège, C. (2009). Proposition de caractérisation et de typage des expressions temporelles en contexte. *In* (Nazarenko et Poibeau, 2009).

Li, W., Li, W., Lu, Q. et Wong, K.-F. (2005). A Preliminary Work on Classifying Time Granularities of Temporal Questions. *In Proceedings of Second international joint conference in NLP (IJCNLP 2005)*, Jeju Island, Korea.

Nazarenko, A. et Poibeau, T., éditeurs (2009). *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis. ATALA, LIPN.

Parent, G., Gagnon, M. et Muller, P. (2008). Annotation d'expressions temporelles et d'événements en français. *In* (Bechet *et al.*, 2008).

Pustejovsky, J., Lee, K., Bunt, H. et Romary, L. (2010). ISO-TimeML : An international standard for semantic annotation. *In* Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M. et Tapias, D., éditeurs : *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Pustejovsky, J. et Verhagen, M. (2010). SemEval-2010 Task 13 : Evaluating Events, Time Expressions, and Temporal Relations (TempEval-2).

Teissèdre, C., Battistelli, D. et Minel, J.-L. (2011). Recherche d'information et temps linguistique : une heuristique pour calculer la pertinence des expressions calendaires. *In Actes de TALN 2011 (Traitement automatique des langues naturelles)*, Montréal. ATALA.

Wang, Y., Zhu, M., Qu, L., Spaniol, M. et Weikum, G. (2010). Timely YAGO : Harvesting, Querying, and Visualizing Temporal Knowledge from Wikipedia. *In Proceedings of the* 13^{th} *International Conference on Extending Database Technology (EDBT), Lausanne, Switzerland, March 22-26*, pages 697–700.

Vers le FDTB : French Discourse Tree Bank

Laurence Danlos Diégo Antolinos-Basso Chloé Braud Charlotte Roze
ALPAGE, Université Paris Diderot, 175 rue du Chevaleret, 75013 Paris
prenom.nom@linguist.jussieu.fr

RÉSUMÉ

Nous présentons les premiers pas vers la création d'un corpus annoté en discours pour le français : le French Discourse TreeBank enrichissant le FTB. La méthodologie adoptée s'inspire du Penn Discourse TreeBank (PDTB) mais elle s'en distingue sur au moins deux points à caractère théorique. D'abord, notre objectif est de fournir une couverture totale d'un texte du corpus, tandis que le PDTB ne fournit qu'une couverture partielle, qui ne peut donc pas être qualifiée d'analyse discursive comme celle faite en RST ou SDRT, deux théories majeures sur le discours. Ensuite, nous avons été amenés à définir une nouvelle hiérarchie des relations de discours qui s'inspire de RST, de SDRT et du PDTB.

ABSTRACT

Towards the FDTB : French Discourse Tree Bank

We present the first steps towards creating an annotated corpus for discourse in French : the French Discourse Treebank enriching the FTB. Our methodology is based on the Penn Discourse Treebank (PDTB), but it differs in at least two points of a theoretical nature. First, our goal is to provide full coverage of a text, while the PDTB provides only partial coverage, which can not be described as discourse analysis such as the one made in RST or SDRT, two major theories on discourse. Second, we were led to define a new hierarchy of discourse relations which is based on RST, SDRT and PDTB.

MOTS-CLÉS : Discours, corpus annoté manuellement, analyse discursive, PDTB, RST, SDRT.

KEYWORDS: Discourse, manually annotated corpus, discourse analysis, PDTB, RST, SDRT.

1 Introduction

Dans l'idée de disposer de corpus annotés pour le français, nous avons l'objectif de développer le FDTB (French Discourse Tree Bank), un corpus annoté pour l'analyse discursive. Le FDTB s'inspire du PDTB (Penn Discourse Tree Bank, (PDTB Group, 2008)) qui ajoute une couche d'annotation discursive (manuelle) sur le PTB-v2 (Penn Tree Bank, (Marcus *et al.*, 1999)), corpus anglais tiré du *Wall Street Journal* annoté manuellement pour la morpho-syntaxe. De même, le FDTB ajoute une couche d'annotation discursive sur le FTB (French Tree Bank, (Abeillé *et al.*, 2003)), corpus français tiré du journal *Le Monde* annoté manuellement pour la morpho-syntaxe [1].

Le PDTB est présenté à la Section 2. Si le FDTB s'en inspire, il s'en départit néanmoins sur

1. Le corpus Annodis (Péry Woodley *et al.*, 2009) est aussi un corpus français annoté manuellement pour l'analyse discursive. Le FDTB s'en distingue principalement sur deux points : (i) le fait de disposer d'une analyse syntaxique manuelle du corpus et (ii) de s'appuyer fortement sur les connecteurs de discours (explicites et implicites).

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 471–478,
Grenoble, 4 au 8 juin 2012.

certains choix méthodologiques (Section 3), entre autres, sur le fait que nous voulons une "couverture totale" des articles journalistiques du corpus et non une 'couverture partielle" comme obtenue dans le PDTB. Enfin, nous concluons sur quelques remarques concernant les premières annotations effectuées. Soulignons que celles-ci ne sont que préliminaires, la réalisation d'un corpus comme le FDTB étant un travail de longue haleine qui n'a commencé que depuis six mois (pour l'instant au sein d'une seule équipe).

2 Présentation du PDTB

Le PDTB s'appuie sur trois principes de base qui font consensus dans la communauté travaillant sur le Discours :

(i) Un connecteur de discours est un prédicat sémantique à deux arguments dénotant des "objet abstraits" (notés AO dans (Asher, 1993) qui a introduit cette notion). Pour une annotation discursive, il faut donc identifier dans le texte d'entrée les connecteurs de discours - ceux-ci appartiennent à une liste fermée d'éléments regroupant principalement conjonctions de coordination et de subordination et certains adverbiaux - , vérifier qu'ils sont bien employés comme connecteurs de discours [2], et enfin délimiter les empans de texte correspondant aux arguments de ces connecteurs.

(ii) Un connecteur de discours lexicalise une "relation de discours" ("relation rhétorique") qui appartient à une liste fermée d'éléments organisés dans une hiérarchie arborescente, où les feuilles sont les relations de discours, les nœuds intermédiaires des classes de relations de plus en plus générales en montant vers la racine. Il faut donc pour chaque connecteur de discours indiquer quelle(s) relation(s) de discours il lexicalise en précisant une feuille de la hiérarchie en cas de certitude ou en remontant dans la hiérarchie en cas de doute.

(iii) Les relations de discours ne sont pas forcément lexicalisées par un connecteur de discours (explicite) : elles sont alors inférables [3]. On parle alors de "connecteur implicite" (ou de "connecteur vide Ø"). Pour une annotation discursive, il faut donc identifier les positions où on doit insérer un connecteur implicite, et appliquer aux connecteurs implicites les traitements décrits en (i) et (ii) pour les connecteurs explicites [4].

En plus des annotations venant des trois principes de base décrits ci-dessus, sont annotées dans le PDTB pour chaque connecteur de discours (explicite ou implicite) la source de la relation de discours en jeu et la source de chacun de ses arguments. Ainsi, pour l'exemple (3a) de (Prasad *et al.*, 2006) où le connecteur *while* est souligné, ses argument Arg1 et Arg2 repérés par les segments de texte respectivement en italiques et en gras, le "segment attributif" *purchasing agent said*, placé dans une boîte, se contente d'indiquer que la source de Arg2 n'est pas le locuteur

2. Un même mot ou groupe de mots peut avoir un emploi comme connecteur de discours et un emploi non discursif, comme illustré pour *à ce moment-là* en (1) : en (1a), il s'agit d'un emploi discursif mais pas en (1b), (Roze, 2009).
(1) a. Tu as l'air de penser qu'elle n'est pas honnête. A ce moment-là, ne lui raconte rien.
b. Il a commencé à pleuvoir. Marie est arrivée à ce moment-là.

3. Ainsi en (2a), le locuteur demande à son interlocuteur d'inférer que le contenu propositionnel de la seconde phrase est la cause du contenu propositionnel de la première phrase sans que la relation de discours $Explication$ liant ces deux AO soient explicitement indiqués par un connecteur, comme elle l'est en (2b).
(2) a. Fred n'était pas en forme aujourd'hui. Il a mal dormi la nuit dernière.
b. Fred n'était pas en forme aujourd'hui parce qu'il a mal dormi la nuit dernière.

4. En plus des connecteurs explicites ou implicites, le PDTB fait aussi appel aux notions $AltLex$, $EntRel$, et $NoRel$, mais nous n'avons pas la place de les présenter dans cet article.

mais les purchasing agents. Les annotations basiques du PDTB sont donc complétées par d'autres informations, représentées ici sous forme de tableau donné en (3b). Ce tableau indique la valeur du trait [Source] pour la relation REL (marquée par *while* et identifiée comme étant $Contraste$), pour Arg1 et pour Arg2. La valeur "Wr" est utilisée pour l'auteur ("writer") du texte, "Inh" indique que la valeur de [Source] est héritée de celle de REL, "Ot" ("other") est utilisée pour un (ou des) individu(s) autre(s) que l'auteur (il s'agit des *purchasing agents* pour Arg2). Le tableau comporte d'autres informations relatives à la factivité et la polarité, les traits [Type], [Polarity] et [Determinacy], mais nous ne les détaillerons pas ici [5].

(3) a. *Factory orders and construction outlays were largely flat in December* while purchasing agents said **manufacturing shrank further in October**.

b.

	REL	Arg1	Arg2
[Source]	Wr	Inh	Ot
[Type]	Comm	Null	Comm
[Polarity]	Null	Null	Null
[Determinacy]	Null	Null	Null

Quelques données quantitatives concernant le PDTB issues de (Prasad *et al.*, 2008) : 18459 connecteurs explicites — appartenant à une liste fermée de 100 éléments — et 16224 connecteurs implicites ont été annotés. Les relations de discours sont au nombre de 30 réparties dans quatre classes majeures ($TEMPORAL, CONTINGENCY, COMPARISON$ et $EXPANSION$). Ce travail d'annotation fut mené sur plusieurs années par des chercheurs senior et des doctorants [6]. L'accord inter-annotateur est : pour les relations de discours de 77 % sur les feuilles de la hiérarchie et de 90% sur les quatre classes majeures ; pour les empans des arguments de 90.2 % pour les connecteurs explicites et de 85.1 % pour les implicites. Un outil d'annotation ANNOTATOR a été développé ainsi qu'une interface de visualisation des annotations, qui propose entre autres de voir pour chaque argument de connecteur l'analyse syntaxique proposée dans le PTB.

L'exploitation des annotations du PDTB a donné lieu à multe voix de recherche (en plus des nombreuses statistiques données en annexes du manuel d'annotation (PDTB Group, 2008)), citons le parsing de discours (Lin *et al.*, 2011), la classification des connecteurs (Pitler *et al.*, 2010), et d'autres récapitulées dans (Webber *et al.*, 2011).

Des projets analogues pour d'autres langues (dont le chinois, le turc et l'hindi) sont en cours de développement. Et nous nous attaquons maintenant au français, en effectuant quelques choix différents de ceux du PDTB, non pas tant à cause de différences entre les langues concernées — l'anglais et le français sont relativement très proches — mais pour des raisons méthodologiques ou théoriques [7].

5. Comme expliqué dans (Danlos, 2011), les informations de factivité utilisées dans FactBank (Saurí et Pustejovsky, 2009) sont plus élaborées que celles du PDTB et nous projetons donc d'utiliser les informations de factivité à la FactBank plutôt que celles du PDTB.

6. D'après A. Joshi (pc) l'idée de départ était de confier ce travail à des étudiants undergraduate, comme ce fut le cas pour l'annotation morpho-syntaxique du PTB. Elle fut abandonnée car l'analyse discursive est trop difficile : elle repose en effet beaucoup sur de l'interprétation qui peut déraper dans une subjectivité totale et improductive si elle n'est pas guidée par des principes coercitifs reposant sur des connaissances solides en morpho-syntaxe, sémantique et pragmatique.

7. Cette position ne va pas sans inconvénients, par exemple le fait de ne pas pouvoir comparer directement le FDTB au PDTB qui fait référence. Néanmoins, ces inconvénients sont contre-balancés par le fait qu'il semble nécessaire de faire avancer les connaissances sur l'analyse discursive.

3 Différences entre le FDTB et le PDTB

Nous ne sommes pas d'accord avec tous les choix méthodologiques — et les principes théoriques sous-jacents — du PDTB, et nous nous orientons donc vers d'autres choix décrits ci-dessous.

3.1 Couverture partielle versus totale

Notre objectif dans le FDTB est d'avoir une "couverture totale" du texte annoté (i.e. un article *du Monde*). La couverture totale s'oppose à la "couverture partielle" réalisée dans le PDTB. L'équipe du PDTB s'est fixée comme objectif d'annoter les arguments de certains connecteurs (plus précisément, les connecteurs explicites appartenant à une liste d'une centaine d'éléments, plus certains connecteurs implicites mais pas tous (PDTB Group, 2008)) : leur objectif n'est pas d'obtenir une analyse discursive complète du texte comme celle obtenue en RST (Mann et Thompson, 1987; Taboada et Mann, 2006) ou SDRT (Asher et Lascarides, 2003) où un graphe discursif connexe couvre **tous** les segments du texte (au même titre qu'une analyse syntaxique d'une phrase couvre **tous** les mots de la phrase). Cette position du PDTB est clairement revendiquée (Webber *et al.*, 2011), mais ce n'est pas la nôtre car seule une analyse complète d'un texte permet de rendre compte de sa cohérence et d'en extraire des informations ou de le résumer adéquatement, par exemple.

Une couverture totale demande de se départir du PDTB sur les points suivants :

– Annoter tous les connecteurs explicites et non pas simplement ceux appartenant à une liste de 100 éléments définis comme les plus fréquents [8]. Le lexique LEXCONN répertorie une liste aussi exhaustive que possible des connecteurs du français : il comporte plus de 300 éléments (Roze, 2009; Roze *et al.*, 2010). Le FDTB va annoter tous les connecteurs de LEXCONN — après désambiguation entre emplois discursifs versus non discursifs.

– Annoter plus de connecteurs implicites que dans le PDTB. Le PDTB a fortement restreint les positions où il était licite de poser que deux AO sont reliés par un connecteur vide (PDTB Group, 2008). Par exemple, les connecteurs vides des discours en (4) ne sont pas considérés alors qu'il est clair qu'une relation de discours existe entre les deux AO à gauche et à droite du connecteur Ø. Les positions où un connecteur vide doit être inséré demandent une étude linguistique qui débouche sur une insertion automatique du connecteur Ø demandant le moins possible de révision manuelle (Antolinos-Basso, 2012). Signalons que dans le PDTB, il est demandé aux annotateurs de préciser le sens d'un connecteur implicite en insérant un (ou deux) connecteur(s) explicite(s) ainsi qu'en indiquant la ou les relation(s) de discours exprimée(s). Il semble inutile d'insérer des connecteurs explicites à la place des connecteurs vides : on peut simplement demander aux annotateurs de préciser la ou les relation(s) de discours exprimée(s) (après avoir éventuellement testé si la présence de tel ou tel connecteur explicite n'altère pas la sémantique du texte).

(4) a. Fred a tiré sur Marie, Ø la tuant.
b. Fred a tué Marie, Ø en lui tirant dessus.
c. Fred a fait la vaisselle, Ø passé l'aspirateur, et lavé les carreaux.

8. Nous ne disposons pas pour le français de la liste des 100 connecteurs les plus fréquents. Certes, on peut effectuer un simple comptage des occurrences des (suites de) mots correspondant aux connecteurs, mais ce comptage ne prendrait pas en compte qu'une même suite de mots peut correspondre à un emploi discursif ou non, voir *à ce moment là* dans (1). Pour déterminer les 100 connecteurs les plus fréquents du français, il faut disposer d'annotations comme celle du FDTB.

– Supprimer les notions de Sup1 et Sup2. Citons (Joshi *et al.*, 2006) qui arguent pour un "principe de minimalité" : "Only as many clauses and/or sentences should be included as are minimally required for interpreting the relation. Any other span of text that is perceived to be relevant (but not necessary) should be annotated as supplementary information : Sup1 for material supplementary to Arg1, Sup2 for material supplementary to Arg2." Ce principe de minimalité ne peut donner lieu qu'à des interprétations plutôt subjectives, dont la subjectivité est cependant limitée par des consignes relevant de la syntaxe, par exemple : "An argument includes any non-clausal adjuncts, prepositions, connectives, or complementizers introducing or modifying the clause" (Joshi *et al.*, 2006) [9]. La question suivante se pose : quel est le rôle de Sup1 ou Sup2 par exemple dans l'exploitation du PDTB pour les techniques d'apprentissage supervisé reposant sur les données annotées ? Nous n'avons trouvé aucune information sur la question dans les nombreux articles exploitant les résultats du PDTB (Section 2) et nous subodorons que les segments annotés Sup1 et Sup2 sont tout bonnement et simplement ignorés. De plus, Sup1 et Sup2 ne sont pas pris en compte dans l'accord inter-annotateurs sur les empans des arguments des connecteurs (Prasad *et al.*, 2008). Il semble donc que les notions de Sup1 et Sup2 n'ont été utilisées que pour le confort des annotateurs. C'est une raison valable dans l'objectif d'une couverture partielle que s'est fixé le PDTB, mais qui ne trouve guère de justification dans l'objectif d'une couverture totale où les segments de texte qui auraient été identifiés comme Sup1 ou Sup2 doivent être intégrés dans l'analyse discursive totale du texte.

3.2 Hiérarchie des relations des connecteurs

La hiérarchie des relations de connecteurs établie dans (PDTB Group, 2008, page 27) est discutable, ne serait-ce que parce qu'elle ne semble pas avoir été établie sur la base des relations de discours (et de leur organisation en classes) définie en RST (Mann et Thompson, 1987; Taboada et Mann, 2006) ou SDRT (Asher et Lascarides, 2003), les deux grandes théories sur le discours. Or ces théories présentent des classifications intéressantes sur des dimensions orthogonales :

– RST distingue les relations de discours 'informationnelle versus présentationnelle" en s'appuyant sur le fait que certaines relations n'établissent qu'une relation sémantique (informationnelle) entre les contenus propositionnels des ses arguments tandis que d'autres (présentationnnelles) demandent de faire appel à des actes de parole. Ainsi la relation de discours lexicalisée par *donc* est informationnelle en (5a), tandis qu'elle est présentationnelle en (5b) puisqu'elle sous-entend l'acte de parole *J'en déduis (donc qu'il a plu)*.

(5) a. Il a plu. Donc, les routes sont mouillées.
b. Les routes sont mouillées. Donc, il a plu.

– SDRT distingue les relations véridicales des relations non véridicales. La notion de véridicalité y est définie par la formule $R(\alpha, \beta) \Rightarrow K_\alpha \wedge K_\beta$, qui se glose ainsi : si $R(\alpha, \beta)$ est vraie alors les contenus propositionnels des arguments, notés K_α et K_β, sont vrais. Comme montré dans (Danlos et Rambow, 2011), cette définition ne tient pas dès qu'on quitte les assertions de l'auteur, voir (6) avec $\textsc{Narration}(\alpha, \beta)$ où NARRATION, qualifiée de véridicale en SDRT, n'implique pas que K_β soit vrai.

9. Nous sommes actuellement en train de rédiger des consignes analogues pour le français, avec à nouveau des différences par rapport au PDTB. Ainsi, contrairement au PDTB, nous excluons tout connecteur des empans pour les arguments Arg1 et Arg2 d'un connecteur, ce qui va de pair avec l'objectif d'obtenir une couverture totale.

(6) [Fred ira à Dax pour Noël.]$_\alpha$ Jane pense qu'*ensuite* [il ira à Pau.]$_\beta$

Il n'empêche que la définition de véridicalité posée en SDRT tient quand on considère que α et β sont des assertions de l'auteur, et qu'il semble pertinent de distinguer les relations véridicales (e.g. $Narration$) des autres (e.g. $Condition$ illustrée dans *[Marie viendra à ma fête]$_\alpha$ si [Fred vient]$_\beta$).*)

Nous avons donc défini une nouvelle hiérarchie des relations de discours qui s'appuie sur les classifications orthogonales des relations de discours définies en RST et SDRT et sur celle du PDTB. Cette hiérarchie, donnée en appendice et décrite dans (Danlos et Roze, 2011), distingue cinq classes de relations véridicales (*CAUSALE*, *TEMPORELLE*, *COMPARAISON*, *EXPANSION* et *ADDITIVE*) et une classe de relations non-véridicales. Dans chaque classe, des sous-classes distinguent éventuellement les relations informationnelles/sémantiques (e.g RÉSULTAT) des relations présentationnelles/pragmatiques (distinguées par le signe * e.g RÉSULTAT*). Cette hiérarchie contient 31 feuilles auxquelles il faut ajouter la relation UNKNOWN qui est utile pour certains connecteurs dont le sens est très spécifique (e.g. *au fur et à mesure que* qui s'emploie dans un énoncé comme *Son débit de parole augmentait au fur et à mesure qu'il vidait la bouteille*) ou lorsqu'un annotateur n'est satisfait par aucune relation de la hiérarchie [10].

4 Premières expériences et perspectives futures

Pour nous familiariser avec l'outil d'annotation — ANNOTATOR adapté pour les besoins du FDTB —, identifier les problèmes récurrents d'annotation en vue de rédiger un manuel d'annotation complet et précis (voir note 6), nous avons mené quelques expériences sur certains connecteurs explicites. Ces premières expériences ont montré une grande divergence dans ce qui est annoté Sup1 ou Sup2, ce qui nous conforte dans l'idée d'abandonner tout bonnement et simplement ces notions. Elles ont aussi montré qu'il n'était pas adéquat de travailler connecteur par connecteur. Dans la suite du développement du FDTB, nous demanderons aux annotateurs de travailler article par article, en annotant les connecteurs explicites et implicites avec l'objectif d'une analyse discursive totale.

Nous invitons toute personne intéressée par les connecteurs de discours et/ou l'analyse discursive à nous contacter afin de participer à cette entreprise de longue haleine aux résultats attendus.

Remerciements

Nous remercions M. Djemaa, C. Dordonne, A. Hu, M. Majdoub, L. Patris, C. Ribeyre, F. Sari, et Y. Sun, étudiants de Master2 dans le cursus de Linguistique Informatique de l'Université Paris Diderot, qui ont servi de cobayes (en binôme) pour une annotation expérimentale des connecteurs *afin de, lorsque, puisque* et *en revanche*. Le FTB comporte aux alentours de 80 occurrences de chacun de ces connecteurs.

10. Un annotateur peut choisir deux éléments de la hiérarchie (mais pas plus que deux, est-ce une limite ?) pour un même connecteur (explicite ou implicite). Par exemple, pour *alors que* en (7) il peut choisir OVERLAP **et** CONTRASTE sans avoir à trancher arbitrairement entre les deux.

(7) En ce moment, Fred est complètement déprimé, *alors que* Marie est en pleine forme.

Références

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for french. *In* Abeillé, A., éditeur : *Treebanks*. Kluwer Academic Publishers, Dordretch.

Antolinos-Basso, D. (2012). Les connecteurs implicites dans le FDTB. Mémoire de Master, Université Paris Diderot.

Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Kluwer, Dordrecht.

Asher, N. et Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press, Cambridge.

Danlos, L. (2011). Analyse discursive et informations de factivité. *In Actes de TALN 2011*, Montpellier, France.

Danlos, L. et Rambow, O. (2011). Veridicality of discourse relations and factuality information. *In Proceedings of the fourth workshop on Constraints in Discourse (CID 2011)*, Agay, France.

Danlos, L. et Roze, C. (2011). Hiérarchie des relations de discours dans le FDTB. Rapport technique, ALPAGE, Université Paris Diderot.

Joshi, A., Prasad, R. et Webber, B. (2006). Discourse annotation : Discourse connectives and discourse relations. *In Tutorial at the Association for Computational Linguistics*, Sydney, Australia.

Lin, Z., Ng, H. T. et Kan, M.-Y. (2011). Automatically evaluating text coherence using discourse relations. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (ACL-HLT)*, Portland, OR.

Mann, W. et Thompson, S. (1987). Rhetorical structure theory. *In* Kempen, G., éditeur : *Natural Language Generation*, pages 85–95. Martinus Nijhoff Publisher, Dordrecht.

Marcus, M., Santorini, B., Marcinkiewicz, M. A. et Taylor, A. (1999). Building a treebank for french. *In Treebank-3*. Linguistic Data Consortium, Philadelphie.

PDTB Group (2008). The Penn Discourse Treebank 2.0 annotation manual. Rapport technique, Institute for Research in Cognitive Science, University of Philadelphia.

Péry Woodley, M.-P., Asher, N., Enjalbert, P., Benamara, F., Bras, M., Fabre, C., Ferrari, S., Ho Dac, L.-M., Le Draoulec, A., Mathet, Y., Muller, P., Prévot, L., Rebeyrolle, J., Tanguy, L., Vergez Couret, M., Vieu, L. et Widlöcher, A. (2009). ANNODIS : une approche outillée de l'annotation de structures discursives. *In Proceedings of TALN 2009*, pages 190–196, Senlis, France.

Pitler, E., Louis, A. et Nenkova, A. (2010). Automatic sense prediction for implicit discourse relations in text. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*, Suntec, Singapore.

Prasad, R., Dinesh, N., Lee, A., Joshi, A. et Webber, B. (2006). Attribution and its annotation in the Penn Discourse Treebank. *Revue TAL*, 47(2).

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. et Webber, B. (2008). The penn discourse treebank 2.0. *In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Roze, C. (2009). LEXCONN : Base lexicale des connecteurs discursifs du français. Mémoire de Master, Université Paris Diderot.

Roze, C., Danlos, L. et Muller, P. (2010). LEXCONN : a French lexicon of discourse connectives. *In Proceedings of Multidisciplinary Approaches to Discourse (MAD 2010)*, Moissac, France.

Saurí, R. et Pustejovsky, J. (2009). FactBank : A corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268.

Taboada, M. et Mann, W. (2006). Rhetorical Structure Theory : Looking back and moving ahead. *Discourse Studies*, 8(3):423–459.

Webber, B., Egg, M. et Kordoni, V. (2011). Discourse structure and language technology. *Natural Language Engineering*, 1(1):1–54.

Hiérachie des relations de discours

Classe CAUSALE
- — Type sémantique
  - — Résultat
  - — Explication
  - — Moyen
  - — But
- — Type pragmatique
  - — Résultat*
  - — Explication*
  - — But*
- — Type sémantico-pragmatique
  - — Evidence

Classe TEMPORELLE
- — Type sémantique
  - — Overlap
  - — Précédence
  - — Succession

Classe COMPARAISON
- — Polarités opposées
  - — Contraste
  - — Concession
    - — Concession-avant
    - — Concession-arrière
- — Polarités égales
  - — Parallèle
- — Type pragmatique
  - — Contraste*

Classe EXPANSION
- — Type Elaboration
  - — Elab-entité
  - — Elab-événement
  - — Elab-instance
  - — Exception
- — Résumé
- — Evaluation

Classe ADDITIVE
- — Type Liste
  - — Enumeration
  - — Continuation
- — Disgression

Classe NON VERIDICALE
- — Alternatives
  - — Alternative
  - — Alternative*
- — Correction
- — Sémantiques Conditionnelles
  - — Unless
  - — Otherwise
  - — Condition
  - — Conséquence-hypothétique
- — Pragmatiques Conditionnelles
  - — Condition*

Combinaison de ressources générales pour une contextualisation implicite de requêtes

Romain Deveaud[1] Patrice Bellot[2]
(1) LIA - Université d'Avignon
romain.deveaud@univ-avignon.fr
(2) LSIS - Université Aix-Marseille
patrice.bellot@lsis.org

RÉSUMÉ

L'utilisation de sources externes d'informations pour la recherche documentaire a été considérablement étudiée dans le passé. Des améliorations de performances ont été mises en lumière avec des corpus larges ou structurés. Néanmoins, dans ces études les ressources sont souvent utilisées séparément mais rarement combinées. Nous présentons une évaluation de la combinaison de quatre différentes ressources générales, standards et accessibles. Nous utilisons une mesure de distance informative pour extraire les caractéristiques contextuelles des différentes ressources et améliorer la représentation de la requête. Cette évaluation est menée sur une tâche de recherche d'information sur le Web en utilisant le corpus ClueWeb09 et les *topics* de la piste Web de TREC. Les meilleurs résultats sont obtenus en combinant les quatre ressources, et sont statistiquement significativement supérieurs aux autres approches.

ABSTRACT

Query Contextualization and Reformulation by Combining External Corpora

Improving document retrieval using external sources of information has been extensively studied throughout the past. Improvements with either structured or large corpora have been reported. However, in these studies resources are often used separately and rarely combined together. We present an evaluation of the combination of four different scalable corpora over a web search task. An informative divergence measure is used to extract contextual features from the corpora and improve query representation. We use the ClueWeb09 collection along with TREC's Web Track topics for the purpose of our evaluation. Best results are achieved when combining all four corpora, and are significantly better than the results of other approaches.

MOTS-CLÉS : Combinaison de ressources, RI contextuelle, recherche web.

KEYWORDS: Resources combination, contextual IR, web search.

1 Introduction

La recherche d'information a pour but de satisfaire le besoin d'information d'un utilisateur. En effet, lorsqu'un utilisateur effectue une recherche dans une base documentaire, il fournit au système une représentation de son besoin d'information. Le rôle du système est alors de prendre en compte cette représentation et de présenter à l'utilisateur un ensemble de documents pertinents par rapport au besoin d'information initial. Ces documents sont généralement présentés

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 479–486,
Grenoble, 4 au 8 juin 2012.

sous une forme de liste et ordonnés par ordre décroissant de pertinence. Il existe des modèles de recherche d'information qui permettent de récupérer efficacement des documents par rapport à une requête, qui joue le rôle de la représentation d'un besoin d'information. La difficulté réside donc dans la capacité de l'utilisateur à représenter son besoin d'information d'une façon adéquate pour le système. Seulement, les requêtes formulées par ces utilisateurs ne décrivent pas toujours parfaitement ce besoin, et des connaissances additionnelles sont parfois nécessaires pour compléter cette description manquante. Une des manières de mieux définir le sujet d'une recherche est d'enrichir la requête originale avec des informations supplémentaires. Celles-ci consistent traditionnellement en des mots que l'on va ajouter à la requête formée par l'utilisateur. Typiquement, ces mots sont extraits de documents récupérés en utilisant la requête initiale. Les documents peuvent provenir de la collection cible (la base de documents au sein de laquelle le système effectue la recherche) (Harman, 1992) ou de collections externes.

Les collections externes utilisées peuvent être de types très différents. Elles peuvent être générales ou spécifiques à un domaine précis, structurées ou non, ou encore construites automatiquement ou manuellement. L'utilisation de ressources externes a été considérablement étudiée dans le passé, et elle a prouvé son efficacité à améliorer les performances des systèmes de recherche d'information lorsqu'ils choisissent les données appropriées. Ces études se concentrent principalement sur la manière dont une ressource individuelle peut améliorer les performances d'un système de recherche d'information, mais proposent rarement d'utiliser ces ressources conjointement. Des sources de données telles que Wikipédia (Li *et al.*, 2007; Suchanek *et al.*, 2007), WordNet (Liu *et al.*, 2004; Suchanek *et al.*, 2007; Fang, 2008), des articles journalistiques ou encore le web lui-même (Diaz et Metzler, 2006) ont été utilisées. Dans leur étude, (Diaz et Metzler, 2006) expérimentent l'utilisation de ressources externes larges et générales. Ils présentent un modèle qui permet d'incorporer des données additionnelles à la façon d'un retour de pertinence simulé (Lavrenko et Croft, 2001), et ils l'évaluent en considérant un corpus d'actualité et deux corpus de pages web comme ressources externes. Ils démontrent que chaque ressource améliore les performances du système de recherche d'information indépendamment, mais ils ne reportent pas d'expériences sur une combinaison de ces ressources. D'un autre côté, (Mandala *et al.*, 1999) présentent dans leur travail une méthode d'enrichissement qui combine des caractéristiques extraites de WordNet et de deux thesaurus spécifiques créés à partir de la collection de documents. Le premier a pour but d'identifier les relations sémantiques entre deux mots en calculant ses co-occurrences. Le second se concentre sur la pondération de paires de mots liés par leur relation syntaxique. Cette étude est une des seules qui rapporte des améliorations de performance en combinant plusieurs ressources.

Dans cet article, nous évaluons les performances d'un système de recherche pouvant combiner un nombre quelconque de ressources externes. Cette évaluation est menée sur une tâche de recherche de pages web et nous utilisons pour cela la collection ClueWeb09, qui est à ce jour la représentation statique la plus complète du web. Les requêtes utilisateurs et les jugements de pertinence proviennent de la piste Web de TREC (Clarke *et al.*, 2009).

Nous commençons par détailler le modèle de recherche d'information que nous utilisons dans la section 2, puis nous présentons notre approche de combinaison de ressources dans la section 3. La section 4 présente une évaluation étendue ainsi qu'une discussion sur les résultats obtenus et des perspectives sur nos travaux futurs.

2 Modèles de langue pour la recherche d'information

Nous avons choisi de suivre une approche par modèle de langue pour la recherche d'information et nous rappelons ici les principes du modèle état-de-l'art que nous utilisons. Plusieurs travaux ont en effet démontré l'efficacité de ce modèle à intégrer des informations provenant de différentes sources, qu'elles soient intra-collection ou extra-collection (Diaz et Metzler, 2006).

Le modèle de dépendance séquentielle (ou Sequential Dependence Model, SDM) est un cas particulier du modèle MRF (Markov Random Field) pour la recherche d'information. Il a été introduit par Metzler et Croft (Metzler et Croft, 2005) et a montré des performances état-de-l'art concernant plusieurs contextes de recherche dont celui sur le web (Allan *et al.*, 2008; Metzler *et al.*, 2006). Ce modèle n'agit que sur les mots de la requête et consiste à modéliser les dépendances entre les mots adjacents. Suivant le modèle SDM, la fonction calculant le poids d'un mot de la requête q dans un document D est donnée par l'équation :

$$f_T(q,D) = \log\left[\frac{c(q,D)+\mu\cdot\frac{c(q,\mathscr{C})}{|\mathscr{C}|}}{|D|+\mu}\right]$$

avec $c(q,\mathscr{C})$ the nombre d'occurrences du mot de la requête q dans la collection cible $\mathscr{C}$, $|\mathscr{C}|$ la taille de la collection et $|D|$ la taille du document D. μ est le paramètre du lissage de Dirichlet, nous fixons sa valeur à 2500 comme le recommande (Zhai et Lafferty, 2004) pour les requêtes constituées de mots-clés. C'est l'estimation par maximum de vraisemblance de l'unité lexicale q dans le document D.

Le modèle propose deux fonctions supplémentaires pour deux autres types de dépendances qui agissent sur les bigrammes de la requête. La fonction $f_O(q_i,q_{i+1},D)$ considère la correspondance exacte de deux mots de la requête adjacents. Elle est dénotée par l'indice O. La seconde, $f_U(q_i,q_{i+1},D)$, est dénotée par l'indice U et considère la correspondance non ordonnée de deux mots au sein d'une fenêtre de 8 unités lexicales. Finalement, le score d'appariement requête-document qui utilise les fonctions ci-dessus définies par le modèle de dépendance séquentielle revient à :

$$score_{SDM}(Q,D) = \lambda_T \sum_{q\in Q} f_T(q,D) + \lambda_O \sum_{i=1}^{|Q|-1} f_O(q_i,q_{i+1},D) + \lambda_U \sum_{i=1}^{|Q|-1} f_U(q_i,q_{i+1},D) \qquad (1)$$

où λ_T, λ_O et λ_U sont des paramètres libres. Dans nos expériences nous fixons ces paramètres en suivant les recommandations des auteurs ($\lambda_T = 0,85$, $\lambda_O = 0,10$ et $\lambda_U = 0,05$). Plus loin, nous nous réfèrerons à la fonction de score définie par l'équation (1) par l'acronyme SDM.

3 Combinaison de ressources générales

Certains besoins en information sont parfois trop complexes pour être représentés par des requêtes constituées d'un petit nombre de mots. De plus, le processus de création de la requête peut nécessiter un effort cognitif de la part de l'utilisateur, et mettre en jeu des connaissances qu'il ne possède pas ou qu'il souhaite acquérir au terme de sa recherche. L'utilisation de ressources externes permet de pallier ces manques, dans un contexte de recherche connu uniquement de

l'utilisateur. Ce contexte peut être interprété en utilisant la masse de connaissances contenue dans ces ressources, mais il faut pour cela se réduire à des sous-ensembles contenant uniquement des informations contextuelles par rapport à la requête.

Considérant une ressource $\mathscr{R}$, nous formons un sous-ensemble contextuel $\mathscr{R}_Q$ à partir des N premiers documents renvoyés par le modèle SDM pour une requête Q. On peut alors calculer la distance informative entre le modèle de langue $\theta_{\mathscr{R}_Q}$ du sous-ensemble contextuel et le modèle de langue θ_D de chaque document D de la collection cible. Cette distance agît naturellement comme un processus de contextualisation : plus la distance entre les deux modèles de langue est importante, moins le document D est lié au contexte de recherche latent de la requête Q. Dans ce travail nous utilisons la divergence de Kullback-Leibler, ce qui nous permet de mesurer à quel point une ressource et un document donné sont proches. Formellement, la divergence de KL entre le modèle de langue $\theta_{\mathscr{R}_Q}$ d'un sous-ensemble contextuel $\mathscr{R}_Q$ et le modèle de langue θ_D d'un document D s'exprime par :

$$\begin{aligned} KL(\theta_{\mathscr{R}_Q}||\theta_D) &= \sum_{w\in V} p(w|\theta_{\mathscr{R}_Q}) \log \frac{p(w|\theta_{\mathscr{R}_Q})}{p(w|\theta_D)} \\ &= \sum_{w\in V} p(w|\theta_{\mathscr{R}_Q}) \log p(w|\theta_{\mathscr{R}_Q}) - \sum_{w\in V} p(w|\theta_{\mathscr{R}_Q}) \log p(w|\theta_D) \\ &\propto -\sum_{w\in V} p(w|\theta_{\mathscr{R}_Q}) \log p(w|\theta_D) \end{aligned} \tag{2}$$

La dernière simplification de l'équation ci-dessus peut être réalisée car son premier membre est l'entropie de la ressource et n'affecte pas le classement des documents.

Ici la contextualisation est effectuée à partir des informations provenant d'une unique source externe d'information, mais cette source peut-être incomplète ou imprécise pour certains sujets. Nous choisissons donc de combiner les connaissances de plusieurs ressources différentes en calculant toutes les divergences possibles. Ainsi, le contexte de la requête peut être interprété d'autant de manières qu'il y a de ressources et gagner en précision. Formellement, le score d'un document D par rapport à une requête Q est donné par :

$$score(Q,D) = SDM(Q,D) - \frac{1}{|\mathscr{S}|} \sum_{\mathscr{R}_Q \in \mathscr{S}} KL(\theta_{\mathscr{R}_Q}||\theta_D) \tag{3}$$

où $\mathscr{S}$ est un ensemble de ressources. Nous utilisons ici le score de la divergence de KL pour dégrader un document ; en effet, plus la distance est importante, plus le score du document va être réduit. Ainsi, la combinaison de plusieurs ressource agît intuitivement comme une généralisation du contexte de recherche : plus le nombre de ressources utilisées augmente, meilleure est la représentation contextuelle du besoin d'information. Il est à noter que le modèle de recherche d'information ainsi obtenu est très proche d'une précédente méthode qui avait montré son efficacité dans le cadre d'une recherche de passages précis en utilisant Wikipédia comme ressource externe (Deveaud *et al.*, 2011).

4 Evaluation et résultats

Nous évaluons notre approche en utilisant le corpus ClueWeb09 [1], qui est à ce jour la plus grande collection de test mise à disposition de la communauté de recherche d'information. Ce corpus a servi de support à de nombreuses tâches de TREC comme la Web Track, Blog Track, Million Query Track... Nous ne considérons ici que la catégorie B du ClueWeb09, constituée d'environ 50 millions de pages web. Nous utilisons pour notre évaluation la catégorie B ainsi que les *topics* et les jugements de pertinence officiels mis à disposition des participants de la Web Track.

Concernant les ressources utilisées, nous avons souhaité modéliser plusieurs contextes de recherche fréquemment rencontrés sur le web, tels que la recherche de connaissances ou d'actualités. Nous avons donc choisi Wikipédia comme source encyclopédique, le New York Times ainsi que le corpus GigaWord comme source journalistiques et un sous-ensemble du ClueWeb09 composé uniquement de pages non spammées comme source web. Le corpus GigaWord anglais de LDC [2] est constitué de dépêches journalistiques provenant de quatre sources d'actualités distinctes, dont le New York Times. Le corpus New York Times de LDC [3] comprend quant à lui des articles publiés dans ce journal entre 1987 et 2007. La ressource Web est issue de la catégorie B du ClueWeb09 à laquelle nous avons soustrait toutes les pages web considérées comme spam. Nous utilisons pour cela l'ensemble "Fusion" de scores de spam pour le ClueWeb09 distribué par (Cormack *et al.*, 2010) [4]. Cette liste attribue à chaque document un score qui représente le pourcentage de documents de la collection qui sont plus spammés que lui. Ainsi, plus le score est grand, moins la probabilité que le document soit un spam est importante. Pour la construction de notre ressource, nous n'avons conservé que les documents dont le score est supérieur à 70, comme le préconisent les auteurs (Cormack *et al.*, 2010). Pour finir, notre corpus Wikipédia contient tous les articles anglais contenu dans l'encyclopédie en ligne au mois de juillet 2011 [5].

Ressource	Type	Nb documents	Nb mots uniques	Nb mots total
GigaWord (GW)	Journalistique (dépêches)	4 111 240	1 288 389	1 397 727 483
New York Times (NYT)	Journalistique (articles)	1 855 658	1 086 233	1 378 897 246
Wikipédia (Wiki)	Encyclopédique	3 214 014	7 022 226	1 033 787 926
ClueWeb09 non spammé (Web)	Web	29 038 220	33 314 740	22 814 465 842

TABLE 1: Récapitulatif des ressources utilisées.

Les processus d'indexation et de recherche de documents sont réalisés en utilisant le moteur de recherche Indri [6]. La liste de mots-outils employée est celle fournie par défaut avec Indri, elle comporte 417 mots communs en langue anglaise. Pour la racinisation nous utilisons l'implémentation d'Indri du raciniseur standard de Krovetz. Nous avons indexé le corpus ClueWeb09 ainsi que les trois ressources externes en utilisant chaque fois ces mêmes paramètres. Lors de la recherche de documents nous résolvons le problème des probabilités nulles avec un lissage de Dirichlet, pour lequel nous fixons le paramètre μ à 2500. Cette méthode de lissage est en effet recommandée lors des recherches par mots-clés (Zhai et Lafferty, 2004), ce qui est notre cas avec les requêtes de TREC. Les documents sont ordonnés en utilisant la formule donnée dans l'équation (3). Nous

1. http://boston.lti.cs.cmu.edu/clueweb09/wiki/
2. http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2003T05
3. http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2008T19
4. http://plg.uwaterloo.ca/~gvcormac/clueweb09spam/
5. http://dumps.wikimedia.org/enwiki/20110722/
6. http://lemurproject.org/

comparons les performances de l'approche d'enrichissement contextuel que nous proposons avec celles de deux systèmes de base. Le premier est le modèle de dépendance séquentielle (SDM) introduit dans la section 2, et le second est le traditionnel retour de pertinence simulé (ou Pseudo-Relevance Feedback, PRF) (Lavrenko et Croft, 2001) avec $\lambda = 0,5$. Dans cette évaluation, nous utilisons les *topics* 1 à 50 de la Web Track de TREC. Nous considérons les 10 premiers documents renvoyés par une requête SDM pour chaque ressource externe $\mathscr{R}$. Nous calculons alors les probabilités $p(w|\theta_{\mathscr{R}})$ et nous reformulons la requête originale en lui ajoutant les 20 mots possédant les meilleurs probabilités d'apparition dans la ressource. Ces mots ajoutés sont également pondérés par la probabilité précédemment calculée afin de refléter leur informativité au sein de la ressource. Nous nous servons de la nouvelle requête ainsi formée pour classer les documents de la catégorie B du ClueWeb09. Pour chaque requête, nous renvoyons jusqu'à 1000 documents. Nous reportons les résultats de ces expériences en terme de gain cumulé à 10 documents (nDCG@10), de précision moyenne (MAP) et de précision à 10 documents dans le tableau 2.

Ressource	nDCG@10	P@10	MAP
Aucune	0,2746	0,3714	0,1837
PRF	0,2486	0,3667	0,2147*
GW	0,2974	0,4014	0,1834
Wiki	0,2996	0,4255	0,2298*
Web	0,3014	0,4480*	0,2369*
NYT	0,3071	0,4395*	0,2118*
Web + NYT	0,3004	0,4195	0,2257*
Wiki + GW	0,3034*	0,4253	0,2298**
Web + Wiki	0,3088*	0,4521*	0,2374**
NYT + GW	0,3114	0,4405*	0,2075*
Wiki + NYT	0,3119	0,4500*	0,2329**
Web + GW	0,3120*	0,4318*	0,2241*
Wiki + NYT + GW	0,3067*	0,4366*	0,2320**
Web + NYT + GW	0,3100*	0,4359*	0,2205**
Web + Wiki + GW	0,3202**	0,4563*	0,2331**
Web + Wiki + NYT	0,3246***	0,4563**	**0,2395*****
Web + Wiki + NYT + GW	**0,3268*****	**0,4665****	0,2353***

TABLE 2: Résultats sur la catégorie B du ClueWeb09 pour les *topics* 1 à 50 de la Web Track de TREC. Evaluation des combinaisons de Wikipédia (Wiki), le New York Times (NYT), le GigaWord (GW) et le ClueWeb09 non spammé (Web) comme ressources externes. Nous utilisons le test apparié de Student ($^{*} : p < 0,1$; $^{**} : p < 0,05$; $^{***} : p < 0,01$) pour déterminer les différences statistiquement significatives avec le système de base.

L'observation principale que l'on peut faire est que la combinaison des quatre ressources est quasiment tout le temps plus performante que toutes les autres combinaisons, à l'exception de la mesure MAP. Contrairement aux autres, cette combinaison complète tire parti de chaque ressource individuellement, et les améliorations observées sont toujours très statistiquement significatives pour toutes les métriques. Il est d'ailleurs intéressant de voir que certaines combinaisons de 3 ressources (Wiki+NYT+GW par exemple) obtiennent des résultats inférieurs à certaines

combinaisons de 2 ressources (NYT+GW par exemple), mais où les performances sont plus significatives. On observe le même comportement entre les combinaisons de 2 ressources et les ressources seules. La combinaison de plusieurs ressources apporte donc une certaine stabilité au modèle de RI, tout en augmentant substantiellement les résultats.

Nous observons également que les résultats décroissent uniformément lorsque l'on baisse le nombre de ressources utilisées dans les combinaisons. Il est intéressant de voir que le corpus NYT utilisé seul améliore significativement les performances de recherche par rapport au corpus GigaWord seul (t-test p-value : 0,081 pour la mesure MAP). En effet le GigaWord contient des dépêches provenant du NYT, on pourrait donc instinctivement penser que leurs performances pourraient être comparables. La principale différence réside dans le fait que les articles du NYT ont été écrits par des journalistes utilisant un vocabulaire spécialisé et augmenté, contrairement aux dépêches qui sont très courtes et factuelles. De plus, le corpus GigaWord est deux fois plus gros en nombre de documents que le NYT, mais les dépêches sont très courtes (340 mots par dépêche en moyenne, contre 743 mots par article NYT en moyenne) et ont pour but d'être directes. De plus le vocabulaire employé est bien plus varié dans les articles du NYT. Ainsi, le grand nombre de documents contenus par le corpus GigaWord n'arrive pas à contrebalancer la qualité d'écriture et la complétude du NYT.

Nous avons également expérimenté différentes valeurs de lissage, différentes pondérations entre les ressources et nous avons fait varier le nombre de pages sélectionnées pour chacune des ressources. Les performances observées étaient comparables à celles reportées dans cette étude, notre système peut se passer d'une étape de paramétrage ou d'apprentissage.

5 Conclusions

Nous avons présenté dans cet article une approche permettant de contextualiser implicitement une requête utilisateur à l'aide de plusieurs ressources externes. Cette approche permet de pénaliser les documents qui sont trop éloignés d'un ensemble de ressources en calculant une distance entre les distributions de mots dans le document et dans ces ressources. Les résultats de nos expérimentations montrent qu'une combinaison de toutes les ressources étudiées permet d'améliorer substantiellement et très significativement les performances d'un système de recherche d'information état-de-l'art.

Nous avons également noté que la qualité d'écriture des ressources est essentielle. Ainsi, choisir une ressource complète et correctement écrite semble plus important que choisir une ressource de grande taille sans considérer son contenu textuel. Nous prévoyons d'étendre cette étude avec un plus grand nombre de ressources et d'autres méthodes de contextualisation, ainsi que plusieurs modèles de recherche d'information. En effet nous pouvons imaginer employer n'importe quel modèle probabiliste qui pourrait s'interpoler avec une combinaison de ressources. Nous planifions également de traduire les requêtes et d'adapter les jugements de pertinence afin de valider ces expériences sur d'autres langues que l'anglais.

Remerciements Ces recherches ont bénéficié du soutien financier de l'Agence Nationale de la Recherche (ANR 2010 CORD 001 02) en faveur du projet CAAS.

Références

ALLAN, J., CARTERETTE, B., ASLAM, J. A., PAVLU, V. et KANOULAS, E. (2008). Million Query Track 2008 Overview. *In Proceedings of the Seventeenth Text REtrieval Conference (TREC)*.

CLARKE, C. L. A., CRASWELL, N. et SOBOROFF, I. (2009). Overview of the TREC 2009 Web Track. *In Proceedings of the Eighteenth Text REtrieval Conference (TREC)*.

CORMACK, G. V., SMUCKER, M. D. et CLARKE, C. L. A. (2010). Efficient and Effective Spam Filtering and Re-ranking for Large Web Datasets. *CoRR*, abs/1004.5168.

DEVEAUD, R., SANJUAN, E. et BELLOT, P. (2011). Ajout d'informations contextuelles issues de Wikipédia pour la recherche de passages. *In Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles*, TALN 2011.

DIAZ, F. et METZLER, D. (2006). Improving the estimation of relevance models using large external corpora. *In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 154–161.

FANG, H. (2008). A Re-examination of Query Expansion Using Lexical Resources. *In Proceedings of ACL-08 : HLT*, pages 139–147, Columbus, Ohio. Association for Computational Linguistics.

HARMAN, D. (1992). Relevance feedback revisited. *In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '92, pages 1–10.

LAVRENKO, V. et CROFT, W. B. (2001). Relevance based language models. *In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 120–127.

LI, Y., LUK, W. P. R., HO, K. S. E. et CHUNG, F. L. K. (2007). Improving weak ad-hoc queries using wikipedia as external corpus. *In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 797–798.

LIU, S., LIU, F., YU, C. et MENG, W. (2004). An effective approach to document retrieval via utilizing WordNet and recognizing phrases. *In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 266–272.

MANDALA, R., TOKUNAGA, T. et TANAKA, H. (1999). Combining multiple evidence from different types of thesaurus for query expansion. *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 191–197.

METZLER, D. et CROFT, W. B. (2005). A markov random field model for term dependencies. *In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 472–479.

METZLER, D., STROHMAN, T. et CROFT, B. W. (2006). Indri at TREC 2006 : Lessons Learned From Three Terabyte Tracks. *In Proceedings of the Fifteenth Text REtrieval Conference (TREC)*.

SUCHANEK, F. M., KASNECI, G. et WEIKUM, G. (2007). Yago : a core of semantic knowledge. *In Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 697–706.

ZHAI, C. et LAFFERTY, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22:179–214.

Repérage des entités nommées pour l'arabe : adaptation non-supervisée et combinaison de systèmes

Souhir Gahbiche-Braham[1,2] Hélène Bonneau-Maynard[1,2] Thomas Lavergne[1]
François Yvon[1,2]
(1) LIMSI-CNRS, (2) Université Paris-Sud
{souhir.gahbiche,helene.maynard,thomas.lavergne,francois.yvon}@limsi.fr

RÉSUMÉ
La détection des Entités Nommées (EN) en langue arabe est un prétraitement potentiellement utile pour de nombreuses applications du traitement des langues, en particulier pour la traduction automatique. Cette tâche représente toutefois un sérieux défi, compte tenu des spécificités de l'arabe. Dans cet article, nous présentons un compte-rendu de nos efforts pour développer un système de repérage des EN s'appuyant sur des méthodes statistiques, en détaillant les aspects liés à la sélection des caractéristiques les plus utiles pour la tâche ; puis diverses tentatives pour adapter ce système d'une manière entièrement non supervisée.

ABSTRACT
Named Entity Recognition for Arabic : Unsupervised adaptation and Systems combination

The recognition of Arabic Named Entities (NE) is a potentially useful preprocessing step for many Natural Language Processing Applications, such as Machine Translation. This task is however made very complex by some peculiarities of the Arabic language. In this paper, we present a summary of our recent efforts aimed at developing a statistical NE recognition system, with a specific focus on feature engineering aspects. We also report several approaches for adapting this system in an entirely unsupervised manner to a new domain.

MOTS-CLÉS : Adaptation non supervisée, Repérage des entités nommées.

KEYWORDS: Unsupervised domain adaptation, named entity recognition.

1 Introduction

La détection des Entités Nommées (EN) est un élément essentiel à de nombreuses tâches de TAL, qu'elles soient mono ou multilingues, comme la recherche d'information ou la traduction automatique. En témoignent les nombreuses campagnes d'évaluation internationales (MUC, CoNLL, ACE) ou nationales (ESTER) organisées au cours des 15 dernières années.

Nous nous intéressons à cette question dans un contexte de traduction automatique statistique depuis l'arabe vers le français (ou l'anglais). Il est fréquent, en effet, qu'une EN à traduire ne soit pas présente dans les corpus parallèles qui servent à entraîner les systèmes de traduction. Dans (Gahbiche-Braham *et al.*, 2011), nous avons observé que parmi les 1% des formes inconnues, environ un quart correspondent à des EN. Dans ce cas, le système recopie, par défaut, le mot inconnu *verbatim* dans la sortie, alors qu'il serait préférable, pour une EN, de produire une forme

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 487–494,
Grenoble, 4 au 8 juin 2012.

translittérée en alphabet latin (Hermjakob *et al.*, 2008; Zhang *et al.*, 2011), ou encore de consulter des dictionnaires. (Daumé III et Jagarlamudi, 2011) adaptent leur système de traduction en créant des dictionnaires à partir de mots fréquents dans le domaine cible. L'étiquetage en EN apparait donc comme un prétraitement potentiellement utile à la traduction.

L'arabe est une langue morphologiquement riche et complexe. L'analyse automatique des mots arabes est compliquée par l'absence de voyellation dans les textes écrits d'une part (Habash, 2010), et d'autre part par l'existence de nombreuses variantes orthographiques, notamment sur les noms propres, ce qui multiplie les formes inconnues dans les textes. L'étiquetage en EN en langue arabe représente de nombreux défis intéressants : l'arabe se caractérise par le manque de ressources dictionnairiques et surtout par l'absence de distinction majuscule/minuscule qui est un indicateur très utile pour identifier les noms propres dans les langues utilisant l'alphabet latin.

À la suite de nombreux travaux, nous abordons cette tâche avec des outils d'apprentissage automatique et utilisons le modèle des champs markoviens conditionnels (ou CRF (Lafferty *et al.*, 2001)), avec l'implémentation présentée dans (Lavergne *et al.*, 2010), qui permet de construire des modèles intégrant un très grand nombre de descripteurs. Cette démarche pose la question de la pertinence des corpus d'apprentissage au regard des données de test. Nous traitons cette question en explorant les possibilités d'une adaptation non-supervisée. Nous proposons enfin une hybridation entre un système statistique et un système symbolique. Le reste de l'article est organisé comme suit. Dans la section 2, nous passons en revue des travaux sur le repérage des EN dans les textes arabes, et sur l'adaptation de modèles statistiques. Nous présentons dans la section 3 les expériences qui ont conduit au développement de notre système de base. L'adaptation de notre système est décrite à la section 4. La section 5 conclut ces travaux.

2 État de l'art

2.1 Étiquetage en EN pour l'arabe

Les premiers travaux sur la reconnaissance des EN pour l'arabe datent de 1998 et reposent sur des méthodes à base de règles (Maloney et Niv, 1998), voir également le travail plus récent de (Shaalan et Raza, 2009) ou de (Zaghouani *et al.*, 2010). (Samy *et al.*, 2005) utilisent un corpus parallèle pour extraire des EN en arabe. Ils utilisent un étiqueteur à base de règles enrichies avec un lexique monolingue espagnol pour extraire les EN en espagnol qui sont, par la suite, translittérées vers l'arabe. (Zitouni *et al.*, 2005) utilisent des techniques d'apprentissage automatique (des *Maximum Entropy Markov Models*) en considérant des jeux de descripteurs idoines, et parviennent à de très bons résultats.

Ces travaux ont été prolongés en particulier par Benajiba et ses co-auteurs, et ont donné lieu notamment à la construction du corpus ANER (voir section 3). Dans une première approche (Benajiba et Rosso, 2007), un étiquetage fondé sur le maximum d'entropie est exploré. Cette approche est étendue ensuite en décomposant la prédiction en deux temps : d'abord les frontières de l'EN en introduisant des catégories morpho-syntaxiques (POS), puis à la détermination de son type. Une seconde approche, fondée sur l'utilisation des CRF (Benajiba et Rosso, 2008) a permis d'explorer l'intégration de l'ensemble des traits dans un modèle unique, amenant à de meilleures performances. (Benajiba *et al.*, 2008) montrent également l'efficacité d'un prétraitement des textes pour séparer les différents constituants du mot (préfixes, lemme, et suffixes). (Abdul Hamid et Darwish, 2010) intègrent des traits intra-mot (n-grammes de caractères) dans

une modélisation CRF. Cette approche permet de capturer implicitement les caractéristiques morphosyntaxiques, introduites explicitement dans les expériences de (Benajiba et Rosso, 2008).

2.2 Adaptation et combinaison de systèmes

En apprentissage automatique, l'adaptation consiste à développer un système de traitement pour un domaine cible à partir de données et/ou d'un système de traitement développé pour un domaine source. D'un point de vue statistique, cela implique que les distributions des exemples observés sont différentes au moment de l'apprentissage et au moment du test.

Cette problématique a fait l'objet de multiples propositions en modélisation statistique des langues (par exemple l'étude de (Bellagarda, 2001) pour les modèles statistiques de langue), utilisation de pondérations différentielles pour les exemples de la source et de la cible (Jiang et Zhai, 2007), utilisation de descripteurs spécifiques pour les exemples source et cible (Daume III, 2007), etc. (Daume III *et al.*, 2010) présentent des travaux plus récents. Dans un cadre non supervisé, la stratégie la plus commune est l'auto-apprentissage (*self-training*) générant automatiquement des données d'apprentissage pour le domaine cible à partir du système source (Mihalcea, 2004).

Concernant le repérage des EN, le problème de l'adaptation se pose avec une acuité particulière, dûe au fait que les EN (i) sont souvent associées avec un thème particulier et (ii) ont également des distributions d'occurences très variables dans le temps. Cette problématique est étudiée en particulier par (Béchet *et al.*, 2011) qui (i) combinent deux approches d'étiquetage en EN pour le français : une approche symbolique avec une approche probabiliste et (ii) adaptent le système probabiliste fondé sur un processus discriminant à base de CRF, au domaine des données de test.

3 Étiquetage en entités nommées : systèmes de base

Dans cette section, nous décrivons les expériences réalisées pour développer des systèmes de base et en particulier pour identifier les descripteurs linguistiques utilisés. Tous ces modèles sont entraînés avec l'implémentation des CRF réalisée dans l'outil Wapiti[1] (Lavergne *et al.*, 2010). Cette implémentation permet (i) d'utiliser de très gros modèles incluant nominalement des centaines de millions de descripteurs, et (ii) de sélectionner les descripteurs les plus utiles par le biais d'une pénalité L_1 (Sokolovska *et al.*, 2009).

3.1 Protocole expérimental

Les expériences ont été réalisées sur le corpus ANER[2] (Benajiba *et al.*, 2007) constitué à partir d'articles de presse, et composé de plus de 150 000 occurrences de mots (4 871 phrases). Le corpus distingue 4 types d'EN : localisation (LOC : 40% des EN observées), personne (PERS : 32%), organisation (ORG : 18%) et une classe « divers » regroupant tous les autres types (MISC : 10%)[3], et peut être considéré comme le corpus de référence pour la tâche. Il utilise le schéma d'annotation IOB-2 et distingue 9 étiquettes. Les expériences sont produites à partir de données translittérées[4], sans faire d'analyse morphologique. Les scores sont calculés en utilisant l'outil

[1] Wapiti est librement disponible à l'adresse `http://wapiti.limsi.fr`.
[2] `http://users.dsic.upv.es/~ybenajiba/downloads.html`
[3] Seuls les trois premiers types sont utilisés dans nos évaluations.
[4] `http://www.qamus.org/transliteration.htm`

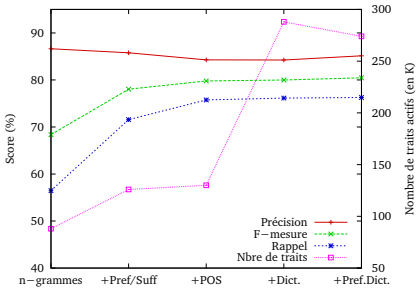


FIG. 1 – Précision (en %), rappel (en %), F-mesure et nombre de traits actifs pour des modèles de complexité croissante (à chaque nouveau modèle, de nouveaux traits sont ajoutés).

d'évaluation de CoNLL 2002[5]. Les modèles sont évalués par validation croisée à 10 partitions, sur des tests d'environ 25 000 mots chacun.

3.2 Sélection de caractéristiques et comparaison à l'état de l'art

Différentes versions du modèle de base ont été développées, qui incluent des jeux de descripteurs de richesse croissante. Nous décrivons ci-dessous les principales familles de descripteurs ; chaque réalisation x d'un élément d'une de ces familles donne lieu à un ensemble de fonctions booléennes testant x avec chaque étiquette et avec chaque bigramme d'étiquettes possibles.
N-grammes de mots : ces caractéristiques testent tous les unigrammes, les bigrammes, les trigrammes et les quadrigrammes dans, respectivement, des fenêtres de tailles 5, 3, 4 et 5.
Préfixes et suffixes : chaque séquence d'une, deux, ou trois lettres observée à l'initiale ou à la finale d'un mot du corpus d'apprentissage donne lieu à un nouveau descripteur. L'apparition de ces préfixes et suffixes est testée dans une fenêtre de taille 5 centrée sur le mot courant.
POS-tags : ce trait concerne les étiquettes morpho-syntaxiques prédites en utilisant un modèle entraîné par Wapiti sur l'*Arabic Tree Bank*[6] (Gahbiche-Braham *et al.*, 2012). Les tests évaluent les unigrammes et bigrammes d'étiquettes respectivement dans des fenêtres de tailles 5 et 3.
Ponctuation et nombres : ce trait teste la présence de caractères de ponctuations et de chiffres dans le mot courant ainsi que dans les deux mots voisins.
Dictionnaires : Ces dictionnaires proviennent de l'ANERGazet [2], d'extraits de Wikipedia et de la base de noms propres distribuée par JRC[7]. Pour chaque mot w, on teste s'il figure dans le dictionnaire (Dict sur la figure 1), ou s'il y figure précédé de préfixes (Pref.Dict). Ces dictionnaires contiennent 3 798 noms de lieux, 386 noms d'organisation et 13 648 noms de personnes.

Les résultats de ces expériences sont reportés sur la figure 1, qui représente la variation de la précision, du rappel et de la F-mesure, ainsi que le nombre de traits actifs. On constate qu'au fur et à mesure que de nouveaux traits sont ajoutés au modèle précédent, le rappel et la F-mesure augmentent, parfois au prix d'une légère dégradation de la précision.

[5] `http://bredt.uib.no/download/conlleval.txt`
[6] `http://www.ircs.upenn.edu/arabic/`
[7] *Joint Research Center* de la Communauté européenne : `http://langtech.jrc.it/JRC-Names.html`
[8] Le total reporté dans (Benajiba et Rosso, 2008) inclue l'EN MISC. Le total ici a été fait en calculant la moyenne.

Résultats du système de base				Résultats de (Benajiba et Rosso, 2008)[8]			
	Précision	Rappel	$F_{\beta=1}$		Précision	Rappel	$F_{\beta=1}$
LOC	90,59%	85,42%	87,83	LOC	93,03%	86,67%	89,74
ORG	78,67%	61,05%	68,75	ORG	84,23%	53,94%	65,76
PERS	81,31%	73,61%	77,27	PERS	80,41%	67,42%	73,35
Total	85,14%	76,27%	80,46	Total	85,89%	69,34%	76,28

TAB. 1 – Précision, rappel et F-mesure du modèle de base (qui combine tous les traits) sur le corpus ANER en comparaison avec les résultats de (Benajiba et Rosso, 2008)

Le tableau 1 donne une autre vue des performances du modèle le plus complet qui comprend environ 275 000 traits finalement sélectionnés par Wapiti sur un potentiel d'environ 80 millions. Afin de comparer notre système à l'état de l'art, le tableau 1 présente également les résultats obtenus par (Benajiba et Rosso, 2008) sur le corpus ANER avec un système utilisant également les CRF. Notre modèle de base semble donc cohérent avec les performances décrites dans l'état de l'art et atteint des performances globales semblables à celles décrites dans (Abdul Hamid et Darwish, 2010).

4 Adaptation du système d'étiquetage des entités nommées

Les applications étudiées dans ce travail s'inscrivent dans le cadre du projet SAMAR[9], qui vise à développer une plateforme de traitement de dépêches en langue arabe. Les données sont principalement produites par l'Agence France Presse (AFP). L'étiquetage en EN est envisagé ici comme un pré-traitement pour la traduction des données de l'arabe vers le français et l'anglais.

Nous disposons dans ce cadre de ressources supplémentaires pour adapter la détection des EN :
– de données du domaine (AFP), non-annotées (130 000 phrases, 3 500K mots) ;
– d'un étiquetage automatique d'une partie des données réalisé par un système symbolique développé par un des partenaires du projet, TEMIS (Guillemin-Lanne *et al.*, 2007).
– d'un corpus de test annoté manuellement et constitué de 900 phrases issues de l'AFP.

Les dépêches traitées dans notre application diffèrent substantiellement des données du corpus ANER, qui contient à la fois des articles de presse, des données collectées en ligne, en particulier des extraits de Wikipedia. Il existe également un décalage temporel entre la constitution de ce corpus (2007) et les données que nous devons traiter, qui sont postérieures à 2009.

4.1 Adaptation non-supervisée par auto-apprentissage

Le système de base, constitué à partir du corpus ANER, est utilisé pour annoter automatiquement le corpus AFP. Deux systèmes adaptés sont alors obtenus en utilisant comme corpus d'entraînement soit (i) le corpus étiqueté automatiquement seul, soit (ii) les deux corpus. Le tableau 2 donne les résultats des trois systèmes sur les données de test AFP. On constate une baisse sensible des performances du système de base (la F-mesure passe de 80,46 sur les données de test ANER à 72,64 sur les données de test AFP). Après adaptation (AFP et ANER+AFP), on constate une amélioration de la F-mesure pour les noms de lieux et d'organisations.

[9] http://samar.fr

	Modèle de base ANER			Modèle AFP			Modèle ANER+AFP		
	Précision	Rappel	$F_{\beta=1}$	Précision	Rappel	$F_{\beta=1}$	Précision	Rappel	$F_{\beta=1}$
LOC	89,30%	78,85%	83,75	90,81%	77,84%	83,83	91,45%	78,18%	84,29
ORG	50,24%	37,72%	43,09	51,01%	35,94%	42,17	51,76%	36,65%	42,92
PERS	68,07%	66,03%	67,03	70,83%	69,29%	70,05	70,87%	68,75%	69,79
Total	77,61%	68,27%	72,64	79,39%	68,14%	73,34	79,86%	68,34%	73,65

TAB. 2 – Comparaison et Adaptation de système de reconnaissance d'entités nommées

En moyenne, nous gagnons 1 point en F-mesure pour le système adapté. La bonne qualité des performances obtenues avec les annotations automatiques est principalement dûe à une augmentation très sensible de la couverture. Alors que seules 11% environ des EN de type personne du test sont dans le corpus ANER, on en retrouve plus de 60% quand on utilise le corpus automatique AFP pour l'apprentissage. Des écarts similaires, quoique moins importants, sont obtenus pour les organisations, et dans une moindre mesure, pour les lieux. Ceci illustre bien le caractère très localisé des occurrences des EN dont la distribution fluctue en fonction de l'actualité. D'une manière générale, ces améliorations restent toutefois limitées. Il est possible que la sélection des données de test (par l'AFP) conduise à sous-estimer l'apport de l'adaptation : le jeu de test contient majoritairement des dépêches ressortissant aux thèmes « guerre » et « politique », mais aucune de la catégorie « sport », pourtant très présente dans le corpus d'entraînement.

4.2 Un système hybride

Nous présentons ici les performances des trois modèles d'étiquetage décrits à la section 4.1 dans un cadre de combinaison de systèmes. La démarche suivie consiste à étiqueter automatiquement le corpus de test par l'annotateur de Temis, qui atteint une précision de 81% et une F-measure de 74% sur le corpus de test de l'AFP.

Le corpus de test est ensuite étiqueté une seconde fois par Wapiti, en considérant que les EN annotées par Temis sont correctes et en n'utilisant Wapiti que pour prédire les zones qui n'ont pas été détectées comme EN par l'étiqueteur symbolique. Les résultats sont donnés dans le tableau 3.

	Modèle de base ANER			Modèle AFP			Modèle ANER+AFP		
	Précision	Rappel	$F_{\beta=1}$	Précision	Rappel	$F_{\beta=1}$	Précision	Rappel	$F_{\beta=1}$
LOC	94,01%	85,12%	89,35	92,52%	81,31%	86,56	92,76%	81,31%	86,66
ORG	86,26%	66,18%	74,90	84,85%	61,09%	71,04	85,93%	62,18%	72,15
PERS	84,31%	76,01%	79,95	79,44%	72,22%	75,66	80,67%	72,73%	76,49
Total	90,24%	79,39%	84,47	87,80%	75,36%	81,11	88,45%	75,68%	81,57

TAB. 3 – Adaptation et test sur un corpus pré-étiqueté par un analyseur symbolique

Ces résultats montrent dans tous les cas une amélioration très sensible (+8 points) par rapport aux résultats antérieurs, en particulier quand on utilise le modèle non-adapté, qui a de meilleures performances que les modèles adaptés. Ceci est dû au fait que le système ANER a été entraîné sur un corpus annoté manuellement quand les systèmes adaptés utilisent des annotations automatiques potentiellement bruitées. Par comparaison avec le tableau 2, l'hybridation améliore les performances de chacun des systèmes pris séparément. Ceci ouvre des perspectives, en particulier pour mettre en place l'hybridation dès la construction du corpus d'apprentissage.

5 Conclusion

Dans cet article, nous avons présenté un système d'étiquetage en Entités Nommées construit par des méthodes d'apprentissage supervisé. Ce système, qui embarque des centaines de milliers de descripteurs, obtient des performances comparables aux meilleurs systèmes de l'état de l'art. Nous avons ensuite exploré diverses manières de réaliser une adaptation non-supervisée, par auto-apprentissage, de ce système conduisant à une légère amélioration des performances. Nous avons enfin montré qu'une hybridation du système statistique avec un système symbolique pouvait donner lieu à des gains bien supérieurs.

Ce travail ouvre de multiples perspectives portant sur les aspects liés à l'adaptation comme sur les aspects relatifs à la traduction. Concernant l'adaptation, il reste à reprendre les expériences précédentes avec un corpus produit par combinaison d'annotations ; de manière plus fondamentale, il reste également à voir comment *entraîner* Wapiti avec ces pré-étiquetages partiels, en utilisant par exemple des modèles à données latentes. Du point de vue de l'application finale, deux questions restent posées. L'une concerne l'ordre dans lequel effectuer les traitements préalables à la traduction : trois étapes s'enchaînent dans notre pipeline actuel : analyse morpho-syntaxique, détection des EN, puis segmentation des formes complexes. Il n'est pas dit que cet ordre soit optimal, et d'autres architectures devront être explorées. Ensuite, l'impact de la détection des EN sur la qualité de la traduction doit être évalué. Un travail préalable consistera à étudier comment les EN sont transférées d'une langue à l'autre, à partir de corpus parallèles annotés en EN.

Remerciements

Ces travaux ont été partiellement financés par le projet Cap-Digital SAMAR et par le programme Quaero. Merci à TEMIS pour l'annotation des corpus et à l'AFP pour le corpus de référence.

Références

Abdul Hamid, A. et Darwish, K. (2010). Simplified feature set for Arabic named entity recognition. *In Proc. of the 2010 Named Entities Workshop*, pages 110–115, Uppsala.

Béchet, F., Sagot, B. et Stern, R. (2011). Coopération de méthodes statistiques et symboliques pour l'adaptation non-supervisée d'un système d'étiquetage en entités nommées. *In actes de la conférence TALN*, Montpellier, France.

Bellagarda, J. R. (2001). An overview of statistical language model adaptation. *In Proc. of the ISCA Workshop on Adaptation Methods for Speech Recognition*, pages 165–174, Sophia Antipolis.

Benajiba, Y., Diab, M. et Rosso, P. (2008). Arabic named entity recognition using optimized feature sets. *In Proc. of EMNLP*, EMNLP, pages 284–293.

Benajiba, Y. et Rosso, P. (2007). Conquering the NER task for the Arabic language by combining the maximum entropy with POS-tag information. *In Proceedings of Workshop on Natural Language-Independant Engineering*, IJCAI.

Benajiba, Y. et Rosso, P. (2008). Arabic named entity recognition using conditional random fields. *In Proceedings of the Conference on Language Resources and Evaluation*.

Benajiba, Y., Rosso, P. et Benedí, J.-M. (2007). Anersys : An arabic named entity recognition system based on maximum entropy. *In CICLing*, pages 143–153.

DAUME III, H. (2007). Frustratingly easy domain adaptation. *In Proc. of the 45th Annual Meeting of the ACL*, pages 256–263, Prague, Czech Republic.

DAUME III, H., DEOSKAR, T., MCCLOSKY, D., PLANK, B. et TIEDEMANN, J., éditeurs (2010). *Proc. of the 2010 Workshop on Domain Adaptation for NLP*. Uppsala, Sweden.

DAUMÉ III, H. et JAGARLAMUDI, J. (2011). Domain adaptation for machine translation by mining unseen words. *In ACL*, Portland, OR.

GAHBICHE-BRAHAM, S., BONNEAU-MAYNARD, H., LAVERGNE, T. et YVON, F. (2012). Joint segmentation and POS tagging for arabic using a CRF-based classifier. *In Proc. of LREC'12*.

GAHBICHE-BRAHAM, S., BONNEAU-MAYNARD, H. et YVON, F. (2011). Two ways to use a noisy parallel news corpus for improving statistical machine translation. *In Proc. of Workshop on Building and Using Comparable Corpora*, pages 44–51, Portland, OR.

GUILLEMIN-LANNE, S., DEBILI, F., TAHAR, Z. B. et GACI, C. (2007). Reconnaissance des entités nommées en arabe. *In Colloque VSST, Veille Stratégique Scientifique et Technologique*.

HABASH, N. (2010). *Introduction to Arabic Natural Language Processing*. Morgan Claypool.

HERMJAKOB, U., KNIGHT, K. et DAUMÉ III, H. (2008). Name translation in statistical machine translation - learning when to transliterate. *In Proc. of ACL-08 : HLT*, pages 389–397, Ohio.

JIANG, J. et ZHAI, C. (2007). Instance weighting for domain adaptation in nlp. *In Proc. of the 45th Annual Meeting of the ACL*, pages 264–271, Prague, Czech Republic.

LAFFERTY, J., MCCALLUM, A. et PEREIRA, F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *In Proc. ICML*, pages 282–289, San Francisco.

LAVERGNE, T., CAPPÉ, O. et YVON, F. (2010). Practical very large scale CRFs. *In Proceedings the 48th Annual Meeting of the Association for Computational Linguistics*, pages 504–513.

MALONEY, J. et NIV, M. (1998). TAGARAB : a fast, accurate Arabic name recognizer using high-precision morphological analysis. *In Proc. of the Workshop on Computational Approaches to Semitic Languages*, Semitic '98, pages 8–15, Stroudsburg, PA, USA.

MIHALCEA, R. (2004). Co-training and self-training for word sense disambiguation. *In* NG, H. T. et RILOFF, E., éditeurs : *HLT-NAACL Workshop : CoNLL-2004*, pages 33–40, Boston.

SAMY, D., MORENO, A. et MA GUIRAO, J. (2005). A proposal for an Arabic named entity tagger leveraging a parallel corpus. RANLP '05.

SHAALAN, K. et RAZA, H. (2009). NERA : Named entity recognition for arabic. *Journal of the American Society for Information Science and Technology*, 60(9):1652–1663.

SOKOLOVSKA, N., CAPPÉ, O. et YVON, F. (2009). Sélection de caractéristiques pour les champs aléatoires conditionnels par pénalisation l_1. *TAL*, 50(3):139–171.

ZAGHOUANI, W., POULIQUEN, B., EBRAHIM, M. et STEINBERGER, R. (2010). Adapting a resource-light highly multilingual named entity recognition system to arabic. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), pages 563–567.

ZHANG, M., LI, H., KUMARAN, A. et LIU, M. (2011). Report of news2011 machin transliteration shared task. *In Proceedings of the 2011 Named Entities Workshop*.

ZITOUNI, I., SORENSEN, J., LUO, X. et FLORIAN, R. (2005). The impact of morphological stemming on Arabic mention detection and coreference resolution. *In Proc. of Workshop on Computational Approaches to Semitic Languages*, pages 63–70, Ann Arbor, Michigan.

Propagation de polarités dans des familles de mots : impact de la morphologie dans la construction d'un lexique pour l'analyse d'opinions

Núria Gala[1] *Caroline Brun*[2]
(1) LIF-CNRS UMR 7279, 163 av. de Luminy case 901, 13288 Marseille Cedex 9, France
(2) Xerox Research Centre Europe, 6 chemin de Maupertuis 38240 Meylan, France
nuria.gala@lif.univ-mrs.fr, caroline.brun@xrce.xerox.com

RESUME

Les ressources lexicales sont cruciales pour de nombreuses applications de traitement automatique de la langue (par exemple, l'extraction d'opinions à partir de corpus). Cependant, leur construction pose des problèmes à différents niveaux (coût, couverture, etc.). Dans cet article, nous avons voulu vérifier si les informations morphologiques liées à la dérivation pouvaient être exploitées pour l'annotation automatique d'informations sémantiques. En partant d'une ressource regroupant les mots en familles morphologiques en français, nous avons construit un lexique de polarités pour 4 065 mots, à partir d'une liste initiale d'adjectifs annotés manuellement. Les résultats obtenus montrent que la propagation des polarités est correcte pour 78,89% des familles avec un seul adjectif. Le lexique ainsi obtenu améliore aussi les résultats du système d'extraction d'opinions.

ABSTRACT

Spreading Polarities among Word Families: Impact of Morphology on Building a Lexicon for Sentiment Analysis

Lexical resources are essential for many natural language applications (for example, opinion mining from corpora). However, building them entails different problems (cost, coverage, etc.). In this paper, we wanted to verify whether morphological information about derivation could be used to automatically annotate semantic information. Starting from a resource that groups words into morphological families in French, we have built a lexicon with polarities for 4 065 words from an initial seed set of manual annotated adjectives. The results obtained show that spreading polarities is accurate for 78.89% of the families with a unique adjective. The lexicon obtained also improves the results of the opinion mining system on different corpora.

MOTS-CLES : ressources lexicales, morphologie dérivationnelle, analyse de sentiments

KEYWORDS : lexical resources, derivational morphology, opinion mining

1 Introduction

Depuis quelques années, l'analyse de sentiments suscite de l'intérêt dans la communauté du traitement automatique des langues (TAL), comme conséquence d'un réel besoin dans le traitement de grandes masses de données : services web pour le tourisme ou la culture, discours politiques, etc. Par analyse de sentiments, on entend la détection de la polarité d'un texte, c'est-à-dire, l'obtention automatique de la tendance ou de l'opinion qui s'en dégage.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 495–502,
Grenoble, 4 au 8 juin 2012.

Deux approches ressortent dans la littérature. Les approches statistiques supervisées, fondées sur les co-occurrences de mots dans des corpus, et les approches plus linguistiques qui s'appuient, elles, sur des ressources lexicales. L'idée des méthodes statistiques est de calculer, à partir d'un ensemble de co-occurrences annotées, d'autres co-occurences polarisées (Hatzivassiloglous et McKeown 1997 ; Turney 2002). La procédure de classification se fait automatiquement à partir d'exemples, le modèle attribue une polarité en fonction d'un processus inductif (Pang et al. 2002). L'autre type de méthode est considéré plus linguistique, dans la mesure où l'on utilise des ressources comme des thésaurus, des réseaux lexicaux, etc. (Kim et Hovy 2004 ; Esuli et Sebastiani 2005), ce qui permet d'améliorer les performances des systèmes d'analyse (Choi et Cardie 2009 ; Lu et al. 2011). Les méthodes qui utilisent des lexiques présupposent deux hypothèses : toute unité lexicale aurait une orientation sémantique intrinsèque, indépendamment de son contexte d'apparition ; cette orientation peut être exprimée avec une valeur numérique (Taboada et al. 2011).

Notre travail se situe dans cette perspective : nous nous sommes proposées de construire un lexique de polarités. A partir d'une liste initiale de 3 882 adjectifs annotés manuellement par trois annotateurs, nous avons voulu observer l'impact de la morphologie dérivationnelle dans le maintien ou non de la polarité. C'est-à-dire, nous avons voulu tester l'hypothèse selon laquelle la polarité intrinsèque d'un adjectif est la même que celle des unités lexicales de sa famille morphologique. L'idée a été de voir (i) si on pouvait construire une ressource qui capitalise sur les liens morphologiques pour propager des informations sémantiques, et (ii) si une telle ressource améliore les résultats d'un système d'analyse d'opinions.

Si l'estimation de la polarité d'un texte passe par des phénomènes contextuels (intensificateurs, négation, etc.) et syntaxiques (Brun 2011), la qualité du lexique à la base du système reste cruciale. La construction d'un tel lexique demeure donc un aspect important. Des lexiques existants pour l'anglais ont été construits à partir de WORDNET, par exemple, WORDNET-AFFECT (Strapparava and Valitutti 2004), SENTIWORDNET (Esuli and Sebastiani 2006). Pour le français, on peut citer les travaux de Vernier et Monceaux (2010) pour obtenir automatiquement une liste de 982 termes subjectifs à partir de l'indexation de documents sur le Web ou l'application LIKEIT de JEUXDEMOTS (Lafourcade 2007) où l'enrichissement de la liste de mots polarisés se fait de façon contributive. En dehors de ces exemples, à notre connaissance, il n'existe pas de lexique de polarités pour le français.

Dans cet article, nous décrirons la méthodologie de construction de notre lexique (section 2) et dans la section 3, nous évaluerons la qualité des données obtenues au regard des familles morphologiques (propagation des polarités). Nous présenterons, enfin, les résultats d'un système d'analyse d'opinions qui intègre la ressource.

2 Construction de la ressource lexicale

Pour constituer le lexique de polarités, nous avons utilisé la deuxième version de POLYMOTS, une ressource lexicale regroupant 19 009 mots, à ce jour, en 2 069 familles morpho-phonologiques (Gala et Rey 2008). La deuxième version de cette ressource, outre une description plus fine de quelques familles de mots en clusters sémantiques

(Gala et al. 2011), contient des étiquettes grammaticales, ce qui nous a permis d'extraire les 3 785 adjectifs et de les annoter manuellement avec trois valeurs (positif, négatif, neutre). Cette liste initiale d'adjectifs a été complétée avec une centaine d'adjectifs supplémentaires provenant d'un lexique de l'analyseur XIP. Nous totalisons 3 882 adjectifs annotés.

2.1 Accord inter-annotateurs

Afin de prendre en compte l'accord inter-annotateurs, nous avons transformé les étiquettes en pourcentages (*100%neg, 33%neutre_66%neg, 66%pos_33%neutre*, etc.). Sans compter l'accord à 100%, nous obtenons 22 étiquettes différentes que nous avons regroupées en accord majoritaire (75%-25% si il y a eu quatre annotations -les trois annotateurs initiaux plus l'annotation provenant du lexique de XIP- ou 66%-33%). Enfin, nous avons considéré comme non significatif les cas où il y a eu un seul annotateur ou bien les cas où il n'y a pas eu de tendance claire (*33%pos_33%neutre_33%neg, 50%pos_25%neutre_25%neg,* etc.). La distribution des étiquettes en termes d'accord inter-annotateurs est la suivante : 1 341 adjectifs avec accord total (34,5%), 969 accord majoritaire (25%) et 1 572 accord non significatif (40,50%)[1].

2.2 Propagation des polarités vers les familles de mots

Pour chacun des 3 882 adjectifs, nous avons étendu automatiquement sa polarité vers les mots de sa famille morphologique. Trois cas de figure se sont présentés : (i) la famille contient un seul adjectif, (ii) la famille en contient plusieurs, (iii) la famille n'en contient aucun.

Dans le cas où plusieurs adjectifs sont présents dans la famille (36,97% des cas, 765 familles au total), la difficulté réside dans le choix du critère d'attribution de la polarité lorsqu'elle est différente. A ce stade, le choix de l'étiquette à propager devait être arbitraire, nous n'avons donc pas utilisé l'ensemble de ces données.

Le 32,19% des familles de POLYMOTS (666 au total) qui ne contiennent pas d'adjectifs, a également échappé au processus d'annotation automatique. Il s'agit de familles pour lesquelles la dérivation est nulle (*agrume, aisselle, falaise*, *oncle*, *taie*, etc.) ou quasi nulle (*cage/cageot, poutre/poutrelle/poutraison, nid/nidation/nidification/nidifier,* etc.). Il s'agit aussi de familles pour lesquelles des adjectifs ont été rajoutés après avoir constitué notre liste d'adjectifs initiale.

A ce jour, le lexique que nous avons créé par propagation des polarités à partir des adjectifs initialement annotés contient, 4 065 mots correspondant aux 638 familles avec un seul adjectif (30,84%). La moyenne de mots par famille est de 6,4 mots. Le lexique est constitué de 662 adjectifs (16,3%), 2 337 noms (57,4%), 878 verbes (21,6%) et 193 adverbes (4,7%), cf. http://polarimots.lif.univ-mrs.fr.

[1] Cette classification donne une pondération supérieure aux adjectifs pour lesquels une polarité se dégage. Ce poids est encodé dans le lexique : poids = 1 si 100% d'accord, poids = 0.5 si 75% ou 66% d'accord, poids = 0 si les valeurs sont distribuées (50%-50% ou 33%-33%-33%).

2.3 Remarques méthodologiques

La liste d'adjectifs initiale a été annotée par trois annotateurs bénévoles différents. Dans le cas de sens multiples, nous avons considéré le sens propre en priorité en nous appuyant sur les définitions du TLFi. Ainsi, par exemple, lancinant a comme définition « qui se fait sentir par élancements aigus ». Lors de l'annotation des mots de la famille de cet adjectif, la polarité négative a été propagée. Cela ne pose pas de problèmes pour *lance* (« arme ») ou *lanciner* (« se faire sentir par des élancements douloureux ») mais en pose pour d'autres termes de la famille comme *lancement, élan,* etc. qui devraient être annotés comme neutres. Dans notre lexique, les mots polysémiques (ex. *lancement* sens « départ ») ou homonymiques (ex. *élan* sens « animal ») n'ont qu'une polarité correspondant à un seul sens. Le traitement des sens figurés et de la polysémie en général reste un problème crucial qui mérite un travail beaucoup plus approfondi (qui sort du cadre de la construction de notre ressource et, de surcroît, de la présentation dans cet article). Par ailleurs, même si nous faisons l'hypothèse qu'un mot possède une polarité intrinsèque, nous sommes conscientes que celle-ci est susceptible de varier en fonction des mots co-occurrents (par exemple, pour *gorgé*, elle peut osciller entre positif et négatif dans, respectivement, « un fruit gorgé de vitamines » par rapport à « un terrain gorgé d'eau »). Dans ces cas, faute de contexte, nous avons attribué une polarité neutre.

Enfin, une fois la ressource annotée, un certain nombre de polarités attribuées automatiquement ont été modifiées en fonction d'affixes porteurs d'altérations sémantiques. C'est le cas d'affixes de négation/opposition (*anti-, contre- dé-/dés-, i-/im-/in-, mal-/mé-*) par exemple dans *antiatomique*, *contrepoison*, *déboisement, inachevé, malaisé*.... Étant donné la variété de polarités et de cas (opposition nette positif/négatif : *poison/contrepoison, salissure/antisalissure*, etc. ; dégradation ou amélioration par rapport à une marque neutre : *obligeant/désobligeant*, *créditer/discréditer*, etc.) nous avons modifié les polarités des mots avec des affixes de négation au cas par cas.

2.4 Évaluation intrinsèque

Une évaluation manuelle a été faite pour 2 954 mots correspondant à 450 familles annotées automatiquement par propagation de la polarité à partir d'un adjectif (environ 70% du lexique). Les résultats de cette évaluation sont les suivants : 355 familles maintiennent la polarité de l'adjectif (78,89%) et 95 ne la maintiennent pas (21,11%).

L'impact de la taille des familles morphologiques est un facteur essentiel dans le maintien d'une polarité. Ainsi, plus la taille de la famille est réduite, plus la polarité reste identique, étant donné une cohésion sémantique plus forte. C'est la cas de familles de moins de huit dérivés, par exemple : *acariâtre/acariâtreté* (100%neg), *vertèbre/vertébral/vertébré* (100%neutre), *allègre/allégrement/alléger/allégresse* (100%pos), *aromal/aromatique/aromatiser/arôme... (66%pos_33%neutre), etc.*

Dans le cas de familles avec un seul adjectif mais de taille plus grande (surtout au delà d'une dizaine de dérivés), la dispersion sémantique est trop importante, ce qui fait varier considérablement les polarités au sein de la famille (par exemple, *sec/dessécher/séchage/séchoir/...*, *bombé/bombe/bombage/bombardement/...*, *nerveux/nerf/nerveusement/nervure/...*, etc.). Comme évoqué plus haut, les problèmes de polysémie ont également un impact très important. Enfin, dans d'autres cas, les suffixes

formateurs d'adjectifs ou participes (*-é, -ant, -al, -eux*) modifient souvent le sens du mot dérivé et font basculer la polarité généralement de neutre à négatif (*âge/âgé, angle/anguleux, bête/bestial, bouillir/ébouillanté*, *larme/larmoyant*, etc.) mais aussi de neutre à positif (*baraque/baraqué, rayon/rayonnant*).

3 Évaluation par le biais d'un système d'extraction d'opinions

Pour évaluer la qualité de la ressource lexicale, nous l'avons intégrée à un système d'extraction d'opinions afin de mesurer l'impact de cette intégration sur la capacité de ce système à classer correctement des revues en ligne selon l'opinion globale de l'utilisateur.

3.1 Constitution des corpus d'évaluation

Nous avons collecté deux corpus de revues en ligne : l'un concernant des livres (*evene.fr/livres)* et l'autre des restaurants (*www.linternaute/restaurant)*. Ces revues sont au format html et semi-structurées. Elles ont été converties au format XML en filtrant les informations pertinentes : titre du livre/nom du restaurant, auteur du commentaire, date, note globale et commentaire libre. Le corpus de revues de livres contient 3 110 revues, le corpus de revues de restaurants contient 99 373 revues.

3.2 Le système d'extraction d'opinions

Notre système d'extraction d'opinions utilise les résultats de l'analyse syntaxique profonde fournie par l'analyseur syntaxique XIP (Ait-Mokthar et al. 2002). Une version de l'extracteur d'opinions a été conçue pour l'anglais (Brun 2011), nous l'adaptons pour le français : l'analyseur est enrichi avec un lexique contenant les polarités associées aux mots, et par un ensemble de règles syntactico-sémantiques qui extraient de relations d'opinions à « grain fin », c'est-à-dire des opinions associées aux éléments sur lesquelles elles portent. Par exemple, pour la phrase suivante, le système extrait :

« *Ce livre est très prenant, l'histoire est vraiment bien racontée.* »

OPINION[POSITIVE](prenant, livre) & OPINION[positive](vraiment bien racontée, histoire).

Où le premier élément de la relation est le prédicat porteur de la polarité et le deuxième élément est la cible de l'opinion. La première relation est déduite de la relation attributive détectée par XIP entre *livre* et *prenant*, adjectif de polarité positive. La deuxième relation découle du fait que le sujet passif de la phrase est en relation avec un prédicat modifié par un adverbe de polarité positive.

Actuellement, outre le développement des règles d'extraction de relations d'opinions, l'emphase a été mise sur le traitement des phénomènes de négation. Du point de vue des ressources lexicales, nous disposions initialement d'un lexique de polarités préliminaire, construit à la fois manuellement et en adaptant à notre système une partie du lexique de Blogoscopie, (Vernier et Monceaux 2010). Il était constitué de 714 mots dont 418 adjectifs (58,54%), 163 noms (22,83%), 111 verbes (15,55%) et 22 adverbes (3,08%). L'objectif des présents travaux est l'extension de ce lexique de polarités en utilisant les familles de mots.

3.3 Expériences et résultats

Pour évaluer la qualité de la ressource, nous utilisons une méthode similaire à celle utilisée dans (Brun 2011) qui a été conçue pour évaluer la performance de l'extracteur d'opinions pour l'anglais. Ici, le système est testé afin d'évaluer l'apport de la ressource lexicale. Nous avons donc exécuté une série des tests concernant l'intégration des lexiques de polarité. Les deux corpus constitués précédemment sont utilisés pour évaluer les performances du système de détection d'opinions sur la classification des revues. Ces corpus peuvent être considérés comme annotés pour la classification, l'auteur assignant une note globale à la revue. Nous utilisons les relations d'opinions extraites par le système pour entraîner et tester un classifieur SVM binaire (*SVMLight*, (Joachims 1999)) afin de classer les revues comme positives (note entre 3 et 5) ou négatives (note entre 0 et 2). La référence de cette évaluation est calculée avec le lexique initial, avant intégration de la ressource lexicale constituée. Le protocole d'évaluation utilise 200 revues du corpus « livre » (100 négatives, 100 positives, choisies aléatoirement) pour entrainer, valider et tester et 4000 revues du corpus « restaurant » (2000 négatives, 2000 positives, choisies aléatoirement) pour entrainer, valider et tester.

Les traits utilisés pour le SVM correspondent au nombre d'occurrences d'une relation d'opinion portant sur un une cible donnée : par exemple, si le système extrait, pour une revue donnée, 1 fois la relation *OPINION[positive](vraiment bien racontée, histoire),* le trait du SVM et sa valeur sont OPINION_POSITIVE_CIBLE_histoire = 1 ; si le système extrait les relations *OPINION[negative](décevant, cuisine)* et *OPINION[negative](mauvais, cuisine)*, le trait et sa valeur sont OPINION_NEGATIVE_CIBLE_cuisine = 2, etc.

	Taux d'accord	Exactitude classif. Corpus « livres »	Exactitude classif. Corpus « restaurant »
Référence	*(1)*	79,80%	74,80%
Adjectifs	100% *(2)*	81,50%	76,20%
	100% + 75%/66% *(3)*	82,50%	77,10%
Lexiques obtenus par propagation	100% *(4)*	**83,00%**	**77,90%**
	100% + 75%/66% *(5)*	82,40%	77,20%

TABLE 1 – Résultats de la classification des revues selon le type d'information lexicale

Pour les deux corpus, nous avons calculé l'exactitude de la classification par le SVM (« accuracy ») selon l'intégration des données lexicales. Nous avons distingué les ressources pour lesquelles l'accord entre annotateurs était de 100% lors de l'annotation initiale des adjectifs et celle pour lesquels l'accord était majoritaire (75% et 66%[2]). Nous

[2] Nous n'avons pas intégré les ressources pour lesquelles l'accord initial était plus faible.

avons donc mené 5 jeux de tests pour les deux corpus : (1) avec les ressources initiales, (2) en intégrant les adjectifs annotés avec un accord de 100%, puis (3) en intégrant les adjectifs dont le taux d'accord d'annotation manuelle était majoritaire, enfin les lexiques obtenus par propagation de la polarité des adjectifs aux familles de mots. Nous avons distingué celles pour lesquelles le taux d'accord inter-annotateur était à 100% (4) et celles pour lesquelles l'accord était entre 75% et 66% (5).

Les résultats sur les deux corpus de test sont synthétisés dans le tableau 1. Nous constatons que dans la plupart des cas, il y a un gain, et que la configuration optimale est obtenue dans le cas du test (4). Seule l'intégration du lexique obtenu automatiquement par propagations aux familles de mots pour les adjectifs dont l'accord inter-annotateurs était de 75-66% fait régresser le taux d'exactitude, ce qui montre l'importance de cet accord. En outre, on observe que les tendances sont les mêmes que le corpus traite de littérature ou de restaurants.

4 Conclusion

Dans cet article, nous nous sommes intéressées à la création automatique d'un lexique pouvant améliorer les résultats d'un système d'extraction d'opinions. Pour le construire, nous avons capitalisé sur une ressource regroupant les mots en familles morphologiques et avons observé si les informations sémantiques concernant la polarité des mots pouvaient se maintenir ou non au sein d'une même famille. Nous avons ainsi propagé automatiquement les polarités d'une liste d'adjectifs annotés manuellement vers les mots de leurs familles.

Les résultats de cette expérience montrent que, dans 78,89% des familles avec un seul adjectif, notre hypothèse se vérifie, à condition que certains affixes porteurs de modifications sémantiques (négation) soient pris en compte. L'utilisation de ce lexique dans un système d'extraction d'opinions montre que les résultats de la classification des opinions s'améliorent. Les améliorations plus significatives concernent les mots pour lesquels l'accord inter-annotateur est de 100%, ce qui implique que l'étape d'annotation initiale conditionne fortement les résultats finaux.

En perspective, différentes pistes sont envisageables. Concernant la construction du lexique, notre intérêt porte sur le traitement de la polysémie ainsi que sur la recherche de critères plus fins pour propager les polarités dans le cas de familles à plusieurs adjectifs. Concernant l'utilisation du lexique par le système d'extraction d'opinions, nous poursuivrons les expériences avec ces ressources améliorées ainsi que sur d'autres corpus.

Références

AIT-MOKTHAR, S., CHANOD, J.P. (2002). Robustness beyond Shallowness: Incremental Dependency Parsing. *Special Issue of NLE Journal.*

BRUN C. (2011). Detecting opinions using Deep Syntactic Analysis. *Proceedings of Recent Advances in Natural Language Processing (RANLP 2011)*, Hissar, Bulgaria.

HU M. ET LIU B. (2004). Mining and summarizing customer reviews. *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004)*, Seattle,

Washington, USA.

CHOI Y. ET CARDIE C. (2009). Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. *In Proceedings of the conference on empirical methods in NLP, EMNLP-09*, pages 741-748, Edinbourg, Ecosse.

ESULI ET SEBASTIANI (2005). Determining semantic orientation of terms through gloss classification. *In Proceedings of CIKM,* pages 617-624,

ESULI A. ET SEBASTIANI F. (2006). SentiWordNet: a publicly available lexical resource for opinion mining. *Dans les actes de LREC-06,* Gêne, Italie.

GALA, N., HATHOUT, N., NASR, A., REY, V. ET SEPPÄLÄ, S. (2011) Création de clusters sémantiques à partir du TLFi. Actes de *Traitement Automatique des Langues Naturelles* (TALN 2011). Montpellier, juin 2011.

GALA N. ET REY V. (2008). Polymots : une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques. Actes de *Traitement Automatique des Langues Naturelles* (TALN 2008), Avignon, juin 2008.

HATZIVASSILOGLOUS V. ET MCKEOWN K. (1997). Predicting the semantic orientation of adjectives. Actes de ACL-97, pages 174-181, Madrid, Espagne.

JOACHIMS T. (1999). Making large-Scale SVM Learning Practical. *Advances in Kernel Methods – Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (ed.), MIT Press.

KIM S. M. ET HOVY E. (2004). Determining the sentiment of opinions. Actes de *COLING-04*, pp. 1367-1373, Barcelone, Espagne.

LAFOURCADE M. (2007) Making people play for Lexical Acquisition. Actes de *7th Symposium on Natural Language Processing*, Pattaya, Thailand.

LU Y., CASTELLANOS M., DAYAL U., ZHAI CH. (2011). Automatic construction of context-aware sentiment lexicon: an optimization approach. Actes de *WWW Conference, session semantic analysis*. Hyderabad, India, pp. 347-356.

PANG B., LEE L. ET VAITHYANATHAN S. (2002) Thumbs up? Sentiment classification using machine learning techniques. *In Proceedings of the conference on empirical methods in NLP, EMNLP-02*, pp. 79-86, Philadelphia, USA.

STRAPPARAVA C. ET VALITUTTI A. (2004). WordNet Affect : an affective extension of WordNet. Actes de *4th International conference on Language Resources and Evaluation (LREC-04)*, Lisbon, Portugal.

TABOADA M., BROOKE J., TOFILOSKI M., VOLL K., STEDE M. (2011) Lexicon-based methods for sentiment analysis. Dans *Computational Linguistics,* Volume *37 (2)*.

TURNEY P. (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Actes de *ACL-02*, pages 417-424, Philadelphia, USA.

VERNIER M. ET MONCEAUX L. (2010) Enrichissement d'un lexique de termes subjectifs à partir de tests sémantiques. *Revue TAL volume 51 (1),* pp. 125-149.

Transitions thématiques : Annotation d'un corpus journalistique et premières analyses

Alexandre Labadié[1] Patrice Enjalbert[2] Stéphane Ferrari[2]
(1) GETALP, LIG, BP 53, 38041 Grenoble Cedex 9, France (2) Laboratoire GREYC, Université de Caen & CNRS, Bd Maréchal Juin, BP 5186 F-14032 Caen Cedex, France Alexandre.Labadie@imag.fr, Patrice.Enjalbert@unicaen.fr, Stephane.Ferrari@unicaen.fr

RÉSUMÉ

Le travail présenté dans cet article est centré sur la constitution d'un corpus de textes journalistiques annotés au niveau discursif d'un point de vue thématique. Le modèle d'annotation est une segmentation classique, à laquelle nous ajoutons un repérage de *zones de transition* entre unités thématiques. Nous faisons l'hypothèse que dans un texte bien construit, le scripteur fournit des indications aidant le lecteur à passer d'un sujet à un autre, l'identification de ces indices étant susceptible d'améliorer les procédures de segmentation automatique. Les annotations produites ont fait l'objet d'analyses quantitatives mettant en évidence un ensemble de propriétés des transitions entre thèmes.

ABSTRACT

Manual thematic annotation of a journalistic corpus : first observations and evaluation.

The work presented in this paper focuses on the creation of a corpus of journalistic texts annotated at dicourse level, more precisely on a topic level. The annotation model is a classic segmentation one, to which we add *transition zones* between topical units. We assume that in a well-structured text, the author provides information helping the reader to move from one topic to another, where an identification of these clues is likely to improve automatic segmentation. The produced annotations have been subject of several quantitative analyses showing a set of linguistic properties of topical transitions.

MOTS-CLÉS : Structure du discours, segments thématiques, transitions thématiques, annotation.

KEYWORDS: Discourse structure, topical segments, topical transitions, annotation.

1 Introduction

On peut observer actuellement un intérêt croissant pour l'analyse de la structure du discours dans la communauté TALN, tant à des fins applicatives (indexation de documents, résumé automatique, ...), que pour des études de linguistique de corpus. Cet article est consacré à celles de ces approches qui tentent de saisir l'organisation du texte en termes de "blocs homogènes" successifs selon des critères thématiques (succession de thèmes tout au long du texte). Notre contribution concerne tant le modèle discursif lui-même (accent mis sur les zones de transition) que la constitution d'un corpus annoté en conséquence. Des ressources de ce type sont clairement nécessaires, comme "référence" pour l'évaluation de traitements automatiques ou pour des observations d'ordre linguistique. De fait, ce travail a été initié au sein du projet ANR Annodis, visant à l'annotation de textes selon plusieurs modèles discursifs (Péry-Woodley *et al.*, 2009) [1]. Le

1. http://w3.erss.univ-tlse2.fr/annodis/

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 503–510,
Grenoble, 4 au 8 juin 2012.

corpus est composé d'articles de journaux (articles de fond de politique et d'économie).
La segmentation thématique est une tâche populaire en TALN (voir, par exemple (Morris et Hirst, 1991; Hearst, 1997; Choi, 2000)). Dans la plupart de ces travaux, les segments sont identifiés par leurs champs lexicaux et, malgré des résultats intéressants, on peut regretter que ces approches purement lexicales ne prennent pas en compte les signaux que les auteurs insèrent souvent dans le but d'informer le lecteur de changements de sujet. Dans notre travail, nous faisons l'hypothèse de l'existence de *zones de transition*, entre segments thématiques, ayant cette fonction, dans l'intention d'étudier leurs propriétés. Outre un intérêt théorique (mis en évidence par de nombreux travaux en linguistique) l'étude des indices de transition thématique est susceptible d'améliorer les performances d'outils de segmentation automatique [2]. Afin de tester notre hypothèse, nous avons donc demandé aux annotateurs d'indiquer de telles zones.
L'article est organisé comme suit. Nous présentons d'abord le corpus et la procédure d'annotation, de concert avec le modèle d'annotation thématique. Nous détaillons ensuite une mesure d'accord inter-annotateurs originale : aucune mesure réellement satisfaisante (permettant en particulier de comparer la segmentation de plusieurs annotateurs, et non le simple accord avec une annotation de référence) n'étant à ce jour disponible. Puis nous considérons la question des "signaux" indiquant les transitions entre segments thématiques.

2 Corpus et procédure d'annotation

2.1 Textes, outils et annotateurs

Le corpus en vue est composé de textes journalistiques du journal *Le Monde*, année 1994. Ce choix s'explique à la fois par les objectifs applicatifs (résumé, indexation, *etc.*) et par la qualité linguistique du journal. Nous avons choisi au hasard 30 textes (principalement en politique et en économie) de différentes tailles. Le corpus totalise 46 689 mots et a été réparti entre trois catégories : Petits : moins de 1000 mots (15 textes) ; Moyens : entre 1000 et 3000 mots (10 textes) et Grands : plus de 3000 mots (5 textes). Chaque texte a été annoté par 3 annotateurs au sein d'un groupe de 5 annotateurs "naïfs" [3] en utilisant la plateforme Glozz [4](Widlöcher et Mathet, 2009). Les textes ont été présentés aux annotateurs avec un minimum de mise en page (titre, sous-titres et paragraphes).

2.2 Modèle d'annotation thématique

Il existe une grande variété de méthodes de segmentation thématique. La plupart admettent qu'un segment thématique est une unité de texte qui est thématiquement cohérente et thématiquement distincte de la précédente et de la suivante. Certaines se concentrent sur la recherche des frontières (Beeferman *et al.*, 1999), d'autres tentent de regrouper des portions de texte selon leur cohérence lexicale (Choi, 2000). Mais, que ce soit pour l'apprentissage ou l'évaluation, les expériences s'appliquent surtout à des corpus obtenus artificiellement à partir de l'agrégation de petits textes, l'hypothèse étant que la frontière entre deux textes peut être assimilée à une frontière thématique. Cette hypothèse est régulièrement critiquée (Bestgen et Pirard, 2006; Labadié et Prince, 2008) car elle ne prend pas en compte la question de la cohérence globale du texte.
Divers travaux s'intéressent néanmoins aux marques linguistiques de ruptures thématiques. On pourra citer des courants de recherche intéressés à la notion de segment cadratif, dans la lignée de (Charolles, 1997), ou à la typologie des expressions référentielles (Asher *et al.*, 2006). (Piérard et

2. (Couto J. *et al.*, 2004), par exemple, exploite divers types "d'introducteurs thématiques".
3. Des étudiants en master d'informatique et de sciences du langage n'ayant pas de connaissance *a priori* de la tâche et rémunérés pour l'occasion.
4. `http://www.glozz.org/`

Bestgen, 2006) étudie la manière dont ces deux types de marques peuvent se combiner avec des phénomènes de cohésion lexicale. Dans le domaine du TAL, (Couto J. *et al.*, 2004), exploite divers types "d'introducteurs thématiques" pour contribuer à la segmentation, tandis que (Widlöcher *et al.*, 2006) intègre divers indices linguistiques dans une procédure de fouille de texte.
C'est à cette recherche des marques de ruptures que notre travail se propose de contribuer, en s'appuyant sur la constitution d'un corpus annoté susceptible de les mettre en évidence. Notre modèle est basé sur l'hypothèse que, dans un texte bien construit, les ruptures thématiques brutales sont l'exception plutôt que la règle. Le plus souvent, des zones de transition aident le lecteur à passer d'un thème à l'autre. La zone de transition "parfaite", telle que nous la définissons, devrait comprendre à la fois une conclusion du thème précédent et une introduction du nouveau. Mais, même dans un texte correctement structuré, on peut trouver des transitions incomplètes, voir des frontières abruptes. La tâche d'annotation consiste alors à délimiter quatre types d'unités qui, par souci de simplification, doivent avoir la phrase comme unité atomique :
Les segments thématiques : Même si nous avons supposé que les transitions entre les thèmes sont, en principe, "floues" il convient, pour des raisons pratiques, de définir des frontières. Ces segments représentent les thèmes principaux et, le cas échéant, les sous-thèmes du texte. Par souci de simplification, nous nous sommes limités à un seul niveau de sous-thème.
Les introductions et conclusions thématiques : Elles correspondent à des portions de textes où l'annotateur repère des indices clairs indiquant que l'auteur introduit ou conclut un thème.
Les transitions "floues" : Si l'annotateur perçoit une transition, mais sans qu'une introduction ou une conclusion s'impose.
Nous n'avons volontairement pas donné aux annotateurs d'instructions sur ce que les "indices" de ces zones peuvent être, notre but étant précisément de les identifier.

3 Une mesure d'accord inter-annotateurs pour une tâche de segmentation

3.1 Contexte

Différentes mesures existent pour comparer des segmentations multiples portant sur un même texte. Ainsi, dans (Bestgen, 2009), l'auteur teste comparativement les efficacités de *WindowDiff* et de la distance de Hamming généralisée. Mais de telles mesures sont essentiellement adaptées pour comparer les résultats d'une segmentation automatique à une segmentation de référence. Elles ne peuvent être aisément généralisées à plus de 2 annotations concurrentes. Il existe par ailleurs la mesure classique Kappa (Fleiss, 1971), un standard qui permet de mesurer un taux d'accord à propos de la sémantique d'objets bien définis, mais ne prend pas en considération la notion de segment en tant que telle. Nous proposons de ce fait une nouvelle mesure, conçue pour répondre à notre problème de segmentation multi-annotateurs, directement applicable sur notre corpus.

3.2 La mesure d'accord

Notre mesure vise à proposer une évaluation adéquate des différences entre segments issus d'annotateurs distincts, ces segments ne constituant pas nécessairement un pavage complet du texte (cas des zones de transition par exemple).
Considérons dans un premier temps le cas de deux annotateurs, en supposant qu'une correspondance (alignement) ait déjà été effectuée entre leurs segments respectifs, que ce soit de manière automatique ou manuelle [5]. Lorsque nous comparons deux segments alignés, nous considérons

5. Le problème de l'alignement automatique de plusieurs segmentation est une question à part entière, qui dépasse le cadre de cet article.

les écarts-types (*StDev*) entre les bornes de début et les bornes de fin des deux segments. La somme de ces écart-types est ramenée proportionnellement à la moyenne des tailles des deux segments (*Avg*), par analogie avec le coefficient de variation (StDev/Avg) :

$$x = \frac{\text{StDev}(pos_{\text{init}}) + \text{StDev}(pos_{\text{end}})}{\text{Avg}(size)}$$

. Ce calcul produit une valeur entre 0 et ∞ : 0 correspond à un alignement parfait (positions de début et de fin identiques) ; ∞ est une limite théorique correspondant à des positions divergentes (éloignement infini d'au moins l'une des bornes pour une taille de segment limitée).
Du fait de la difficulté d'interpréter un résultat entre 0 et ∞, ce résultat est projeté sur un intervalle borné. Par analogie avec la mesure κ, nous avons choisi $]-1,1]$. 0, le meilleur des cas, a pour image 1 et ∞, le pire des cas, est projeté sur -1. Nous avons accordé une attention particulière à la valeur $x = 1$. Cette valeur est atteinte lorsque la différence entre les positions est du même ordre de grandeur que la taille des segments elle-même, ce qui correspond, dans notre interprétation, à une limite entre accord et désaccord [6] , et nous conduit à projeter 1 sur la valeur "neutre" 0 de l'intervalle $]-1,1]$. Ces contraintes sont satisfaites par l'utilisation pour la projection d'une fonction monotone, continue, construite à partir de la fonction tangente hyperbolique :

$$y = 1 - 2 \times \sqrt[2]{\frac{e^{\frac{x}{2}} - 1}{e^{\frac{x}{2}} + 1}}$$

. La mesure peut être facilement généralisée à n annotateurs en utilisant toujours les mesures statistiques écarts-types et moyennes entre n segments alignés (un par annotateur). Nous utilisons actuellement une moyenne arithmétique de ces résultats pour l'ensemble des annotations alignées à l'échelle d'un texte dans son intégralité.

3.3 Interprétation, alignement et usage

La projection sur $]-1,1]$ avec par construction le 0 comme limite entre accord et désaccord permet une interprétation plus facile des résultats. Ainsi, toute valeur inférieure à 0 correspond à un désaccord qui sera plus important à mesure que la valeur se rapproche de -1. À l'opposé, les valeurs supérieures à 0 représentent un accord qui ira en s'améliorant en avançant vers 1. Nous considérons que les valeurs supérieures à $0,3$ sont satisfaisantes et deviennent très satisfaisantes au delà de $0,6$. En effet, la fonction de normalisation possède une tangente infinie en 0, des écarts faibles font décroître rapidement les valeurs obtenues [7].
Dans notre expérience, l'alignement a été effectué à la main de telle sorte que chaque segment de chaque annotateur soit aligné avec un segment de chacun des autres annotateurs [8]. Ainsi, si par exemple un annotateur "englobe" deux segments d'un autre annotateur, chacun de ces dernier sera aligné avec le premier. De plus, l'alignement a été fait "en aveugle", c'est-à-dire que la personne chargée d'aligner les segments n'a pris en compte que les position de début et de fin des segments, sans avoir connaissance du texte. Notre objectif n'est donc pas ici d'optimiser la mesure d'accord (comme (Mathet et Widlöcher, 2011)), mais de proposer un alignement continu de tous les annotateurs entre eux, même au détriment du score final.

6. Ce choix découle d'une observation préalable de différentes configurations de segments ; la normalisation de la mesure est donc construite pour reproduire cette interprétation de manière automatique.

7. Par exemple, si sur un segment annoté par dix annotateurs, un seul annotateur fait varier une seule des deux bornes du segment de 10% de la longueur du segment (les autres annotateurs étant en accord parfait), l'accord tombe à $0,83$.

8. Seuls les scores d'accord des segments thématiques sont présentés, le faible nombre de zones de transition rendant l'interprétation d'un tel score sur ces dernières au mieux hasardeuse.

Le tableau 1 présente les scores obtenus en fonction de la catégorie de texte et pour l'ensemble du corpus. Pour chaque texte, ce score est la moyenne sur le texte des scores pour chaque segment (comme précisé en 3.2). Avec des scores allant de 0,16 à 0,7, mais majoritairement entre 0,3

	Petits	Moyens	Grands	Global
Minimum	0,36	0,18	0,16	0,16
Maximum	0,7	0,65	0,4	0,7
Moyenne	0,5	0,41	0,27	0,43
σ	0,11	0,12	0,1	0,13

TABLE 1 – Accord sur la segmentation du corpus

et 0,56, nous pouvons considérer, selon notre "grille de lecture", qu'il existe un accord global satisfaisant sur la structure thématique des textes. Un examen plus précis montre que si l'accord local parfait n'existe pas, la plupart des frontières proposées par les annotateurs ne sont pas éloignées de plus d'un paragraphe. En outre, il n'y a pas de cas de texte avec un désaccord moyen (< 0), ce qui laisse supposer que les lecteurs sont sensibles à des facteurs d'une nature relativement "objective". L'expérience menée sur les zones de transition, présentée plus loin, tend à confirmer ce point.
Une deuxième observation concerne la taille des textes : plus les textes sont petits, plus les scores sont élevés. Une raison évidente est la combinatoire des partitions possibles du texte, qui augmente avec la taille. Un facteur humain peut aussi intervenir, lié à la maintenance de l'attention de l'annotateur, qui peut être source d'erreurs pour les textes longs – au dire des annotateurs eux-mêmes. Enfin, une cause majeure de désaccord concerne la structuration en thèmes/sous-thèmes, i.e. la hiérarchisation des segments produits. Cette dernière conduit en effet certains annotateurs à diviser en plusieurs segments thématiques des portions de texte que d'autres considèrent comme un seul segment, ce qui fait, à dessein, considérablement chuter notre score d'accord.

4 Transitions thématiques

Notre objectif central concerne, rappelons-le, l'étude des transitions entre segments thématiques. Deux séries d'observations peuvent être faites à ce sujet, concernant d'une part leur perception par les annotateurs et d'autre part le repérage d'indices "de surface" ce ces transitions.

4.1 Les transitions thématiques sont bien perçues par les annotateurs

Comme indiqué ci-dessus, les annotateurs ont dû délimiter les segments thématiques mais aussi des introductions, des conclusions et des transitions "floues". Une fois la phase d'annotation terminée, durant le processus de débriefing, nous avons demandé aux cinq annotateurs, entre autres choses, comment ils avaient procédé. La réponse a été claire et unanime : tout d'abord, chercher dans le texte où se situent les *changements de thèmes* ; ensuite, utiliser cette information pour identifier les segments thématiques. Cette façon de procéder leur semblait naturelle, et a confirmé nos hypothèses : les indications de changement de thème sont essentielles pour le lecteur, et clairement perçues par lui.
Les résultats chiffrés vont dans le même sens : 84% des segments thématiques ont une introduction, 39% une conclusion et 16% une zone de transition "floue". Sur l'ensemble des textes, presque tous les changements de thèmes sont donc accompagnés par une zone de transition, une introduction la plupart du temps.

4.2 Indices "de surface" pour les transitions thématiques

La question se pose alors de déterminer si ces zones possèdent des caractéristiques linguistiques spécifiques que le lecteur détecte, et quelles sont-elles. Outre l'intérêt théorique, l'identification de telles caractéristiques est susceptible de fournir des informations utiles pour un segmenteur automatique. Le problème est complexe, impliquant des mécanismes sémantiques et interprétatifs profonds, objet de divers travaux en linguistique comme rappelé dans l'introduction. Mais on peut aussi s'interroger sur la présence de formes *de surface* "reflétant" ces mécanismes. Afin d'étudier cette question nous avons fait les expériences suivantes.

Comme on pouvait s'y attendre, les introductions annotées coïncident presque toujours avec des changements de bloc, où "bloc" signifie "paragraphe" ou "sous-titre" [9]. Ainsi, en observant le corpus, nous avons considéré un certain nombre de traits distinctifs possibles et, pour chacun, nous avons établi une comparaison entre les introductions annotées et le début des paragraphes ne contenant pas d'introduction [10]. Trois hypothèses différentes ont été examinées dont les résultats sont présentés dans le tableau 2.

Les phrases introductives sont plus courtes. L'intuition est qu'elles peuvent opérer comme de véritables titres et transmettre l'information portée par le segment sous forme condensée. Par ailleurs il peut sembler judicieux à l'auteur (de manière consciente ou non) de signaler une nouvelle partie par une rupture de rythme dans le flot textuel. Les résultats montrent une réduction de 25% environ de la longueur des phrases des introductions thématiques par rapport à la première phrase des paragraphes classiques.

Les introductions thématiques aiment les segments détachés. Il est communément admis que les segments détachés en début de phrase ou, mieux, de paragraphe sont des positions thématiques privilégiées (Ho-Dac, 2008). Ces segments sont souvent séparés du reste du texte par une virgule, aussi avons-nous considéré une virgule en position pré-verbale comme une approximation acceptable pour identifier un détachement. Le ratio des détachements dans les introductions est de $0,4$ contre $0,3$ pour les autres paragraphes.

Les introduction thématiques ont des premiers mots préférés. La table 2 donne pour chaque catégorie morpho-syntaxique un ratio de ses occurrences en première position d'introduction thématique, par rapport aux autres paragraphes. On notera que la balise "adj" (adjectif) correspond en fait aux démonstratifs, "adv" (adverbe) généralement aux connecteurs du discours et "pun" (la ponctuation) presque toujours aux guillemets de citation. Les autres abréviations désignent respectivement les conjonctions, les déterminants, les noms, les prépositions, les pronoms et les verbes. Ces résultats suscitent plusieurs remarques : les adverbes (connecteurs de discours) sont sur-représentés, ce qui peut être interprété comme une influence de la structure rhétorique sur la structure thématique ; les conjonctions au contraire, et sans surprise, sont des marques de continuité, de même que les démonstratifs et les pronoms. Les déterminants et les noms sont sur-représentés, probablement en raison de leur fonction de focalisation sur un référent de discours (qui peut être nouveau en cas de changement thématique). Enfin, il n'y a dans notre corpus aucune introduction commençant par une citation.

Commentaires. Ces observations recoupent partiellement certaines hypothèses formulées dans la littérature : par exemple les segments cadratifs vont apparaître dans notre première expérience en tant que constituants détachés et on retrouve l'idée commune selon laquelle les pronoms et les GN démonstratifs sont des marques de continuité. Ces éléments sont toutefois repris dans

9. Dans *Le Monde* un sous-titre n'est généralement pas le début d'une sous-partie, mais plutôt une sorte d'indicateur survenant *en son sein*, de sorte que les annotateurs ne les ont pas nécessairement étiquetés comme introductions.

10. 196 introductions pour 524 paragraphes.

un ensemble plus vaste de "candidats" (pour lesquels un examen linguistique attentif serait intéressant). Une différence est l'accent mis ici sur l'identification de marques de surface là où les études linguistiques proposent (avec pertinence, mais une exploitation informatique moins immédiate) des caractérisations sémantiques. Enfin notons que l'expérience semble plaider pour une certaine spécificité des paragraphes introducteurs de thèmes, ce qui peut questionner l'hypothèse formulée dans (Bestgen et Pirard, 2006) selon laquelle il serait possible d'utiliser de manière fiable le changement de paragraphe pour étudier les ruptures thématiques.

	Introductions	Autres paragraphes	Différence en %
Longueur des phrases	17	24	−23%
Détachement	0,4	0,3	+25%
Premier mot			
adj	0,06	0,08	−25%
adv	0,09	0,05	+45%
con	0,01	0,04	−75%
det	0,4	0,33	+20%
nom	0,21	0,16	+25%
pun	0	0,04	−100%
pre	0,12	0,15	−20%
pro	0,9	0,11	−18%
ver	0,03	0,04	−25%

TABLE 2 – Indices de changements thématiques. Introductions contre les autres paragraphes.

5 Conclusion et travaux futurs

Dans cet article nous avons présenté un ensemble d'expériences relatives à la structuration discursive thématique, expériences appuyées sur la constitution d'un corpus annoté. De fait, le premier résultat positif est, selon nous, ce corpus lui-même, le manque de ressources de ce type étant généralement reconnu [11].

Nous avons également proposé une mesure d'accord multi-annotateurs adaptée à l'évaluation de l'accord entre segmentations multiples. Le défaut de telles mesures rend la constitution de tels corpus difficile et, bien qu'imparfaite (avec un l'alignement qui doit être fait *a priori* notamment), notre mesure à le double avantage d'être facile à mettre en œuvre et à interpréter.

L'accord inter-annotateurs semble suffisant (eu égard au caractère éminemment interprétatif du type de structure considéré) pour en faire un outil de recherche et les observations réalisées semblent confirmer la pertinence de la notion de zone de transition. Différents types d'indices de surface ont été mis en évidence, susceptibles d'aider au repérage de ces zones, donc à la segmentation automatique. L'intégration de ces indices dans une procédure de segmentation automatique à base lexicale pourrait également en constituer une validation "indirecte".

Une autre perspective consiste à étendre le corpus annoté afin de confirmer nos observations et, à terme, autoriser la mise en œuvre de procédures d'apprentissage.

11. Les annotations Glozz sont déportées, et peuvent être obtenues sur simple demande auprès des auteurs, charge au demandeur de posséder les droits sur le corpus du *Monde*

Références

ASHER, N., DENNIS, P. et REESE, B. (2006). Names and pops and discourse structure. *In Workshop on Constraints in Discourse*, pages 11–18, Maynooth.

BEEFERMAN, D., BERGER, A. et LAFFERTY, J. (1999). Statistical models for text segmentation. *In Machine Learning*, volume 34, pages 177–210.

BESTGEN, Y. (2009). Quel indice pour mesurer l'efficacité en segmentation de textes ? *Actes de TALN'09*.

BESTGEN, Y. et PIRARD, S. (2006). Comment évaluer les algorithmes de segmentation automatiques ? essai de construction d'un matériel de référence. *Actes de TALN'06*.

CHAROLLES, M. (1997). L'encadrement du discours : univers, champs, domaines et espaces. *Cahier de Recherche Linguistique*, 6:1–73.

CHOI, F. Y. Y. (2000). Advances in domain independent linear text segmentation. *Proceeding of NAACL-00*, pages 26–33.

COUTO J., Ferret, O., GRAU, B., HERNANDEZ, N., JACKIEWICZ, A., MINEL, J. et PORHIEL, S. (2004). Régal, un système pour la visualisation sélective de documents. *Revue d'Intelligence Artificielle*, 18(4):481–514.

FLEISS, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

HEARST, M. A. (1997). Texttiling : Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics.*, 23(1):33–64.

HO-DAC, L.-M. (2008). Exploring discourse organization through theme position. *19th European Systemic Functional Linguistics Conference and Workshop (ESFLCW) : Data and interpretation*.

LABADIÉ, A. et PRINCE, V. (2008). Finding text boundaries and finding topic boundaries : two different task ? *Proceedings of GoTAL 2008*.

MATHET, Y. et WIDLÖCHER, A. (2011). Une approche holiste et uniée de l'alignement et de la mesure d'accord inter-annotateurs. *In TALN 2011*, volume 1, pages 247–258, Montpellier, France.

MORRIS, J. et HIRST, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:20–48.

PÉRY-WOODLEY, M.-P., ASHER, N., BENAMARA, F., BRAS, M., ENJALBERT, P., FABRE, C., FERRARI, S., HO-DAC, L.-M., LE DRAOULEC, A., MATHET, Y., MULLER, P., PRÉVOT, L., REBEYROLLE, J., VERGEZ-COURET, M., VIEU, L. et WIDLÖCHER, A. (2009). Annodis : une approche outillée de l'annotation de structures discursives en corpus. *Actes de TALN'09*.

PIÉRARD, S. et BESTGEN, Y. (2006). Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de textes. *TAL*, 2(47):89–110.

WIDLÖCHER, A., BILHAUT, F., HERNANDEZ, N., RIOULT, F., CHARNOIS, T., FERRARI, S. et ENJALBERT, P. (2006). Une approche hybride de la segmentation thématique : collaboration du traitement automatique des langues et de la fouille de texte. *In DEFT'06*, Fribourg.

WIDLÖCHER, A. et MATHET, Y. (2009). La plate-forme glozz : environnement d'annotation et d'exploration de corpus. *Actes de TALN'09*.

La "multi-extraction" comme stratégie d'acquisition optimisée de ressources (non) terminologiques

Blandine Plaisantin Alecu[1], Izabella Thomas[2], Julie Renahy[1]
(1) Prolipsia SAS, TEMIS Innovation, 18 r Alain Savary, 25000 Besançon
(2) Centre Tesnière, Université de Franche-Comté, UFR SLHS, 30 r Mégevand, 25030 Besançon
{blandine.alecu, julie.renahy}@prolipsia.com,
izabella.thomas@univ-fcomte.fr

RÉSUMÉ

A partir de l'évaluation d'extracteurs de termes menée initialement pour détecter le meilleur outil d'acquisition du lexique d'une langue contrôlée, nous proposons dans cet article une stratégie d'optimisation du processus d'extraction terminologique. Nos travaux, menés dans le cadre du projet ANR Sensunique, prouvent que la « multi-extraction », c'est-à-dire la coopération de plusieurs extracteurs de termes, donne des résultats significativement meilleurs que l'extraction via un seul outil. Elle permet à la fois de réduire le silence et de filtrer automatiquement le bruit grâce à la variation d'un indice relatif au potentiel terminologique.

ABSTRACT

Multi-extraction as a strategy of optimized extraction of terminological and lexical resources

Based on the evaluation of terminological extractors, initially to find the best tool for building a controlled language lexicon, we propose a strategy of optimized extraction of terminological resources. Our work highlights that the cooperation of several extraction tools gives better results than the use of a single one. It both reduces silence and automatically filters noise thanks to a variable related to termhood.

MOTS-CLÉS : terminologie, extraction, langue contrôlée, potentiel terminologique, filtrage de termes.
KEYWORDS : terminology, extraction, controlled language, termhood, term filtering.

1 Introduction

1.1 Contexte, problématique et objectifs

Le présent article fait état de recherches effectuées dans le cadre du projet Sensunique (ANR-EMMA-2010-039), qui fait suite au projet LiSe (ANR-06-SECU-007) dans lequel nous avions conçu une méthodologie de contrôle de la langue française et développé un prototype expérimental d'aide à la rédaction en Langue Contrôlée (LC). Une des ambitions du projet Sensunique est d'alléger et de fiabiliser la tâche de conception d'une LC par un linguiste, en mettant à sa disposition des outils de recensement du Lexique d'une LC (désormais LLC) ; c'est sur ce point que nous concentrons cet article.

Dans notre approche, une LC doit être circonscrite à un domaine et à un environnement de rédaction précis, c'est-à-dire pour un public restreint et pour un type de textes particulier ; on peut dès lors parler d'une LC « sur mesure » ; la précision des facteurs

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 511–518,
Grenoble, 4 au 8 juin 2012.

influençant le texte à générer, et donc de la LC sous-jacente, différencie notre approche (Renahy et al., 2009) de celles des travaux menés sur le français[1]. En outre, elle est issue d'une analyse de corpus (généralement de petite taille), lequel doit recenser l'ensemble des textes en vigueur pour l'activité et le public concernés.

A notre connaissance, aucun outil dédié au recensement d'un LLC n'existe à ce jour. Cependant, puisque l'acquisition d'un LLC peut être comparée jusqu'à un certain point à l'acquisition terminologique, nous avons choisi de nous appuyer sur les Extracteurs de Termes (EdT), tout en tentant d'améliorer leurs résultats pour qu'ils répondent à nos besoins. Plus précisément, nous avons évalué le bénéfice que nous pourrions tirer de la coopération de plusieurs EdT. De multiples travaux fondés sur la coopération d'outils ont démontré son intérêt : en premier pour la reconnaissance vocale avec le système ROVER (Fiscus, 1997), repris, pour n'en citer que quelques uns[2], pour des analyseurs syntaxiques (Brunet-Manquat, 2004) ou des étiqueteurs morphosyntaxiques (Serp, 2008). Mais ce principe n'a jamais été appliqué, à notre connaissance, aux EdT.

1.2 LLC et Ressource Terminologique (RT)

Les notions de RT, résultat de l'acquisition terminologique, et LLC se recoupent et se distinguent à la fois. La principale différence concerne leurs périmètres respectifs. Parce que tous deux visent à couvrir un domaine particulier, renvoyant à des concepts spécifiques[3], RT et LLC recensent chacun des termes. Mais le périmètre d'une RT s'arrête aux unités terminologiques spécifiques du domaine alors qu'un LLC doit contenir l'ensemble du lexique nécessaire à la rédaction d'un texte dans sa globalité. Møller *et al.* (2006) parlent de « mots » (référant alors à des unités monolexémiques comme multilexémiques) afin de ne pas confondre les unités d'un LLC avec des unités terminologiques. Nous choisissons, quant à nous, de considérer comme Unités Lexicales[4] (UL) toutes les unités d'un LLC.

Pour être exhaustif, un LLC doit contenir des UL non spécifiques (par extension, non terminologiques). On ne peut pas appliquer une dichotomie spécifique-non spécifique : il s'agit, comme le dit Camlong (1996) d'un continuum allant du vocabulaire terminologique du domaine au vocabulaire général. Un texte écrit en LC peut inclure différents types d'UL, illustrés ici avec des exemples de protocoles d'immunobiologie : les termes du domaine (*anticorps monoclonaux*) ; ceux d'un autre domaine (*fenêtres informatiques*) ; les UL du lexique général prenant un sens spécifique dans le domaine traité (*population* (*bactérienne*)) ; celles du lexique général rentrant dans la composition de termes (*anticorps de chèvre*) ou gardant leur sens courant (*échantillons divers*) et les mots grammaticaux. Une RT ne suffit donc pas à l'exhaustivité d'un LLC.

Puisqu'un LLC ne contient pas que des UL terminologiques, nous avons imaginé une

[1] Dont l'unique exemple est Le Français Rationalisé, du GIFAS (1990).

[2] Voir Brunet-Manquat (2004) pour une liste plus exhaustive.

[3] En admettant que ces concepts sont dénommés par des termes.

[4] Nous reprenons ici la notion d'unité lexicale telle que définie par (L'Homme, 2005)

stratégie de mise en exergue du statut terminologique des candidats, extraits par les EdT, basée sur leur potentiel terminologique (*termhood*), c'est-à-dire le degré de spécialisation de leur sens dans le domaine à l'étude (Kageura et Umino, 1996).

1.3 Hypothèse

Nous posons l'hypothèse (H1) que l'utilisation simultanée de plusieurs EdT (que nous appellerons désormais la « multi-extraction ») est plus profitable (qu'un seul) au recensement du LLC, hypothèse que l'on peut subdiviser en :

- (H1.1) : les résultats proposés par plusieurs EdT sont les candidats-termes (CT) les plus pertinents, et peuvent être considérés comme UL terminologiques (la multi-extraction permet de déterminer le statut terminologique d'une UL en faisant ressortir son potentiel terminologique).
- (H1.2) : les CT non valides sont des UL candidates non terminologiques potentiellement pertinentes (le bruit des EdT, dans leur fonction initiale de recensement des termes, peut diminuer le silence, dans la fonction détournée de recensement des UL d'un LLC). Si cela est confirmé, la tâche de recensement d'un LLC peut être organisée, en classant les résultats par poids terminologique[5] (Pt) et/ou comme aide au filtrage du bruit pour l'acquisition de RT.

2 Matériel et méthodes

2.1 Corpus et lexique de référence

Nous avons constitué un corpus de référence de 14 modes opératoires d'immunobiologie (10 064 mots) de l'Établissement Français du Sang (EFS) Bourgogne Franche-Comté[6]. Notons que la méthodologie pensée est indépendante du domaine. Nous avons construit manuellement sur la base de ce corpus un LLC de référence, grâce à des critères linguistiques, la consultation de ressources terminologiques[7] et d'experts métier[8]. Le lexique de référence obtenu contient 1 512 UL (lemmes) pour 1 729 formes fléchies (utilisées en corpus), 7 catégories syntaxiques fonctionnelles distinctes (distinction minimale nécessaire pour un LLC : Adjectif, Adverbe, Nom, Nom propre, Verbe au participe passé, Verbe au participe présent, Verbe hors participes), 92 matrices morphosyntaxiques distinctes (exemple : *Nom Prep Det Nom Prep Det Nom Prep Nom* pour *fraction de l'immunoglobuline de l'antisérum de lapin*) et 2 statuts lexico-terminologiques distincts (terminologique et général).

2.2 Pré-sélection des EdT comme outils de recensement du LLC

Nous avons établi les critères suivants, afin d'estimer l'utilisabilité et l'adéquation

[5] Pt est un indice de fiabilité d'une ULC en tant que terme (relatif à son potentiel terminologique).

[6] Documents décrivant le déroulement détaillé et structuré des différentes étapes d'une manipulation.

[7] Le Grand Dictionnaire Terminologique, Termium Plus, le dictionnaire médical Masson 5ème édition.

[8] EFS Bourgogne Franche-Comté, partenaire Santé dans le projet Sensunique.

technique des EdT à nos besoins et de nous limiter à 3 EdT (coût raisonnable de la tâche d'évaluation) : langue (français), méthode (non purement statistique [9] : linguistique ou hybride), disponibilité (de suite), licence (libre ou commerciale : dans ce cas, coût faible ou nul), maturité de l'outil (non prototype), environnement informatique (Unix), modalité d'exécution (service web ou appel en ligne de commande), temps d'exécution (respectant le seuil d'appel en web service) et domaine d'application (non spécifique).

Ces critères nous ont menés aux EdT Acabit (Daille, 1994), TermoStat (Drouin, 2003) et YaTeA (Aubin et al., 2006). Acabit procède par identification de groupes nominaux complexes sur des matrices syntagmatiques pour extraction de bi-termes, regroupement de variantes (à partir de ces bi-termes) puis filtrage statistique. YaTeA enchaîne identification de groupes nominaux à partir de frontières morphosyntaxiques, calcul de leurs structures en tête et modifieur, puis exploitation de ces structures pour l'analyse des groupes nominaux restants. Enfin, TermoStat fonctionne par détection de CT sur patrons morphosyntaxiques puis pondération et filtrage selon la spécificité de chaque CT (méthode de mise en opposition de corpus spécialisés et non spécialisés). En outre, YaTeA et TermoStat ont l'avantage d'extraire des termes simples en plus des termes complexes ; et TermoStat est le seul à extraire également des termes non nominaux.

2.3 Evaluation

Deux des tâches du linguiste lors de la conception d'un LLC ont été évaluées :

1. Recensement des UL (de l'ensemble des UL d'un LLC) ;
2. Recensement des termes (des UL de statut terminologique d'un LLC).

Pour chacune de ces tâches, nous avons procédé à 3 expérimentations :

1. Évaluation des résultats de chaque EdT pris séparément ;
2. Évaluation des résultats cumulés de tous les EdT (union) ;
3. Évaluation des résultats consolidés, ou communs (intersection).

Pour chaque évaluation, nous avons calculé les mesures suivantes :

- Précision : P = (formes extraites correctes) / (formes extraites) ;
- Rappel : R = (formes extraites correctes) / (formes de référence) ;

2.3.1 Appariement des résultats

Notre objectif étant d'estimer la capacité des EdT à recenser les UL (terminologiques ou non) et non leur capacité de lemmatisation ou de variation terminologique, nous avons opté pour comparer les formes fléchies. Contrairement à Hamon (2000), nous ne cherchons pas les différents types d'erreurs, mais évaluons la présence d'une forme extraite dans le lexique de référence. Pour ne pas comparer les regroupements opérés par les EdT, nous avons extrait une liste des formes fléchies de tous les CT. Nous avons ainsi pu apparier les formes fléchies candidates avec celles du lexique de référence.

[9] A cause de la taille estimée des corpus utilisés pour concevoir une LC.

3 Résultats et discussion

3.1 Tâche 1 : Recensement des UL

Expérimentations	Outil(s)	P	R
Résultats d'un EdT	TermoStat	64 %	40 %
	YaTeA	43 %	**52 %**
	Acabit	44 %	17 %
Résultats cumulés (union)	TermoStat ∪ YaTeA	44 %	68 %
	TermoStat∪ ACABIT	55 %	48 %
	YaTeA ∪ ACABIT	41 %	59 %
	TermoStat ∪ YaTeA ∪ ACABIT	**42 %**	**72 %**
Résultats communs (intersection)	TermoStat ⋂ YaTeA	**74 %**	22 %
	TermoStat ⋂ ACABIT	63 %	9 %
	YaTeA ⋂ ACABIT	62 %	11 %
	(TermoStat ⋂ YaTeA) ou (TermoStat ⋂ ACABIT) ou (YaTeA ⋂ ACABIT)[10]	**69 %**	29 %

TABLE 1 – Tâche de recensement des UL

La première expérimentation montre que, dans le meilleur des cas, en utilisant un seul EdT, 52 % (valeur en gras, Table 1) des UL du LLC sont recensées, ce qui est loin de satisfaire le critère d'exhaustivité.

Pour la deuxième expérimentation, le cumul des résultats des 3 EdT permet de couvrir quasiment ¾ du lexique de référence (rappel de 72 %, en gras, Table 1). Ceci confirme l'hypothèse H1 : la multi-extraction permet de mieux couvrir le LLC que l'utilisation d'un seul EdT. En revanche, dans ce cas, il reste à filtrer manuellement près de 60 % des propositions et il devient nécessaire de filtrer automatiquement le bruit.

La troisième expérimentation démontre que la combinaison d'EdT obtenant la meilleure précision est TermoStat + YaTeA (74 %, Table 1). Cependant, il apparaît également que n'importe quelle combinaison de 2 EdT donne une précision de 69 % (donc légèrement plus faible). Nous proposons de filtrer le bruit sur cette dernière combinaison en considérant que ce cas de figure sera plus généralisable (à d'autres domaines) dans la mesure où il « suffit » qu'une UL soit proposée par 2 EdT pour être estimée pertinente. L'opération consisterait à augmenter la valeur d'un indice relatif au potentiel terminologique des UL concernées (proposées par 2 EdT), et creuser ainsi l'écart avec celles qui ne sont pas proposées que par un EdT. Cela revient à distinguer les ULC à fort potentiel terminologique de celles à faible potentiel terminologique en les

[10] Sur l'ensemble des résultats communs à (proposés par) au moins 2 EdT, quels qu'ils soient.

classant et non en supprimant ces dernières.

3.2 Tâche 2 : Recensement des termes

Expérimentations	Outil	P	R
Résultats d'un EdT	TermoStat	28 %	52 %
	YaTeA	16 %	58 %
	ACABIT	14 %	17 %
Résultats cumulés (union)	TermoStat ∪ YaTeA	16 %	76 %
	TermoStat∪ ACABIT	23 %	60 %
	YaTeA ∪ ACABIT	14 %	63 %
	TermoStat ∪ YaTeA ∪ ACABIT	15 %	**79 %**
Résultats communs (intersection)	TermoStat ⋂ YaTeA	**37 %**	33 %
	TermoStat ⋂ ACABIT	24 %	11 %
	YaTeA ⋂ ACABIT	26 %	14 %
	(TermoStat ⋂ YaTeA) ou (TermoStat ⋂ ACABIT) ou (YaTeA ⋂ ACABIT)	31 %	39 %
	TermoStat ⋂ YaTeA ⋂ ACABIT	32 %	9 %

TABLE 2 – Tâche de recensement des termes

La mesure de précision de 37 % (TermoStat ⋂ YaTeA, Table 2) pour les résultats communs permet de valider l'hypothèse H1.1. : la multi-extraction aide à déterminer le statut terminologique d'une UL en faisant ressortir son potentiel terminologique. La différence (même faible) de rappel entre les résultats cumulés des 3 EdT pour le recensement des termes (79 %, Table 2) et le recensement des UL (72 %, Table 1) démontre qu'une partie des candidats proposés ne sont pas des termes mais sont, pour le LLC, des UL correctes, de statut non terminologique (hypothèse H1.2). Bien que les résultats soient moindres que ceux escomptés, ils demeurent satisfaisants et il est possible qu'ils soient meilleurs sur des corpus plus conséquents.

En résumé, cumuler les résultats de tous les EdT permet de couvrir 79 % des termes (rappel TermoStat ∪ YaTeA ∪ ACABIT, Table 2), et le meilleur moyen d'aider à déterminer le statut d'une UL est, non pas de se baser sur les résultats communs aux 3 EdT (contrairement à ce que nous attendions), mais de se baser sur les résultats communs aux 2 EdT TermoStat et YaTeA (précision de 37 % dans la Table 2). Ceci valide tout de même l'hypothèse selon laquelle la multi-extraction aide à recenser et à organiser la validation d'un LLC.

3.3 Stratégie basée sur les observations

Le fait que la multi-extraction permette à la fois de réduire le silence et le bruit des propositions nous incite à introduire un indice, relatif au potentiel terminologique,

variable en fonction des résultats des EdT. Nous proposons d'attribuer à chaque UL candidate un poids terminologique Pt ; puis de faire varier ce Pt initialement nul en fonction des résultats de chaque EdT. Nous proposons une stratégie de variation du Pt consistant à augmenter du Pt d'une UL en fonction du nombre d'EdT qui la proposent comme candidate. Ce principe traduit bien les faits suivants :

- un EdT propose un candidat « terme » donc ayant un potentiel terminologique ;
- un candidat a d'autant plus de probabilité d'être un terme fiable qu'il y a d'EdT le proposant (comme candidat) ;
- les résultats pourront être classés et validés selon la valeur du Pt.

Notons que les 3 EdT utilisés intègrent *a priori* (TermoStat) ou *a posteriori* (Acabit et YaTeA) des indices statistiques afin de cerner les termes les plus pertinents. Bien qu'il puisse être intéressant de coupler les valeurs de ces indices (différents pour chaque candidat) au Pt, nous avons fait le choix de ne pas le faire expressément, afin de ne pas rendre l'algorithme de pondération dépendant des EdT utilisés (et puisque le calcul de Pt repose déjà indirectement sur l'efficacité des EdT utilisés).

4 Conclusion

L'exploitation des résultats des expérimentations menées nous a permis de proposer une méthode d'acquisition d'un LLC et d'optimisation de l'acquisition terminologique. Elle repose sur la coopération de plusieurs EdT et permet de faire ressortir le potentiel terminologique des candidats, de réduire le silence obtenu avec un seul EdT et de filtrer le bruit en classant les candidats sur un indice de potentiel terminologique.

Outre concevoir un outil dédié au recensement d'un LLC, l'originalité de ces travaux réside dans le fait que nous proposons de faire coopérer plusieurs EdT pour améliorer leurs résultats et mettre en place un système de filtrage, alors que les travaux antérieurs d'évaluation d'EdT visaient leur mise en opposition (ou classement) (Grabar, 2004).

Nous avons mis au point, pour améliorer l'extraction terminologique, un système à base de vote, sur la méthode dite du "vote à la majorité" (Brunet-Manquat, 2004) où plus un terme est proposé par différents EdT, plus sa fiabilité est renforcée.

Nous avons conçu une plateforme implémentant cette méthode. Elle intègre les étiqueteurs morphosyntaxiques Brill[11] et TreeTagger, le lemmatiseur Flemm (Namer, 2000) pour les analyses préalables et nécessaires à l'extraction, et les EdT Acabit, TermoStat et YaTeA. Elle permet de procéder à l'extraction et à l'organisation de lexique terminologique et non-terminologique à partir d'un corpus français au format XML TEI P5. La plateforme est paramétrée par défaut sur le principe du "vote à la majorité" mais l'utilisateur peut ajuster le poids attribué à chaque EdT, en fonction de ses besoins, afin de rendre cette plateforme aussi flexible que possible. Nous avons également intégré un module d'interrogation de ressources terminologiques ou lexicales existantes, ce qui permet de renforcer, une nouvelle fois, la fiabilité du potentiel

[11] Avec le lexique et le fichier de règles fournis par l'ATILF-CNRS, de Nancy.

terminologique des candidats.

Références

AUBIN, S. et HAMON, T. (2006). Improving Term Extraction with Terminological Resources. *In : Advances in Natural Language Processing, 5th International Conference on NLP (FinTAL'2006),* Springer, 2006, p. 380-387.

BRUNET-MANQUAT, F. (2004). Fusionner pour mieux analyser : Conception et évaluation de la plate-forme de combinaison. *In Actes de TALN-2004 (Traitement automatique des langues naturelles).* Fez, Maroc, 19-22 avril 2004. vol. 1/1, p. 111-120.

CAMLONG, A. (1996). Méthode d'analyse lexicale textuelle et discursive, Paris, Orphrys.

DAILLE, B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. *In : The Balancing Act: Combining Symbolic and Statistical Approaches to Language. Workshop at the 32nd Annual Meeting of the ACL (ACL'94),* Las Cruces, New Mexico, USA.

DROUIN, P. (2003). Term Extraction Using non-Technical Corpora as Point of Leverage. *In : Terminology,* vol.9, n°1, *John Benjamins Publishing Company: Amsterdam/Philadelphia*, p. 99-115.

FISCUS, J.G. (1997), A post-processing system to yield reduced error word rates: Recognizer output voting error reduction (ROVER). *In IEEE Workshop on Automatic Speech Recognizer and Understanding,* p. 347-354.

GRABAR N. (2004). Terminologie médicale et morphologie : Acquisition de ressources morphologiques et leur utilisation pour le traitement de la variation terminologique, Thèse de Doctorat en Informatique Médicale, Université Paris 6.

HAMON, T. (2000). Variation sémantique en corpus spécialisé : Acquisition de relations de synonymie à partir de ressources lexicales, Thèse de Doctorat en Informatique, Université Paris Nord.

KAGEURA, K. et UMINO, B. (1996). Methods for automatic term recognition: A review. *In : Terminology*, 3(2), p. 259–289.

L'HOMME, M.-C. (2005). Sur la notion de terme. *In Meta : journal des traducteurs*, vol. 50, n° 4, p. 1112-1132.

MØLLER, M. H., CHRISTOFFERSEN, E., HANSEN, M. (2006). Building a Controlled Language Lexicon for Danish. *In LSP and Professional Communication*, vol. 6, Nr. 1, p. 12-38.

NAMER, F. (2000). FLEMM : un analyseur flexionnel du français à base de règles. *In Traitement Automatique des Langues* ;vol. 41/2, p. 523-547.

RENAHY, J., DEVITRE, D., THOMAS, I., DZIADKIEWICZ, A., (2009). Controlled language norms for the redaction of security protocols: finding the median between system needs and user acceptability. *In Proceedings of the 11th International Symposium on Social Communication, Santiago de Cuba,* Cuba, 19-23 January 2009, p. 289-293.

Une approche de recherche d'information structurée fondée sur la correction d'erreurs à l'indexation des documents

Arnaud Renard[1, 2] Sylvie Calabretto[1, 2] Béatrice Rumpler[1, 2]
(1) Université de Lyon, CNRS
(2) INSA-Lyon, LIRIS, UMR 5205, F-69621 Villeurbanne Cedex
arnaud.renard@insa-lyon.fr, sylvie.calabretto@insa-lyon.fr,
beatrice.rumpler@insa-lyon.fr

RÉSUMÉ

Dans cet article, nous nous sommes intéressés à la prise en compte des erreurs dans les contenus textuels des documents XML. Nous proposons une approche visant à diminuer l'impact de ces erreurs sur les systèmes de Recherche d'Information (RI). En effet, ces systèmes produisent des index associant chaque document aux termes qu'il contient. Les erreurs affectent donc la qualité des index ce qui conduit par exemple à considérer à tort des documents mal indexés comme non pertinents (resp. pertinents) vis-à-vis de certaines requêtes. Afin de faire face à ce problème, nous proposons d'inclure un mécanisme de correction d'erreurs lors de la phase d'indexation des documents. Nous avons implémenté cette approche au sein d'un prototype que nous avons évalué dans le cadre de la campagne d'évaluation INEX.

ABSTRACT

Structured Information Retrieval Approach based on Indexing Time Error Correction

In this paper, we focused on errors in the textual content of XML documents. We propose an approach to reduce the impact of these errors on Information Retrieval (IR) systems. Indeed, these systems rely on indexes associating each document to corresponding terms. Indexes quality is negatively affected by those misspellings. These errors makes it difficult to later retrieve documents (or parts of them) in an effective way during the querying phase. In order to deal with this problem we propose to include an error correction mechanism during the indexing phase of documents. We achieved an implementation of this spelling aware information retrieval system which is currently evaluated over INEX evaluation campaign documents collection.

MOTS-CLÉS : Recherche d'information, dysorthographie, correction d'erreurs, xml.

KEYWORDS: Information retrieval, misspellings, error correction, xml.

1 Introduction

Les documents produits dans un cadre professionnel doivent satisfaire à un niveau minimum de qualité et font l'objet de multiples cycles de relecture et correction permettant d'y parvenir. Cela constituait auparavant le principal mode de production d'informations néanmoins cette pratique a fortement évolué et à l'échelle d'Internet, il s'agit désormais d'un mode de production de l'information qui peut être considéré comme marginal. En effet, la plupart des documents sont créés par des utilisateurs hors de tout cadre professionnel. Ces derniers sont donc davantage

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 519–526,
Grenoble, 4 au 8 juin 2012.

susceptibles de commettre des erreurs en employant un lexique qu'ils ne maîtrisent pas toujours et qui peut s'avérer inadapté au sujet traité. Par ailleurs, le contenu publié sur Internet n'est pas soumis à un contrôle de qualité : les blogs ont popularisé l'auto-publication de masse à la fois gratuite et immédiatement disponible. Il est donc légitime dans ce cas d'émettre des réserves sur la qualité des documents et autres informations produits dans ce cadre (Subramaniam *et al.*, 2009). Les systèmes de RI constituent les principaux points d'accès aux informations d'Internet. Ils sont affectés par les erreurs (Kantor et Voorhees, 2000) dont la correction constitue un axe d'amélioration important qu'il convient d'étudier (Varnhagen *et al.*, 2009).

Dans la section 2 nous présenterons la RI dans les documents (semi-)structurés XML ainsi que les travaux tentant de mêler RI et correction d'erreurs. Dans la section 3, nous présenterons notre approche intégrant la gestion de la correction des erreurs durant la phase d'indexation des documents. Nous analyserons les résultats de l'évaluation de notre système de RI sans et avec prise en charge des erreurs sur la campagne d'évaluation INEX dans la section 4. Enfin, nous conclurons et nous présenterons nos perspectives d'évolution en section 5.

2 Contexte général et positionnement

2.1 Recherche d'information structurée

Les documents XML constituent un des formats de diffusion de l'information les plus répandus sur internet. Nous allons dans un premier temps modéliser ces documents dont la structure explicite est plus complexe que de simples documents textuels "plats". Un document XML structuré ds peut être représenté par un arbre dans lequel on peut distinguer 3 types de nœuds différents : les nœuds feuilles nf_i représentant le contenu textuel, les nœuds internes ni_i correspondant aux éléments ainsi que leurs attributs na_i.

```
<?xml version="1.0" encoding="UTF−8"?>
<article>
  <name id="1337">Lorem...</name>
  <body>
    ...ipsum dolor sit amet,
    <emph>consectetur</emph>
    adipiscing elit.
  </body>
</article>
```

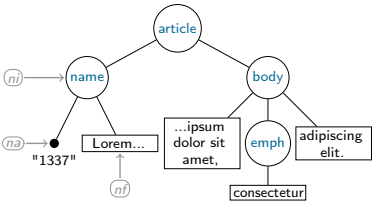


FIGURE 1 – Document XML (à gauche) et sa représentation arborescente (à droite).

Les informations textuelles sont présentes principalement dans les nœuds feuilles qui sont les nœuds à indexer en priorité et qui constitueront le niveau de granularité le plus fin de notre système de RI. Il diffère en cela des systèmes de RI classiques dont la granularité correspond au document. Plusieurs approches de la littérature (Kamps *et al.*, 2009) permettent la prise en compte de cette granularité plus fine mais aussi de la structure des documents. Nous proposons de nous appuyer sur une adaptation du modèle vectoriel de (Salton, 1971) ainsi que sur l'approche employée par XFIRM (Sauvagnat et Boughanem, 2005) qui introduit une méthode de propagation de la pertinence au travers de la structure des documents.

2.1.1 Pondération des nœuds feuilles (orientée contenu)

Lors de l'évaluation d'une requête le score relatif à la pertinence des nœuds feuilles est calculé directement, tandis que les scores des nœuds internes sont propagés dynamiquement à partir des nœuds feuilles à travers l'arborescence du document. Cela permet de retourner une liste ordonnée des nœuds (sous-arbres) les plus pertinents pour la requête.

Le score d'un nœud feuille s_{nf} vis-à-vis d'une requête textuelle r composée d'une séquence de n termes (ou mots-clés) $t_1, ..., t_n$ se calcule selon la formule suivante :

$$s_{nf}(r) = \sum_{i=1}^{n} p_{t_i}^{r} \times p_{t_i}^{nf} \tag{1}$$

dans laquelle, $p_{t_i}^{r}$ et $p_{t_i}^{nf}$ sont respectivement les poids du i-ème terme t_i dans la requête r (évalué lors de l'interrogation), et dans le nœud feuille nf (évalué lors de l'indexation). Afin d'adapter le modèle vectoriel de Salton aux documents XML structurés, nous avons choisi un système de pondération qui reflète l'importance locale des termes dans les nœuds feuilles (tf) [1] et globale dans les documents (idf) [2] ainsi que dans les éléments (ief) [3] de la collection.

$$p_{t_i}^{r} = tf_{t_i}^{r} \qquad p_{t_i}^{nf} = tf_{t_i}^{nf} \times idf_{t_i} \times ief_{t_i} \tag{2}$$

où, $tf_{t_i}^{r}$ et $tf_{t_i}^{nf}$ sont respectivement la fréquence du terme t_i dans la requête r et dans le nœud feuille nf. La fréquence correspond au nombre d'occurences du terme t_i respectivement dans la requête r (dénoté par $|t_i^r|$) et dans nf (dénoté par $|t_i^{nf}|$), divisé par le nombre de termes respectivement dans la requête r (dénoté par $|r|$) et dans le nœud feuille nf (dénoté par $|nf|$).

$$tf_{t_i}^{r} = \frac{|t_i^r|}{|r|} \qquad tf_{t_i}^{nf} = \frac{|t_i^{nf}|}{|nf|} \tag{3}$$

et, idf_{t_i} (resp. ief_{t_i}) représente la fréquence inverse du terme t_i dans les documents (resp. les nœuds feuilles). $|D|$ (resp. $|NF|$) est le nombre total de documents (resp. nœuds feuilles) de la collection et $|d_{t_i}|$ (resp. $|nf_{t_i}|$) le nombre de documents (resp. nœuds feuilles) qui contiennent le terme t_i.

$$idf_{t_i} = log\left(\frac{|D|}{|d_{t_i}|+1}\right) + 1 \qquad ief_{t_i} = log\left(\frac{|NF|}{|nf_{t_i}|+1}\right) + 1 \tag{4}$$

2.1.2 Pondération des nœuds internes (orientée structure)

Lorsqu'un nœud feuille est pertinent vis-à-vis d'une requête, les nœuds internes ancêtres de ce dernier le sont également dans une certaine mesure du fait qu'ils englobent ce dernier. Le score des nœuds feuilles peut ainsi être propagé de proche en proche à leurs nœuds ascendants (selon une fonction d'agrégation) jusqu'au nœud racine qui représente le document dans son intégralité.

$$s_{ni}(r) = |NF_{ni}^{s_{nf}(r)>0}|. \sum_{nf_k \in NF_n} \alpha^{dist(ni,nf_k)-1} \times s_{nf_k}(r) \tag{5}$$

1. tf : term frequency (fréquence du terme dans un contexte : requête, élément, ou document).
2. idf : inverse document frequency (fréquence inverse du terme dans les documents).
3. ief : inverse element frequency (fréquence inverse du terme dans les éléments).

où, α compris dans l'intervalle $[0..1]$ représente le facteur d'atténuation de l'importance du nœud feuille nf_k vis-à-vis du nœud interne ni et $dist(ni, nf_k)$ représente la distance entre le nœud interne ni et le nœud feuille nf_k dans la structure arborescente du document. Ainsi, les termes qui apparaissent près de la racine d'un sous-arbre sont plus pertinents pour l'élément racine que ceux qui apparaissent à un niveau plus profond du sous-arbre.

Et $|NF_{ni}^{s_{nf}(q)>0)}|$ représente le nombre de nœuds feuilles du nœud interne qui sont pertinents car un nœud qui contient plus de nœuds pertinents peut être considéré comme plus pertinent.

En présence d'erreurs, le calcul des scores des nœuds feuilles et notamment le facteur $p_{t_i}^{nf}$ de la formule 1 est impacté car le $tf_{t_i}^{nf}$ (cf. formule 2) est amoindri voire annulé dans certain cas ce qui diminue la pertinence du nœud. Il est donc important de considérer la correction des erreurs.

2.2 Correction des erreurs dans les systèmes de RI

La plupart des approches de correction d'erreurs associées aux systèmes de RI considèrent uniquement la correction des requêtes. tels que les travaux de (Sitbon *et al.*, 2007), ou encore le "Did you mean..." introduit par Google qui n'est donc pas adapté. Certains des travaux liés à la campagne d'évaluation TREC [4] considèrent la correction des documents.

La campagne TREC-5 Confusion track a rendu disponibles différentes versions d'une collection de plus de 55000 documents contenant respectivement des taux d'erreurs de 0%, 5%, et 20%. L'article de synthèse de la campagne (Kantor et Voorhees, 2000) présente les différentes approches pour la gestion des erreurs suivies par 5 des participants. Néanmoins, ils ont pu constater une dégradation des performances de tous les systèmes de RI en présence de documents corrompus contenant des erreurs essentiellement dues à la non correspondance entre les termes de la requête et les termes par lesquels les documents ont été indexés. Le même phénomène d'augmentation des silences à l'interrogation et de perte de précision même à de faibles taux de corruption des documents (taux d'erreurs de 3%) a été observé par (Ruch, 2002). La campagne TREC-6 Spoken document retrieval track (Voorhees *et al.*, 2000) considère des documents issus de transcriptions de même que (Gravier *et al.*, 2011).

Dans le cadre de TREC-5, trois systèmes s'appuient sur l'expansion de requêtes en y ajoutant des versions altérées des termes qui la composaient. Cela présente l'inconvénient d'introduire du bruit supplémentaire dans les résultats du système de RI lorsque le nombre de variations des termes de la requête ajoutées à la requête initiale augmente. Deux autres systèmes ont suivi des approches différentes et ont essayé de corriger directement les erreurs présentes dans les documents ce qui semble apporter un gain plus important. Cela constitue un point de départ intéressant dans l'étude de la robustesse des systèmes de RI face aux erreurs.

3 Proposition : Construction d'index corrigés

Notre approche consiste à corriger les erreurs lors de la phase d'indexation du système de RI et plus précisément pendant l'analyse du contenu textuel des documents. Notre proposition s'appuie

4. TREC : Text REtrieval Conference

donc sur deux sous-systèmes : un *système de RI XML* fondé sur le modèle XFIRM présenté en section 2.1, et un *système de correction d'erreurs* qui y est intégré.

En effet, le modèle XFIRM ne permet pas la prise en compte des erreurs, c'est pourquoi les fonctions de calcul de la pondération des termes doivent être modifiées pour en tenir compte. De plus, un système de correction d'erreurs est nécessaire afin d'identifier les termes erronés et d'identifier les termes qui doivent leur être substitués dans l'index.

Supposons les deux phrases suivantes *p1* et *p2* appartenant respectivement à deux documents XML *ds1* et *ds2* simples car comportant un seul élément à leur racine respectivement *nf1* et *nf2* :

p1 : "The trees are green." ***p2*** : "Green paper is made of teer."

Lors de la construction de l'index, les termes sont lemmatisés (mis sous une forme standard : noms au singulier, ...) puis filtrés en fonction d'une liste de mots non significatifs ("stop-words"). L'index construit à partir de ces deux documents est ainsi représenté dans la table 1.

Terme	Document	Élément	tf	idf	ief
green	ds1	nf1	0,5	0,82	0,82
	ds2	nf2	0,33		
paper	ds2	nf2	0,33	1	1
teer	ds2	nf2	0,33	1	1
tree	ds1	nf1	0,5	1	1

TABLE 1 – Index des documents *ds1* et *ds2* (les facteurs *idf* et *ief* sont égaux car les documents ne comportent chacun qu'un seul élément).

Ainsi, une recherche comportant les mots-clés ***tree*** et ***paper*** aboutira aux scores suivants pour chacun des nœuds des deux documents :

$$\begin{aligned} s_{ds1}(r) = s_{nf1}(r) &= p^r_{tree} \times p^{nf1}_{tree} + p^r_{paper} \times p^{nf1}_{paper} = 0,25 \\ s_{ds2}(r) = s_{nf2}(r) &= p^r_{tree} \times p^{nf2}_{tree} + p^r_{paper} \times p^{nf2}_{paper} = 0,165 \end{aligned} \tag{6}$$

Comme cela peut être constaté sur cet exemple, le document *ds1* obtient un score de $0,25$ supérieur au score de $0,165$ obtenu par *ds2*. Néanmoins, on s'aperçoit bien en lisant les 2 phrases que *p1* (et donc *nf1* et *ds1*) devrait moins bien répondre à la requête que *p2* car elle ne contient pas paper alors que c'est le cas de *p2* (et donc *nf2* et *ds2*). Pour pallier cela, le système de correction d'erreurs est utilisé afin d'associer chaque erreur à sa correction avec un degré de confiance δ tel que $t_{err} \overset{\delta}{\longmapsto} t_{cor}$. Cela permet ainsi de détecter que le terme teer noté t_{err} constitue un terme erroné et qu'il doit être remplacé par le terme original tree noté t_{cor}.

Afin de prendre en compte les occurrences potentielles des termes issus de la correction, il est nécessaire de modifier les formules permettant l'obtention de la pondération des termes dans les nœuds des documents (cf. formule 2) à savoir : le tf (cf. formule 3), l'idf et l'ief (cf. formule 4).

$$tf^{nf}_{t_i} = \frac{|t^{nf}_i| + \sum_{e=1}^{|t^{nf}_{cor}|} \delta_e}{|nf|} \tag{7}$$

où, $\sum_{e=1}^{|t^{nf}_{cor}|} \delta_e$ est le nombre (pondéré par la confiance δ_e) de termes erronés t^{nf}_{err} dont la correction t^{nf}_{cor} est égale au terme original t^{nf}_i.

$$idf_{t_i} = log\left(\frac{|D|}{|d_{t_i}|+\sum_{e=1}^{|d_{t_{cor}}|}\delta_e+1}\right)+1 \qquad ief_{t_i} = log\left(\frac{|NF|}{|nf_{t_i}|+\sum_{e=1}^{|nf_{t_{cor}}|}\delta_e+1}\right)+1 \tag{8}$$

où, $\sum_{e=1}^{|d_{t_{cor}}|}\delta_e$ (resp. $\sum_{e=1}^{|nf_{t_{cor}}|}\delta_e$) est le nombre (pondéré par la confiance δ_e) de documents $d_{t_{err}}$ (resp. d'éléments $nf_{t_{err}}$) contenant des termes erronés dont la correction $d_{t_{cor}}$ (resp. $nf_{t_{cor}}$) est égale au terme original d_{t_i} (resp. nf_{t_i}).

Ainsi, si on reprend l'exemple précédent en considérant un degré de confiance plutôt modéré δ de 60% dans la correction (en pratique ce degré est déterminé par le score attribué à t_{cor} par le système de correction d'erreurs), on obtient l'index corrigé selon les formules 7 et 8 présenté dans le tableau 2 :

Terme	Document	Élément	tf	idf	ief
green	ds1	nf1	0,5	0,82	0,82
	ds2	nf2	0,33		
paper	ds2	nf2	0,33	1	1
tree	ds1	nf1	0,5	0,89	0,89
	ds2	nf2	0,2		

TABLE 2 – Index modifié des documents *ds1* et *ds2* (les facteurs *idf* et *ief* sont égaux car les documents ne comportent chacun qu'un seul élément).

Une recherche comportant les mots-clés ***tree*** et ***paper*** aboutira aux scores suivants pour chacun des nœuds des deux documents :

$$\begin{aligned} s^{cor}_{ds1}(r) = s^{cor}_{nf1}(r) = p^r_{tree} \times p^{nf1}_{tree} + p^r_{paper} \times p^{nf1}_{paper} = 0,19 \\ s^{cor}_{ds2}(r) = s^{cor}_{nf2}(r) = p^r_{tree} \times p^{nf2}_{tree} + p^r_{paper} \times p^{nf2}_{paper} = 0,25 \end{aligned} \tag{9}$$

Par conséquent, le document *ds2* sera mieux classé que le document *ds1* (et cela bien que le degré de confiance dans les corrections qui a été choisi soit relativement faible), ce qui est correct compte tenu du fait que c'est ce premier qui est le plus pertinent des deux documents. L'approche proposée a servi de support à l'implémentation de nos prototypes *SnAIRS/SAIRS (Spelling (non-)Aware Information Retrieval System)* évalués ci-dessous.

4 Évaluation

Nos prototypes *SnAIRS*/*SAIRS* ont été évalués sur la collection de documents du track ad-hoc de la campagne d'évaluation INitiative for the Evaluation of XML retrieval (INEX) de 2008. Cette campagne comporte une collection de 659387 documents XML issus de Wikipedia associée à 70 requêtes évaluées. Le track ad-hoc est composé de 3 taches : *focused*, *relevant in context* et *best in context*, qui sont associées à différentes métriques permettant de les évaluer. L'objectif poursuivi suite à la participation à de telles campagnes est d'évaluer le système de RI complet *sans* puis *avec* correction d'erreurs (s'appuyant sur Aspell (Atkinson, 2011)) lors de l'indexation des documents. De cette façon, il est possible d'obtenir à la fois des indicateurs globaux sur les performances de notre système de RI (en comparant ses résultats à ceux obtenus par d'autres systèmes évalués lors de la campagne), mais aussi des indicateurs locaux nous permettant d'estimer l'impact relatif de la correction d'erreurs sur les résultats de notre système de RI.

Prototype	*SnAIRS*	*SAIRS*	Δ (%)
Volume index (Go)	*8,0*	*6,9*	*-13,75*
Durée req. min. (ms)	*2*	*1*	*-50*
Durée req. max. (ms)	13320	27279	+104,80
Durée req. moy. (ms)	605	1139	+88,26
Durée req. 1er quartile (ms)	*4*	*4*	*0*
Durée req. mediane. (ms)	5	6	+20
Durée req. 3e quartile (ms)	16	32	+100
Durée req. total (ms)	41775	78657	+88,29

TABLE 3 – Propriétés de *SnAIRS* (sans correction) et *SAIRS* (avec correction).

Bien que la collection de documents INEX ne contienne qu'un faible taux d'erreurs (les documents issus de Wikipedia sont de relativement bonne qualité), on peut constater dans la table 3 que le volume occupé par l'index est beaucoup plus important pour les mêmes documents lorsque ces derniers contiennent des erreurs même en faible quantité. Ce comportement peut s'expliquer par le fait que les erreurs constituent autant de variations des termes qui viennent augmenter le nombre d'entrées différentes dans l'index. On pourrait penser qu'un index plus petit devrait permettre une exécution plus rapide des requêtes. Bien que cela ne soit pas visible (on constate une dégradation et non pas un gain) dans la table 3, c'est effectivement le cas mais cela est contrebalancé par le fait qu'il y a un nombre plus important de correspondances dans l'index et donc un nombre plus important de résultats à retourner ce qui demande plus de temps.

La taille de l'index et le temps de réponse ne sont pas les seuls facteurs impactés par les erreurs, c'est aussi le cas de la pertinence des résultats. Les systèmes ont ainsi été évalués sur la tache *focused* qui est la plus classique car elle est dédiée à la recherche des éléments (parties de documents) les plus pertinents dans les premiers rangs des résultats de la requête. Cette tache est évaluée en fonction de la précision interpolée à 1% de rappel (iP[.01], la métrique principale), mais aussi de la moyenne des précisions interpolées (MAiP) sur les 101 points de rappel.

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
SnAIRS	0.3073	0.2894	0.1788	0.1501	0.0499
SAIRS	0.3592	0.3141	0.1967	0.1694	0.0598

TABLE 4 – Résultats de *SnAIRS* (sans correction), *SAIRS* (avec correction).

On peut observer sur la table 4 que *SnAIRS* obtient une précision inférieure à *SAIRS* aux différents niveaux de rappels considérés par la campagne INEX et notamment pour la mesure officielle d'iP[.01]. La correction des erreurs à l'indexation permet donc d'obtenir une précision accrue dans les premiers niveaux de rappels. Ces résultats sont prometteurs (de nombreux paramètres peuvent être améliorés) bien qu'ils soient pour l'instant relativement éloignés du Top-10 d'INEX (Kamps *et al.*, 2009) dont les systèmes plus aboutis intègrent des mécanismes tel que l'expansion de requêtes leur permettant de mieux satisfaire aux requêtes "pauvres" composées d'un seul mot-clé.

5 Conclusion et perspectives

Dans cet article nous avons considéré un problème qui touche de façon transverse l'ensemble des applications amenées à manipuler des informations de qualité variable. Nous avons ainsi

considéré le cas des informations textuelles qui sont souvent considérées de fait comme étant "propres". Nous avons proposé une solution à ce problème pour les systèmes de RI structurés en section 3 qui pourrait être étendue à la plupart des systèmes de RI car cette dernière consiste à y intégrer un *système de correction d'erreurs* lors du processus d'indexation. Nous avons dans un premier temps identifié les contraintes spécifiques imposées par les systèmes de RI vis-à-vis des systèmes de corrections d'erreurs, et nous les avons évalués dans (Renard *et al.*, 2011). La correction d'erreurs à l'indexation présente des avantages (cf. table 3) et permet de construire des index plus représentatifs du contenu réel des documents ce qui aboutit à de meilleurs résultats (cf. table 4) que sans correction d'erreurs. De plus, la collection de documents basée sur Wikipedia ne contient que peu d'erreurs et il serait intéressant de corrompre volontairement cette dernière afin de mieux mettre en lumière l'apport de notre proposition.

Références

Atkinson, K. (2011). Correcteur Aspell. `http://aspell.net`. [consulté le 15/01/2012].

Gravier, G., Guinaudeau, C., Lecorvé, G. et Sébillot, P. (2011). Exploiting speech for automatic TV delinearization : From streams to cross-media semantic navigation. *EURASIP JIVP*, 2011(0).

Kamps, J., Geva, S., Trotman, A., Woodley, A. et Koolen, M. (2009). Overview of the INEX 2008 Ad Hoc Track. *In* Geva, S., Kamps, J. et Trotman, A., éditeurs : *Advances in Focused Retrieval*, volume 5631 de *Lecture Notes in Computer Science*, pages 1–28. Springer-Verlag.

Kantor, P. B. et Voorhees, E. M. (2000). TREC-5 Confusion Track : Comparing Retrieval Methods for Scanned Text. *Information Retrieval*, 2(2):165–176.

Renard, A., Calabretto, S. et Rumpler, B. (2011). An evaluation model for systems and resources employed in the correction of errors in textual documents. *In* Morvan, F., Tjoa, A. M. et Wagner, R. R., éditeurs : *8th International Workshop on Text-based Information Retrieval in conjunction with the 22nd International Conference DEXA 2011*, pages 160–164, Toulouse, France. IEEE Computer Society.

Ruch, P. (2002). Using contextual spelling correction to improve retrieval effectiveness in degraded text collections. *In 19th international conference on Computational linguistics-Volume 1*, volume 1, page 7. Association for Computational Linguistics.

Salton, G. (1971). *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall.

Sauvagnat, K. et Boughanem, M. (2005). Using a Relevance Propagation Method for Adhoc and Heterogeneous Tracks at INEX 2004. *In* Fuhr, N., Lalmas, M., Malik, S. et Szlavik, Z., éditeurs : *Advances in XML Information Retrieval*, volume 3493 de *Lecture Notes in Computer Science*, pages 499–532. Springer-Verlag.

Sitbon, L., Bellot, P. et Blache, P. (2007). Traitements phrastiques phonétiques pour la réécriture de phrases dysorthographiées. *In 14ème conférence TALN*, Toulouse, France.

Subramaniam, L. V., Roy, S., Faruquie, T. A. et Negi, S. (2009). A Survey of Types of Text Noise and Techniques to Handle Noisy Text. *Language*, pages 115–122.

Varnhagen, C. K., McFall, G. P., Figueredo, L., Takach, B. S., Daniels, J. et Cuthbertson, H. (2009). Spelling and the Web. *Journal of Applied Developmental Psychology*, 30(4):454–462.

Voorhees, E. M., Garofolo, J. et Sparck Jones, K. (2000). TREC-6 Spoken Document Retrieval Track. *Bulletin of the American Society for Information Science and Technology*, 26(5):18–19.

Post-édition statistique pour l'adaptation aux domaines de spécialité en traduction automatique

Raphaël Rubino, Stéphane Huet, Fabrice Lefèvre, Georges Linarès
LIA-CERI, Université d'Avignon et des Pays de Vaucluse, Avignon, France
{prénom.nom}@univ-avignon.fr

RÉSUMÉ
Cet article présente une approche de post-édition statistique pour adapter aux domaines de spécialité des systèmes de traduction automatique génériques. En utilisant les traductions produites par ces systèmes, alignées avec leur traduction de référence, un modèle de post-édition basé sur un alignement sous-phrastique est construit. Les expériences menées entre le français et l'anglais pour le domaine médical montrent qu'une telle adaptation *a posteriori* est possible. Deux systèmes de traduction statistiques sont étudiés : une implémentation locale état-de-l'art et un outil libre en ligne. Nous proposons aussi une méthode de sélection de phrases à post-éditer permettant d'emblée d'accroître la qualité des traductions et pour laquelle les scores oracles indiquent des gains encore possibles.

ABSTRACT
Statistical Post-Editing of Machine Translation for Domain Adaptation

This paper presents a statistical approach to adapt generic machine translation systems to the medical domain through an unsupervised post-edition step. A statistical post-edition model is built on statistical machine translation outputs aligned with their translation references. Evaluations carried out to translate medical texts from French to English show that a generic machine translation system can be adapted *a posteriori* to a specific domain. Two systems are studied : a state-of-the-art phrase-based implementation and an online publicly available software. Our experiments also indicate that selecting sentences for post-edition leads to significant improvements of translation quality and that more gains are still possible with respect to an oracle measure.

MOTS-CLÉS : Traduction automatique statistique, post-édition, adaptation aux domaines de spécialité.

KEYWORDS: Statistical Machine Translation, Post-editing, Domain Adaptation.

1 Introduction

La traduction automatique statistique basée sur l'alignement sous-phrastique (Koehn *et al.*, 2003) est une approche très populaire qui mène à des performances intéressantes. Les modèles statistiques sous-jacents sont construits à l'aide de phrases en relation de traduction, d'où sont extraites des probabilités d'alignements entre les séquences de mots. Les ressources linguistiques nécessaires à l'estimation de ces probabilités sont les corpus parallèles, éléments indispensables dans le processus de construction du modèle de traduction. Cependant, ces ressources sont coûteuses à

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 527–534,
Grenoble, 4 au 8 juin 2012.

produire et limitent l'approche. Ce manque de données parallèles se fait ressentir plus fortement pour des domaines spécialisés. En effet, beaucoup d'activités humaines impliquent l'utilisation d'une langue spécifique comportant des particularités syntaxiques et terminologiques (Sager *et al.*, 1980). Ainsi, construire des systèmes de traduction pour tous les domaines semble hors de portée et nous pensons que l'adaptation d'un système « générique » représente une solution viable à la prise en charge des domaines dans leur diversité.

Même dans les cas où la traduction automatique peut atteindre de bonnes performances, il est possible d'améliorer manuellement la sortie des systèmes. Mais cela implique une étape de post-édition qui peut se révéler coûteuse selon l'effort à fournir pour produire une traduction de qualité (Martínez, 2003). Il paraît donc intéressant d'automatiser ce processus d'édition *a posteriori* afin de contrôler et d'améliorer les traductions produites par un système. Nous proposons, dans cet article, d'utiliser une approche de post-édition statistique (*statistical post-edition*, SPE) basée sur l'alignement sous-phrastique afin d'adapter des systèmes de traduction automatique à un domaine de spécialité. Le domaine étudié est celui de la médecine et la paire de langues est français-anglais. Plusieurs systèmes de traductions statistiques sont utilisés et nous proposons différentes méthodes afin d'introduire une petite quantité de données spécialisées pendant le processus de traduction. Nous étudions aussi l'impact de la post-édition en l'appliquant systématiquement à toute traduction, puis en sélectionnant les phrases à post-éditer à l'aide d'un classifieur.

L'organisation de cet article est la suivante. La section 2 présente l'approche de SPE par segments sous-phrastique. Puis, dans la section 3, nous proposons un cadre expérimental et donnons des détails sur les données utilisées ainsi que les différentes configurations évaluées. La section 4 contient les résultats en terme de traduction automatique, suivie de la section 5 présentant les résultats de notre approche pour la post-édition. La sélection des phrases à post-éditer est détaillée dans la section 6.

2 L'adaptation aux domaines par post-édition statistique

La post-édition d'une traduction automatique consiste à générer un texte T'' à partir d'une hypothèse de traduction T' provenant d'un texte source S. Ainsi, ne sont nécessaires à la SPE que des données monolingues dans la langue cible. Le corpus parallèle utilisé pour la construction du modèle de post-édition peut être constitué d'hypothèses de traductions, de leurs références de traduction, d'hypothèses post-éditées manuellement, etc.

Parmi les premiers travaux à relater de l'efficacité de la SPE, Simard *et al.* (2007a) proposent d'éditer automatiquement des hypothèses de traduction produites par un système à base de règles. Une étude détaillée de cette approche (traduction par règles et SPE par alignement sous-phrastique) est proposée par (Dugast *et al.*, 2007) et les gains observés sur la mesure d'évaluation BLEU (Papineni *et al.*, 2002) peuvent atteindre jusqu'à 10 points.

L'amélioration de la qualité des traductions produites par un système à base de règles en utilisant la SPE est donc possible. Certains auteurs y voient la possibilité d'adapter le système de traduction à des domaines de spécialité. Selon (Isabelle *et al.*, 2007; Simard *et al.*, 2007b), en introduisant des données spécialisées lors de la phase de post-édition, il est possible d'adapter un système de traduction *a posteriori*. Cette technique a aussi été appliquée par de Ilarraza *et al.* (2008), à partir de traductions basées sur des règles post-éditées par alignement sous-phrastique, en y ajoutant des informations morphologiques. Parmi ces travaux, il est important de noter que

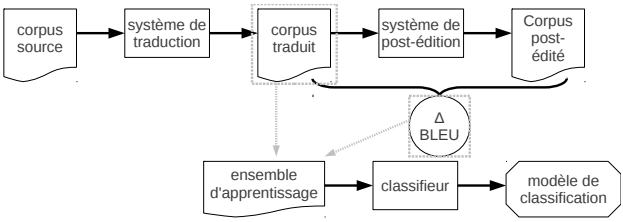


FIGURE 1: Architecture générale combinant la traduction et la SPE, ainsi que la sélection des phrases à post-éditer par estimation des gains $\Delta BLEU$.

certains auteurs ont suggéré de combiner deux systèmes statistiques pour la traduction et la post-édition, sans toutefois poursuivre ces expérimentations (Isabelle *et al.*, 2007; Oflazer et El-Kahlout, 2007).

Plus récemment, Béchara *et al.* (2011) sont les premiers à étudier la combinaison de la traduction et la post-édition par alignement sous-phrastique. Des améliorations sont observées entre le français et l'anglais, en introduisant des informations sur le contexte source dans la phase de post-édition. Mais il n'y a, à notre connaissance, aucune publication sur l'utilisation de l'approche pour l'adaptation aux domaines de spécialité. Nous proposons donc de l'étudier dans cet article, en relation avec la sélection automatique des phrases à post-éditer. Ce dernier aspect est abordé dans (Suzuki, 2011), où les auteurs décrivent une sélection de phrases à post-éditer manuellement basée sur l'estimation de la qualité de traduction au niveau des phrases.

Les travaux présentés dans cet article sont centrés sur l'adaptation de systèmes de traduction génériques en utilisant une petite quantité de données parallèles du domaine. La figure 1 illustre l'architecture générale de notre approche, combinant la traduction, la post-édition et la classification. Dans un premier temps, une traduction automatique d'un texte spécialisé est produite. Puis, en utilisant la référence de traduction, un corpus parallèle monolingue est construit. Ce corpus est utilisé pour construire le modèle de post-édition. Lorsqu'un nouveau texte est traduit, nous proposons alors deux méthodes : la post-édition *naïve* de toutes les phrases ou la sélection des phrases à post-éditer. Cette sélection est faite sur l'estimation des gains possibles par post-édition en recourant à la métrique automatique BLEU.

3 Cadre expérimental

L'idée générale des travaux présentés dans cet article est d'accroître la qualité de traductions de textes spécialisés produites par un système générique par une étape de post-édition. Nous considérons deux types de systèmes de traduction permettant de passer de la langue source à la langue cible : MOSES avec son implémentation de l'alignement sous-phrastique (Koehn *et al.*, 2007) et le système en ligne GOOGLE TRADUCTION[1]. Nous utilisons MOSES pour post-éditer les hypothèses de traduction produites par les deux types de systèmes de traduction individuellement.

Ressources Les données génériques (non spécialisées) sont présentées dans le tableau 1a. Les

1. `http://translate.google.com/`, en utilisant l'API gratuite pendant les mois de Juin et Juillet 2011.

Corpus	Phrases	Mots
Données parallèles		
Europarl v6	1,8 M	50 M
Nations Unies	12 M	300 M
EMEA (Médical)	160 k	4 M
Données monolingues		
News Commentary v6	181 k	4 M
Shuffled News Corpus (2007–2011)	25 M	515 M

(a) Taille des données utilisées en nombre de phrases.

Système		BLEU (%)	p-valeur
MT_g	ML_g	29,9	
MT_g	ML_{g+m}	38,2	0,002
MT_g	ML_m	39,2	0,002
google		44,9	0,002
MT_m	ML_m	46,4	0,007
MT_{g+m}	ML_m	47,2	0,001
MT_{g+m}	ML_{g+m}	47,3	0,75

(b) Scores BLEU obtenus sur le corpus de test du domaine médical.

TABLE 1: Données utilisées (a) et scores BLEU obtenus avec les différents systèmes de traduction (b).

données bilingues, utilisées pour la construction des modèles de traduction, sont composées de la sixième version du corpus Europarl et du corpus issu des Nations Unies. Les données monolingues, utilisées pour la construction des modèles de langue, sont composées des corpus d'actualités *News Crawl* et de la partie langue cible de *News Commentary*. Toutes ces données génériques ont été mises à disposition lors de la campagne d'évaluation *WMT11* [2]. Les données spécialisées sont issues quant à elles de documents provenant de l'agence européenne de médecine (corpus EMEA (Tiedemann, 2009)). Trois catégories médicales sont concernées : des rapports d'évaluation concernant des traitements effectués sur des humains, d'autres effectués sur des animaux, et des documents de médecine générale. Certains documents composant ce corpus sont des prescriptions médicales. Une des caractéristiques de ces documents est une forte redondance au niveau des phrases. Nous décidons de retirer les phrases répétées car elles ne représentent pas un grand intérêt pour la post-édition. En effet, traduire une phrase se trouvant dans l'ensemble d'apprentissage revient a consulter une mémoire de traduction au niveau des phrases. Aussi, nous ne conservons dans le corpus que les phrases d'une longueur inférieure à 80 mots car les alignements obtenus sur des longues phrases sont généralement de qualité moindre. Puis, trois sous-ensembles de phrases sont constituées, formant un corpus d'entraînement (156k phrases), un corpus de développement (ou optimisation, 2k phrases) et un corpus de test (2k phrases).

Systèmes de traduction Parmi les systèmes de traduction utilisés, l'outil accessible en ligne, noté *google* dans les expériences présentées dans cet article, ne peut être modifié. Nous pouvons cependant post-éditer et évaluer les traductions produites par ce système. La boîte à outils de traduction MOSES permet, quant à elle, de construire un modèle de traduction à partir des corpus parallèles et de contrôler chaque étape de ce processus. Ainsi, nous pouvons faire varier les données monolingues et bilingues utilisées.

Pour les modèles de langue (ML), trois modèles 5-grammes sont construits. Un premier est établi sur les données monolingues génériques (ML_g), un second est construit sur la partie langue cible du corpus d'entraînement médical (ML_m). Ces deux modèles sont combinés par interpolation linéaire (ML_{g+m}) selon des poids estimés par le calcul de la perplexité sur le corpus de développement spécialisé. Pour ce dernier modèle, le vocabulaire générique est limité à 1 million de mots les plus fréquents. Le poids optimal associé au modèle de langue spécialisé est de $0,9$ malgré sa petite taille, ce qui montre le niveau de spécificité du domaine médical.

2. `http://www.statmt.org/wmt11/`

Pour les modèles de traduction (MT), MOSES permet de construire une table de traduction et un modèle de réordonnancement en choisissant les données à utiliser. Un premier modèle de traduction est construit avec les données bilingues génériques (MT_g), un second avec les données médicales (MT_m). La combinaison de ces deux modèles permet d'obtenir un modèle mixte (MT_{g+m}). Pour ces trois configurations, les données bilingues sont alignées au niveau des mots avec l'outil MGIZA++ (Gao et Vogel, 2008). Les poids associés aux éléments composant les modèles de traduction sont optimisés sur le corpus de développement médical pour la métrique BLEU selon la méthode *MERT* (Och, 2003).

Afin de construire des systèmes de post-édition pour l'adaptation au domaine médical, nous utilisons les traductions du corpus d'entraînement spécialisé produites par chaque système de traduction individuellement. Chaque sortie est alignée avec la référence de traduction puis utilisée pour construire un modèle de post-édition à l'aide de MOSES (avec les paramètres par défaut). Le corpus de développement spécialisé, une fois traduit par chaque système de traduction, est utilisé afin d'optimiser les poids des composants du modèle de post-édition.

4 Traduction de textes spécialisés

La première série d'expériences porte sur la traduction du corpus de test spécialisé. Les résultats obtenus selon les différentes configurations de systèmes sont présentés dans le tableau 1b. La comparaison entre les systèmes est effectuée selon la méthode d'approximation par sous-échantillonnage aléatoire implémentée dans l'outil FASTMTEVAL (Stroppa *et al.*, 2007). Ces résultats indiquent que la meilleure configuration, avec un score BLEU de 47,3%, est celle combinant les données génériques et spécialisées ($MT_{g+m}ML_{g+m}$). Cependant, ces scores ne sont pas significativement supérieurs à ceux obtenus par $MT_{g+m}ML_m$ (p-valeur=0,75). Nous pouvons donc en conclure que l'intégration des données génériques dans le modèle de langue ne permet pas d'améliorer les performances du système de traduction, ce qui démontre encore une fois la forte spécificité du domaine médical. Cette constatation est plus marquée encore lors de l'utilisation du modèle de traduction générique (MT_g). Introduire des données génériques dans le modèle de langue (MT_gML_{g+m}) dégrade d'un point de BLEU les performances en comparaison avec le score obtenu par MT_gML_m (de 39,2% à 38,2%). Le système de traduction en ligne obtient quant à lui 44,9% de BLEU, soit 1,5 points de moins que le système construit uniquement sur les données spécialisées.

5 Post-édition des traductions

Afin de post-éditer les hypothèses de traduction produites par les systèmes étudiés, le corpus d'entraînement spécialisé est traduit par les différentes configurations et aligné avec sa référence de traduction. Ainsi, un système de post-édition est construit pour chaque système de traduction. Lorsque le corpus de test est traduit, il peut être post-édité dans son intégralité, ou uniquement sur les phrases avec une amélioration possible. Deux scores sont donc calculés : le premier représente l'application *naïve* de la post-édition, le second indique les gains maximums possibles de notre approche (score *oracle*).

Système en ligne Le système de traduction en ligne obtient des résultats convenables lors de la phase de traduction et les résultats obtenus en post-édition sont présentés dans le tableau 2. La comparaison entre les systèmes montre des différences significatives, avec des $p\text{-}valeur$ de 0,001 pour la post-édition *naïve* et 0,05 pour les scores *oracles*. Deux systèmes de post-édition sont

	base	$+ SPE_m ML_m$	$+ SPE_m ML_{g+m}$
google	44,9	46,8 (53,3)	47,9 (53,5)
$MT_g ML_g$	29,9	43,4 (44,2)	45,6 (47,0)
$MT_g ML_m$	39,2	42,7 (44,2)	42,5 (44,4)

TABLE 2: Scores BLEU (%) après post-édition des traductions produites par les systèmes en ligne (*google*) et *Moses* sur le corpus de test médical, utilisant un modèle de langue générique ou spécialisé. Scores *oracle* entre parenthèses.

construits, utilisant les données spécialisées pour le modèle de post-édition et se différenciant selon les modèles de langue, spécialisé ($SPE_m ML_m$) ou mixte ($SPE_m ML_{g+m}$). Les meilleurs résultats sont obtenus par $SPE_m ML_{g+m}$ avec un score BLEU de 47,9%. Le score oracle indique pour cette configuration un score maximum de 53,5%, ce qui motive notre approche de sélection de phrases.

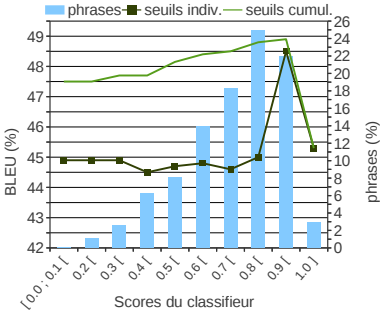
Système générique Nous proposons à présent de post-éditer les sorties d'un système de traduction statistique par alignement sous-phrastique ne disposant pas de données spécialisées pour le modèle de traduction. Les résultats de la post-édition sont présentés dans le tableau 2 en utilisant ($MT_g ML_m$) ou non ($MT_g ML_g$) les données spécialisées pour construire un ML. Le recours à la post-édition montre un gain significatif pour $MT_g ML_g$ en faisant progresser le score BLEU de 29,9% à 45,6% pour la meilleure configuration ($SPE_m ML_{g+m}$). L'augmentation observée de 3,5 points de BLEU à partir de $MT_g ML_m$ est moins importante, tout en restant statistiquement significative ($p\text{-}valeur = 0,001$).

Système spécialisé et mixte Afin de couvrir l'ensemble des expériences possibles selon les différents systèmes de traduction construits, nous procédons à présent à l'évaluation de la post-édition de sorties issues des systèmes utilisant les données spécialisées. Deux systèmes de traduction sont concernés, MT_m et MT_{g+m}, chacun pouvant utiliser un modèle de langue spécialisé ou mixte. Cependant, nous ne détaillons pas les résultats obtenus après post-édition des traductions, car aucune amélioration n'a été observée lors des expérimentations. Seuls les scores *oracles* indiquent qu'un gain est possible, si la sélection des phrases à post-éditer est correctement effectuée.

6 Sélection des phrases à post-éditer

La sélection des traductions à post-éditer est motivée par les scores *oracles* mesurés dans les expériences précédentes. Nous proposons d'utiliser un classifieur afin de détecter les traductions pouvant être améliorées grâce à la post-édition. Le score $\Delta BLEU$ (avant et après post-edition) permet d'associer une classe aux phrases d'un corpus d'entraînement. Nous choisissons le corpus de développement spécialisé comme ensemble d'entraînement pour le classifieur. Ce dernier est de type Séparateur à Vaste Marge (SVM (Boser *et al.*, 1992)), implémenté dans l'outil LIBSVM (Chang et Lin, 2011). Les phrases traduites sont utilisées sous la forme de vecteurs de n-grammes (avec $n \in [1;3]$ dans notre cas).

Nous évaluons notre approche de sélection de phrases à post-éditer en utilisant la configuration dont le score *oracle* est le plus élevé, c-à-d le système de traduction en ligne (*google*) avec le système de post-édition $SPE_m LM_{g+m}$ (*oracle* atteignant 53,5%). Le corpus de développement spécialisé permet de construire le modèle de classification, puis le corpus de test spécialisé est soumis au SVM afin d'en extraire les phrases à post-éditer. Les performances du classifieur



(a) Scores BLEU et nombre de phrases étiquetées « à post-editer » selon les seuils individuels et cumulés des scores issus du classifieur sur le corpus de test médical.

google	initial	+ SPE	+ SPE_selection
TER	42,3	40,4	39,7
BLEU	44,9	47,9	48,9

(b) Scores TER et BLEU après traduction, sélection des phrases et post-édition (p(à post-éditer) $>= 0,8$) sur le corpus de test médical.

FIGURE 2: Analyse sur le corpus de test des résultat de post-édition avec sélection de phrases pour le système de traduction en ligne.

atteignent 79,5% de rappel et 40,1% de précision. Le classifieur produit un score de confiance pour chaque classe attribuée, correspondant à la probabilité qu'une phrase appartienne à la classe prédite. Ceci nous permet de définir des seuils d'acceptation des phrases classées dans la catégorie à post-éditer. Nous évaluons notre approche de sélection selon ces seuils, individuellement ou cumulés. Les résultats sont présentés dans la figure 2a. En cumulant les seuils au dessus de 0,8 pour la classe « à post-éditer », 1 point de BLEU est gagné par rapport à l'application *naïve* de la post-édition. Nous remarquons qu'une quantité plus importante de phrases sont post-éditées entre les seuils 0,5 et 0,8, et seulement 60 phrases le sont avec un score supérieur à 0,9. Cet aspect influence les résultats en terme de score BLEU. L'évaluation globale du corpus de test après post-édition des phrases sélectionnées montre une amélioration selon les deux métriques utilisées (tableau 2b). L'utilisation d'une sélection apparaît donc comme une méthode apportant des gains en comparaison à l'application *naïve* de la post-édition (avec une p-valeur égale à 0,004).

7 Conclusion et perspectives

Nous avons présenté dans cet article une approche de post-édition statistique fondée sur les segments pour l'adaptation aux domaines de spécialité en traduction automatique. Les expériences menées montrent qu'un système de traduction générique peut être adapté *a posteriori* par l'introduction de données spécialisées dans une étape de post-édition. L'application *naïve* de la SPE permet d'améliorer la qualité de traduction dans certains cas. Les scores *oracles* indiquent que pour toutes les configurations étudiées, des gains en terme de score BLEU sont possibles. Les meilleurs résultats sont obtenus par le système de traduction en ligne, couplé avec un système de post-édition construit sur des données mixtes, et en utilisant un classifieur pour sélectionner les traductions à post-éditer. Comparé au système de base, le score BLEU est amélioré de 4 points. L'apprentissage du classifieur est toutefois limité aux hypothèses de traduction et nous envisageons, dans des travaux futurs, d'enrichir les paramètres d'apprentissage pour mieux tenir compte du contexte de traduction afin de s'approcher des scores *oracles* mesurés.

Références

BÉCHARA, H., MA, Y. et van GENABITH, J. (2011). Statistical post-editing for a statistical MT system. *In MT Summit XIII*, pages 308–315.

BOSER, B., GUYON, I. et VAPNIK, V. (1992). A training algorithm for optimal margin classifiers. *In 5th annual workshop on Computational learning theory*, pages 144–152.

CHANG, C.-C. et LIN, C.-J. (2011). LIBSVM : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27 :1–27 :27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

de ILARRAZA, A., LABAKA, G. et SARASOLA, K. (2008). Statistical postediting : A valuable method in domain adaptation of RBMT systems for less-resourced languages. *In MATMT*, pages 35–40.

DUGAST, L., SENELLART, J. et KOEHN, P. (2007). Statistical post-editing on Systran's rule-based translation system. *In WMT*, pages 220–223.

GAO, Q. et VOGEL, S. (2008). Parallel implementations of word alignment tool. *In ACL Workshop : Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57.

ISABELLE, P., GOUTTE, C. et SIMARD, M. (2007). Domain adaptation of MT systems through automatic post-editing. *In MT Summit XI*, pages 255–261.

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R. *et al.* (2007). Moses : Open source toolkit for statistical machine translation. *In ACL*, pages 177–180.

KOEHN, P., OCH, F. et MARCU, D. (2003). Statistical phrase-based translation. *In NAACL-HLT*, volume 1, pages 48–54.

MARTÍNEZ, L. (2003). *Human Translation versus Machine Translation and Full Post-Editing of Raw Machine Translation Output*. Thèse de doctorat, Dublin City University.

OCH, F. (2003). Minimum error rate training in statistical machine translation. *In ACL*, volume 1, pages 160–167.

OFLAZER, K. et EL-KAHLOUT, I. (2007). Exploring different representational units in English-to-Turkish statistical machine translation. *In WMT*, pages 25–32.

PAPINENI, K., ROUKOS, S., WARD, T. et ZHU, W. (2002). BLEU : A method for automatic evaluation of machine translation. *In ACL*, pages 311–318.

SAGER, J., DUNGWORTH, D. et MCDONALD, P. (1980). English special languages : principles and practice in science and technology. *Wiesbaden : Oscar BrandstetterVerlay*, pages 2–35.

SIMARD, M., GOUTTE, C. et ISABELLE, P. (2007a). Statistical phrase-based post-editing. *In NAACL-HLT*, pages 508,515.

SIMARD, M., UEFFING, N., ISABELLE, P. et KUHN, R. (2007b). Rule-based translation with statistical phrase-based post-editing. *In WMT*, pages 203–206.

STROPPA, N., OWCZARZAK, K. et WAY, A. (2007). A cluster-based representation for multi-system MT evaluation. *In TMI*, pages 221–230.

SUZUKI, H. (2011). Automatic post-editing based on SMT and its selective application by sentence-level automatic quality evaluation. *In MT Summit XIII*, pages 156–163.

TIEDEMANN, J. (2009). News from OPUS—a collection of multilingual parallel corpora with tools and interfaces. *In RANLP*, volume V, pages 237–248.

Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées

Benoît Sagot[1] Marion Richard[1,2] Rosa Stern[1,3]
(1) Alpage, INRIA Paris–Rocquencourt & Université Paris Diderot, 175 rue du Chevaleret, 75013 Paris
(2) ISHA, Université Paris Sorbonne, 7 rue Victor Cousin, 75006 Paris
(3) AFP MediaLab, 2 place de la Bourse, 75002 Paris
benoit.sagot@inria.fr, ma.rih.on75@gmail.com, rosa.stern@afp.com

RÉSUMÉ

Le Corpus Arboré de Paris 7 (ou French TreeBank) est le corpus de référence pour le français aux niveaux morphosyntaxique et syntaxique. Toutefois, il ne contient pas d'annotations explicites en entités nommées. Ces dernières sont pourtant parmi les informations les plus utiles pour de nombreuses tâches en traitement automatique des langues et de nombreuses applications. De plus, aucun corpus du français annoté en entités nommées et de taille importante ne contient d'annotation référentielle, qui complète les informations de typage et d'empan sur chaque mention par l'indication de l'entité à laquelle elle réfère. Nous avons annoté manuellement avec ce type d'informations, après pré-annotation automatique, le Corpus Arboré de Paris 7. Nous décrivons les grandes lignes du guide d'annotation sous-jacent et nous donnons quelques informations quantitatives sur les annotations obtenues.

ABSTRACT

Referential named entity annotation of the Paris 7 French TreeBank

The French TreeBank developed at the University Paris 7 is the main source of morphosyntactic and syntactic annotations for French. However, it does not include explicit information related to named entities, which are among the most useful information for several natural language processing tasks and applications. Moreover, no large-scale French corpus with named entity annotations contain referential information, which complement the type and the span of each mention with an indication of the entity it refers to. We have manually annotated the French TreeBank with such information, after an automatic pre-annotation step. We sketch the underlying annotation guidelines and we provide a few figures about the resulting annotations.

MOTS-CLÉS : Résolution d'entités nommées, Corpus annoté, Corpus arboré de Paris 7.

KEYWORDS: Named entity resolution, Annotated corpus, French TreeBank.

1 Introduction et état de l'art

La notion d'entité nommée (EN) est au cœur d'un nombre considérable de travaux en traitement automatique des langues depuis plusieurs décennies. Elle a notamment fait l'objet des conférences MUC (Marsh et Perzanowski, 1998) et des campagnes associées, puis de campagnes CoNNL (Sang et Meulder, 2003) et ACE (Doddington *et al.*, 2004). Traditionnellement, et notamment dans les campagnes MUC, les EN sont classées en noms de personnes, noms de lieux, noms

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 535–542,
Grenoble, 4 au 8 juin 2012.

d'organisations, et parfois « autres noms propres ». Au-delà de cette définition simple et restrictive, l'usage s'accorde de plus en plus à étendre la notion d'EN à d'autres types, comme les noms de marques et de produits, qui sont également le plus souvent des noms propres, mais également les noms d'œuvres, les dates, les montants, voire les adresses, les URLs, les nombres, voire tout token [1] ou séquence de tokens qui n'a pas vocation à faire partie d'un lexique et qui respecte une grammaire dite locale, spécifique à la nature de ce qu'elle dénote (Sekine et Nobata, 2004; Grouin *et al.*, 2011). Ce que tous ces types d'EN ont en commun est leur caractère référentiel, qui se décline toutefois de façon différente suivant les types, voire suivant les corpus. C'est ainsi qu'Ehrmann (2008) définit une entité nommée comme suit : *Étant donné un modèle applicatif et un corpus, on appelle EN toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.*

La tâche de **reconnaissance d'EN** est généralement conçue comme ayant pour objectif d'identifier automatiquement en corpus les *mentions* d'EN, c'est-à-dire les séquences de tokens qui réfèrent à une entité, mais également de typer ces mentions, par exemple avec des catégories telles que `Person`, `Location`, `Organization`. De très nombreux travaux ont été publiés depuis longtemps sur cette tâche, qui a été abordée entre autres avec des méthodes à base de règles et, souvent, de ressources lexicales (« gazetteers », ou lexiques de mentions d'entités associées à des types) (Sekine et Nobata, 2004; Rosset *et al.*, 2005; Stern et Sagot, 2010), des techniques statistiques d'apprentissage reposant sur des corpus annotés (Finkel *et al.*, 2005; Bechet et Charton, 2010), et des techniques hybrides. Ces travaux accordent souvent, sous une forme ou sous une autre, une importance particulière au problème du typage en lien avec des phénomènes tels que la métonymie : en effet, on peut distinguer deux façons de typer une entité, soit de façon intrinsèque (*la France* dénote un lieu), soit en contexte (dans *La France a signé le traité*, *France* peut être typé comme une organisation).

Qu'il s'agisse de permettre l'évaluation de ces systèmes, ou leur entraînement dans le cas de systèmes statistiques, des corpus annotés ont été développés pour diverses langues. Pour le français, on peut citer notamment les corpus ESTER et ESTER2 (60 plus 150 heures d'émissions transcrites orthographiquement et annotées en EN) (Galliano *et al.*, 2009), ainsi que le corpus Quaero (Grouin *et al.*, 2011). Tous ces corpus reposent sur des données orales transcrites (émissions de radio). On notera que le corpus Quaero repose sur une définition originale, très riche et structurée de la notion d'EN (Rosset *et al.*, 2011). De plus, il contient des annotations de typage à la fois en termes de type absolu et de type en contexte, contrairement aux corpus ESTER qui ne contiennent que le type en contexte.

Toutefois, la seule tâche de reconnaissance des EN ne suffit pas à extraire les informations nécessaires pour des applications comme l'extraction d'informations. Seul un système permettant d'associer à ces mentions un référent extra-linguistique permet d'exploiter le résultat de la détection grâce au sens qui lui est ainsi conféré (Blume, 2005). Cette tâche dite de **résolution des EN** (REN, *entity linking*) consiste ainsi à associer à chaque mention d'EN l'entrée adéquate dans une base d'entités qui sert de référence, en traitant notamment les cas d'homonymie : une mention de *Michael Jordan* réfère-t-elle au footballeur, au joueur de basket ou à l'économiste ? Une mention d'*Orange* réfère-t-elle à une ville (laquelle ?) ou à l'entreprise ? Outre les difficultés liées à l'homonymie des mentions, mais également aux phénomènes de métonymie, on est également confronté à la diversité des mentions pour une même entité (variantes graphiques : *Jacques Chirac* et *J. Chirac*, surnoms : *Ali le chimique* pour *Ali Hassan al-Majid*).

1. Un token, contrairement à une (occurrence de) forme, est une unité typographique (Sagot et Boullier, 2008).

La résolution automatique d'EN fait l'objet de travaux depuis quelques années sur l'anglais (Bunescu et Pasca, 2006; Cucerzan, 2007; McNamee et Dang, 2009). C'est souvent Wikipedia qui est utilisé comme base d'entités de référence. Pour le français, Stern et Sagot (2010) proposent un système à base de règles intégré à la chaîne de traitements de surface SxPipe(Sagot et Boullier, 2008), nommé NP. Ce système repose sur la base d'entités Aleda (Sagot et Stern, 2012), extraite automatiquement à partir de Wikipedia et de la base de noms de lieux Geonames. Cependant, peu de corpus annotés en référence, associés à une base d'entités, sont disponibles pour évaluer ou entraîner de tels systèmes. Pour l'anglais, on peut citer le corpus rendu disponible pour la tâche de peuplement de base de données de la campagne TAC 2009 (McNamee et Dang, 2009). Pour le français, le seul corpus disponible est constitué de 100 dépêches de l'Agence France-Presse de 300 mots chacune en moyenne, qui utilise comme référence la base Aleda (Stern et Sagot, 2010). Outre les informations référentielles, ce corpus inclut naturellement pour chaque mention annotée les informations d'empan et de type, plus précisément de type absolu, en cohérence avec le type de l'entité tel qu'il est indiqué (ou devrait l'être) dans Aleda.

Dans cet article, nous décrivons un travail d'annotation des entités nommées du Corpus Arboré de Paris 7, ou French TreeBank (Abeillé *et al.*, 2003), avec les mêmes principes que le petit corpus de Stern et Sagot (2010). L'objectif est triple :

– Préciser les conventions d'annotation en les confrontant à un corpus plus important et moins contemporain des ressources utilisées pour construire Aleda ;
– Fournir à la communauté un corpus écrit de taille importante dont les EN soient annotées en empan, type et référence ;
– Ajouter une couche d'annotation à un corpus pour lequel d'autres niveaux d'annotations sont disponibles (à ce jour, annotations morphosyntaxiques, arbres de constituance et fonctions syntaxiques ; à terme, annotations de type FrameNet et annotations discursives).

Nous décrivons donc succinctement les conventions d'annotations utilisées (section 2), le processus d'annotation guidé par une pré-annotation effectuée à l'aide du système NP mentionné ci-dessus (section 3) et les résultats obtenus (section 4).

2 Conventions d'annotation

2.1 Principes généraux

Nous avons annoté le Corpus Arboré de Paris 7 en indiquant l'empan, le type absolu [2] parfois complété d'un sous-type et l'identifiant du référent dans la base Aleda de toute mention d'EN sous forme de nom propre, à l'exclusion de tout autre type d'expression référentielle (descriptions définies, pronoms...). Nous nous sommes restreints aux noms de personnes, de lieux, d'organisations, d'entreprises, et à certains noms de produits. Une annotation systématique des noms de produits et des noms d'œuvres sera à effectuer ultérieurement. En revanche, nous ne nous sommes intéressés ni aux EN moins standard (URL, pourcentages...) ni aux expressions temporelles. Enfin, nous n'avons annoté aucune EN enchâssée dans une autre.

Plus précisément, nous avons fait usage de 7 types de base : `Person`, `Location`,

2. L'entité *France* sera donc toujours une entité `Location` avec pour sous type `Pays`, comme indiqué dans Aleda pour le référent correspondant, même si le sens contextuel réfère à l'organisation politique, au peuple français, à l'équipe de foot, etc.

`Organization`, `Company`, `Product`, `POI` (Point of Interest) et `FictionChar` (personnage de fiction). Ces types sont parfois précisées par un sous-type. Les types et sous-types ont été organisés via une ontologie, dans laquelle les types sont des classes qui sont à la tête d'une hiérarchie de sous-classes qui correspondent à des sous- types. Pour sous-typer une entité il n'est donc pas nécessaire de préciser la totalité des sous-classes qui lui correspondent.

Notre définition de ce qu'est une EN conduit à des cas limites, notamment pour les mentions qui n'ont pas de référent autonome en soi, mais qui en acquièrent un en contexte, comme par exemple *banque centrale*. Dans ce type de cas, nous avons considéré qu'il y avait bien mention d'EN, et nous avons donc annoté, pour peu que le contexte donné permette d'établir quelle est la banque précise dont il est question, à la condition (arbitraire) supplémentaire que la mention commence par une majuscule. Ainsi, une mention comme *banque centrale* sera systématiquement ignorée. En revanche, les mentions primaires d'entités qui ne dépendent pas du contexte sont annotées qu'elles aient ou non des majuscules, comme par exemple *banque mondiale*. Cette situation se retrouve par exemple également dans le cas de l'annotation des noms d'universités. Nous considérons ainsi qu'*université de Nantes* dénote une université située à Nantes, et nous n'annotons que la ville de Nantes, alors qu'*Université de Nantes* fait directement référence à l'organisation qu'est cette université, et nous annotons donc l'ensemble comme une organisation. Il en va de même, par exemple, pour *Université de Montpellier*, puisqu'il n'existe pas d'organisation unique qui corresponde à ce terme : dans ce cas, seul *Montpellier* est annoté, en tant que ville.

Les mentions annotées peuvent correspondre au nom normalisé (*Jacques Chirac*), à une variante (*Chirac* dans *M. Chirac*, cf. plus bas concernant *M.*) ou à un surnom (comme dans *l'Hexagone*. Ainsi, la description définie *l'avocat de M. Chirac* entraînera une annotation de la mention *Chirac*, en ignorant la référence à l'avocat de ce dernier. Par ailleurs, les mots grammaticaux ou contextuels entourant la mention de l'entité sont ignorés. Ainsi les déterminants ne sont pas pris en compte, ni les titres, professions ou adjectifs pouvant apparaître pour qualifier l'entité. Ainsi, dans *Chine méridionale*, seul *Chine* est annoté comme un nom de lieu, et dans *M. Bill Clinton* seul *Bill Clinton* est annoté comme un nom de personne.

Les balises utilisées pour l'annotation contiennent les informations suivantes :
– l'identifiant de l'EN dans Aleda (attribut `eid`) ; dans le cas d'une entité non présente dans la base l'identifiant est marqué `null`.
– le nom normalisé de l'entité, tel qu'indiqué dans Aleda ; pour les lieux il s'agit du nom donné dans GeoNames et pour les autres entités du titre de l'article dans la Wikipedia française.
– un type, ainsi qu'un sous-type dans le cas où l'entité entre dans une catégorie sous-typée (cf. section ci-dessous).

Voici deux exemples d'annotation :

```
<ENAMEX type="Organization" eid="1000000000016778" name="Confédération
française démocratique du travail">CFDT</ENAMEX>
<ENAMEX type="Location" sub_type="Country" eid="2000000001861060"
name="Japan">Japon</ENAMEX>
```

Dans certaines balises se trouve une information supplémentaire pour les cas de fusions d'entreprises, de changement de nom d'une entreprise et de lieux géographiques qui n'existent plus en tant que tels :
– `current_eid` est l'identifiant dans Aleda d'une entité actuelle correspondant à une entité qui n'existe plus mais qui est le référent réel (par exemple en cas de rachat d'une entreprise par une autre) ;
– `current_name` est le nom normalisé de l'entité `current_eid`.

Si une entreprise a changé de nom sans que l'on puisse considérer qu'il y a eu changement de référent, on ajoute un attribut `former_name` qui permet d'indiquer que la mention de l'entité réfère à l'ancien nom de l'entreprise. Enfin un attribut `former_location` initialisé à `True` permet d'annoter un lieu géographique disparu.

2.2 Types et sous-types

L'annotation des **noms de personnes** ne pose pas de problèmes particuliers. Deux précisions toutefois concernant deux cas :
- Les groupes de personnes : Les références à des groupes de personnes, telles que *la famille Agnelli* ou *les frères Maxwell*, ne sont pas annotées. Il en est de même pour les populations : le groupe de personnes désigné par Les Français n'est pas pris en compte dans l'annotation.
- Les fonctions ou titres : Les fonctions ne sont pas annotées. Une expression telle que *le Premier ministre M. Fillon* permettra d'annoter *Fillon* mais ne tiendra pas compte de la mention *Premier ministre* en tant qu'EN (cela suit le fait que nous ne retenons pas les mentions sous forme de description définie). Dans la lignée, l'expression *Général de Gaulle* donnera lieu à l'annotation de la mention *de Gaulle*, mais *Général* ne sera pas inclus dans l'empan.

Le type `FictionChar` permet d'annoter toutes les mentions qui font référence à des **noms de personnes ou d'animaux fictifs**, telles que *McGyver* ou *Zorro*.

Parmi les **noms de lieux**, les sous-types utilisés sont les suivants :
- Sont sous-typés `Country` les états indépendants ;
- Les divisions territoriales des pays sont annotées `CountryDivision` (chaque pays ayant sa propre gestion du territoire, des sous-types tels que *département*, *etat fédéral*, *canton*, *district*, *comté* ne semblaient pas pertinents) ;
- le sous-type `Region` permet d'annoter des parties du monde non liées à une gestion politique. Ce sous-type regroupe les continents et parties de continents, ainsi que des lieux géographiques sans frontières mais dans le vocabulaire courant tel que la région du Golfe ;

Le type « Point Of Interest » (`POI`) est utilisé pour les entités telles que les ports, les salles de spectacle, les stades, les quartiers, etc.

Nous considérons comme des **noms d'organisations** et typons `Organization` toutes les références à des organisations qu'elles soient politique, éducative, économique, etc, à l'exclusion des noms d'entreprises. Un seul sous-type est utilisé, `PoliticalGroup` pour les organisations politiques. Enfin, les **noms d'entreprises**, typés `Company`, ne sont pas sous-typés.

2.3 Difficultés génériques

L'annotation est parfois rendue difficile du fait de l'ambiguïté de l'entité ou de sa catégorisation. En contexte journalistique, les principaux cas de difficulté sont liés à la temporalité du corpus. En effet les informations présentes dans un corpus journalistique dépendent essentiellement du contexte temporel. Ainsi, en 2011, année dont datent les informations ayant servi à construire la base de référence Aleda, la Tchécoslovaquie, l'URSS ou l'entreprise Thomson CSF n'existent plus, mais en 1990, date de rédaction des textes constituant le corpus, ces pays existaient encore. Faire l'impasse sur ces entités serait faire l'impasse sur une partie de l'information passée, nous avons donc défini des règles d'annotation pour ces entités particulières.

Les mentions d'entreprises peuvent ne plus avoir de référence actuelle. Si l'entreprise a simplement disparu, son référent est annoté `null`. Mais cela peut aussi être dû à un changement de nom, à un rachat ou à une fusion avec une autre entreprise. Dans les cas d'un changement de nom, nous avons inclus l'ancien nom dans un attribut `former_name`.

```
<ENAMEX type="Organization" eid="1000000000036708" name="France 2"
former_name="Antenne 2">Antenne 2</ENAMEX>
```

Dans le cas du rachat d'une entreprise ou d'une fusion, nous ajoutons deux sous-types pour permettre d'identifier la référence actuelle de l'entreprise en opposition à la référence du texte, (datant ici des années 90). Ainsi nous conservons la valeur temporelle du texte dans l'annotation tout en actualisant la référence. Par exemple à l'époque de la rédaction des dépêches du corpus le groupe UAP n'avait pas encore fusionné avec le groupe AXA, les attributs `current_eid` et `current_name` permettent d'actualiser la référence en donnant l'information que le groupe s'appelle désormais AXA.

```
<ENAMEX type="Company" eid="null" name="Union des assurances de Paris"
current_eid="1000000000201762" current_name="AXA">UAP </ENAMEX>
```

Quant aux filiales, elles sont annotées seules c'est à dire sans référence à leur entreprise mère. Lorsqu'il s'agit de petites filiales qui n'ont pas de référent, comme c'est le cas de nombreuses fois, l'identifiant sera annoté `null`. Les filiales étrangères connues d'Aleda des grosses entreprises sont annotées en tant que telles.

Plusieurs mentions faisant référence à une entreprise ou une usine sont citées via leur marque (ex. : *Mamie nova*). Dans ce cas, l'annotation est faite sur la marque. Dans une expression comme *un tracteur John Deer*, l'entité *John Deer* est annotée en tant qu'entreprise, pour rester cohérent avec l'idée de départ qui est de considérer les EN dans leur sens absolu. Le choix d'annoter une EN présentant une ambiguïté entre nom de produit ou d'entreprise est guidé par le type associé à l'entité dans la base de données Aleda (par exemple, *Nike* et *Adidas* sont annotés `Company`, mais *Evian* ou *Lessieur* sont annotés `Product`).

Pour les dénominations géographiques qui ne sont plus d'actualité à la date d'extraction d'Aleda mais qui sont utilisées dans le corpus, nous ajoutons un sous-type `former-location` de type `boolean` qui est annoté `true` par défaut. Ces lieux ne possédant pas d'identifiant dans la base de données ils sont identifiés `null`, mais cela permet leur annotation et reconnaissance dans le corpus.

```
<ENAMEX type="Location" sub_type="Country" eid="null" former_location="true"
name="Tchécoslovaquie">Tchécoslovaquie </ENAMEX>
```

Outre ces difficultés relativement générales, certaines entités ou mentions ont posé des problèmes spécifiques sur lesquels il a fallu faire des choix. C'est le cas de *Trésor (public)* [3], *Hachette* [4] et *Thomson* [5] qui correspondent à des organisations qui ont fortement évolué entre la date de rédaction du corpus et la date de création d'Aleda.

3. Certains pays disposant de plusieurs organismes se divisant les tâches associées à la notion générale de trésor public. Dans ce cas, nous choisissons le ministère des finances comme référent de la mention *Trésor (public)*.

4. Toutes les occurrences de *Hachette* sont annotées sous la référence au Groupe Lagardère, propriétaire de *Hachette livres* et de *Hachette Filippachi Médias*. *Hachette* ne correspond plus aujourd'hui à une entité unique.

5. À l'époque de rédaction du corpus, la mention *Thomson* pouvait faire référence à deux entreprises : Thomson CSF aujourd'hui devenu Thalès et Thomson SA aujourd'hui Technicolor. Dans les cas les plus simples la mention est suivi de CSF ou SA, et on donne donc le référent Thalès ou Technicolor, en indiquant le `former_name`. Les occurrences de *Thomson* seul sont désambiguïsées en contexte.

3 Processus d'annotation

L'intégralité du Corpus Arboré de Paris 7 dans sa version 2007 (à l'exclusion des phrases n'ayant pas reçu d'annotations fonctionnelles), soit 12 351 phrases contenant 350 931 tokens, a été annoté en EN (empan, type, référence) conformément aux directives esquissées ci-dessus. L'annotation a consisté en une validation ou correction manuelle dans un éditeur XML du résultat de SxPipe/NP utilisé comme pré-annotateur. Le résultat de cette campagne d'annotation est à considérer comme préliminaire, puisqu'une seule personne a annoté le corpus. Nous sommes bien conscients des problématiques liées à toute tâche d'annotation manuelle, et en particulier à l'annotation en EN, problématiques analysées par exemple par Fort *et al.* (2009). L'effort d'annotation devrait donc être poursuivi, notamment en obtenant une deuxième annotation, qui permettrait la mesure d'accords inter-annotateurs ainsi qu'une adjudication des cas de désaccord.

4 Résultats et perspectives

Au total, 5 890 des 12 351 phrases contiennent au moins une mention d'EN. Au total, 11 636 mentions ont été annotées, qui se répartissent en 3761 noms de lieux, 3357 noms d'entreprises, 2381 noms d'organisations, 2025 noms de personnes, 67 noms de produits, 29 noms de personnages de fiction et 15 POI.

Outre une amélioration de la qualité de l'annotation, comme évoqué ci-dessus, nous prévoyons d'utiliser ce nouveau corpus annoté de multiples façons. Tout d'abord, il pourra être utilisé pour entraîner et évaluer des systèmes de reconnaissance et/ou de résolution d'entités nommées. Une comparaison avec le petit corpus de Stern et Sagot (2010), composé de textes relevant d'un domaine proche mais distinct (dépêches d'agence), sera en ce sens utile. Mais la disponibilité d'informations morphosyntaxiques et syntaxiques permettra des expériences intéressantes impliquant par exemple des modèles joints de reconnaissance d'EN et d'étiquetage morphosyntaxique, ou de reconnaissance d'EN et d'analyse syntaxique.

Remerciements

Ce travail a été financé par le projet ANR EDyLex (ANR-009-CORD-08).

Références

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for French. *In* Abeillé, A., éditeur : *Treebanks*. Kluwer, Dordrecht.

Bechet, F. et Charton, E. (2010). Unsupervised knowledge acquisition for extracting named entities from speech. *In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing*.

BLUME, M. (2005). Automatic entity disambiguation : Benefits to ner, relation extraction, link analysis, and inference. *International Conference on Intelligence Analysis*.

BUNESCU, R. et PASCA, M. (2006). Using encyclopedic knowledge for named entity disambiguation. *In Proceedings of EACL*, volume 6, pages 9–16.

CUCERZAN, S. (2007). Large-scale named entity disambiguation based on wikipedia data. *In Proceedings of EMNLP-CoNLL*, volume 2007, pages 708–716.

DODDINGTON, G., MITCHELL, A., PRZYBOCKI, M., RAMSHAW, L., STRASSEL, S. et WEISCHEDEL, R. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. *In Proceedings of LREC - Volume 4*, pages 837–840.

EHRMANN, M. (2008). *Les Entités Nommées, de la Linguistique au TAL - Statut Théorique et Méthodes de Désambiguïsation*. Thèse de doctorat, Université Paris 7 Denis Diderot.

FINKEL, J. R., GRENAGER, T. et MANNING, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

FORT, K., EHRMANN, M. et NAZARENKO, A. (2009). Towards a methodology for named entities annotation. *In Proceedings of the Third Linguistic Annotation Workshop*, ACL-IJCNLP '09, pages 142–145, Stroudsburg, PA, USA. Association for Computational Linguistics.

GALLIANO, S., GRAVIER, G. et CHAUBARD, L. (2009). The Ester 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts. *In Interspeech 2009*.

GROUIN, C., ROSSET, S., ZWEIGENBAUM, P., FORT, K., GALIBERT, O. et QUINTARD, L. (2011). Proposal for an extension of traditional named entities : From guidelines to evaluation, an overview. *In Proceedings of the Fifth Linguistic Annotation Workshop (LAW-V)*, pages 92–100, Portland, OR. Association for Computational Linguistics.

MARSH, E. et PERZANOWSKI, D. (1998). Muc-7 evaluation of ie technology : Overview of results. *In Proceedings of the Seventh Message Understanding Conference (MUC-7) - Volume 20*.

MCNAMEE, P. et DANG, H. (2009). Overview of the tac 2009 knowledge base population track. *In Text Analysis Conference (TAC)*.

ROSSET, S., GROUIN, C. et ZWEIGENBAUM, P. (2011). Entités nommées structurées : guide d'annotation Quaero. Notes et Documents 2011-04, LIMSI, Orsay, France.

ROSSET, S., ILLOUZ, G. et MAX, A. (2005). Interaction et recherche d'information : le projet Ritel. *Traitement Automatique des Langues*, 46(3):155–179.

SAGOT, B. et BOULLIER, P. (2008). SxPipe 2 : architecture pour le traitement présyntaxique de corpus bruts. *Traitement Automatique des Langues (T.A.L.)*, 49(2):155–188.

SAGOT, B. et STERN, R. (2012). Aleda, a free large-scale entity database for French. *In Proceedings of LREC*. To appear.

SANG, E. F. T. K. et MEULDER, F. D. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. *In Proceedings of CoNLL-2003*, pages pp. 142–147.

SEKINE, S. et NOBATA, C. (2004). Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. *In Proceedings of LREC 2004*, Lisbon, Portugal.

STERN, R. et SAGOT, B. (2010). Resources for named entity recognition and resolution in news wires. *In Proceedings of LREC 2010 Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management*, La Valette, Malte.

Utilisation des fonctions de croyance pour l'estimation de paramètres en traduction automatique

Christophe Servan Simon Petitrenaud
LIUM, Le Mans
christophe.servan@lium.univ-lemans.fr
simon.petit-renaud@lium.univ-lemans.fr

RÉSUMÉ

Cet article concerne des travaux effectués dans le cadre du 7ème atelier de traduction automatique statistique et du projet ANR COSMAT [1]. Ces travaux se focalisent sur l'estimation de paramètres contenus dans une table de traduction. L'approche classique consiste à estimer ces paramètres à partir de fréquences relatives d'éléments de traduction. Dans notre approche, nous proposons d'utiliser le concept de masses de croyance afin d'estimer ces paramètres. La théorie des fonctions de croyances est une théorie très adaptée à la gestion des incertitudes dans de nombreux domaines. Les expériences basées sur notre approche s'appliquent sur la traduction de la paire de langue français-anglais dans les deux sens de traduction.

ABSTRACT

Feature calculation for Statistical Machine Translation by using belief functions

In this paper, we consider the translation of texts within the framework of the 7th Workshop of Machine Translation evaluation task and the COSMAT corpus using a statistical machine translation approach. This work is focused on the translation features calculation of the phrase contained in a phrase table. The classical way to estimate these features are based on the direct computation counts or frequencies. In our approach, we propose to use the concept of belief masses to estimate the phrase probabilities. The Belief Function theory has proven to be suitable and adapted for the management of uncertainties in many domains. The experiments based on our approach are focused on the language pair English-French.

MOTS-CLÉS : Traduction automatique statistique, fonctions de croyance, apprentissage automatique, estimation de paramètres.

KEYWORDS: Statistical machine Translation, belief function, machine learning, feature estimation.

1 Introduction

Il est classique d'utiliser une table de traduction comme élément d'un modèle de traduction automatique statistique (TAS). Dans un système de traduction automatique fondé sur des segments (ou séquences de mots), une table de traduction contient les traductions alternatives et ses probabilités pour un segment en une langue source donnée. Chaque ligne ou événement d'une table de traduction comprend deux segments, l'un en langue source et l'autre en langue cible.

1. http://www.cosmat.fr

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 543–550,
Grenoble, 4 au 8 juin 2012.

On suppose que les événements sont indépendants les uns des autres. Les tables de traduction peuvent contenir beaucoup de paramètres comme les probabilités de traduction des segments ou les probabilités lexicales. Afin d'estimer ces paramètres, les systèmes de TAS utilisent de grands corpus appelés bitextes, qui sont composés de phrases en langue source et en langue cible qui sont la traduction l'une de l'autre. Pour chaque phrase, les mots des deux langues sont alignés en fonction de la traduction.

Dans l'approche classique, l'estimation des probabilités est faite par l'utilisation des fonctions de compte simples, sur la base de fréquences relatives. Dans de nombreux travaux, la théorie des fonctions de croyance (initialement théorie de Dempster-Shafer) permet une représentation à la fois plus souple et plus précise de différents types d'incertitude que les modèles probabilistes (Smets, 1988; Cobb et Shenoy, 2006). Par exemple, les modèles probabilistes peuvent difficilement prendre en compte le conflit entre deux hypothèses différentes de traduction, en particulier dans le cas des exemples rares. Il est également délicat d'estimer le degré de confiance global que l'on a sur l'ensemble des éléments de traduction pour une source donnée.

La théorie des fonctions de croyance est capable de traiter ce genre de situations en fournissant des degrés de conflit quand il y a des hypothèses contradictoires, ainsi que des mesures de confiance globale. Dans cet article, nous proposons une méthode originale pour estimer les paramètres associés aux événements constitués de paires de segments à l'aide des fonctions de croyance.

Cet article présente nos premiers travaux et leurs résultats réalisés avec cette nouvelle approche. Il est organisé comme suit : tout d'abord, nous rappelons brièvement la théorie de la traduction automatique statistique. Dans la section 3, nous détaillons notre approche basée sur les fonctions de croyance. Ensuite, nous présentons des expériences sur différents corpus de traduction français-anglais. Enfin, nous concluons cet article et proposons quelques perspectives.

2 Estimation de paramètres en traduction automatique statistique

Soit une phrase source s traduite en un certain nombre de phrases cibles $t \in T_s$, où T_s est l'ensemble de toutes les traductions observées de s dans la table de traduction. Le modèle de traduction automatique statistique (TAS) utilise un ensemble de n fonctions $f_i, i = 1 \ldots, n$, qui dépendent des séquences de mots sources et cibles, afin de déterminer la meilleure traduction. Les fonctions que l'on considère habituellement comprennent le modèle de traduction, le modèle de distorsion, le modèle de langage cible et différentes pénalités. Parmi toutes les traductions possibles, celle qui sera choisie maximise la probabilité *a posteriori*, et peut s'exprimer de la façon suivante :

$$t^* = \arg \max_{t=1 \in T_s} \log \left(\prod_{i=1}^{n} f_i(t,s)^{\lambda_i} \right), \quad (1)$$

où chaque paramètre λ_i est un coefficient de pondération pour chaque fonction f_i (Och, 2003). Ces poids sont généralement optimisés de façon à maximiser la performance de traduction sur

langue source (s) - fr	langue cible (t) - en
...	...
étant donné un	given a
étant donné un	starting from an
étant donné	given
étant donné	given
étant donné	starting from
étant donné	starting
étant	starting
...	...

Tableau 1 – Exemple de paires de segments extraits d'un bitexte

des ensembles de données de développement. Le travail présenté dans cet article se focalise sur les caractéristiques utilisées pour estimer le modèle de traduction.

Dans l'outil de traduction classique « Moses » (Koehn *et al.*, 2007), la table de traduction contient cinq caractéristiques (Koehn, 2010) : les paramètres de traduction des segments, la pondération lexicale des traductions et la pénalité des segments. Les paramètres de traduction des segments sont estimés à l'aide de la règle de décision de Bayes dans les deux sens de traduction. Les poids lexicaux sont estimés à partir du modèle IBM 1 basés sur les mots de chaque paire de segment. Enfin, la pénalité d'apparition du segment est définie par l'utilisateur. Cette fonction permet de privilégier les segments en fonction de leur longueur et prend une valeur constante ρ pour tous les segments. Si $\rho > e$, on préférera des segments longs aux segments courts. Inversement, si $\rho < e$, les segments courts seront privilégiés. Une fois que ces paramètres sont définis, les poids de l'ensemble de ces caractéristiques sont optimisés au cours du processus d'entraînement par le taux d'erreur minimum (MERT) (Och, 2003).

Dans cet article, nous nous concentrons sur l'estimation des paramètres associés aux segments de traduction dans les deux sens de traduction, nous estimons les probabilités lexicales de manière classique et, enfin, nous fixons la pénalité d'apparition ρ à la valeur e.

Le tableau 1 donne un exemple de paires de segments extraits d'un bitexte. A partir de cet exemple, nous obtenons la table de traduction présenté dans le tableau 2 contenant l'estimation des différentes probabilités. Ainsi, la probabilité de traduction de « starting » sachant « étant donné » est une simple fréquence conditionnelle égale à 0,25 et la probabilité de « given » sachant « étant donné » est égale à 0,5. La probabilité conditionnelle inverse du segment de traduction est estimée de la même manière.

Cette façon d'estimer les paramètres a quelques inconvénients. Lorsque certaines paires de segments apparaissent plusieurs fois, comme la paire « *la maison blanche*|*the white house* », et n'ont pas d'occurrences concurrentes, l'estimation de la probabilité du segment est égale à 1, mais dans d'autres situations, des événements peuvent survenir très rarement et être ambigus. Par exemple, supposons que pour le mot français « *chien* » (qui devrait se traduire par « *dog* » en anglais),

segment source (s) - fr	segment cible (t) - en	$p(t\|s)$	$lex(t\|s)$	$p(s\|t)$	$lex(s\|t)$	ρ
...	...					
étant donné	given	0,5	0,060147	0,333333	0,306373	2,718
étant donné	starting	0,25	7,15882e-06	0,333333	5,19278e-05	2,718
étant donné	starting from	0,25	7,15882e-06	0,333333	0,0277778	2,718
...	...					

Tableau 2 – Exemple de table de traduction avec les différents paramètres

deux occurrences contradictoires soient disponibles dans la table de traduction : « $chien|cat$ » et « $chien|dog$ ». Pour chacun de ces deux événements, l'estimation de la probabilité peut être égale à 1, car ils n'ont été observés qu'une seule fois. Par exemple, dans le cadre de notre expérience sur le corpus COSMAT, il existe 13 480 cas correspondant à 33 900 entrées sur les 363 324 que compte la table de traduction, soit un peu moins de 10%, dans le sens de traduction français-anglais.

Même si l'estimation de la probabilité de la traduction des paires *inversées* « $cat|chien$ » et « $dog|chien$ » peut équilibrer ce problème, si l'événement n'est observé qu'une seule fois dans les deux sens de traduction, l'estimation des probabilités conditionnelles inversées est inutile. Il existe la possibilité de lisser les probabilités de l'ensemble des événements (Foster *et al.*, 2006). Cependant, les approches de lissage optent pour une redistribution des estimations afin de donner, notamment, une probabilité non nulle aux événements non observés (Chen et Goodman, 1996; Goodman, 2001). Notre but n'est pas celui-ci, mais plutôt de proposer une approche différente de l'estimation des paramètres des événements observés.

L'utilisation de théories alternatives à la théorie des probabilités permet de mieux ajuster ces estimations. L'une d'elles est particulièrement adaptée à la gestion de différents types d'incertitudes : la théorie des fonctions de croyance, qui a été proposée puis développée depuis une trentaine d'années. Cette théorie a été appliquée avec succès à de nombreux domaines tels que l'identification du locuteur (Petitrenaud *et al.*, 2010) ou la classification en général (Elouedi *et al.*, 2000). Dans nos travaux, nous nous utilisons certains concepts fondamentaux de cette théorie pour notre problème d'estimation de paramètres.

3 Fonctions de croyances pour les systèmes de TAS

Dans cette section, nous présentons brièvement quelques notions de la théorie des fonctions de croyance (Shafer, 1976; Smets et Kennes, 1994) et nous l'appliquons au problème d'estimation de paramètres de modèles de traduction. Dans cet article, nous adoptons le point de vue proposé par Smets : le modèle de croyances transférables (MCT) (Smets et Kennes, 1994). L'objectif de ce modèle est de déterminer la croyance concernant différentes propositions, à partir d'informations disponibles.

Soit Ω un ensemble fini, appelé cadre de discernement de l'expérience. La représentation de l'incertitude est faite par le biais de la notion de fonction de croyance, définie comme une fonction m de 2^Ω sur $[0,1]$ telle que $\sum_{A\subseteq\Omega} m(A) = 1$. La quantité $m(A)$ représente la croyance allouée à la proposition A, et à aucune proposition plus restrictive. Une des opérations les plus importantes dans le MCT est la procédure d'agrégation des informations, c'est-à-dire la combinaison de plusieurs fonctions de croyance définies dans un même cadre de discernement (Smets et Kennes, 1994). En particulier, la combinaison de deux fonctions de croyance m_1 et m_2 indépendantes définies sur Ω est faite en utilisant l'opérateur binaire conjonctif $\cap$, tel que $m' = m_1 \cap m_2$ (Smets et Kennes, 1994) :

$$\forall A \subseteq \Omega, m'(A) = \sum_{B\cap C=A} m_1(B)m_2(C) \tag{2}$$

Cet opérateur est associatif et commutatif, il est alors possible de définir la combinaison de n fonctions $m_1, \ldots, m_n$ sur Ω par la fonction de croyance $m = m_1 \cap \ldots \cap m_n$. Cette dernière fonction m capture l'information globale sur l'ensemble des expériences connues.

Ici, nous proposons d'utiliser le MCT pour estimer les paramètres de traduction des segments. Tout d'abord, pour une source s, chaque cible $t_i \in T_s$ donne une information particulière pour la traduction qui peut être décrite par une fonction de croyance m_s^i, telle que :

$$\begin{cases} m_s^i(\{t_i\}) = p(t_i|s) \\ m_s^i(T_s) = \overline{p(t_i|s)} \end{cases}, \tag{3}$$

où $\overline{p(t_i|s)} = 1 - p(t_i|s)$. Si nous combinons les informations définies par toutes les hypothèses disponibles dans la table concernant la traduction de s, à partir de l'opérateur conjonctif défini dans l'équation 2, nous obtenons alors une fonction de croyance $m_s = \cap_{t \in T_s} m_s^i$. La masse de t_i est obtenue par la formule suivante :

$$m_s(\{t_i\}) = p(t_i|s). \prod_{t_k \in T_s \setminus \{t_i\}} \overline{p(t_k|s)}. \tag{4}$$

Notons que généralement $\sum_{t_i \in T_s} m_s(\{t_i\}) = 1 - m(T_s) - m(\emptyset) < 1$. Les masses $m(T_s)$ et $m(\emptyset)$ peuvent être respectivement interprétées comme le degré d'ignorance et le degré de conflit d'informations concernant la traduction de s. Même si $m(\emptyset)$ n'entre pas directement dans notre modèle de traduction, quand il y a un conflit important entre plusieurs hypothèses de traduction, les masses de croyance sur chacun des singletons $t_i \in T_s$ s'affaiblissent. Nous obtenons alors une estimation de la fonction définie dans l'équation 2 par : $f(t_i, s_j) = m_{s_j}(\{t_i\})$. De la même manière, l'estimation de la fonction inverse est obtenue par l'équation suivante :

$$m_t^I(\{s_j\}) = p(s_j|t). \prod_{s_k \in S_t \setminus \{s_j\}} \overline{p(s_k|t)}\,, \tag{5}$$

où S_t est l'ensemble des sources possibles de la cible t. Si nous appliquons ces formules à l'exemple du tableau 1, une nouvelle estimation des paramètres associés aux différentes paires de segments est calculée dans le tableau 3.

$m_s(\text{starting}) = p(\text{starting}\|\text{étant donné}) \cdot \overline{p(\text{given}\|\text{étant donné})} \cdot \overline{p(\text{starting from}\|\text{étant donné})}$
$m_s(\text{starting}) = 0,09375$
$m_s(\text{starting from}) = 0,09375$
$m_s(\text{given}) = 0,28125$
$m_s(T_s) = 0,28125$
$m_s(\emptyset) = 0,25$

Tableau 3 – Exemple d'estimation de paramètres de paires de segments à l'aide du MCT ($s =$ « étant donné »)

Notons que si $p(t_i|s) = 1$, les masses de croyance pour les autres hypothèses deviennent nulles (cf. équation 4). La masse de croyance indiquée dans cette équation peut alors être modifiée de

corpus	AbsTrain		AbsDev		AbsTest		nc7		eparl7		nwtst2010		nwtst2011	
langue	fr	en	fr	en	fr	en	fr	en	fr	en	fr	en	fr	en
# de phrases	5141		1083		1102		137k		2M		2489		3003	
# de mots	135K	120K	28K	25K	28K	25K	4M	3,4M	61,7M	55,7M	62k	70k	75k	84,5k

Tableau 4 – Description des bitextes.

façon suivante :

$$m_s(\{t_i\}) = \frac{1}{1+\frac{1}{|s|}}\ , \tag{6}$$

où $|s|$ désigne le nombre d'occurrences de s. Ainsi, $m_s(\{t_i\}) < 1$ mais plus on a d'information sur s, plus $m_s(\{t_i\})$ tendra vers 1. Enfin, les phrases cibles choisies sont obtenues par le processus de décision défini par l'équation 2.

4 Expériences

		nc7		eparl7-nc7	
	approche	BLEU	TER	BLEU	TER
Sens de la traduction : fr↦en					
newstest2010	prob.	**24,58 (0,13)**	**57,53 (0,03)**	**27,22 (0,05)**	**57,52 (0,10)**
	MCT	24,56 (0,08)	57,66 (0,07)	27,10 (0,10)	57,74 (0,10)
newstest2011	prob.	**25,92 (0,11)**	**54,48 (0,09)**	**29,52 (0,12)**	**55,08 (0,12)**
	MCT	25,83 (0,17)	54,61 (0,08)	29,47 (0,14)	55,28 (0,13)
Sens de la traduction : en↦fr					
newstest2010	prob.	24,75 (0,06)	60,17 (0,26)	**28,04 (0,07)**	**53,77 (0,14)**
	MCT	**24,74 (0,04)**	**60,07 (0,18)**	28,00 (0,03)	53,76 (0,03)
newstest2011	prob.	26,84 (0,19)	57,75 (0,29)	28,60 (0,25)	52,85 (0,34)
	MCT	**26,93 (0,09)**	**57,63 (0,17)**	**28,60 (0,04)**	**52,74 (0,04)**

Tableau 5 – Résultats obtenus suivant les métriques BLEU et TER avec deux systèmes entraînés sur les corpus : News-Commentary 7 (*nc7*) ; Europarl 7 - News-Commentary 7 (*eparl7-nc7*).

Afin de valider notre méthode, plusieurs expériences ont été réalisées. Tout d'abord, nous avons utilisé le corpus COSMAT, qui est un ensemble de bitextes de résumés de thèses de doctorat en français et en anglais. Puis, nos expériences ont été placées dans le contexte de l'évaluation du septième atelier sur la traduction automatique statistique (WMT12).

4.1 Le Corpus COSMAT

Le projet ANR COSMAT est composé de nombreux résumés de thèse de doctorat en français et en anglais. Ces résumés ont été classés en fonction de plusieurs thèmes. Dans nos expériences, nous n'avons retenu que le domaine associé à l'informatique. Les corpus d'apprentissage, de développement et de tests sont décrits dans le tableau 4.

Sur le corpus de développement, la perplexité des modèles de langage cible est de 122 pour le français et de 196 pour l'anglais. Les modèles sont adaptés à la tâche grâce à l'utilisation du corpus d'entraînement (AbsTrain) et des modèles de langage.

Sens de traduction		fr→en		en→fr	
corpus	approche	BLEU	TER	BLEU	TER
AbsDev	prob.	34,78 (0,09)	48,24 (0,29)	**32,28 (0,02)**	52,82 (0,25)
	belief	**34,85 (0,06)**	**48,25 (0,11)**	32,28 (0,01)	**52,40 (0,18)**
AbsTest	prob.	40,03 (0,44)	44,35 (0,12)	**38,80 (0,19)**	47,76 (0,36)
	belief	**40,44 (0,10)**	**44,18 (0,12)**	38,43 (0,12)	**47,66 (0,10)**

Tableau 6 – Résultats obtenus avec le corpus COSMAT suivant les métriques BLEU et TER.

4.2 Le corpus WMT12

Le cadre utilisé pour l'évaluation de WMT12 contient plusieurs corpus. Ceux que nous avons utilisés dans nos expériences sont décrits dans le tableau 4. Les corpus d'apprentissage sont Europarl 7 (eparl7) et News-Commentary 7 (nc7). Les modèles employés quand la langue cible est le français et l'anglais ont respectivement une perplexité de 123 et de 169.

4.3 Résultats

Les tableaux 6 et 5 contiennent les résultats obtenus avec l'approche classique et avec notre approche basée sur les fonctions de croyance. Les métriques utilisées sont le score BLEU (Papineni *et al.*, 2002) et la métrique TER (Snover *et al.*, 2005). Afin de garantir une certaine robustesse des résultats, trois optimisations de MERT ont été faites. Le résultat présenté correspond à une moyenne de ces trois optimisations et la valeur indiquée entre parenthèses est l'écart-type. La pénalité de brièveté (ou de longueur de phrase) associée au score BLEU est d'environ 0,99 (0,01) pour les deux approches, dans les deux sens de traductions et pour chacune des expériences.

Les expériences menées sur COSMAT et sur WMT12 montrent que notre nouvelle approche semble avoir des résultats similaires à ceux de l'approche classique. Toutefois, le score BLEU a tendance à être plus faible dans notre approche lorsque le sens de la traduction est de l'anglais vers le français dans l'expérience avec le corpus WMT12 mais à l'inverse, dans l'expérience COSMAT, notre nouvelle approche est légèrement moins performante dans le sens français vers anglais. Malgré ce constat, ces premiers résultats sont encourageants et nous poussent à poursuivre dans cette direction.

5 Conclusions et perspectives

Cet article présente les premiers résultats sur l'utilisation du Modèle des Croyances Transférables (MCT) en traduction automatique statistique. Cette théorie a été utilisée pour estimer différemment les paramètres des paires de segments de traduction. Les résultats obtenus dans la traduction français-anglais, dans les deux directions, sur les corpus COSMAT et WMT12 sont encourageants. Prochainement, nous prévoyons d'appliquer le MCT en traduction de manière plus approfondie. D'abord, nous allons étendre cette approche à l'estimation des paramètres de pondération lexicale. Nous allons également orienter nos recherches vers une stratégie de prise en compte de la proximité linguistique des différentes hypothèses de traduction pour une phrase donnée. Pour reprendre l'exemple du tableau 1, « $starting$ » serait notamment plus proche de « $starting\ from$ » que de « $given$ ». Le MCT permet d'intégrer ce genre de situations avec une certaine souplesse.

6 Remerciements

Ce travail a été financé par l'Agence Nationale de la Recherche dans le cadre du projet COSMAT et par la Commission Européenne à travers le projet EUROMATRIXPLUS.

Références

CHEN, S. F. et GOODMAN, J. (1996). An empirical study of smoothing techniques for language modeling. *In* JOSHI, A. et PALMER, M., éditeurs : *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310–318, San Francisco, Californie, États-Unis d'Amérique. Morgan Kaufmann Publishers.

COBB, B. R. et SHENOY, P. P. (2006). A comparison of methods for transforming belief function models to probability models. *International Journal of Approximate Reasoning*, 41(3):255–266.

ELOUEDI, Z., MELLOULI, K. et SMETS, P. (2000). Classification with belief decision trees. *In Proceedings of the 9th International Conference on Artificial Intelligence : Methodology, Systems, Architectures.* AIMSA 2000, Springer Lecture Notes on Articial Intelligence.

FOSTER, G., KUHN, R. et JOHNSON, H. (2006). Phrasetable smoothing for statistical machine translation. *In Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 53–61.

GOODMAN, J. T. (2001). A bit of progress in language modeling. *Computer Speech &amp ; Language*, 15(4):403 – 434.

KOEHN, P. (2010). *Statistical MAchine Translation*. Cambridge University Press.

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A. et HERBST, E. (2007). Moses : Open source toolkit for statistical machine translation. *In Proc. of the 45th Annual Meeting of the Association for Computational Linguistics (Demo and Poster Sessions)*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

OCH, F. J. (2003). Minimum error rate training in statistical machine translation. *In Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

PAPINENI, K., ROUKOS, S., WARD, T. et ZHU, W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. *In Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, USA.

PETITRENAUD, S., JOUSSE, V., MEIGNIER, S. et ESTÉVE, Y. (2010). Automatic named identification of speakers using belief functions. *In Information Processing and Management of Uncertainty (IPMU'10)*.

SHAFER, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

SMETS, P. (1988). Belief functions versus probability functions. pages 17–24.

SMETS, P. et KENNES, R. (1994). The transferable belief model. *Artificial Intelligence*, 66:191–234.

SNOVER, M., DORR, B., SCHWARTZ, R., MAKHOUL, J., MICCIULA, L. et WEISCHEDEL, R. (2005). A study of translation error rate with targeted human annotation. Rapport technique LAMP-TR-126,CS-TR-4755,UMIACS-TR-2005-58, University of Maryland, College Park and BBN Technologies.

La longueur des tours de parole comme critère de sélection de conversations dans un centre d'appels

Philippe Suignard[1] *Frederik Cailliau*[2] *Ariane Cavet*[2]
(1) EDF R&D, 1, avenue du Général de Gaulle, 92141, Clamart
(2) Sinequa, 12 rue d'Athènes, 75009 Paris
Philippe.Suignard@edf.fr, cailliau@sinequa.com, cavet@sinequa.com

RESUME

Cet article s'intéresse aux conversations téléphoniques d'un Centre d'Appels EDF, automatiquement découpées en « tours de parole » et automatiquement transcrites. Il fait apparaître une relation entre la longueur des tours de parole et leur contenu, en ce qui concerne le vocabulaire qui les compose et les sentiments qui y sont véhiculés. Après avoir montré qu'il y a un intérêt à étudier ces longs tours, l'article analyse leur contenu et liste quelques exemples autour des notions d'argumentation et de réclamation. Il montre ainsi que la longueur des tours de parole peut être un critère utile de sélection de conversations.

ABSTRACT

Turn-taking length as criterion to select call center conversations

This article focuses on telephone conversations collected in an EDF Call Center, automatically segmented in "turn-taking" and automatically transcribed. It shows a relationship between the length of the turns and their content regarding the vocabulary and the feelings that are conveyed. After showing that there is an interest in studying these long turns, the article analyzes their content and lists some examples around the notions of argumentation and claim. It shows that the length of turns can be a useful criterion for selecting conversations.

MOTS-CLES : Centre d'appels, Conversation, Tour de parole, Reconnaissance de Parole.
KEYWORDS : Call Center, Conversation, Turn Taking, Automatic Speech Recognition.

1 Introduction

Avec plus de 30 millions de clients et plusieurs milliers de conseillers en ligne, les Centres d'Appels constituent, pour EDF, un maillon important de la Gestion de la Relation Client et font l'objet d'un suivi permanent avec un focus sur la « professionnalisation des conseillers », consistant à améliorer leur pratique professionnelle afin de toujours mieux répondre aux clients. Cette amélioration passant par des analyses qualitatives, seul un faible pourcentage peut être retenu pour écoute, d'où l'importance des critères de sélection.

C'est dans cet esprit qu'EDF R&D a participé aux projets Infom@gic/Callsurf et Voxfactory. Le projet Callsurf (Garnier *et al.*, 2008 ; Bozzi *et al.*, 2009) consistait à enregistrer et transcrire automatiquement les conversations entre clients et conseillers, pour ensuite les analyser. Le projet Voxfactory y ajoute la détection automatique de l'émotion véhiculée, à partir du texte (Cailliau et Cavet, 2010), et par le signal (Devillers *et al.*, 2010).

Cet article s'intéresse à la notion de « tours de parole » (Sacks *et al.*, 1974) et à leur longueur mesurée en seconde, une information qui ne semble pas encore avoir fait l'objet d'étude,

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 551–558,
Grenoble, 4 au 8 juin 2012.

contrairement à la longueur des phrases, en nombre de mots, dans des données textuelles plus conventionnelles comme l'étude du théâtre du XVII[e] siècle par (Labbé et Labbé, 2010).

La suite de l'article s'intéresse à la relation entre la longueur des tours de parole et le vocabulaire qui les compose, ainsi qu'à la relation entre la longueur de ces tours et les sentiments qu'ils véhiculent. Elle montre que lorsque la longueur du tour a tendance à augmenter, les informations trouvées semblent plus chargées émotionnellement et le vocabulaire employé a tendance à devenir plus pertinent d'un point de vue métier.

La partie 2 présente la notion de tour de parole et la partie 3 présente le corpus. La partie 4 s'intéresse à la relation entre longueur du tour de parole et vocabulaire, tandis que la partie 5 décrit la relation entre longueur et sentiment. La dernière partie analyse leur contenu et liste quelques exemples autour des notions d'argumentation et de réclamation.

2 Notion de tour de parole

La notion de « tour de parole » (TDP) correspond à la prise de parole par un locuteur et désigne le temps pendant lequel il garde cette parole. Au final, la suite ordonnée des TDP va constituer une conversation (Vincent, 2002). Le tour de parole semble donc être une notion assez simple, mais bien que beaucoup utilisé en « analyse conversationnelle », il reste sujet à interprétation et beaucoup d'interrogations à son sujet subsistent. Différents travaux se demandent encore « Qu'est-ce que vraiment un TDP ? » (Laforest, 2011).

Sur la figure 1, l'exemple (1) semble présenter trois TDP, mais l'intervention de « l'agent » n'est qu'un simple « back-channel » pour manifester son attention au client. En analyse conversationnelle, on pourrait regrouper ces trois tours en un seul et les considérer comme étant une seule unité. Dans l'exemple (2), le client a du mal à trouver le nom d'un terme métier, que lui souffle l'agent. Ici aussi on peut considérer qu'il s'agit d'un seul tour de parole et qu'il est co-construit par les deux locuteurs.

(1) Client : « … oui, je vous appelle… » Agent : « oui » Client « … pour un problème… »	(2) Client : « … et quand je vais la recevoir, la … » Agent : « la facture rectificative ? »

FIGURE 1 - Deux exemples de tours de parole

Sans vouloir éluder la difficulté de définir précisément ce « tour de parole », nous nous en affranchissons un peu dans cette étude, dans la mesure où nous nous intéressons à des conversations automatiquement découpées en TDP par un segmenteur et retranscrites par un procédé de transcription automatique de la parole (Adda *et al.*, 2011). Au final, le découpage en TDP est imparfait mais reflète la réalité des conversations (répétitions, disfluences, parole superposée, etc.) ainsi que les conditions d'enregistrement (voix sur un seul canal, bruit de fond, téléphone portable, etc.).

Plusieurs raisons peuvent expliquer la variation de longueur des TDP : un client peut monopoliser la parole pour exprimer un problème, une réclamation, une insatisfaction, etc., mais un conseiller peut également monopoliser la parole et prendre du temps pour répondre au client (parce qu'un problème peut être compliqué à résoudre par exemple). Enfin, une discussion peut aussi contenir des passages « serrés » ou « tendus » (énervements, émotions…), ce qui peut empêcher le segmenteur de détecter le changement

d'interlocuteur. D'un certain point de vue, cela peut être considéré comme un point faible du système, mais de l'autre c'est également un marqueur du fait qu'il se passe quelque chose de particulier et qu'il s'agit donc d'un phénomène intéressant à étudier.

3 Présentation du corpus

Pour réaliser notre étude, un corpus de conversations téléphoniques a été enregistré dans le Centre d'Appels Bleu Ciel d'Aix-en-Provence. Les enregistrements ont eu lieu entre janvier et février 2010, auprès d'une quinzaine de conseillers volontaires. Comme la plupart des enregistreurs du marché, l'enregistreur de conversation utilisé est mono-canal, ce qui signifie que les deux signaux de parole du client et du conseiller se superposent quand ils parlent en même temps. Une fois enregistrés, les appels font l'objet d'une série de traitements. D'abord intervient le « segmenteur », qui a pour but de séparer le signal en segments qui, idéalement, correspondent aux tours de parole du client et du conseiller. Ensuite, les TDP sont transcrits par un processus de transcription automatique (avec un taux d'erreur d'environ 30 %).

Au total, ce corpus est constitué de 8 551 conversations, composées de 800 596 tours de parole. La durée moyenne d'un TDP est de 3,7 s, le plus long dure plus de 2 min. La répartition du nombre de tours en fonction de leur durée est présentée en figure 2 (échelle linéaire à gauche et logarithmique à droite) :

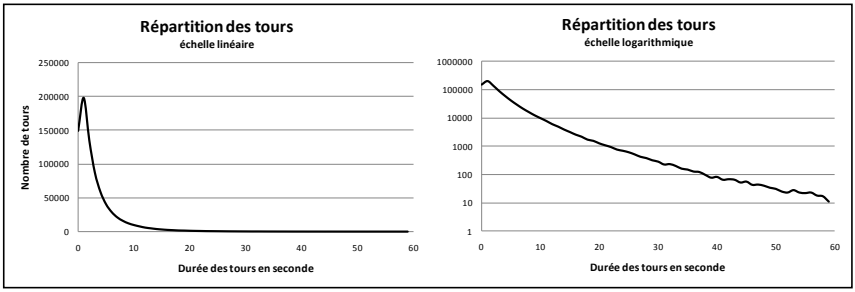


FIGURE 2 - Distribution des tours de parole en fonction de leur durée

Ces courbes montrent que la très grande majorité des TDP dure moins de 10 s et que leur nombre chute fortement en fonction de la durée. Néanmoins, il y a quand même des tours longs, voire très longs. Si l'on considère les TDP dont la durée est supérieure à 20 s, il n'y en a « que » 1,2 % en nombre mais 9 % en durée. Pour un seuil de 40 s, on en trouve 0,12 % en nombre et 1,7% en durée.

Quand on regarde ces données au niveau des conversations, on s'aperçoit que le nombre de conversations ayant au moins un tour de plus de 20 s est de 3 904 soit 45 % et que le nombre de conversations ayant au moins un tour de plus de 40 s est de 674 soit 7,9 %. En conclusion, on peut dire que les longs TDP sont peu fréquents par rapport à l'ensemble des TDP, mais non négligeables si on prend en compte leur durée totale.

4 Longueur des tours et vocabulaire : une relation ?

De manière assez intuitive, les TDP très courts semblent peu porteurs d'information. Dans ceux-ci, on trouvera beaucoup de : « oui », « non », « EDF bleu ciel bonjour », etc. Par contre, dans des TDP plus longs, une conversation peut s'installer, le client peut présenter la raison de son appel, le conseiller va répondre à sa problématique, faire des propositions, etc. La question qu'on se pose ici peut se résumer de la manière suivante : « Est-ce que l'on parle des mêmes choses dans les tours longs que dans les autres ? »

Pour répondre à cette question, la méthode va être la suivante. Pour une durée d donnée, on va constituer deux corpus : $T_{inf}(d)$, le corpus constitué des TDP dont la durée est inférieure à d et $T_{sup}(d)$, le corpus constitué des TDP dont la durée est supérieure à d. Puis, nous calculons la distance entre ces deux corpus comme décrit dans (Labbé et Labbé, 2003) :

$$dist(d) = \sum_{m \in T_{inf}(d) \cup T_{sup}(d)} \left| f\left(m, T_{inf}(d)\right) - f\left(m, T_{sup}(d)\right) \right| \quad et \quad f\left(m, T(d)\right) = \frac{occ(m, T(d))}{|T(d)|}$$

Avec $f(m, T(d))$ la fréquence du mot m dans le corpus T(d), $occ(m, T(d))$, le nombre d'occurrences du mot m dans le corpus $T(d)$ et $|T(d)|$, le nombre total de mots dans le corpus $T(d)$. Pour le calcul des mots m, deux alternatives sont retenues : les unigrammes et les bigrammes étendus.

Pour les unigrammes, on constitue la liste de tous les mots du corpus sauf ceux faisant partie d'une « stop liste », comme « le », « la », « les », etc. Pour les bigrammes étendus, on constitue la liste de toutes les suites de deux mots dont aucun des deux n'appartient à la « stop liste » et de toutes les suites de trois mots composées d'une préposition au milieu et dont les deux mots extrémités sont absents de la « stop liste ». On détecte ainsi les bigrammes comme « heures pleines », « relevé de compteur », « pompe à chaleur », etc.

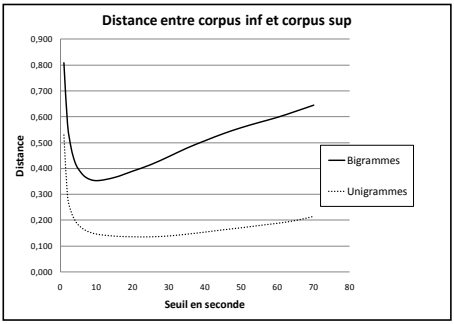


FIGURE 3 – Distance entre corpus inférieur et supérieur à une certaine durée

La figure 3 montre que les deux courbes de distance (pour les unigrammes et les bigrammes) ont globalement la même allure. Quand le seuil de séparation entre les deux corpus est très petit, la distance entre les deux corpus est élevée ce qui semble normal, puisque dans les tours les plus courts, on trouve les « oui », « non », « bonjour », etc. Ensuite cette distance décroît pour atteindre un minimum situé entre 10 s et 20 s. Puis l'écart croit

de nouveau.

Ce graphique montre qu'il y a bien une relation entre la longueur des TDP et le vocabulaire de ceux-ci. C'est vers 15 s que se situe une sorte d'équilibre, c'est-à-dire que c'est pour un seuil de 15 s que le corpus des TDP inférieurs à cette durée ressemble le plus au corpus des TDP supérieurs à cette durée. Cela justifie également l'idée d'aller étudier les tours ayant une durée supérieure à 20 s pour y trouver des particularités ou spécificités de vocabulaire.

5 Longueur des tours et sentiment : une relation ?

5.1 Trouver du sentiment dans un tour de parole

Notre analyse du sentiment est faite en trois phases : détection et normalisation des expressions, calcul d'un poids positif et négatif pour chaque tour de parole et calcul des zones positif et négatif. Les deux premières phases sont détaillées dans (Cailliau et Cavet, 2010) et seront brièvement rappelées ici.

Nous appliquons tout d'abord un ensemble de grammaires faites manuellement à partir d'une fouille approfondie du corpus. Elles détectent les mots et expressions porteurs de sentiment dans le cadre typique de la conversation téléphonique pour un total de 30 types d'entités. A chaque type nous avons attribué un poids déterminé de façon empirique, ainsi qu'une orientation positive, négative ou neutre. Ensuite chaque tour de parole obtient un score de polarité positif et un score de polarité négatif correspondant à la somme des poids des entités repérées dans le tour. Les poids des extractions neutres sont ajoutés au plus haut score positif ou négatif. L'exemple suivant d'un tour de parole contient plusieurs entités, avec le calcul des poids positif et négatif, illustré en figure 4.

« Oui oui mais il non mais c'est ça **OK** **tant mieux** ailleurs parce que **sinon** ça serait **dur** à sortir »

Extraction	Classe et sous-classe	Poids positif	Poids négatif
OK	Acceptation – Refus : acceptation	2	-
tant mieux	Appréciation : favorable (émotif)	4	-
sinon	Accord – Désaccord : rectificatif	-	1
dur	Appréciation : défavorable (émotif)	-	4
Poids total du TDP		6	5

FIGURE 4 – Poids des entités repérées dans le tour de parole

5.2 Du tour de parole au passage

Comme le montrent indépendamment (Cailliau et Cavet, 2010) et (Danesi et Clavel, 2010), le taux d'erreur de la transcription automatique d'environ 30 % impacte directement les extractions. Pour minimiser cet impact, nous privilégions le passage au TDP, en partant du principe qu'un TDP sentimentalement marqué apparaît rarement isolé. Pour ce faire, nous avons mis en place un lissage par une fenêtre coulissante de 5 tours de parole.

Un algorithme de zonage nous permet de transformer la courbe en zones neutre (▭), positive (▩), négative (▬), et très négative (■). Les seuils ont été définis de façon empirique et la barre colorée est obtenue par projection des zones sur l'échelle temporelle.

La figure 5 présente les courbes positive et négative obtenues sur une conversation après lissage sur l'échelle du tour de parole et sur l'échelle de temps, avec la barre colorée correspondante. Le pic négatif apparaissant au début de la conversation occupe relativement peu de tours de parole (17 TDP sur 55, soit 30,9 %), mais il occupe beaucoup de temps sur la conversation (environ 300 s sur 590 s, soit 50,8 %).

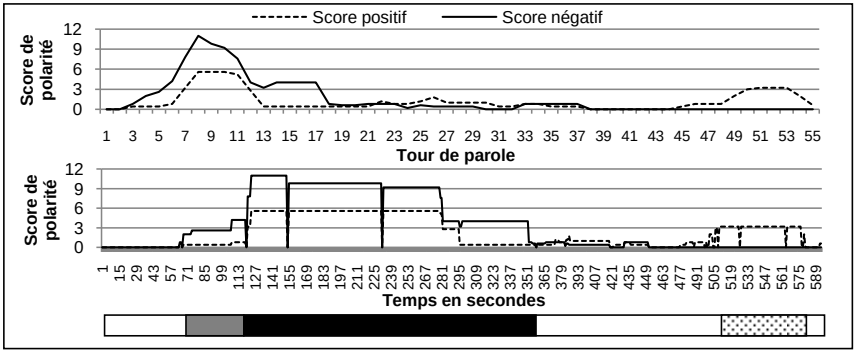


FIGURE 5 – Courbes positive et négative d'une conversation et zones correspondantes

5.3 Longueur des tours de parole et couleur associée

Si on conserve les seuils de 20 s et 40 s utilisés en partie 3, on constate, en figure 6, que les TDP courts apparaissent très majoritairement dans des zones non marquées (95 %). Le nombre de TDP dans des zones non neutres augmente fortement avec la longueur de ces TDP, au détriment du non marqué. Les TDP longs contiennent donc généralement plus d'expressions émotionnelles que les TDP courts. Ces résultats ne sont pas si surprenants : notre méthode de calcul privilégie les TDP plus longs car nous utilisons la somme des poids des entités extraites pour calculer les poids positif et négatif des TDP. On remarque également qu'il ne suffit pas d'avoir un TDP long pour avoir du sentiment : 45 % d'entre eux sont neutres, et moins de 8 % sont marqués comme « très négatifs ».

Durée des TDP	Neutre	Positif	Négatif	Très négatif	
< 20 s	95,3 %	3,0 %	1,6 %	0,1 %	100 %
entre 20 s et 40 s	74,8 %	13,1 %	10,4 %	1,7 %	100 %
> 40 s	45,0 %	25,5 %	21,6 %	7,8 %	100 %

FIGURE 6 – Longueur des TDP et émotion détectée

6 Trouver la spécificité de vocabulaire par régression linéaire

Les deux parties précédentes ont permis de montrer qu'il y avait une spécificité des longs TDP par rapport aux autres : ils sont plus « chargés » en émotion et ont une spécificité en termes de vocabulaire. Dans cette partie, nous appliquons une méthode basée sur la fréquence des mots pour identifier lesquels sont à l'origine de cette spécificité.

On commence par éliminer les tours très courts (inférieurs à 10 s, parce que très nombreux et très peu porteurs d'informations) et très longs (supérieurs à 80 s, parce que très peu nombreux et pouvant donc perturber fortement les résultats) pour constituer 7 sous-corpus : $T_{10\text{-}20}$, $T_{20\text{-}30}$, $T_{30\text{-}40}$, $T_{40\text{-}50}$, $T_{50\text{-}60}$, $T_{60\text{-}70}$, $T_{70\text{-}80}$, $T_{10\text{-}20}$ étant le corpus constitué des TDP dont la durée est comprise entre 10 s et 20 s. Pour chaque sous-corpus, on calcule les fréquences des mots. On pratique une régression linéaire pour trouver les mots pour lesquels l'augmentation de fréquence est la plus forte. Une analyse des mots présentant la plus forte augmentation permet de discerner deux sphères : la sphère de l'argumentation et la sphère de la réclamation.

La sphère de l'argumentation est caractérisée par une forte présence d'adverbes comme « effectivement », « exactement », « normalement », « maintenant », « directement » et « justement ». Cette présence des adverbes s'explique par le fait que les longs TDP ont tendance à être d'avantage consacrés aux problèmes les plus complexes et qu'ils donnent lieu à des échanges argumentés à la fois du côté du client pour expliquer son problème mais également du côté du conseiller pour justifier une réponse. Voici quelques extraits :

Agent : « Avant, vous étiez dans une période où **effectivement** le client pouvait choisir… **effectivement** ses heures creuses… »

Agent : « L'intervention coûtera xx €, somme que vous n'aurez pas à payer, elle sera prélevée **effectivement** sur la facture… »

Agent : « Une somme de x €, **effectivement** ce n'est pas négligeable… »

Client : « … je suis propriétaire mais la la locataire a dû elle avait téléphoné **justement** à edf pour résilier le contrat et moi en fait c'est quelqu'un qui a signé de le jour que j'ai emménagé et **justement** pour couper edf j'avais dit que **justement** j'avais fait l'ouverture sur internet… »

La seconde sphère est celle de la réclamation, caractérisée par la présence de mots comme « problème », « réponse », ou directement « réclamation », comme par exemple :

Agent : « … d'accord avec tous les points où il y a eu un **problème** d'accord d'accord donc j'ai j'ai bien noté que vous souhaitiez une **réponse** par écrit. »

Ces réclamations vont porter à la fois sur des problèmes techniques et sur des problèmes relationnels. Concernant les aspects techniques, on trouve des mots comme : « technicien », « technique », « raccordement », « disjoncteur », « énergie », « chauffage », « hiver », « index », « chantier », « câble », etc. Ces extraits, concernant l'installation ou le raccordement d'un nouveau compteur, peuvent donner lieu à des passages assez longs.

Agent : « … elle est divisé en 2 parties l'installation donc le réseau edf qui va jusqu'à la partie haute du **disjoncteur** donc tout ce qui est **compteur** et **disjoncteur** c est edf par contre… »

Concernant les aspects relationnels, on trouve des mots comme : « courrier », « fournisseur », « mail », « réponse », « client », « rendez-vous », « montant », etc.

Client : « … donc premièrement on a on a on a insisté d'avoir un **rendez** de **rendez-vous** téléphonique premièrement… mais quand même bon alors on a attendu 2 semaines pour ce **rendez-vous** téléphonique le monsieur il m'a appelé… »

Client : « … ça suffit je vais faire une **réclamation** concernant le le **rendez-vous** qui n'a pas écoutez je n'habite pas sur place … »

Agent : « … en attendant la **réponse** de notre service national consommateurs… »

7 Conclusion

Comme les Centres d'Appels représentent un maillon important de la Gestion de la Relation Client pour EDF, l'amélioration continue de leur performance est un enjeu majeur. Cette amélioration passant par des analyses qualitatives, seul un faible pourcentage peut être retenu pour écoute, d'où l'importance des critères de sélection. En s'appuyant sur un corpus de conversations enregistrées dans un Centre d'Appels EDF et transcrites automatiquement, nous avons montré une relation entre la longueur des tours de parole et leur contenu, à travers le vocabulaire qui les compose et les sentiments qui y sont véhiculés. Par conséquent, la longueur des tours de parole est un critère utile de sélection de conversations. Il peut s'ajouter de façon complémentaire aux autres stratégies de sélection que sont les mots clés (nom d'offre, entreprise concurrente, etc.), les thématiques, les sentiments, etc.

Références

ADDA, G., CAILLIAU, F., DAQUO, A-L, GARNIER-RIZET, M., GUILLEMIN-LANNE, S., SUIGNARD, P. et WAAST-RICHARD, C. (2011). La transcription automatique et la fouille de données conversationnelles pour l'analyse de la relation client. In M. Campedel et P. Hoogstoël (Ed.), *Sémantique et multimodalité en analyse de l'information*. Hermes Lavoisier, Paris.

BOZZI, L., SUIGNARD, P., WAAST-RICHARD, C. (2009). Segmentation et classification non supervisée de conversations téléphoniques automatiquement retranscrites. In *Actes de TALN*, Senlis.

CAILLIAU, F., et CAVET, A. (2010). Analyse des sentiments et transcription automatique : modélisation du déroulement de conversations téléphonique. In *TAL*, 51-3, ATALA.

DANESI, C. et CLAVEL, C. (2010). Impact of spontaneous speech features on business concept detection: a study of call-centre data. In *Proc. of SSCS '10*, ACM, New York, NY, USA.

DEVILLERS, L., VAUDABLE, C et CHASTAGNOL, C. (2010). Real-life emotion-related states detection in call centers: a cross-corpora study. In *INTERSPEECH-2010*, pages 2350-2353.

GARNIER-RIZET, M., ADDA, G., CAILLIAU, F., GUILLEMIN-LANNE, G. et WAAST-RICHARD, C. (2008). CallSurf - Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content. In *Actes de LREC 2008*. Marrakech.

LABBE, C. et LABBE, D. (2003). La distance intertextuelle. In *Corpus*, décembre 2003, `http://corpus.revues.org/index31.html` [consulté le 20/12/2011].

LABBE, C. et LABBE, D. (2010). Ce que disent leurs phrases. In *Proceedings of 10th International Conference Statistical Analysis of Textual Data*, Rome, Italie.

LAFOREST, M. (2011). Trois petits tours et puis s'en vont ou qu'est-ce qu'un tour de parole ? In *Langues et linguistique*, numéro spécial Journées de linguistique, pages 34-42.

SACKS, H., SCHEGLOFF, E.A. et JEFFERSON, G. (1974). A simplest systematics for the organization of turn-taking for conversation. In *Language*, 50 (4), pages 696-735.

VINCENT, D. (2002). Les enjeux de l'analyse conversationnelle et les enjeux de la conversation. In *Revue québécoise de linguistique*, 30-1, pages 177-198.

Enjeux méthodologiques, linguistiques et informatiques pour le traitement du français écrit des sourds

Tristan Vanrullen[1] *Leïla Boutora*[2] *Jean Dagron*[3]
(1) TVSI, 13009 Marseille
(2) LPL, UMR 7309 CNRS/Univ. d'Aix-Marseille, 13100 Aix-en-Provence
(3) Assistance publique - Hôpitaux de Marseille, 13005 Marseille
tristan.vanrullen@gmail.com, leila.boutora@lpl-aix.fr,
jean.dagron@ap-hm.fr

RESUME

L'ouverture du Centre National de Réception des Appels d'Urgence (CNRAU) accessible aux sourds et malentendants fait émerger des questions linguistiques qui portent sur le français écrit des sourds, et des questions informatiques dans le domaine du traitement automatique du langage naturel. Le français écrit des sourds, pratiqué par une population hétérogène, comporte des spécificités morpho-syntaxiques et morpho-lexicales qui peuvent rendre problématique la communication écrite entre les personnes sourdes appelantes et les agents du CNRAU. Un premier corpus de français écrit sourd élicité avec mise en situation d'urgence (FAX-ESSU) a été recueilli dans la perspective de proposer des solutions TAL et linguistiques aux agents du CNRAU dans le cadre de ces échanges écrits. Nous présentons une première étude lexicale, morphosyntaxique et syntaxique de ce corpus reposant en partie sur une chaîne de traitement automatique, afin de valider les phénomènes linguistiques décrits dans la littérature et d'enrichir la connaissance du français écrit des sourds.

ABSTRACT

Methodological, linguistic and computational challenges for processing written French of deaf people

With the setup of a national emergency call-center for deaf people in France (CNRAU), some questions arise in linguistics and natural language processing about the written expression of deaf people. It is practiced by an heterogeneous population and shows morpho-syntactic, lexical and syntactic specificities which increase the difficulty, over the emergency situation, to successfully communicate between the deaf callers and the call-center operators. A first corpus (FAX-ESSU) of written French of deaf people was built with emergency conditions in order to provide linguistic and NLP solutions to the call center operators. On this corpus, we present a first study realized with the help of a natural language processing toolbox, in order to validation linguistic phenomenons described in the scientific literature and to enrich the knowledge of written French of deaf people.

MOTS-CLES : Français écrit des sourds, TAL, Français Langue Etrangère, linguistique de corpus, lexique, syntaxe, méthodologie.
KEYWORDS : Written French of deaf people, NLP, French as a foreign language, corpus linguistics, lexicon, syntax, methodology.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 559–566,
Grenoble, 4 au 8 juin 2012.

1 Contexte et motivations de l'étude

1.1 Contexte institutionnel, sociolinguistique et technologique

L'ouverture en septembre 2011 d'un Centre National de Réception des Appels d'Urgence (CNRAU, numéro d'urgence 114) pour les personnes sourdes et malentendantes résulte de l'application du décret publié le 16 avril 2008 (prévu par l'article 78 de la loi du 11 février 2005). Le centre d'appels concerne trois types d'urgence : police, pompiers, SAMU. Il vise à rendre accessibles dans une modalité autre qu'audio-vocale les services d'appels d'urgence à une population spécifique, vulnérable socialement, et hétérogène tant au niveau du potentiel auditif des personnes sourdes, que de leur parcours éducatif et linguistique (Gillot 1998). Selon leur profil, les personnes sourdes ou malentendantes peuvent utiliser, avec une maîtrise inégale, différents modes de communication reposant sur un canal visuo-gestuel, audio-vocal ou mixte : la Langue des Signes Française (LSF), langue visuelle-gestuelle pratiquée par les personnes sourdes signantes en France ; le français écrit ; ou le français oral dans certaines conditions, seul ou accompagné du Langage Parlé Complété (LPC, code manuel utilisé par une partie des sourds oralisants, pour désambigüiser la lecture labiale).

Pendant les premières années de fonctionnement du centre d'appels, seules sont disponibles les modalités de communication écrite par fax et par SMS. Or, selon le rapport Gillot (1998), 80 % de la population sourde adulte entretient un rapport difficile au français écrit. Ce rapport à l'écrit des personnes sourdes peut se traduire par des productions individuelles qui peuvent mener, pour les agents du 114, à une compréhension approximative ou lacunaire des messages écrits reçus, voire à une incompréhension des situations décrites par les appelants. Or, les objectifs fixés au CNRAU nécessitent avant toute chose que **les agents sourds et entendants comprennent rapidement les messages entrants et formulent des réponses écrites dans un français adapté, c'est-à-dire interprétable par l'appelant sourd.**

Cette situation nous amène à envisager des solutions techniques et scientifiques pour que le CNRAU puisse assurer sa mission. En premier lieu, il s'agit de **caractériser les phénomènes linguistiques et les types d'erreurs propres au français écrit des sourds** afin d'alimenter la formation des opérateurs, mais également pour permettre le développement ultérieur des outils TAL envisagés en soutien à la prise de décision des opérateurs et à la formulation de réponses écrites adaptées. Dans cette perspective, il importera à terme de déterminer l'outillage formel et technique permettant la *correction*, la *complétion*, la *reformulation* et la *traduction* automatiques des messages transmis au centre d'appels.

1.2 Etat de l'art des descriptions du français écrit par des sourds

Le français écrit des sourds est un domaine d'étude peu documenté. Les rares études menées ces 20 dernières années révèlent des phénomènes morpho-syntaxiques et morpho-lexicaux que l'on retrouve dans les productions écrites des sourds québécois (Nadeau 1993, repris dans Nadeau et Machabée, 1998), suisses (Niederberger, 2004) et français métropolitains (Tuller, 2000 et Périni, 2007).

Les personnes sourdes apprennent le français écrit à partir d'un accès au français oral soit lacunaire (restes auditifs, lecture labiale), soit inexistant. La LSF est langue première (L1) dans 10% des cas seulement. Les productions écrites des sourds se caractérisent par une grande variabilité inter-individuelle. Nadeau et Machabée (1998) ont cependant identifié des phénomènes propres au français écrit des sourds qui sont absents des productions des entendants apprenants du français écrit L1 et L2. Ces phénomènes touchent entre autres les rapports localisant/localisé, possesseur/possédé, l'absence de marques morphologiques pour le temps, mais également les redondances lexicales, les confusions de catégories lexicales, et l'utilisation des pronoms - omission illicite des pronoms objets et/ou sujet; confusion entre les 1ère et 3e personnes. Tuller (2000) ajoute l'ellipse de la préposition.

La question suivante se pose tout de même : ces productions constituent-elles une catégorie identifiable (le français écrit des sourds) possédant des caractéristiques communes et des frontières bien délimitées, ou au contraire s'agit-il d'un continuum de productions dont certaines sont plus proches de productions de français écrit d'entendants L1 francophones que d'une autre production de sourd L2. L'écrit des entendants comporte lui-même un certain degré de variabilité inter- et intra- individuel mais ces productions restent dans le champ des « possibles » c'est-à-dire de la grammaire du français écrit. L'étude de notre corpus soutenue par les outils TAL vise également à répondre à cette question.

2 Présentation du corpus FAX-ESSU et de la méthodologie

2.1 Le corpus FAX-ESSU

Quelques extraits choisis : [1] *je ne vais pas lé drocteur* ; [2] *un voiture est brule* ; [3] *elle est tombé son lit* ; [4] *elle tomber tu lit* ; [5] *tombé sur l'escalier* ; [6] *difficile respiratoire*

Le corpus FAX-ESSU (fax Ecrits par des personnes Sourdes en Situation d'Urgence) est un corpus élicité sous forme de fax d'urgence rédigés en français écrit par des personnes sourdes, mises en situation d'urgence. Il a été recueilli dans la perspective d'alimenter la formation des agents du 114. L'élicitation a pris la forme d'un jeu de rôle présenté par une personne sourde en LSF, comportant diverses situations qui nécessitent un appel d'urgence. La consigne donnée aux 17 participants sourds signants et non signants était de rédiger un appel d'urgence manuscrit type fax, sur un temps limité de trente secondes après avoir eu connaissance de la situation critique. 23 situations ont été présentées et ont donné lieu à la production d'énoncés en français. Les productions des situations qui ont posé problème (manque d'entraînement pour la situation 1, erreurs dues à une mauvaise compréhension des situations 7 et 8) ont été éliminées, ainsi que les productions des locuteurs qui comportaient des dessins non traitables à ce stade. Le corpus étudié comporte donc 300 productions correspondant à 15 locuteurs et 20 situations.

Nous présentons ici une première étude lexicale, morphosyntaxique et syntaxique du corpus FAX-ESSU réalisée à la fois manuellement et avec l'aide d'une chaîne de traitement automatique, afin de valider les phénomènes linguistiques décrits dans la

littérature et d'enrichir la connaissance du français écrit sourd.

2.2 Méthodologie

La première étude réalisée sur le corpus FAX-ESSU a pour objet de faire émerger à l'aide d'outils TAL des spécificités du français écrit des sourds relevées ou non dans la littérature. Elle vise également à mettre en évidence les limites d'une chaîne de traitement TAL, dédiée initialement au français standard, dans le traitement du français écrits des sourds.

Pour cela, nous avons travaillé sur trois niveaux : **lexical**, **morphosyntaxique** et **syntaxique**, en procédant à des analyses manuelles et des analyses automatiques corrigées manuellement. Pour le niveau syntaxique (section 3.1), l'analyse du corpus a été réalisée manuellement, en s'appuyant sur le formalisme des Grammaires de Propriétés (Blache, 2000). Ce choix a permis de caractériser les erreurs rencontrées et de les classer par type de contrainte (Propriétés) : **obligation** et **exigence** (omission du noyau syntagmatique et d'autres catégories exigibles), **dépendance** (accord), **précédence** (ordre attendu entre les syntagmes). Une analyse complémentaire a été menée, faisant ressortir des problèmes de sélection de catégorie impliquant régulièrement le niveau lexical, et des phénomènes propres au lexique (section 3.2).

Les analyses automatiques portent sur les niveaux lexical (en cours d'étude) et morpho-syntaxique (section 3.3) ; elles ont été menées avec la chaine de traitement LPLSuite, conçue au Laboratoire Parole et Langage (Blache, Vanrullen, Balfourier, 2006). Cette chaîne a été entraînée sur un corpus de français standard, et évaluée au cours de la campagne nationale d'évaluation des analyseurs syntaxiques EASY 2004-2006, ce qui donne une mesure de sa performance sur le traitement du français écrit standard en comparaison avec les solutions industrielles actuelles. La validation manuelle qui a suivi le traitement automatique a impliqué le travail de trois personnes durant plusieurs mois.

Pour les trois niveaux, l'annotation et la correction manuelles, qui s'appuient sur la segmentation et l'étiquetage automatique du corpus, ont permis de (1) détecter, corriger et comptabiliser les erreurs d'étiquetage morphosyntaxique ; et (2) mettre en évidence différents phénomènes syntaxiques spécifiques au français écrit des sourds, et faire émerger les patrons syntaxiques correspondants.

3 Premiers résultats de l'analyse du corpus FAX-ESSU

3.1 Niveau syntaxique : apport des grammaires formelles

L'analyse syntaxique du corpus repose sur le formalisme des Grammaires de Propriétés (Blache 2000). Cette analyse a permis de mettre en évidence 54 phénomènes syntaxiques dont la fréquence dans les énoncés produits varie de 0,3 % à 62%. Une analyse complémentaire portant sur la sélection de la catégorie morpho-syntaxique a permis de relever en plus 26 phénomènes dont la fréquence va de 0,3 à 16 %.

Dans la figure 1, nous indiquons le risque de rencontrer un type de phénomène donné dans un énoncé, 100% correspondant à une erreur de ce type par énoncé (17 mots en moyenne par énoncé). Sont concernés : **l'omission du noyau syntagmatique**, obligatoire,

(65%) ou d'une autre **catégorie exigée** dans le syntagme (105%) ; les erreurs dans la sélection de la **catégorie** (39%) ou du **lexème** (50%) au sein d'une même catégorie ; une **erreur d'accord** (24%) – genre, nombre, personne – qui concerne un énoncé sur quatre ; et un problème de **précédence** portant sur un mot ou un syntagme mal positionné (8%).

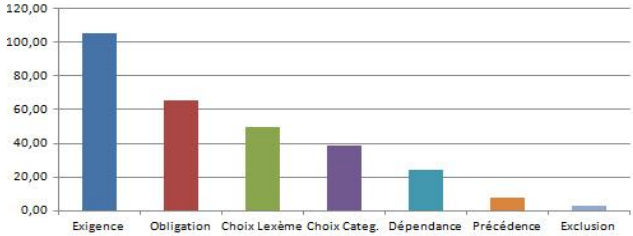


FIGURE 1 – % d'apparition des types de phénomènes rencontrés dans le corpus

Le tableau 1 regroupe nos observations syntaxiques sur les faits les plus remarquables qui apparaissent dans les proportions suivantes (ordre décroissant) : **l'omission du déterminant** (62%), de la **préposition** (35%) et du **verbe** (28%) ; la **sélection de la forme verbale adéquate** (39%) qui concerne pour 28% la catégorie 'verbe' avec un participe passé ou une forme infinitive au lieu d'un verbe fléchi ; et pour 11 % une confusion intra-catégorielle avec un nom à la place d'un verbe fléchi ou d'un participe passé.

Types d'erreurs en % des énoncés	**Omission**		**Accord**	**Sélection de catégorie ou de lexème**
	Obligation	**Exigence**	**Dépendance**	
Verbe	28%	11% (auxiliaire)	5% (genre) 3% (personne) 4% (nombre)	13% (part. passé) 11% (auxiliaire) 4% (infinitif)
Préposition	35%			11% (choix préposition)
Déterminant		62%	5% (genre) 6% (nombre)	3% (err. choix déterminant)
Nom	(1%)	17% (nom ou pronom sujet)		2% (adj) 6% (verbe) 5% (part. passé)
Pronom				8% (err. choix pronom)
Adverbe (nég.)		4% (« ne »)		

TABLEAU 1 –% d'apparition sur le corpus des phénomènes syntaxiques les plus fréquents

Le modèle des Grammaires de Propriétés a permis de rendre compte des phénomènes syntaxiques les plus massifs en termes de fréquence dans les énoncés produits. L'analyse complémentaire a contribué à affiner l'analyse sur la sélection des catégories et des lexèmes, qui ne sont pas prises en compte par les GP.

Cette première analyse fait ressortir des patrons syntaxiques exploitables en TAL. Dans le cadre des GP, ces patrons peuvent prendre la forme d'un assouplissement de certaines contraintes (obligation, exigence, dépendance). A titre d'exemple, un tel assouplissement permettrait de caractériser des **syntagmes verbaux** dont le noyau ou l'auxiliaire seraient absents ou encore dont le noyau serait une catégorie nominale ; de caractériser des

syntagmes prépositionnels sans préposition (!) ; de même pour l'absence du déterminant dans les **syntagmes nominaux** ou encore l'absence du pronom sujet relatif et donc du SN dans les **propositions relatives**.

3.2 Analyse manuelle du lexique : premiers résultats

Les phénomènes qui concernent le **lexique** apparaissent 1,5 fois par énoncé (154%), sans compter les phénomènes de *choix du lexème* que nous avons décrits en section 3.1 (emploi d'un mot pour un autre au sein de la même catégorie). Les erreurs d'accentuation (60%) ne constituent pas seulement un problème lexical, mais peuvent provoquer une confusion de la catégorie au niveau morphosyntaxique. Par exemple, l'absence d'accent sur un participe passé (*blesse/blessé*) qui concerne plus de 15% des énoncés entraîne une mauvaise identification de la catégorie concernée par l'étiqueteur automatique.

En plus de ces ambiguïtés d'origine accentuelle, les problèmes lexicaux tels que les mots inventés (*cacatouches/cacahuètes*) ou ceux orthographiés de façon approximative (*drocteur/docteur*), abrégée ou tronquée, ont également un impact sur la qualité de l'étiquetage automatique, et par suite sur l'analyse syntaxique automatique.

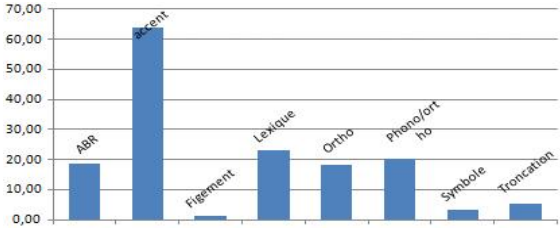


FIGURE 2 – % d'apparition des phénomènes lexicaux (hors choix du lexème)

La situation d'urgence peut être directement mise en cause dans l'usage d'abréviations (*Dr/docteur ; bcp/beaucoup*), de troncations en fin d'énoncé (*doct/docteur* ; *enf/enfant*) ou de symboles (*+/plus*). Ce phénomène peut être comparé aux productions observées dans les SMS. Enfin, les phénomènes concernant les figements ou expressions idiomatiques telles que *perdre les eaux*, *prendre feu*, *tomber dans les pommes*, qui n'ont pas été relevés de manière systématique, devront faire l'objet d'une étude détaillée. Une étude complémentaire du lexique qui s'appuie sur la segmentation automatique est en cours. Elle porte plus particulièrement sur la variabilité lexicale au sein du corpus FAX-ESSU chez les scripteurs sourds.

3.3 Etiquetage morphosyntaxique automatique

Nous avons souhaité vérifier la couverture et la pertinence d'un étiqueteur morpho-syntaxique automatique (POS-tagger) pour l'étude du corpus FAX-ESSU, sachant que le processus d'étiquetage est basé sur l'apprentissage d'un corpus de français écrit standard. Les données quantitatives issues de l'étiquetage et de la segmentation automatiques permettent de dessiner un premier contour très général du corpus. Les 346 phrases reconnues contiennent en tout 4434 mots dont 681 formes distinctes. 181 de ces formes

(soit 26,8%) n'appartiennent pas au lexique du français mais correspondent à des erreurs orthographiques et à des mots inconnus. La figure 3 indique la répartition des catégories présentes dans le corpus sur la base de l'étiquetage automatique (vert) effectué avec l'étiqueteur de la chaîne d'outils LPLSuite. puis corrigé manuellement (rouge).

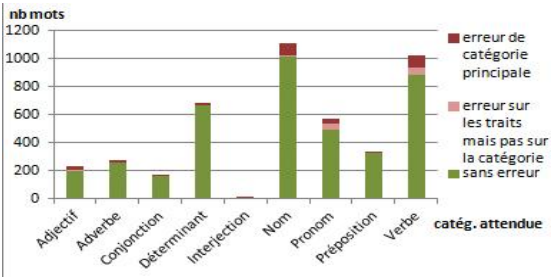


FIGURE 3– Couverture du POS-tagger sur le corpus FAX-ESSU

Une étude plus détaillée des erreurs de catégorisation de l'étiqueteur montre que le nom, le verbe, le pronom et l'adjectif sont moins bien étiquetées que le déterminant, l'adverbe et la préposition. En dehors de l'erreur déjà constatée en français standard sur l'ambigüité adjectif/participe passé, on constate au moins trois types d'erreurs d'étiquetage imputables à la structure syntaxique du français écrit des sourds : verbes interprétés comme noms, pronoms reconnus comme prépositions, noms reconnus comme adjectifs. De fait, le taux important d'omissions et d'erreurs lexicales dans le corpus a un impact significatif sur l'étiquetage. Par exemple, la catégorie Nom est étiquetée 34 fois en tant qu'adjectif , 21 fois comme déterminant, 23 fois en tant que verbe et 13 fois en tant que nom avec cependant des traits erronés (par ex. masc. au lieu de fém.)

4 Discussion des premiers résultats et perspectives

Les phénomènes que nous avons pu relever se recoupent avec les phénomènes pointés par la littérature (cf section 1.2). Les problèmes de compréhension relevés par les lecteurs et annotateurs humains se retrouvent au niveau du traitement automatique sous la forme d'erreurs de catégorisation des lexèmes et des syntagmes. Dès l'étape de l'étiquetage morphosyntaxique, et de façon encore plus proéminente au niveau syntaxique, apparaissent des difficultés et des erreurs de traitement liées aux spécificités du français écrit des sourds. Le TAL offre sur ce sujet un éclairage précis, notamment grâce à un outillage linguistique formel.

Après plusieurs mois de fonctionnement du numéro 114, les messages SMS semblent constituer l'essentiel des appels reçus au centre 114. La constitution d'un corpus SMS d'urgence nécessitera l'anonymisation des données et une gestion des droits et agréments permettant l'exploitation d'un corpus du domaine de la santé (HAS). Dans ces conditions, plusieurs études sont envisagées : dans un premier temps, la méthodologie linguistique et TAL développée et ajustée pour l'étude du corpus FAX-ESSU pourra être appliquée à l'étude des SMS d'urgence. Puis le corpus SMS sourds d'urgence pourra être comparé aux productions SMS sourds métropolitains et de la Réunion récemment recueillies (Blondel *et al.*, 2011), ainsi qu'aux SMS francophones recueillis dans le cadre de la campagne

SMS4Science (Antoniadis *et al.*, 2011). L'étude comparative de ces différents corpus permettra de mettre évidence dans les corpus ESSU ce qui tient à la communication médiée par SMS, ce qui tient au français écrit des sourds et ce qui provient de la situation d'urgence elle-même.

Remerciements

Nous remercions les personnes qui ont contribué à la réalisation de ce travail : Christian, Christine, Joëlle et Roberto pour nourrir nos réflexions sur le français écrit des sourds, les participants sourds au corpus élicité ; Samia et Patricia pour leur travail préparatoire sur le corpus ; Elodie, Fanny et Joy pour leur contribution à l'analyse des données.

Références

ANTONIADIS, G., CHABERT, G., ZAMPA, V. (2011). Alpes4science : Constitution d'un corpus de SMS réels en France métropolitaine. *79eme congrès de l'ACFAS. 9-13 mai 2011, Sherbrook*.

BLACHE, Ph. (2000). Le rôle des contraintes dans les théories linguistiques et leur intérêt pour l'analyse automatique : les Grammaires de Propriétés. *Actes de TALN 2000 (Traitement Automatique des Langues Naturelles)*.

BLACHE Ph., T. VANRULLEN & J.-M. BALFOURIER (2006). Constraint-Based Parsing as an Efficient Solution: Results from the Parsing Evaluation Campaign EASy. *Proceedings of LREC 2006*.

BLONDEL, M., GONAC'H, J., LEDEGEN G. & J. SEELI (2011). Ecriture-sms en Métropole et à La Réunion : « Zones instables et flottantes » du français ordinaire et spécificités du contexte de surdité. Gilles Col (dir.), *Transcrire, Écrire, Formaliser (1). Travaux linguistiques du CERLICO*, 23.

GILLOT, D. (1998). *Le droit des sourds : 115 propositions*. Rapport parlementaire au Premier Ministre.

NADEAU, M. (1993). Peut-on parler de "français sourd"? *Revue de l'Association canadienne de linguistique appliquée* (ACLA), vol. 15, no 2, pages 97-117.

NADEAU, M & D. MACHABEE (1998). Dans quelle mesure les erreurs des sourds sont-elles comparables à celles des entendants ? in Dubuisson C. & Daigle D. (Dir.), *Lecture, écriture et surdité*, Montréal, Logiques édition, pages 169-195.

NIEDERBERGER, N. (2004). *Capacités langagières en langue des signes française et en français chez l'enfant sourd bilingue : quelles relations ?* Thèse de Psychologie, Université de Genève.

PERINI, M. (2007). *La remédiation de l'illettrisme chez les adultes sourds locuteurs de la LSF : travail préparatoire à l'élaboration d'une méthodologie et de supports pédagogiques adaptés.* Mémoire de Master 2, Université Paris 8.

TULLER, L. (2000). Aspects de la morphosyntaxe du français des sourds. *Recherches Linguistiques de Vincennes* 29, pages 143-156, PUV.

Index

Adda-Decker, Martine, 359
Afli, Haithem, 447
Alfared, Ramadan, 71
Antolinos-Basso, Diégo, 471
Asher, Nicholas, 343

Baker, Michael, 351
Barcellini, Flore, 351
Barrault, Loïc, 447
Bechet, Denis, 71
Bédaride, Paul, 155
Bellot, Patrice, 479
Ben Abacha, Asma, 15
Ben Yahia, Yacine, 455
Benamara, Farah, 343
Benzitoun, Christophe, 99
Bernhard, Delphine, 211
Bertels, Ann, 239
Besançon, Romaric, 29
Bittar, Andre, 463
Blache, Philippe, 307
Bonneau-Maynard, Hélène, 487
Bouamor, Houda, 197, 267
Boudabous, Mohamed Mahdi, 225
Boudin, Florian, 281
Boutora, Leïla, 559
Bouzid, Maroua, 423
Braffort, Annelies, 375
Braud, Chloé, 471
Brouwers, Laetitia, 211
Brun, Caroline, 495
Brunessau, Stephan, 423
Brunet-Manquat, Francis, 335

Cadilhac, Anaïs, 343
Cailliau, Frederik, 551
Calabretto, Sylvie, 519
Candito, Marie, 321
Cavet, Ariane, 551
Cellier, Peggy, 253
Charnois, Thierry, 253, 423

Claveau, Vincent, 85, 383
Clavel, Chloé, 359
Constant, Matthieu, 57
Courmet, Arnaud, 431

Dagron, Jean, 559
Daille, Béatrice, 141
Danlos, Laurence, 471
de Groc, Clément, 183
De Hertog, Dirk, 239
de Loupy, Claude, 183
Denis, Alexandre, 351
Detienne, Francoise, 351
Deveaud, Romain, 479
Dikovsky, Alexander, 71
Diwersy, Sascha, 399
Duchier, Denys, 431
Dutrey, Camille, 359

Eensoo, Egle, 367
Enjalbert, Patrice, 503
Eshkol, Iris, 431

Fabre, Cécile, 169
Ferrari, Stephane, 503
Ferret, Olivier, 29, 391
Filhol, Michael, 375
Fort, Karën, 99, 383
François, Thomas, 211
Freard, Dominique, 351

Gahbiche-Braham, Souhir, 487
Gala, Nuria, 495
Garcia-Fernandez, Anne, 391
Goulian, Jérôme, 335
Grau, Brigitte, 1
Grilheres, Bruno, 423
Guillaume, Bruno, 293

Hadrich Belguith, Lamia, 225, 455
Hagege, Caroline, 463

Hernandez, Nicolas, 281
Heylen, Kris, 239
Houda, Saadane, 127
Huet, Stéphane, 527

Illouz, Gabriel, 197

Jean-Louis, Ludovic, 29

Keskes, Iskandar, 225
Kraif, Olivier, 399

Labadié, Alexandre, 503
Lardilleux, Adrien, 113
Laroche, Audrey, 407
Lavergne, Thomas, 487
Lefeuvre, Anaïs, 43
Lefèvre, Fabrice, 527
Legallois, Dominique, 253
Lepage, Yves, 113
Ligozat, Anne-Laure, 1, 211
Linarès, Georges, 527

Maaloul, Mohamed Hédi, 225
Martinet, Mathieu, 431
Max, Aurélien, 15, 197, 267
Mezghani Hammami, Souha, 455
Minard, Anne-Lyse, 1
Moot, Richard, 43
Morin, Emmanuel, 141
Morlane-Hondère, François, 169

Nasredine, Semmar, 127

Perrier, Guy, 293
Petitrenaud, Simon, 543
Plaisantin Alecu, Blandine, 511
Planas, Emmanuel, 415
Popescu, Vladimir, 343

Quignard, Matthieu, 351
Quiniou, Solen, 253

Rauzy, Stéphane, 307
Renahy, Julie, 511
Renard, Arnaud, 519
Rétoré, Christian, 43
Richard, Marion, 535
Rosset, Sophie, 359
Roze, Charlotte, 471
Rubino, Raphaël, 527
Rumpler, Béatrice, 519

Sagot, Benoît, 99, 535
Sandillon-Rezer, Noémie-Fleur, 43
Seck, Mohamadou, 343
Seddah, Djamé, 321
Serrano, Laurie, 423
Servan, Christophe, 543
Shwenk, Holger, 447
Sigogne, Anthony, 57
Stern, Rosa, 535
Suignard, Philippe, 551

Tanguy, Ludovic, 439
Tannier, Xavier, 183
Tellier, Isabelle, 431
Thomas, Izabella, 511
Tulechki, Nikola, 439

Valette, Mathieu, 367
Vanrullen, Tristan, 559
Vasilescu, Ioana, 359
Vilnat, Anne, 197, 267

Watrin, Patrick, 57

Yvon, François, 113, 487

Zweigenbaum, Pierre, 15