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 5 : démonstrations

Éditeurs
Laurent Besacier
Hervé Blanchon
Gilles Sérasset

4 – 8 Juin 2012
Grenoble, France

© 2012 Association Francophone pour la Communication Parlée (AFCP) et
Association pour le Traitement Automatique des Langues (ATALA)

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 (1$^{\text{ères}}$ JEP), 1979 (10$^{\text{èmes}}$ JEP), et avec en 2000 un détour à Aussois (23$^{\text{èmes}}$ 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 29$^{\text{ème}}$ 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

Grew : un outil de réécriture de graphes pour le TAL

Bruno Guillaume[1, 2] Guillame Bonfante[1, 3] Paul Masson
Mathieu Morey[4,5] 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
(4) Laboratoire Parole et Langage, Aix-Marseille Université
(5) Linguistics and Multilingual Studies, Nanyang Technological University

RÉSUMÉ
Nous présentons un outil de réécriture de graphes qui a été conçu spécifiquement pour des applications au TAL. Il permet de décrire des graphes dont les nœuds contiennent des structures de traits et dont les arcs décrivent des relations entre ces nœuds. Nous présentons ici la réécriture de graphes que l'on considère, l'implantation existante et quelques expérimentations.

ABSTRACT
Grew: a Graph Rewriting Tool for NLP

We present a Graph Rewriting Tool dedicated to NLP applications. Graph nodes contain feature structures and edges describe relations between nodes. We explain the Graph Rewriting framework we use, the implemented system and some experiments.

MOTS-CLÉS : réécriture de graphes, interface syntaxe-sémantique.

KEYWORDS: graph rewriting, syntax-semantics interface.

1 Réécriture de graphes

Même si la structure d'arbre est mise en avant pour les modélisations linguistiques, elle est souvent insuffisante, notamment pour les représentations sémantiques. Même en syntaxe, autour d'une structure d'arbre, des mécanismes comme la coindexation ou les structures de traits réentrantes font apparaître des structures de graphes. Partant de ces observations, pour avoir un cadre unique d'étude, nous avons pris le parti de considérer toutes ces structures comme des graphes. Évidemment, un cadre naturel pour décrire les calculs et les transformations sur les graphes est la réécriture de graphes. En effet, la réécriture est un modèle de calcul qui permet de décrire n'importe quelle transformation ; c'est également un domaine très actif en informatique théorique et de nombreux résultats de confluence ou de terminaison existent.

Par rapport à la réécriture de mots ou de termes, la réécriture de graphes est plus délicate et il n'existe pas de définition canonique de cette réécriture. Les travaux théoriques, utilisant la théorie des catégories, décrivent deux types de réécriture (SPO et DPO) ; cependant ces définitions, mathématiquement élégantes, sont peu pratiques pour écrire effectivement des règles. Nous avons donc utilisé une autre présentation qui décrit explicitement les transformations à appliquer au graphe à l'aide d'une suite d'actions élémentaires.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 5: Démonstrations, pages 1–2,
Grenoble, 4 au 8 juin 2012. ©2012 ATALA & AFCP

2 le logiciel Grew

Nous donnons quelques détails et particularités de notre système ci-dessous. Plus d'informations dont disponibles sur le site `grew.loria.fr`.

Graphes. Les graphes que nous considérons sont composés d'un ensemble de nœuds qui contiennent des structures de traits non récursives et d'un ensemble d'arcs étiquetés. Ces arcs codent en fait des relations, ils vérifient donc toujours la contrainte qu'il n'y peut pas y avoir deux arcs avec la même étiquette, la même source et le même but. Dans les expérimentations, les relations utilisées sont des dépendances mais le système permet d'exprimer n'importe quel type de relations (comme la dominance dans un arbre syntagmatique par exemple).

Règles. Pour manipuler les graphes, nous utilisons des règles de réécriture. Une règle de réécriture est définie en trois parties : un *patron positif* qui est un graphe qu'on va chercher à apparier avec une partie du graphe à réécrire ; un ensemble éventuellement vide de *patrons négatifs* qui peuvent bloquer l'application de la règle ; une liste de *commandes* qui décrivent les transformations à apporter au graphe. Les commandes permettent d'ajouter, de modifier ou de supprimer des traits, des nœuds ou des arcs.

Modules. Dans les applications, le nombre de règles à considérer peut être grand (plusieurs centaines). Pour contrôler le comportement global du calcul et pour faciliter le développement et la maintenance d'un système, les règles sont réparties dans différents *modules*. À l'intérieur d'un module, les règles ne sont pas ordonnées et tous les résultats sont calculés. En revanche, les modules sont ordonnés et la réécriture de l'ensemble du système se fait en appliquant chaque module sur l'ensemble des résultats du module précédent. Un système de règles organisées en modules s'appelle un GRS (pour *Graph Rewriting System*).

Interfaces utilisateurs. Le logiciel GREW permet de définir un GRS et de l'appliquer à des graphes. D'une part, une interface graphique permet de visualiser le détail d'une réécriture : il est possible de visualiser l'ensemble des étapes de la réécriture, règle par règle. D'autre part, il est possible d'appliquer un GRS sur un ensemble de graphes (éventuellement sur un cluster de calcul) et d'obtenir des statistiques sur l'ambiguïté du calcul et sur les fréquences d'utilisation des règles ou des modules.

Notre système a été expérimenté (Guillaume et Perrier, 2012) sur des données de tailles réelles : réécriture de 12 000 graphes à l'aide d'un système de 34 modules contenants plus de 400 règles. Parmi des fonctionnalités récentes, on peut noter la possibilité de paramétrer les règles avec des informations lexicales et l'ajout un nouveau type de règles qui filtrent les résultats d'un module pour ne garder que ceux qui respectent certains motifs. Ces deux nouveautés permettent de mieux factoriser les règles et donc facilitent le développement de plus grands systèmes de règles.

Références

GUILLAUME, B. et PERRIER, G. (2012). Annotation sémantique du French Treebank à l'aide de la réécriture modulaire de graphes. *In Actes de TALN 2012 (Traitement automatique des langues naturelles)*, Grenoble. ATALA.

Interfaces de navigation dans des contenus audio et vidéo

Géraldine Damnati
(1)France Telecom, Orange Labs, Lannion
geraldine.damnati@orange.com

RESUME

Deux types de démonstrateurs sont présentés. Une première interface à visée didactique permet d'observer des traitements automatiques sur des documents vidéo. Plusieurs niveaux de représentation peuvent être montrés simultanément, ce qui facilite l'analyse d'approches multi-vues. La seconde interface est une interface opérationnelle de "consommation" de documents audio. Elle offre une expérience de navigation enrichie dans des documents audio grâce à une visualisation de métadonnées extraites automatiquement.

ABSTRACT

Navigation interfaces through audio and video contents

Two types of demonstrators are shown. A first interface, with didactic purposes, allows automatic processing of video documents to be observed. Several representation levels can be viewed simultaneously, which is particularly helpful to analyse the behaviour of multi-view approaches. The second interface is an operational audio document "consumption" interface. It offers an enriched navigation experience through the visualisation of automatically extracted metadata.

MOTS-CLES : Traitements multi-vues, navigation enrichie.
KEYWORDS : Multi-view processing, enriched navigation.

1 Interface didactique

Il s'agit d'un démonstrateur qui permet d'illustrer les traitements automatiques réalisés sur des contenus vidéo. Le principe est de visualiser sous forme de timeline des informations de structuration extraites automatiquement. Pour chaque segment, un onglet permet de visualiser des informations issues du canal audio (typiquement la transcription automatique synchronisée avec le player) et un onglet permet de visualiser des informations liées au canal vidéo (typiquement des images clé ou *key frames*). L'interface offre des fonctionnalités de navigation d'un segment à l'autre. Au-delà de ces fonctionnalités de base, l'intérêt de l'outil est de pouvoir cumuler plusieurs timeline et observer ainsi l'apport de traitement multi-niveaux. Plusieurs résultats de travaux de recherche seront montrés via cette interface.

Reconnaissance du rôle du locuteur

La capture d'écran ci-contre représente une analyse en rôle des tours de parole dans des Journaux Télévisés. Elle illustre une approche multi-vue qui consiste à fusionner une analyse purement acoustique modélisant l'intonation

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 5: Démonstrations, pages 3–4,
Grenoble, 4 au 8 juin 2012. ⓒ2012 ATALA & AFCP

des locuteurs en fonction de leur rôle et une analyse purement linguistique basée sur une analyse de la transcription automatique du contenu parlé (Damnati et Charlet, 2011). L'interface permet de visualiser les résultats de chacune des analyses ainsi que de leur fusion, afin de mieux analyser leur complémentarité.

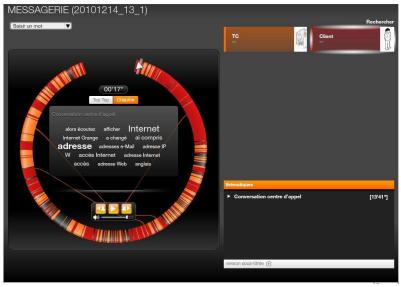
Détection de personnes dans des documents vidéo

Les travaux réalisés dans le cadre du défi REPERE (Béchet *et al.*, 2012) seront également montrés via cette interface. Ce projet a pour but d'identifier les personnes dans des contenus télévisés en exploitant conjointement le canal audio (contenu parlé et analyse en locuteurs) et le canal vidéo (texte incrusté et analyse de visages). L'interface permet de visualiser les informations extraites dans les différentes modalités ainsi que le résultat de la fusion.

2 Interface de navigation enrichie

Cette interface à pour vocation de proposer aux utilisateurs une expérience de navigation enrichie dans des contenus purement audio, en s'appuyant sur des métadonnées produites automatiquement. Elle propose en quelque sorte de "visualiser" des contenus audio. Elle est déclinée à Orange Labs dans différents domaines, allant de la consommation de podcast de radio à l'écoute de conversations issues des centres d'appels.

La capture d'écran ci-contre illustre une interface de visualisation conversations client/téleconseiller, et s'inscrit dans le domaine plus large du *Speech Analytics*. Elle permet d'avoir une vue synthétique du déroulé de la conversation, structurée en locuteurs, une visualisation d'expressions clés extraites des transcriptions automatiques, un filtrage des conversations par motif d'appel, etc…



Références

DAMNATI, G., CHARLET, D. (2011). Multi-view approach for speaker turn role labeling in TV Broadcast News shows, *Proc. Interspeech'11,* Florence, 2011.

BECHET, F., AUGUSTE, R., AYACHE, S., CHARLET, D., DAMNATI, G., FAVRE, B., FREDOUILLE, C., LEVY, C. (2012). Percol0 - un système multimodal de détection de personnes dans des documents vidéo. *Proc. JEP'12*, Grenoble, 2012.

Synthèse de texte avec le logiciel Syntox

Lionel Clément
Université Bordeaux 1 – LaBRI (UMR 5800) – CLEE-ERSS (UMR 5263)
lionel.clement@labri.fr

RÉSUMÉ

Le logiciel **Syntox**, dont une interface utilisateur en ligne ce trouve à cette *URL* : `http://www.syntox.net`, est une mise en application d'un modèle basé sur les grammaires attribuées, dans le cadre de la synthèse de texte. L'outil est une plateforme d'expérimentation dont l'ergonomie est simple. **Syntox** est capable de traiter des lexiques et des grammaires volumineux sur des textes ambigus à partir de la description explicite de phénomènes linguistiques.

ABSTRACT

Automated generation of text with Syntox

Syntox, which includes an online user interface at URL `http://www.syntox.net`, is an implementation of a model based on attribute grammars, in the context of automated generation of text. The sofware is intended as a platform for experimentation with an ergonomic interface. **Syntox** is usable with large vocabularies and grammars to produce ambiguous texts from an explicit description of linguistic phenomena.

MOTS-CLÉS : Synthèse de texte, Grammaire attribuée, Syntaxe.

KEYWORDS: Text generation, Attribute Grammars, Syntax.

Introduction

Le programme Syntox demande une grammaire, un lexique (ou une connexion à un serveur qui délivre ce lexique), et une structure de traits censée correspondre à une représentation de la syntaxe d'un texte. Le logiciel produit une forêt partagée. L'ensemble des textes engendrés par cette forêt est ensuite affiché à l'écran avec des éléments permettant la mise au point des grammaires (forêts partielles, messages d'erreurs, étapes de synthèse).

Synthèse de texte à l'aide de grammaires attribuées

Contrairement à ce qui se fait dans les modèles habituels «d'unification», comme PATR-II (Shieber, 1984), il n'est jamais question de marquer une contrainte d'égalité $A = B$ entre deux éléments dont l'un entre dans la structure compositionnelle de l'autre. Nous rejetons systématiquement cette approche qui pose un problème de complexité algorithmique (dont l'indécidabilité dans le cas général) pour l'analyse ou la synthèse lorsque A et B sont dans des rapports pluri-univoques (homonymie, polysémie, synomynie). A cette approche qui se résume à résoudre par unification

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 5: Démonstrations, pages 5–6,
Grenoble, 4 au 8 juin 2012. ©2012 ATALA & AFCP

un système d'équations et qui donne son nom aux grammaires «d'unification», nous opposons un modèle où A et B sont dans une relation de dépendance, plus précisément de copie.

Le modèle utilisé par Syntox est celui, bien connu et assez ancien des grammaires attribuées(Knuth, 1968). En conformité avec celui-ci, Syntox impose la non circularité des calculs sur les attributs pour la raison évoquée immédiatement. Ceci sera essentiel pour garantir la décidabilité des algorithmes qui mettent en correspondance les textes et les représentations profondes de ces textes (ici la synthèse de texte à partir d'une représentation de la syntaxe).

Pour nous approcher des formalismes utilisés en linguistique, notamment des grammaires d'unification, les structures de traits, censés contenir les représentations profondes des textes, ont été choisies comme attributs de la grammaire. L'analyse syntagmatique est supportée par la grammaire hors-contexte sous-jacente de la grammaire attribuée.

La grammaire décline pour chaque syntagme, deux attributs, l'un synthétisé, que je note $\Uparrow$, et l'autre hérité que je note $\uparrow$. Le premier est calculé en fonction des attributs des termes qui composent le syntagme. Le second est calculé en fonction des attributs du terme qui le contient. D'évidence les propriétés lexicales seront synthétisées, y compris celles qui ne sont pas locales, les propriétés globales [1] seront héritées.

Écrire une grammaire jouet du français

Grâce à quelques exemples simples, on montre l'intérêt de manipuler les deux types d'attributs (synthétisé et hérités). Les phénomènes lexicaux peuvent être complexes. Je montre comment faire avec Syntox pour qu'un prédicat nominal s'articule avec un verbe support ou encore comment un prédicat impose des constantes dans le lemme ou les catégories grammaticales de ses compléments.

Sur l'exemple des adjectifs épithètes, je montre comment utiliser les listes d'ajouts et comment récupérer les attributs synthétisés de façon récursive pour traiter l'accord.

Au delà de la phrase En conclusion, je propose quelques exemples aux limites de l'outil avec des textes qui mettent en oeuvre une ou plusieurs propositions. En aucune manière, Syntox ne permet de traiter la résolution des anaphores ou l'ellipse. Il conviendra alors de savoir articuler Syntox dans un système de génération de texte pour ces problèmes en particulier.

Références

KNUTH, D. E. (1968). Semantics of context-free languages. *Mathematical Systems Theory*, 2(2):127–145.

SHIEBER, S. M. (1984). The design of a computer language for linguistic information. *In Proceedings of the Tenth International Conference on Computational Linguistics*, pages 362–366, Stanford University, Stanford, California.

1. Par propriétés globales, j'entends par exemple celles qui ne sont pas sujettes à la variation.

Un segmenteur-étiqueteur et un chunker pour le français

Isabelle Tellier[1,2], Yoann Dupont[1,2], Arnaud Courmet[2]
(1) LaTTiCe, université Paris 3 - Sorbonne Nouvelle
(2) LIFO, université d'Orléans
isabelle.tellier@univ-paris3.fr, yoann.dupont@etu.univ-orleans.fr,
arnaud.coumet@gmail.com

RÉSUMÉ

Nous proposons une démonstration de deux programmes : un segmenteur-étiqueteur POS pour le français et un programme de parenthésage en "chunks" de textes préalablement traités par le programme précédent. Tous deux ont été appris à partir du French Tree Bank.

ABSTRACT

A Segmenter-POS Labeller and a Chunker for French

We propose a demo of two softwares : a Segmenter-POS Labeller for French and a Chunker for texts treated by the first program. Both have been learned from the French Tree Bank.

MOTS-CLÉS : étiquetage POS, chunking, apprentissage automatique, French Tree Bank, CRF.

KEYWORDS: POS tagging, chunking, Machine Learning, French Tree Bank, CRF.

1 Introduction

Nous proposons de faire une démonstration de plusieurs programmes appris automatiquement à partir du French Treebank(Abeillé *et al.*, 2003) :
– un segmenteur combiné avec un étiqueteur morphosyntaxique (Constant *et al.*, 2011)
– un "chunker" ou analyseur syntaxique superficiel (Abney, 1991; Sha et Pereira, 2003; Tellier *et al.*, 2012)

Les deux programmes sont utilisables en séquence, le chunker s'appuyant pour fonctionner sur le résultat fourni par l'étiqueteur. Ils ont tous les deux été appris automatiquement par un CRF (Conditional Random Fields) (Lafferty *et al.*, 2001; Tellier et Tommasi, 2011). Ils sont libres et gratuits, disponibles en téléchargement mais ont surtout été testés sous Debian, Ubuntu et Mac (résultats non garantis sous Windows). Il faut pour les utiliser disposer des logiciels suivants :
– un interpréteur Python : http ://www.python.org/download/
– Wapiti, une implémentation des CRF linéaires : http ://wapiti.limsi.fr/
– Bazaar, un gestionnaire de versions donnant accès au serveur où ils sont stockés : http ://wiki.bazaar.canonical.com/

Nous décrivons brièvement ci-dessous les différentes options disponibles pour ces programmes et les résultats de leur évaluation.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 5: Démonstrations, pages 7–8,
Grenoble, 4 au 8 juin 2012. ©2012 ATALA & AFCP

2 Les programmes

Pour télécharger le segmenteur-étiqueteur, il faut saisir l'instruction suivante :
bzr branch lp : yoann-dupont/crftagger/stand-alone-tagger
Le processus ayant permis de l'apprendre est décrit dans (Constant *et al.*, 2011). Son originalité est de permettre plusieurs segmentations possibles :
– soit une segmentation "maximale" réalisée à l'aide de règles écrites manuellement
– soit une segmentation qui cherche à identifier les unités multimots du texte, en tenant compte de celles présentes dans le French Treebank ainsi que dans le Lefff (Sagot, 2010).

L'étiqueteur s'appuie sur la segmentation choisie et distingue 29 étiquettes : en validation croisée, il atteint une exactitude de 97,3% sans tenir compte des unités multimots, 95,2% avec elles.
Pour télécharger le "chunker", il faut saisir l'instruction suivante :
bzr branch lp : yoann-dupont/crftagger/chunker_models
Le processus ayant permis de l'apprendre est décrit dans (Tellier *et al.*, 2012). Il s'appuie sur les étiquettes fournies par le programme précédent et fonctionne suivant deux variantes possibles :
– une variante qui se concentre sur la seule identification de tous les "groupes nominaux simples" (i.e. non récursifs) NP. En supposant un étiquetage morphosyntaxique parfait et en réquérant la stricte égalité des frontières, ils sont identifiés en validation croisée avec une précision de 97,49%, un rappel de 97,40%, et donc une F-mesure de 97,45.
– une variante qui cherche à réaliser un parenthésage complet des phrases, en distinguant 6 types de chunks possibles. En supposant un étiquetage parfait, la "micro-average" (moyenne des F-mesures de chaque groupe pondérées par leur effectif) vaut 79,73, tandis que la "macro-average" (moyenne des F-mesure sans pondération) vaut 73,37.

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.

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.

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)*.

Sha, F. et Pereira, F. (2003). Shallow parsing with conditional random fields. *In Proceedings of HLT-NAACL 2003*, pages 213 – 220.

Tellier, I., Duchier, D., Eshkol, I., Courmet, A. et Martinet, M. (2012). Apprentissage automatique d'un chunker pour le français. *In Actes de TALN'12, papier court (poster)*.

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.

SPPAS : segmentation, phonétisation, alignement, syllabation

Brigitte Bigi
Laboratoire Parole et Langage, CNRS & Aix-Marseille Université,
5 avenue Pasteur, BP80975, 13604 Aix-en-Provence France
brigitte.bigi@lpl-aix.fr

RÉSUMÉ

SPPAS est le nouvel outil du LPL pour l'alignement texte/son. La segmentation s'opère en 4 étapes successives dans un processus entièrement automatique ou semi-automatique, à partir d'un fichier audio et d'une transcription. Le résultat comprend la segmentation en unités inter-pausales, en mots, en syllabes et en phonèmes. La version actuelle propose un ensemble de ressources qui permettent le traitement du français, de l'anglais, de l'italien et du chinois. L'ajout de nouvelles langues est facilitée par la simplicité de l'architecture de l'outil et le respect des formats de fichiers les plus usuels. L'outil bénéficie en outre d'une documentation en ligne et d'une interface graphique afin d'en faciliter l'accessibilité aux non-informaticiens. Enfin, SPPAS n'utilise et ne contient que des ressources et programmes sous licence libre GPL.

ABSTRACT

SPPAS : a tool to perform text/speech alignment

SPPAS is a new tool dedicated to phonetic alignments, from the LPL laboratory. SPPAS produces automatically or semi-automatically annotations which include utterance, word, syllabic and phonemic segmentations from a recorded speech sound and its transcription. SPPAS is currently implemented for French, English, Italian and Chinese There is a very simple procedure to add other languages in SPPAS : it is just needed to add related resources in the appropriate directories. SPPAS can be used by a large community of users : accessibility and portability are importants aspects in its development. The tools and resources will all be distributed with a GPL license.

MOTS-CLÉS : segmentation, phonétisation, alignement, syllabation.

KEYWORDS: segmentation, phonetization, alignement, syllabification.

Dans le vaste domaine de la segmentation texte/son, on trouve sur le web de nombreuses « boîtes à outils » pour réaliser différents niveaux de segmentations de la parole et l'apprentissage des modèles sous-jacents. Des ressources (dictionnaires, modèles, etc.) sont également disponibles pour pouvoir les exploiter. Pourtant, lorsqu'il s'agit d'effectuer des alignements (en quantité raisonnable), la plupart des phonéticiens choisissent de le faire manuellement même si plusieurs heures sont souvent nécessaires pour n'aligner qu'une seule minute de signal. La raison principalement évoquée concerne le fait que très peu d'outils sont à la fois disponibles librement, utilisables *de façon simple et ergonomique*, multi-plateforme et, bien sûr, qui traite la langue souhaitée. Ainsi, bien qu'elles soient très utilisées par les informaticiens, des boîtes à outils telles que, par exemple, HTK (Young, 1994), Sphinx (Carnegie Mellon University, 2011)

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 5: Démonstrations, pages 9–10,
Grenoble, 4 au 8 juin 2012. ©2012 ATALA & AFCP

ou Julius (Lee *et al.*, 2001), ne bénéficient toujours pas d'un développement qui permette une accessibilité à une communauté plus large d'utilisateurs. Développer un outil d'alignement automatique, s'appuyant uniquement sur des ressources libres (outils et données) et regroupant les critères nécessaire à son accessibilité à des non-informaticiens n'est pas uniquement un défi technique. On suppose en effet que si tel était le cas, cet outil existerait déjà.

Le logiciel présenté lors de la session de démonstration s'appelle *SPPAS* :

`http://www.lpl-aix.fr/~bigi/sppas/`

La segmentation en IPUs consiste à aligner les macro-unités d'un texte (segments, phrases, etc) avec le son qui lui correspond. L'algorithme implémenté dans SPPAS s'appuie sur la recherche des pauses dans le signal et leur alignement avec les unités proposées dans la transcription (en supposant qu'une pause sépare chaque unité).

La phonétisation consiste à représenter les unités (mots, syllabes) d'un texte par des symboles phonétiques. SPPAS s'appuie uniquement sur la consultation d'un dictionnaire de prononciations et traite les deux cas suivants : 1/ une entrée peut se prononcer de différentes manières (homographes hétérophones, accents régionaux, phénomènes de réductions propres à l'oral...). Dans ce cas, c'est l'aligneur qui choisit la phonétisation. 2/ une entrée peut être absente du dictionnaire. SPPAS produit une phonétisation automatique dont l'algorithme, indépendant de la langue, cherche les segments les plus longs dans le dictionnaire. Dans un processus semi-automatique, l'utilisateur peut choisir/modifier la phonétisation.

L'alignement en phonème consiste à déterminer la localisation temporelle de chacun des phonèmes d'une unité. SPPAS fait appel à Julius pour réaliser l'alignement. Pour réaliser l'alignement, Julius a besoin d'une grammaire et d'un modèle acoustique. La grammaire contient la (ou les) prononciation(s) de chaque mot et l'indication des transitions entre les mots. L'alignement requiert aussi un modèle acoustique qui doit être au format HTK-ASCII.

Concernant la syllabation, SPPAS encapsule le syllabeur du LPL (Bigi *et al.*, 2010). Il consiste à définir un ensemble de règles de segmentation entre phonèmes. Les phonèmes sont regroupés en classes et des règles de segmentation entre ces classes sont établies.

Références

Bigi, B., Meunier, C., Nesterenko, I. et Bertrand, R. (2010). Automatic detection of syllable boundaries in spontaneous speech. *In Language Resource and Evaluation Conference*, pages 3285–3292, La Valetta, Malta.

Carnegie Mellon University (2011). CMUSphinx : Open Source Toolkit For Speech Recognition. http ://cmusphinx.sourceforge.net.

Lee, A., Kawahara, T. et Shikano., K. (2001). Julius — an open source real-time large vocabulary recognition engine. *In European Conference on Speech Communication and Technology*, pages 1691–1694.

Young, S. (1994). The HTK Hidden Markov Model Toolkit : Design and Philosophy. *Entropic Cambridge Research Laboratory, Ltd*, 2:2–44.

Solution Proxem d'analyse sémantique verticale : adaptation au domaine des Ressources Humaines

François-Régis Chaumartin[1]
(1) Proxem, 19 bd de Magenta, 75010 Paris
frc@proxem.com

RESUME

Proxem développe depuis 2007 une plate-forme de traitement du langage, Antelope, qui permet de construire rapidement des applications sémantiques verticales (par exemple, pour l'e-réputation, la veille économique ou l'analyse d'avis de consommateurs). Antelope a servi à créer une solution pour les Ressources Humaines, utilisée notamment par l'APEC, permettant (1) d'extraire de l'information à partir d'offres et de CVs et (2) de trouver les offres d'emploi correspondant le mieux à un CV (ou réciproquement). Nous présentons ici l'adaptation d'Antelope à un domaine particulier, en l'occurrence les RH.

ABSTRACT

How to adapt the Proxem semantic analysis engine to the Human Resources field

Proxem develops since 2007 the NLP platform, Antelope, with which one can quickly build vertical semantic applications (for e-reputation, business intelligence or consumer reviews analysis, for instance). Antelope was used to create a Human Resources solution, notably used by APEC, making it possible (1) to extract information from resumes and offers and (2) to find the most relevant jobs matching a given resume (or vice versa). We present here how to adapt Antelope to a particular area, namely HR.

MOTS-CLES : entités nommées, extraction de relations, création d'ontologies, similarité
KEYWORDS : named entities, information extraction, ontologies development, matching

1 La plate-forme de traitement linguistique Antelope

Antelope (Chaumartin, 2008) est une plate-forme de TAL intégrant des composants de traitement syntaxique et sémantique, ainsi qu'un lexique de large couverture pour l'anglais et le français. Elle permet de créer rapidement des applications d'analyse sémantique, en enchaînant plusieurs opérations au sein d'une chaîne de traitement.

La qualité des documents traités étant très variable, une correction orthographique est souvent nécessaire ; néanmoins, cette opération doit être effectuée avec une connaissance du contexte métier ; par exemple, les noms propres (qui ne figurent pas dans le lexique intégré) ne doivent pas être « corrigés » vers un mot proche.

La reconnaissance d'entités nommées vise classiquement à identifier des personnes, lieux et organisation. Dans un contexte d'enseigne de grande distribution, les entités intéressantes à détecter sont plutôt les produits, marques et concurrents cités, ainsi que des concepts liés au métier (risque sanitaire, risque juridique…). Dans le domaine des Ressources Humaines, il s'agira plutôt de reconnaître des métiers, des compétences, des talents, des diplômes…

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 5: Démonstrations, pages 11–12,
Grenoble, 4 au 8 juin 2012. ©2012 ATALA & AFCP

2 Adaptation d'Antelope au domaine RH

Nous avons développé une nouvelle approche d'acquisition à large échelle d'entités nommées. (a) Une première phase d'extraction terminologique permet d'amorcer la liste des concepts du domaine. (b) Une seconde phase utilise des ressources de large couverture (la Wikipédia et un WordNet pour le français) pour créer des gazettes ; en cas d'ambiguïté (le métier d'*architecte* relève par exemple du BTP de l'informatique), les termes des gazettes sont automatiquement associés à des mots clés activateurs ou inhibiteurs. (c) L'application de ces gazettes permet de constituer un premier corpus annoté selon les entités nommées du domaine. Un apprentissage (par CRF) est alors effectué sur le corpus, pour identifier de nouvelles instances d'entités. Les concepts correspondant aux entités nommées sont organisés sous forme de taxonomie. La figure 1 montre par exemple, sur la partie de gauche, l'organisation des métiers, compétences et talents sous forme d'arborescence. Cette information est utilisée lors de la recherche de documents similaires ; concrètement, elle permet de déterminer que –toutes choses égales par ailleurs– la compétence « développement Java » est plus proche d'une compétence « développement en langage objet » que de « développement COBOL ». Ce point améliore fortement la pertinence des documents trouvés lors d'une recherche.

FIGURE 1 – Une capture d'écran de la solution d'analyse sémantique RH de Proxem.

Remerciements

Le projet SIRE (FEDER) a partiellement financé l'adaptation d'Antelope au domaine RH.

Références

CHAUMARTIN F.-R. (2008). ANTELOPE, une plate-forme industrielle de traitement linguistique. *Traitement Automatique des Langues* 49:2.

Nomao : un moteur de recherche géolocalisé spécialisé dans la recommandation de lieux et l'e-réputation

Estelle Delpech[1,2] Laurent Candillier[1,2]
(1) NOMAO, 1 avenue Jean Rieux 31500 Toulouse, www.nomao.com
(2) EBUZZING GROUP, 97 rue du cherche-midi 75006 Paris, www.ebuzzing.com
{prenom}@nomao.com, {prenom.nom}@ebuzzing.com

RÉSUMÉ
Cette démonstration présente NOMAO, un moteur de recherche géolocalisé qui permet à ses utilisateurs de trouver des lieux (bars, magasins...) qui correspondent à leurs goûts, à ceux de leurs amis et aux recommandations des internautes.

ABSTRACT
Nomao : a geolocalized search engine dedicated to place recommendation and e-reputation

This demonstration showcases NOMAO, a geolocalized search engine which recommends places (bars, shops...) based on the user's and its friend's tastes and on the web surfers' recommendations.

MOTS-CLÉS : recherche d'information, analyse d'opinion, génération de texte, fouille du web.

KEYWORDS: information retrieval, opinion mining, text generation, web mining.

1 Nomao : recherche locale, recommandation et e-réputation

Nomao est un moteur de recherche géolocalisé spécialisé dans la recommandation de lieux et l'e-réputation. Il répond à des requêtes portant sur un type de lieux (restaurant, médecin. . .) et une zone géographique donnée, par exemple : « hôtel bon marché à Grenoble ». Dans sa page de résultats, Nomao favorise les lieux correspondant aux goûts de l'utilisateur (filtrage collaboratif), aux recommandations de ses amis et des internautes.

Nomao extrait ses données de sites de contenu générés par les utilisateurs et d'annuaires de lieux. Il les agrège, les enrichit et en présente une synthèse personnalisée à ses utilisateurs. Chaque lieu est associé à une fiche descriptive résumant les opinions exprimées sur Internet à son propos ainsi que diverses informations factuelles : descripteurs (*cuisine italienne*, *wifi gratuit*), adresse, n° de téléphone, etc. Nomao fonctionne actuellement en 5 langues (allemand, anglais, espagnol, français, italien).

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 5: Démonstrations, pages 13–14,
Grenoble, 4 au 8 juin 2012. ©2012 ATALA & AFCP

2 Traitement automatique des langues pour la recherche locale, la recommandation et l'e-réputation

Nos axes de recherche et développement en TAL émargent sur plusieurs domaines :

• **Analyse d'opinion** Nous cherchons à analyser l'opinion à deux niveaux de granularité. À granularité moyenne, chaque commentaire doit être associé à une tonalité positive, plutôt positive, négative, plutôt négative ou neutre. À granularité fine, les termes extraits des commentaires doivent être rattachés à une catégorie et leur tonalité identifiée :

"mousse au chocolat excellente" ⇒ {CATÉGORIE : dessert, TONALITÉ : positive}

• **Génération automatique de textes** Les informations associées à un lieu sont présentes en base de données sous la forme de paires attribut/valeur. Afin d'en faciliter la lecture à nos utilisateurs, nous générons automatiquement des descriptifs en langue naturelle :

{NOM : La Braisière, TYPE : restaurant, DESCRIPTEURS : sud-ouest, tapas}
⇒ "La Braisière est un restaurant spécialisé dans la cuisine du sud-ouest et les tapas."

• **Interprétation de requêtes** Nomao reçoit deux types de requêtes : soit l'utilisateur cherche un type de lieu dans une zone géographique donnée (« restaurant à Paris »), soit il cherche des informations sur un lieu qu'il connait déjà (« restaurant "La Gare", Paris »). Les requêtes sont donc ambiguës et il faut déterminer quels mots réfèrent au nom du lieu, au type du lieu, à sa localisation et à d'éventuels descripteurs :

"restaurant l'auberge en gascogne"
⇒ { TYPE : restaurant, NOM : auberge en gascogne, LOCALISATION : ∅, DESCRIPTEURS : ∅ }
|| { TYPE : restaurant, NOM : ∅, LOCALISATION : gascogne, DESCRIPTEURS : auberge }

• **Normalisation graphique et morphosyntaxique** Les descripteurs associés aux lieux sont collectés à partir de plusieurs sources et apparaissent sous des formes différentes qu'il convient de normaliser (correction orthographique, lemmatisation) :

{grillade, grillades, grilade} ⇒ grillades

Il en est de même pour les termes extraits des commentaires qui comportent des variations morphosyntaxiques que nous souhaitons pouvoir identifier :

{"attente trop longue", "attendre longtemps"} ⇒ attente longue

• **Identification de relations sémantiques** Le développement d'un thésaurus doit permettre d'améliorer la normalisation des descripteurs et des termes :

{boîte, discothèque} ⇒ discothèque

Les relations sémantiques peuvent être utilisées pour étendre/affiner les requêtes et améliorer la recommandation :

"restaurant végétarien" ⇒ "restaurant végétarien OR végétalien"

Enfin, les sources de données n'étant pas fiables à 100%, il faut pouvoir détecter automatiquement les descripteurs incohérents pour un lieu :

{gastronomique, étoilé, haut de gamme, fast-food} ⇒ {gastronomique, étoilé, haut de gamme}

Le DictAm : Dictionnaire électronique des verbes amazighs-français

Samira MOUKRIM
Université Sidi Mohamed Ben Abdellah
samiramoukrim@yahoo.fr

RESUME

Le DictAm est un dictionnaire électronique des verbes amazighs-français. Il vise à rendre compte de l'ensemble des verbes dans le domaine berbère : *conjugaison, diathèse et sens*. Le DictAm comporte actuellement près de 3000 verbes dans une soixantaine de parlers berbères. C'est un travail qui est en cours de réalisation et qui a pour ambition de répertorier tous les verbes berbères ainsi que leurs équivalents en français.

ABSTRACT

The DictAm : An electronic dictionary of Amazigh-French verbs

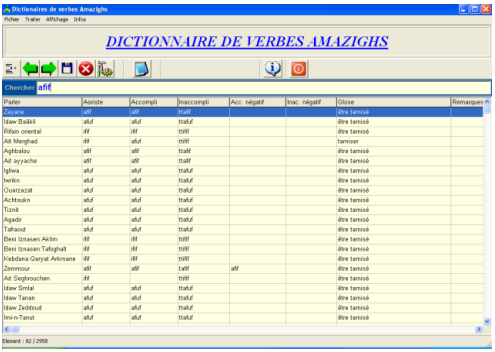
In the frame of promoting the linguistic diversity among knowledge society, we suggest to elaborate an electronic dictionary of Amazigh-French verbs (DictAm). This dictionary aims at accounting for the verbs in the Amazigh sphere as a whole: conjugation, diathesis and meaning. The DictAm also adopts a comparative perspective in the sense that it collects the lexical materials of different dialectal varieties and makes them reachable. Now, the DictAm comprises around 3000 verbs deriving from about sixty Amazigh speeches. This work is currently in progress as well as it aspires to set up a repertoire of all Amazigh verbs and their French equivalents.

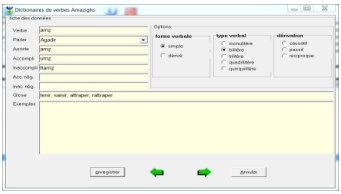
MOTS-CLES: Dictionnaire électronique, dimension bilingue, diversité linguistique, verbes.
KEYWORDS: Electronic dictionary, bilingual dimension, linguistic diversity, verbs.

Dans le cadre de la promotion de la diversité linguistique dans la société de la connaissance, nous nous proposons d'élaborer un dictionnaire électronique des verbes amazighs-français (DictAm). Ce dictionnaire vise à rendre compte de l'ensemble des verbes dans le domaine berbère : *conjugaison, diathèse et sens*. Le DictAm a également une visée comparative dans la mesure où il rassemble et rend accessible les matériaux lexicaux des différentes variétés dialectales.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 5: Démonstrations, pages 15–16,
Grenoble, 4 au 8 juin 2012. ©2012 ATALA & AFCP

La structure de la base de données a été déterminée en prenant en compte toutes les caractéristiques du verbe en amazighe. Et pour faciliter la consultation, les verbes sont classés par ordre alphabétique de leur forme aoriste-impératif. La structuration et le format des données ont été pensés de manière à permettre un transfert des données sélectionnées vers un document Word ou Excel (et prochainement HTML). Quant à la dimension bilingue du DictAm, elle se manifeste au travers de l'association pour chaque entée lexicale berbère d'un équivalent en langue française. Pour la programmation du DictAm, nous avons fait appel à El Amrani M., informaticien en Allemagne, qui nous a aidé à concrétiser ce projet.





L'alimentation de la base de données s'est faite à partir des sources documentaires existantes : les dictionnaires classiques (version papier) ; les lexiques et glossaires ainsi qu'une exploitation systématique des textes publiés. Actuellement, le DictAm comporte près de *3000 verbes* dans une *soixantaine* de parlers amazighs. C'est un travail qui est en cours de réalisation et qui a pour ambition de répertorier tous les verbes berbères ainsi que leurs équivalents en français.

Ainsi conçu, le DictAm répondra à trois types de besoins : *i)* les besoins relatifs à la collecte et à l'organisation des données lexicales issues des différentes variétés de l'amazighe ; *ii)* les besoins des apprenants et des chercheurs qui travaillent sur l'amazighe et enfin ; *iii)* les besoins des comparatistes. En préservant toute la richesse héritée des différentes variétés dialectales et en intégrant l'amazighe dans les nouvelles technologies de l'information, le DictAm va sûrement contribuer à la *promotion* de cette langue.

Vizart3D : Retour Articulatoire Visuel pour l'Aide à la Prononciation

Thomas Hueber[1] *Atef Ben-Youssef*[1]
Pierre Badin[1] *Gérard Bailly*[1] *Frédéric Eliséi*[1]
(1) GIPSA-lab, UMR 5216/CNRS/INP/UJF/U.Stendhal, Grenoble, France
(prénom.nom)@gipsa-lab.grenoble-inp.fr

RESUME

L'objectif du système Vizart3D est de fournir à un locuteur, en temps réel, et de façon automatique, un retour visuel sur ses propres mouvements articulatoires. Les applications principales de ce système sont l'aide à l'apprentissage des langues étrangères et la rééducation orthophonique (correction phonétique). Le système Vizart3D est basé sur la tête parlante 3D développée au GIPSA-lab, qui laisse apparaître, en plus des lèvres, les articulateurs de la parole normalement cachés (comme la langue). Cette tête parlante est animée automatiquement à partir du signal audio de parole, à l'aide de techniques de conversion de voix et de régression acoustico-articulatoire par GMM.

ABSTRACT

Vizart3D: Visual Articulatory Feedack for Computer-Assisted Pronunciation Training

We describe a system of visual articulatory feedback, which aims to provide any speaker with a real feedback on his/her own articulation. Application areas are computer-assisted pronunciation training (phonetic correction) for second-language learning and speech rehabilitation. This system, named Vizartd3D, is based on the 3D augmented talking head developed at GIPSA-lab, which is able to display all speech articulators including usually hidden ones like the tongue. In our approach, the talking head is animated automatically from the audio speech signal, using GMM-based voice conversion and acoustic-to-articulatory regression.

MOTS-CLES : retour visuel, aide à la prononciation, GMM, temps réel, tête parlante
KEYWORDS : visual feedback, pronunciation training, GMM, real-time, talking head

Plusieurs études semblent montrer que fournir à un locuteur un retour visuel sur ses propres mouvements articulatoires pouvait s'avérer utile pour la rééducation orthophonique et l'apprentissage des langues (Badin, 2010). Ce retour visuel peut notamment s'effectuer via une tête parlante *augmentée*, c'est-à-dire un clone orofacial virtuel qui laisse apparaître l'ensemble des articulateurs, externes (lèvres, mâchoire) comme internes (langue, voile du palais). Dans (Engwall, 2008), Engwall propose un paradigme expérimental du type « magicien d'Oz » pour montrer l'efficacité d'une telle approche (système ARTUR): un phonéticien expert évalue la nature du défaut de prononciation du sujet, et lui fait visualiser le geste articulatoire cible en sélectionnant l'animation adéquate parmi un ensemble d'animations pré calculées. Dans (Ben Youssef, 2011), nous avons proposé un système de retour articulatoire visuel également basé sur l'utilisation d'une tête parlante augmentée. Dans notre approche, la tête parlante augmentée est animée *automatiquement* à partir du

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 5: Démonstrations, pages 17–18,
Grenoble, 4 au 8 juin 2012. ©2012 ATALA & AFCP

signal audio, par inversion acoustico-articulatoire. Cependant, dans ce système, l'animation de la tête parlante ne peut débuter qu'une fois la phrase entièrement produite (approche par HMM, décodage acoustico-phonétique basé sur l'algorithme de Viterbi). C'est cette limitation que le système Vizart3D tente de lever, en proposant une version temps-réel de notre système de retour articulatoire visuel. Un schéma général du système Vizart3D est présenté à la Figure 1.

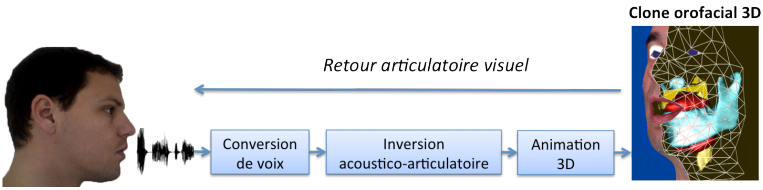


Figure 1 : Schéma général du système Vizart3D

Le système Vizart3D est basé sur la tête parlante augmentée, développée au GIPSA-lab à partir de données IRM, CT et vidéo, acquises sur un locuteur de référence. L'animation de cette tête parlante à partir de la voix d'un locuteur λ s'effectue en 3 étapes (exécutées toutes les 10 ms) : (1) Conversion de voix : l'enveloppe spectrale du locuteur λ, extraite par analyse mel-cepstrale, est transformée en une enveloppe spectrale dite « cible », qui peut être vue comme l'enveloppe qui aurait été obtenue si la même phrase avait été prononcée par le locuteur de référence ; dans notre implémentation, nous utilisons une approche basée sur une régression par GMM (*Gaussian Mixture Model*) – (2) Inversion acoustico-articulatoire : une cible articulatoire est estimée à partir de l'enveloppe spectrale cible (position de la langue (3 points), des lèvres (2 points), et de la mâchoire (1 point)); cette étape d'inversion est également basée sur une modélisation par GMM, à partir d'un corpus de données audio et articulatoires, acquises sur le locuteur de référence par articulographie électromagnétique 2D) – (3), les paramètres de contrôle de la tête parlante sont inférés par régression linéaire, à partir de la cible articulatoire estimée à l'étape 2.

Références

BADIN, P., BEN YOUSSEF, A., BAILLY, G., ELISEI, F., HUEBER, T. (2010) Visual articulatory feedback for phonetic correction in second language learning, Actes de SLATE, P1-10.

ENGWALL, O. (2008) Can audio-visual instructions help learners improve their articulation? - An ultrasound study of short term changes, Actes d'Interspeech, Brisbane, Australie, pp. 2631-2634.

BEN YOUSSEF A., HUEBER T., BADIN P., BAILLY G. (2011) Toward a multi-speaker visual articulatory feedback system, Actes d'Interspeech, Florence, Italie, pp. 489-492.

ROCme! : logiciel pour l'enregistrement et la gestion de corpus oraux

Emmanuel Ferragne[1] *Sébastien Flavier*[2] *Christian Fressard*[2]
(1) CLILLAC-ARP, Université Paris 7
(2) Dynamique Du Langage, UMR 5596, CNRS-Université de Lyon
rocme@ish-lyon.cnrs.fr

RÉSUMÉ

ROCme! permet une gestion rationalisée, autonome et dématérialisée de l'enregistrement de corpus oraux. Il dispose notamment d'une interface pour le recueil de métadonnées sur les locuteurs totalement paramétrable via des balises XML. Les locuteurs peuvent gérer les réponses au questionnaire, l'enregistrement audio, la lecture, la sauvegarde et le défilement des phrases (ou autres types de corpus) en toute autonomie. ROCme! affiche du texte, avec ou sans mise en forme HTML, des images, du son et des vidéos.

ABSTRACT

ROCme!: software for the recording and management of oral corpora

ROCme! has been designed to allow a sensible, autonomous, and dematerialized management of speech recordings. Users can create interfaces for metadata collection thanks to XML tags. Speakers autonomously fill in questionnaires, record, play, and save audio; and browse sentences (or other types of corpora). ROCme! can display text, optionally with HTML formatting, images, sounds, and video.

MOTS-CLÉS : corpus, oral, linguistique, logiciel
KEYWORDS : corpus, oral, linguistics, software

1 Résumé

Le développement du logiciel ROCme![1] – Recording of Oral Corpora Made Easy – intervient en réponse à un besoin de la communauté des linguistes. Il s'agit de : i) permettre à un locuteur d'enregistrer de façon autonome des phrases qui s'affichent à l'écran, ii) recueillir des métadonnées par le biais d'un questionnaire dématérialisé, iii) assurer la cohérence de la procédure à travers un projet d'enregistrement.

ROCme! s'appuie sur le moteur d'exécution Adobe AIR pour offrir une interface graphique conviviale compatible Windows et Mac. Les locuteurs gèrent l'enregistrement de leurs productions orales via les boutons de l'interface ou les touches du clavier. Les phrases à afficher sont incluses dans un fichier texte UTF-8.

Afin de garantir une collecte cohérente des données orales, il convient de définir un ensemble de contraintes favorisant l'homogénéité d'une série d'enregistrements ; c'est le but de la notion de projet. Ainsi, à la création d'un projet, on pourra choisir de présenter les stimuli en ordre aléatoire, et de faire apparaître un masque entre le déclenchement de l'enregistrement et l'apparition d'un stimulus (pour éviter que le début ne soit coupé).

[1] www.ddl.ish-lyon.cnrs.fr/rocme

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 5: Démonstrations, pages 19–20,
Grenoble, 4 au 8 juin 2012. ©2012 ATALA & AFCP

On peut aussi interdire la sauvegarde d'un signal écrêté, prévoir un mot de passe empêchant le participant de quitter le prompteur, afficher une horloge (utile pour des locuteurs payés à l'heure), passer dans un mode où chaque locuteur ne lit qu'une partie du corpus, ou encore enregistrer plusieurs fois chaque phrase. Concernant la collecte des métadonnées, le projet peut prévoir que l'utilisateur soit invité à remplir lui-même son questionnaire. Enfin, les options audio (taux d'échantillonnage, résolution d'amplitude et nombre de canaux) sont paramétrables.

ROCme! permet la mise en place d'un questionnaire pour le recueil de métadonnées à partir de balises XML. À la création, le code XML peut être directement entré dans une boîte de l'interface pour pré-visualiser la version graphique du questionnaire. Un exemple est présenté dans la Figure 1, qui inclut les 6 valeurs de types de données prises en compte par ROCme ! (age, date, number, single, choice, multiple). Spécifier un type de données est optionnel, mais vivement recommandé car ROCme! dispose d'un module qui traite les valeurs de type pour proposer en temps réel des statistiques descriptives concernant le projet, permettant ainsi à l'investigateur de repérer rapidement certains déséquilibres et donc d'ajuster sa stratégie de recrutement.

```
<user>
    <field id="Nom" required="required"></field>
    <field id="Sexe" type="single" required="required">
        <field id="Homme"></field>
        <field id="Femme"></field>
        </field>
        <field id="Âge" type="age" required="required"></field>
        <field id="Date_de_naissance" type="date"></field>
        <field id="Nombre_d'enfants" type="number"></field>
        <field id="Lieu_de_résidence" lbl="Lieu de résidence
                des 10 dernières années" type="choice">
            <field id="Nord"></field>
            <field id="Sud"></field>
        </field>
        <field id="Langues_parlées" type="multiple">
        <group>
            <field id="Langue"/>
         </group>
    </field>
</user>
```

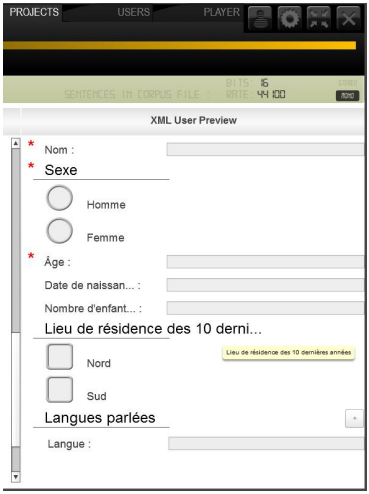


FIGURE 1 – Questionnaire en version XML et interface graphique correspondante

Remerciements

Le projet ROCme! a reçu le soutien financier du Bonus Qualité Recherche de l'Université Paris 7 Diderot et le soutien technique du Laboratoire Dynamique Du Langage.

Index