RELAX

Extraction de relations sémantiques dans les contextes biographiques

Michaela Geierhos — Olivier Blanc — Sandra Bsiri

Centrum für Informations- und Sprachverarbeitung (CIS)
Ludwig-Maximilians-Universität München
Oettingenstraße 67, D-80538 München, Allemagne
{michaela.geierhos,olivier.blanc,sandra.bsiri}@cis.uni-muenchen.de

RÉSUMÉ. *L'extraction d'informations biographiques dans les textes est une tâche complexe. Ce papier présente le contexte linguistique et la modélisation de notre système RELAX (RELation eXtraction), moteur d'extraction de relations sémantiques dans les textes permettant de relier les personnes aux événements qui sont survenus au cours de leur vie. La notion de relation biographique est décrite ici comme une relation entre deux personnes ou entre une personne et un événement biographique, tel que la naissance, le mariage, le divorce ou le parcours professionnel, etc. Notre approche consiste à caractériser les relations biographiques à l'aide de grammaires locales. Nos résultats montrent que notre approche est viable et que notre système peut être utilisé comme base pour l'implémentation de systèmes de questions-réponses.*

ABSTRACT. *The automatic extraction of biographical information from business news is a complex task. This approach deals with the characterization of the so called biographical relations by means of local grammars. In order to provide a well-founded proceeding, it is necessary to give a complete and accurate definition of the notion "relation" seen here as a social relationship between human beings or expressing the binding of person to a biographical event, such as birth, marriage, divorce, professional career etc. This paper outlines the linguistic background and modeling method of our system RELAX doing semantic RELAtion eXtraction and linking persons to their corresponding biographical information. Results show that software applications like question-answering systems can successfully deal with this method and efficiently retrieve semantic knowledge out of text corpora.*

MOTS-CLÉS : *extraction d'informations biographiques, relations sémantiques, grammaires locales, entités nommées, questions-réponses.*

KEYWORDS: *biographical information extraction, semantic relations, local grammars, named entities, question-answering.*

1. Introduction

La masse toujours croissante de documents disponibles sur Internet rend l'accès à l'information difficile même avec l'usage indispensable des moteurs de recherche. Les informations relatives aux personnalités qui font l'actualité dans le monde ne font pas exception. Il est, en effet, très rare de trouver des biographies complètes sur les personnes d'intérêt public. Les textes journalistiques donnent de nouvelles informations biographiques concernant les personnes citées, mais celles-ci demeurent néanmoins souvent partielles. Ainsi pour répondre à un besoin informationnel sur un personnage de notoriété mondiale, l'utilisateur est contraint de lancer une requête sur un moteur de recherche à partir du nom de la personne et de mots-clés associés à l'événement biographique d'intérêt, de parcourir ensuite les multiples documents résultats, susceptibles de contenir l'information recherchée, pour enfin rassembler les différents passages associés à son besoin et retracer ainsi la biographie souhaitée. Cette méthode s'avère répétitive et peu fiable, car l'information présente dans les diverses sources électroniques n'est pas indexée sur des critères sémantiques par les moteurs de recherche mais les documents sont prétraités uniquement sur la base des mots qu'ils contiennent. C'est donc à l'utilisateur que revient la tâche de classer les différents documents donnés en réponse par le système et de retrouver dans ceux-ci les segments pertinents pour ses besoins personnels.

L'extraction des données biographiques contenues dans les diverses ressources textuelles diffusées sur Internet est une tâche complexe, fortement dépendante de la langue et des phénomènes linguistiques associés (Poibeau, 2003). De nombreux auteurs ont reconnu la complexité du problème (Kanzaki, 2003–2007 ; Davis et Galbraith, 2004) et se sont intéressés à l'identification des indicateurs spécifiques internes et externes utiles à la reconnaissance des événements biographiques (Agichtein et Gravano, 2000 ; Duboué *et al.*, 2003). Les quelques nomenclatures accessibles sont malheureusement incomplètes, voire inconsistantes. Une étude approfondie élaborée par (Sparck-Jones, 1993) montre l'importance de la distinction des faits sémantiques les plus pertinents comme base pour les résumés automatiques et dans ce cas pour la reconstitution de résumés biographiques à partir de plusieurs documents (Schiffman *et al.*, 2001).

L'extraction automatique des relations sémantiques à partir de documents textuels non structurés est utile pour plusieurs domaines applicatifs distincts comme les systèmes de résumés automatiques, les systèmes de classification automatique, ou aussi les systèmes de questions-réponses (Tsur *et al.*, 2004). Ces travaux se concentrent essentiellement sur le sous-langage économique et financier de la presse en langue anglaise. La performance des extractions est d'autant plus satisfaisante que les documents analysés se restreignent à ce domaine spécialisé. Une étude semblable a été élaborée pour les documents en langue française par (Kevers, 2006).

Les travaux présentés dans cet article proposent une méthode de découverte des contextes riches en connaissances, notion connue en anglais sous *« knowledge rich contexts (KRCs) »* (Meyer, 2001) et dépeignent les patrons lexicosyntaxiques (*« know-*

ledge patterns » – KPs), indicateurs sémantiques permettant de repérer les événements biographiques liés à une personne. Outre l'enrichissement automatique d'une nomenclature de faits biographiques, les KPs nous ont permis d'augmenter sensiblement et d'une manière automatique les différents dictionnaires thématiques que nous utilisons pour l'extraction des contextes au moyen de grammaires locales (M. Gross, 1993 ; M. Gross, 1997).

Les énoncés que nous considérons décrivent des relations entre des personnes d'un côté et des événements biographiques de l'autre. La notion d'information biographique est définie dans la seconde section de cet article, dans laquelle nous détaillons les entités susceptibles d'intervenir dans de telles relations. Dans les sections 3 et 4 nous présentons les ressources linguistiques (grammaires locales et dictionnaires terminologiques) que nous avons construites pour l'extraction de ce type de relations. Nous y présentons également nos méthodes utilisant des patrons lexicosyntaxiques, pour l'acquisition semi-automatique de nouvelles instances terminologiques et la découverte de locutions verbales synonymiques à partir de verbes initiaux qui nous ont permis d'enrichir de manière conséquente ces deux types de ressources. À la section 5, nous présentons, sur un exemple de prédicat (*« to be born »*), les différentes relations biographiques extraites par notre système. Nous donnons, dans la section 6, une évaluation de la qualité d'extraction de notre système et nous concluons dans les sections 7 et 8 en présentant le système de questions-réponses `LaolaWeb`, qui constitue une première application directe de ces travaux.

2. L'information biographique

Retracer la biographie d'une personne consiste à établir la liste de l'ensemble des événements survenus tout au long de sa vie, tels que sa naissance, son parcours scolaire et professionnel, ses relations privées et professionnelles établies avec d'autres individus, sa mort. Pour délimiter formellement quels sont les énoncés en langue naturelle dont la sémantique porte une information biographique, nous utilisons le modèle des classes d'objets (G. Gross, 1994 ; Le Pesant et Mathieu-Colas, 1998) : les classes d'objets sont « des classes sémantiques construites à partir de critères syntaxiques » (Le Pesant et Mathieu-Colas, 1998) ; celles-ci sont définies par des prédicats définitionnels sémantiquement homogènes de type verbes, adjectifs ou noms auxquels correspondent des domaines d'arguments. Par exemple, la classe ⟨Profession⟩ étudiée par (Buvet et Foucou, 2001), est l'ensemble des noms simples et composés répondant essentiellement aux prédicats *« gagner sa vie comme »* et *« exercer la profession de »* et contient des instances telles que *« ingénieur, instituteur, second de cuisine »*.

Dans ce contexte, nous définissons une information biographique comme une relation prédicative à plusieurs arguments dont l'un est nécessairement une entité de la classe ⟨Personne⟩. Cette classe peut se subdiviser à son tour en la sous-classe ⟨Nom Propre⟩ et la sous-classe ⟨Rôle Social⟩. Il n'y a pas de restriction de sélection sur les autres intervenants de la relation pour que celle-ci portent une information biographique. Cependant, dans les différentes relations que nous avons étudiées

dans le cadre de ces travaux, les autres arguments sont typiquement des instances de classes ⟨Personne⟩, ⟨Lieu⟩, ⟨Date⟩, ⟨Organisation⟩, ⟨Secteur d'activité⟩, ⟨Matière⟩ ou ⟨Profession⟩.

L'« information biographique » est ainsi une relation entre une personne et un événement biographique, qui peut être exprimée à l'aide d'un verbe prédicatif, retraçant le portrait de quelqu'un (1a) à (1c).

(1) a. *Sigman, born in Brooklyn in 1909.*
b. *Andrew Gilligan graduated from Cambridge with a degree in history.*
c. *Jim Sweeney will also be joining AmeriQuest as Vice President.*

La sémantique d'événements biographiques n'est pas nécessairement portée par un verbe ; elle peut également être portée par d'autres catégories grammaticales prédicatives comme le montrent les phrases à verbes supports suivantes :

(2) a. *Elizabeth gave birth to a little girl in May 2004.*
b. *Paul and Claire became man and wife in 1998.*
c. *John Smith was born in Florida on February 12, 1965.*
d. *Jacob McCandles is six-feet under.*
e. *Gov. Greenhalge breathed his last at his home in Lowell, Mass.*

Ces exemples montrent que la fonction de prédicat peut aussi être remplie par un nom prédicatif (2a) et (2b), un adjectif prédicatif (2c) ou une expression idiomatique (2d) et (2e), nous parlons dorénavant de locution verbale pour désigner n'importe lequel de ces types de prédicats.

Dans le cadre de ces travaux, nous nous sommes essentiellement intéressés à douze types d'événements biographiques : six événements que nous classifions dans les événements personnels : la naissance, l'enfance, la formation, le mariage, le divorce et le décès ; six événements relatifs à la carrière professionnelle d'une personne : l'obtention d'un emploi, l'occupation d'un poste, le licenciement, le succession, la démission et le départ en retraite.

3. Les grammaires locales et sous-langages

Nous partons de l'hypothèse que l'ensemble des énoncés décrivant une information biographique se caractérisent par un lexique de taille finie et un nombre de schémas de phrases limité. En ce sens, nous considérons qu'ils constituent un sous-langage, dans le sens de (Harris, 1968). (Hunston et Sinclair, 2000) montrent qu'il est possible de considérer les grammaires locales comme des petits sous-langages et que par

conséquent, pour un domaine donné (ici, l'information biographique), il est possible d'élaborer un ensemble de grammaires locales étendu couvrant au mieux la totalité du sous-langage.

Selon (Sager, 1986) « Le caractère distinctif d'un sous-langage est que pour certains sous ensembles des phrases du langage, il existe des restrictions de sélection, pour lesquelles on ne peut pas fournir de règles dans le cas général ».

Les grammaires locales que nous considérons sont des réseaux de transitions récursifs (Woods, 1970), représentées par des graphes dont la construction et la manipulation sont facilitées par le logiciel libre `Unitex`[1] (Paumier, 2004 ; Silberztein, 1993). Elles n'ont pas pour vocation de décrire l'ensemble de la grammaire d'une langue, mais décrivent les structures syntaxiques et lexicales des phénomènes linguistiques (Nakamura, 2005) propres à ce langage spécialisé (Harris, 1988 ; Harris, 1991).

4. Enrichissement du lexique à travers les contextes riches en connaissances

Nous présentons dans cette section les ressources lexicales dont nous disposons ainsi que la méthode par laquelle il nous a été possible d'enrichir automatiquement nos bases lexicales associées aux classes d'objets identifiées comme intervenant dans les relations biographiques. Nous définissons, dans la deuxième section de cet article, la notion d'« information biographique » comme une relation prédicative entre plusieurs arguments dont l'un est une entité de la classe ⟨Personne⟩, les autres intervenants pouvant alternativement appartenir aux classes d'objets ⟨Lieu⟩, ⟨Date⟩, ⟨Organisation⟩, ⟨Branche⟩, ⟨Matière⟩ ou ⟨Profession⟩. Chacune de ces classes a été traduite par un dictionnaire électronique de la forme DELA (Courtois et Silberztein, 1990 ; Courtois, 2004) dont les entrées lexicales reprennent les instances hyponymes du nom de la classe d'objet source.

La pertinence de l'analyse contextuelle et par conséquent la qualité du système d'extraction automatique des faits biographiques est d'autant plus satisfaisante que l'on dispose de bases de connaissances riches en entrées lexicales. Ainsi, plus les classes d'objets sont riches en instances, plus les extractions sont pertinentes et les analyses contextuelles nécessaires à la levée d'ambiguïté sont rudimentaires.

4.1. *Acquisition automatique de nouvelles entrées lexicales*

Nous avons ainsi constitué un lexique de spécialités propres au monde de l'entreprise (activité professionnelle, secteur d'activité, etc.), ainsi qu'un lexique de noms propres pour les personnes, les toponymes et les organisations à partir de di-

1. `http://www-igm.univ-mlv.fr/~unitex`

verses ressources disponibles sur Internet (Wikipedia[2], WordNet[3], Biography.com[4], SpecialistInfo.com[5], ZoomInfo.com[6], Guide to the World of Occupations[7], LabourMarket[8], MapPlanet.com[9], Occupational Outlook Handbook[10], Prospects.ac.uk[11], etc.) et des lexiques du laboratoire CIS.

Cependant, lors de nos tests préliminaires d'analyse sur corpus, nous avons observé qu'il existait de nombreuses unités lexicales, hyponymes des classes d'objets présentées ci-dessus qui n'étaient pas encore recensées dans nos dictionnaires électroniques. C'est là qu'interviennent les patrons lexicosyntaxiques (KPs) (Meyer, 2001) pour découvrir de nouvelles instances associées à ces différentes catégories. Nous montrons sur un exemple de la classe ⟨Secteur d'activité⟩ comment un tel gain automatique de connaissances est possible. La désignation du secteur d'activité respecte une structure syntaxique où le contexte droit est toujours représenté par un descripteur comme *« industry »*, *« sector »* ou *« company »* qui peut agir en qualité de déclencheur (KPs) pour reconnaître de nouveaux noms de secteurs.

(3) a. *administration sector*
b. *automobile industry*
c. *arts and leisure sector*

Ainsi un KP recensant ces descripteurs et décrivant la structure syntaxique interne d'un nom de secteur permet d'identifier de nouvelles instances de cette classe dans les textes analysés. Les noms des secteurs de l'exemple (3) peuvent être également retrouvés dans les textes accompagnés par des descripteurs différents : il ne s'agit pas de séquences figées comme le montre le terme *automobile* de l'exemple (3b) qui peut aussi apparaître dans d'autres séquences comme *« automobile industry»* ou *« automobile business »*. Pour ce genre de termes, nous ne retenons dans les classes d'objets correspondantes que la séquence sans le descripteur. Ce qui revient ici à ajouter le terme *« automobile »* aux instances de la classe ⟨Secteur d'activité⟩ et non pas les deux séquences *« automobile industry »* et *« automobile business »* qui seront reconnues ultérieurement dans les textes à travers les grammaires descriptives dotées des déclencheurs *« business »* et *« industry »*. Certains noms de secteurs sont cependant moins flexibles ; c'est notamment généralement le cas des termes associés au descripteur *« service »*, tels que *« reparation service »* ou *« animal physiotherapy services »*.

2. http://en.wikipedia.org/wiki/
3. http://wordnet.princeton.edu/
4. http://www.biography.com
5. http://www.specialistinfo.com
6. http://www.zoominfo.com
7. http://www.occupationsguide.cz/en/abecedni/abecedni.htm
8. http://www.labourmarket.co.nz/labourmarket.htm
9. http://mapplanet.com/ix/
10. http://www.umsl.edu/services/govdocs/ooh20002001/1.htm
11. http://www.prospects.ac.uk

De telles séquences, où *« service »* ne peut être remplacé par aucun autre descripteur, viennent enrichir les classes d'objets en tant que mots composés figés. Cette méthode nous a permis de découvrir plus de 40 000 nouvelles entrées à partir de 10 000 instances initiales pour la classe ⟨Secteur d'activité⟩ vérifiées manuellement.

Une seconde méthode adoptée pour l'augmentation automatique des instances des classes d'objets fut déjà considérée par (Mallchok, 2004), pour enrichir sa classe des noms d'organisations. Il s'agit de trouver les instances respectant une certaine structure interne et précédées par un contexte gauche formulé à partir d'un adjectif spécifique aux organisations. Une liste de 110 adjectifs spécifiques (*« newly formed, privatized, profitable, worldwide »*) a alors été extraite à partir du corpus Reuter et des contextes des organisations déjà classifiés. Une deuxième liste de preuves plus génériques a également été sélectionnée à partir de ces mêmes contextes, comme par exemple *« available from, award from the, awarded to »* que nous avons reprise et appliquée à nos corpus d'apprentissage pour enrichir les deux classes d'objets ⟨Nom d'organisation⟩ et ⟨Type d'organisation⟩.

Classe d'objet	***Sous-classe d'objet***	***Balise sémantique***	***Nombre d'instances***	***Exemples pour des instances de classe d'objet***
Nome Propre	Titre	⟨Title⟩	370	*Queen, Lord, PhD, Mr.*
	Prénom	⟨FirstName⟩	38 500	*Lara, Marie-Luise, Ben*
	Nom de Famille	⟨Surname⟩	1 250 000	*Oltay-Smith, Yildiz*
	Nom de Personne (Complet)	⟨LongName⟩	8 300 000	*Henna Nordqvist*
Rôle Social	Famille	⟨Human⟩	6 400	*daughter, son, aunt*
	Profession	⟨JobDescriptor⟩	45 000	*cook, kitchen helper*
	Habitant	⟨Citizen⟩	600	*Aucklander, Brooklyner*
Secteur d'activité	Matière	⟨Discipline⟩	580	*art history*
	Branche	⟨Sector⟩	38 000	*life insurance, farming*
Organisation	Type d'organisation	⟨CompanyDescriptor⟩	23 800	*car manufacturer*
	Nom d'organisation	⟨Company⟩	516 000	*Fujitsu Siemens*
	Forme juridique d'entreprise	⟨LegalForm⟩	115	*ltd, inc, plc, AG, GmbH, b.v., S.A., s.a.r.l., LLC*
Lieu	Pays et Continent	⟨Nation⟩ ⟨Continent⟩	430	*South America, France, Germany, Europe*
	Ville	⟨City⟩	327 400	*'s-Gravenhage, Paris*
Date	Mois	⟨Month⟩	24	*June, Jun., July*
	Jour de semaine	⟨DayOfWeek⟩	7	*Saturday, Sunday*

Tableau 1. *Bref aperçu des classes d'objets et leurs structures*

Le tableau 1 résume l'ensemble des classes d'objets que nous avons identifiées comme étant *a priori* nécessaires au bon déroulement de l'extraction de l'information biographique.

La somme des entrées dépasse, à ce jour, les 10 millions d'unités lexicales dont environ 15 % étaient découvertes à l'aide de nos méthodes de Bootstrapping. Chaque unité est représentée dans le dictionnaire, accompagnée d'une étiquette sémantique correspondant au nom de la classe d'objet associée. Nous montrons, section 6.3, que ces nouvelles entrées apprises par *Bootstrapping* nous ont permis d'améliorer de manière conséquente les résultats de notre système d'extraction d'informations, au niveau du rappel notamment.

4.2. *Bootstrapping et extraction des prédicats synonymiques*

Pour chacun des douze types d'événements biographiques étudiés (*cf.* tableaux 2 et 3), nous avons sélectionné une liste restreinte de verbes à partir desquels nous avons extrait des locutions verbales synonymiques de manière semi-automatique.

La découverte des relations synonymiques associées aux verbes initiaux fut effective à l'aide de techniques de Bootstrapping (Senellart, 1998 ; M. Gross, 1999a) appliquées de manière itérative. Le processus complet d'extraction de candidats synonymiques est assez complexe et implique un enrichissement des contextes internes par la détection de contextes externes spécifiques et inversement. Nous en présentons un exemple dans la section suivante décrivant la découverte de nouveaux prédicats verbaux apparaissant régulièrement dans les structures de type $N_{hum}\ V\ as\ N_{profession}$. Pour une description complète de l'ensemble du processus nous renvoyons à (Geierhos, 2007).

Nous donnons dans les tableaux 2 et 3 des échantillons de verbes et de locutions synonymiques extraites au moyen des grammaires locales développées à partir d'un corpus renfermant un an de dépêches du *Financial Times* apparues en 2004.

Le tableau 2 présente des exemples de constructions prédicatives relatives aux informations biographiques dites « personnelles », nous traitons en totalité 50 prédicats différents pour ce type de relations.

Les prédicats du tableau 3 concernent les informations professionnelles et décrivent les différentes relations pouvant exister entre une personne, son activité professionnelle et son employeur. Nous avons recensé un total de 95 prédicats pour ce type de relations, englobant les structures verbales autour de la date d'embauche, la date de départ, le type de l'activité exercée et la position hiérarchique dans l'entreprise.

Schéma initial	*Formes synonymiques*	*Classe d'objet comme sujet*	*Classe d'objet comme objet*
Naissance			
X was born in D/L	X saw the light of day in D/L	X : Personne	D : Date, L : Lieu
X was born as N	X saw the light of day as N	X : Personne	N : Nom Propre
Naissance/Enfance			
X was born and raised (up) in L	X was born and brought up in L	X : Personne	L : Lieu
Enfance			
X was raised (up) in L	X was brought up in L	X : Personne	L : Lieu
	X spent X's childhood in L	X : Personne	L : Lieu
	X grew up in L	X : Personne	L : Lieu
Obtention d'un diplôme			
X graduated in D/L	X become a graduate from O in D/L	X : Personne	D : Date, L : Lieu O : Organisation
	X took X's degree in M from O in D/L	X : Personne	D : Date, L : Lieu M : Matière O : Organisation
	X received X's degree in M from O in D/L	X : Personne	D : Date, L : Lieu M : Matière O : Organisation
	X got X's degree in M from O in D/L	X : Personne	D : Date, L : Lieu M : Matière O : Organisation
	X completed X's studies in M from O in D/L	X : Personne	D : Date, L : Lieu M : Matière O : Organisation
Mariage			
X married Y	X and Y became man and wife	X : Personne	Y : Personne
	X joint in marriage with Y	X : Personne	Y : Personne
	X plighted X's troth to Y	X : Personne	Y : Personne
	X pledged X's troth to Y	X : Personne	Y : Personne
	X took Y to wife/husband	X : Personne	Y : Personne
	X wedded Y	X : Personne	Y : Personne
	X led Y to the altar	X : Personne	Y : Personne
	Y was married to X	X : Personne	Y : Personne
	Y got married to X	X : Personne	Y : Personne
	Y was wedded to X	X : Personne	Y : Personne
Divorce			
X was divorced from Y	X filed a divorce from Y	X : Personne	Y : Personne
	X sued for divorce from Y	X : Personne	Y : Personne
	X got a divorce from Y	X : Personne	Y : Personne
	X parted from Y	X : Personne	Y : Personne
	X separated from Y	X : Personne	Y : Personne
	X split from Y	X : Personne	Y : Personne
	X split up with Y	X : Personne	Y : Personne
	X broke up with Y	X : Personne	Y : Personne
	X ended X's marriage to Y	X : Personne	Y : Personne
	X annulled X's marriage to Y	X : Personne	Y : Personne
	X dissolved X's marriage to Y	X : Personne	Y : Personne
	X parted company with Y	X : Personne	Y : Personne
Décès			
X died in D/L	X breathed X's last in D/L	X : Personne	D : Date, L : Lieu
	X deceased in D/L	X : Personne	D : Date, L : Lieu
	X departed X's life in D/L	X : Personne	D : Date, L : Lieu
	X laid down X's life in D/L	X : Personne	D : Date, L : Lieu
	X lost X's life in D/L	X : Personne	D : Date, L : Lieu
	X met X's death in D/L	X : Personne	D : Date, L : Lieu
	X passed away in D/L	X : Personne	D : Date, L : Lieu
	X perished in D/L	X : Personne	D : Date, L : Lieu

Tableau 2. *Nomenclature d'informations personnelles*

Schéma initial	***Formes synonymiques***	***Classe d'objet comme sujet***	***Classe d'objet comme objet***
Obtention d'emploi			
X was appointed (as) P	X was adopted as P	X : Personne	P : Profession
	X was commissioned as P	X : Personne	P : Profession
	X was designated as P	X : Personne	P : Profession
	X was elected as P	X : Personne	P : Profession
	X was installed as P	X : Personne	P : Profession
	X was named as P	X : Personne	P : Profession
	X was nominated as P	X : Personne	P : Profession
	X was selected as P	X : Personne	P : Profession
X joint O as P (of B) in D	X became member of O as P (of B) in D	X : Personne	P : Profession O : Organisation B : Branche D : Date
Occupation d'un poste			
X was employed as P	X was engaged as P	X : Personne	P : Profession
	X was hired as P	X : Personne	P : Profession
	X was recruited as P	X : Personne	P : Profession
X was paid as P by O	X drew salary by O	X : Personne	P : Profession O : Organisation
X worked as P for O	X served as P for O	X : Personne	P : Profession O : Organisation
	X jobbed as P for O	X : Personne	P : Profession O : Organisation
	X laboured as P for O	X : Personne	P : Profession O : Organisation
Licenciement			
X was dismissed as P	X was fired as by O/P	X : Personne	P : Profession O : Organisation
	X was dismissed as P	X : Personne	P : Profession
	X was removed as P of O	X : Personne	P : Profession O : Organisation
Succession			
X was replaced as P by Y	X was succeeded as P by Y	X : Personne	P : Profession Y : Personne
	X was followed as P by Y	X : Personne	P : Profession Y : Personne
Démission			
X resigned as P of O	X quitted as P of O	X : Personne	P : Profession O : Organisation
	X left job as P of O	X : Personne	P : Profession O : Organisation
Départ en retraite			
X retired as P in D	X stopped working as P	X : Personne	P : Profession
	X stopped work as P in D	X : Personne	P : Profession O : Organisation
	X gave up work as P in D	X : Personne	P : Profession O : Organisation
	X reached retirement age in D	X : Personne	D : Date

Tableau 3. *Nomenclature d'informations professionnelles*

Exemples de Bootstrapping

Il est difficile d'évaluer avec précision l'apport du Bootstrapping pour l'enrichissement de nos dictionnaires et grammaires, étant donné que nous n'avons malheureusement pas conservé toutes les information sur l'origine de chaque entrée présente dans les ressources linguistiques que nous utilisons qui sont les résultats du travail collaboratif de plusieurs personnes. Cependant, nous tentons de donner quelques chiffres significatifs dans cette section en prenant l'exemple d'extraction de nouvelles instances associées à la classe d'objet ⟨Profession⟩ dans les textes ainsi que celui de la découverte de nouveaux prédicats verbaux relatifs à l'occupation d'un poste. Notre approche est basée sur l'utilisation de patrons lexicosyntaxiques (ou KPs (Meyer, 2001)) décrits dans des grammaires locales.

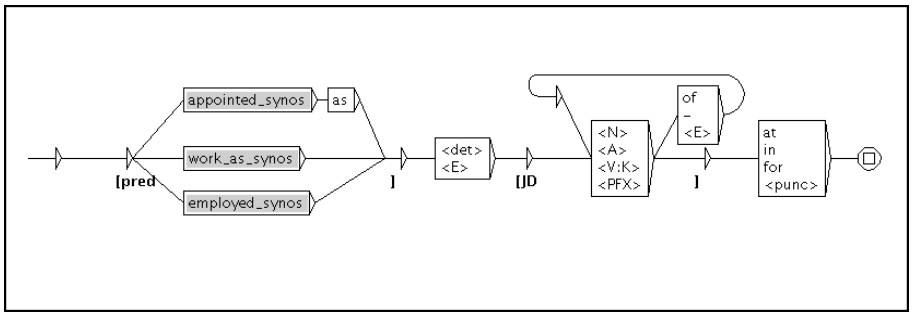

Figure 1. *Grammaire locale pour la reconnaissance de nouvelles instances de la classe ⟨Profession⟩*

Ainsi, par exemple, le graphe de la figure 1 décrit un contexte dans lequel sont susceptibles d'apparaître des noms de profession : le nom apparaît à droite d'un verbe conjugué synonyme de *« work as »* (la liste de ces verbes est décrite dans le sous-graphe `work_as_synos` et est composée des différents prédicats synonymiques que nous avons recensés). Pour notre expérience, nous avons imposé que la phrase nominale extraite soit suivie par un signe de ponctuation (étiquette ⟨punc⟩) ou par une des prépositions *« at »*, *« in »* ou *« for »* qui, dans un tel contexte, tendent à introduire un nom de compagnie ou d'organisation. La forme exacte de la phrase nominale est décrite comme étant une séquence de longueur variable de noms ou d'adjectifs éventuellement séparés par un trait d'union ou la préposition *« of »*. Ce motif permet de reconnaître des noms composés de différentes structures internes tels que : *« vice-president »* (PFX-N), *« assistant manager »* (NN), *« director of marketing »* (N of N).

De même, nous avons utilisé le graphe de la figure 2 de manière à découvrir de nouveaux prédicats relatifs à l'occupation d'un poste. Nous nous sommes limités ici à la recherche des prédicats verbaux simples acceptant un sujet humain (décrit dans le sous-graphe `:NOhum`) et un argument nominal introduit par le préposition *« as »* dont la tête est une instance de la classe ⟨Profession⟩ recensée dans notre dictionnaire.

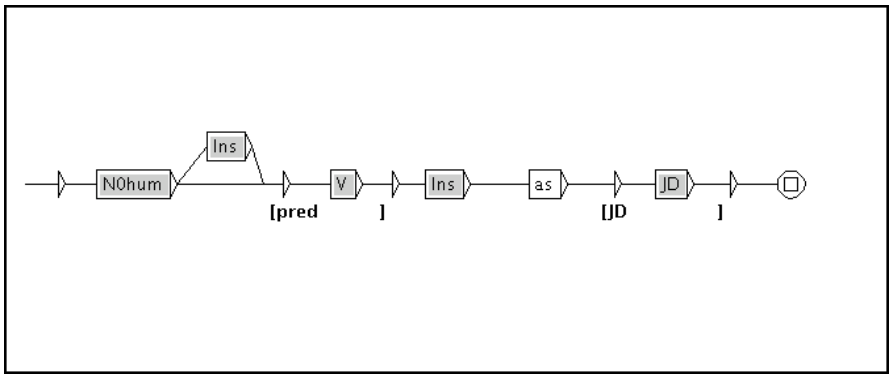

Figure 2. *Grammaire locale pour la découverte de prédicats verbaux relatifs à l'occupation d'un poste*

En faisant varier le motif décrit entre le nom humain et le nom de profession, nous pouvons trouver d'autres prédicats acceptant des schémas de phrases différents.

Nous avons appliqué ces deux grammaires à l'année 2004 du *Financial Times* à l'aide du logiciel `Outilex` (Blanc *et al.*, 2006). Le corpus a été prétraité en annotant les verbes modifiés par différents types d'auxiliaires à l'aide de la grammaire de lemmatisation de (M. Gross, 1999b) afin de traiter ces séquences reconnues comme de simples unités verbales. Nous avons obtenu un total de 6 345 concordances pour la grammaire de la figure 1 et de 6 640 concordances pour la grammaire de la figure 2. Nous avons extrait les différents candidats à partir de ces concordances en les triant par fréquence d'apparition. Les figures 3 et 4 présentent les 10 meilleurs candidats obtenus pour chacune de ces catégories.

1.	*president*
2.	*consultant*
3.	*assistant manager*
4.	*preferred bidder*
5.	*journalist*
6.	*prostitues*
7.	*defendants*
8.	*security guard*
9.	*company*
10.	*teacher*

Figure 3. *Les 10 meilleurs candidats pour des instances de la classe* ⟨*Profession*⟩

En considérant uniquement les termes apparaissant un minimum de trois fois au total, nous obtenons 560 candidats que nous avons vérifiés manuellement ; 151 ont ainsi été validés comme étant des termes appartenant effectivement à la classe ⟨Profession⟩.

1.	*to serve as*
2.	*to act as*
3.	*to be appointed as*
4.	*to continue as*
5.	*to remain as*
6.	*to join as*
7.	*to operate as*
8.	*to be employed as*
9.	*to be elected as*
10.	*to be hired as*

Figure 4. *Les 10 meilleurs candidats pour des prédicats relatifs à la carrière professionnelle*

De même, sur les 50 candidats verbaux ayant été reconnus au moins six fois, nous en avons comptabilisé 23 comme étant effectivement des prédicats décrivant des événements relatifs au parcours professionnel.

Les tableaux des figures 5 et 6 présentent respectivement le nombre de nouvelles instances de la classe ⟨Profession⟩ et les prédicats verbaux relatifs à l'activité professionnelle découverts pour chaque mois de l'année 2004 dans le corpus du *Financial Times*, c'est-à-dire les termes n'ayant pas été découverts dans un précédent mois.

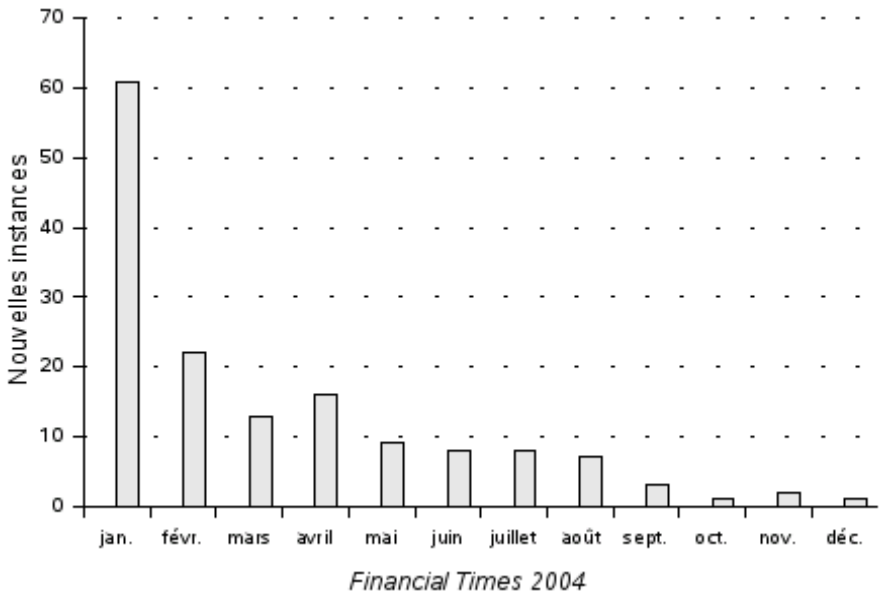

Figure 5. *Nouvelles instances de la classe ⟨Profession⟩ découvertes chaque mois*

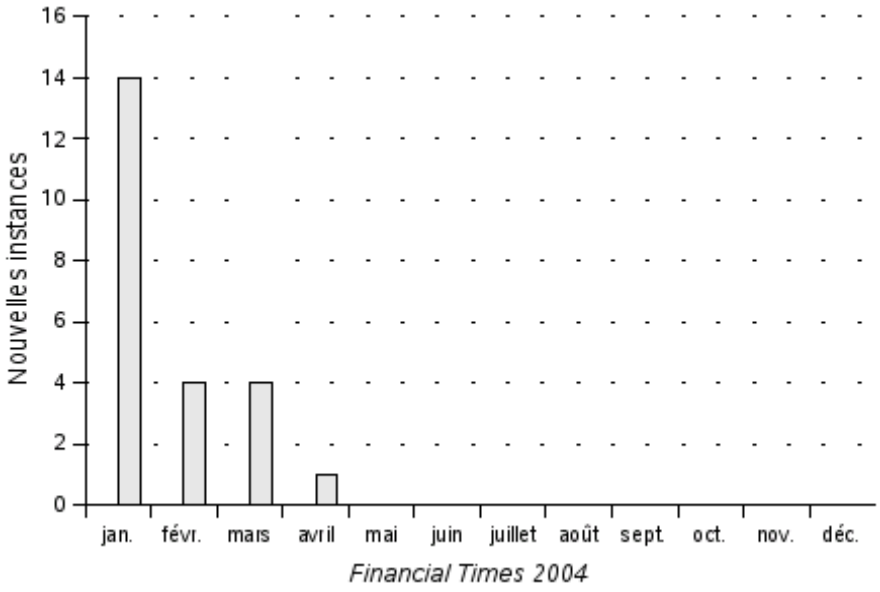

Figure 6. *Nombre de nouveaux prédicats verbaux découverts chaque mois*

5. Exemple de relation : *« to be born »*

La naissance est l'un des événements les plus décrits dans les curriculum vitæ. On y retrouve, dans la plupart des cas, les date et lieu de naissance et éventuellement les noms des parents. L'exemple 4 présente les différentes locutions verbales que nous avons recensées permettant d'exprimer cet événement. Seul le prédicat *« to be born »* était initialement connu, les autres (4b) à (4d) sont des locutions verbales synonymiques ayant été découvertes par Bootstrapping.

(4) a. *to be born*
b. *to be born and raised (up)*
c. *to be born and brought up*
d. *to see the light of day*

Dans le cas où l'individu est né au même endroit où il a passé son enfance, les deux prédicats *« to be born »* et *« to be raised (up) »* sont conjointement liés.

Dans la figure 7, nous présentons le graphe principal de notre grammaire décrivant les différents énoncés de la relation *« to be born »*.

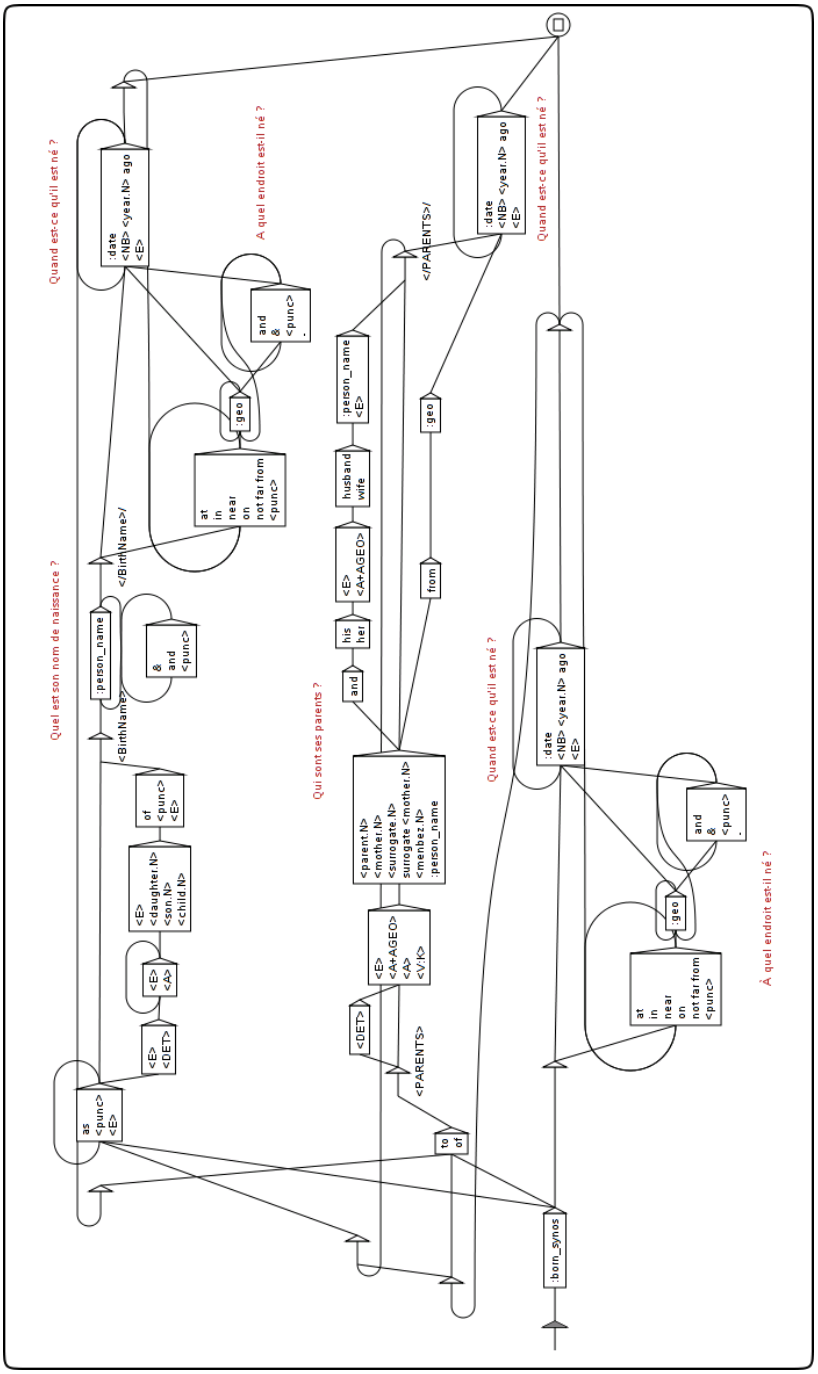

Figure 7. *Grammaire locale de la structure syntaxique de* « to be born »

Les transductions qui décorent notre grammaire permettent d'insérer des balises de type XML autour des différentes entités reconnues lors de son application sur un texte, produisant ainsi un texte annoté dans lequel les différentes informations biographiques et leurs arguments ont été identifiés.

L'emploi de cette grammaire (*cf.* figure 7) permet ainsi de répondre à un certain nombre de questions en relation directe avec la naissance d'un individu, à savoir :

(5) Quand est-ce qu'il est né ?
   a. *Tunku Abdul Rahman was born <Date>in 1903</Date>*
   b. *Cecil Beaton, who was born <Date>in 1904</Date>*
   c. *heir first child, Aidan Gering Pollack, born <Date>last week</Date>*

(6) À quel endroit est-il né ?
   a. *Sigman, born in <GEO>Brooklyn</GEO> in 1909*
   b. *the former mayor of Tel Aviv was born in <GEO>Germany</GEO>*
   c. *McQueen was born in <GEO>London</GEO>*

(7) Qui sont ses parents ?
   a. *a child was born to <FN>Basir</FN> and his <Hum>wife</Hum>*
   b. *Prince Michael II was born to an unknown <Hum>surrogate</Hum>*
   c. *twin baby boys born to an <Hum>American surrogate mother</Hum>*
   d. *Jones, who was born to <Hum>Welsh parents</Hum> in Ghana*
   e. *She was born as daughter of <LN>Matheus Klaas</LN> and <LN>Ida Lysse</LN>*

(8) Quel est son nom de naissance ?
   a. *1934 born in Poland as <LN>Manya Sklodowska</LN>*
   b. *was born as <LN>Francisco Gutierrez</LN> in Havana*
   c. *Martika was born as <LN>Marta Marrero</LN> in Whittier*

(9) Quel est son nom actuel ?
   a. *Born and raised in Queens, <SN>Woodbridge</SN> had attended the School of Visual Arts*
   b. *<LN>Lycia Danielle Trouton</LN>, who was born in Belfast*
   c. *<LN>James Saunders</LN> was born in London in 1925*

Pour analyser les questions des exemples (5) à (9) il faut les transformer dans le langage d'interrogation présenté dans la section 7. Par exemple, pour une question comme *« Who was born in 1925 ? »*, la requête doit être formulée comme suit :

```
[semantic=WHO] IN ([semantic=BORN] CONTAINING ([semantic=WHEN]
                        CONTAINING 1925))
```

6. Évaluation

Pour évaluer la qualité de notre système d'extraction, nous l'avons appliqué sur un corpus composé de 23 345 notices biographiques issues du site `biography.com`. Ce corpus compte 1 153 131 mots répartis en 68 646 phrases, ce qui représente une moyenne d'environ 47 mots et 3 phrases par notice.

6.1. *Résolution* ad hoc *des coréférences pronominales*

Le système `RELAX` présenté ici, n'a pas pour vocation de résoudre les coréférences qui apparaissent dans les textes. Il nous faudra toutefois traiter ce problème à moyen terme afin de rendre notre système viable dans un contexte applicatif réel. Pour cette première évaluation, nous avons donc mis au point un traitement *ad hoc*, qui nous a permis de réduire de manière conséquente le problème du silence causé par les nombreuses reprises anaphoriques.

Nous avons profité du fait que chaque notice biographique ne traite en général que d'une unique personne et qu'il n'y a, par conséquent, que très peu d'ambiguïtés pour les pronoms personnels. Pour chacune des biographies, nous avons ainsi compté le nombre d'occurrences des formes *« he »* et *« she »* et nous avons annoté la forme apparaissant le plus souvent dans chaque texte par le nom de la personne décrite dans celui-ci (qui correspond au titre de la page). De plus, lorsque la forme *« he »* est la forme la plus fréquente nous annotons de la même manière les pronoms *« him »* et le déterminant possessif *« his »*. Dans le cas où la forme *« she »* apparaît plus fréquemment nous annotons également les occurrences de la forme *« her »*, qui peut recevoir les deux analyses *pronom personnel* ou *déterminant possessif*. Le moteur d'analyse `Outilex` acceptant en entrée un texte étiqueté sous la forme d'un automate acyclique, nous représentons naturellement ce type d'ambiguïtés par deux transitions parallèles.

Le tableau 4 présente le nombre total d'occurrences de chaque pronom personnel, ainsi que le nombre d'occurrences qui ont été annotées. Pour des raisons de complétude, nous donnons également le nombre d'apparition des pronoms *« they »* et *« their »* que nous ne traitons pas. Comme l'indique le tableau, notre méthode nous a permis d'annoter environ 95 % de la totalité des occurrences des pronoms personnels apparaissant dans notre corpus de test.

Formes	*he*	*she*	*his*	*her*	*him*	*they*	*their*	**Total**
Occurrences totales	35 225	4 597	17 412	3 096	1 682	552	744	64 634
Occurrences annotées	35 155	4 545	17 231	2 904	1 613	0	0	61 448

Tableau 4. *Résolution* ad hoc *des coréférences pronominales*

Nous avons manuellement vérifié ces annotations pour 50 biographies sélectionnées au hasard. Nous avons calculé une précision de 100 % et un rappel de 98 % sur cet échantillon.

Formes	*he*	*she*	*his*	*her*	*him*	**Total**
Total	68	3	24	3	1	99
Correctement annotées	68	2	24	2	1	97
Erreur (ou absence) d'annotation	0	1	0	1	0	2
Précision	100 %	100 %	100 %	100 %	100 %	100 %
Rappel	100 %	66,6 %	100 %	66,6 %	100 %	98 %

Tableau 5. *Évaluation des résolutions d'anaphores sur 50 biographies*

Ainsi, cette méthode *ad hoc* donne de très bons résultats pour ce corpus en particulier. Pour le traitement de corpus plus généraux, il sera néanmoins nécessaire d'intégrer un module plus robuste, capable de résoudre également les reprises anaphoriques nominales.

6.2. *Protocole d'évaluation*

Les principaux points du protocole d'évaluation sont les suivants :

– nous considérons chaque relation entre un prédicat biographique et un de ses arguments comme unité pour le calcul du rappel et de la précision, et non l'événement biographique dans son ensemble. Ainsi, pour le texte suivant [12] :

```
 <RELATION type="job obtainment"> <Person> he=Croke </Person>
became Roman catholic <JD> bishop </JD> of <ORG> Auckland
</ORG> </RELATION>, New Zealand (1870)
```

nous comptons quatre paires (prédicat, argument) au total dans cet énoncé : le prédicat désigne l'événement d'obtention d'un poste et ses quatre arguments désignent respectivement la personne (*« Croke »*), l'intitulé du poste obtenu (*« bishop »*), le lieu (*« Auckland, New Zealand »*) et la date (*« 1870 »*). Par conséquent, nous comptabilisons deux annotations correctes, une erreur d'annotation et deux absences d'annotation, ce qui correspond à une précision de 66,6 % et un rappel de 50 % pour cet exemple ;

12. Tous les exemples sont tirés des résultats obtenus sur notre corpus d'évaluation.

– lorsque un segment décrivant un argument d'un événement biographique n'est annoté que partiellement par notre système d'extraction, nous considérons cela comme une erreur de rappel mais pas de précision. Par exemple, considérons l'annotation suivante :

```
 <RELATION type="born"> <RELATION type="job"> <Person> Abegg
</Person>, <JD> Chemist </JD> </RELATION> , was born in
<GEO> Gda-sk, N Poland </GEO> </RELATION> (formerly Danzig,
Germany).
```

On observe que l'argument locatif *« Gda-sk, N Poland (formerly Danzig, Germany) »* n'a été que partiellement reconnu. Pour cet exemple, nous calculons donc une précision de 100 % et un rappel de 75 %. Nous comptons ici quatre paires *prédicat argument* car l'énoncé contient deux informations biographiques identifiées par les prédicats *« born »* et *« job »* et le segment *« Abegg »* est simultanément argument de ces deux prédicats ;

– nous ne comptabilisons pas comme erreurs de rappel les énoncés non annotés décrivant des événements biographiques qui ne font partie des douze types de relations que nous avons étudiées dans le cadre de ce travail tels que, par exemple, l'acquisition d'un prix Nobel, la réalisation d'une œuvre ou d'un exploit sportif, etc.

6.3. *Résultats de l'évaluation*

Nous avons extrait de ce corpus annoté 150 biographies au hasard que nous avons vérifiées manuellement. Afin d'évaluer également l'apport du *Bootstrapping* pour la qualité d'extraction, nous avons appliqué notre grammaire deux fois sur ce même corpus d'évaluation : une première fois en utilisant nos lexiques terminologiques d'origine, une seconde fois en utilisant nos lexiques dans leur état actuel, c'est-à-dire, les mêmes lexiques enrichis à l'aide de nos méthodes de *Bootstrapping*. Concrètement, nous avons fait varier pour ces deux annotations les lexiques décrivant les trois classes d'objets suivantes : ⟨Compagnie⟩ (286 453 instances avant *Bootstrapping*, 516 028 instances après *Bootstrapping*), ⟨Secteur d'activité⟩ (16 591 instances avant, 38 573 après *Bootstrapping*) et ⟨Profession⟩ (20 751 instances avant, 44 896 après). Nous donnons dans le tableau 6, les taux de rappel et de précision obtenus avec ces deux types de ressources.

	Avant	**Après**
Précision	96,2 %	97,1 %
Rappel	72,4 %	93,3 %

Tableau 6. *Apport du Bootstrapping pour l'extraction d'informations*

Le tableau 7 présente plus en détail, pour chacun des douze types d'événements, le nombre de relations que nous avons recensées dans le textes ainsi que le nombre d'arguments qui ont été correctement ou non annotés par notre système.

Type d'événement	Nombre d'événements	Nombre d'arguments	Annotations correctes	Erreurs d'annotation	Précision	Rappel
Naissance	291	583	571	0	100 %	97,9 %
Enfance	1	2	2	0	100 %	100 %
Formation	55	222	213	9	95,8 %	89,2 %
Mariage	12	28	27	0	100 %	96,4 %
Divorce	2	2	2	0	100 %	100 %
Décès	122	254	236	2	99,2 %	92,1 %
Obtention d'un emploi	68	287	273	14	94,9 %	74,2 %
Occupation d'un poste	147	448	431	26	94,0 %	78,6 %
Licenciement	4	9	9	0	100 %	100 %
Succession	2	7	5	0	100 %	71,4 %
Démission	3	6	5	0	100 %	83,3 %
Départ en retraite	3	6	6	0	100 %	100 %
Total	710	1 854	1 780	51	97,1 %	89,8 %

Tableau 7. *Résultats détaillés pour chaque type d'événement biographique*

Ces résultats sont très satisfaisants. Les principales causes d'erreurs de rappel sont dues à la couverture de nos dictionnaires et, plus rarement, à la présence de certaines structures syntaxiques plus ou moins complexes non décrites dans nos grammaires :

```
Van Maanel forced him=Falck to resign (1823).
```

Notons que le fait d'imposer que chaque argument soit recensé dans nos dictionnaires comme instance d'une classe d'objet particulière pour être correctement identifié, nous permet d'obtenir un taux de précision proche de 100 %. D'ailleurs, les seules causes d'erreurs de précision que nous avons recensées, sont dues à la présence de formes ambiguës qui apparaissent dans nos lexiques comme des instances de plusieurs classes d'objets différentes. Par exemple, dans la phrase suivante, la forme *« Sydney »* a été incorrectement identifiée comme un argument locatif :

```
 <RELATION type="death"> <Person> Gypsy </Person> died of a
broken heart after <GEO> Sydney </GEO> </RELATION> Chaplin
[...]
```

Nous pensons pouvoir réduire ce type d'erreurs en assignant dans nos grammaires des priorités à certaines analyses en fonction de l'événement biographique qui y est décrit.

Nous devons toutefois relativiser la signification de ces chiffres en précisant que le corpus d'évaluation que nous avons choisi est très particulier et n'est pas représentatif d'un corpus général de la langue anglaise. En effet, il est constitué exclusivement de phrases très courtes et certaines structures syntaxiques apparaissent de manière récurrente tout au long du texte, ce qui facilite beaucoup l'analyse automatique. On peut donc prévoir des résultats moins bons, notamment au niveau du rappel, pour l'analyse de textes qui n'auraient pas ces spécificités. Du fait de l'absence de corpus annotés manuellement par des relations biographiques, il nous est très difficile de procéder à une telle évaluation. Nous avons d'ailleurs choisi le corpus `biography.com` parce qu'il contient de façon très dense l'information biographique, ce qui nous a facilité amplement le tâche du calcul du rappel.

Malgré ces quelques réserves, nous considérons ces résultats très satisfaisants et nous pensons qu'ils tendent à prouver que notre système a un potentiel applicatif réel.

7. LaolaWeb : système de réponses à des questions biographiques

Nous présentons dans cette section le moteur de recherche sémantique `LoalaWeb` (Schömmer, 2007), qui constitue une première application concrète de nos travaux sur l'extraction de l'information biographique. Nous décrivons les principales fonctionnalités du système et nous montrons comment il bénéficie de notre système d'annotations des relations biographiques par grammaires locales.

Le système de questions-réponses `LaolaWeb` a été initialement conçu dans le cadre du projet `LAOLA` (*Linguistic Analysis Of Large corporA*) et un prototype est actuellement en ligne à l'adresse `http://schoemmer.de:8080/laolaWeb/`. L'interface Web permet à l'utilisateur d'interroger en ligne le système sur des questions biographiques concernant les personnes mentionnées dans les textes prétraités. La base de connaissances en informations biographiques de la version actuelle a été construite à partir de 25 000 biographies qui ont été récupérées depuis la version anglaise du site Wikipedia. La phase d'extraction des informations biographiques dans ces textes a été essentiellement effectuée par une application en cascade de nos graphes d'annotations, qui permet d'assigner des étiquettes hiérarchiques aux différents segments reconnus par la grammaire. Les figures 8 et 9 montrent des exemples de telles annotations. Les étiquettes les plus pertinentes pour les systèmes de questions-réponses sont celles qui définissent des traits sémantiques (tels `WHO`, `WHOM`, `WHEN`, `WHERE`, etc.).

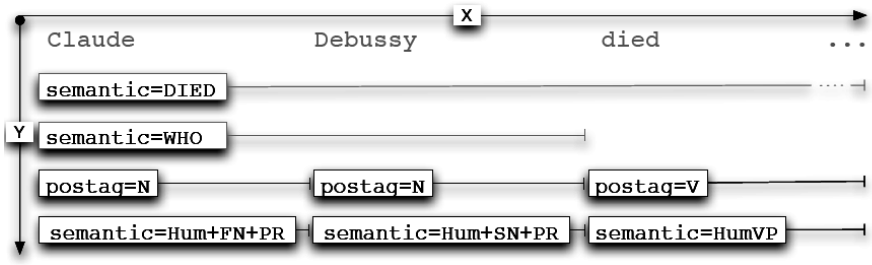

Figure 8. *Annotations horizontales et verticales*

Pour l'instant, seules les relations suivantes sont considérées :

- *X married Y*
- *X divorced from Y*
- *X was born in Y*
- *X died on Y*
- *X graduated from Y*
- *X works as Y*

Le système comprend une syntaxe spéciale pour formuler les requêtes de la forme d'une algèbre sur les segments annotés. Par exemple, pour une question comme *« Quand Claude Debussy est-il mort ? »*, la requête doit être formulée comme suit :

```
[semantic=WHEN] IN ([semantic=DIED] CONTAINING Claude Debussy)
```

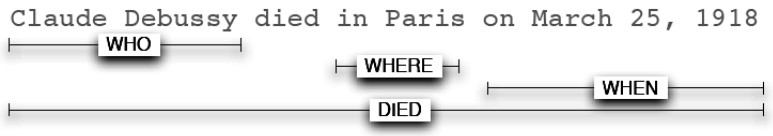

Figure 9. *Annotations sémantiques*

Ainsi, le segment annoté par l'étiquette `WHEN` présente comme réponse :

« on March 25, 1918 ».

Le langage de requête implémente les relations binaires `CONTAINING` et `IN` sur des segments textuels. L'opérateur `IN` permet de contraindre que le segment décrit dans son opérande gauche soit une partie de celui décrit dans son opérande droit. L'opérateur `CONTAINING` permet de spécifier des termes qui doivent apparaître dans les segments retournés de manière à préciser les résultats obtenus de différentes façons. `LaolaWeb` est toujours en cours de développement et il est prévu d'implémenter un système de conversion des requêtes en langue naturelle vers ce métalangage. Pour que le système soit vraiment opérationnel, il sera nécessaire d'y intégrer un système de résolution des anaphores plus avancé que celui présenté précédemment.

8. Conclusion et perspectives

L'extraction automatique de l'information biographique est une tâche très complexe qui dépend fortement des connaissances acquises sur le domaine à analyser. Nous avons montré dans cet article l'intérêt des grammaires locales comme formalisme de représentation des variabilités syntaxiques existantes entre les instances des classes d'objets et nous avons présenté notre processus itératif basé sur l'utilisation de patrons lexicosyntaxique et des méthodes de Bootstrapping utiles pour la reconnaissance et l'extraction d'informations dans les textes ainsi que pour l'acquisition automatique de terminologie nouvelle.

Une application directe de ces travaux a été implémentée pour servir un système de question-réponse (Schömmer, 2007). Les résultats de cette application ainsi que nos résultats d'évaluation nous incitent à poursuivre cette étude en l'étendant à d'autres événements biographiques non encore étudiés ainsi qu'à de nouveaux genres textuels pour proposer un système `RELAX` capable de générer automatiquement la biographie générale d'une personne en extrayant l'information biographique pertinente dispersée dans une multitude de documents.

9. Bibliographie

Agichtein E., Gravano L., « Snowball : Extracting Relations from Large Plain-Text Collections », *Proceedings of the Fifth ACM International Conference on Digital Libraries*, San Antonio, Texas, USA, p. 85-94, 2000.

Blanc O., Constant M., Laporte É., « Outilex, plate-forme logicielle de traitement de textes écrits », *in* P. Mertens, C. Fairon, A. Dister, P. Watrin (eds), *Verbum ex machina. Actes de la 13e conférence sur le Traitement automatique des langues naturelles (Cahiers du Cental 2)*, Presses universitaires de Louvain, Louvain-la-Neuve, Belgique, p. 83-92, 2006.

Buvet P.-A., Foucou P.-Y., « Classes d'objets et recherche sur le web », *Lingvisticæ Investigationes*, vol. 23, p. 219-228, 2001.

Courtois B., « Dictionnaires électroniques DELAF anglais et français », *in* C. L. et Éric Laporte et Mireille Piot et Max Silberztein (ed.), *Lexique, syntaxe et lexique-grammaire ; syntax, lexis & lexicon-grammar*, John Benjamins, Amsterdam/Philadelphia, p. 113-123, 2004.

Courtois B., Silberztein M., « Dictionnaires électroniques du français », *Langues française*, vol. 87, p. 11-22, 1990.

Davis I., Galbraith D., « BIO : A vocubulary for biographical information », 2004, `http://purl.org/vocab/bio/`.

Duboué P., McKeown K., Hatzivassiloglou V., « ProGenIE : Biographical descriptions for Intelligence Analysis », *Proceedings of the NSF/NIJ Symposium on Intelligence and Security Informatics*, vol. 2665 of *Lecture Notes in Computer Science*, Springer, Tucson, Arizona, USA, p. 343-345, juin, 2003.

Geierhos M., Grammatik der Menschenbezeichner in biographischen Kontexten, Rapport Technique, Centrum für Informations- und Sprachverarbeitung (CIS), Ludwig-Maximilians-Universität, Munich, Allemagne, 2007.

Gross G., « Classes d'objets et description des verbes », *Langages*, 1994.

Gross M., « Local grammars and their representation by finite automata », *in* M. Hoey (ed.), *Data, Description, Discourse : Papers on the English Language in honour of John McH Sinclair*, Harper-Collins, London, p. 26-38, 1993.

Gross M., « The Construction of Local Grammars », *in* E. Roche, Y. Schabès (eds), *Finite-State Language Processing*, MIT Press, Cambridge, Massachusetts, p. 329-354, 1997.

Gross M., « A bootstrap method for constructing local grammars », *Contemporary Mathematics : Proceedings of the Symposium, University of Belgrad*, Belgrad, Serbie, p. 229-250, 1999a.

Gross M., « Lemmatization of compound tenses in English », *Lingvisticæ Investigationes*, vol. 22, n° 2, p. 71-122, 1999b.

Harris Z. S., *Mathematical Structures of Language*, John Wiley & Sons, New York, 1968.

Harris Z. S., *Language and Information*, Columbia University Press, New York, 1988.

Harris Z. S., *A theory of language and information : A mathematical approach*, Clarendon Press - Oxford, New York, 1991.

Hunston S., Sinclair J., « A local grammar of evaluation », *in* S. Hunston, G. Thompson (eds), *Evaluation in Text : authorial stance and the construction of discourse*, Oxford University Press, Oxford, England, p. 74-101, 2000.

Kanzaki M., « Who's who description vocabulary », 2003–2007, http://www.kanzaki.com/ns/whois.

Kevers L., « L'information biographique : modélisation, extraction et organisation en base de connaissances », *in* P. Mertens, C. Fairon, A. Dister, P. Watrin (eds), *Verbum ex machina. Actes de la 13e conférence sur le Traitement automatique des langues naturelles (Cahiers du Cental 2)*, Presses universitaires de Louvain, Louvain-la-Neuve, Belgique, p. 680-689, 2006.

Le Pesant D., Mathieu-Colas M., « Introduction aux classes d'objets », *Langages*, vol. 131, p. 6-33, 1998.

Mallchok F., Automatic Recognition of Organization Names in English Business News, Thèse de Doctorat, Ludwig-Maximilians-Universität, Munich, Allemagne, 2004.

Meyer I., « Extracting knowledge-rich contexts for terminography », *in* D. Bourigault, C. Jacquemin, M.-C. L'Homme (eds), *Recent Advances in Computational Terminology*, John Benjamins, Amsterdam, p. 279-302, 2001.

Nakamura T., « Analysing Texts in a Specific Domain with Local Grammars : The Case of Stock Exchange Market Reports », *Linguistic Informatics – State of the Art and the Future*, vol. 1, p. 76-98, 2005.

Paumier S., *Manuel d'utilisation d'Unitex*. 2004, http://wwwigm.univmlv.fr/~unitex/.

Poibeau T., *Extraction automatique d'information, du texte brut au web sémantique*, Lavoisier, 2003.

Sager N., « Sublanguage : Linguistic phenomenon, computational tool », *in* R. Grishman, R. Kittredge (eds), *Analyzing Language in Restricted Domains : Sublanguage Description and Processing*, Lawrence Erlbaum Associates, Hillsdale, NJ, USA, p. 1-18, 1986.

Schiffman B., Mani I., Concepcion K. J., « Producing Biographical Summaries : Combining Linguistic Knowledge with Corpus Statistics », *Proceedings of the* 39^{th} *Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, p. 450-457, 2001.

Schömmer C., *« Grammatikentwicklung im Rahmen lokaler Grammatiken. Eine semantische Suchmaschine für biographische Prädikate »*, Rapport de DEA, Ludwig-Maximilians-Universität, Munich, Allemagne, 2007.

Senellart J., « Locating noun phrases with finite state transducers », *Proceedings of the 17th International Conference on Computational Linguistics*, Montréal, Canada, p. 1212-1219, 1998.

Silberztein M., « Dictionnaire électroniques et analyse automatique de textes - Le système INTEX », 1993, Paris, Masson.

Sparck-Jones K., « What might be in a summary ? », *in* G. Knorz, J. Krause, C. Womser-Hacker (eds), *Information Retrieval '93 : Von der Modellierung zur Anwendung*, Universitätssverlag Konstanz, p. 9-26, 1993.

Tsur O., de Rijke M., Sima'an K., « BioGrapher : Biography Questions as a Restricted Domain Question Answering Task », *Proceedings ACL 2004 Workshop on Question Answering in Restricted Domains*, Barcelone, Espagne, p. 23-30, 2004.

Woods W. A., « Transition network grammars for natural language analysis », *Commun. ACM*, vol. 13, n° 10, p. 591-606, 1970.