Elaboration d'une cascade de transducteurs pour l'extraction des noms de personnes dans les textes

Nathalie Friburger, Denis Maurel

Laboratoire d'Informatique de Tours
64 avenue Jean Portalis, 37000 Tours
{friburger, maurel}@univ-tours.fr

Résumé – Abstract

Cet article décrit une cascade de transducteurs pour l'extraction de noms propres dans des textes. Après une phase de pré-traitement (découpage du texte en phrases, étiquetage à l'aide de dictionnaires), une série de transducteurs sont appliqués les uns après les autres sur le texte et permettent de repérer, dans les contextes gauches et droits des éléments "déclencheurs" qui signalent la présence d'un nom de personne. Une évaluation sur un corpus journalistique (journal *Le Monde*) fait apparaître un taux de précision de 98,7% pour un taux de rappel de 91,9%.

This article describes a finite-state cascade for proper nouns extraction in texts. After a preprocessing (division of the text in sentences, tagging with dictionaries, etc.), a series of finite state transducers cascade is applied one after the other to the text and locate left and right contexts which indicate presence of a person name. An evaluation on a journalistic corpus (*Le Monde*) gives a rate of precision of 98,7% for a rate of recall of 91,9 %.

Mots clefs – keywords

Transducteur, noms propres, extraction de motifs

Transducer, proper nouns, pattern extraction

1 Introduction

Les automates à nombre fini d'états, et tout particulièrement les transducteurs, sont de plus en plus utilisés pour le traitement automatique des langues [Roche, Schabes, 1997]. Dans cet article, nous proposons d'utiliser des transducteurs en cascade pour localiser les noms propres [Coates-Stephens, 1993] dans les textes journalistiques. Ce type de textes a déjà été étudié dans de nombreux travaux depuis le système Frump [Dejong, 1982] jusqu'aux programmes américains Tipster et MUC d'évaluation des systèmes d'extraction d'information, mais le

problème de la détection des noms propres reste largement non résolu. Les informations extraites s'intégrerons dans un travail de classification de textes à partir de noms propres.

Nous décrirons tout d'abord une phase de pré-traitement nécessaire des textes (découpage du texte en phrases, étiquetage à l'aide de dictionnaires), puis nous décrivons en détail les transducteurs utilisés. Enfin, nous présenterons les résultats d'une évaluation sur un corpus d'environ 165 000 mots. Nous discuterons des principales difficultés rencontrées et des problèmes qui restent à résoudre.

2 Pré-traitements des textes

Avant d'appliquer la cascade de transducteurs sur un texte, nous le soumettons à un certain nombre de pré-traitements afin d'améliorer les résultats ultérieurs. Le texte est tout d'abord découpé en phrases [Friburger et al. , 2000] (ce qui permet d'éliminer l'ambiguïté du point de fin de phrase avec les noms propres en contenant), puis étiqueté du point de vue morpho-syntaxique. Nous pourrions utiliser un étiqueteur tel que celui de [Brill, 1992] qui donne des résultats corrects à plus de 95% ou le système Cordial[1] qui offre des résultats encore meilleurs. Nous avons préféré utiliser l'étiquetage par les dictionnaires du système Intex [Silberztein, 1993]. Les mots sont étiquetés par toutes leurs formes présentes dans tous les dictionnaires utilisés sans désambiguation. Parmi les différentes étiquettes d'un même mot, la bonne étiquette est la plupart du temps présente ce qui évite un mauvais étiquetage. Ce mauvais étiquetage peut en effet gêner et diminuer les performances d'un système [Morin, 1999].

Nous utilisons des dictionnaires qui contiennent les mots (lemmes et formes fléchies), ainsi que des informations grammaticales (nom, verbe, etc.) et sémantiques (humain, concret, toponymes, prénoms, sigles, etc.).

- Delas : dictionnaire des mots simples et de leurs formes fléchies [Courtois, Silberztein, 1990]
- Prolintex : dictionnaire de toponymes réalisés dans le cadre du projet Prolex [Piton, Maurel, 1997]
- Prenom-prolex : Dictionnaire de prénoms (plus de 6500 entrées) et Sigles-prolex : Dictionnaire de sigles avec leurs extensions (environ 3300 entrées)[2]
- Dictionnaire des professions[3]

L'avantage de ces dictionnaires est double :

- Chaque mot est donné avec sa forme lemmatisée, ce qui permet de ne pas avoir à décrire toutes les flexions des mots dans les transducteurs pour les détecter.

---

[1] Proposé par Synapse Développement

[2] Ces dictionnaires ont été crée au Laboratoire d'Informatique de Tours dans le cadre du projet Prolex de dictionnaire des noms propres.

[3] Dictionnaire élaboré par Cédric Fairon, LADL [Fairon, 2000].

- Les dictionnaires utilisés contiennent des informations syntaxiques (Nom, verbe, pronom, etc.) et sémantiques (humain, prénom, toponyme, etc.) qui peuvent aider à la localisation des motifs de noms propres.

3 Une cascade de transducteurs pour l'extraction des noms de personnes

3.1 Les transducteurs

Les transducteurs sont des automates qui possèdent un alphabet d'entrée et un alphabet de sortie : c'est cette propriété que nous pouvons utiliser pour extraire des motifs et les catégoriser. L'alphabet d'entrée contient les motifs que l'on souhaite repérer dans les textes tandis que l'alphabet de sortie contient, dans notre cas, des informations balisés dans un langage inspiré de XML. Les motifs que nous recherchons sont les noms propres ainsi qu'une partie de leurs contextes lorsqu'ils sont exploitables et repérables. Voici un exemple de nom de personne trouvé dans un texte et balisé par le transducteur qui l'a repéré :

Le Juge Renaud Van Ruymbeke → *<profession> juge <\profession> <person> <prenom> Renaud <\prenom> <nom>Van Ruymbeke<\nom> <\person>.*

Le système Intex nous permet d'exploiter les transducteurs sur des textes. Nous avons réalisé un programme qui complète les possibilités d'Intex et permet de générer une cascade de transducteurs.

Dans de nombreuses applications, les transducteurs sont utilisés en cascade. Cette technique peut être utilisée pour réaliser l'étiquetage syntaxique d'un texte (comme le font Xerox ou le système Fastus [Hobbs et al., 1996]). La cascade est basée sur une idée simple : passer les transducteurs sur le texte dans un ordre précis pour le transformer ou en extraire des motifs. Chaque motif que nous découvrons est remplacé dans le texte par une étiquette qui nous permet de le retrouver dans un index. Nous éliminons les motifs découverts du texte pour éviter qu'un transducteur passé ultérieurement ne les extraie à nouveau.

3.2 Etude des formes présentes dans les contextes droit et gauche des noms de personnes

Avant de créer la cascade, nous avons fait un inventaire des formes existant en contexte droit et gauche des noms de personnes dans les articles de journaux. En effet, ce sont les contextes qui aident à repérer les noms propres et en particulier une grande majorité de noms de personnes. Nous avons remarqué que le contexte gauche permet de détecter plus de 90% des noms de personnes dans les textes de style journalistique : ceci est certainement dû à des impératifs stylistiques propres à ce type de textes qui se veulent la plupart du temps objectif et qui doivent décrire au mieux les faits. Une étude sur un texte tirés du journal *Le Monde* d'environ 165000 mots nous a permis de déterminer les catégories de contextes les plus fréquents.

- Cas 1 : 25,9% des noms de personne sont précédés d'un contexte contenant un titre ou un nom de profession suivis d'un prénom et d'un patronyme. Ex : ***M. Jean-Pierre Soisson*** *déclarait : ... Ce regain de violence a coïncidé avec la visite officielle du* ***président péruvien Alberto Fujimori*** *en Equateur, qui a pris fin samedi.*

- Cas 2 : 19,1% des noms de personnes sont précédés d'un contexte déclencheur contenant un titre ou un nom de profession suivi, soit d'un patronyme seul, soit d'un prénom inconnu (i.e. absent de notre dictionnaire des prénoms) et d'un patronyme. Ex : *... et qui qualifiait la démission du* ***président Chadli*** *d'" événement important et lourd de conséquences ", ...*

- Cas 3 : C'est le cas le plus fréquent. 43,4% des noms de personnes n'ont pas de contextes descriptibles mais sont composés du prénom de la personne (connu de notre dictionnaire) et suivi du nom de la personne. Ex : ***Pierre Bourdieu*** *est sans conteste l'une des figures majeures de la sociologie contemporaine.*

- Cas 4 : 5,2% des formes sont repérables grâce à la présence d'un verbe utilisé pour désigner une action mettant en jeu une personne (*dire, expliquer, etc*.) comme dans "***Wieviorka*** *est décédé, le 28 décembre, à Paris***"** ou "***Jelev*** *a dit*" ou par la présence d'un titre ou d'une profession en apposition (contexte droit) ex : *"...* ***Jospin, premier Ministre*** *..."*. Cependant des verbe comme *dire, expliquer* … peuvent être employés avec un sujet non humain, des indices tels que ceux ci sont à utiliser avec une très grande prudence.

- Cas 5 : Les 6,4% de noms de personnes restants n'ont aucun contextes, même complexes qui puissent les distinguer à coup sûr d'autres noms propres. Ces noms de personnes sans contextes (pas même un prénom qui pourrait les distinguer d'autres noms propres) sont principalement des noms de personnes très connues pour lesquels l'auteur du texte estime qu'il n'est pas nécessaire de préciser le prénom ni le titre ou la profession, ex : *Picasso n'est pas le premier à passer à la postérité commerciale.* Cependant nous avons remarqué que 49% de ces noms de personnes restants sont détectables en réalisant une seconde passe dans laquelle on recherche les patronymes qui ont été découvert un autre endroit du texte par un des transducteurs. Ce qui réduit à 3,3% le nombre de formes indétectables. Ce pourcentage peut être encore réduit en créant un dictionnaire des noms de personnes célèbres.

La même étude a été menée sur des articles du journal *Ouest France* (67 000 mots environ). Les résultats sont présentés dans le Tableau 1.

Cas 1 : 17,1%	Cas 2 : 16,3%	Cas 3 : 59 %	Cas 4 : 2,2%	cas 5 : 5,4%

Tableau 1 : Proportion des différents cas sur le journal *Ouest France*

On remarque que le cas 3 est le plus commun et est bien plus fréquent dans *Ouest France* que dans le journal *Le Monde*. Par contre, on peut remarquer que le cas 1 (nom précédé d'un titre ou d'une profession) chute de 25,9% dans *Le Monde* à 17,1% dans *Ouest France* : ceci est certainement dû à un souci de rigueur du journal *Le Monde*.

3.3 Etude combinatoire des différentes formes de noms de personnes

Nous avons aussi étudié quelles formes pouvaient prendre les noms de personnes. On trouve en majorité des prénoms suivis d'un patronyme ou des noms seuls. Comme l'ont déjà remarqué aussi [Kim, Evens, 1996], l'auteur d'un article de journal donne en général une première fois la forme complète du nom de personne, puis des formes abrégées ; c'est pourquoi la majorité des noms de personnes trouvés le sont assez souvent avec leur prénom et leur nom.

Les formes de prénoms à reconnaître sont les suivantes :

- prénoms simples, ex : *Jean*,
- prénoms composés, ex : *Jean-Pierre, Charles Edouard*, ou en partie abrégés ex : *Pierre-J.*
- prénoms abrégés simples, ex : *J.* pour *Jean, Th.* pour *Thierry*,
- prénoms composés abrégés, ex : *J.P., J.-P., J-P, J-P.*
- prénoms composés en partie inconnus : prénom composé dont une des parties est dans le dictionnaire des prénoms et l'autre est inconnue.

La reconnaissance des prénoms se fait sur la base d'indices morphologiques ainsi que sur le dictionnaire des prénoms. Les prénoms totalement inconnus du dictionnaire ne sont pas reconnus, ils font alors partie intégrante du nom de personne. Nous avons fait l'hypothèse que les personnes sont plus souvent citées par les journalistes en donnant d'abord le prénom puis le nom. Cette règle n'est évidemment pas absolue, puisque, si dans le corpus étudié du journal *Le Monde* aucun forme nom-prénom n'a été détectée, nous avons observé l'ordre nom-prénom dans 3% des noms de personnes des articles étudiés dans le journal *Ouest France.*

Les différentes formes de patronymes sont les suivantes :
- Patronymes "composés" (surtout des noms d'origine étrangère),
  ex : *Mac Donald, Mac Donnell-Douglas, O'Ryan, L'Huillier, Le Falch'un, von Bulow, El Amra, Da Silva ...*
- Patronymes "simples" (composés d'un ou plusieurs mots commençant par une majuscule),
  ex : *Dupont, Durand-Pérec*
- Patronymes français à particules, ex : *Dupont de Nemours, de Neuville, de la Fontaine*

Nous avons distingué les patronymes français à particules des patronymes composés car la particule française (*de, du*) est très ambiguë dans les textes avec la préposition *de* ce qui n'est pas le cas pour les particules étrangères. Nous devrons donc tenir compte de ces différences dans l'ordre de passage des transducteurs.

Les contextes déclencheurs qui décrivent le majorité des contextes gauches sont simplement les civilités (ex: *Mme, Monsieur, etc.*), les titres de toutes sortes : politiques (ex : *ministre, député, etc*.), militaires (ex : *général, lieutenant, etc*.) religieux (ex : *cardinal, évêque, etc*.), administratifs (ex *: inspecteur, agent, etc*.) … ainsi que les noms de professions (ex : le juge, l'architecte, etc.). Les noms de professions sont les termes déclencheurs les moins fréquents. Le dictionnaire des toponymes permet de repérer les adjectifs de nationalités dans des expressions telles que "*le président américain Clinton*", "*l'allemand Helmut Kohl* ".

3.4 Description de la cascade de transducteurs

Ordre de passage des transducteurs	Recherche de contextes gauches déclencheurs (oui/non)	Prénoms	Patronymes	Exemples
1	Oui	Prénoms composés et abrégés Prénoms simples	Patronymes composés Patronymes à particules	Ex : *le président Richard von Weizsõcker*
2	Oui	Prénoms composés et abrégés	Patronymes simples Patronymes à particules	Ex : *M. J.-P. de Fonsac*
3	Oui	Prénoms composés en partie inconnus	Patronymes composés	Pas d'exemple
4	Oui	Prénoms composés en partie inconnus	Patronymes simples	Ex : *Roger-Pol Droit (Pol n'était pas dans notre dictionnaire de prénoms)*
5	Oui	Prénoms simples	Patronymes simples	Ex : *M. Guy Fleury*
6	Oui		Patronymes composés	Ex : *M. Strauss-Kahn*
7	Oui sans les professions et les titres.		Patronymes à particules	Ex : *M. de Neuville*
8[4]	Oui	<CNP>[5] <CNP>	<CNP> - <CNP>[6] <CNP> <CNP>	Ex : *M. Amnon Lipkin-Shahak*
9	Oui	<CNP>	<CNP>	Ex : *le général Veljko Kadijevic*
10	Oui		Patronymes simples	Ex : *Mme Bouchardeau*
11	Interdit les déterminants avant le nom de personne	Prénoms composés et abrégés Prénoms simples	Patronymes composés	Ex *Philippine Leroy-Beaulieu* *Ex : Maria da Graca Meneghel*
12	Interdit les déterminants avant le nom de personne	Prénoms composés et abrégés Prénoms simples	Patronymes simples	Ex : *P. Bourdieu*
13	Interdit les déterminants avant le nom de personne	Prénoms composés en partie inconnus	Patronymes composés	Pas d'exemple
14	Interdit les déterminants avant le nom de personne	Prénoms composés en partie inconnus	Patronymes simples	Pas d'exemple

Tableau 2 : Description d'une partie des transducteurs reconnaissant les noms de personnes (cas 1, 2 et 3)

[4] Les transducteurs 8 et 9 permettent de reconnaître des formes composées de mots commençant par une majuscule sans prénom connu et précédées de déclencheurs de la présence de noms de personne.

[5] <CNP> est l'étiquette qui désigne un candidat nom propre, i.e. un mot qui commence par une lettre majuscule.

[6] <CNP>-<CNP> désigne 2 candidats noms propres séparés d'un tiret.

D'après les différentes constatations faites lors de notre étude des formes des noms de personnes et de leur contexte, nous avons défini une cascade de transducteurs dans laquelle nous prenons en compte les impératifs liés aux transducteurs eux mêmes.

Nous avons donné priorité aux motifs les plus longs afin de repérer les noms entiers. Par exemple, si nous passons un transducteur qui reconnaît *M.* suivi d'un mot commençant par une majuscule avant le transducteur qui reconnaît *M*. suivi d'un prénom puis d'un nom, et que l'on a un texte contenant le motif *M. Jean Dupont*, on découvre le motif *<person> <nom> Jean Dupont </nom> </person>* au lieu du motif *<person> <prénom> Jean <prénom> <nom> Dupont </nom> </person>*.

Le Tableau 2 décrit la partie la plus importante de la cascade de transducteurs : c'est-à-dire la description des cas 1, 2 et 3 présentés dans la section 3.2. Chaque transducteur est numéroté dans son ordre de passage.

Exemple de lecture du tableau :

Le transducteur 1 reconnaît des noms de personnes composés par:

- un prénom simple, composé ou abrégé,
- puis un patronyme composé ou à particule (ex : *0'Reilly, de La Fontaine*),
- et précédé d'un mot déclencheur.

Avant de passer les transducteurs qui repèrent les noms de personnes sans mots déclencheurs (juste avec les prénoms), il faut passer les transducteurs qui détectent les noms d'association afin d'éviter des problèmes tels que ceux présentés dans la partie 4, car les transducteurs qui reconnaissent des noms de personnes sont ambigus avec d'autres transducteurs. Par exemple, le *motif "L'association Hugues Aircraft"* risque d'être découvert comme un nom de personne à cause de la présence de *Hugues* dans le dictionnaire des prénoms : le graphe 14 reconnaît ce motif.

3.5 Evaluation

Voici maintenant un exemple de résultats obtenus sur un extrait d'article du journal *Le Monde* n°1619. Chaque motif trouvé est remplacé par une étiquette qui porte le nom du transducteur qui l'a trouvé et contient la position du motif repéré dans un fichier dans lequel il est indexé.

Le texte initial est le suivant[7] :

> *Invité à un séminaire sur la crise des fusées de 1962 organisé à La Havane,* ***M. Robert McNamara****, qui fut le secrétaire à la défense du* ***président Kennedy****, a estimé de son côté que les deux pays devaient normaliser leurs relations, minées depuis trente ans par " la peur et l'hostilité ".{S} Au cours de ce séminaire,* ***le président cubain Fidel Castro*** *a révélé que l'URSS avait déployé en 1962 trente-six ogives nucléaires à Cuba, dont neuf avaient été installées sur des missiles.{S}*

Les amorces de noms propres sont ensuite repérées par la cascade de transducteurs. Le texte devient alors :

---

[7] Les symboles {S}, présents dans le texte, signalent les marques de séparation de phrases

> *Invité à un séminaire sur la crise des fusées de 1962 organisé à La Havane,* ***<$titCiv:1358$> Robert McNamara****, qui fut le secrétaire à la défense du* ***<$titPolit:7035$> Kennedy****, a estimé de son côté que les deux pays devaient normaliser leurs relations, minées depuis trente ans par " la peur et l'hostilité ".{S} Au cours de ce séminaire,* ***<$titPolit:1375$> Fidel Castro*** *a révélé que l'URSS avait déployé en 1962 trente-six ogives nucléaires à Cuba, dont neuf avaient été installées sur des missiles.{S}*

On lance ensuite les transducteurs qui repèrent des noms de personnes

> *Invité à un séminaire sur la crise des fusées de 1962 organisé à La Havane,* ***<$person6:12663$>****, qui fut le secrétaire à la défense du* ***<$person11:33051$>****, a estimé de son côté que les deux pays devaient normaliser leurs relations, minées depuis trente ans par " la peur et l'hostilité ".{S} Au cours de ce séminaire,* ***<$person11a:26686$>*** *a révélé que l'URSS avait déployé en 1962 trente-six ogives nucléaires à Cuba, dont neuf avaient été installées sur des missiles.*

Nous obtenons finalement un fichier dans lequel nous avons un index de tous les motifs trouvés :

<Civ:ms> **M.** <\Civ>
<titPolit> **président** <\titPolit>
<titPolit> **président** <nation> **cubain** <\nation> <\titPolit>
…
<Person> <$titCiv:1358$> <prénom> **Robert** <\prénom> <Nom> **McNamara** <\Nom> <\Person>
<Person> <$titPolit:7035$> <Nom> **Kennedy** <\Nom> <\Person>
<Person> <$titPolit:1375$> <Prénom> **Fidel** <\Prénom> <Nom> **Castro** <\Nom> <\Person>

Afin de connaître les résultats obtenus après cette cascade de transducteurs étaient corrects, nous avons vérifié une partie (80 000 mots environ) de notre corpus du journal *Le Monde* (Tableau 3). Nous avons utilisé les mesures classiques de **rappel** et de **précision**. Le rappel est le nombre de noms de personnes correctement trouvés par la cascade de transducteurs divisé par le nombre de noms de personnes réellement présents dans le texte ; le rappel calcule donc la proportion de noms de personnes trouvés. La précision représente le nombre de noms de personnes correctement trouvés divisé par le nombre de noms de personnes correctes et incorrectes trouvés par la cascade.

	Cas 1	**Cas 2**	**Cas 3**	**Cas 4**	**Cas 5**	**Totaux**
Nombre d'occ. de noms de personnes présents	253	187	424	50	64	977
Nombre d'occ. de noms de personnes trouvés	245	187	413	32	32	909
Nombre d'occ. de noms de personnes correctement trouvés	242	186	410	30	31	899
Rappel	95,7%	99,5%	96,7%	60,0%	48,4%	91,9%
Précision	98,8%	99,5%	99,3%	93,8%	96,9%	98,7%

Tableau 3 : Résultats obtenus sur un extrait du journal *Le Monde*

Nous avons dénombré 977 noms de personnes dans ce corpus de 80 000 mots. Dans le tableau 3, nous indiquons le nombre d'occurrences de noms de personnes dans cette portion de textes

et leur nombre de noms de personnes trouvés par notre système. Puis nous indiquons les pourcentages de rappel et de précision pour chaque cas.

Les résultats obtenus sur les trois premiers cas sont très bons ; nous obtenons 97% de rappel et 99,2% de précision sur les noms de personnes qui sont précédés d'un déclencheur et/ou d'un prénom. Nous remarquons que le rappel est nettement moins bon dans les cas 4 et 5. Comme nous l'avons expliqué en 3.2, le cas 4 représente les noms de personnes que l'on peut détecter par des indices parfois trop ambigus : on ne peut donc pas tous les repérer. Le cas 5 est constitué de noms de personnes sans aucun indice : seul le sens de la phrase ou la connaissance du monde d'un lecteur humain permet de déterminer qu'ils sont un nom de personne ; 48,4% des noms de personnes du cas 5 peuvent être trouvés, grâce à leur présence sous une forme détectable ailleurs dans le texte.

Les résultats obtenus lors de la recherche des noms de personnes pourront certainement être améliorés lors de la recherche des autres noms propres.

4 Problèmes et solutions

Voici les problèmes principaux que nous avons repérés et essayés d'éviter au cours de ce travail :

- Ex : *L'orchestre a joué le Carmen de Bizet* : Carmen est un prénom connu du dictionnaire et si, dans notre cas, on n'interdit pas les déterminants, on obtient l'interprétation erronée suivante : *le <person> <prenom> Carmen <\prenom> <nom> de Bizet <\nom> <\person>*. Nous observons le même problème sur le motif *"la France de Vichy "*. Nous avons donc interdit la présence de déterminants ou de noms communs devant un prénom ou un nom, ce qui évite ces erreurs. Un travail sur les noms propres déterminés a été réalisé par [Garric, Maurel, 2000]

- Dans les expressions "*la Duchesse de Windsor*", "*le maire de Paris*", *de Paris* et *de Windsor* ne sont pas des noms de personnes à particules. Nous interdisons donc les noms de personnes à particules trouvés derrière un titre ou un nom de profession car ce sont rarement des noms de personnes. Par contre *M. de Neuville* n'est certainement pas ambigu.
- On constate également qu'avec nos outils, la phrase :"*Bull a négocié un ensemble de crédits bilatéraux avec des banques étrangères (l'allemande Commerzbank, la japonaise Tokai, ...)"* donne comme résultat *<person> Commerzbank </person> , <person> Tokai </Person>*.

5 Conclusion

Le principe de la cascade de transducteurs est assez simple et efficace ; par contre, la description des motifs à trouver s'avère fastidieuse si on veut obtenir le meilleur résultat possible. Les combinaisons et interactions possibles sont complexes. Mais les résultats obtenus sont prometteurs. Les autres noms propres (toponymes, noms d'organisations) sont plus difficiles à repérer car leurs contextes sont beaucoup plus variés.

Les motifs découverts peuvent, au delà du sujet de l'extraction de motifs et de la classification de textes, servir dans de nombreux domaines. On peut ainsi imaginer la création d'un système

d'écriture de fichiers XML semi-automatisé ou encore prévoir un enrichissement semi-automatique de dictionnaires.

Références

Brill E. (1992), A Simple Rule-Based Part of Speech Tagger, *Proceedings of the third conference on Applied Natural Languages Processing,* Trento, Italy, ACL, pp. 152-155.

Coates-Stephens, S. (1993), The Analysis and Acquisition of Proper Names for the Understanding of Free Text, *Computers and the Humanities,* 26 (5-6), pp. 441-456.

Courtois B., Silberztein M. (1990), *Dictionnaire électronique des mots simples du français*, Paris, Larousse.

Dejong G. (1982), An Overview of the frump System, dans *W.B. Lehnert et M.H. Ringle éd., Stratégies for Natural Language Processing*, ErlBaum, pp. 149-176.

Fairon C. (2000), *Structures non-connexes. Grammaire des incises en français : description linguistique et outils informatiques*, Thèse de doctorat en informatique, Université Paris 7.

Friburger N., Dister A., Maurel D. (2000), *Améliorer le découpage des phrases sous INTEX, Actes des journées Intex 2000, RISSH*, Lièges, Belgique, à paraître.

Garric N., Maurel D. (2000), *Désambiguïsation des noms propres déterminés par l'utilisation de grammaires locales*, colloque AFLA 2000 (Association française de linguistique appliquée), Paris, 6-8 juillet.

Hobbs, J.R., Appelt D.E., Bear J., Israel D., Kameyama M., Stickel M., Tyson M. (1996), FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. Dans *Finite State Devices for Natural Language Processing*. MIT Press, Cambridge, MA.

Kim J.S., Evens M.W. (1996), Efficient Coreference Resolution for Proper Names in the Wall Street Journal Text, dans *Online Proceedings of Annual MAICS Conference (Midwest Artificial Intelligence and cognitive Science Conference)*, Bloomington, USA.

Morin E. (1999), Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique, dans t.a.l. Traitement Automatique des Langues, 40(1), pp. 143-166

Piton O., Maurel D. (1997), Le traitement informatique de la géographie politique internationale, *Colloque Franche-Comté Traitement automatique des langues (FRACTAL 97)*, Besançon, 10-12 décembre, Bulag, numéro spécial, pp. 321-328.

Roche E., Schabes Y. (1997), *Finite State Language Processing*, Cambridge, Massachussets, MIT Press.

Silberztein M. (1993), *Dictionnaire électroniques et analyse automatique de textes - Le système INTEX*, Paris, Masson.