LIZARD, un assistant pour le développement de ressources linguistiques à base de cascades de transducteurs

Antonio Balvet

Université Paris X Nanterre / UMR MoDyCo
200, Avenue de la République 92 001 Nanterre
antonio.balvet@u-paris10.fr

Mots-clefs – Keywords

TALN, assistant linguistique, ressources linguistiques, agents autonomes, études sur corpus

NLP, linguistic wizard, linguistic resources, autonomous agents, corpus-based studies

Résumé – Abstract

Nous présentons un outil visant à assister les développeurs de ressources linguistiques en automatisant la fouille de corpus. Cet outil, est guidé par les principes de l'analyse distributionnelle sur corpus spécialisés, étendue grâce à des ressources lexicales génériques. Nous présentons une évaluation du gain de performances dû à l'intégration de notre outil à une application de filtrage d'information et nous élargissons le champ d'application de l'assistant aux études sur corpus menées à l'aide de cascades de transducteurs à états finis.

We present a tool providing linguistic resources developers with automated corpus analysis features. Our tool implements specialized corpora distributional analysis principles, extended by the integration of external generic lexical resources. We present an evaluation of the gain in performance attributable to our tool, for a text filtering task. We also widen our tool's scope of applications to transducer cascades-based corpus processing activities.

1 Introduction

Les approches basées sur les analyses partielles ont fait la preuve de leur efficacité en TALN, y compris pour des applications en vraie grandeur[1]. Il en va ainsi de tâches aussi diverses que

[1] Voir, par exemple, (Mohri, 2001) pour un exposé d'une techniques d'analyse partielle, appliquée à la reconnaissance vocale, ou encore (Abney, 1996) pour l'analyse syntaxique.

l'étiquetage syntaxique, les études sur corpus, la terminologie et la lexicologie, mais également d'applications de techniques de TALN à des champs connexes tels que la recherche d'information. Toutefois, toutes les analyses partielles n'offrent pas les mêmes potentiels de réutilisabilité et d'extension, ni la même transparence. Ainsi, les ressources développées pour une application sont rarement directement exploitables pour d'autres applications, voire d'autres types de corpus (documentation technique, e-mail etc…). Par ailleurs, chaque outil demande un apprentissage de la part des utilisateurs.

Comment faciliter la tâche des utilisateurs, éventuellement non linguistes et non informaticiens, de systèmes à visée industrielle ? Autrement dit, comment garantir des performances intéressantes tout en délestant les utilisateurs d'une partie de la charge de développement ? Nous pensons apporter une réponse à cette question en assurant l'intégration de connaissances spécifiques (tirées du corpus) et génériques (hors-corpus) de façon automatique, par le biais d'un assistant linguistique : LIZARD (LInguistic wiZARD). Cet assistant est guidé par les principes de l'analyse distributionnelle, appliquée aux corpus de spécialité. Nous présenterons tout d'abord les fonctionnalités principales de LIZARD, puis nous examinerons un exemple d'application à la recherche d'information. Nous présenterons, une évaluation chiffrée de l'apport de LIZARD pour une tâche de filtrage d'information par cascades de transducteurs.

2 LIZARD, fonctionnalités principales

2.1 Une architecture modulaire

La Figure 1 donne un aperçu de l'architecture de l'assistant linguistique, dans laquelle les composants sont représentés sous la forme de boîtes rectangulaires.

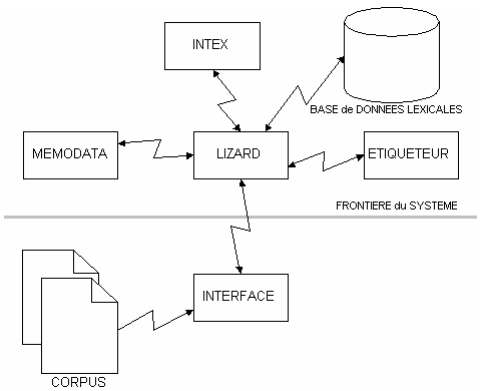

Figure 1 : vue globale de l'architecture de LIZARD

Les composants retenus sont : Intex, pour les opérations liées aux corpus[2], effectuées grâce à des cascades de transducteurs à états finis, Memodata[3], pour les opérations sémantiques telles que le

[2] Voir (Silberztein *et al*., 2001) pour plus de précisions sur la plate-forme Intex.

calcul de la distance sémantique entre deux expressions, la comparaison de mots, expressions et phrases. Par ailleurs, un étiqueteur syntaxique générique, en l'occurrence QTag de O. Mason de l'université de Sheffield est intégré au système, ainsi qu'une interface utilisateur fournissant à l'utilisateur un accès distant (applet Java). Les échanges entre composants sont représentés par des flèches brisées uni- ou bidirectionnelles en fonction des propriétés de chaque objet. La ligne grisée représente la frontière du système ; seule l'interface est accessible à l'utilisateur. L'entrée du système est constituée par les corpus de textes bruts, la sortie est une base de données lexico-grammaticales, codant quelques propriétés syntaxiques de surface ainsi que quelques propriétés lexicales (liens sémantiques) de verbes extraits des corpus : nombre et type de compléments habituels de chaque verbe, transformations possibles, termes sémantiquement reliés. Cette base est le résultat d'une expansion sémantique réalisée par le composant Memodata à partir de schémas de sous-catégorisation rudimentaires extraits des corpus.

Dans cette conception modulaire, chaque composant peut être remplacé si l'application le demande : ainsi, on peut envisager de remplacer Memodata par Wordnet, Intex par d'autres outils d'exploration des textes, ou encore d'inclure un nouveau module. Chaque module peut être aisément transformé en agent autonome en suivant les spécifications de la plate-forme Open Agent Architecture, développée au Stanford Research Institute[4]. En effet, la déclaration d'un agent OAA passe schématiquement par la spécification des services qu'il assure en termes de requêtes et de réponses, les échanges normalisés entre agents OAA étant contrôlés par un agent superviseur. L'intérêt majeur de la plate-forme OAA est la possibilité de faire cohabiter des agents hétérogènes, en l'occurrence, pour LIZARD, les agents Memodata, Interface et Étiqueteur sont écrits en Java, alors que l'agent Intex est développé en C/C++. L'ensemble des échanges entre agents ont lieu sous la forme de requêtes adressées au superviseur, qui les aiguille vers le bon service. LIZARD est, de fait, un système multiagents distribué : les modules gourmands en ressources (tels que Memodata et Intex) peuvent être localisés sur des serveurs dédiés, pour ne laisser que l'interface sur le poste client.

2.2 Extraction de signatures thématiques

La fonctionnalité principale de LIZARD est l'extraction d'expressions typiques d'un domaine, que nous appelons signatures thématiques. Cette extraction repose sur les modules Intex et Memodata et vise à produire des bases de données lexicales proches, dans leur format, des tables du lexique-grammaire. Les signatures thématiques recherchées[5] se distinguent des termes (Bourigault, 1993), des unités lexicales complexes (Habert *et al.*, 1997), ou encore des collocations en ce qu'elles ne valent que pour un domaine, un corpus de spécialité et une application donnée. Ces signatures sont essentiellement construites autour de prédicats, réalisés soit par des verbes pleins à l'actif et au passif

---

[3] Les opérations sémantiques sont assurées par le Dictionnaire Intégral (DI), décrit dans (Dutoit, 2000).

[4] Voir (Martin et *al*., 1999).

[5] Par exemple : *Thales rachète la filiale EADS de Dassault*.

(ex : *racheter la filiale XY*), soit par des formes nominalisées éventuellement associées à des verbes-support (ex : *se porter acquéreur de la filiale XY*).

2.2.1 Application de cascades de transducteurs pour l'extraction de schémas de sous-catégorisation

LIZARD, par le biais de l'agent Intex, permet de prétraiter, étiqueter puis explorer des corpus de textes de façon efficace. En effet, Intex repose sur la technique des cascades de transducteurs à états finis, qui assurent des temps de traitement compatibles avec des applications en vraie grandeur. Par ailleurs, Intex a adopté une représentation graphique pour les transducteurs utilisés tant dans les étapes de prétraitement que dans celles de fouille des textes à partir de requêtes des utilisateurs, qui assurent une meilleure lisibilité aux grammaires décrites sous cette forme par rapport à des expressions régulières classiques (voir Figure 2) ou des règles de type clauses Prolog. De plus, chaque transducteur graphique peut faire appel à d'autres transducteurs, ce qui confère à Intex des qualités intéressantes en termes de réutilisabilité et de modularité. Le transducteur graphique de la Figure 2 ci-dessous décrit une grammaire locale très simple reconnaissant un groupe verbal dont la tête est conjuguée au passé composé, à la voix passive[6] (la reconnaissance d'une insertion non bornée entre l'auxiliaire et le verbe plein est rendue possible grâce à l'état <MOT>*).

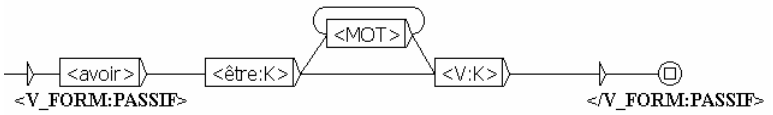

Figure 2 : transducteur graphique Intex codant une grammaire locale simplifiée

Les transducteurs graphiques Intex représentent visuellement les alternatives par autant de chemins divergents, les éléments à repérer sont décrits dans les boîtes, la sortie de chaque transducteur est matérialisée en dessous de l'état qui la produit. Intex permet de rechercher des séquences de caractère particulières, mais également des constructions syntaxiques particulières : le système reconnaît une quarantaine d'étiquettes syntaxiques telles que N, Det, V etc…

LIZARD fonctionne comme une surcouche au-dessus des composants particuliers intégrés, destinée autant à faciliter la tâche de développeurs experts dans l'élaboration de ressources linguistiques, que d'utilisateurs non experts. Dans l'état actuel, LIZARD n'offre que les services orientés vers les experts : la fonctionnalité principale de l'assistant est l'extraction de patrons de sous-catégorisation rudimentaires à partir de textes étiquetés et désambiguï sés. Cette extraction repose sur plusieurs phases dites de « généralisation », elle vise à fournir un ensemble d'expressions typiques et non ambiguës d'un domaine de spécialité, en fonction d'un corpus et d'une application particulière visée :

---

[6] Cette grammaire locale reconnaît les formes suivantes : *ont été vraiment très choqués, a été atteint* etc… Elle permet de marquer dans le texte analysé le début et la fin de la séquence reconnue, grâce aux balises <V_FORM :PASSIF></V_FORM :PASSIF> introduites en sortie de transducteur. Ce marquage est susceptible d'être utilisé lors de phases d'analyse ultérieures, selon le principe des cascades de transducteurs (Abney, 1996).

les signatures thématiques. Ces phases ont pour but de ne sélectionner que les unités potentiellement intéressantes au regard de l'application visée, de façon paramétrable. Ainsi, la Figure 3 donne un aperçu d'une phase de généralisation visant à ne conserver que la forme lemmatisée des entrées verbales, suivie d'un certain nombre de compléments essentiels[7]. Par ailleurs, les mots mal étiquetés sont conservés tels quels.

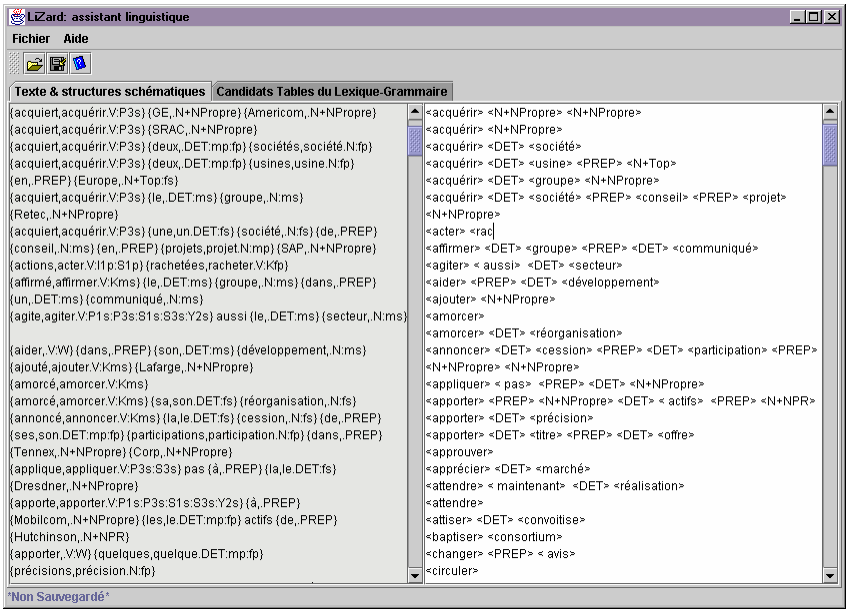

Figure 3 : fouille de corpus avec LIZARD, première phase

À ce stade, un certain nombre d'observations sur les préférences de sélection des verbes extraits sont possibles : on voit que pour le corpus considéré (un corpus financier), au moins deux constructions sont possibles pour le verbe acquérir : acquérir + Nom Propre (un nom de société), et acquérir + groupe nominal (Det + (usine, société, groupe, nom propre)). Cette première phase de généralisation est suivie d'une deuxième phase, qui vise à ne produire que des schémas de sous-catégorisation tels que : V + Det + N, V + Prep + Det + N etc … Cette deuxième phase sert essentiellement à proposer des candidats-signatures thématiques, qui seront stockées dans la base de données lexico-grammaticales sous une forme proche des tables du lexique-grammaire[8]. La Figure 4 donne un aperçu de la seconde phase. Cette figure présente deux tables correspondant aux deux schémas suivants : V + Prep + NPropre et V + Det + N. La phase de validation des candidats-signatures thématiques permet de ne sélectionner que les entrées pertinentes. En l'état actuel, cette

[7] Principalement des Noms, des Déterminants, des Prépositions, quelques Adverbes.

[8] Une entrée lexicale suivie de traits binaires codant un certain nombre de propriétés syntaxiques et sémantiques, telles que le type des compléments possibles, les transformations valides etc…

validation est réalisée manuellement, toutefois nous envisageons de l'automatiser en utilisant les fonctions de calcul de distance sémantique de Memodata[9].

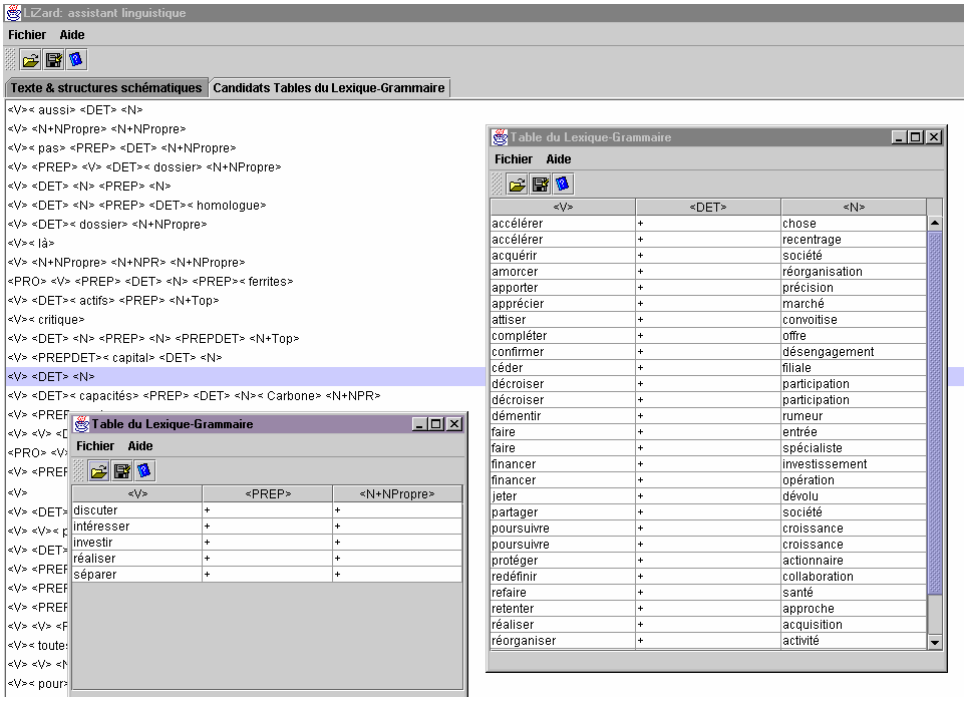

Figure 4 : fouille de corpus avec LIZARD, validation des candidats-signature thématique

2.2.2 Interrogation d'un réseau sémantique

Une fois les candidats-signatures extraits, une phase d'expansion permet de compléter les schémas de sous-catégorisation par l'apport de connaissances hors-corpus. Le DI constitue un réseau où chaque entrée est accessible par le biais d'une relation sémantique. Les entrées du dictionnaire sont des mots simples ou composés, mais également des locutions plus complexes. L'expansion du noyau de signatures thématiques est réalisée de façon interactive, en proposant à l'utilisateur des termes sémantiquement proches de ceux trouvés dans la base : synonymes, génériques, spécifiques, ainsi que locutions proches et formes transformées (ex : formes nominalisées d'une entrée verbale[10]).

---

9 Le DI intègre des algorithmes de calcul de distance sémantique qui permettent de trouver, par exemple, que *acheter une société* et *acheter une entreprise* sont plus proches l'un de l'autre que de *acheter des fleurs*. Nous envisageons de mettre en œuvre ces algorithmes afin de proposer à l'utilisateur un regroupement des candidats-signatures, en fonction de leur profil sémantique (ex., un classement tel que : *racheter DET filiale NPropre > racheter DET société NPropre > PRO racheter DET conduite*).

10 Ex : le DI permet de calculer *acheteur* et *achat* à partir de *acheter*, par les relations « personne qui V » et « action de V ». L'algorithme de parcours du réseau lui-même est décrit de façon extensive dans (Dutoit, 2000). (Poibeau, 2002) donne un exemple de paramétrage de cet algorithme pour une tâche d'acquisition de patrons lexicaux utilisés pour l'extraction d'information.

2.2.3 Une base de données lexico-grammaticales

Le résultat des opérations de fouille de texte et d'expansion des candidats-signatures thématiques est une base de données lexicales, codant le comportement syntaxique de chaque entrée, ainsi qu'un certain nombre d'informations sémantiques (ex : termes proches). La Figure 5 donne un aperçu d'une base de signatures thématiques extraites d'un corpus financier, destinées à être utilisées par un système de filtrage d'information par cascades de transducteurs à états finis : CORAIL[11].

N0 =: Nspec	N1 =: Nspec	N2	PPV	V	N0 V	N0 V N1	N0 V Prep N1	N0 V Const N1	N0 V N1 Prep N2	Const	Complt	VN	Actif	Passif	Nominalisation
:ntrepris	ntrepris	ntrepri	<E>	acheter	-	+	-	-	+	<E>	<E>	<achat>	+	+	+
:ntrepris	ntrepris	ntrepri	<E>	acquérir	-	+	-	-	+	<E>	<E>	<acquisition>	+	+	+
:ntrepris	ntrepris	ntrepri	<E>	augmenter	-	-	-	+	-	<E>	:Capital	<augmentation>	+	+	+
:ntrepris	ntrepris	ntrepri	<E>	échanger	-	+	-	-	+	<E>	<E>	<échange>	+	+	+
:ntrepris	ntrepris	ntrepri	:Refl	engager	-	-	+	+	-	<E>	:Capital	<engagement>	+	-	+
:ntrepris	ntrepris	ntrepri	<E>	entrer	-	-	-	+	-	<E>	:Capital	<entrée>	+	-	+
:ntrepris	ntrepris	ntrepri	<E>	fusionner	-	-	+	-	-	<E>	<E>	<fusion>	+	+	+
:ntrepris	ntrepris	ntrepri	<E>	investir	+	-	+	+	-	<E>	:Capital	<investissement>	+	+	+
:ntrepris	ntrepris	ntrepri	:Refl	marier	-	-	+	-	-	<E>	<E>	<mariage>	+	-	+
:ntrepris	ntrepris	ntrepri	<E>	mettre	-	-	-	+	-	la main sur	<E>	<E>	+	-	-
:ntrepris	ntrepris	ntrepri	:Refl	porter	-	-	-	+	-	acquéreur de	<E>	<E>	+	-	-
:ntrepris	ntrepris	ntrepri	<E>	prendre	-	-	-	+	-	<E>	:Capital	<prise>	+	+	+
:ntrepris	ntrepris	ntrepri	<E>	racheter	-	+	-	-	+	<E>	<E>	<rachat>	+	+	+
:ntrepris	ntrepris	ntrepri	<E>	racheter	-	-	-	+	+	<E>	:Capital	<rachat>	+	+	+

Figure 5 : base de signatures thématiques extraites d'un corpus financier

Cette base est le résultat d'une quinzaine d'heures de travail par l'auteur du système, elle regroupe une vingtaine d'entrées lexicales, et représente une partie des contraintes de sélection et de construction associées à chaque entrée (ex : nombre, type de compléments, transformations autorisées, formes nominalisées). Les candidats-signatures ont été validés à la main, intégrés et complétés par consultation du DI. Le format de la base elle-même est libre, bien que les informations contenues doivent être, en l'état actuel, compatibles avec Intex. On peut envisager une représentation XML de ces données, traduites par la suite dans les formats compatibles avec d'autres plateformes.

3 Application au filtrage d'information

Nous évoquons ici une application du système LIZARD dans le domaine de la recherche d'information, en mettant l'accent sur le filtrage d'information par cascades de transducteurs.

[11] Voir (Balvet *et al*., 2001).

3.1 Description de l'expérience

Le filtrage d'information consiste à écouter un flux de documents et à attribuer automatiquement chaque nouveau document au bon utilisateur, en fonction de son besoin en information, ce de façon binaire[12]. Un prototype de filtrage d'information par cascades de transducteurs simples est décrit dans (Balvet *et al.*, 2001) ; on peut envisager l'extension de ce système aux signatures thématiques telles que décrites plus haut. La **Erreur ! Source du renvoi introuvable.** présente une mesure du gain observé lors d'une migration vers l'approche par signatures thématiques, par rapport à une approche manuelle et une expansion sémantique non systématique.

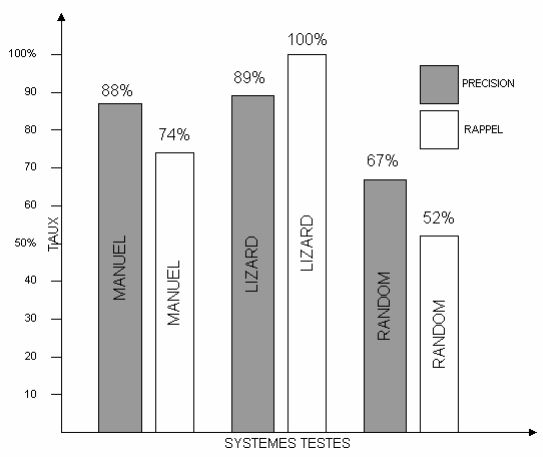

Figure 6 : comparaison de trois approches pour le filtrage d'information par cascades de transducteurs

Le système « random » nous sert à fixer la borne inférieure de qualité pour l'évaluation, il permet, selon nous, de relativiser les performances des systèmes évalués, eu égard à la tâche considérée (classification binaire) : une grande partie du problème (environ 50% en rappel) est couverte par cette approche aléatoire, sans aucune intelligence. Le système dit « manuel » est le résultat d'une étude classique sur corpus, effectuée par S. Bizouard en vue d'une expérience d'évaluation sur une tâche d'extraction d'information. Les grammaires locales développées au cours de cette expérience visaient à alimenter une base de données de scénarios de cessions et acquisitions de sociétés[13] (acheteur, entreprise achetée, montant de la transaction etc…). Ces grammaires peuvent également servir à une tâche de filtrage : elles repèrent, de fait, des documents traitant du thème « cessions et acquisitions ».

Le corpus employé pour cette évaluation consiste en un ensemble de dépêches du domaine financier, il nous a été communiqué par la société Firstinvest, propriétaire d'un portail de services financiers sur Internet. Il est constitué de 2,6 Mo de dépêches faiblement balisées[14], rédigées par des experts

---

[12] Chaque document entré est classé comme pertinent/non pertinent en fonction de chaque profil.

[13] Voir (Bizouard, 2001).

[14] Les documents sont découpés en unités textuelles majeurs telles que : titre, corps de dépêche, date etc… Les noms propres et les noms de société sont respectivement repérés par les balises : et .

financiers en français, traitant 19 thèmes (de « rumeurs » à « profit-warning » en passant par « cessions/acquisitions de sociétés »). Les documents dont nous disposons ne constituent que des exemples positifs, en termes d'apprentissage, étant donné que l'ensemble du fonds documentaire nous est parvenu indexé en fonction des thèmes évoqués plus haut. Nous avons donc dû élaborer manuellement un corpus d'exemples négatifs (50 dépêches), constitué de documents attribués à d'autres thèmes que les cessions/acquisitions de sociétés. Pour notre évaluation, nous nous limiterons au thème 19 : cessions/acquisitions de sociétés, qui regroupe 303 documents, répartis en 2/3 pour le corpus de paramétrage et 1/3 pour le corpus de test.

L'apport essentiel d'une approche telle qu'implémentée par LIZARD semble résider dans une meilleure couverture du domaine (100% de taux de rappel) tout en gardant une précision comparable à l'approche manuelle. Par ailleurs, le recours à un assistant linguistique permet de réduire le temps de développement : d'environ 3 hommes-mois, pour l'approche manuelle, à une quinzaine d'heures pour l'approche semi-automatique.

4 Conclusion et perspectives

Nous avons présenté un outil destiné à alléger la tâche des développeurs de ressources linguistiques basées sur les analyses partielles et les cascades de transducteurs à états finis. Cet outil, LIZARD, est conçu comme un système modulaire, paramétrable, construit autour de composants linguistiques génériques, ayant fait la preuve de leur efficacité. Nous avons souligné l'intérêt d'un tel dispositif, au regard de la diffusion des approches dites partielles et de celles basées sur des cascades de transducteurs. Nous avons montré quel gain le recours à un tel assistant pouvait apporter sur une tâche de recherche d'information : le filtrage d'information. Cet outil permet également de valider l'hypothèse sous-jacente à notre travail exposé : les textes de spécialité renferment des régularités linguistiques susceptibles d'être mises à jour par une analyse des propriétés distributionnelles des objets linguistiques considérés (*i.e.* les signatures thématiques). Les objets ainsi mis en évidence fournissent une sorte de profil thématique des textes, dont la précision permet d'envisager des applications en vraie grandeur, telles que le filtrage d'information. Par ailleurs, des applications à des domaines tels que l'enseignement des langues, la terminologie et la lexicographie nous paraissent pouvoir également bénéficier de l'approche décrite ici.

Le travail présenté ici suscite toutefois des interrogations quant à l'extension d'une approche distributionnelle reposant sur une analyse partielle (grammaires locales), complétée par des ressources externes, à d'autres domaines, d'autres corpus ou encore d'autres applications. Par ailleurs, nous comptons compléter l'approche décrite ici, reposant uniquement sur des ressources explicites (règles symboliques), par la prise en compte des liens entre distribution et probabilité d'occurrence, tels qu'explorés dans (Daille, 1994).

Références

Abney S. (1996), Partial parsing via finite-state cascades, *Actes de ESSLLI'96 Robust Parsing Workshop*.

Balvet A., Meunier F., Poibeau T., Viard D., Vichot F., Wolinski F. (2001), Filtrage de documents et grammaires locales : le projet CORAIL, Actes du *Troisième Congrès du Chapitre Français de l'ISKO (International Society for Knowledge Organisation) : Filtrage et résumé automatique de l'information sur les réseaux*, 5-6 juillet 2001, Université de Nanterre-Paris X.

Bizouard S. (2001), *Évaluation d'outils d'acquisition de ressources linguistiques pour l'extraction*, Mémoire de DESS en Ingénierie Multilingue, CRIM, INALCO.

Bourigault D. (1993), Analyse syntaxique locale pour le repérage de termes complexes dans un texte, *TAL* n°34(2).

Daille B. (1994), *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*, Thèse de doctorat, Université Paris VII.

Dutoit D. (2000), *Quelques opérations Texte ® Sens et Sens ® Texte utilisant une sémantique linguistique universaliste apriorique*, Thèse de doctorat, Université de Caen.

Habert B., Nazarenko A., Salem A. (1997), *Les linguistiques de corpus*, Armand Colin.

Martin D., Cheyer A.J., Moran D.B., (1999), The Open Agent Architecture: a framework for building distributed software systems, *Applied Artificial Intelligence*, vol. 13, p 91-128.

Mohri M., (2001), Language processing with weighted transducers, *Actes de la huitième conférence sur le Traitement Automatique des Langues Naturelles*, 2-5 juillet 2001, pp.5-14, Tours.

Poibeau T. (2002), *Extraction d'information à base de connaissances hybrides*, Thèse de doctorat, Université Paris XIII.

Riloff E. (1994), *Information Extraction as a Basis for Portable Text Classification Systems*, Thèse de doctorat, Université du Massachussets Amherst.

Silberztein M., Poibeau T., Balvet A., (2001). Tutoriel : Intex et ses applications informatiques, *Actes de la huitième conférence sur le Traitement Automatique des Langues Naturelles*, 2-5 juillet 2001, pp.145-174, Tours.