

TALN 2007, Toulouse, 12–15 juin 2007

Evaluer SynLex

Ingrid Falk¹ Gil Francopoulo² Claire Gardent³

(1) CNRS/ATILF, Nancy

(2) INRIA/LORIA, Nancy

(3) CNRS/LORIA, Nancy

Ingrid.Falk@loria.fr

Gil.Francopoulo@wanadoo.fr

Claire.Gardent@loria.fr

Résumé SYNLEX est un lexique syntaxique extrait semi-automatiquement des tables du LADL. Comme les autres lexiques syntaxiques du français disponibles et utilisables pour le TAL (LEFFF, DICOVALENCE), il est incomplet et n'a pas fait l'objet d'une évaluation permettant de déterminer son rappel et sa précision par rapport à un lexique de référence. Nous présentons une approche qui permet de combler au moins partiellement ces lacunes. L'approche s'appuie sur les méthodes mises au point en acquisition automatique de lexique. Un lexique syntaxique distinct de SYNLEX est acquis à partir d'un corpus de 82 millions de mots puis utilisé pour valider et compléter SYNLEX. Le rappel et la précision de cette version améliorée de SYNLEX sont ensuite calculés par rapport à un lexique de référence extrait de DICOVALENCE.

Abstract SYNLEX is a syntactic lexicon extracted semi-automatically from the LADL tables. Like the other syntactic lexicons for French which are both available and usable for NLP (LEFFF, DICOVALENCE), it is incomplete and its recall and precision wrt a gold standard are unknown. We present an approach which goes some way towards addressing these shortcomings. The approach draws on methods used for the automatic acquisition of syntactic lexicons. First, a new syntactic lexicon is acquired from an 82 million words corpus. This lexicon is then used to validate and extend SYNLEX. Finally, the recall and precision of the extended version of SYNLEX is computed based on a gold standard extracted from DICOVALENCE.

Mots-clefs : Lexique syntaxique, Evaluation

Keywords: Syntactic lexicon, Evaluation

1 Introduction

Un lexique syntaxique décrit les propriétés syntaxiques des mots d'une langue. En particulier, un lexique syntaxique associe à chaque foncteur syntaxique un *cadre de sous-catégorisation* spécifiant le nombre et le type (catégorie syntaxique, marqueur introductif, mode, etc.) de ses arguments.

Comme l'ont montré (Carroll & Fang, 2004), un lexique syntaxique exhaustif et détaillé permet d'améliorer les performances des analyseurs syntaxiques. Un tel lexique est également une composante essentielle de tout réalisateur de surface puisqu'il permet de réaliser un contenu sémantique donné par une phrase bien formée et en particulier, une phrase où chaque foncteur syntaxique a le nombre et le type d'arguments requis par son régime. Plus généralement, un lexique syntaxique est une composante de base pour tout système faisant intervenir soit l'analyse, soit la réalisation.

Pour le français, il existe à l'heure actuelle trois lexiques syntaxiques disponibles librement et utilisables par des systèmes de traitement automatique des langues : Proton récemment renommé DicoValence, (van den Eynde & Mertens, 2003), Lefff (Clément *et al.*, 2004) et SYNLEX (Gardent *et al.*, 2006). Néanmoins aucun de ces lexiques n'est entièrement satisfaisant pour deux raisons.

Premièrement, aucun de ces lexiques ne couvre l'ensemble des verbes du français. Ainsi pour 8 790 verbes identifiés pour le français dans Morphalou (Romary *et al.*, 2004), DicoValence inclut 3 700 verbes, Lefff 6 798 et SYNLEX 5244.

Deuxièmement, la qualité de leur contenu et plus précisément, leur rappel et leur précision restent inconnus : Pour l'ensemble des entrées contenues dans chacun de ses dictionnaires, on ne connaît ni quelle proportion des entrées correctes est présente (rappel) ni quelle est la proportion d'entrées incorrectes (précision).

Dans cet article, nous considérons SYNLEX et présentons une approche qui vise à pallier ces lacunes. L'approche s'appuie sur les méthodes mises au point en acquisition automatique de lexique. Un lexique syntaxique (CORLEX) distinct de SYNLEX est acquis à partir d'un corpus de 82 millions de mots. Ce lexique est ensuite utilisé pour valider et compléter SYNLEX. Le rappel et la précision de SYNLEX, de la version améliorée de SYNLEX et de CORLEX sont ensuite calculés par rapport à un lexique de référence extrait de DICOVALENCE.

L'article est structuré comme suit. La section 2 décrit le processus de création de SYNLEX et présente son format et son contenu. La section 3 présente les travaux visant à valider et à étendre SYNLEX puis commente les résultats obtenus. La section 4 conclut en indiquant les directions de recherche futures.

2 Synlex

Synlex est un lexique créé à partir des tables du LADL (Gross, 1975; Guillet & Leclère, 1992; Boons *et al.*, 1976). Le processus de création a été décrit dans (Gardent *et al.*, 2005b; Gardent *et al.*, 2006; Gardent *et al.*, 2005a) et peut être résumé comme suit :

1. une représentation du contenu des colonnes des tables et de leurs interdépendance est

créée manuellement sous la forme d'un graphe *et/ou* dont les noeuds contiennent à la fois des conditions et des pointeurs vers le contenu des colonnes

2. ce graphe *et/ou* est ensuite utilisé en conjonction avec les tables pour produire de façon automatique un lexique syntaxique représentant leur contenu
3. ce lexique est ensuite simplifié pour ne contenir que le type d'information habituellement présente dans un lexique syntaxique (i.e., nombre et types de syntagmes sous-catégorisés par les verbes)

Le format des entrées de SYNLEX est spécifié dans la figure 1 et peut être décrit comme suit. Une entrée se compose d'un verbe, d'une liste d'arguments syntaxiques ayant un rôle sémantique, d'une liste optionnelle d'*associés* c-à-d, d'arguments régis par le verbe mais ne remplissant pas de rôle sémantique (e.g., l'explétive *il* dans *il pleut*) et d'une liste de *macros* donnant des informations supplémentaires sur les propriétés syntaxiques du verbes (e.g., contrôle, passivisation). Les *associés* et les *macros* sont des listes finies d'atomes. Un argument en revanche est défini par un triplet de la forme $F : M - C$ où F est une fonction grammaticale, M un marqueur optionnel (une préposition ou un clitique indiquant la cliticisation d'un argument en cas d'ambiguïté comme par exemple les arguments en *à* qui peuvent se cliticiser soit en *y*, soit en *lui*) et C est une catégorie syntaxique.

<i>Entree</i>	::=	<i>Verb</i> : $\langle Arg^+ \rangle$, <i>Associe</i> [*] , <i>Macro</i> [*]	(1)
<i>Arg</i>	::=	<i>Fonction</i> : <i>Marqueur</i> – <i>Categorie</i>	(2)
<i>Fonction</i>	::=	<i>subj</i> <i>obj</i> <i>obja</i> <i>objde</i> <i>obl</i> <i>attr</i>	(3)
<i>Marqueur</i>	::=	<i>Prep</i> <i>Clitic</i> <i>Compl</i>	(4)
<i>Categorie</i>	::=	<i>sn</i> <i>pinf</i> <i>pcompl</i> <i>qcompl</i>	(5)
<i>Associe</i>	::=	<i>ilimp</i> <i>cln</i> <i>cla</i> <i>cld</i> <i>clg</i> <i>pron</i>	(6)
<i>Macro</i>	::=	<i>CtrlArgXArgY</i> <i>passivable</i> <i>nonPassivable</i>	(7)

FIG. 1 – SynLex Format

Seules 60% des tables du LADL étant disponibles, nous avons complété manuellement le lexique extrait des tables disponibles avec environ 2 000 verbes et leurs cadres de base. Le lexique SYNLEX résultant contient 5244 verbes et 19127 entrées (paires verbe - cadre) faisant intervenir 726 cadres de sous-catégorisation en considérant les associés et 538 cadres de sous-catégorisation sans associés.

3 Evaluation

Comme nous l'avons mentionné, SYNLEX est produit à partir des tables du LADL par un processus de conversion faisant intervenir une représentation intermédiaire. Or l'information contenue dans les tables peut être inexacte et la conversion dans le format SYNLEX peut introduire des erreurs. Enfin, le lexique produit ne couvre ni l'ensemble des verbes du français, ni nécessairement, l'ensemble des entrées d'un verbe. Il est donc nécessaire à la fois de valider et de compléter le lexique obtenu.

Au cours des 15 dernières années, des travaux (Brent, 1991; Briscoe & Carroll, 1997; Manning, 1993) ont montré qu'il est possible d'extraire un lexique syntaxique d'un corpus en utilisant d'abord un analyseur puis un filtre statistique. L'idée est la suivante. Dans un premier temps, un analyseur déterministe est utilisé pour produire à partir d'un corpus des hypothèses sur les cadres de sous-catégorisation des verbes présents dans ce corpus. Plus précisément, l'analyse produite pour chaque proposition par l'analyseur est utilisée pour associer au verbe de la proposition une description des syntagmes maximaux (groupe nominal, groupe prépositionnel, proposition infinitive, etc.) apparaissant avec ce verbe. Dans un deuxième temps, les hypothèses sont soumises à un calcul statistique et seules sont conservées les hypothèses pour lesquelles la probabilité d'erreur est suffisamment basse. Le lexique ainsi obtenu est ensuite évalué (rappel et précision) par rapport à un lexique de référence validé manuellement.

Nous utilisons ici les idées issues de ces travaux pour évaluer la qualité de SYNLEX. D'une part, nous montrons comment un lexique extrait d'un corpus (CORLEX) peut être utilisé pour valider SYNLEX et l'enrichir. Le lexique résultant est appelé XSYNLEX. D'autre part, nous comparons les trois lexiques ainsi créés (SYNLEX, XSYNLEX et CORLEX) avec un corpus de référence (REFLEX) extrait de DICOVALENCE.

3.1 Comparaison et fusion avec un lexique acquis à partir de corpus

Afin d'évaluer la précision et la couverture de SYNLEX, nous commençons par le comparer avec un lexique acquis automatiquement à partir d'un corpus. Ce lexique (CORLEX) est acquis selon la méthodologie décrite ci-dessus : un corpus et un analyseur sont d'abord utilisés pour émettre des hypothèses sur les entrées lexicales (association verbe - cadre) possibles. Ensuite, ces hypothèses sont soumises à un calcul statistique permettant de classer les hypothèses en hypothèses plausibles et hypothèses non plausibles. Dans ce qui suit, nous détaillons chacun de ces procédés.

Création des hypothèses. Le corpus exploité est un corpus de 82 millions de mots avec 65% d'articles de presse, 30 % de compte rendus de débats parlementaires et 5% de textes littéraires.

L'analyseur (TAGPARSER) est un analyseur robuste ascendant qui exploite des connaissances très fines sur la combinaison des mots grammaticaux classifiés en 300 classes de mots simples ou composés (Francopoulo, 2005). Dans la version actuelle (version 1), mise à part, une catégorisation binaire des adjectifs et l'indication comme quoi le verbe accepte ou non, une complétive, l'analyseur n'utilise pas d'information portant sur la sous-catégorisation des verbes et des noms prédicatifs. La technologie mise en oeuvre combine un automate et une matrice statistique induite à partir d'un corpus de 77 000 mots annotés en syntaxe de surface.

Enfin notons que pour cette première expérience, nous nous sommes limités aux cadres qui sont relativement faciles à détecter pendant l'analyse syntaxique i.e., les cadres ne faisant intervenir ni la fonction oblique, ni la fonction attribut. En outre, les associés (e.g., réflexif intrinsèque, clitique figé) et les macros qui concernent des propriétés syntaxiques non détectables par un analyseur (e.g., phénomènes de contrôle, acceptation ou non pour les verbes transitifs de la forme passive, etc.) ne sont pas pris en compte.

L'analyse du corpus par TAGPARSER permet d'extraire 38 550 hypothèses où chaque hypothèse est l'association d'un verbe, d'un cadre et d'une fréquence d'apparition de cette association dans le corpus.

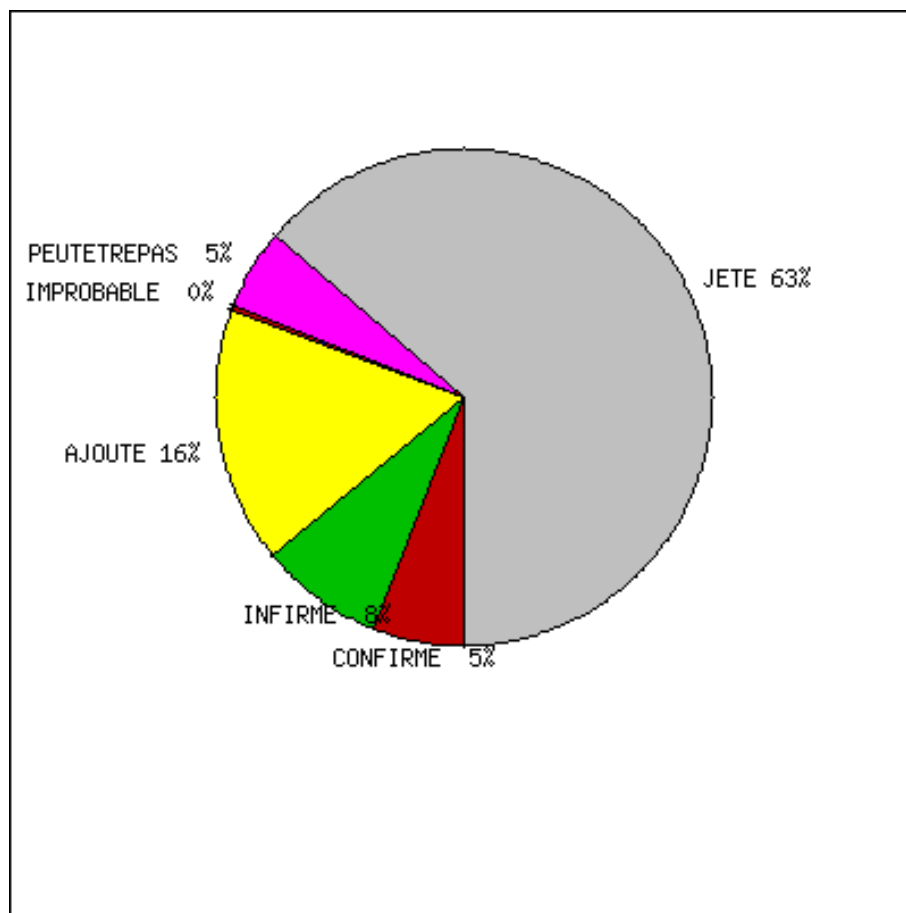


FIG. 2 – Résultats

Filtrage des hypothèses. Afin d'évaluer la plausibilité des hypothèses émises, nous utilisons un test souvent mis en oeuvre (Brent, 1991; Briscoe & Carroll, 1997; Manning, 1993) par les approches portant sur l'acquisition automatique de lexiques à savoir le test binomial sur les hypothèses (BHT). Ce test calcule la probabilité que m occurrences du cadre c apparaissent avec un verbe v n'acceptant pas ce cadre, étant donné n occurrence de ce verbe. Plus la probabilité est basse, plus l'hypothèse est douteuse et par conséquent, plus il est probable que c est un cadre valide de v .

En pratique, nous fixons à 0.05% le seuil utilisé pour déterminer si ou non une association verbe-cadre apparait suffisamment peu fréquemment pour être une erreur. En d'autres termes, toutes les hypothèses pour lesquelles la probabilité d'erreur donnée par le test BHT est en dessous de 0.05% sont acceptées comme valides – les autres sont rejetées. Pour calculer la probabilité d'erreur des hypothèses émises, nous utilisons le UCS toolkit (<http://www.collocations.de/>). Après filtrage, le lexique syntaxique obtenu (CORLEX) comporte 8 742 entrées.

Comparaison et fusion des deux lexiques (SYNLEX et CORLEX). La figure 2 donne une analyse détaillée des résultats obtenus à partir de l'analyse de corpus. Plus généralement, on peut diviser et classifier les données suivant les critères suivants¹ :

CONFIRMÉ : les entrées présentes dans SYNLEX et dans CORLEX et pour lesquelles la pro-

¹Les pourcentages sont donnés par rapport à l'union de CORLEX et SYNLEX.

tabilité d'erreur est inférieure à 0.05% .

INFIRMÉ : les entrées présentes dans SYNLEX et dans CORLEX et pour lesquelles la probabilité d'erreur est supérieure à 0.05% .

AJOUTÉ : les entrées absentes dans SYNLEX qui sont présentes dans CORLEX et pour lesquelles la probabilité d'erreur est inférieure à 0.05% .

JETÉ : les entrées absentes dans SYNLEX qui sont présentes dans CORLEX et pour lesquelles la probabilité d'erreur est supérieure à 0.05% .

IMPROBABLE : les entrées présentes dans SYNLEX absentes dans CORLEX et pour lesquels le verbe impliqué apparaît plus de 5 000 fois dans le corpus.

PEUTÊTREPAS : les entrées présentes dans SYNLEX absentes dans CORLEX et pour lesquels le verbe impliqué apparaît moins de 5 000 fois dans le corpus.

La classe **CONFIRMÉ** permet de valider la partie de SYNLEX trouvée en corpus et validée par les statistiques. Inversement, la classe **INFIRMÉ** permet de détecter les entrées de SYNLEX qui sont sans doute incorrectes. Les données montrent en particulier, que sur la base de cette analyse, plus de la moitié des entrées de SYNLEX sont jugées incorrectes.

Par ailleurs, la classe **AJOUTÉ** permet d'étendre SYNLEX avec les entrées jugées fiables par l'analyse de corpus mais non contenues par SYNLEX. Ceci permet d'augmenter le nombre d'entrées de SYNLEX de 34.56%.

Enfin, les classes **IMPROBABLE** et **PEUTÊTREPAS** regroupent les entrées de SYNLEX qui n'apparaissent pas dans les données extraites du corpus. Les **IMPROBABLE** sont des cas où le verbe considéré apparaît plus de 5 000 fois dans le corpus mais jamais avec le cadre prescrit par SYNLEX. Ils sont éliminés de SYNLEX. Si le verbe apparaît moins de 5 000 fois dans le corpus, l'entrée est conservée mais étiquetée comme peu fiable (**PEUTÊTREPAS**).

En résumé, la fusion **xSYNLEX** de SYNLEX avec CORLEX peut être définie par l'union de **CONFIRMÉ** avec **AJOUTÉ** :

$$\mathbf{xSYNLEX} = \mathbf{CONFIRMÉ} \cup \mathbf{AJOUTÉ} \cup \mathbf{PEUTÊTREPAS}$$

Cependant, cette fusion ne garantit pas un lexique parfait. En effet, la validation statistique reste imparfaite. Par exemple, les meilleurs lexiques extraits pour l'anglais avec des méthodes similaires à celle utilisée ici ont une F-mesure maximum tournant autour de 80 % . La deuxième étape a donc consisté à évaluer les différents lexiques (SYNLEX, CORLEX et xSYNLEX) en mesurant leur rappel et précision par rapport à un lexique de référence REFLEX. L'objectif est de déterminer si l'extension de SYNLEX par les données issues de CORLEX accroît non seulement le nombre d'entrées mais également la qualité du lexique résultant.

3.2 Evaluation de SYNLEX sur un lexique de référence

Une façon de déterminer la qualité d'un lexique consiste à calculer son rappel et sa précision par rapport à un lexique de référence. Soit *Acquis* le contenu du lexique à évaluer et *Ref* celui du lexique de référence, précision et rappel sont définis de la façon suivante :

Précision

$$P = \frac{Acquis \cap Ref}{Acquis}$$

La précision indique la proportion d'entrées correctes dans le lexique acquis (combien d'entrées sont correctes ?)

Rappel

$$R = \frac{Acquis \cap Ref}{Ref}$$

Le rappel indique la proportion entre entrées correctes présentes dans le lexique acquis et entrées présentes dans le lexique de référence (combien d'entrées correctes ont été trouvées ?).

Calcul du rappel et de la précision. Pour l'évaluation, nous avons sélectionné 100 verbes présents dans tous les lexiques (i.e., SYNLEX, XSYNLEX, DICOVALENCE et CORLEX) et distribués de façon régulière sur l'échelle du nombre d'apparition dans le corpus.

Pour chacun de ces 100 verbes, nous avons créé un lexique de référence REFLEX à partir de DICOVALENCE. Les entrées de ces verbes ont été épurées des entrées non prises en compte dans CORLEX (c-à-d, les entrées faisant intervenir des arguments obliques ou attributifs) puis traduites dans le format SYNLEX (cf. Figure 1) afin de permettre une comparaison automatique avec SYNLEX, XSYNLEX et CORLEX.

Les performances des statistiques ont été évaluées sur ces 100 verbes à travers quatre expériences visant à mesurer l'impact de la fréquence d'un cadre sur ces performances.

Etant donné C le nombre total d'entrées présentes dans CORLEX, la fréquence f_c d'un cadre c est dite HAUTE si c apparaît dans plus de 1% des entrées de CORLEX ($f_c \geq 0.01 \times C$); MOYENNE si $0.001 \times C \leq f_c \leq 0.01 \times C$; et BASSE si $f_c \leq 0.0001 \times C$.

Pour chaque lexique (SYNLEX, XSYNLEX et REFLEX), quatre (sous-)lexiques sont créés : un premier contenant toutes les entrées du lexique (TOUT) et trois autres contenant uniquement les entrées faisant intervenir des cadres de haute (HF), moyenne (MF) et basse (BF) fréquence. La référence minimum (baseline) est fixée comme étant le lexique acquis à partir du corpus sans filtrage statistique (toutes les entrées trouvées par TAGPARSER sont prises en compte).

Le rappel et la précision pour chacun des 5 cas considérés sont donnés dans la Figure 3.

Discussion. Ces premiers résultats montrent que pour l'échantillon de cadres considérés (les cadres ne faisant pas intervenir d'obliques ou d'attributs), la couverture et la précision de SYNLEX sont relativement bas. La couverture faible n'est pas surprenante et s'explique du fait de l'incomplétude inhérente aux tables du LADL puisque seules 60% des tables sont disponibles.

La mauvaise précision est en revanche plus surprenante mais peut, peut être, être expliquée par la relative permissivité des tables du LADL : si une construction est possible pour un verbe donné, elle sera marquée comme telle même si elle est très rare.

Un autre facteur contribuant à diminuer la précision concerne la décision de ne pas prendre en compte les associés c-à-d, les arguments régis par le verbe mais ne remplissant pas de rôle

		TOUT	HF	MF	LF
SYNLEX	P	0.30	0.63	0.16	0.02
	R	0.44	0.45	0.47	0.3
	F	0.37	0.54	0.31	0.16
xSYNLEX	P	0.58	0.69	0.23	0.29
	R	0.63	0.66	0.56	0.5
	F	0.59	0.67	0.4	0.4
xSYNLEX + INFIRMÉ	P	0.49	0.61	0.21	0.27
	R	0.76	0.78	0.67	0.5
	F	0.62	0.70	0.44	0.38
BASELINE	P	0.22	0.29	0.07	0.15
	R	0.89	0.95	0.70	0.5
	F	0.56	0.62	0.39	0.32

FIG. 3 – Précision et rappel

sémantique. Or parmi ces associés, on trouve le clitique réfléchi intrinsèque (e.g., *se* dans *s'évanouir*). En conséquence, toutes les entrées faisant intervenir un clitique intrinsèque (l'associé CLR) sont traitées de façon incorrecte comme des entrées sans ce clitique.

Malgré tout, un examen plus approfondi des cas fautifs reste à faire pour déterminer les causes précises de ce manque de précision et éventuellement, y remédier.

Le rappel et la précision de xSYNLEX, le lexique enrichi à partir du corpus, sont relativement bas mais proches de certains résultats obtenus dans la littérature pour des langues autres que l'anglais. (Fast & Przepiórkowski, 2005) par exemple, cite un rappel de 47% et une précision de 49% pour une expérience similaire sur le polonais. Pour ce lexique, le rappel et la précision sont meilleurs que pour SYNLEX. En d'autres termes, le lexique extrait du corpus permet de valider et d'étendre la partie de SYNLEX faisant intervenir les cadres considérés pour l'acquisition automatique.

Enfin, les données concernant xSYNLEX+ INFIRMÉ montrent qu'ignorer la plausibilité statistique des hypothèses (i.e., conserver les entrées de SYNLEX qui sont infirmées par les statistiques) permet d'améliorer le rappel (0.76 contre 0.63 dans xSYNLEX) au détriment bien sûr de la précision (0.49 contre 0.58 dans xSYNLEX).

4 Conclusion et perspectives

Comme nous l'avons mentionné dans l'introduction, trois lexiques syntaxiques sont actuellement disponibles et utilisables dans le domaine du traitement automatique des langues. Cependant, ils sont tous incomplets et leur contenu n'a pas fait l'objet d'une évaluation permettant de déterminer rappel et précision.

Le travail présenté dans cet article est un premier pas vers la définition d'une procédure d'évaluation et de fusion de ces lexiques.

Il montre en particulier que DICOVALENCE peut servir de base à la création d'un lexique de référence permettant ainsi de calculer le rappel et la précision de lexiques créés de façon automatique ou semi-automatique.

Il montre également, qu'un lexique acquis à partir d'un corpus peut permettre d'améliorer la couverture et la précision d'un lexique existant ; et plus généralement, que la comparaison et la fusion de plusieurs lexiques pourrait permettre à relativement court terme de produire un lexique syntaxique du français complet et de bonne qualité.

Néanmoins, plusieurs aspects méritent d'être approfondis.

Tout d'abord, notons que l'évaluation de SYNLEX présentée ici est très partielle puisqu'elle ne porte que sur 33 des 726 cadres présents dans SYNLEX. Une évaluation plus extensive prenant en compte les obliques et les attributs est donc nécessaire.

Un second point concerne la procédure d'acquisition automatique. En effet, l'approche présentée ici est une approche préliminaire qui peut être améliorée sur au moins deux points à savoir, la qualité des hypothèses émises d'une part et la qualité du filtre statistique d'autre part.

Les hypothèses émises peuvent être affinées par l'emploi d'un analyseur plus performant – par exemple, en utilisant une information de sous-catégorisation pour informer l'analyseur ou encore en utilisant un analyseur profond plutôt que local. Une autre possibilité que nous entendons explorer prochainement, est d'utiliser plusieurs analyseurs en parallèle et de comparer/fusionner leurs résultats par un système de vote.

Les travaux fait sur l'anglais suggèrent en outre que le filtre statistique peut être amélioré de deux façons. Ainsi (Briscoe & Carroll, 1997) montre que le seuil permettant de déterminer l'acceptabilité d'une hypothèse doit être fixé différemment suivant le type de cadre considéré plutôt que de façon uniforme pour l'ensemble des hypothèses comme nous l'avons fait ici. Et (Korhonen, 2002) montre que l'utilisation de techniques de lissages informées par les classes sémantiques de verbes permet d'améliorer les résultats. L'exploitation de ces résultats devrait permettre d'améliorer la qualité du lexique extrait.

Une troisième point, plus ouvert celui-là, concerne l'élargissement des méthodes explorées à l'ensemble du lexique et en particulier au traitement des macros. Comme nous l'avons vu, SYNLEX, LEFFF et DICOVALENCE contiennent outre des informations portant sur la valence (arguments régis par le verbe remplissant ou non un rôle sémantique), des informations portant sur les phénomènes de contrôle, la passivation, la possibilité pour un verbe d'être utilisé dans une tournure impersonnelle, etc. Si elles sont utiles pour le traitement automatique des langues et en particulier, pour l'analyse et la réalisation de surface, ces informations ne peuvent pas être extraites à partir des corpus par les techniques utilisées en acquisition automatique de lexique. Elles sont en revanche partiellement présentes dans les lexiques existants (LEFFF, DICOVALENCE et SYNLEX). Une question intéressante est donc de savoir comment cette information peut être utilisée pour informer la complétion d'un lexique partiellement sous-spécifié dans cette dimension. Ou en d'autres termes, comment un lexique acquis à partir de corpus peut être fusionné avec un ou des lexiques acquis par des méthodes «symboliques» (LEFFF, SYNLEX) de façon à enrichir la partie acquise statistiquement avec l'information additionnelle contenue dans les lexiques symboliques.

Dans tous les cas, la précision relativement basse des lexiques produits suggère qu'une phase de

validation manuelle est nécessaire. Dans cette optique, une approche qui consiste à privilégier (dans une juste mesure) le rappel plutôt que la précision est sans doute préférable (il est plus facile d'éliminer que d'ajouter). Ce qui suggère en particulier, que XSYNLEX+ INFIRMÉ est préférable à XSYNLEX et plus spécifiquement, que l'extraction de SYNLEX à partir des tables est utile.

Références

- BOONS J.-P., GUILLET A. & LECLÈRE C. (1976). *La structure des phrases simples en français. I : Constructions intransitives*. Droz, Genève.
- BRENT M. (1991). Automatic acquisition of subcategorisation frames from untagged text. In *Proceedings of the 29th Meeting of the ACL*, p. 209–214, Berkeley.
- BRISCOE T. & CARROLL J. (1997). Automatic extraction of subcategorisation from corpora. In *Proceedings of the 5th ANLP conference*, p. 356–363.
- CARROLL J. & FANG A. (2004). The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP)*, p. 107–114, Sanya City, China.
- CLÉMENT L., SAGOT B. & LANG B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of LREC'04*, Lisbonne.
- FAST J. & PRZEPIÓRKOWSKI A. (2005). Automatic extraction of polish verb subcategorisation. an evaluation of common statistics. In *Proceedings of the 2nd Language and Technology conference*, p. 191–195.
- FRANCOPOULO G. (2005). Tagparser et technolangu-easy. In *Actes de l'atelier Easy, TALN*.
- GARDENT C., GUILLAUME B., PERRIER G. & FALK I. (2005a). Extracting subcategorisation information from Maurice Gross' Grammar Lexicon. *Archives of Control Sciences*, 15(LI), 253–264.
- GARDENT C., GUILLAUME B., PERRIER G. & FALK I. (2005b). Maurice gross' grammar lexicon and natural language processing. In *Proceedings of the 2nd Language and Technology Conference*, Poznan, Poland.
- GARDENT C., GUILLAUME B., PERRIER G. & FALK I. (2006). Extraction d'information de sous-catégorisation à partir des tables du ladl. In *Actes de La 13ème édition de la conférence sur le Traitement Automatique des Langues Naturelles (TALN 2006)*.
- GROSS M. (1975). *Méthodes en syntaxe*. Hermann.
- GUILLET A. & LECLÈRE C. (1992). *La structure des phrases simples en français. Constructions transitives locatives*. Droz, Genève.
- KORHONEN A. (2002). *Subcategorization Acquisition*. PhD thesis, University of Cambridge.
- MANNING C. (1993). Automatic acquisition of a large subcategorisation dictionary from corpora. In *Proceedings of the 31st Meeting of the ACL*.
- ROMARY L., SALMON-ALT S. & FRANCOPOULO G. (2004). Standards going concrete : from LMF to morphalou. In *Workshop on Electronic Dictionaries*, Geneva, Switzerland.
- VAN DEN EYNDE K. & MERTENS P. (2003). La valence : l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies* 13, 63-104.