Méthodologie lexicographique de constitution d'un lexique syntaxique de référence pour le français

Laurence Danlos, Benoît Sagot

Méthodologie lexicographique de constitution d'un lexique syntaxique de référence pour le français

LEXICOGRAPHIE ET INFORMATIQUE : BILAN ET PERSPECTIVES

Colloque international à l'occasion du 50e anniversaire du lancement du projet du *Trésor de la Langue Française*

Dates : 23-25 janvier 2008

Lieu : Nancy, Campus Lettres et Sciences Humaines

Organisation : UMR ATILF (CNRS/Nancy-Université)

Benoît Sagot (1) benoit.sagot@inria.fr

Laurence Danlos (1) danlos@linguist.jussieu.fr

(1) Projet ALPAGE – INRIA Rocquencourt et Université Paris 7

Mots-clés : Lexiques syntaxique, lexique de référence, validation manuelle, validation sur corpus

Keywords: Syntactic lexicon, reference lexicon, manual validation, corpus validation

Résumé : Cet article présente une méthodologie lexicographique originale pour la constitution d'un lexique syntaxique de référence pour le français. Cette méthodologie se propose de fusionner les informations lexicales issues des principales ressources existant à l'heure actuelle et ensuite de les valider manuellement et sur corpus, dans la continuité de travaux récents déjà engagés sur trois d'entre elles. L'ensemble du processus fait donc appel à une double validation - à l'aide d'interfaces adaptées - reposant à la fois sur des travaux de lexicographie effectués par des linguistes et sur l'exploitation des résultats d'analyseurs de surface et d'analyseurs syntaxiques profonds traitant des corpus volumineux, textuels ou dictionnairiques.

Abstract : This paper introduces an novel lexicographic methodology whose goal is the development of a reference syntactic lexicon for French. This methodology proposes to gather and validate both manually and on corpora lexical information coming from the main existing resources, following recent works on three of them. The proposed process relies on a double validation with appropriate user interfaces: on the one hand thanks to lexicographic work by linguists, and on the other hand by the exploitation of automatically parsed textual and dictionary corpora.

Introduction

Qu'il s'agisse de travaux en linguistique ou en traitement automatique des langues (TAL), l'analyse et l'exploitation manuelle ou automatique de corpus de textes rédigés en français souffrent de la non-existence d'un lexique électronique couvrant pour le français courant, disposant d'informations linguistiques riches et qui soit librement distribué. Un certain nombre de ressources lexicales pour le français existent pourtant, mais elles ne sont que partiellement satisfaisantes. Certaines se limitent aux informations morphologiques (e.g Multex, Morphalou, DELAS) ou à des informations syntaxiques limitées (TLFi). D'autres (présentées dans les Sections 1 et 2) contiennent des informations syntaxiques plus riches qui ont été soit compilées par des linguistes sur la base de l'introspection soit acquises (semi-)automatiquement à partir de corpus. Les ressources issues de travaux linguistiques présentent les défauts inhérents à la méthode introspective (couverture et tolérance relativement arbitraires) et ne sont souvent pas directement exploitables dans des applications automatiques. À l'inverse, les ressources développées à partir de techniques d'apprentissage sur corpus pâtissent d'un manque de formalisation linguistique et de validation manuelle.

Nous proposons ici une méthodologie lexicographique pour la constitution d'un lexique morphologique et syntaxique de référence du français qui cherche à pallier les limites des deux types d'approches rappelées ci-dessus. L'idée est de fusionner dans un modèle lexical consensuel des informations extraites des principales ressources déjà existantes, et de confronter le résultat à la fois à une validation linguistique manuelle et à une validation sur corpus.

Nous montrons l'intérêt de cette démarche en décrivant (Section 1) des travaux déjà effectués sur la conception d'un modèle lexical et sur son utilisation pour rassembler au sein du Le*fff* (Lexique des Formes Fléchies du Français) des informations provenant d'autres ressources. Nous décrivons ensuite la façon dont nous envisageons d'étendre et de compléter cette démarche à un éventail plus large de ressources (Section 2). Nous montrons dans la Section 3 la façon dont interagiront des travaux linguistiques de validation manuelle et des procédures de validation grâce à des systèmes automatiques d'analyse de corpus textuels et dictionnairiques.

1 Modèle lexical et enrichissement du Le*fff*

La méthodologie lexicographique que nous proposons repose sur un modèle lexical unique et consensuel dans lequel on peut représenter de façon linguistiquement satisfaisante et automatiquement exploitable les informations jugées nécessaires. Nous décrivons ici une version préliminaire de ce modèle, qui est utilisée dans le Le*fff*. Nous présentons ensuite des travaux effectués sur la comparaison et la fusion d'informations lexicales en vue de l'enrichissement du Le*fff*, travaux qui préfigurent la méthodologie proposée dans cet article.

1.1 Modèle lexical

Le modèle lexical utilisé dans la version actuelle du Le*fff*, lexique électronique du français courant (520 000 entrées) librement disponible (Sagot et al. 2006, et www.lefff.net), est une version préliminaire du modèle lexical que nous souhaitons exploiter pour mettre en œuvre la méthodologie décrite dans cet article. Il est issu pour partie de travaux réalisés au sein du projet LexSynt, mené au sein de l'ILF (Institut de Linguistique Française) et dirigé par Sylvain Kahane de 2005 à 2007. L'objectif de LexSynt était de commencer à faire coopérer plusieurs équipes de recherche francophones spécialistes de lexicologie, de modélisation d'informations linguistiques (tant pour les lexiques que pour les formalismes grammaticaux) et de TAL, ces derniers cherchant à coupler dans un même programme informatique lexiques et grammaires.

Parmi les avancées de LexSynt, et grâce à une comparaison entre diverses ressources lexicales [Danlos et Sagot 2007, Sagot et Danlos 2007], il a été mis à jour un consensus sur les informations syntaxiques caractérisant un emploi de verbe simple et plein du français courant. Ces informations syntaxiques donnent le cadre de sous-catégorisation (la valence, la rection) d'un emploi de verbe plein sous forme de liste de fonctions syntaxiques (Suj, Obj, Objà, etc.)[1] assorties de leurs diverses réalisations de surface (réalisations sous forme de groupe nominal, de groupe adjectival, de complétive, d'interrogative indirecte, d'infinitive ou de clitique, chacune de ces réalisations pouvant être précédées d'une préposition). Soulignons bien que ce modèle lexical se veut indépendant des choix théoriques de ses utilisateurs, et en particulier des théories syntaxiques. C'est ce qui permet au Le*fff* d'être utilisé dans divers outils de TAL (e.g. analyseurs syntaxiques reposant sur TAG [Thomasset et de La Clergerie, 2005] ou sur LFG [Boullier et Sagot, 2005]).

À titre d'illustration, voici deux entrées extraites du Le*fff* pour le verbe *apprendre* dans une construction active standard, ainsi qu'une entrée pronominale (dite « à agent fantôme ») :

Luc apprend la conduite / à conduire à Marie
apprend v [pred='apprendre_teach_1<**Suj**:sn|cln, **Obj**:(sn|cla|à-sinf|scompl|qcompl), **Objà**:(à-sn|cld)>',
cat=v, @personnel, @CtrlObjàObj, @CompInd, @P3s]

Marie apprend la conduite / à conduire
apprend v [pred='apprendre_learn_2<**Suj**:sn|cln, **Obj**:(sn|cla|à-sinf|scompl|qcompl)>',
cat=v, @personnel, @CtrlObjàObj, @CompInd, @P3s]

La conduite / conduire s'apprend facilement
apprend v [pred='apprendre_learn_2<**Suj**:sn|cln|sinf>',
cat=v, @personnel, @pronominal, @P3s]

On y distingue un identifiant sémantique de l'entrée « intensionnelle » correspondante (i.e. l'entrée factorisée de niveau lemme à partir de laquelle sont générées automatiquement les entrées pour chaque forme fléchie du

[1] L'inventaire des fonctions syntaxiques ainsi que leurs critères définitoires ont de nombreux points communs avec les « paradigmes » de Dicovalence, ressource décrite dans la Section 2.

lemme[2]), la liste (entre chevrons) de fonctions syntaxiques assorties de leurs réalisations, et un ensemble de couples attributs-valeurs — souvent factorisés par le biais d'une « macro » préfixée par @ — pour décrire les autres informations morphosyntaxiques (catégorie syntaxique, pronominalité, impersonnalité, contrôles, attributifs, traits morphologiques, etc.).

Ce modèle consensuel issu du projet LexSynt est limité à plusieurs titres. En particulier, la modélisation des entrées autres que les verbes pleins et simples n'a pas été étudiée en détail. Toutefois, c'est en partie grâce à ce modèle que le Le*fff* est utilisé dans divers outils de TAL, et qu'il a pu être enrichi comme décrit ci-dessous.

1.2 Enrichissement du Le*fff* à l'aide d'autres ressources

Des travaux récents sur le Le*fff* ont initié le projet décrit ici. À ce jour, ces travaux ont couvert les constructions impersonnelles [Sagot et Danlos 2007] suite aux travaux de [Danlos 2005], certaines expressions verbales figées (de façon préliminaire) [Danlos *et al.* 2006], et les adverbes en *–ment* [Sagot et Fort 2007]. Ils se sont penchés sur l'extraction, manuelle ou automatique, d'informations du lexique-grammaire (qui est décrit brièvement dans la Section 2). L'intégration de ces informations dans le Le*fff* a été validé par comparaison avec Dicovalence (voir Section 2) en ce qui concerne les constructions impersonnelles [Danlos et Sagot 2007].

Ces travaux, dont les résultats sont encourageants, ont montré qu'il est nécessaire, malgré la richesse du lexique-grammaire, de procéder à un double travail de linguistique et de modélisation, afin d'exploiter son contenu dans un lexique destiné au TAL. Ces travaux préfigurent les phases de fusion et de validation manuelle décrites ci-dessous. Mais il leur manque à la fois une plus grande variété de ressources de départ, une modélisation plus fine des phénomènes, une validation manuelle plus poussée, et une validation sur corpus plus systématique.

2 Intégration de ressources lexicales

2.1 Principales ressources existantes disponibles

La méthodologie décrite dans cet article repose sur la disponibilité (c'est-à-dire l'existence et la libre diffusion) d'un certain nombre de ressources lexicales pour le français, qui, bien que de nature, d'origine, de couverture et de qualité variées, constituent toutes ensemble un corpus de données morphologiques et syntaxiques très important. Parmi les principales ressources lexicales disponibles pour le français, outre le Le*fff* décrit ci-dessus, on peut citer :

- Dicovalence : Ce dictionnaire [van den Eynde & Mertens 2006] est une ressource informatique qui répertorie les cadres de valence de plus de 3 700 verbes simples du français et qui contient plus de 8 000 entrées. Sa particularité réside dans le fait que les informations valencielles sont définies selon les principes de « l'Approche Pronominale » [van den Eynde & Blanche-Benveniste 1978]. Pour chaque place de valence (appelée « paradigme »), Dicovalence précise le paradigme de pronoms qui y est associé et qui couvre « en intention » les lexicalisations possibles. Ensuite, la délimitation d'un cadre de valence, appelé « formulation », repose non seulement sur la configuration (nombre, nature, caractère facultatif, composition) de ces paradigmes pronominaux, mais aussi sur les autres propriétés de construction associées à cette configuration, comme les « reformulations » passives.
- SynLex : Cette ressource [Gardent *et al.* 2006] est issue d'une conversion automatique du sous-ensemble du lexique-grammaire des verbes qui est disponible[3] vers un format mieux approprié pour le TAL. Le lexique Synlex contient 28 000 cadres de sous-catégorisation pour 4 100 verbes.
- DiCo: Le Dictionnaire de Combinatoire [Polguère 2003] est une base de données lexicale du français, développée à l'OLST (Observatoire de Linguistique Sens-Texte de l'Université de Montréal) par Igor Mel'čuk et Alain Polguère. La finalité première de cette ressource est de décrire chaque entrée (« lexie ») selon deux axes : les dérivations sémantiques (relations sémantiques fortes) qui la lient à d'autres lexies de la langue et les collocations (expressions semi-idiomatiques) qu'elle contrôle. Cette description s'accompagne d'une modélisation des structures syntaxiques régies par la lexie et d'une modélisation de son sens, sous

[2] Ce processus de génération comporte une phase de flexion, en fonction de la classe morphologique associée à l'entrée intensionnelle, puis une phase de construction de la structure syntaxique associée à chacune des formes fléchie obtenues (les informations syntaxiques variant d'une forme à une autre, en particulier pour les formes infinitives et participiales).

[3] Le lexique-grammaire a été développé au LADL (Laboratoire d'Automatique Documentaire et Linguistique) sous la direction de Maurice Gross. Il contient des données électroniques extensives sur les propriétés morphosyntaxiques des foncteurs du français (verbes, noms, adjectifs, adverbes). Il est actuellement maintenu et développé à l'IGM (Institut Gaspard Monge) sous la direction d'Eric Laporte. Seules 60% des données du lexique-grammaire sont actuellement diffusées librement.

forme d'étiquetage sémantique. Actuellement, le DiCo inclut 1075 lexies (acceptions) et 25 540 liens lexicaux.

- DicoLPL : Ce lexique du LPL (Laboratoire Parole et Langage) [van Rullen *et al.* 2005] contient 580 000 formes fléchies. Il décrit pour chaque entrée ses traits morphologiques, sa forme phonétisée, sa fréquence et le lemme sous-jacent. Les verbes contiennent quant à eux quelques informations concernant la sous-catégorisation. DicoLPL a été constitué sur la base d'un lexique interne au LPL et complété par croisement de ressources existantes et vérification sur corpus.

2.2 Fusion des informations lexicales

L'intégration des ressources que nous venons de décrire pour former un point de départ au lexique syntaxique de référence se déroule en deux étapes successives : l'adaptation de ces ressources au modèle lexical décrit à la section 1.1 et la fusion de ces ressources dérivées en une ressource préliminaire unique.

L'étape de conversion est délicate, car elle nécessite une interprétation des données présentes dans chaque ressource. Ainsi, notre modèle lexical fait appel à la notion de fonction syntaxique alors que la plupart des ressources lexicales citées ne font pas directement appel à cette notion. Il faut donc « traduire » les informations codées dans ces ressources en termes de fonctions syntaxiques.

Une fois les conversions effectuées, il faut fusionner les ressources dérivées en un seul lexique qui constitue une version préliminaire du lexique de référence. La difficulté majeure de cette étape est la gestion des conflits. Certains devraient pouvoir se résoudre assez facilement. Par exemple, si un actant d'un verbe est marqué comme obligatoire dans une ressource et facultatif dans une autre, il est aisé d'indiquer que le statut obligatoire ou facultatif de cet actant est non déterminé dans cette version préliminaire. D'autres conflits vont demander plus de réflexion. Nous pensons entre autres à la séparation des entrées pour un même lemme. Si un lemme a un certain nombre d'entrées dans une des ressources et un autre nombre d'entrées dans une autre, l'identification des entrées similaires est aisée, mais que faire des autres ? À partir de quel moment va-t-on considérer que deux entrées sont à fusionner, ou, à l'inverse, à distinguer ? Une différence telle que le statut obligatoire ou facultatif d'un actant ne devrait pas compter (voir ci-dessus) ; à l'inverse, une différence de fonctions syntaxiques devrait compter, mais qu'en est-il d'une différence de réalisations de surface des fonctions syntaxiques ? En résumé, cette phase de fusion demande un gros travail collaboratif de lexicologie permettant de mettre au point des heuristiques de résolution automatique des conflits les plus courants.

3 Double validation

Un des objectifs majeurs de la méthodologie proposée est de construire une ressource lexicale dont la pertinence linguistique soit ancrée à la fois dans l'introspection linguistique *et* dans la validation sur corpus. Nous considérons en effet que les avantages et inconvénients de ces deux types d'approches sont complémentaires.

La validation par introspection est confiée à des linguistes qui corrigent les entrées de la version préliminaire du lexique de référence, par exemple en établissant des critères pour gérer les conflits mal traités automatiquement lors de la phase de fusion et en appliquant ces critères de façon systématique. Ce travail linguistique sera guidé par des informations extraites de corpus étiquetés ou analysés superficiellement, ainsi que par des recherches dans des dictionnaires électroniques comme le TLFi. L'objectif de cette validation manuelle est de savoir si telle ou telle entrée est bien attestée en corpus ou, au contraire, de s'apercevoir qu'une entrée est manquante.

L'objectif de la seconde validation est de vérifier et de probabiliser les entrées du lexique de référence par son utilisation dans différents analyseurs syntaxiques. En effet, les recherches actuelles en analyse syntaxique, en particulier dans les pays anglo-saxons, montrent qu'il est indispensable que les descriptions linguistiques utilisées en TAL soient ancrées dans la réalité de corpus concrets, notamment à des fins de validation, de probabilisation et d'évolution dynamique permanente. Dans un premier temps, il est raisonnable de se restreindre aux deux premiers de ces aspects. Pour chaque analyseur syntaxique concerné, il faut en premier lieu adapter l'analyseur afin qu'il puisse fonctionner avec un lexique au format de la ressource. Il faut ensuite procéder à l'analyse syntaxique de corpus variés et volumineux (plusieurs dizaines de millions de mots) pour :

- identifier les entrées lexicales qui semblent superflues, car n'entrant (presque) jamais dans l'analyse d'aucune phrase par (presque) aucun analyseur ;
- identifier les entrées lexicales probablement erronées ou incomplètes, à l'aide de techniques de fouille d'erreurs dans les résultats d'analyseurs syntaxique telles que celles de [Sagot et de La Clergerie, 2006] ;
- probabiliser les entrées lexicales et les structures syntaxiques qu'elles comportent (fréquence de réalisation des fonctions syntaxiques, fréquences de chaque type de réalisation, etc.) ;

– fournir (si besoin est) des exemples en corpus pour illustrer les entrées lexicales.

La double validation décrite ci-dessus ne peut se faire de façon précise et efficace qu'à l'aide d'un environnement de lexicographie permettant de visualiser aisément les entrées d'un lexique, de les corriger selon des protocoles bien établis, et de connaître leur degré et type de validation. Il faut en effet permettre l'accès à toutes les informations permettant d'accélérer la validation manuelle tout en visualisant ou intégrant les résultats des analyseurs syntaxiques. De plus, dans un contexte de travail collaboratif, il est important de garder une trace des modifications effectuées, de leurs auteurs, voire de commentaires associés pour justifier une décision.

4 Conclusion

Nous avons présenté dans cet article une méthodologie lexicographique originale, qui fait suite à des travaux prometteurs, et dont le but est d'exploiter au mieux l'existant afin de faire franchir un cap décisif aux ressources disponibles pour le français. Les travaux lexicographique sur le français sont en effet nombreux, mais leur exploitation est en retard par rapport à ce qui se passe pour d'autres langues — et pas seulement pour l'anglais — en raison de l'éparpillement des informations lexicales entre différentes ressources et, parfois, de leur manque de formalisation, en particulier dans une optique de TAL. La méthodologie proposée sera mise en œuvre dans les prochaines années et devrait déboucher sur une vraie ressource syntaxique de référence pour le français, qui sera précise, couvrante, linguistiquement pertinente et directement exploitable et exploitée, tant par des linguistes que dans des outils automatiques.

Bibliographie

[Boullier et Sagot, 2005] Boullier, P. et Sagot, B. (2005) : *Analyse syntaxique profonde à grande échelle : SXLFG,* Traitement Automatique des Langues n°46/2.

[Danlos *et al.*, 2006] Danlos, L., Sagot, B. et Salmon-Alt, S. (2006) : *French frozen verbal expressions: from lexicon-grammar tables to NLP applications,* Actes du Colloque Lexique Grammaire 2006, Palerme, Italie.

[Danlos et Sagot, 2007] Danlos, L. et Sagot, B. (2007) : *Comparaison du Lexique-grammaire et de Dicovalence : vers une intégration dans le Lefff,* Actes de TALN 2007, Toulouse, France.

[Danlos, 2005] Danlos, L. (2005) : *ILIMP : Outil pour reprérer les occurrences du pronom impersonnel* il*,* Actes de TALN 2005, Dourdan, France.

[Gardent *et al.*, 2006] Gardent, G., Guillaume, B., Perrier, G. et Falk, I. (2006) : *Extraction d'information de sous-catégorisation à partir des tables du LADL,* Actes de TALN 2006, Louvain, Belgique.

[Polguère, 2003] Polguère, A. (2003) : *Étiquetage sémantique des lexies dans la base de données DiCo,* Traitement Automatique des Langues n°44/2.

[Sagot *et al.*, 2006] Sagot, B., Clément, L., de La Clergerie, É et Boullier, P. (2006) : *The Lefff 2 syntactic lexicon for French : architecture, acquisition, use,* Actes de LREC 2006, Gênes, Italie.

[Sagot et Danlos, 2007] Sagot, B. et Danlos, L. (2007) : *Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire. Constructions impersonnelles,* Cahiers du Cental, Louvain, Belgique.

[Sagot et de La Clergerie, 2006] Sagot, B. et de La Clergerie, É. (2006) : *Error mining in parsing results,* Actes de ACL-CoLing 2006, Sydney, Australie.

[Sagot et Fort, 2007] Sagot, B. et Fort, K. (2007) : *Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire. Adverbes en* -ment*,* Actes du Colloque Lexique Grammaire 2007, Bonifacio, France (à paraître).

[Thomasset et de La Clergerie, 2005] Thomasset, F. et de La Clergerie, É. (2005) : *Comment obtenir plus des méta-grammaires,* Actes de TALN 2005, Dourdan, France.

[van den Eynde et Blanche-Benveniste, 1978] van den Eynde, K. et Blanche-Benveniste, C. (1978) : *Syntaxe et mécanismes descriptifs : présentation de l'approche pronominale,* Cahiers de Lexicologie n°32 (pages 3–27).

[van den Eynde et Mertens, 2006] van den Eynde, K. et Mertens, P. (2007) : *Le dictionnaire de valence Dicovalence : manuel d'utilisation (version 1.2),* en ligne à l'adresse http://bach.arts.kuleuven.be/dicovalence/.

[van Rullen *et al.*, 2005] van Rullen, T., Blache, P., Portes, C., Rauzy, S., Maeyheux, J.-F., Guénot, M.-L., Balfourier, J.-M. et Bellengier, E. (2005) : *Une plateforme pour l'acquisition, la maintenance et la validation de ressources lexicales,* Actes de TALN 2005, Dourdan, France.