Les tables du Lexique-Grammaire au format TAL

Elsa Tolone[1]

1 : Institut Gaspard-Monge, Université Paris-Est, Cité Descartes, 5 bd Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2 - France.

Contact : elsa.tolone@univ-paris-est.fr

Résumé
Les tables du Lexique-Grammaire constituent un lexique syntaxique très riche pour le français. Cette base de données linguistique est cependant inexploitable informatiquement car elle est incomplète et manque de cohérence. Notre objectif est d'adapter les tables pour les rendre utilisables dans diverses applications de Traitement Automatique des Langues (TAL). Nous expliquons les problèmes rencontrés et les méthodes adoptées pour permettre de les intégrer dans un analyseur syntaxique.

Abstract
Lexicon-Grammar tables are a very rich syntactic lexicon for the French language. This linguistic database is nevertheless not suitable for use by computer programs, as it is incomplete and lacks consistency. Our goal is to adapt the tables, so as to make them usable in various Natural Language Processing (NLP) applications. We describe the problems we encountered and the approaches we followed to enable their integration into a parser.

Mots-clés : Traitement Automatique des Langues, lexique, Lexique-Grammaire, analyse syntaxique

Keywords: Natural Language Processing, lexicon, Lexicon-Grammar, parsing

1. Introduction

Dans le domaine du Traitement Automatique des Langues (TAL), l'analyse syntaxique constitue un point clé dans un grand nombre de traitements automatiques tels que la compréhension de texte, l'extraction d'information ou la traduction. Le but d'un analyseur syntaxique est de pouvoir construire la structure grammaticale d'une phrase, ce qui est une tâche difficile, en raison de la complexité et de la richesse de la langue. Pour simplifier, on peut classer les différentes approches en deux catégories :

– analyseurs symboliques utilisant une grammaire et/ou un lexique développés manuellement ;
– analyseurs probabilistes reposant sur un modèle acquis à partir d'un corpus annoté manuellement.

L'approche symbolique bien que laborieuse puisque les ressources sont développées entièrement à la main, permet de construire une base très riche d'informations linguistiques. Il s'agit notamment de décrire les caractéristiques grammaticales des mots, même si représenter toutes ces données est difficile. Cela induit des modifications coûteuses si les bons choix de formalisation ne sont pas réalisés dès le départ. C'est dans ce contexte que nous nous plaçons, l'objectif étant de montrer comment nous avons rendue cohérente et complété une base de données lexicale, les tables du Lexique-Grammaire, afin d'en faire une ressource utilisable à terme dans les applications de TAL.

L'article est organisé de la façon suivante. Dans la section 2, nous présentons tout d'abord les tables du Lexique-Grammaire, qui sont les données sur lesquelles nous travaillons. Nous listons dans la section 3 les problèmes rencontrés liés à ces tables, ainsi que les différentes solutions apportées et celles restant à faire. La section 4 explique comment ces tables sont effectivement utilisées dans un processus d'analyse syntaxique symbolique complet. Enfin, nous terminons par une discussion à la section 5 sur les avantages de cette méthode par rapport à l'approche probabiliste.

2. Contexte

Les tables du Lexique-Grammaire constituent aujourd'hui une des principales sources d'informations lexicales syntaxiques pour le français. Leur développement a été initié dès les années 1970 par Maurice Gross, au sein du LADL (Laboratoire d'Automatique Documentaire et Linguistique) puis de l'IGM (Université Paris-Est) [2, 13, 16]. Ces informations se présentent sous la forme de *tables*. Chaque table correspond à une *classe* qui regroupe les éléments lexicaux d'une catégorie grammaticale donnée (verbes, noms, adjectifs, etc.) partageant certaines propriétés. Une table se présente sous forme de matrice : en lignes, les éléments de la classe correspondante ; en colonnes, les propriétés qui ne sont pas forcément respectées par tous les éléments de la classe ; à la croisée d'une ligne et d'une colonne le signe + ou − selon que l'entrée lexicale décrite par la ligne accepte ou non la propriété décrite par la colonne.
À titre d'exemple, la Fig. 1 montre un extrait de la classe 33 des verbes qui se construisent avec un argument introduit par la préposition à. Si un verbe a deux sens distincts, il possède deux entrées lexicales puisque chaque sens n'accepte pas le même ensemble de propriétés. Un des exemples qui figure dans la classe 33 est le verbe se rendre : Le caporal s'est rendu à l'ennemi et Max s'est rendu à mon opinion.

N0 =: Nhum	N0 =: N-hum	N0 =: Nnr	Ppv	Ppv =: se figé	Ppv =: en figé	Ppv =: les figé	Nég	<ENT>	N0 V	N0 être V-ant	N1 =: Nhum	N1 =: N-hum	N1 =: le fait Qu P	Ppv =: lui	Ppv =: y	N0hum V W sur ce point	[extrap]	<OPT>
+	-	-	<E>	-	-	-	-	renaître	+	+	-	+	-	-	+	-	-	Max renaît au bonheur de vivre
+	-	-	se	+	-	-	-	rendre	+	-	+	+	+	-	+	+	+	Max s'est rendu à mon opinion
+	-	-	se	+	-	-	-	rendre	+	-	+	-	-	-	-	-	-	Le caporal s'est rendu à l'ennemi
+	-	-	<E>	-	-	-	-	renoncer	-	-	+	+	-	-	+	-	-	Max renonce à son héritage

FIG. 1 – Extrait de la classe 33 des verbes

Il existe 61 tables (et donc classes) de verbes simples, la catégorie la mieux décrite, 32 tables d'adverbes (adverbes en *-ment* et locutions adverbiales), 59 tables de noms prédicatifs (noms avec argument(s) qui sont étudiés avec leur verbe support)[1] et 65 tables d'expressions figées. Une partie est téléchargeable sous une licence libre (LGPL-LR)[2].
Nous avons commencé par rassembler les informations sur les tables actuellement développées, avec leurs auteurs principaux, le nombre de tables, leur support papier et les différents formats auxquels elles ont été converties, dans un inventaire le plus complet possible donné à la Tab. 1.
À l'heure actuelle, la plupart des tables constituent autant de fichiers *Excel*, qui ne forment pas véritablement un tout cohérent. Outre les entrées non encore codées (des signes ~ remplacent alors les + et les −), ces tables souffrent de diverses formes d'incohérence et d'incomplétude :
– certains intitulés de colonnes diffèrent d'une table à l'autre bien qu'ils dénotent la même propriété linguistique ;
– certains intitulés de colonnes ne dénotent pas la même propriété d'une table à l'autre ;
– certains intitulés de colonnes laissent implicites des informations pourtant nécessaires à leur exploitation automatique ;
– les propriétés définitoires ne sont pas représentées dans les tables.

[1] Dans les phrases à verbe support, ce n'est pas le verbe qui remplit la fonction de prédicat de la phrase, mais un nom prédicatif (Luc monte une attaque contre le fort), un adjectif prédicatif (Luc est fidèle à ses idées), etc. La distribution du sujet et éventuellement des compléments essentiels dépend de cet élément prédicatif.
[2] `http://infolingu.univ-mlv.fr`, Données Linguistiques > Lexique-Grammaire > Visualisation

Auteurs et références	Nom des tables	Nb	Origine	Format Excel	Table des classes	Format *lglex*	Format Le*fff*
(a) Verbes							
M. Gross [13]	1 à 18	18	Livre	OK	OK	OK	OK
J.-P. Boons, A. Guillet, C. Leclère (BGL) [1] [2] [16]	31H à 39	42	Livres/ Rapport	OK	OK	OK	OK
(b) Noms prédicatifs							
J. Giry-Schneider [7] [8]	F1A à F9 et FN à FNPN	31	Livres	OK	OK	OK	/
A. Meunier [19]	AN01 à AN06 et ANSY	7	Thèse	OK	OK	OK	/
G. Gross [12]	AD à IS2	15	Thèse	OK	OK	OK	/
J. Giry-Schneider, A. Balibar-Mrabti [11]	AN07 à AN10	4	Rapport	OK	OK	OK	/
J. Giry-Schneider [10]	ANDN et ANSN	2	Revue	OK	OK	OK	/
D. de Négroni-Peyre [5]	PSY et SYM	2	Revue	En cours	En cours	En cours	/
R. Vivès [26]	APE1 à APE22	6	Thèse	En cours	En cours	En cours	/
J. Labelle [17]	ANA à ANSU	8	Thèse	En cours	En cours	En cours	/
(c) Adverbes (simples et figés)							
C. Molinier, F. Levrier [20]	ADVMF à ADVPS	16	Livre	OK	En cours	/	/
M. Gross [15]	PAC à PV	16	Livre	OK	OK	/	/
(d) Expressions figées							
J.-P. Boons, A. Guillet, C. Leclère (BGL) [2]	31I	1	Livre	OK	En cours	/	/
M. Gross [14]	A12 à YA	64	Non publié	OK	En cours	/	/
L. Danlos [4]	Z à ZS	8	Thèse	En cours	En cours	/	/
(e) Adjectifs non prédicatifs							
E. Laporte [18]	ADJLOCTABLE	1	Revue	OK	/	/	/
J. Giry-Schneider [9]	DEDJA et DADJI	2	Non publié	En cours	/	/	/
(f) Adjectifs prédicatifs							
J. Giry-Schneider	ADJ01 à ADJSYM	38	Non publié	/	/	/	/

TAB. 1 – Inventaire des tables du Lexique-Grammaire et avancement des travaux de conversion vers des formats électroniques exploitables

Un travail de fond est actuellement réalisé à l'IGM (projet *LGTag*) pour traiter ces différents problèmes. L'objectif est de rendre les tables du Lexique-Grammaire exploitables dans un analyseur syntaxique. L'explicitation en détail de ce travail, ainsi que les méthodes adoptées pour résoudre ces problèmes constituent l'objet du présent article.

3. Problèmes et solutions

Les problèmes rencontrés sont dus au fait que ces tables ont été créées durant plus de 30 ans par différentes personnes. Elles font souvent partie d'annexes de thèses, et certaines ont été reprises ensuite dans la publication de livres mais pas toutes. D'autres figurent dans des revues (telles que Lingvisticæ Investigationes ou les Cahiers de Lexicologie) ou des rapports (Rapport de recherche ou Rapport technique du LADL). Chaque personne y a contribué dans le cadre de son travail de recherche, avec sa propre vision et ses propres notations, tout en respectant le même système de codage (+ et −) représentant l'acceptation ou non de propriétés par les entrées.
Des travaux de mise en cohérence et d'explicitation des propriétés inventoriées dans les tables du Lexique-Grammaire ont été mis en place à l'IGM. Nous allons lister les problèmes qui se sont posés lors de cette étape d'homogénéisation des tables.

3.1. Découpage en classes

Chaque classe regroupe un certain nombre d'entrées jugées similaires car elles acceptent des propriétés communes, que l'on appelle les *propriétés définitoires*. Elles sont en général constituées d'au moins une *construction*, dite « de base ». Par exemple, la construction `N0 V à N1` indique que la phrase contient un sujet, suivi du verbe correspondant à l'entrée et d'un complément indirect introduit par la préposition `à`. Par ailleurs, la *propriété distributionnelle* `N0 = : Nhum` spécifie que le sujet doit être de type humain. Avoir les deux propriétés précédentes acceptées en même temps revient à admettre la construction `N0hum V à N1`.

Notons tout d'abord que ce découpage en classes, c'est-à-dire le regroupement de certaines entrées, est en partie arbitraire. En effet, il est possible de prendre en compte des propriétés plus ou moins précises, qui englobent plus ou moins de verbes, et d'obtenir alors un nombre de classes différent. Ainsi, la construction N0hum V à N1 concerne moins d'entrées que la construction N0 V à N1 qui n'a pas de restriction sur le sujet. De plus, d'autres propriétés auraient pu être prises en compte, ce qui aurait amené à un découpage totalement différent. Par ailleurs, plusieurs classes ont la même construction de base N0 V N1 de N2 mais ont été découpées selon d'autres propriétés définitoires pour éviter qu'une même classe possède un trop grand nombre d'entrées. C'est le cas des tables 37M1 à 37M6 de BGL (cf. Tab. 1(a)).
Le choix des propriétés définitoires a été fait individuellement par chaque auteur pour correspondre à la thématique étudiée sans forcément prendre en compte les autres travaux déjà réalisés. Cela a conduit à ce que plusieurs entrées identiques se retrouvent en doublons dans différentes classes. Par exemple, pour les noms (cf. Tab. 1(b)), les tables FR1 à FR3 de G. Gross contiennent des doublons avec les tables de J. Giry-Schneider, ainsi que la table AA de G. Gross avec les tables d'A. Meunier. Pour les verbes (cf. Tab. 1(a)), certaines tables de BGL reprennent parfois des entrées déjà présentes dans les tables de M. Gross.
Les classes qui possèdent des entrées redondantes sont difficiles à scinder, car elles n'étudient pas forcément les mêmes propriétés. Nous avons décidé d'en faire deux entrées distinctes, et donc deux sens distincts, puisqu'un même mot a autant d'entrées que de sens différents, bien que cela ne soit pas forcément le cas. Nous devons donc trouver comment combiner les deux ensembles de propriétés, ou alors choisir de ne prendre en compte qu'une entrée.

3.2. Format Excel

Dans un premier temps, nous avons cherché à obtenir les classes au format électronique. En effet, les classes n'ont pas été créées en un format permettant directement leur utilisation par un programme informatique tel qu'un analyseur syntaxique. D'ailleurs, certaines tables étaient seulement disponibles dans la littérature (dans l'annexe d'ouvrages ou de thèses). Nous les avons scannées (cf. Fig. 2)[3] et leur avons appliqué un outil de reconnaissance de caractères (OCR), qui a nécessité un travail de correction manuelle. Cela a permis d'avoir 16 tables de noms supplémentaires : les tables de D. de Négroni, R. Vivès et J. Labelle (cf. Tab. 1(b)).
Par ailleurs, les colonnes représentant diverses propriétés sont parfois regroupées en familles, voire mises en dépendance les unes par rapport aux autres, comme illustré à la Fig. 2. Cette structuration n'est pas exploitable informatiquement de façon simple, et a été éliminée lors du passage au format Excel. Pour cela, les colonnes concernées ont été renommées lors de leur regroupement, le but étant d'organiser en un tout cohérent l'ensemble des données existantes.

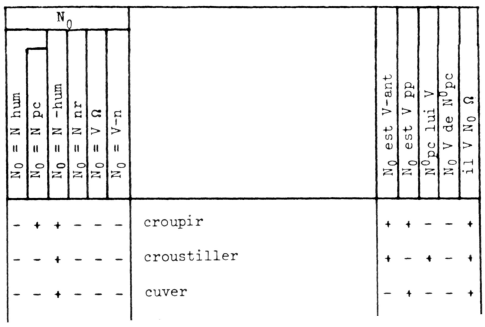

FIG. 2 – Extrait de la table 31R telle que publiée dans la littérature [2]

[3] La table 31R présentée à la Fig. 2 était déjà disponible au format électronique, il n'a donc pas été nécessaire de la scanner, mais elle permet d'illustrer entre autres ce à quoi ressemble une table dans la littérature.

3.3. Propriétés homogènes avec documentation

Notre objectif est de rendre l'ensemble des classes cohérent, car certains intitulés de colonnes peuvent être différents d'une table à l'autre alors qu'ils dénotent la même propriété linguistique. Après les avoir repérés, nous avons choisi une notation commune et effectué les transformations nécessaires.

Certaines différences sont dues simplement à des erreurs d'inattention, ou à des détails qui n'ont pas été comparés à l'existant. C'est ainsi que l'intitulé Det = : E a été remplacé par Det = : <E> pour être identique aux autres tables de noms (cela concerne les tables AN01, ANDN et F1A à F9, cf. Tab. 1(b)). D'autre part, l'intitulé [extrap] présent dans les tables de M. Gross et l'intitulé il V N0 W utilisé par BGL ont la même signification (cf. Tab. 1(a)). Nous avons donc renommé les intitulés de toutes ces colonnes en [extrap]. Il fallait cependant veiller à ne pas abuser de ce type de regroupement pour ne pas perdre une information sous-jacente. Ainsi, les deux intitulés [passif de] et N0 est Vpp W peuvent paraître identiques alors que le deuxième représente les constructions en être avec participe passé à valeur adjectivale statique, ce qui n'est pas le cas du passif.

De plus, certains intitulés de colonnes laissaient implicites des informations pourtant nécessaires à leur exploitation automatique. Ainsi, l'intitulé [pc z.] ("Prép (ce) = zéro") signifie que la préposition et le ce de la complétive (s'il est présent) peuvent être effacés. Le problème est que cet intitulé ne précise pas quel est l'argument concerné par le fait de pouvoir prendre la forme Qu P en plus de la forme Prép (ce) Qu P. Par exemple, dans la table 16 de M. Gross (cf. Tab. 1(a)), les deux arguments sont des complétives (construction de base : N0 V Prép Qu P Prép ce Qu P). Nous avons créé les intitulés Prép N1 = : Prép (ce) Qu P = Qu P et Prép N2 = : Prép (ce) Qu P = Qu P pour expliciter le fait que la propriété concerne l'argument N1 ou N2.

La problématique réside dans le fait que certains intitulés de colonnes ne sont pas clairs :

– soit il manque de la documentation comme c'est le cas des classes d'expressions figées de M. Gross (cf. Tab. 1(d)) puisqu'il n'a pas eu le temps de les publier ;
– soit un même intitulé peut avoir différentes interprétations et représenter une propriété linguistique différente en fonction des tables.

Nous devons donc vérifier pour toutes les classes à quelle signification chaque intitulé fait référence. L'objectif à terme est qu'un intitulé dénote une seule propriété linguistique, qui elle-même ne soit désignée que par un seul intitulé dans l'ensemble des tables. Même si les propriétés sont documentées dans leur ensemble, cette documentation est difficilement accessible : il faut se procurer par exemple la thèse de la table correspondante. C'est pourquoi une documentation des propriétés pour les verbes est en cours de construction à l'IGM.

3.4. Découpage des classes pour établir les propriétés définitoires

En théorie, les entrées sont regroupées dans une classe parce qu'elles acceptent les mêmes propriétés définitoires, mais ce n'était pas toujours le cas. Nous avons procédé à l'éclatement en plusieurs classes quand cela était nécessaire. C'est le cas de la table 2 des verbes de M. Gross (cf. Tab. 1(a)) qui acceptait un N1 pour certaines entrées et pour d'autres non. Or, une propriété dite définitoire pour une classe est considérée comme étant acceptée pour toutes les entrées, sans exception. Nous avons donc créé une nouvelle table 2T regroupant toutes les entrées transitives (acceptant un N1) tout en les supprimant de la table 2. Cela a été fait manuellement pour chaque entrée car aucune indication n'était donnée dans la littérature.

3.5. Tables des classes

Les tables n'étaient pas exploitables informatiquement, c'est-à-dire ne pouvaient pas être intégrées dans les applications de TAL, car les propriétés définitoires n'étaient décrites que dans la littérature. Elles ne figuraient pas dans les tables alors que ce sont des informations essentielles. C'est la raison pour laquelle un travail est en cours à l'IGM : la création de *tables des classes* pour rendre explicites ces informations [3].

Ces tables sont au nombre d'une par catégorie grammaticale. Une table des classes regroupe en colonnes l'ensemble de toutes les propriétés répertoriées pour la catégorie concernée, et liste en lignes l'ensemble des classes définies pour cette même catégorie. À l'intersection d'une ligne et d'une colonne, le signe $+$ (resp. $-$) indique que la propriété correspondante est vérifiée (resp. non vérifiée) par tous les éléments de la classe (c'est-à-dire par toutes les entrées de la table correspon-

dante). Le signe o indique que la propriété est explicitement codée dans la table concernée, car elle est vérifiée par certaines de ses entrées mais pas toutes. Enfin, le signe ? indique une cellule non encore renseignée.
Par exemple, la table des classes des verbes regroupe les 61 classes de verbes et l'ensemble des 490 propriétés. Un extrait de cette table est donné Fig. 3.

table	N0 =: Nhum	N0 =: N-hum	N0 =: Nnr	N0 =: V1-inf W	<ENT>	Ppv =: se figé	N0 V	N0 V N1	zone 1	N0 V à N1	N1 =: Nhum	N1 =: N-hum	N0 V Prep N1 V0-inf W	N0 V N1 V0-inf W	N0 V V0-inf W
V_2	+	-	-	-	o	o	-	-	-	-	-	+	o	o	+
V_4	-	-	+	+	o	-	o	+	-	-	o	o	-	-	-
V_31R	o	o	-	-	o	o	+	-	-	-	-	-	-	-	-
V_31H	+	-	-	-	o	o	+	-	-	-	-	-	-	-	-
V_33	o	o	o	-	o	o	o	-	-	+	o	o	-	-	-
V_32H	o	-	o	-	o	o	-	+	-	-	+	-	-	-	-

FIG. 3 – Extrait de la table des classes des verbes

Ceci fait apparaître toutes les propriétés définitoires et permet également de coder toutes les propriétés pour chaque entrée. N'oublions pas que le choix des propriétés codées dans chaque table a été en partie arbitraire, ce qui implique que toutes les propriétés ne sont pas codées dans toutes les tables. D'autres encore sont seulement décrites dans la littérature et ne sont pas exploitables alors qu'elles peuvent être pertinentes. C'est le cas, par exemple, des *redistributions* [passif par] et [passif de], qui sont fréquentes en français.

3.6. Travail restant à faire

Le codage des tables n'est pas terminé :
– les ? dans la tables des classes correspondent à des propriétés non encore étudiées pour certaines tables. L'importance de certaines d'entre elles empêche l'analyse des constructions correspondantes par un analyseur ;
– des entrées dans les tables des verbes n'ont pas encore été codées et contiennent des signes $\sim$ à la place des $+$ et des $-$. Ces emplois de verbes ne seront pas pris en compte dans un analyseur, alors qu'ils sont parfois essentiels ;
– d'autres entrées sont tout simplement manquantes car tout n'a pas été encore été étudié, surtout dans certaines catégories. C'est le cas des adjectifs prédicatifs, pour lesquels un travail de découpage est en cours d'achèvement (cf. Tab. 1(f)). Cela a permis d'établir 38 classes, mais il reste ensuite à coder un ensemble de propriétés (à définir) pour les entrées retenues. Pour les adjectifs non prédicatifs (cf. Tab. 1(e)), 3 tables seulement existent au format Excel.

En ce qui concerne les verbes, un travail est actuellement en cours pour mettre à jour les entrées. Certains choix ont été fait à une époque où les données linguistiques étaient répertoriées sur des fiches cartonnées, mais n'ont jamais été répertoriés dans les versions électroniques ultérieures. De plus, un index électronique recense toutes les entrées et indique les tables dans lesquelles elles apparaissent en donnant plusieurs exemples de phrases. Il s'agit de mettre en correspondance les tables et cet index.
Cependant, le découpage des verbes très fréquents et pour lesquels il est difficile d'identifier clairement tous les sens a été mis de côté. Une de nos priorités est de les ajouter. Notons également que l'on ne peut pas établir de liste définitive puisque de nouveaux mots apparaissent chaque jour.

Remarquons que les différentes personnes ayant codé les tables peuvent avoir des différences d'interprétation et surtout une rigueur variable. Seulement deux codages sont possibles ($+/-$) mais on peut être laxiste pour les $+$ (c'est-à-dire permettre des phrases presque inacceptables, ce qui permet d'être plus couvrant mais augmente l'ambiguïté) ou au contraire trop intransigeant pour les $-$ (ce qui pose plus de problèmes car ces formes ne pourront jamais être reconnues). Ceci est le problème du codage binaire. Le codage aurait pu être de la forme $++/+/?/-/--$, mais il aurait été plus difficile d'avoir des données cohérentes d'un auteur à l'autre. Une solution possible serait d'indiquer une probabilité d'apparition de chaque construction pour chaque entrée.
Une autre différence de notation pourrait être envisagée pour la table des classes. En effet, tous les signes $-$ n'ont pas la même valeur, mais il n'est pas toujours évident de les distinguer rigoureusement :
– certaines propriétés ont une vraie valeur $-$ car elles ne sont pas valides pour toutes les entrées de la table ;
– d'autres sont codées $-$ lorsqu'elles sont inappropriées car elles dénotent un élément qui n'est pas pertinent pour la table.
Cependant, l'ajout d'un signe / dans la table des classes pour coder une propriété inappropriée ne serait qu'une information linguistique supplémentaire qui n'aurait pas de réel impact dans un analyseur.
L'ensemble de ces travaux de récupération et de mise en cohérence des données linguistiques a permis d'obtenir, du moins pour les verbes, les noms et les adverbes, une nouvelle version des tables du Lexique-Grammaire, qui, combinée avec les tables des classes, constitue un ensemble complet et synthétique de données linguistiques.

4. Explicitation du processus complet d'utilisation des tables

Cette version des tables a permis d'envisager une utilisation de ces données lexicales dans des outils de traitement automatiques. A cette fin, une version textuelle structurée des tables, nommée *lglex*, a été développée pour les verbes et les noms [3].
La sous-partie de ce lexique qui reproduit les entrées verbales et nominales des tables librement distribuées est elle-même librement distribuée, également sous licence LGPL-LR[1]. Des travaux similaires ont été faits dans [6] et une comparaison a déjà été réalisée dans [3].
Le travail étant plus avancé pour les verbes, qui sont par ailleurs le lexique le plus indispensable dans un analyseur, nous avons pu les intégrer dans un analyseur syntaxique à grande échelle, l'analyseur FRMG [25]. Cette intégration a été possible grâce au travail décrit dans [24], par la conversion des tables au format L*efff* [21]. L'analyseur syntaxique FRMG couplé à ce lexique a ensuite été évalué sur le corpus de référence de la campagne EASy [23]. Cela valide l'ensemble de l'approche malgré le caractère préliminaire et partiel des résultats.

5. Conclusion

Après avoir résolu la plupart des problèmes, nous obtenons des données, qui tout en étant les plus correctes possibles, sont riches et détaillées. En effet, malgré les erreurs et les différences de jugements entre personnes, aucune approximation n'a été faite. Bien entendu, les systèmes par acquisition automatique fonctionnent bien et peuvent plus facilement prendre en compte des nouveaux mots. Mais ils comportent beaucoup d'erreurs du fait qu'ils omettent des phénomènes peu ou pas présents dans les corpus d'apprentissage. Il existe des moyens de corriger ces erreurs, notamment au moyen de techniques automatiques telles que celles décrites dans [22], mais cela prend du temps et ne suffit pas pour tout détecter. Remarquons que ces techniques peuvent être également employées pour améliorer les ressources manuelles, en rattrapant des erreurs dans le processus de conversion des tables ou dans les tables elles-mêmes.
L'utilisation d'une ressource lexicale la plus riche possible reste donc un moyen efficace pour améliorer la qualité d'un analyseur syntaxique, comme l'ont montré les travaux décrits dans [22]. Notons tout de même que le fait d'avoir des données aussi complètes augmente l'ambiguïté (c'est-à-dire le nombre d'entrées par lemme) et que les temps d'analyse sont plus élevés. D'autres problèmes restent à résoudre, bien que la méthode manuelle soit longue, nous ne pouvons qu'être encouragés à poursuivre.

Bibliographie

1. Jean-Pierre Boons, Alain Guillet, et Christian Leclère. La structure des phrases simples en français : Classes de constructions transitives. Technical report, LADL, CNRS, Paris 7, 1976.
2. Jean-Pierre Boons, Alain Guillet, et Christian Leclère. *La structure des phrases simples en français : Constructions intransitives*. Droz, Genève, Suisse, 1976.
3. Matthieu Constant et Elsa Tolone. A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables. In *Actes du 27ème Colloque Lexique et Grammaire*, L'Aquila, Italie, 2008.
4. Laurence Danlos. *Représentation d'informations linguistiques : les constructions N être Prép X*. Thèse de doctorat, Université Paris 7, 1980.
5. Dominique de Negroni-Peyre. Nominalisations par être en et réflexivation (admiration, opposition, révolte et rage). *Lingvisticæ Investigationes*, 2(1) :127–164, 1978.
6. Claire Gardent, Bruno Guillaume, Guy Perrier, et Ingrid Falk. Extraction d'information de sous-catégorisation à partir des tables du LADL. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN'06)*, Louvain, Belgique, 2006.
7. Jacqueline Giry-Schneider. *Les nominalisations en français : L'opérateur faire dans le lexique*. Droz, Genève, Suisse, 1978.
8. Jacqueline Giry-Schneider. *Les prédicats nominaux en français : Les phrases simples à verbe support*. Droz, Genève, Suisse, 1987.
9. Jacqueline Giry-Schneider. Les adjectifs intensifs : syntaxe et sémantique. *Cahiers de Lexicologie*, 86(1) :163–178, 2005.
10. Jacqueline Giry-Schneider. Les noms épistémiques et leurs verbes supports. *Lingvisticæ Investigationes*, 27(2) :219–238, 2005.
11. Jacqueline Giry-Schneider et Antoinette Balibar-Mrabti. Classes de noms construits avec avoir. Technical report, LADL, Université Paris 7, 1993.
12. Gaston Gross. *Les constructions converses du français*. Droz, Genève, Suisse, 1989.
13. Maurice Gross. *Méthodes en syntaxe : Régimes des constructions complétives*. Hermann, Paris, France, 1975.
14. Maurice Gross. Une classification des phrases "figées" du français. *Revue Québécoise de Linguistique*, 11(2) :151–185, 1982.
15. Maurice Gross. *Grammaire transformationelle du français : Syntaxe de l'adverbe*, volume 3. Paris : ASSTRIL, 1986.
16. Alain Guillet et Christian Leclère. *La structure des phrases simples en français : Les constructions transitives locatives*. Droz, Genève, Suisse, 1992.
17. Jacques Labelle. *Etude de constructions avec opérateur avoir (nominalisations et extensions)*. Thèse de doctorat, LADL, Université Paris 7, 1974.
18. Éric Laporte. Une classe d'adjectifs de localisation. *Cahiers de Lexicologie*, 86 :145–161, 2005.
19. Annie Meunier. *Nominalisations d'adjectifs par verbes supports*. Thèse de doctorat, LADL, Université Paris 7, 1981.
20. Christian Molinier et Françoise Levrier. *Grammaire des adverbes : description des formes en -ment*. Droz, Genève, Suisse, 2000.
21. Benoît Sagot, Lionel Clément, Éric de La Clergerie, et Pierre Boullier. The Le*fff* 2 syntactic lexicon for French : architecture, acquisition, use. In *Proceedings of the 5th Language Resource and Evaluation Conference (LREC'06)*, Genova, Italie, 2006.
22. Benoît Sagot et Éric de La Clergerie. Fouille d'erreurs sur les sorties d'analyseurs syntaxiques. *Traitement Automatique des Langues (T.A.L.)*, 49(1), 2008.
23. Benoît Sagot et Elsa Tolone. Exploitation des tables du Lexique-Grammaire pour l'analyse syntaxique automatique. In *Actes du 28ème Colloque Lexique et Grammaire*, Bergen, Norvège, 2009. À paraître.
24. Benoît Sagot et Elsa Tolone. Intégrer les tables du Lexique-Grammaire à un analyseur syntaxique robuste à grande échelle. In *Actes de TALN'09 (session poster)*, Senlis, France, 2009.
25. François Thomasset et Éric de La Clergerie. Comment obtenir plus des méta-grammaires. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN'05)*, Dourdan, France, juin 2005.
26. Robert Vivès. *Avoir, prendre, perdre : constructions à verbe support et extensions aspectuelles*. Thèse de doctorat, LADL, Université Paris 7, 1983.