Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire
Constructions impersonnelles

Benoît Sagot[1], Laurence Danlos[2]
INRIA Futurs (projet Signes), Université Paris 7 (Lattice)

Abstract

We intend to develop a large-coverage morphological and syntactic lexicon for French which can be directly used in Natural Langage Processing (NLP) applications, in particular in those involving deep parsing, regardless of the underlying grammatical framework. This lexicon, named Le*fff* (Lexique des Formes Fléchies du Français — Lexicon of French inFlected Forms), has been under development since 2004. At the beginning, this lexicon contained only verbal morphological information, mostly automatically induced from corpora. It now covers all parts of speech, and is progressively enriched with syntactic information. In this paper, we show how we used the lexicon-grammar tables, whose development has been initiated by M. Gross, to enrich the Le*fff*. These tables are a valuable starting point. However, it is necessary to achieve both a linguistic and formal modeling work, in order to exploit their content in a NLP lexicon such as the Le*fff*. We illustrate this approach on one kind of non-standard verbal and adjectival entries : impersonal structures.

Keywords : Lexicon-grammar, Le*fff*, impersonal constructions

Résumé

Nous avons le projet de développer un lexique morphologique et syntaxique du français à large couverture qui soit directement utilisable dans des applications de Traitement Automatique du Langage (TAL), en particulier celles nécessitant une analyse syntaxique profonde, et ce, quel que soit le cadre théorique utilisé. Ce lexique, baptisé Le*fff* (Lexique des Formes Fléchies du Français), est en cours de développement depuis 2004. Au départ, ce lexique ne comportait que des informations morphologiques verbales, principalement induites automatiquement à partir de corpus. Il couvre maintenant toutes les catégories, et est progressivement enrichi d'informations syntaxiques. Nous montrons ici comment nous l'avons enrichi à partir des tables du lexique-grammaire, initialement développées autour de M. Gross. Celles-ci constituent un point de départ d'une valeur inestimable. Il est néanmoins nécessaire de procéder à un double travail de linguistique et de modélisation, afin d'exploiter leur contenu dans un lexique TAL tel que le Le*fff*. Nous illustrons cette approche sur un type particulier d'entrées verbales et adjectivales non standard : les constructions impersonnelles.

Mots-clés : Lexique-grammaire, Le*fff*, constructions impersonnelles

[1]INRIA Futurs, projet Signes, benoit.sagot@inria.fr
[2]Université Paris 7, Institut Universitaire de France, Lattice, laurence.danlos@linguist.jussieu.fr

1. Introduction

L'analyse syntaxique profonde ne peut se faire qu'à la condition de disposer d'informations lexicales riches qui forment un lexique morphologique et syntaxique à large couverture. De plus, un tel lexique doit représenter ces informations d'une façon adaptée à l'utilisation dans les systèmes de traitement automatique. Une telle ressource n'est pas disponible pour le français, malgré de nombreuses initiatives indépendantes, qui ont atteint des degrés d'avancement divers et qui ont privilégié des aspects différents. Deux d'entre elles ont donné naissance à des ressources à large couverture que tout sépare :

- Les tables du lexique-grammaire, dont le développement a été initié par Maurice Gross au LADL (Gross 1975) et se poursuit à l'IGM autour d'Éric Laporte,
- Le Le*fff* (Lexique des Formes Fléchies du Français), lexique morphologique et syntaxique du français à large couverture (520 000 entrées) initié par Lionel Clément, développé par l'un des auteurs et utilisé dans divers systèmes de TAL (Sagot *et al.* 2006).

Il existe, entre autres, une troisième ressource lexicale pour le français, qui ne couvre que les verbes simples, mais dont les fondements linguistiques sont solides. Il s'agit du lexique DICOVALENCE, successeur du dictionnaire PROTON, développé par Karel van den Eynde et Piet Mertens (van den Eynde et Mertens 2006) dans le cadre de l'approche pronominale initiée par Claire Blanche-Benvéniste et Karel van den Eynde (van den Eynde et Blanche-Benvéniste 1978), et mise à la disposition de la communauté. Nous en reparlons plus bas.

L'intérêt principal des tables du lexique-grammaire réside dans leur qualité : elles sont le résultat de travaux minutieux et précis d'investigation linguistique menés depuis les années 1970 dans le cadre d'une équipe CNRS, le LADL. Cependant, les tables ne sont pas directement exploitables dans des systèmes d'analyse, une partie importante des informations n'étant compréhensible que moyennant de nombreuses connaissances implicites. À l'inverse, l'intérêt principal du Le*fff* est que sa structure et son format sont spécifiquement adaptés à une utilisation dans des systèmes de TAL. Mais la qualité, la richesse et la couverture du Le*fff* n'atteignent pas, loin s'en faut, celles des tables du lexique-grammaire.

Nous avons donc étudié le moyen de profiter simultanément de la couverture et de la précision des tables du lexique-grammaire et de l'adéquation du Le*fff* aux traitements automatiques. Parallèlement à d'autres travaux complémentaires, qui cherchent à comparer le Le*fff* à un lexique au même format généré automatiquement à partir des tables du lexique-grammaire (Sagot et Gardent, *en cours*), nous avons travaillé et travaillons à l'utilisation manuelle des tables du lexique-grammaire pour valider, invalider, ou compléter certaines informations du Le*fff*. Ici, nous nous concentrons sur les informations concernant les constructions impersonnelles.

Après une description des tables du lexique-grammaire et des données que nous en avons extraites concernant les constructions impersonnelles, nous présenterons le L*efff* plus en détails et la modélisation qui y est faite des constructions impersonnelles.

2. Le lexique-grammaire : constructions impersonnelles

2.1. Brève introduction au lexique-grammaire des verbes standard

Le lexique-grammaire est composé d'un ensemble de tables : 61 tables pour les verbes (la catégorie la mieux décrite), environ 30 tables pour les adjectifs (travail en cours d'achèvement) et plusieurs tables pour les noms dits « prédicatifs » (noms avec argument(s) qui sont étudiés avec leur verbe support). Chaque table regroupe les éléments d'une catégorie donnée partageant une « propriété définitoire ». Une table se présente sous forme de matrice : en lignes, les éléments lexicaux de la table ; en colonnes, les propriétés qui ne sont pas forcément respectées par tous les éléments de la table ; à la croisée d'une ligne et d'une colonne le signe + ou – en suivant une sémantique évidente.

Les propriétés définitoires relèvent généralement du cadre de sous-catégorisation. Ainsi, les critères les plus communément utilisées dans les propriétés définitoires sont le nombre de compléments, la nature prépositionelle ou non des compléments (pour les compléments prépositionnels, sont distingués ceux qui sont introduits par les préposition *à*, *de*, *avec*, Loc, et autres prépositions), la nature de la réalisation du sujet et des compléments (sont distinguées les réalisations sous forme de complétive, notée Que P, d'infinitive, notée V-inf, et de syntagme nominal, la valeur par défaut, notée N). Par exemple, la propiété définitoire de la Table 9 est N_0 V (Que P)$_1$ à N_2 : cette table regroupe des verbes comme *dire*, *dissimuler* et *ordonner*, dont le cadre de sous-catégorisation peut se caractériser par une complétive objet et un complément nominal introduit par la préposition *à* (*Luc a dit/dissimulé/ordonné à Marie que Zoé chante*)[3].

Une propriété définitoire peut aussi indiquer qu'un élément de la table entre dans deux constructions qui sont généralement reliées par un lien de paraphrase. Ainsi la Table 35S regroupe les verbes intransitifs « symétriques » qui se caractérisent par deux constructions, N_0 V avec N_1, et N_0 et N_1 V — voir *Luc flirte avec Zoé* et *Luc et Zoé flirtent (ensemble)*. Enfin de nombreuses propriétés définitoires incluent des informations sémantiques élémentaires. Par exemple, des informations sur les classes des noms têtes des syntagmes nominaux (humain, concret, pluriel, etc.). Ou encore, des informations sur la sémantique des procès : ainsi les verbes entrant dans la construction N_0 V N_1 de N_2 ont été divisés en deux paquets : la Table 37E regroupe les procès d'enlèvement (*Luc a débarassé le grenier de ses caisses* signifie que *Luc a enlevé les caisses du grenier*), tandis que les tables 37M (tables 37M1, 37M2, …,

[3] Dans la propriété définitoire de la Table 9, le complément indirect apparaît après la complétive. Il n'empêche que cet ordre peut être inversé dans une phrase respectant la propriété définitoire.

37M6) regroupent les procès d'ajout (*Luc a muni la porte d'un verrou* signifie que *Luc a mis un verrou sur la porte*). Les tables 37M*i* se distinguent par des propriétés très diverses (morphologiques, sémantiques ou autres) qui ne relèvent pas du cadre de sous-catégorisation et qui auraient pu/dû figurer en colonne dans une unique table 37M. La raison de ce découpage est principalement numérique : la table 37M aurait regroupé 890 verbes, et il a été considéré que la consultation manuelle d'une matrice de 890 lignes était difficile, d'où sa division en six sous-tables.

Les tables se présentent donc sous forme de matrice de + et – où les colonnes indiquent les propriétés qui varient d'un élément à l'autre. Ainsi, dans la Table 9, une colonne intitulée « de V^2 W » permet de coder si un verbe appartenant à cette table autorise que son complément direct (de position 1) soit une infinitive introduite par le complémentiseur *de* et contrôlée par N_2 (*Luc a ordonné/dit à Zoé de chanter*, versus **Luc a dissimulé à Zoé de chanter*). Une autre colonne intitulée « Aux V^0 W » permet de coder si un verbe appartenant à la Table 9 autorise que son complément direct soit une infinitive directe à un temps composé et contrôlée par N_0[4] (*Luc a dit/dissimulé à Zoé avoir chanté*, versus **Luc a ordonné à Zoé avoir chanté*). En fait, la situation est plus compliquée car il peut y avoir une structure hiérarchique entre colonnes. Ainsi la colonne « de V^2 W » dépend d'une colonne qui indique que la complétive est au subjonctif (*Luc a dit/ordonné à Léa que Zoé parte demain*) tandis que la colonne « Aux V^0 W » dépend d'une colonne qui indique que la complétive est à l'indicatif (*Luc a dit/dissimulé à Léa que Zoé part demain*). De ce fait, pour convertir en un format tel que celui du Le*fff* les informations codées dans la Table 9, il faut comprendre que cette table regroupe (au moins) deux ensembles de verbes : l'un composé de verbes, comme *ordonner*, dont la complétive est au subjonctif et qui permettent une infinitive en de V^2 W, l'autre, comme *dissimuler*, composé de verbes dont la complétive est à l'indicatif[5] et qui permettent une infinitive en Aux V^0 W. Le verbe *dire* appartient à ces deux ensembles.

Ces dépendances complexes entre colonnes, c'est-à-dire entre propriétés syntaxiques, ont d'ailleurs été modélisées (manuellement) pour certaines tables par (Gardent *et al.* 2006), sous forme de graphes. Ces graphes, qui modélisent également les nombreuses informations implicites dans les tables, sont destinées à l'extraction automatique d'un lexique TAL, nommé SynLex, à partir des tables traitées[6].

Nous terminons cette section sur les colonnes en indiquant que diverses colonnes permettent d'indiquer les propriétés de pronominalisation et de cliticisation des différents compléments. Rappelons que ces propriétés peuvent être considérées comme des propriétés définitoires dans le lexique DICOVALENCE (van den Eynde et Mertens 2006). On se reportera à (Danlos et Sagot 2007) pour une comparaison entre le Lexique-Grammaire et DICOVALENCE.

---

[4] Une autre colonne intitulée « V^0 W » permet de coder des phrases comme *Luc dit/prétend être le Messie* qui sont plus naturelles sans complément de type à N_2.

[5] Sans parler de l'induction du subjonctif en mode non-assertif.

[6] On pourra se reporter à la page Internet du projet : http://libresource.inria.fr//projects/SynLex

2.2. Les constructions impersonnelles et l'outil ILIMP

Nous nous intéressons ici aux phrases dont le sujet est le pronom impersonnel *il* (ce pronom est aussi appelé pléonastique ou explétif). Comme la plupart des phénomènes linguistiques, les constructions impersonnelles reposent sur des conditions tant lexicales que syntaxiques. Par exemple, l'adjectif *violet* ne peut jamais être la tête lexicale d'une phrase impersonnelle, (1a), l'adjectif *probable* ancre une phrase impersonnelle lorsqu'il est suivi d'un complément phrastique, (1b), l'adjectif *difficile* ancre une phrase impersonnelle (resp. personnelle) lorsqu'il est suivi d'une infinitive introduite par la préposition *de* (resp. *à*), (1c) et (1d).

(1)a Il est violet

b Il est probable que Fred viendra

c Il est difficile **de** résoudre ce problème

d Il est difficile **à** résoudre [ce problème][7]

De ce fait, le lexique-grammaire du français est une ressource linguistique appropriée pour répertorier l'ensemble des constructions impersonnelles. Ce travail a été effectué lors de la réalisation d'ILIMP (Danlos 2005). ILIMP est un outil qui prend en entrée un texte brut (sans annotation linguistique) rédigé en français et qui fournit en sortie le texte d'entrée où chaque occurrence du pronom *il* est décorée de la balise [ANA] pour anaphorique ou [IMP] pour impersonnel. Cet outil a été conçu en vue de la résolution des anaphores : il permet de distinguer les occurrences anaphoriques du pronom *il*, pour lesquelles un système de résolution des anaphores doit chercher un antécédent, des occurrences où *il* est un pronom impersonnel pour lequel la recherche d'antécédent ne fait pas sens. Il donne 97,5% de bons résultats évalués sur 10.000 occurrences de *il* extraites du journal *Le Monde*[8]. Nous allons voir comment il peut être utilisé dans un lexique syntaxique comme le Le*fff*. Auparavant, présentons-le brièvement.

2.2.1. ILIMP : aspects informatiques

Pour l'aspect informatique, ILIMP repose sur UNITEX[9] qui est un logiciel permettant d'écrire des patrons linguistiques (expressions régulières ou automates récursifs) qui sont localisés dans le texte d'entrée, avec un éventuel ajout d'annotations lorsque les automates sont en fait des transducteurs. Pour ILIMP, l'idée de base consiste à écrire

[7] Un GN entre crochets indique un antécédent possible du sujet anaphorique *il*.

[8] Il existe d'autres outils permettant de reprérer les pronoms impersonnels, en particulier des travaux sur le pronom anglais *it*. Ces outils, qui n'atteignent pas le taux de précision d'ILIMP, utilisent généralement des techniques d'apprentissage, cf. (Boyd et al. 2006). Pour le français en tout cas, les techniques par apprentissage semblent inadéquates dans la mesure où l'ensemble des constructions impersonnelles est stable d'un domaine/genre à l'autre et ne relève pas d'un phénomène productif. Par exemple, on ne va pas observer de nouvelles constructions impersonnelles parce que l'on passe du genre journalistique au domaine aéronautique. On peut donc espérer obtenir une liste complète des constructions impersonnelles, ce qui a été presque réalisé dans ILIMP.

[9] UNITEX est un logiciel sous licence GPL, dont l'ancêtre est INTEX. La documentation et le téléchargement de UNITEX se trouvent sur le site http://ladl.univ-mlv.fr.

(manuellement) un ensemble de transducteurs comme celui présenté en (2) sous une forme linéaire simplifiée. La balise [IMP] est l'ajout d'information amenée par l'aspect transducteur de (2). Les éléments entre chevrons de (2) se glosent de la façon suivante : <être.V:3s> correspond à toutes les formes du verbe *être* conjugué à la troisième personne du singulier, <Adj1:ms> correspond aux adjectifs masculins singuliers de la classe Adj1 qui regroupe des adjectifs se comportant comme *difficile*, <V:W> correspond aux verbes à l'infinitif.

(2) Il [IMP] <être.V:3s> <Adj1:ms> de <V:W>

La balise [IMP] vient décorer les occurrences de *il* qui apparaissent dans les phrases correspondant au patron de (2). Cette balise vient donc décorer *il* dans (1c). La balise [ANA] est la balise par défaut : elle vient décorer les occurrences de *il* qui n'ont pas été balisées par [IMP]. Cette balise vient décorer *il* dans (1d). Néanmoins, la situation est un peu plus complexe, car il existe une troisième balise [AMB] — abréviation de « ambigu » — qui sera expliquée ci-dessous.

2.2.2. Aspects linguistiques : les diffférentes constructions impersonnelles

Pour l'aspect linguistique, ILIMP repose sur le lexique-grammaire. Plus précisément, nous avons extrait **manuellement** du lexique-grammaire tous les items lexicaux qui peuvent ancrer une phrase impersonnelle avec leur complémentation syntaxique. On peut distinguer les constructions intrinsèquement impersonnelles, qui ne peuvent avoir comme sujet que *il,* des constructions avec un « sujet profond extraposé ».

Parmi les premières, on trouve 45 verbes météorologiques de la table 31I de (Boons *et al.*, 1976a) *(Il pleut, Il vente)*, 21 verbes de la table 17 de (Gross, 1975) *(Il faut du pain /que Fred vienne)* et 38 expressions figées de (Gross, 1993) (*Il était une fois, quoi qu'il en soit.)*.

Pour les constructions impersonnelles à sujet profond extraposé, on peut distinguer celles à sujet phrastique de celles à sujet uniquement nominal. Parmi les premières, on trouve 682 adjectifs[10] *(Il est probable que Fred viendra)*, 88 expressions être Prép X des Tables Z5P et Z5D de (Danlos 1980) *(Il est de règle de porter un chapeau)*, 21 verbes de la Table 5 de (Gross 1975) *(Il plaît à Zoé que Luc vienne)*, 140 verbes de la Table 6 et 92 verbes de la Table 9 de (Gross 1975) construits au passif ou au *se*-moyen *(Il a été dit/se raconte que Fred viendra)*, et enfin plus quelques verbes des tables 7 et 8 (voir ci-dessous).

Les constructions impersonnelles à sujet extraposé nominal ont pour tête lexicale des verbes qui sont dispersés dans les tables élaborées par (Boons *et al.* 1976a, 1976b)[11]. On peut distinguer d'un côté des verbes comme *manquer* ou *rester* dont l'emploi en

[10] Le lexique-grammaire des adjectifs n'est pas complet, loin s'en faut. Nous avons extrait manuellement les adjectifs à construction impersonnelle des tables de (Picabia, 1978) et (Meunier, 1981) et complété ces données au fur et à mesure de la réalisation d'ILIMP, sans toutefois atteindre une couverture exhaustive de ces adjectifs.

[11] Dans ces tables, la possibilité d'une construction impersonnelle n'est pas codée.

construction impersonnelle est tout à fait courant *(Il manque/reste du pain)*, et de l'autre côté des verbes « inaccusatifs » (*Il est venu trois personnes)* ou des verbes construits au passif *(Il a été mangé trois gâteaux)*, dont l'emploi dans une construction impersonnelle relève d'un niveau de langue châtié. Seuls les verbes du type *manquer* ou *rester* ont été recensés. Pour ces verbes, le statut impersonnel on non du sujet dépend du déterminant introduisant le GN sujet extraposé, voir la paire en (3), ou du nom tête de ce GN, voir la paire en (4).

(3)a Il manque **du** poivre (dans cette maison)

b Il manque **de** poivre [ce rôti]

(4)a Il reste la **valise** du chef (dans la voiture)

b Il reste la **priorité** du chef [le chômage]

Dans un très petit nombre de cas (une dizaine), un item lexical peut ancrer une construction impersonnelle ou personnelle avec le même cadre de sous-catégorisation. C'est le cas pour l'adjectif *certain* construit avec un complément phrastique, comme illustré dans la phrase en (5a). Comme les deux lectures de (5a) semblent également fréquentes, *il* dans le patron (5b) reçoit la balise [AMB].

(5)a Il est certain que Fred viendra (*Jean/Cela est certain que Fred viendra)*

b Il [AMB] est certain que P

Soulignons bien que l'extraction des constructions impersonnelles à partir du lexique-grammaire a été manuelle, et non automatique comme dans le travail présenté dans (Gardent *et al.* 2006). Cette différence est de taille, comme illustré dans l'exemple qui suit : le cas des verbes à sujet phrastique extraposable qui proviennent des tables de (Gross 1975). Une extraction automatique consisterait à prendre :

- toutes les entrés de la table 5, dont la propriété définitoire est justement l'existence de la construction $(\text{Que P})_0$ V Prép N_1 associée à la construction impersonnelle Il V Prép N_1 Que P (voir *Que Luc parte plaît à Zoé* versus *Il plaît à Zoé que Luc parte*),
- toutes les entrées de la table 17, dont la propriété définitoire est Il V (Prép ce) Que P Prép N_2 (voir *Il semble à Luc que Zoé est partie*),
- les quelques entrées des Tables 7 (N_0 V à ce que P) et 8 (N_0 V de ce que P) où il existe une colonne intitulée [extrap] (dépendant de la colonne intitulé Sujet) qui code la possibilité d'extraposer un sujet phrastique (voir *Que Zoé soit partie découle de ce que Luc est arrivé* versus *Il découle de ce que Luc est arrivé que Zoé soit partie*).

Une telle extraction automatique aurait induit du bruit et du silence. Commençons par le bruit. La table 5 contient un verbe comme *galoper* dont l'entrée dans la Table 5 est justifiée par l'exemple en (6) (avec *Prép* = Loc).

(6) Il a galopé dans l'esprit de Luc que Zoé devait tondre

L'exemple (6) est clairement métaphorique. De ce fait, nous ne voulons pas compter *galoper* comme tête lexicale d'une possible construction impersonnelle. La raison est que nous ne voulons pas multiplier **inutilement** l'ambiguïté dans les traitements automatiques. Que ce soit dans ILIMP ou dans les analyseurs syntaxiques qui reposent sur le Le*fff*, considérer *galoper* comme tête lexicale d'une possible construction impersonnelle amènerait à considérer (6) et (7) ci-dessous comme ambiguës entre une lecture impersonnelle et personnelle, alors qu'elles ne le sont nullement. Dans la lecture impersonnelle de (6) ou (7), l'objet direct de *tondre* n'est pas réalisé. Dans la lecture personnelle, il est réalisé sous forme de pronom relatif (*que)*, la phrase ayant la structure N_0 galoper Loc N_1 (Table 35L).

(7) Il a galopé dans le champ de Luc que Zoé devait tondre [il = le cheval]

Certes, on peut arguer que les exemples (6) et (7) se différencient par les noms *esprit* et *champ*, l'un à caractère abstrait (*esprit*) l'autre à caractère concret (*champ*), et que cette différence peut permettre de désambiguïser ces phrases. Néanmoins, il est bien connu que ces traits sémantiques sont difficilement codables. Aussi, nous pensons que nous ne pourrons pas désambiguïser (6) ou (7) par manque d'informations sémantiques, et nous prenons le parti de ne pas considérer que *galoper* a une entrée dans la Table 5, entrée qui serait à la fois rare et métaphorique. Ceci pour ne pas ajouter une ambiguïté artificielle dans les analyseurs, qui doivent déjà faire le départ entre un nombre d'analyses exponentiel en la longueur de la phrase.

Passons au silence. La possibilité d'avoir un sujet phrastique extraposable dans une construction passive ou pronominale n'est pas codée dans les tables à complétive. Pourtant, elle varie d'un verbe à l'autre, y compris dans la même table. A titre d'illustration, considérons la table 6 de propriété définitoire N_O V (Que P)$_1$, dont font partie les verbes *spéculer* et *sentir* (*Luc spécule/sent que Zoé va partir*). Le verbe *spéculer* autorise une construction impersonnelle passive, mais pas *sentir*, voir (8).

(8) a Il a été spéculé que Zoé partira

b *Il a été senti que Zoé partira[12]

Une colonne intitulée [extrap-passif] et une autre colonne [extrap-pronominale] auraient donc dues être inclues dans la table 6 et dans d'autres tables. En l'absence d'un tel codage, seul un travail manuel de linguiste permet de rattraper des constructions impersonnelles non codées dans le lexique-grammaire.

C'est donc bien à partir des bases linguistiques de ILIMP (et non directement à partir de celles du lexique-grammaire) qu'ont été renseignées les constructions impersonnelles dans le Le*fff*, en suivant la modélisation décrite dans la Section 4. 1.

[12] L'acceptabilité de cet exemple semble augmenter si l'on rajoute un élément en *par* qui introduit un syntagme nominal indéfini ou collectif (*il a été senti [par un grand nombre de personnes / par tout l'assistance / ??par Luc] que Marie partira*).

3. Le L*efff*

Le L*efff* est un lexique électronique de la langue française, librement disponible[13]. Il associe à chaque forme fléchie des informations morphologiques (lemme, étiquette morphologique) et syntaxiques (dont le cadre de sous-catégorisation). Il est destiné à être directement utilisé dans des applications de TAL, en essayant toutefois d'être indépendant des choix théoriques de ses utilisateurs, et en particulier des théories syntaxiques. Ce lexique présente la particularité de chercher un équilibre entre la pertinence de la modélisation linguistique et l'adéquation aux besoins opérationnels. Ceci se traduit dans la façon dont l'information linguistique est représentée, mais également dans la façon dont elle est ajoutée au L*efff*. En effet, en plus de méthodes manuelles, des techniques d'acquisition automatique d'information ont été utilisées pour compléter et corriger le L*efff*. Cependant, ces techniques sont toujours suivies d'étapes de validation manuelle, pour permettre la préservation d'un niveau satisfaisant de qualité.

Le L*efff* est une ressource à large couverture : il rassemble plus de 110 000 lemmes, auxquels correspondent plus de 520 000 entrées[14]. Parmi ces entrées, les formes verbales, un certain nombre de formes nominales et adjectivales (mais pas toutes), ainsi que les prépositions et d'autres types d'entrées (y compris des constructions à verbe support) sont associées à des cadres de sous-catégorisation spécifiques[15], ainsi qu'à des informations syntaxiques complémentaires (contrôle, attributif,…).

3.1. Le développement du L*efff* : historique et architecture

Le développement du L*efff* a commencé en 2003, à partir du constat suivant : à cette époque, il n'existait pas de lexique syntaxique pour le français librement utilisable et à couverture importante. Le développement d'un tel lexique a donc été lancé, avec le double objectif qu'il soit adapté au TAL tout en restant linguistiquement pertinent.

La première étape dans le développement d'un lexique syntaxique est celui du lexique morphologique sous-jacent. Dans (Clément, Sagot et Lang, 2004), les auteurs décrivent la première version d'une technique d'acquisition automatique de lexique morphologique, à partir d'un corpus brut et d'une description morphologique de la langue étudiée. Cette technique, dont une version plus aboutie est présentée dans (Sagot, 2005), repose sur la variabilité morphologique. Pour cette raison, elle n'a été appliquée pour le français que pour l'acquisition d'un lexique verbal. C'est ainsi qu'a été mis à disposition de la communauté le L*efff* 1, lexique morphologique verbal du français, acquis automatiquement et validé manuellement.

---

[13] On se référera au site Internet du L*efff*, à l'adresse suivante : www.lefff.net

[14] De plus, certaines de ces entrées sont encore factorisées : par exemple, pour un verbe du premier groupe, les première et troisième personne du singulier du présent de l'indicatif et du subjonctif sont regroupés en une seule entrée factorisée, avec une étiquette morphologique incluant deux disjonctions

[15] Par opposition aux cadres de sous-catégorisation génériques (et très tolérants) attribués pour l'instant aux noms et aux adjectifs pour lesquels les informations manquent.

Par la suite, des cadres de sous-catégorisation ont été ajoutés à ce lexique de formes verbales. Pour environ deux tiers des verbes, des premiers travaux non publiés de Lionel Clément constituaient déjà un premier lexique syntaxique du français, développé à des fins d'analyse syntaxique. Ces informations relativement préliminaires ont servi de base à la constitution d'un graphe de classes syntaxiques, définies par héritage de propriétés syntaxique atomiques, propriétés elles-mêmes définies de façon indépendante de la définition des classes. Ces classes permettent de définir un **lexique intensionnel** formé de triplets *(lemme, classe morphologique, classe syntaxique)*. La classe morphologique permet de fléchir les lemmes et d'associer aux formes obtenues, outre leur lemme et sa classe syntaxique de leur lemme, leur étiquette morphologique et leur « type morphosyntaxique » (*Defaut*, *Infinitif*, *ParticipePasséActif*, *ParticipePasséPassif*) qui spécifie pour chaque forme des transformations éventuelles à opérer sur la structure syntaxique de base (*Defaut*)[16].

Une deuxième phase permet, à partir de la classe syntaxique associé au lemme et de cette propriété morphosyntaxique, de construire le **lexique extensionnel** complet, qui associe à chaque forme fléchie une catégorie (ou partie du discours), éventuellement un poids (calculé selon des heuristiques ou renseigné manuellement), et une structure syntaxique complète, y compris un cadre de sous-catégorisation.

Un lexique syntaxique extensionnel des formes verbales du français a donc été constitué et mis à disposition de la communauté : il s'agit du L*efff* 2.0.

En parallèle à ces développements, le développement d'un lexique syntaxique couvrant toutes les catégories était en cours. Alors que les catégories fermées ont été renseignées principalement à la main, les autres catégories ouvertes ont été constituées par divers moyens complémentaires : acquisition automatique (avec validation manuelle) à l'aide de techniques déjà citées (Clément, Sagot et Lang 2004, Sagot 2005) , acquisition automatique (avec validation manuelle) d'informations syntaxiques atomiques (*cf.* Sagot 2006 : ch. 7), pour certains noms, adjectifs et adverbes, exploitation du lexique morphologique Multext pour le français (Véronis 1998), dont la libre exploitation nous a été autorisée explicitement par son principal auteur , corrections et ajouts manuels ou guidés par des techniques automatiques, comme par exemple la fouille d'erreurs dans les sorties d'analyseurs syntaxiques (Sagot et de La Clergerie 2006).

C'est donc aujourd'hui un lexique syntaxique à large couverture pour le français, qui ne se restreint ni aux seules formes verbales ni aux seules informations morphologiques, qui est mis à disposition. Le L*efff*, actuellement[17] en version 2.5, est entièrement téléchargeable sous une licence libre (LGPL-LR), sur www.lefff.net.

[16] Ainsi, *Infinitif* rend le sujet facultatif ; ou encore, *ParticipePasséPassif*, comme nous le verrons plus bas, applique au cadre de sous-catégorisation le changement de diathèse. C'est ce mécanisme qui permet de créer des entrées spécifiques pour les différents types de constructions impersonnelles, dont les structures syntaxiques sont obtenues là aussi par transformation des structures de base.

[17] Avril 2007

3.2. Modélisation des informations syntaxiques

Une entrée simple du Le*fff* ressemble à ce qui suit :

mange v [pred='manger$_1$<Suj:sn|cln,Obj:(sn|cla)>', cat=v, @PS13s]

On distingue bien la catégorie et la structure syntaxique, présentée entre crochets. Cette dernière comporte un « pred » à la LFG, composé d'un identifiant sémantique, souvent identique au lemme[18], et d'un cadre de sous-catégorisation que nous étudions en détails ci-dessous. Ici, il s'agit d'un sujet nominal ou clitique obligatoire et d'un objet nominal ou clitique facultatif (comme indiqué par les parenthèses). Enfin, la catégorie morphosyntaxique est indiquée par l'attribut cat[19], ainsi qu'une macro résumant l'étiquette morphologique[20] (les macros sont introduites par « @ »).

3.2.1. Le cadre de sous-catégorisation : les fonctions syntaxiques et leurs réalisations

Un cadre de sous-catégorisation est constitué d'une liste (éventuellement vide, sinon présentée entre chevrons) de *fonctions syntaxiques*, chacune d'entre elle se voyant attribuer un certain nombre de *réalisations* (de surface) possibles[21]. Ces réalisations peuvent être des clitiques ou des syntagmes (nominaux, adjectivaux, etc). Toutefois, une fonction syntaxique peut n'être que facultativement réalisée, ce qui est indiqué par une mise entre parenthèses. La position d'une fonction dans la liste de fonctions que constitue un cadre de sous-catégorisation est le *rang* de cette fonction dans ce cadre[22].

Nous avons pris en compte les arguments et les conclusions de divers auteurs, et plus particulièrement les travaux de Karel van den Eynde et Piet Mertens pour le lexique de valence Proton, aujourd'hui DICOVALENCE, développé dans l'approche pronominale (van den Eynde et Mertens 2006). Ceci nous a conduit à la liste de fonctions syntaxiques ci-dessous, indiquées avec leurs critères définitoires.

---

[18] Pour distinguer les homonymes, un double mécanisme a été mis en place : numérotation des prédicats homonymes, et possibilité d'attribuer à un prédicat un identifiant explicite. Ainsi, on peut distinguer quatre homonymes pour *passer*, mais également deux pour *voler* (correspondant respectivement aux verbes anglais *steal* et *fly*.). Dans les fichiers texte du Le*fff* extensionnel, ceci est noté de la façon suivante : *lemme___(identifiant explicite)__identifiant numérique*. Ainsi, on a *passer_____1* à *passer_____4*, ainsi que *voler___steal__1* et *voler___fly__2*. Dans les exemples, par souci de lisibilité, nous avons remplacé le quintuple symbole « _ » par une mise en indice.

[19] Cette dernière est souvent identique à la catégorie (syntaxique), mais pas toujours. Ainsi, le préfixe *ex-* a une catégorie (syntaxique) *adjPref*, mais sa catégorie morphosyntaxique est cat=adj.

[20] Cette étiquette, comme indiqué dans une note précédente, est ici à double disjonction (personne 1 et 3, temps P=présent de l'indicatif et S=présent du subjonctif)

[21] Le découplage entre fonctions syntaxiques et réalisations résoud un grand nombre de difficultés théoriques et pratiques qui se posent dans un formalisme comme LFG. Ceci permet par exemple de représenter correctement des cadres de sous-catégorisation où deux « fonctions grammaticales » LFG identiques peuvent coexister, et en particulier deux compléments indirects ou obliques introduits par la même préposition (cf. *Le taux de chômage a été divisé par deux par les dernières réformes*).

[22] Les rangs 1, 2, 3, 4 correspondent aux positions 0, 1, 2, 3 qui indicent les positions du sujet et des compléments dans les cadres syntaxiques définitoires des tables du lexique-grammaire (*cf.* section 2).

Suj : Fonction *sujet*. La forme clitique est celle d'un clitique nominatif personnel. À l'actif, elle est réalisée canoniquement en position pré-verbale[23] (avec accord). Elle correspond au paradigme P0 du DICOVALENCE.

Obj : Fonction *objet (direct)*. La forme clitique est celle d'un clitique accusatif, ou d'un clitique génitif à sens partitif. Un verbe sous-catégorisant une fonction objet est dit transitif (direct). Si le verbe est passivable[24], cette fonction est translatée, pour devenir fonction sujet. Elle est proche du paradigme P1 du DICOVALENCE.

Objà : Fonction *objet indirect introduit par à*, ou fonction *à-objet*. Est substituable un syntagme prépositionnel de la forme *à + pronom non-clitique*[25]. La cliticisation est possible à l'aide du clitique datif (dans tous les cas, ou bien seulement dans le cas humain), peut être possible dans certains cas seulement (non humain) à l'aide du clitique locatif *y*, ou ne pas être possible du tout. Elle se distingue de la fonction locative par la non-substituabilité des pronoms *là, ici, là-bas*. Elle correspond au paradigme P2 du DICOVALENCE.

Objde : Fonction *objet indirect introduit par de*, ou fonction *de-objet*. Est substituable un syntagme prépositionnel de la forme *de + pronom non-clitique*. La cliticisation est possible à l'aide du clitique génitif. Elle se distingue de la fonction délocative Dloc par la non-substituabilité avec les locutions pronominales *de là*, *d'ici*. Elle correspond au paradigme P3 du DICOVALENCE.

Loc : Fonction *locative*. Les pronoms *là*, *ici*, *là-bas* sont substituables. La cliticisation, si elle est possible, se fait à l'aide du clitique locatif *y*. Elle correspond au paradigme PL du DICOVALENCE.

Dloc : Fonction *délocative*. Les locutions pronominales *de là*, *d'ici* sont substituables. La cliticisation, si elle est possible, se fait à l'aide du clitique génitif *en*. Elle correspond au paradigme PDL du DICOVALENCE.

Att : Fonction *attributive*. Cette fonction, dont les propriétés de cliticisation sont variables, regroupe les attributs du sujet ou d'un des objets (objet, à-objet). Les situations couvertes sont variées : *prendre Pierre [pour Adj / det N]*$_{Att}$*, nommer Pierre [président]*$_{Att}$*, regarder Pierre [courir]*$_{Att}$*, trouver Pierre [Adj]*$_{Att}$*, voir Pierre [(comme) (det) N / (comme) Adj]*$_{Att}$.

Obl et **Obl2** : Fonctions *obliques*. Ces fonctions, qu'aucun critère ne distingue l'une de l'autre[26], abritent les compléments obliques, jamais cliticisables, qui ne rentrent dans aucune des autres fonctions (y compris le « complément d'agent » de la grammaire traditionnelle dans les constructions passives). Dans un avenir proche,

[23] Ce n'est cependant pas le cas en présence d'un *il* impersonnel avec sujet extraposé.

[24] Ce qui n'est pas automatique, même en cas de transitivité : *Ce problème regarde Marie* / **Marie est regardée par ce problème*.

[25] *Pronom non-clitique* est à prendre ici au sens de (van den Eynde et Mertens 2006).

[26] On n'utilise Obl2 que lorsqu'il y a deux compléments obliques.

grâce à la disponiblité du lexique DICOVALENCE (ex-Proton), ces fonctions obliques pourront être mieux précisées.

Comme dit plus haut, on attribue, dans un cadre de sous-catégorisation donné, une disjonction de réalisations à chaque fonction syntaxique présente. Cette disjonction de réalisations est donnée entre parenthèses si la réalisation de la fonction est facultative.

Les réalisations possibles sont de trois types :

- Un pronom clitique : clitique nominatif (**cln**), clitique accusatif (**cla**), clitique génitif (**en**), clitique locatif (**y**). On notera que le *se* réfléchi ou réciproque est considéré comme une réalisation de type cla ou cld selon les cas (*Les époux se disputent / Pierre se laisse cette possibilité*)[27] ;
- Un syntagme direct : syntagme nominal (**sn**), syntagme adjectival (**sa**), syntagme infinitif (**sinf**), syntagme phrastique fini (**scompl**), interrogative indirecte (**qcompl**). Rien n'exclue la possibilité d'introduire également des syntagmes adverbiaux (**sadv**) ;
- Un syntagme prépositionnel : il s'agit d'un syntagme direct précédé d'une préposition, comme **de-sn**, **à-sinf** ou **pour-sa**[28]. Enfin, les notations **à-scompl** et **de-scompl** représentent les réalisations en *à ce que P* et *de ce que P* respectivement.

À titre d'exemple, les formes du verbe *ordonner*, utilisées dans une construction personnelle active, seront du type :

ordonnât v [pred='ordonner$_1$<Suj:sn|cln,Obj:sn|cla|de-sinf|scompl, Objà:(à-sn|cld)>',cat=v,@T3s]

On notera que les fonctions syntaxiques dont il s'agit ici sont des fonctions syntaxiques de surface, au sens où un changement de diathèse redistribue certaines fonctions syntaxiques, ainsi que la façon dont elles sont réalisées. En revanche, leur rang dans le cadre de sous-catégorisation, qui n'est pas modifié par la diathèse, permet de garder trace de l'identité du rôle sémantique sous-jacent. Ainsi, les deux participes passés (actif et passif) de *manger* ont pour entrées respectives :

mangé v [pred='manger$_1$<Suj:sn|cln,Obj:(sn|cla)>',cat=v,@active,@avoir, @Kms]

mangé v [pred='manger$_1$<Obl:(par-sn),Suj:sn|cln>', cat=v,@passive,@Kms]

Comme on peut le voir, le sujet de la construction active, qui est de rang 1, correspond au complément oblique en **par-sn** de la construction passive, également de rang 1. De

[27] Actuellement, toute fonction syntaxique Obj ou Objà réalisable de façon clitique est donc considérée comme pouvant être réalisée par le clitique *se* (réflexif ou réciproque). Ceci est une approximation, que des travaux ultérieurs devront préciser plus avant.

[28] On notera que nous ne distinguons pas les prépositions des complémentiseurs, cette distinction se déduisant de la fonction syntaxique dont on parle. Un **de-sinf** réalisant une fonction objet met en œuvre le complémentiseur *de* (*Jean ordonne à Marie de partir*) ; un **de-sinf** réalisant une fonction de-objet est prépositionnel (*Jean rêve de partir*).

même pour l'objet actif qui devient sujet passif. On notera que ces deux participes passés, quoique n'ayant pas le même cadre de sous-catégorisation, sont deux formes issues de la même entrée lexicale au niveau intensionnel. En réalité, une même entrée intensionnelle peut se voir associer une classe syntaxique qui dénote elle-même une disjonction de comportements. C'est ainsi qu'un adjectif comme *envisageable*, une fois les constructions impersonnelles prises en compte (*cf.* section 4), n'aura qu'une seule entrée au niveau intensionnel, alors qu'au niveau extensionnel, chacune de ses formes aura deux entrées, l'une pour la construction personnelle (*ceci est envisageable/une chose envisageable*), et l'autre pour la construction impersonnelle (*il est envisageable de Vinf /que P*).

3.2.2. Autres propriétés syntaxiques

D'autres propriétés syntaxiques complètent le cadre de sous-catégorisation. Pour la plupart des réalisations infinitives (directes ou non) dans les cadres de sous-catégorisation verbaux, une information de contrôle est donnée : selon les cas, le sujet de l'infinitive est « égal » (en un sens qui dépend des théories syntaxiques) au sujet, à l'objet direct ou à l'objet indirect (à-objet) du verbe. On notera que certains verbes sous-catégorisent des fonctions pouvant se réaliser par une infinitive sans qu'il y ait contrôle : *le travail consiste à créer un lexique*. De même, toute fonction attributive à réalisation nominale ou adjectivale possible (directe ou prépositionnelle) est précisée par une indication de la fonction à laquelle l'attribut s'applique (sujet, objet, à-objet).

Une autre propriété (mal) renseignée dans le Le*fff* est relative aux contraintes sur le mode des complétives, qu'elles réalisent des fonctions syntaxiques sous-catégorisées par des entrées verbales, nominales, ou autres. Quatre grands cas de figure sont répertoriés : le cas où les deux modes sont possibles, le mode indicatif obligatoire, le mode subjonctif obligatoire, et le mode dit alternant (subjonctif en cas de non-assertion, indicatif sinon).

Ces propriétés sont indiquées, comme pour l'étiquette morphosyntaxique, par des macros (telles que @CtrlSuj pour le contrôle sujet, @AttObj pour l'attribut de l'objet, @ObjSubj pour la complétive objet au subjonctif). L'idée est que chaque utilisateur du Le*fff* est amené à donner à ces macros la signification appropriée, compte tenu du formalisme ou du contexte d'utilisation de cette information. Toutefois, nous disposons d'une définition de ces macros sous forme de structure de traits avec partage possible, qui représente de façon transparente ce qu'elles veulent dire. Ainsi, on a la définition suivante pour le contrôle sujet :

@CtrlSuj := [Suj=[]1, Obj =[Suj = []1]]

Ceci indique que le sujet de l'infinitive objet est partagé (par co-indiciation) avec le sujet principal (le mécanisme qui n'applique cette macro que lorsque la réalisation de l'objet est effectivement infinitive n'est pas indiqué) .

Enfin, si les fonctions syntaxiques correspondent à des arguments syntaxico-sémantiques du prédicat concerné (verbal, adjectival, nominal, etc.), certaines entrées

s'accompagnent de la présence de clitiques qui ne correspondent à aucun argument syntaxico-sémantique. Outre les pronoms impersonnels *il* et *ça*, que nous traiterons plus bas, c'est le cas du clitique réfléxif dans le cas des verbes dits essentiellement pronominaux (*s'évanouir*), du clitique génitif (*en référer à quelque chose, en être quelque part*), du clitique locatif (*y passer*), voir de plusieurs d'entre eux (*s'y connaître, s'en tirer*). Ces clitiques sont requis à l'aide de macros spécifiques dans la structure syntaxique, dont certaines empruntent une terminologie issue de DICOVALENCE (@pseudo-y, @pseudo-en, @négatif, @pronominal…).

3.3. Utilisation en TAL

Le L*efff* est utilisé par au moins deux systèmes d'analyse très différents pour le français. Le premier d'entre eux est FRMG (Thomasset et de La Clergerie 2005), un analyseur TAG qui repose sur une métagrammaire, laquelle génère une TAG factorisée. Les entrées du L*efff* sont utilisées comme « hypertags » pour ancrer les quasi-arbres. Le second analyseur, est l'analyseur du français construit à l'aide de SxLFG, constructeur d'analyseurs LFG (Boullier et Sagot 2005). Il s'agit d'un analyseur LFG efficace qui utilise les entrées du L*efff* comme entrées lexicales LFG. Ces deux systèmes qui reposent sur le L*efff* sont utilisés dans différentes expériences à grande échelle, telles que la campagne EASy d'évaluation des analyseurs syntaxiques (Boullier *et al.* 2005), l'analyse de corpus de plusieurs millions de phrases ou l'apprentissage d'informations (p.ex. d'ontologies) à partir de corpus spécialisés.

Il est difficile de donner un aperçu quantitatif de la couverture et de la précision du L*efff*. Toutefois, nous avons développé un « chunker » à règles et reposant sur le L*efff* qui segmente une phrase en syntagmes non-récursifs (les « constituants » de la campagne EASy), dont les résultats étaient déjà très satisfaisants avant l'enrichissement présenté dans cet article (Sagot 2006 : ch. 12). Nous nous attendons à ce qu'après enrichissement, les résultats de ce chunker (mais également des analyseurs profonds cités ci-dessus) soient significativement meilleurs : ceci devrait nous permettre dans un avenir proche d'évaluer quantitativement l'impact respectif de la prise en compte des constructions impersonnelles.

Par ailleurs, des travaux sont en cours pour comparer le L*efff* à différentes autres ressources, notamment Morphalou (pour le lexique morphologique sous-jacent), mais surtout SynLex, lexique syntaxique cité plus haut, extrait à partir de certaines tables du lexique-grammaire (Gardent *et al.* 2006), et DICOVALENCE, anciennement PROTON (van den Eynde et Mertens 2006).

4. Du lexique-grammaire au L*efff* : constructions impersonnelles

L'extracion des informations du Lexique-Grammaire en vue de leur intégration dans le L*efff* est loin d'être triviale : différents types de travaux sont en effet indispensables pour mener à bien une telle entreprise :

- Compréhension précise des informations répertoriées dans les tables

- Filtrage des informations linguistiques à intégrer au Le*fff* (filtrage des colonnes)
- Filtrage des entrées lexicales à intégrer au Le*fff* (filtrage des lignes)
- Développement pour ces constructions d'un modèle compatible avec les principes sur lesquels reposent le Le*fff*
- Extraction effective des informations et des entrées choisies au format Le*fff*, en respectant le modèle retenu
- Si possible, évaluation dans des analyseurs de l'impact d'un tel travail, au moins en termes de taux de couverture.

C'est à cette succession de tâches que nous nous sommes attelés pour les constructions impersonnelles. Comme nous l'avons vu dans la section 2.2, une partie du travail avait déjà été effectuée au sujet des impersonnelles lors de la construction de l'outil ILIMP (Danlos 2005). En particulier, la compréhension fine des tables et le filtrage des entrées lexicales étaient déjà réalisés.

Cependant, ILIMP se présente sous forme de réseaux de transition récursifs qui reconnaissent des motifs dans un texte brut (i.e. sans annotation linguistique). Ce qui est très différent du point de vue adopté ici, qui est celui de la construction d'un lexique syntaxique, destiné à être utilisé dans un analyseur syntaxique profond. En un sens, notre point de vue est moins complexe, puisque l'on peut faire abstraction (par exemple) de la variabilité du matériau qu'on peut insérer entre une tête verbale et un complément (figé, par exemple), et de la complexité de sa délimitation : c'est en effet aussi le travail de la grammaire, et non seulement du lexique. Mais la contrepartie de cette relative simplification est une plus grande abstraction dans les descriptions. Il ne s'agit plus de décrire des motifs appliquables à des séquences de formes étiquetées, mais des cadres de sous-catégorisation et des contraintes syntaxiques. C'est la raison pour laquelle nous avons construit un modèle des constructions impersonnelles qui n'est pas la transcription directe des graphes d'ILIMP, mais dont le contenu (associations entre lemmes et constructions) en est directement extrait.

La dichotomie entre constructions intrinsèquement impersonnelles et constructions à sujet extraposé (*cf.* 2.2.2) reste naturellement fondamentale dans notre description. Parmi ces dernières, et en raison de l'importance donnée au cadre de sous-catégorisation fonctionnel, la distinction principale ne se fait pas entre verbes à sujet phrastique et verbes à sujet nominal, mais entre constructions à diathèse active (le sujet extraposé est un sujet profond, il réalise une fonction sujet) et constructions à diathèse passive ou moyenne.

Dans tous les cas, une construction impersonnelle est caractérisée par le fait que la position syntaxique sujet, obligatoire en français dans les propositions finies, est occupée par un pronom impersonnel *il* ou *ça* (mais nous n'avons pas encore traité ce dernier cas). Reprenons successivement les deux classes de constructions impersonneles identifiées dans la section 2 : les constructions intrinsèquement impersonnelles et les constructions à sujet extraposé.

4.1. Les constructions intrinsèquement impersonnelles

Les constructions intrinsèquement impersonnelles sont modélisées par l'absence de fonction sujet dans le cadre de sous-catégorisation. Par conséquent, seule une construction impersonnelle est possible : c'est ce qui est requis par la macro @impers. Ces constructions sont de trois types[29] :

Les verbes avec cadre de sous-catégorisation vide (la plupart des verbes météorologiques de la table 31I, comme *il vente*) ;

vente v [pred='venter$_1$',@impers,@PS3s]

Les verbes intrinsèquement impersonnels sous-catégorisant une fonction objet, comme les verbes de la table 17 (comme *il faut* N/Vinf/queP) ou la locution *il y a* N (Loc N) ;

faut v [pred='falloir$_1$<Obj:(sn|cla|sinf|scomp)>',@impers,@ObjSubj[30], @P3s]

a v [pred='y avoir<Obj:sn,Loc:(loc-sn)>',@impers,@pseudo-y,@P3s][31]

Les verbes intrinsèquement impersonnels sous-catégorisant une autre fonction (comme *il s'agit de N/Vinf*) ;

agit v [pred='s'agir$_1$<Objde:de-sn|clg|scompl|de-sinf>',@pron,@impers, @ObjdeSubj,@P3s][32]

4.2. Les constructions à sujet extraposé

Ces constructions se répartissent entre constructions à prédicat verbal et constructions à prédicat adjectival. Au contraire des constructions intrinsèquement impersonnelles, les cadres de sous-catégorisation incluent une fonction sujet. Il y a donc possibilité d'alternance entre constructions impersonnelles et personnelles.

Comme nous l'avons vu, la macro @impers impose une construction impersonnelle, qui est ici avec sujet extraposé, puisque nous traitons dans cette section des entrées sous-catégorisant une fonction sujet. Les constructions personnelles parallèles, inexistantes pour les constructions intrinsèquement impersonnelles, sont indiquées par la macro @pers. On notera que la duplication au niveau extensionnel des constructions impersonnelles et personnelles n'est pas la conséquence d'une duplication au niveau intensionnel : chaque forme d'une même entrée intensionnelle (du même prédicat) peut être impliquée dans plusieurs constructions, certaines personnelles, d'autres impersonnelles. Comme nous l'avons vu plus haut, c'est déjà le cas pour les verbes personnels, dont le participe passé a deux entrées, l'une active, l'autre passive.

Il en est de même pour les adjectifs. Il semble que tout adjectif dont la fonction sujet peut avoir une réalisation complétive ou infinitive admette une construction

[29] Pour simplifier la lecture, nous n'avons pas répété l'information cat=v ou cat=adj dans la structure syntaxique des exemples donnés.

[30] Rappelons que cette macro indique que si la fonction objet est réalisée sous la forme d'une complétive, alors celle-ci doit être au subjonctif.

[31] Une autre entrée couvre le cas (familier) *il y a que je suis malade*.

[32] La réalisation **de-scompl** du Objde de *agir* est exclue : **Il s'agit de ce que Paul parte*.

impersonnelle de la forme il_{imp} est Adj Y (où Y dénote une complétive et/ou une infinitive, suivant les cas). La construction impersonnelle, qui est alors possible, induit une translation du **sinf** en **de-sinf** (*dormir est impossible / il est impossible de dormir*).

On peut regrouper les constructions à sujet extraposé en différentes classes :

Un certain nombre de verbes admettant une construction impersonnelle en parallèle de la construction personnelle correspondante (toute la table 5, quelques entrées des tables 7 et 8) ;

plaît v [pred='plaire$_1$<Suj:sn|cln|sinf|scompl,Objà:(à-sn|cld)>',@pers,@P3s]

plaît v [pred='plaire$_1$<Suj:de-sinf|scompl,Objà:(à-sn|cld)>',@SujSubj, @impers, @P3s]

découle v [pred='découler$_1$<Suj:sn|cln|scompl,Objde:de-sn|clg|de-scompl>', @SujSubj,@ObjdeInd,@pers,@PS13s]

découle v [pred='découler$_1$<Suj:sn|scompl,Objde:de-sn|clg|de-scompl>', @SujSubj,ObjdeInd,@impers,@PS13s]

Les verbes non intrinsèquement impersonnels pour lesquels une construction impersonnelle passive ou moyenne existe, soit de façon exclusive, soit à côté d'une construction personnelle (certains verbes, en particulier des tables 6 et 9). L'exemple du participe passé *raconté* permet d'illustrer tous les cas :

- Participe passé actif (Pierre a raconté un conte à Marie)

raconté v [pred='raconter$_1$<Suj:cln|sn,Obj:sn|cla|sinf|scompl,Objà:(à-sn|cld)>', @CtrlSuj,@ObjInd,@pers,@Kms]

- Participe passé passif en construction personnelle (*Un conte a été raconté par Pierre à Marie*)

raconté v [pred='raconter$_1$<Obl:(par-sn),Suj:sn|scompl,Objà:(à-sn|cld)>', @passif, @pers, @Kms]

- Participe passé passif en construction impersonnelle (*Il a été raconté un conte à Marie par Pierre*)

raconté v [pred='raconter$_1$<Obl:(par-sn),Suj:sn|scompl,Objà:(à-sn|cld)>', @passif, @impers, @Kms]

- Participe passé moyen en construction personnelle (Un (tel) conte (ne) s'est (pas) raconté à (quelqu'un comme) Marie (depuis longtemps))

raconté v [pred='raconter$_1$<Suj:sn|scompl,Objà:(à-sn|cld)>', @pron, @pers, @Kms]

- Participe passé moyen en construction impersonnelle (Il (ne) s'est (pas raconté un (tel) conte à (quelqu'un comme) Marie (depuis longtemps))

raconté v [pred='raconter$_1$<Suj:sn|scompl,Objà:(à-sn|cld)>', @pron, @impers, @Kms]

Un certain nombre d'adjectifs admettant une construction impersonnelle de type ilimp est Adj Y (*il est envisageable de faire cela*). On notera qu'un adjectif sous-catégorise toujours une fonction sujet, et parfois d'autres fonctions également[33] ;

envisageable adj [pred='envisageable$_1$<Suj:(sn|sinf|scompl)>',@pers,@s]

envisageable adj [pred='envisageable$_1$<Suj:(de-sinf|scompl)>',@impers,@s]

Un certain nombre d'expressions de type être Prep X admettant une construction impersonnelle de type ilimp est Prep X Y (*il est de règle de porter un chapeau*). Elles sont similaires à des adjectifs, et catégorisées comme telles.

de règle adj [pred='de règle$_1$<Suj:(sn|sinf|scompl)>',@pers,@s]

de règle adj [pred='de règle$_1$<Suj:(de-sinf|scompl)>',@impers,@s]

à l'actif adj [pred='à l'actif$_1$<Suj:(sn|sinf|scompl),Objde:de-sn>', @pers,@s][34]

à l'actif adj [pred='à l'actif$_1$<Suj:(de-sinf|scompl),Objde:de-sn>',@impers,@s]

4.3. Bilan

Nous avons donc extrait des différents graphes pertinents qui constituent ILIMP des listes de verbes et d'adjectifs associés à chacun de ces cas. Puis nous avons créé de nouvelles classes syntaxiques pour les constructions impersonnelles, ou modifié certaines classes existantes, afin d'ajouter ou de modifier les entrées du Le*fff* extensionnel d'une façon cohérente à la fois avec l'analyse ci-dessus et les principes de représentation présentés en section 3.

Conclusion

Le Le*fff*, lexique syntaxique du français à large couverture, a désormais des fondements linguistiques et formels qui permettent son utilisation dans des analyseurs syntaxiques profonds à grande échelle. Ces fondements nous on permis d'exploiter la source d'informations linguistique que sont les tables du lexique-grammaire, pour modéliser une familles de phénomènes syntaxiques non standard : les constructions impersonnelles. Des travaux préliminaires ont également eu lieu sur les expressions verbales figées, il nous faudra les poursuivre. De plus, la mise à disposition du lexique DICOVALENCE nous permet de disposer d'une autre source d'informations linguistiques, ce qui sera particulièrement utile, entre autres, pour modéliser de façon satisfaisante l'ensemble des constructions pronominales, aujourd'hui à l'état d'ébauche dans le Le*fff*.

[33] On notera que l'information sur le mode de la complétive n'est actuellement pas disponible. Il en est de même ci-dessous pour les expressions en être Prep X.

[34] Anticipant ainsi certains problèmes liés au figement, notons que ces entrées ne couvrent pas le cas *à son actif*. Il est donc nécessaire d'introduire, indépendamment, un lemme *à son actif* dont les formes fléchies (*à mon actif*, *à ton actif*,…) ne sous-catégorisent qu'un sujet.

Références

BOONS J.-P., GUILLET A., LECLERE C. (1976a), La structure des phrases simples en français, Constructions intransitives, Droz, Genève.

BOONS J.-P., GUILLET A., LECLERE C. (1976b), *La structure des phrases simples en français, Classes de constructions transitives*, Rapport de recherches, LADL, CNRS, Univ. Paris 7.

BOULLIER P. et SAGOT B. (2005), « Analyse syntaxique profonde à grande échelle: SXLFG », in *Traitement Automatique des Langues*, n° 46/2.

BOYD A, GEGG-HARRISON W. ET BYRON D. (2006), Identifying non-referential *it*. A machine learning approach incorporating linguistically motivated patterns, revue TAL, vol. 46 n° 1.

CANDITO M.-H. (1999), Représentation modulaire et paramétrable de grammaires électroniques lexicalisées, Thèse de doctorat, Université Paris 7.

CLEMENT L., SAGOT B. et LANG B. (2004), « Morphology based automatic acquisition of large-coverage lexica », in *Proceedings of LREC 2004*, Lisbonne, Portugal.

DANLOS L. (1980), Représentation d'informations linguistiques: les constructions N être Prép X, Thèse de troisième cycle, Université Paris 7.

DANLOS L. (1992), « Support Verb Constructions: linguistic properties, representation, translation », in *Journal of French Linguistic Studies*, n° 2/1, Cambridge University Press, Cambridge.

DANLOS L. (2005), « ILIMP : Outil pour repérer les occurences du pronom impersonnel *il* », in *Actes de TALN 2005*, Dourdan, France.

DANLOS L. et SAGOT B. (2007), « Comparaison du Lexique-grammaire et de DICOVALENCE : vers une intégration dans le Le*fff*», in *Actes de TALN 2007*, Toulouse, France.

VAN DEN EYNDE K. et BLANCHE-BENVENISTE, C. (1978), « Syntaxe et mécanismes descriptifs : présentation de l'approche pronominale », in *Cahiers de Lexicologie* n°32 : 3-27.

VAN DEN EYNDE K. et MERTENS P. (2006), Le dictionnaire de valence DICOVALENCE : manuel d'utilisation, à paraître.

GARDENT C., GUILLAUME B., PERRIER G. et FALK I. (2006), « Extraction d'information de sous-catégorisation à partir des tables du LADL », in *Actes de TALN 2006*, Louvain, Belgique.

GUILLET A. et LECLERE C. (1992), La structure des phrases simples en français, Constructions transitives locatives, Droz, Genève.

SAGOT B. (2005), « Automatic acquisition of a Slovak Lexicon from a Raw Corpus », in *Proceedings of TSD 2005*, Karlovy Vary, Tchéquie (LNAI 3658, © Springer-Verlag).

SAGOT B., CLEMENT L., DE LA CLERGERIE É. et BOULLIER P. (2006), « The Le*fff* 2 syntactic lexicon for French: architecture, acquisition, use », in *Actes de LREC 2006*, Gênes, Italie.

SAGOT B. et DE LA CLERGERIE É. (2006), « Error mining in parsing results », in *Proceedings of ACL-CoLing 2006*, Sydney, Australie.

SAGOT B. (2006), Analyse automatique du français : lexiques, formalismes, analyseurs, Thèse de doctorat, Université Paris 7.

VERONIS J. (1998), Multext-Lexicons, A set of Electronic Lexicons for European Languages.

THOMASSET F., DE LA CLERGERIE É. (2005), « Comment obtenir plus des méta-grammaires », in *Actes de TALN 2005*, Dourdan, Belgique.