Le lexique-grammaire est-il exploitable pour le traitement des langues ?

Eric Laporte

Le lexique-grammaire est-il exploitable pour le traitement des langues ?

Éric Laporte
Université Paris-Est

Abstract

The Lexicon-Grammar of French is a dictionary with structured syntactic-semantic information. In order to assess its exploitability in language processing, we survey four criteria: readability, degree of formalisation, degree of validity of information content, and richness in information. We contribute concrete examples to inform this discussion. We compare the significance of the criteria, in order to evaluate the validity of the priorities retained and of the compromises adopted in the course of the construction of the Lexicon-Grammar.

Keywords : lexicon, syntax, lexicon-grammar, parsing, language resource, language processing

Résumé

Le lexique-grammaire du français est un dictionnaire contenant des informations syntaxico-sémantiques structurées. Pour évaluer son exploitabilité dans le traitement des langues, nous passons en revue quatre critères : sa lisibilité, son degré de formalisation, le degré de validité de son contenu informatif, et sa richesse en informations. Nous proposons des exemples concrets susceptibles d'éclairer le débat sur cette question. Nous pesons l'importance de ces critères, afin d'évaluer la validité des priorités retenues et des compromis adoptés tout au long de l'élaboration du lexique-grammaire.

Mots-clés : lexique, syntaxe, lexique-grammaire, analyse syntaxique, ressource linguistique, traitement des langues

1. Introduction

Le lexique-grammaire (LG) du français est une grande base de données lexicales, syntaxiques et sémantiques. Il n'a pas été conçu pour être exploité dans le traitement des langues dès le début de sa construction, c'est-à-dire vers 1968, en tout cas pas uniquement. Toutefois, par la suite, certains auteurs du LG ont fait état de cette potentialité d'exploitation, qui a été pour beaucoup dans le fait que l'élaboration du LG a rencontré un soutien durable, s'est poursuivie au-delà des premières années, s'est étendue à d'autres langues que le français, et est toujours vivante actuellement. Certains chercheurs ont mis en doute que le LG soit exploitable pour le traitement des langues, généralement en privé. Quelques auteurs ont exploité des portions substantielles du LG des verbes français pour l'analyse syntaxique automatique de phrases (Roche, 1999 ; Paumier, 2001 ; Blanc, 2006). D'autres ont étudié les obstacles

à l'exploitation du LG en traitement des langues (Hathout et Namer, 1997, 1998 ; Gardent *et al.*, 2005, 2006). Dans cet article, nous approfondissons la réflexion sur cette question, et nous apportons au débat des exemples concrets sur quatre points : la lisibilité visuelle du format, le degré de formalisation, le degré de validité et la richesse en contenu informatif [1].

C'est avec une admiration sincère que nous dédions cette étude à Christian Leclère, un des principaux auteurs du LG des verbes distributionnels du français, et un des piliers du Laboratoire d'automatique documentaire de linguistique (LADL).

2. Lisibilité visuelle

Un simple coup d'œil sur un extrait du LG (fig. 1) montre que la lisibilité de son format a été une priorité dans sa conception.

N0 =: Nhum	N0 =: N-hum	N0 =: Nnr	N0 =: V-inf W	N0 être V-n	Ppv	Ppv =: se figé	Ppv =: en figé	Ppv =: y figé	Nég	<ENT>	N0 être V-ant	N0 est Vpp	N0 V de N0pc	[extrap]	Nactif V N0	<OPT>
~	~	~	~	~	<E>	-	-	-	-	barboter	~	~	~	~	~	Le gaz barbote dans l'eau
+	+	-	-	-	<E>	-	-	-	-	barboter	-	-	-	+	~	Max barbote dans l'eau
~	~	~	~	~	<E>	-	-	-	-	basculer	~	~	~	~	~	La chaise bascule
-	+	-	-	-	<E>	-	-	-	-	battre	-	-	+	-	~	Son coeur bat
-	+	-	-	-	<E>	-	-	-	-	béer	+	-	-	+	~	Sa bouche bée
-	+	-	-	-	<E>	-	-	-	-	blouser	+	-	-	-	~	Le chemisier blouse
~	~	~	~	~	<E>	-	-	-	-	boiter	~	~	~	~	~	Cette chaise boite
~	~	~	~	~	<E>	-	-	-	-	bomber	~	~	~	~	~	La voiture bombe
~	~	~	~	~	<E>	-	-	-	-	boucler	~	~	~	~	~	Le programme boucle
-	+	-	-	-	<E>	-	-	-	-	bouffer	+	-	-	-	~	Ses manches bouffent
~	~	~	~	~	<E>	-	-	-	-	bouger	~	~	~	~	~	La dent bouge
~	~	~	~	~	<E>	-	-	-	-	bouillir	~	~	~	~	~	L'eau bout à cent degrés

Figure 1. Extrait de la table 31R (Boons *et al.*, 1976) dans la version 1.1.

Les entrées lexicales sont faciles à identifier visuellement et à comparer : ce sont les lignes de la table. De même, les propriétés syntaxico-sémantiques sont matérialisées par l'alignement vertical des valeurs qu'elles prennent : ce sont les colonnes. Ce format tabulaire permet de croiser sur un même écran des dizaines d'entrées et des dizaines de propriétés. Ainsi, lorsqu'il code une entrée, le lexicologue a sous les yeux la description d'entrées comparables, pour peu que chaque table regroupe une classe d'entrées suffisamment homogène. Cette visualisation facilite le codage. Elle n'est réalisable qu'avec des étiquettes (ou intitulés) de propriétés répétées sur chaque écran, et donc très succincts : chaque intitulé occupe au plus une trentaine de caractères.

S'il existait des lexiques syntaxico-sémantiques complets et satisfaisants pour le traitement des langues, s'il existait un consensus dans la communauté scientifique sur

[1] Nous remercions le CNRS pour son soutien à ce travail par l'intermédiaire du Laboratoire d'Informatique Gaspard-Monge (LIGM).

les propriétés à renseigner ou sur les entrées lexicales à distinguer, ou s'il était prouvé que la perspective de construire de tels lexiques de façon entièrement automatique n'est pas largement utopique, alors le format tabulaire de la figure 1 ne serait peut-être pas d'actualité. Mais nous en sommes loin. Ce format facilite notamment la discussion scientifique sur la construction manuelle de lexiques pour le traitement des langues, discussion qui a besoin de se développer : dans le LG, par exemple, certaines entrées verbales ne sont pas codées, comme celles de *basculer* et *boiter* dans la figure 1 ; certaines constructions, notamment pronominales, ne sont pas représentées ; les entrées adjectivales sont en construction...

Chaque propriété syntaxico-sémantique décrit partiellement une construction. L'intitulé "N0 est Vpp" dans la figure 1 représente une phrase à interprétation statique constituée du sujet N_0 de la construction de base, de la copule *être*, et du participe passé du verbe. Dans le cas de l'entrée *s'évanouir* dont la construction de base est illustrée par *Luc s'évanouit*, la construction à sens statique en question est celle de *Luc est évanoui*. La propriété "N0 =: N-hum" indique que le sujet N_0 de la construction de base peut être occupé par un groupe nominal dénotant une entité non humaine. Pour l'entrée *blouser* décrite dans cette figure, cela correspond à des phrases telles que *Le chemisier blouse*.

Pour que les intitulés de propriétés soient mnémoniques, ils ont été construits à partir de symboles représentant des valeurs de traits : *N* pour substantif, *est* pour le verbe *être*, *pp* pour participe passé ; mais pour qu'ils soient succincts, on n'y a généralement pas précisé les traits correspondants, respectivement ici : catégorie grammaticale, verbe support, temps/mode.

Il serait hasardeux d'exploiter le LG dans un système de traitement des langues sans s'assurer qu'il peut être achevé et mis à jour, et donc qu'il existe sous un format lisible et éditable. Pourtant, peu d'auteurs ont pris en compte ce critère dans leur évaluation du LG. Hathout et Namer (1997, 1998), Gardent *et al.* (2005, 2006), par exemple, n'y font aucune allusion. Ils raisonnent en consommateurs et ne prennent pas en compte le début de la chaîne de production. La majorité des acteurs du traitement des langues considère d'ailleurs l'élaboration manuelle de ressources linguistiques comme un des cauchemars du domaine, un écueil à éviter et une source d'erreurs. Cette attitude découle probablement de goûts intellectuels des informaticiens, mais elle est objectivement irrationnelle, et le domaine devrait à notre avis la remettre en question.

Pour Gardent *et al.* (2005, 2006), le format du LG n'est pas standard, car les constructions ne sont pas sous la forme de structures de traits, avec noms de traits et noms de valeurs, comme celles utilisées par les systèmes actuels. En adoptant de telles conventions, les deux propriétés ci-dessus sont représentées par des formules telles de la figure 2, ou par des formules équivalentes en XML, encore moins concises.

```
construction:[predicate:[part-of-speech="verb", mode="participle",
                         tense="past"],
              support-verb:[part-of-speech="verb", lemma-list:[value="être"]],
```

```
        arguments:(constituent:[position="0",
                distribution:[component:[category="NP"]]])]
    constituent:[position="0",
        distribution:[component:[category="NP", human="false"]]]
```

Figure 2. Deux propriétés sous la forme de structures de traits.

Ces conventions sont manifestement incompatibles avec les exigences de compacité de l'édition manuelle sur format tabulaire. Les structures de traits sont un standard destiné à d'autres usages que la visualisation lisible. Les projets Comlex (Grishman *et al.*, 1994) et FrameNet (Fillmore et Atkins, 1994) n'ont pas non plus adopté un format de structures de traits pour l'édition et la mise à jour des lexiques. Dans la recherche, le respect des standards doit être accompagné d'une remise en question plus ouverte que dans l'ingénierie. C'est au contraire l'expérience issue de la construction du LG — et d'autres projets producteurs de lexiques à grande couverture pour le traitement des langues, mais il en existe peu — qui a vocation à nourrir la construction des standards et des normes. C'est le sens de la réflexion sur le format du LG effectuée par le projet Genelex (Alcouffe et al., 1993), qui fut une des sources du projet de normalisation Eagles. Le projet Lexsynt a également donné l'occasion de tenir compte du LG lors de l'élaboration de la norme LMF (Francopoulo *et al.*, 2006).

3. Degré de formalisation

Un des obstacles à l'utilisation du LG en traitement des langues est son degré de formalisation. Il est plus formalisé que le TLF (Dendien et Pierrel, 2003), dans lequel les propriétés syntaxico-sémantiques sont décrites par du texte ou suggérées par des exemples, et non spécifiées par des intitulés normalisés ; mais il l'est moins qu'un analyseur syntaxique.

3.1. Représentation des propriétés

Les propriétés syntaxico-sémantiques y sont représentées par des intitulés succincts (cf. section 2), moins précis que les formalismes utilisés par les analyseurs syntaxiques et les grammaires pour représenter les constructions syntaxiques. Par exemple, dans l'intitulé "N0 V vers N", qui représente une construction illustrée par *Des animaux divaguent vers le fleuve*, le symbole *N* représente un groupe nominal, déterminant compris, comme *le fleuve*. Dans l'intitulé "N0 V N1 Dnum N", qui représente la construction de *Luc loue son studio 400 euros*, le même symbole *N* représente cette fois-ci un substantif, alors que le déterminant, ici *400*, est symbolisé séparément par *Dnum*. Hathout et Namer (1997) notent ainsi que certaines informations sont implicites, non entièrement spécifiées ou représentées de façon non uniforme.

Dans les années 2000, les projets Lexsynt et LMF ont suscité chez les spécialistes de l'analyse syntaxique un renouveau de l'intérêt pour le LG. Cela a motivé la recherche de solutions à cette insuffisance de formalisation, notamment à travers l'utilisation de

réseaux de transitions récursifs (Paumier, 2001 ; Blanc, 2006) ou de formules plus précises que les intitulés. Cependant, de telles formules ne sauraient être aussi concises que ceux-ci (cf. section 2) : la solution n'est donc pas de substituer simplement ces formules aux intitulés, qui gardent leur raison d'être. Gardent *et al.* (2005) préconisent plutôt que les informations du LG soient rendues utilisables dans des systèmes de traitement des langues par un prétraitement qui les ferait passer à un niveau de formalisation équivalent à celui de la norme LMF, et éventuellement soient encodées en XML[2]. Constant et Tolone (2008) transcodent ainsi les informations du LG sous la forme d'ensembles de structures de traits comparables à celles de la figure 2. Ce traitement relie entre elles les propriétés qui contribuent à décrire une même construction, par exemple les deux propriétés mentionnées dans la section 2 : le LG lui-même ne les relie pas explicitement (Gardent *et al.* 2005), si ce n'est à travers le symbole N_0 contenu dans les deux intitulés. Le traitement de Constant et Tolone (2008) produit des formules plus appropriées au traitement des langues, mais ne résout pas définitivement le problème, car le métalangage syntaxico-sémantique varie de système à système et de théorie à théorie, et LMF ne cherche pas à le normaliser. Il est difficile de représenter les propriétés syntaxico-sémantiques par des formules à la fois complètes et satisfaisantes pour tous les systèmes et toutes les théories. D'autres traitements du même type peuvent donc être envisagés en parallèle.

Par ailleurs, un travail systématique sur les intitulés de propriétés a été engagé au LIGM, afin d'élever légèrement leur degré de formalisation, sans toutefois en modifier substantiellement les conventions, la compacité ou la lisibilité. Ainsi, les cartouches horizontaux qui matérialisent une classification des propriétés dans les versions imprimées des tables ont été supprimés en 2003-2004 : ils contribuaient certes à la lisibilité, et apportaient des informations, mais faisaient des intitulés de colonnes des objets complexes constitués de plusieurs étiquettes. Lors de la suppression de ces cartouches, les informations qu'ils contenaient ont été incorporées aux intitulés. Autre exemple, en 2009, la propriété "(N1)(de V1 W)" codée dans la classe de verbes 12 a été réintitulée "Qu Psubj =: Qu Ni Vsubj W = (Ni) (de Vi-inf W)". Cette propriété relie la construction illustrée par *Le ressort empêche la bague de glisser* à celle de *Le ressort empêche que la bague glisse* ; l'emploi du symbole N_1 pour désigner le sujet qui subit la montée, ici *la bague*, était critiquable car ce symbole désigne déjà par ailleurs l'ensemble de la complétive objet, ici *que la bague glisse* ; c'est pourquoi il a été remplacé par N_i.

3.2. Documentation des propriétés

Les propriétés syntaxico-sémantiques n'étant pas définies avec précision par leurs intitulés, elles sont documentées dans des publications scientifiques (pour les verbes

[2] Ce décalage entre le LG et les formalismes habituels de dictionnaires pour le traitement des langues explique probablement que certains spécialistes du domaine ne considèrent pas le LG comme un dictionnaire. Il a pourtant les caractères les plus fondamentaux d'un dictionnaire, notamment sa structure en entrées lexicales et de son contenu informatif.

distributionnels, Gross, 1975 ; Boons *et al*., 1976a, 1976b ; Guillet et Leclère, 1992). Mais cette documentation n'est pas suffisante :

- aucun de ces quatre ouvrages n'a été publié en traduction anglaise ;

- l'un d'entre eux est peu diffusé (Boons *et al*., 1976b) ;

- les définitions ne sont pas toujours assez précises pour les spécialistes d'analyse syntaxique, qui ne sont pas toujours spécialistes de syntaxe ;

- et un même intitulé recouvre parfois des propriétés différentes suivant les classes ; ainsi "N0 =: N-hum" indique que le sujet N_0 de la construction de base peut être occupé par un groupe nominal dénotant une entité non humaine, le verbe conservant son sens canonique (cf. ci-dessus *Le chemisier blouse*), sauf dans la classe 31H où ce même intitulé indique que la phrase prend alors un sens métaphorique, comme dans *Le paysage sommeille* à comparer avec *Luc sommeille*.

Hathout et Namer (1997) jugent ainsi l'interprétation des tables difficile. Pour remédier à ce problème, la documentation la plus complète, qui est celle de Guillet et Leclère (1992 : 409-430) a été entièrement revue, étendue à toutes les propriétés et traduite en anglais, au cours d'un travail collectif auquel a pris part Christian Leclère.

3.3. Délimitation des classes

Le LG répartit les entrées lexicales dans des classes. L'appartenance d'une entrée à une classe implique certaines propriétés syntaxico-sémantiques caractéristiques, qui délimitent la classe. Les auteurs ont choisi pour cela les propriétés les plus importantes : le nombre de compléments essentiels dans la construction de base, l'introduction de ces compléments par des prépositions, la possibilité qu'ils soient constitués d'une complétive, etc. Ainsi, la classe 9 (Gross, 1975) regroupe les verbes dont la construction de base est N_0 *V* N_1 *à* N_2, où le complément essentiel direct N_1 peut être occupé par une complétive, mais où le complément essentiel N_2 indirect en *à* ne le peut pas, comme dans *Luc dit qu'il pleut à tout le monde*. Cependant, les propriétés définissant les classes sont documentées de façon imprécise, soit informellement dans les textes, soit par des formules du même type que les intitulés. Elles ne figurent pas dans les tables : ainsi, la table 9 n'a pas de colonne intitulée "N0 V N1 à N2". Or cette propriété sert de référence pour la représentation des autres constructions, comme "N0 V à N2" (*Luc téléphone à tout le monde*), et pour les propriétés distributionnelles, comme "N0 =: N-hum" : la numérotation des arguments, ici N_0 et N_2, fait le lien[3]. Ces conventions semblent avoir compliqué la compréhension des propriétés par les utilisateurs. Gardent *et al.* (2005), par exemple, se demandent si les indices font référence à la position du constituant dans la construction de base ou dans une autre.

[3] Dans la version imprimée de la table 34L0 (Boons et al., 1976a), la numérotation des arguments de certaines constructions est indépendante de celle de la construction de base, et des cartouches horizontaux lèvent l'ambiguïté informellement. Lors de la suppression des cartouches horizontaux, les intitulés de cette table ont été rendus conformes aux conventions des autres tables.

Pour formaliser la définition des classes, il a été décidé de renseigner une table des classes, dans laquelle les propriétés sont attribuées non plus à des éléments lexicaux, mais à des classes entières (Constant et Tolone, 2008). Ce travail est en cours.

3.4. Délimitation des entrées lexicales

Comme pour tout lexique au sens linguistique, les objets de base du LG sont les entrées lexicales. En cas de polysémie, les entrées sont séparées : les entrées de *foncer* (dans *Luc fonce au port* et dans *Le pigment fonce les couleurs*) sont distinguées l'une de l'autre de la même façon que *foncer* l'est de *fonder* (dans *Luc fonde une agence*). Plusieurs constructions peuvent relever d'une même entrée ; ainsi, le LG ne consacre pas d'entrée distincte à *Les couleurs foncent* : il rattache cette construction à la même entrée que *Le pigment fonce les couleurs*, à travers une propriété intitulée "N1 V". Certaines classes font exception à ce principe. Ainsi, *Luc saupoudre du sel sur les frites* est décrit dans la classe 38LS avec la construction canonique N_0 *V* N_1 *Loc* N_2, où *Loc* désigne une préposition locative. La construction croisée *Luc saupoudre les frites de sel* est spécifiée dans cette entrée sous l'intitulé "N0 V N2 (de N1)", mais elle est également décrite indépendamment, de façon plus détaillée, dans la classe 37M4, avec une numérotation indépendante des arguments. Dans l'avenir, nous souhaitons rendre ces tables homogènes avec les autres sur ce point.

La spécification de plusieurs constructions dans une même entrée a parfois été mal acceptée par les utilisateurs. Ainsi, Gardent *et al.* (2005), à propos de la classe 1, ajoutent des entrées lexicales correspondant au remplacement d'une infinitive complément (*Max commence par examiner Luc*) par un complément humain (*Max commence par Luc*).

4. Degré de validité

Gardent *et al.* (2006) notent que certaines informations contenues dans le LG peuvent être incorrectes. Plusieurs sources d'erreurs expliquent en effet la présence d'informations invalides.

Il existe d'abord des erreurs matérielles. Des anomalies des programmes de gestion des tables ont inversé tous les signes + et - dans certaines entrées, par exemple *traîner là* [4] dans la classe 1. J'ai moi-même introduit au cours de la révision des intitulés (cf. section 3.1) plusieurs erreurs que j'ai corrigées en 2009.

Ensuite, certains verbes supports ont été codés dans le LG des verbes distributionnels, comme *faire* dans *Luc fait du tennis* ou *subir* dans *Le pétrole subit une hausse*. Comme il existe des tables de noms prédicatifs décrivant *tennis* ou *hausse*, il est nécessaire de collationner ces entrées et d'éliminer les entrées de verbes supports.

[4] Gardent *et al*. (2005) prennent cette entrée comme exemple, mais ne mentionnent pas l'erreur.

Enfin, les auteurs ont cherché à infléchir légèrement leurs jugements d'acceptabilité dans le sens de la tolérance. Gardent *et al.* (2005) trouvent ainsi certaines constructions assez peu probables. La description de *deviner*, par exemple, marque comme acceptable la construction N_0 *deviner N être Adj* (*Luc devine cette question être cruciale*, classe 6). Il faut toutefois noter que *je te devine être capable d'autant de répartie* est attesté [5] sur un blog dans un message de septembre 2008. Les auteurs du LG des verbes distributionnels du français n'ont pas cherché à appuyer leurs décisions sur des attestations dans des corpus. C'était irréalisable à l'époque (Boons *et al.*, 1976 : 37). Un contrôle de la validité plus objectif aurait été lourd et aurait compromis la faisabilité du projet. La couverture en informations a été préférée à l'objectivité. Aujourd'hui, confronter le LG avec un corpus serait un travail intéressant, mais il serait irréaliste de prétendre relier à des attestations toutes les informations contenues dans le LG. Ce dictionnaire représente un balayage du vocabulaire (13 000 entrées verbales, mais toutes n'ont pas été codées) croisé avec un balayage de 460 propriétés syntaxico-sémantiques, au cours duquel on teste les mêmes constructions sur les entrées rares comme *godailler* que sur les entrées fréquentes comme *bouillir*. Un corpus représente également un balayage croisé, mais partiel, sans la garantie que la totalité des combinaisons soit passée en revue ; il n'atteste pas d'inacceptabilités. Le choix de couvrir une grande masse d'informations justifie d'ailleurs en partie aussi celui d'un degré de formalisation limité (cf. section 3) : un appareil formel plus complexe n'aurait-il pas freiné l'application de tant de tests ?

La présence d'erreurs dans le LG ne doit pas faire oublier ses points forts en ce qui concerne la validité.

Le fait que l'on puisse repérer des erreurs est en soi un signe de falsifiabilité du LG au sens épistémologique : il se prononce explicitement sur des points vérifiables. Il reste d'ailleurs assez neutre par rapport aux différentes théories syntaxiques. Ses auteurs se sont concentrés sur des phénomènes relativement vérifiables, c'est-à-dire ceux pour lesquels l'observation est plus reproductible. Ils ont ainsi marqué l'aspect processif ou statique de certaines constructions, comme N_2 *V* N_1, illustré par *Le rideau cache le sac*, statique, à comparer à la construction de base de la même entrée, *Luc cache le rideau derrière le sac*, processive ; mais dans le cas de la construction N_1 *V Loc* N_2, le marquage de ce trait sémantique n'a pas été jugé suffisamment reproductible pour être systématisé : si l'aspect est nettement processif dans *Le volet claque contre le mur*, et nettement statique dans *Le carton tient contre la caisse*, l'intuition sémantique est moins claire dans *Le frein frotte sur la jante*. D'une manière générale, les auteurs du LG se sont entourés de précautions méthodologiques en vue d'assurer la reproductibilité de leurs observations (Laporte, 2008), et le recours à l'intuition y est plus sévèrement encadré que, par exemple, dans Levin (1993), d'où une base empirique plus solide [6].

[5] http://capmetz57.over-blog.com/article-22749572-6.html, 2 décembre 2009.

[6] Beth Levin connaissait pourtant le travail de Boons *et al*. (1976) (communication personnelle).

Il arrive qu'un cadre théorique ait une difficulté à prendre en compte un fait observé dans le LG : c'est probablement ce que Hathout et Namer (1997 : 5) entendent par « *certaines transformations sont linguistiquement incorrectes, dans le cadre théorique considéré* » (HPSG), et illustrent par la construction N_1 *se V auprès de Nhum de ce Qu P* (*Luc se réjouit auprès de Marie de ce que le film sorte*) [7]. Cependant, l'incorrection se situe plutôt du côté du cadre théorique que du phénomène observé. La neutralité par rapport aux théories syntaxiques explique par ailleurs le choix d'un degré de formalisation limité. Un formalisme plus complexe, nécessairement plus dépendant d'une théorie, n'aurait-il pas gêné l'observation éventuelle de faits auxquels cette théorie n'aurait pas été adaptée ?

5. Contenu informatif

Gardent *et al.* (2006 : 145-146) notent que certaines informations sont absentes du LG ou incomplètes, par exemple les fonctions grammaticales et les rôles thématiques, mais que d'autres propriétés, qui ne sont généralement pas utilisées par les analyseurs ou les générateurs, sont présentes, comme l'interprétation temporelle des infinitives. Le contenu informatif du LG est-il suffisant par rapport aux besoins d'un analyseur syntaxique ? Comment se situe-t-il par rapport aux autres lexiques structurés ?

Les fonctions grammaticales ne sont pas toutes codées, car elles recouvrent des propriétés syntaxico-sémantiques, généralement plus factuelles, avec lesquelles elles font en partie double emploi. Ainsi, la notion de complément d'objet direct se fonde sur différentes propriétés qui ne coïncident pas toujours : position après le verbe, absence de préposition, pronominalisation, passivation (Gross, 1969)... Ce sont plutôt ces propriétés qui sont codées dans le LG, ce qui est plus précis. En particulier, les auteurs du LG des verbes ont joué un rôle pionnier dans la délimitation entre les compléments essentiels (objets) et circonstanciels (adjoints, modifieurs). Ainsi, ils ont décrit comme complément essentiel le complément direct des verbes de la classe 32NM (*Luc chausse une grande taille, La pièce sent le jasmin*), souvent considéré comme circonstanciel. Il en est de même du complément indirect de nombreux verbes locatifs (*Luc place sa voiture contre le mur*). Ils ont également recensé de nombreux compléments qui ont un comportement intermédiaire entre ceux d'un complément essentiel et d'un complément circonstanciel, par exemple *sur ce point* dans *Luc se ravise sur ce point*.

En ce qui concerne les rôles thématiques et plus généralement la formalisation du sens, les auteurs du LG se sont limités à des phénomènes dont ils ont pu encadrer l'observation par des tests syntaxiques (cf. section 4).

[7] Cette construction à trois arguments, qui dénote un acte de parole, coexiste avec une construction à deux arguments (*Que le film sorte réjouit Luc*) ; dans celle-ci, on ne peut pas toujours considérer que le troisième argument, formellement absent, est en fait sémantiquement présent. Une telle situation est une anomalie par rapport à la plupart des théories actuelles.

On pourrait citer d'autres lacunes dans le contenu informatif du LG : certaines entrées ne sont pas encore codées, comme *basculer* et *boiter* (figure 1) ; les constructions dont la formation est régulière, comme la négation ou les propositions relatives, ont été négligées sauf lorsqu'elles varient en fonction des éléments lexicaux ; certaines constructions, notamment pronominales, ne sont pas codées ; les tables d'adjectifs sont en construction... Toutes ces informations sont certainement indispensables au bon fonctionnement des analyseurs syntaxiques fondés sur des lexiques et des grammaires.

Malgré ces limitations, il est difficile de contester la richesse du contenu informatif du LG, en comparaison avec d'autres lexiques structurés. Le balayage du lexique et le recensement des constructions sont impressionnants. La délimitation systématique entre constructions figées et constructions libres est difficile à trouver ailleurs, si ce n'est dans les lexiques-grammaires d'autres langues.

De telles avancées auraient-elles été possibles avec d'autres options méthodologiques ?

Conclusion

L'idée que le lexique-grammaire est difficilement exploitable pour le traitement des langues découle en partie de la présence d'erreurs et de lacunes, qui peuvent être corrigées, mais aussi d'un sentiment d'étrangeté que ressentent les spécialistes d'analyse syntaxique devant des choix qui sont peu courants dans la plupart des autres projets dont ils ont connaissance. Une prise en compte des différentes données du problème amène à nuancer cette vue, et à justifier la plupart de ces choix par les caractères originaux du lexique-grammaire : un vaste recensement du lexique et des constructions, la priorité donnée aux données factuelles sur les contraintes liées à des théories spécifiques, une exigence de reproductibilité des observations. Or ce sont justement ces caractères qui ouvrent des perspectives d'exploitation du lexique-grammaire dans des systèmes de traitement des langues.

Références

ALCOUFFE PH., REVELLIN-FALCOZ B., ZAYSSER L. (1993), « Azote : des tables du LADL au format Genelex », *Actes du Colloque Informatiques et Langues naturelles*, IRIN, Université de Nantes.

BLANC O. (2006), *Algorithmes d'analyse syntaxique par grammaires lexicalisées : optimisation et traitement de l'ambiguité*, Thèse de doctorat, I.G.M., Université de Marne-la-Vallée.

BOONS J.P., GUILLET A., LECLERE CH. (1976a), *La structure des phrases simples en français. 1 : Constructions intransitives*. Genève : Droz.

BOONS J.P., GUILLET A., LECLERE CH. (1976b), *La structure des phrases simples en français. Classes de constructions transitives*. Rapport de recherche du LADL n° 6, Université Paris 7.

CONSTANT M., TOLONE E. (2008), « A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables », M. Constant, M. De Gioia, T. Nakamura, S. Vecchiato (Éds.), *Colloque international sur le Lexique et la Grammaire (LGC), L'Aquila, Italie* : 11-18.

DENDIEN J., PIERREL J.M. (2003), « Le Trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence », dans *TAL* 44(2) : 11-37.

FILLMORE CH., ATKINS S. (1994), « Starting where the dictionaries stop: The challenge for computational lexicography », S. Atkins, A. Zampolli (Éds.), *Computational Approaches to the Lexicon*. Oxford University Press : 349-393.

FRANCOPOULO G., GEORGE M., CALZOLARI N., MONACHINI M., BEL N., PET M., SORIA C. (2006), « Lexical Markup Framework (LMF) », *Proceedings of LREC, Genoa* : 233-236.

GARDENT C., GUILLAUME B., PERRIER G., FALK I. (2005), « Extracting subcategorisation information from Maurice Gross' grammar lexicon », dans *Archives of Control Sciences* 2005 : 289-300.

GARDENT C., GUILLAUME B., PERRIER G., FALK I. (2005), « Le lexique-grammaire de M. Gross et le traitement automatique des langues », Journée ATALA : Interface lexique-grammaire et lexiques syntaxiques et sémantiques.

GARDENT C., GUILLAUME B., PERRIER G., FALK I. (2006), « Extraction d'information de sous-catégorisation à partir des tables du LADL », *Verbum ex machina*, Actes de TALN, Collection Cahiers du Cental numéro 2, volume 1 : 139-148, Presses universitaires de Louvain.

GRISHMAN R., MACLEOD C., MEYERS A. (1994), « COMLEX Syntax: Building a Computational Lexicon », *Proceedings of Coling* : 268-272.

GROSS M. (1969),. « Remarques sur la notion d'objet direct en français ». Dans *Langue Française* 1, : 63-73.

GROSS M. (1975), *Méthodes en syntaxe*, Paris, Hermann.

GUILLET A., LECLERE CH. (1992), *La structure des phrases simples en français. 2 : Constructions transitives locatives*. Genève, Droz.

HATHOUT N., NAMER F., (1997), « (Semi-)Automatic Generation of the ALEP Analysis Lexicon », *Proceedings of the 3rd ALEP User Group Workshop*, Saarbrücken.

HATHOUT N., NAMER F., (1997), « Génération (semi)-automatique de ressources lexicales réutilisables à grande échelle. Conversion des tables du LADL », *Actes des 1ères JST FRANCIL*, AUPELF-UREF, Avignon.

HATHOUT N., NAMER F. (1998), « Automatic Construction and Validation of French Large Lexical Resources: Reuse of Verb Theoretical Linguistic Descriptions », *Proceedings of LREC* : 627-636.

LAPORTE É. (2008), « Exemples attestés et exemples construits dans la pratique du lexique-grammaire », J. François (Éd.), *Observations et manipulations en linguistique: entre concurrence et complémentarité*. Mémoires de la Société de linguistique de Paris, nouvelle série, 16. Louvain/Paris/Dudley, Peeters : 11-32.

LEVIN B. (1993), *English Verb Classes and Alternations. A Preliminary Investigation*, The University of Chicago Press.

PAUMIER, S. (2001), « Some remarks on the application of a lexicon-grammar », dans *Lingvisticae Investigationes* 24(2) : 245-256.

ROCHE E. (1999), « Finite state transducers: Parsing free and frozen sentences », A. Kornai (Éd.), *Extended finite state models of language*, Cambridge University Press : 108-120.