Maintenance du Lexique-Grammaire

Formules définitoires et arbre de classement

Elsa Tolone*

* *FaMAF, Universidad Nacional de Córdoba, Argentine & LIGM, Université Paris-Est, France*

elsa.tolone@univ-paris-est.fr

RÉSUMÉ. *Les tables du Lexique-Grammaire, dont le développement a été initié par Gross (1975), constituent un lexique syntaxique très riche pour le français. Nous présentons dans cet article le travail réalisé afin de formaliser la classification existante pour les verbes distributionnels à l'aide de formules logiques, dans le but d'assurer la maintenance du lexique. Nous détaillons les différents types de propriétés définitoires, ainsi que leur codage dans la table des classes, regroupant l'ensemble des propriétés de toutes les classes de verbes. La définition formelle de l'ensemble des propriétés définitoires pour chaque classe a permis de représenter la classification sous forme d'arbre de décision, afin d'aider à trouver la classe associée à toute nouvelle entrée, ce qui permet de garantir la future alimentation du lexique.*

ABSTRACT. *Lexicon-Grammar tables, whose development was initiated by Gross (1975), are a very rich syntactic lexicon for the French language. This paper presents the work done to formalize the existing classification of verbs using logical formulas, in order to maintain the lexicon. We describe the different types of definitional features and their coding in the table of classes, which is composed by all the features of all classes of verbs. The formal definition of all the defining features for each class has to consider the classification with a decision tree to help finding the class associated with a new entry, which ensures the future extension of the lexicon.*

MOTS-CLÉS : *traitement automatique des langues, ressources linguistiques, lexiques syntaxiques, Lexique-Grammaire, table des classes, classification.*

KEYWORDS: *Natural Language Processing, language resources, syntactic lexica, Lexicon-Grammar, table of classes, classification.*

1. Introduction

Les tables du Lexique-Grammaire du français ont été développées à partir des années 1970 par Gross (1975), au sein du Laboratoire d'Automatique Documentaire et Linguistique (LADL), puis du Laboratoire d'Informatique Gaspard-Monge (LIGM) de l'Université Paris-Est (Boons *et al.*, 1976a ; Boons *et al.*, 1976b ; Guillet et Leclère, 1992). Ces informations se présentent sous la forme de *tables*. Chaque table correspond à une *classe* qui regroupe les éléments lexicaux d'une catégorie lexicale donnée (verbes, noms, adjectifs, etc.), partageant certaines propriétés syntaxico-sémantiques.

Une table se présente sous forme de matrice : en lignes, les entrées lexicales de la classe correspondante ; en colonnes, les propriétés syntaxico-sémantiques, qui ne sont pas forcément respectées par tous les éléments de la classe ; à la croisée d'une ligne et d'une colonne le signe $+$ ou $-$ indiquant que l'entrée lexicale décrite par la ligne accepte, ou n'accepte pas la propriété décrite par la colonne.

Actuellement, pour le français, nous disposons de 67 tables (et donc classes) de verbes distributionnels simples, la catégorie la mieux décrite, 81 tables de noms prédicatifs simples et composés, 69 tables d'expressions figées (principalement verbales et adjectivales) et 32 tables d'adverbes (adverbes en *-ment* et locutions adverbiales). Le nombre d'entrées est d'environ 13 800 pour les verbes distributionnels, 14 200 pour les noms prédicatifs, 39 600 pour les expressions figées et de 10 400 pour les adverbes.

Le travail décrit dans (Tolone, 2011) a montré que les informations présentes dans les tables étaient à présent exploitables informatiquement, grâce à l'intégration des tables dans l'analyseur syntaxique FRMG (*French MetaGrammar*) (Thomasset et de La Clergerie, 2005). Une évaluation sur le corpus EASy/Passage et le French TreeBank, ainsi qu'une fouille d'erreurs, ont été réalisées dans (Tolone *et al.*, 2012). La version 3.4 contenant l'ensemble des tables est téléchargeable sous une licence libre (LGPL-LR) sur le site `http://infolingu.univ-mlv.fr` (Données Linguistiques $>$ Lexique-Grammaire $>$ Téléchargement).

Avant d'envisager la conversion des tables en un lexique syntaxique pour le TAL, au format *LGLex* (Constant et Tolone, 2010), puis au format Le*fff* (Tolone et Sagot, 2011), de nombreuses modifications ont été effectuées dans les tables d'origine (Tolone, 2009), notamment concernant la construction de base de chacune des tables du Lexique-Grammaire (Tolone *et al.*, 2010). Les tables ne sont pas définies par une seule construction, mais par un ensemble de propriétés définitoires acceptées ou interdites, ainsi que par des disjonctions de propriétés, dont l'une d'entre elles seulement est obligatoire. Cet ensemble de propriétés est représenté par une formule logique et est appelé formule définitoire, comme nous le verrons tout au long de l'article.

Notre objectif est de formaliser la classification existante pour les verbes distributionnels à l'aide de ces formules définitoires, afin d'assurer la maintenance du lexique. Ceci a permis de représenter la classification sous forme d'arbre de décision, pour

trouver la classe associée à toute nouvelle entrée, ce qui facilite la future alimentation du lexique. Cet arbre reflète le découpage actuel des 13 872 verbes distributionnels en 67 classes, il permet donc de rendre explicite l'ensemble des propriétés déterminantes dans ce classement. Ce travail s'inscrit dans la finalité de l'amélioration des tables de verbes effectuée dans le but de les rendre utilisables pour la communauté de TAL.

Tout d'abord, la section 2 présente les tables elles-mêmes, les différentes versions existantes et la documentation complète de toutes les propriétés des verbes. Ensuite, nous précisons à la section 3, que chaque classe représente une sélection des entrées possédant un ensemble de propriétés définitoires, expression que nous définissons et illustrons à travers toutes ses représentations possibles : constructions de base, propriétés distributionnelles, autres constructions, etc. Dans la section 4, nous définissons ce qu'est une table des classes, en expliquant son rôle et l'homogénéisation des propriétés qui en résulte. Dans la section 5, nous exposons les avantages de cette représentation pour la génération d'un lexique syntaxique, en détaillant les particularités de notre approche et en la comparant avec des travaux antérieurs. Puis, nous introduisons à la section 6, ce que l'on appelle les formules définitoires, qui permettent de justifier la classification des tables. Enfin, la section 7 explique les conventions de lecture de l'arbre de classement des verbes distributionnels, servant à déterminer à quelle classe appartient une entrée verbale donnée.

2. Les tables

À titre d'illustration, le tableau 1 montre un extrait de la table 33 (Boons *et al.*, 1976b) (p. 252)[1] des verbes distributionnels qui se construisent avec un argument introduit par la préposition *à* [2]. Cela signifie qu'elle est composée des verbes ayant la propriété N0 V à N1 vraie.

N0 =: Nhum	N0 =: N-hum	N0 =: Nnr	<ENT>Ppv	Ppv =: se figé	Ppv =: les figé	Ppv =: Neg	<ENT>V	Neg	N0 V	N0 être V-ant	N0 V de N0pc	N1 =: Nhum	N1 =: N-hum	N1 =: le fait Qu P	Ppv =: lui	Ppv =: y	[extrap]	N0idée V Loc N1esprit	<OPT>
+	−	−	***les***	−	+	−	***lâcher*** *Advm*	−	+	−	−	+	−	−	−	−	−	−	***Max les lâche difficilement à Ida***
+	−	−	*<E>*	−	−	−	*renaître*	−	+	+	−	−	+	−	−	+	−	+	*Max renaît au bonheur de vivre*
+	−	−	***se***	+	−	−	***rendre***	−	+	−	−	+	+	+	−	+	+	−	***Max s'est rendu à mon opinion***
+	−	−	***se***	+	−	−	***rendre***	−	+	−	−	+	−	−	−	−	−	−	***Le caporal s'est rendu à l'ennemi***
+	−	−	*<E>*	−	−	−	*renoncer*	−	−	−	−	+	+	−	−	+	−	−	*Max renonce à son héritage*
+	+	+	***ne***	−	−	+	***revenir***	+	−	−	−	+	−	−	+	−	−	−	*La tête de Luc ne revient pas à Max*

Tableau 1. *Extrait de la table 33 des verbes distributionnels*

1. Il s'agit d'un extrait de la version 3.4 et non du livre d'origine.
2. Le tableau 2 explicite certains intitulés et l'annexe A contient l'ensemble des notations.

Propriétés	Description de la propriété
<ENT>V	Forme de l'entrée verbale V. Si un adverbe est figé avec le verbe, il fait partie de la forme. Si des clitiques sont figés avec le verbe, ils ne font pas partie de la forme. Si un adverbe négatif est obligatoirement présent dans la construction, mais a une valeur lexicale libre (*pas*, *point*, *nullement*, *aucunement*, *aucun*, *nul*, *personne*, *rien*, *guère*, *jamais*, *plus*, *nulle part*), il ne fait pas partie de la forme
(lexicale)	Exemple : *Ce film dure longtemps :* **durer Advt***; Luc n'arrête pas d'être dérangé :* **arrêter pas***; Votre geste ne va pas :* **aller**
<OPT>	Phrase d'exemple illustrant le sens de l'entrée verbale V
(lexicale)	Exemple : *Max achève de peindre le mur ; Max achève les blessés*
<ENT>Ppv	Clitiques figés avec le verbe V. Les informations données dans cette propriété doivent concorder avec celles données dans les propriétés binaires intitulées Ppv =: en figé, etc.
(lexicale)	Exemple : *Luc* **n'en** *revient pas de ce culot ; Luc* **se** *réserve pour la nuit*
Ppv =: en figé	Le pronom clitique *en/la/le/les/se/y* est figé avec le verbe V
(binaire)	Exemple : *Je n'en reviens pas de ce culot ; Luc s'en va*
Ppv =: la figé	*Fermez-la*
Ppv =: le figé	*La haine le dispute à la colère*
Ppv =: les figé	*On va les aligner*
Ppv =: se figé	*Luc se réserve pour la nuit ; Luc s'en va*
Ppv =: y figé	*Luc y va ; Luc s'y croit*
Ppv =: Neg	Le pronom clitique *ne* est figé avec le verbe V, même en l'absence d'un adverbe de négation
(binaire)	Exemple : *Luc n'arrête pas d'être dérangé ; Luc ne saurait dormir ici*
Neg	Présence obligatoire d'un adverbe de négation
(binaire)	Exemple : *Luc n'arrête pas d'être dérangé*
Prép1	Prépositions de l'objet N1/N2/N3
(lexicale)	Exemple : *Max va jusqu'à exiger des dommages :* **jusqu'à***; Le verre va tomber :* <**E**>*; Qu'Ida est idiote éclate aux yeux de tous :* **Loc**
Prép2	*Max a accrédité auprès des parents la nouvelle que Luc est mort :* **auprès de***; Max a encadré dans ce texte que Luc était absent :* **Loc***; Max a pour preuve de cela qu'il ne s'est pas montré :* **pour+comme**
Prép3	*Max désigne Luc à Léa pour faire ce travail :* **pour***; Max a reçu de Luc comme garantie qu'il aurait une prime :* <**E**>
autre Loc1	Prépositions de l'objet locatif Loc N1 autres que celles représentées dans les propriétés binaires intitulées Loc N1 =: à N1, etc.
(lexicale)	Exemple : *Les convives farandolent* **autour** *de la table ; Les délinquants se recrutent (***parmi+chez***) les riches*
autre Loc2	Prépositions de l'objet locatif Loc N2 autres que celles représentées dans les propriétés binaires intitulées Loc N2 =: à N2 destination, etc.
(lexicale)	Exemple : *Max a découché* **de chez** *Léa ; Max s'en vient* **chez** *vous*

Tableau 2. *Extrait de la documentation des propriétés lexicales*

Si un verbe a deux sens distincts, il possède deux entrées lexicales puisque chaque sens n'accepte pas le même ensemble de propriétés. Un des exemples qui figure dans la table 33 est le verbe *se rendre* :
Le caporal s'est rendu à l'ennemi.
Max s'est rendu à mon opinion.
On peut voir que *se rendre* (dans le sens d'accepter) possède un complément nominal non humain : la propriété N1 =: N-hum est vraie (codage +), alors qu'elle est fausse (codage −) pour *se rendre* (dans le sens de capituler). Il y a aussi des propriétés dont les valeurs sont des éléments lexicaux. Ainsi, les compléments prépositionnels peuvent nécessiter différentes prépositions qui dépendent de l'entrée. Nous avons d'autres exemples dans le tableau 1 : <ENT>Ppv, <ENT>V et <OPT>, codant respectivement les clitiques figés avec le verbe, la forme de l'entrée verbale et une phrase d'exemple illustrant le sens de l'entrée verbale. Un extrait de la documentation (voir 2.2) de ces propriétés est présenté dans le tableau 2. L'ensemble des notations utilisées dans les intitulés est détaillé dans l'annexe A.

2.1. *Versions des tables*

La version 1 d'origine est celle qui figure sous format papier dans la littérature et dont les plus anciennes tables existaient également à l'époque sous format électronique, sur des cartes perforées (deux par table, l'une contenant les intitulés des propriétés et l'autre les entrées avec leur codage). Au fil du temps, les tables ont subi divers changements de support électronique (cartes perforées, bandes magnétiques, disquettes) et de format (au fur et à mesure que les outils pour les éditer ont évolué)[3]. Certaines ont été perdues par manque d'intérêt de la part des auteurs, des institutions, ou de la communauté scientifique.

La version 2 représente 60 % des tables informatisées qui ont été mises en ligne en 2002 par Nathalie Bely sur le site `http://infolingu.univ-mlv.fr/` (Données Linguistiques > Lexique-Grammaire > Visualisation). Un système a été mis en place, permettant d'effectuer une recherche par verbe et par table, avec la possibilité d'afficher les exemples des verbes sélectionnés dans les tables sélectionnées. De plus, un téléchargement est possible en XML, et une documentation est associée à chaque table. Cette documentation indique les propriétés définitoires et donne des exemples pour une entrée lexicale afin d'illustrer les différentes propriétés.

3. De même qu'avec les cartes perforées, deux fichiers au format texte permettaient de représenter une table, grâce au programme d'édition d'A. Guillet (nommé EDIX et enregistré sur deux disquettes MS-Dos) qui présentait tout dans le bon ordre. Chaque intitulé de propriété était sur une ligne dans le premier fichier et le programme les écrivait de telle sorte qu'on pouvait lire les intitulés en colonne. Le deuxième fichier contenait en ligne le numéro de la ligne, l'entrée, une barre oblique et une série de + ou − dans l'ordre par rapport aux intitulés énumérés dans le premier fichier. Voici par exemple, une ligne de la table 31H :
7 s'agiter/− − + − − + − + + + +

Les récentes modifications ont donné jour à la version 3 en septembre 2008, également téléchargeable sur ce même site (Données Linguistiques > Lexique-Grammaire > Téléchargement), qui est régulièrement actualisée. La version 3.4 (5 octobre 2011) contient la totalité des tables dans toutes les catégories, sous tous les formats disponibles (Tolone, 2011). La version 3.4 contient de plus une documentation exhaustive sur toutes les propriétés syntaxico-sémantiques des verbes, ainsi que la définition formelle de chaque table (voir section 6) et l'arbre de classement (voir section 7).

2.2. *Documentation des propriétés*

Les propriétés syntaxico-sémantiques ne sont pas définies avec précision par leurs intitulés. Elles sont documentées dans des publications scientifiques mais cela reste insuffisant :

– toutes ne sont pas documentées, comme c'est le cas pour certaines classes d'expressions figées de M. Gross qui n'ont même pas été publiées ;

– leur documentation est parfois difficilement accessible, car certains ouvrages sont moins diffusés que d'autres. C'est le cas des rapports ou des thèses n'ayant pas débouché sur une publication : Boons *et al.* (1976a) pour les verbes, Meunier (1981) et Giry-Schneider et Balibar-Mrabti (1993) pour les noms prédicatifs ;

– aucun ouvrage n'a été traduit en anglais ;

– les définitions manquent de précision ;

– un même intitulé peut avoir différentes interprétations et représenter une propriété linguistique différente en fonction des classes ; ainsi N0 =: N-hum indique que le sujet N0 de la construction de base peut être occupé par un groupe nominal dénotant une entité non humaine, le verbe conservant son sens canonique (voir 3.2 avec *Le chemisier blouse*), sauf dans la classe 31H (Boons *et al.*, 1976b) (p. 259) où ce même intitulé indique que la phrase prend alors un sens métaphorique, comme dans *Le paysage sommeille*[4], à contraster avec *Luc sommeille* ;

– deux intitulés similaires peuvent avoir une signification différente dans deux tables distinctes, ou éventuellement dans une même table. Ainsi, la table 36DT (Guillet et Leclère, 1992) (p. 123 et 237) comporte essentiellement des verbes prenant un objet direct non humain concret. La propriété N1 =: Nhum y a deux rôles : d'une part elle marque la possibilité pour certains verbes de produire des métaphores (*Paul emprunte une secrétaire au patron*, par rapport à *Paul emprunte cent francs au patron*) ; d'autre part, elle note une sous-classe particulière de constructions où l'objet échangé est strictement humain (*Paul délègue sa secrétaire au patron*). Ce dernier cas est séparé de l'autre par le codage − de la colonne N1 =: N-hum.

L'interprétation de certains intitulés peut donc être difficile. Pour remédier à ce problème, la documentation des propriétés la plus complète, qui est celle des verbes

4. Nous avons remplacé l'intitulé N0 =: N-hum par N0 =: N-hum métaphore dans la table 31H pour distinguer ce cas.

locatifs (Guillet et Leclère, 1992) (p. 409-430) a été entièrement revue, étendue à toutes les propriétés des verbes distributionnels, et traduite en anglais[5] (cf. tableau 2).

De plus, cela a permis de vérifier pour toutes les classes de verbes à quelle signification chaque intitulé faisait référence, l'objectif étant qu'un intitulé dénote une seule propriété linguistique, qui elle-même n'est désignée que par un seul intitulé dans l'ensemble des tables. Elle est à présent complète et mise à jour dès qu'une modification a lieu dans une table. Elle est incluse dans la version 3.4 et dans l'annexe E de (Tolone, 2011).

3. Découpage en classes

3.1. *Codage des propriétés définitoires*

Les tables du Lexique-Grammaire répartissent les entrées lexicales dans des classes. Chaque classe regroupe un certain nombre d'entrées jugées similaires, car elles acceptent des propriétés syntaxico-sémantiques communes, que l'on appelle les *propriétés définitoires*[6].

Les propriétés définitoires de ces classes relèvent généralement du cadre de sous-catégorisation. Ainsi, les critères les plus communément utilisés dans les propriétés définitoires sont le nombre de compléments, la nature prépositionnelle ou non des compléments (pour les compléments prépositionnels, sont distingués ceux qui sont introduits par les prépositions *à*, *de*, *avec*, Loc, etc.), la catégorie grammaticale du sujet et des compléments (sont distinguées les réalisations sous forme de complétive, notée Qu P, d'infinitive, notée V-inf W, et de syntagme nominal, notée N suivi d'un trait sémantique, comme par exemple Nhum ou N-hum, comme cela est détaillé dans l'annexe A).

Les propriétés définitoires d'une classe sont les propriétés communes à toutes les entrées figurant dans la table correspondante. Elles définissent des constructions possibles pour ces entrées, mais elles ne suffisent en aucun cas à elles seules à définir les constructions d'une entrée. Pour cela, il faut tenir compte, en plus des propriétés définitoires définissant cette classe, du codage dans la table de l'ensemble des propriétés de l'entrée concernée.

Par exemple, la table 9 (Gross, 1975) (p. 190) a parmi ses propriétés définitoires : N0 V N1 à N2 (dans cette construction, N0 représente le sujet, V le verbe, N1 le premier argument, N2 le deuxième), où le complément essentiel direct N1 peut être occupé par une complétive : cette table regroupe des verbes comme *dire*, *dissimuler* et *ordonner*, dont le cadre de sous-catégorisation peut se caractériser par

5. Nous avons réalisé ce travail en collaboration avec Éric Laporte et Christian Leclère en 2008-2011.

6. Nous préférons le terme propriété définitoire à celui de *propriété définitionnelle*, mais ils sont tous les deux employés dans la littérature de manière équivalente.

une complétive objet et un complément nominal introduit par la préposition *à* (*Luc a (dit+dissimulé+ordonné) à Marie que Zoé chante*). On peut remarquer que à N2 apparaît avant N1, en accord avec le fait que les propriétés définitoires n'imposent pas d'ordre sur les compléments. Le fait que l'argument N1 puisse être une complétive est codé dans la table elle-même pour différencier les complétives à l'indicatif (N1 =: Qu Pind), au subjonctif (N1 =: Qu Psubj) et celles qui sont interrogatives (N1 =: si P ou si P), et pour permettre également de reconnaître d'autres catégories grammaticales, sans qu'elles soient obligatoires pour figurer dans cette table.

La table 5 (Gross, 1975) (p. 172) a parmi ses propriétés définitoires : N0 V Prép N1, mais aussi N0 =: Qu P, ce qui signifie que le sujet peut être une complétive. Ici c'est la préposition qui est codée dans la table, car elle varie en fonction des entrées (*Qu'Ida allait partir cheminait* **dans** *sa tête* / *Que Max s'est enfui circule* **sur** *son compte*). Le cadre de sous-catégorisation est donc défini par une complétive sujet et un complément nominal introduit par la préposition spécifiée dans la table.

La table 37M1 (Guillet et Leclère, 1992) (p. 130), a parmi ses propriétés définitoires : N0 V N1 Prép N2 et Prép2 =: de[7]. Contrairement à la table 5, toutes les entrées acceptent la préposition *de* pour l'argument N2 (*Max abrutit ses élèves* **de** *travail* / *On a doté l'hôpital* **de** *scanners*), mais aussi d'autres prépositions qui sont codées dans la table (*Max abrutit ses élèves* **avec** *du travail* / *On a doté l'hôpital* **en** *scanners*). La table 9 n'accepte que la préposition *à* pour l'argument N1, c'est pourquoi la propriété définitoire N0 V N1 à N2 suffit ici pour préciser à la fois la préposition et le nombre d'arguments. Le cadre de sous-catégorisation pour la table 37M1 est un complément nominal objet et un complément nominal introduit par la préposition *de*, mais aussi *avec* ou *en*, etc., selon les entrées.

3.2. *Propriétés définitoires acceptées*

Les propriétés définitoires sont constituées d'au moins une construction, dite *construction de base*. Elle comporte le sujet et tous les compléments essentiels (Boons *et al.*, 1976b) dont dépendent la majorité des constructions (par exemple, lors de l'explicitation de la réalisation d'un argument). Ainsi, N0 V N1 à N2 est la construction de base de la table 9, alors que N0 V Prép N1 est celle de la table 5 et enfin, N0 V N1 Prép N2 celle de la table 37M1. La propriété N0 =: Qu P définitoire de la table 5 est une *propriété distributionnelle*, qui spécifie la catégorie grammaticale de l'argument N0, déjà défini dans une construction, souvent la construction de base. La propriété Prép2 =: de est également une propriété distributionnelle, définitoire de la table 37M1, qui spécifie la valeur de la préposition[8].

7. Remarquons que Prép2 fait référence à la préposition du deuxième complément, même si elle n'est pas numérotée dans la construction.

8. Remarquons que la construction de base de la table 9 pourrait être également N0 V N1 Prép N2, avec Prép2 =: à définitoire, mais nous avons préféré intégrer la préposition dans la construction de base lorsqu'une seule était possible.

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. La deuxième construction est appelée *propriété transformationnelle*, car elle est déductible de la première par une redistribution, la première étant la construction de base. Ainsi la table 35S (Boons *et al.*, 1976b) (p. 207) regroupe les verbes intransitifs *symétriques* qui se caractérisent par la construction de base N0 V Prép N1 et par la deuxième construction définitoire N0 et N1 V (*Luc flirte avec Zoé* / *Luc et Zoé flirtent (ensemble)*).

Enfin, de nombreuses propriétés définitoires incluent des traits sémantiques élémentaires. Par exemple, des informations sur les classes des noms têtes des syntagmes nominaux (humain, concret, pluriel, etc.) : ainsi, la table 31R (Boons *et al.*, 1976b) (p. 262) admet la propriété distributionnelle N0 =: N-hum indiquant que le sujet N0 de la construction de base peut être non humain (*Le chemisier blouse*)[9].

Ou encore, des informations sur la sémantique des procès : ainsi les verbes entrant dans la construction N0 V N1 Prép N2, avec Prép =: de et dont le N1 dénote soit un lieu par rapport auquel est située l'entité à laquelle réfère N2, soit le bénéficiaire ou le détrimentaire du procès, ont été divisés en deux ensembles : la table 37E (Guillet et Leclère, 1992) (p. 123) regroupe les procès d'enlèvement (*Luc a débarrassé le grenier de ses caisses*, qui par la propriété transformationnelle N0 enlever N2 de N1 devient : *Max enlève les caisses du grenier*), tandis que les tables 37M1 à 37M6 (Guillet et Leclère, 1992) (p. 130) regroupent les procès d'ajout (*Luc a muni la porte d'un verrou*, qui par la propriété transformationnelle N0 mettre N2 Loc N1 devient : *Luc met un verrou dans la porte*). Les tables 37M1 à 37M6 se distinguent par des propriétés très diverses (morphologiques, sémantiques ou autres) qui ne relèvent pas du cadre de sous-catégorisation et qui auraient pu 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 laborieuse, d'où sa division en six sous-tables.

3.3. *Propriétés définitoires interdites*

Dans tous les cas cités précédemment, les propriétés définitoires correspondent à des propriétés qui sont vraies pour toutes les entrées d'une table, mais l'inverse est possible également, à savoir des propriétés interdites pour l'ensemble des entrées d'une table.

Reprenons, par exemple, la table 9 qui a comme construction de base N0 V N1 à N2, avec une complétive en N1. Il faut exclure la possibilité d'une complétive en N2, sinon cela inclurait certaines entrées de la table 16 (Gross, 1975) (p. 208) par exemple, qui a comme construction de base N0 V Prép N1 Prép N2, avec

9. La propriété N0 =: N-hum ne signifie pas que le sujet doit obligatoirement être non humain, puisque cela dépend du codage de la colonne N0 =: Nhum qui figure dans la table.

une complétive en N1 et en N2. En effet, les prépositions des arguments sont codées dans la table 16 et peuvent être <*E*> pour N1 et *à* pour N2 (*Max assimile que Luc fasse grève à ce qu'il désobéisse*).

De même, la délimitation de la table 37M1 (N0 V N1 Prép N2, avec Prép =: de) inclut la négation de propriétés de complétives, sinon par exemple une partie de la table 10 (Gross, 1975) (p. 193), qui a comme construction de base N0 V N1 Prép N2, avec une complétive en N1 (*Le maire requiert du préfet que la police intervienne*), se trouve incluse dans la table 37M1.

C'est le cas également de la table 32NM (Boons *et al.*, 1976a) (p. 73) qui n'accepte pas la *redistribution* passive (*Cette valise pèse 10 kilos* / **10 kilos sont pesés par cette valise*). Par exemple, on peut interdire les propriétés [passif par] et [passif de], utilisées selon que le complément d'agent est introduit par la préposition *par* ou *de*, et qui indiquent que tous les autres objets sont conservés au passif. Ces deux propriétés font partie des propriétés transformationnelles, puisque ce sont des redistributions à partir de construction de base.

Enfin, nous pouvons voir le cas de la table 31H, qui admet la propriété distributionnelle N0 =: Nhum obligatoirement, ce qui signifie que les entrées doivent (et non peuvent) avoir un sujet humain (sinon les entrées se trouvent dans la table 31R, voir 3.2) : la propriété N0 =: Nhum est donc toujours vraie alors que la propriété N0 =: N-hum n'est jamais acceptée.

3.4. *Codage des propriétés définitoires*

Cependant, les propriétés définitoires vraies pour toutes les entrées d'une table ne figurent pas dans les tables : ainsi, la construction de base de la table 9 est N0 V N1 à N2, mais la table n'a pas de colonne intitulée N0 V N1 à N2, car c'est une information implicite qui est uniquement décrite dans la littérature. Or, cette propriété sert de référence pour la représentation des autres constructions, comme la propriété transformationnelle N0 V à N2 dans la table 9 (effacement du N1 : *Luc téléphone à tout le monde*), et pour les propriétés distributionnelles, comme N0 =: N-hum dans la table 31R (sujet de type non humain : *Le chemisier blouse*, la construction de base étant N0 V).

Citons Vivès (1983), qui résume bien la préoccupation qui à cette époque était liée à la taille des données : « Dans les tables que nous avons établies, nous avons cherché à fournir les indications les plus significatives en essayant de limiter le nombre des colonnes. Lorsqu'une propriété est générale ou quasi-générale pour l'ensemble de la table, nous ne l'avons pas représentée, nous contentant de la signaler dans les commentaires consacrés à chaque table. Lorsqu'une propriété n'est vérifiée que par un nombre réduit d'éléments dans une table, nous avons adopté la même solution : cette propriété fait l'objet d'un commentaire. »

Nous nous sommes donc référée aux commentaires de chaque table dans la littérature, afin d'ajouter toutes les propriétés définitoires vraies pour toutes les entrées d'une table.

4. Tables des classes

4.1. *Création des tables des classes*

Le fait que les tables ne codent pas explicitement leurs propriétés définitoires, est un problème important pour leur exploitation automatique, puisque ces propriétés sont parmi les plus importantes. Les critères de découpage en classes et les propriétés définitoires ne sont décrits que dans la littérature associée aux tables. Ce constat a motivé depuis quelques années le développement au LIGM d'un nouveau type de tables, appelées *tables des classes* (Constant et Tolone, 2010). La notion de table des classes a été définie suivant Paumier (2003)[10]. Son rôle est d'assigner des propriétés syntaxico-sémantiques à une classe quand cela est possible, c'est-à-dire, quand leur valeur est constante pour toute une classe (par exemple, les propriétés définitoires d'une classe).

Ces tables sont au nombre d'une par catégorie grammaticale, donc quatre au total : une pour les verbes, une pour les noms prédicatifs, une pour les expressions figées et une pour les adverbes. Une table des classes regroupe en lignes l'ensemble de toutes les propriétés syntaxico-sémantiques répertoriées pour la catégorie concernée, et liste en colonnes l'ensemble des classes définies pour cette même catégorie. À l'intersection d'une ligne et d'une colonne, le signe $+$ (respectivement, $-$) indique que la propriété correspondante est vérifiée (respectivement, non vérifiée) par tous les éléments de la classe (c'est-à-dire par toutes les entrées de la table correspondante). De plus, le signe $+$ est doublé lorsqu'il s'agit de la construction de base. Le signe o indique que la propriété est explicitement codée dans la table concernée, car elle est vérifiée par certaines de ses entrées mais pas toutes[11]. Le signe O indique que la propriété n'est pas codée dans la table concernée mais devrait l'être, car elle dépend des entrées. Enfin, le signe ? indique une cellule non encore renseignée.

Par exemple, la table des classes des verbes regroupe les 67 classes de verbes distributionnels et l'ensemble des 552 propriétés syntaxico-sémantiques. Un extrait de cette table est donné dans le tableau 3. La notation V_ indique qu'il s'agit de classes

10. Elle était appelée au départ *super-table* et elle comportait quelques variantes puisque son objectif était de l'utiliser avec le logiciel Unitex afin de générer un graphe par table à partir d'un graphe générique faisant appel à cette super-table, puis de créer un graphe lexical pour chaque entrée à partir de ces graphes et des tables elles-mêmes.

11. Dans Paumier (2001), on faisait référence à la propriété codée dans la table par une variable : @A pour la première colonne contenant la première propriété à droite, @B pour la deuxième, etc. De plus, on pouvait utiliser la négation d'une propriété avec la variable !@A pour récupérer le codage inverse sans devoir créer une colonne le contenant.

Propriété \ table	V_1	V_2	V_4	V_5	V_9	V_31R	V_31H	V_33	V_32H	V_37M1
N0 =: Nhum	o	+	o	o	o	o	+	o	o	+
N0 =: N-hum	−	O	−	+	−	+	−	o	−	?
N0 =: N-hum métaphore	?	?	?	?	?	?	o	?	?	?
N0 = : Qu P	O	−	+	+	?	?	?	O	−	?
<ENT>V	o	o	o	o	o	o	o	o	o	o
Ppv =: se figé	o	o	−	o	−	o	o	o	o	−
N0 V	o	o	o	o	o	++	++	o	?	?
N0 V N1	−	−	++	?	o	−	−	?	++	o
N0 V à N1	O	−	?	?	?	−	−	++	?	?
N0 V Prép N1	−	−	?	++	−	−	−	?	?	?
N0 V Prép V0-inf W	++	−	−	−	−	−	−	−	−	−
N0 V N1 à N2	−	−	−	−	++	−	−	−	−	−
N0 V N1 Prép N2	−	−	?	?	−	−	−	?	?	++
N0 V Loc N1 V0-inf W	−	++	?	?	?	−	−	?	?	?
N0 V à N2	−	−	?	?	o	−	−	?	?	?
N1 =: Nhum	O	−	+	o	o	−	−	o	+	o
N1 =: N-hum	O	−	?	o	o	−	−	o	−	o
N1 =: Qu Pind	−	−	O	−	o	−	−	−	−	−
N1 =: Qu Psubj	−	−	O	−	o	−	−	−	−	−
N1 =: si P ou si P	−	−	O	?	o	−	−	−	−	−
N2 =: Qu Pind	−	−	?	−	−	−	−	−	−	−
N2 =: Qu Psubj	−	−	?	−	−	−	−	−	−	−
Prép2 = : à	−	−	?	?	−	−	−	?	?	−
Prép2 = : de	−	−	?	?	−	−	−	?	?	+

Tableau 3. *Extrait de la table des classes des verbes distributionnels*

de verbes. Dans cette table, on peut voir la construction de base de la classe 33 (cf. section 1) : la construction N0 V à N1 est codée + car elle vaut + pour l'ensemble des entrées de la table 33. Les propriétés distributionnelles N1 =: Nhum et N1 =: N-hum se voient assigner o pour la table 33 parce qu'elles dépendent des entrées lexicales. En revanche, pour la table 32H (Boons *et al.*, 1976a) (p. 75), la propriété distributionnelle N1 =: Nhum est codée + alors que la propriété distributionnelle N1 =: N-hum est codée − car l'objet est obligatoirement humain. Les deux propriétés distributionnelles sont donc définitoires de la table, la deuxième étant constante − pour la table (cas similaire à la table 31H avec le sujet, voir 3.3).

Remarquons que le codage − dans la table des classes désigne réellement dans ce cas une propriété que n'accepte aucune des entrées. Dans d'autres cas, il peut désigner simplement une information non pertinente (par exemple, une propriété de N1 pour une entrée qui n'a pas de N1). Cela vient du fait que certaines propriétés sont spécifiques à une table donnée. Elles ne seraient pas pertinentes pour une autre ; ainsi,

N2 bénéficiaire ne figure qu'en 36DT. L'absence d'une propriété dans une table peut donc signifier qu'elle est constante + ou − pour toutes les entrées de cette table, ou qu'elle n'a pas été jugée intéressante, ou encore qu'elle n'est pas pertinente pour la table en question.

4.2. *Homogénéisation et codage des propriétés*

La table des classes fait apparaître les propriétés définitoires, toutes celles qui sont constantes + ont été codées. Notons que dès que l'on peut coder une information avec différentes combinaisons de propriétés, un choix arbitraire a été fait [12] et les autres possibilités ont été codées − dans la table des classes pour ne pas engendrer de redondance. Par exemple, pour la table 9, la construction de base est N0 V N1 à N2, mais aurait pu également être N0 V N1 Prép N2 avec Prép2 =: à définitoire, mais nous avons systématiquement intégré la préposition dans la construction de base lorsqu'une seule était possible. Les deux propriétés N0 V N1 Prép N2 et Prép2 =: à sont donc codées − dans la table des classes.

La table des classes permet également de coder toutes les propriétés syntaxico-sémantiques pour chaque table, y compris celles dont on ne trouve la description que dans la littérature et qui de ce fait, ne sont pas exploitables alors qu'elles peuvent être pertinentes. C'est le cas, par exemple, des propriétés transformationnelles [passif par] et [passif de], qui sont fréquentes en français. Ces propriétés fondamentales ont en général été étudiées et décrites dans les thèses, ce qui signifie que lorsqu'elles ne sont pas codées dans une table, elles sont constantes + ou − (avec peut-être quelques exceptions répertoriées dans les thèses), sans pour autant être définitoires si elles n'ont pas participé au découpage de la classe en question. Cet ensemble de propriétés est à coder prioritairement dans la table des classes, mais il n'est pas toujours facile à repérer.

En outre, le choix des propriétés codées dans chaque table ayant été arbitraire, certaines propriétés peuvent, après réflexion, se révéler pertinentes, soit constante + pour la table, soit variant selon les entrées. Dans ce dernier cas, cela signifie qu'il s'agit d'une propriété qui devrait être codée o dans la table des classes et codée dans la table. En attendant d'être effectivement codée dans la table, elle est codée O dans la table des classes.

La construction de ces tables des classes a permis une homogénéisation importante des tables et en particulier des intitulés de colonne. C'est ainsi que de nombreuses colonnes ont changé de nom (Tolone, 2009). Cela a permis également de revoir la notation des constructions de base de chaque table (Tolone *et al.*, 2010).

12. Le critère de notation principal utilisé pour définir les propriétés définitoires est de noter les arguments N0, N1, etc. sans indication sur leur distribution et d'ajouter ensuite des propriétés ditributionnelles, telles que N1 =: Qu P, afin que chacun des arguments soit numéroté et que l'on puisse s'y référer dans les autres propriétés.

5. Avantages pour la génération d'un lexique

5.1. *Génération d'un lexique syntaxique*

La version actuelle des tables accompagnées des tables des classes nous a permis d'envisager une utilisation de ces données lexicales dans des outils de TAL (Tolone, 2009). En effet, pour construire un lexique syntaxique, il est nécessaire pour une entrée avec un sens donné, de connaître de façon exhaustive les propriétés syntaxiques de toutes ses occurrences en contexte dans des corpus. Or, l'ensemble de ces propriétés syntaxiques est formé des propriétés codées + dans la table des classes pour la table où figure l'entrée, qu'elles soient définitoires ou non, et des propriétés codées + dans la table pour cette entrée[13]. L'outil *LGExtract* (Constant et Tolone, 2010) permet ainsi de regrouper l'ensemble des propriétés syntaxiques pour chaque entrée et de générer le lexique syntaxique *LGLex* à partir des tables du Lexique-Grammaire [14].

L'outil *LGExtract* utilise une approche globale. D'abord, il est relié à la table des classes, qui code les informations qui ne sont pas définies dans les classes d'origine. Ensuite, comme une propriété syntaxique a exactement une interprétation pour l'ensemble des classes (après notre travail d'homogénéisation), notre script d'extraction assigne une seule fois un ensemble d'opérations pour chaque propriété. Ces opérations permettent de combiner les propriétés entre elles, principalement, toutes les propriétés distributionnelles, dont la pertinence dépend des constructions. En revanche, les constructions sont juste listées, puisqu'elles ne sont pas dépendantes les unes des autres. En effet, les tables n'autorisent pas la présence simultanée de deux constructions. Les principes de fonctionnement de *LGExtract* sont expliqués dans (Constant et Tolone, 2010). L'outil permet par exemple de factoriser les propriétés N1 =: Qu Pind et N1 =: Qu Psubj en indiquant dans le lexique *LGLex* que deux réalisations possibles de l'argument N1 sont : *complétive à l'indicatif* et *complétive au subjonctif*. Celles-ci s'ajoutent à la liste des réalisations déjà présentes pour l'argument N1, cette liste pouvant être vide.

En ce qui concerne les différents cadres de sous-catégorisation, ils sont factorisés lors de la conversion au format L*efff*, comme cela est détaillé dans (Tolone et Sagot, 2011)[15]. Parmi les constructions listées dans *LGLex*, nous identifions les nouveaux arguments qui rallongent la construction de base afin de calculer la construction de base maximalement étendue (CBME). Par exemple, l'entrée *rassembler* dans la table 32PL (Boons *et al.*, 1976a) (p. 5 et 24) a pour construction de base la construction

13. Certaines entrées ne sont pas codées (codage ∼). C'est pourquoi l'amélioration du lexique dépend de la poursuite du codage des tables.

14. L'outil *LGExtract* ainsi que le lexique généré *LGLex* sont disponibles sur le site `http://infolingu.univ-mlv.fr/` (Données Linguistiques > Lexique-Grammaire > Téléchargement).

15. Le programme de conversion *LGLex2ilex* ainsi que le lexique généré *LGLex*-L*efff* sont disponibles sur le site `http://infolingu.univ-mlv.fr/` (Données Linguistiques > Lexique-Grammaire > Téléchargement).

transitive simple N0 V N1, mais la possibilité d'ajouter un complément introduit par la préposition *en* codée dans la table conduit à une CBME de la forme N0 V N1 en N2 (*Max a rassemblé ses articles (E+en un ouvrage)*). C'est en effet en partant de la construction de base (d'où l'importance de l'avoir définie) qu'un alignement est fait avec les autres constructions pour identifier la CBME et les variantes de la construction de base. Ces variantes sont obtenues par effacement d'un ou de plusieurs arguments ou par changement de type de réalisation. Dans l'exemple précédent, la construction N0 V N2 codée dans la table indique que l'effacement de l'argument N1 est possible avec suppression de la préposition introduisant l'argument N2 (*Max a rassemblé un ouvrage*).

Les dépendances au sein d'une construction (par exemple, tel argument est effaçable seulement si tel autre l'est aussi) sont perdues : tout effacement d'un argument rend l'argument facultatif sans condition, alors même qu'il peut ne venir que d'une seule construction. De même, toute réalisation d'un argument autorisée par une construction devient possible quelles que soient les réalisations des autres arguments. Il s'agit là formellement d'une approximation des données linguistiques présentes dans les tables (et dans le lexique *LGLex*). Cette approximation a le mérite de permettre de diminuer au maximum le nombre d'entrées. Cependant, il reste à démontrer qu'elle n'a pas de conséquences pour l'utilisation du lexique produit, puisqu'en théorie les dépendances entre effacements permettent d'éliminer des analyses et donc de lever des ambiguïtés. Cette approximation est également adoptée par le modèle de la valence mis en œuvre par le lexique DICOVALENCE (van den Eynde et Mertens, 2006).

5.2. *Particularité du lexique LGLex*

LGExtract a vocation à transformer les tables en un lexique syntaxique indépendant d'une théorie particulière, reposant sur les mêmes concepts linguistiques que ceux qui sont à l'œuvre dans les tables. C'est-à-dire que le format du lexique *LGLex* est ce qu'on appelle un format d'échange. Il a vocation à décrire les tables avec les concepts manipulés par celles-ci, en un format directement exploitable dans les applications de TAL. L'une des utilisations informatiques possibles est la conversion en un autre format, par exemple au format L*efff*, ce qui suppose de manipuler les concepts linguistiques du L*efff*.

En effet, contrairement à *LGLex*, qui liste toutes les constructions acceptées par une entrée telles qu'elles existent dans les tables, le L*efff* regroupe dans une représentation unique des constructions qu'il considère comme étant des variantes l'une de l'autre. Ainsi, l'effacement d'un argument dans une construction est considéré comme une variante de cette construction. Cette représentation est adaptée pour certains formalismes d'analyseurs syntaxiques, tel que TAG utilisé dans l'analyseur FRMG (Thomasset et de La Clergerie, 2005) et LFG dans l'analyseur SXLFG (Boullier et Sagot, 2005).

Il y a donc deux étapes bien distinctes, puisqu'il ne semble pas souhaitable que le format *LGLex* manipule des concepts issus du L*efff* ou de DICOVALENCE : le format *LGLex* doit être utilisable par tous les connaisseurs des tables du Lexique-Grammaire, y compris ceux qui refusent totalement la notion de fonction syntaxique (du L*efff*) ou de paradigme (de DICOVALENCE). La construction du lexique *LGLex* n'a rien à voir avec ses utilisations, comme par exemple sa conversion en un autre format.

LGExtract se concentre sur l'explicitation de certaines colonnes, sans utiliser d'autres concepts que ceux des tables du Lexique-Grammaire. On obtient ainsi dans *LGLex*, une représentation plus explicite des tables, ce qui sert (entre autres) à produire plus simplement une représentation au format L*efff*.

5.3. *Comparaison avec deux autres méthodes*

Hathout et Namer (1998) proposent également une méthode d'extraction à partir des tables du Lexique-Grammaire. Celle-ci permet de produire trois formats de lexique : un lexique intermédiaire, un lexique TAG et un lexique HPSG. Ce premier lexique est généré à partir d'un tiers des tables du Lexique-Grammaire, soit 18 tables, décrivant 2 589 verbes et 3 485 entrées, alors que les deux autres lexiques ne tiennent compte que de 4 tables seulement (4, 36DT, 38L, 1).

Gardent *et al.* (2006) proposent une autre méthode d'extraction, qui permet de produire le lexique SynLex. Comme seules 60 % des tables du Lexique-Grammaire étaient disponibles lors de sa création, ils ont complété manuellement leur lexique avec environ 2 000 verbes et leurs cadres de base, avec notamment la construction de base du Lexique-Grammaire. Le lexique contient 5 244 verbes et 19 127 entrées (paires verbe - cadre).

À titre de comparaison, le lexique *LGLex* contient autant d'entrées que dans le Lexique-Grammaire actuel, c'est-à-dire 13 872 entrées verbales, dont 5 738 verbes distincts. Le lexique *LGLex* converti au format L*efff* contient 5 736 verbes et 22 128 entrées. De plus, les noms prédicatifs, les expressions figées et les adverbes extraits des tables font également partie du lexique *LGLex*. Une comparaison détaillée entre *LGLex* et SynLex ou un des lexiques de Hathout et Namer (1998) n'est pas possible puisque ces derniers ne sont pas disponibles.

La différence principale entre l'approche de Gardent *et al.* (2006) et celle de Hathout et Namer (1998) réside dans la façon de coder l'information structurelle. Gardent *et al.* (2006) codent cette information dans des graphes (construits manuellement afin de rendre la structure d'une table explicite et traduire les intitulés de colonne en structures de traits) à partir desquels le lexique est calculé. Dans Hathout et Namer (1998), l'information est insérée directement dans les tables (les constructions de base sont ajoutées dans chacune des tables avec le codage + pour toutes les entrées) et les intitulés de colonne sont traduits automatiquement dans des spécifications de trait.

Une autre différence concerne le traitement des intitulés de colonne. Alors que dans Hathout et Namer (1998) ces intitulés de colonne sont traduits semi-automatiquement, Gardent *et al.* (2006) utilisent une traduction manuelle : le traitement de chaque propriété est défini pour chaque table dans le graphe associé. Cependant, les tables du Lexique-Grammaire sont continuellement mises à jour pour être améliorées (par exemple, l'ajout ou le renommage de propriétés), ce qui rend cette approche pénible à maintenir. Par exemple, si une même propriété est ajoutée dans plusieurs tables, tous les graphes associés à ces tables doivent être modifiés.

En revanche, dans notre approche, l'information est insérée dans la table des classes et y est beaucoup plus complète puisqu'elle ne se contente pas d'y ajouter la construction de base mais la formule définitoire de chaque table (voir section 6). De plus, la table des classes permet de coder l'ensemble des propriétés existantes pour chaque table, ce qui permet au fur et à mesure de compléter certaines informations, comme vu en 4.2[16]. Ensuite, le traitement des intitulés de colonne utilise une traduction manuelle mais généralisée, c'est-à-dire que chaque propriété n'est définie qu'une seule fois et pas pour chaque table, étant donné que leur signification est la même après leur harmonisation. Au lieu d'interpréter ou de corriger après coup les propriétés, nous les avons entièrement revues pour qu'elles soient cohérentes dans l'ensemble des tables, ce qui a élevé la qualité des tables.

Une comparaison plus détaillée entre notre approche et celles de Gardent *et al.* (2006) ou de Hathout et Namer (1998) n'est pas possible puisque les programmes de ces deux approches ne sont pas disponibles. En particulier, ni les constructions de base insérées par Hathout et Namer (1998), ni les graphes de Gardent *et al.* (2006) ne sont disponibles. De plus, la maintenance de ces programmes n'ayant pas été assurée et la version actuelle des tables ayant été améliorée, ils ne sont plus utilisables tels quels.

5.4. *Suppression des cartouches horizontaux*

Notons enfin que dans la version 1 des tables, des *cartouches* horizontaux matérialisent une classification des propriétés, autrement dit, certaines propriétés sont regroupées en familles, voire mises en dépendance les unes par rapport aux autres (Tolone, 2009). Ce problème était donc au centre des travaux de Gardent *et al.* (2006)[17], ce qui n'est plus le cas actuellement.

En effet, cette structuration n'est pas exploitable informatiquement de façon simple, car les intitulés de colonne deviennent des objets complexes constitués de plusieurs étiquettes. De plus, même si le format Excel permet de garder les cartouches horizontaux grâce à des fusions de colonnes, ceci n'est pas conforme avec la contrainte d'avoir tous les intitulés sur la première ligne permettant d'utiliser les tables avec le

16. Tout ce qui n'a pas encore été codé contient le codage ?.

17. Les graphes étaient construits manuellement pour pouvoir notamment interpréter certains intitulés de colonne en fonction du contexte puisqu'ils n'étaient pas traduits de la même façon dans toutes les tables où ils apparaissaient.

logiciel Unitex (Paumier, 2003). Cette contrainte est d'autant plus valable aujourd'hui avec l'utilisation de l'outil *LGExtract*. Les cartouches horizontaux de la version 1 ont donc été supprimés, bien qu'ils aient contribué à la lisibilité tout en apportant des informations[18]. Lors de la suppression de ces cartouches et de leurs dépendances[19], les informations qu'ils contenaient ont été incorporées aux intitulés. Ceci a parfois introduit des informations implicites, c'est pourquoi les intitulés ont été complétés lors de leur harmonisation (Tolone, 2009).

Un exemple est la table 36DT, qui contenait l'intitulé Ppv =: lui dépendant de la colonne N2 =: N-hum. Il a été renommé Prép N2-hum = Ppv =: lui, la construction de base étant N0 V N1 Prép N2.

Ces conventions semblent avoir compliqué la compréhension des propriétés par certains utilisateurs. Ainsi, dans Gardent *et al.* (2005), on se demande si les indices (d'un argument dans une construction ou dans une propriété distributionnelle) font référence à la position du constituant dans la construction de base ou dans une autre.

Reprenons le cas de la table 32PL pour expliquer ce problème : cette table a pour construction de base N0 V N1, mais dès la deuxième colonne codée dans la table, un argument numéroté N2 apparaît, par exemple dans l'intitulé N2 V N1 ou N0 V N2. On ne peut pas relier ces redistributions à la construction de base, mais on peut les relier à une autre construction codée dans la table quelques colonnes après, intitulée N0 V N1 en N2 (cf. 5.1). Il faut donc considérer la table dans son ensemble pour donner un sens à chaque constituant. Tous les intitulés étant à présent homogènes, les numéros d'ordre des constituants figurant dans les propriétés soit font référence à ceux figurant dans la construction de base, soit sont des arguments supplémentaires. Dans ce cas, ils font référence à une autre construction les contenant qui est plus longue que la construction de base (cf. 5.1).

6. Formules définitoires

Si l'on reprend à nouveau la table 9, la possibilité d'avoir une complétive en N1 est codée dans la table par les trois colonnes N1 =: Qu Pind, N1 =: Qu Psubj et N1 =: si P ou si P. La disjonction de ces trois propriétés fait donc partie de la définition de la table, ce qui n'est pas visible dans la table des classes (codage o) puisque seule une de ces propriétés est obligatoire et est codée dans la table elle-même[20]. Cela signifie qu'il faut prendre en compte cet ensemble de propriétés séparées par des

18. Ce travail a été réalisé par Éric Laporte en 2003-2004 (Laporte, 2010). Nous avons reproduit le même travail en 2009 pour les nouvelles tables de noms numérisées, ainsi que les nouvelles tables d'expressions figées numérisées (Tolone, 2009).

19. Les dépendances ne pouvant figurer dans le format Excel.

20. On aurait pu par exemple créer le nouvel intitulé de colonne suivant : N1 =: Qu Pind+Qu Psubj+si P ou si P, et indiquer que cette propriété est définitoire. Mais cela est en contradiction avec la réutilisabilité des colonnes dans les différentes tables et la volonté de garder des intitulés succincts (Tolone, 2011).

ou logiques, c'est-à-dire dont au moins une des propriétés parmi l'ensemble est vraie. Cela correspond pour la table 9, si l'on écrit formellement cet ensemble, à la formule booléenne suivante : (N1 =: Qu Pind) ou (N1 =: Qu Psubj) ou (N1 =: si P ou si P)[21].

De même, si l'on définit la possibilité d'avoir une complétive en N2 par les deux propriétés N2 =: Qu Pind et N2 =: Qu Psubj, l'exclusion de cette possibilité est représentée par la négation de chacune des deux propriétés : non (N2 =: Qu Pind), ainsi que non (N2 =: Qu Psubj). La conjonction de ces deux négations de propriétés fait également partie de la définition de la table (chaque propriété est codée $-$ dans la table des classes) et peut être définie comme suit : non (N2 =: Qu Pind) et non (N2 =: Qu Psubj).

Nous pouvons ainsi définir formellement la délimitation d'une table par une *formule définitoire*, composée d'un ensemble de disjonctions, conjonctions et négations de propriétés, autrement dit, un ensemble de propriétés séparées par des ou logiques, des et logiques et des négations non. Elle y inclut aussi bien les propriétés définitoires vraies que fausses (introduites par la négation non) pour l'ensemble d'une table et codées dans la table des classes (codage $+$ ou $-$), ainsi que les disjonctions de certaines propriétés codées dans la table elle-même (ensemble de propriétés séparées par des ou logiques).

Les formules définitoires, de même que les propriétés définitoires, ne définissent qu'un ensemble de constructions possibles pour les entrées figurant dans la table concernée, mais ne représentent en aucun cas l'ensemble des constructions d'une entrée. En effet, toutes les propriétés codées dans la table ne figurent pas dans les formules définitoires, mais uniquement celles qui participe à la définition d'une condition obligatoire justifiant la présence d'une entrée dans une table.

À titre d'exemple, reprenons l'ensemble des propriétés que l'on vient de citer pour la table 9 :

– sa construction de base est N0 V N1 à N2 ;

– l'exclusion de la possibilité d'avoir une complétive en N2 s'écrit :
non (N2 =: Qu Pind) et non (N2 =: Qu Psubj) ;

– la possibilité d'avoir une complétive en N1 s'écrit :
(N1 =: Qu Pind) ou (N1 =: Qu Psubj) ou (N1 =: si P ou si P).

On peut écrire formellement la définition de la table 9 (incomplète ici) par la conjonction de ces trois ensembles, ce qui donne la formule définitoire suivante :
(N0 V N1 à N2)
et non (N2 =: Qu Pind) et non (N2 =: Qu Psubj)
et ((N1 =: Qu Pind) ou (N1 =: Qu Psubj) ou (N1 =: si P ou si P)).

Ce sont ces formules définitoires qui permettent de délimiter les différentes classes et d'avoir donc des classes disjointes. Elles ont été formalisées pour toutes les tables de

21. Il faut noter que la liste des réalisations du N1 n'est pas exhaustive dans la définition de la table. Par exemple, les propriétés N1 =: Nhum et N1 =: N-hum permettent de coder un groupe nominal.

verbes distributionnels[22] et sont incluses dans la version 3.4, ainsi que dans l'annexe F de (Tolone, 2011).

Afin de savoir dans quelle table est incluse (ou doit être ajoutée) une entrée, nous avons factorisé les formules définitoires en créant un arbre de classement des verbes distributionnels. Ses conventions de lecture sont expliquées dans la section suivante. Sans cet arbre, pour chaque nouvelle entrée, il faudrait vérifier une par une les formules définitoires de chaque table pour savoir si oui ou non l'entrée accepte la formule en question. Ceci impliquerait de vérifier plusieurs fois l'acceptation d'une même propriété, si elles ne sont pas factorisées. L'arbre de classement a pour but de hiérarchiser les propriétés faisant partie des formules définitoires des tables et ainsi de visualiser les propriétés qui différencient les tables. Sa présentation sous forme d'arbre est plus lisible et plus concise que l'ensemble des formules définitoires, et permet d'être facilement manipulable par des linguistes.

7. Arbre de classement

L'arbre de classement, dont un extrait est montré à la figure 1, figure dans son ensemble[23] dans l'annexe B et est inclus dans la version 3.4, ainsi que dans l'annexe G de (Tolone, 2011)[24]. Il sert à déterminer à quelle classe appartient une entrée verbale donnée. Il est conçu pour un utilisateur qui connaît les propriétés de l'entrée et qui applique successivement les critères indiqués dans l'arbre.

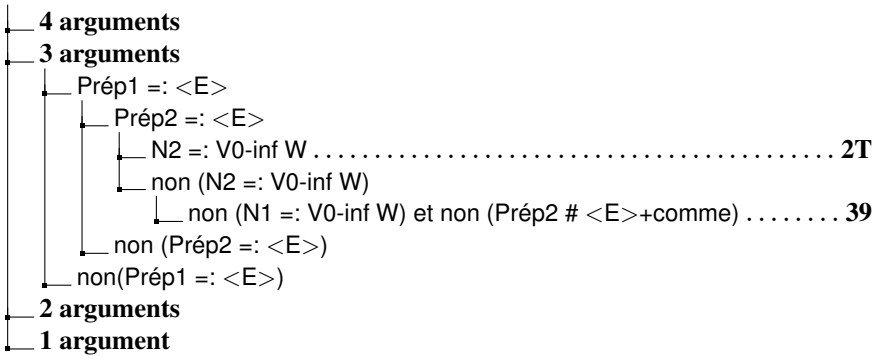

Figure 1. *Extrait de l'arbre de classement des verbes*

22. Nous avons réalisé ce travail en collaboration avec Éric Laporte et Christian Leclère en 2010-2011.

23. Les figures 2, 3, 5 et 8 représentent les 4 sous-arbres principaux.

24. Toute modification future des formules définitoires et donc de l'arbre de classement entraîne des changements de table de certaines entrées et *vice versa*. L'arbre de la version 3.4 ne fait que refléter la classification actuelle, après harmonisation et explicitation des propriétés.

Par exemple, supposons que l'on veuille classer le verbe *élire*, illustré par la phrase *On a élu Guy président*. La construction correspondant à cet exemple est N0 V N1 N2, avec 3 arguments (N0 V Prép N1 Prép N2), dont les prépositions introduisant l'argument N1 et N2 sont absentes (Prép1 =: <E> et Prép2 =: <E>). Le premier critère se situe au niveau du nombre des arguments, qui peuvent aller de 1 à 4. Puis, d'après notre extrait, on peut accepter ou rejeter l'absence de prépositions. Ensuite, la propriété qui distingue la table 2T (Gross, 1975) (p. 165)[25] et la table 39 (Boons *et al.*, 1976a) (p. 62), est N2 =: V0-inf W. Notre verbe n'acceptant pas cette propriété, et de plus, vérifiant la formule booléenne non (N1 =: V0-inf W) et non (Prép2 # <E>+comme), doit figurer dans la table 39, ce qui est effectivement le cas.

Nous détaillons à présent la procédure à suivre pour trouver la classe associée à une nouvelle entrée à l'aide de l'arbre. Nous expliquons comment déterminer la construction de base, tout en précisant les constructions qui sont prises en compte et comment gérer la numérotation des arguments. Puis, nous présentons quels sont les critères à appliquer, en précisant leur notation et comment choisir la formulation dans l'arbre lorsqu'il y a des variantes. Enfin, le résultat du classement est illustré en reprenant l'exemple précédent.

7.1. *Détermination de la construction de base*

Tout d'abord, nous devons savoir quelles sont les différentes constructions qui sont prises en compte. Presque tous les critères concernent la construction qui comporte le sujet et tous les compléments essentiels (Boons *et al.*, 1976b), c'est-à-dire la construction de base (cf. 3.2). Par exemple, pour *transvaser*, la construction de base est illustrée par la phrase *Luc transvase le vin de la cuve dans le tonneau*, qui a un complément direct et deux compléments prépositionnels. Quelques critères concernent des variantes de la construction de base (*On élit Luc comme président / On élit Luc président*), ou encore des constructions totalement différentes dans laquelle les arguments occupent d'autres positions (*Le vin est transvasé par Luc de la cuve dans le tonneau*).

Avant d'utiliser l'arbre de classement sur une entrée, l'utilisateur doit déterminer sa construction de base, car c'est celle-ci qui servira de référence pour l'application de la plupart des critères, comme par exemple le premier concernant le nombre d'arguments de la construction de base. La construction de base est choisie parmi les différentes constructions qui relèvent de l'entrée, et dans lesquelles le verbe conserve le même sens. Ce choix peut être délicat et même en partie arbitraire, mais il repose sur des priorités :

– priorité à la construction qui comporte le plus de compléments essentiels (*Luc conjugue le verbe au futur* par rapport à *Luc conjugue le verbe*) ;

– priorité à l'actif par rapport au passif, même lorsqu'il est moins employé (*Le paysage éberlue Max* par rapport à *Max est éberlué par le paysage*) ;

25. L'éclatement de la table 2 en deux tables (2 et 2T) est détaillé dans Tolone (2009).

– priorité à la construction qui comporte une préposition par rapport à celle sans préposition (*Les soldats patrouillent dans la ville* par rapport à *Les soldats patrouillent la ville*) ;

– priorité à la complétive sur l'infinitive (*Luc craint qu'il ne pleuve* par rapport à *Luc craint d'être mouillé*), et donc complément direct ;

– lorsqu'une construction contient deux compléments essentiels dont l'un dénote un lieu, et situe par rapport à ce lieu une entité dénotée par l'autre complément, priorité à la construction dans laquelle le complément de lieu est prépositionnel (*Luc peint un portrait sur la cloison* par rapport à *Luc peint la cloison d'un portrait*) ;

– lorsqu'il existe un complément direct et un complément avec une des prépositions *en* ou *entre*, la construction dans laquelle apparaît ce dernier a priorité par rapport aux éventuelles autres constructions qui comportent un autre complément prépositionnel (*Luc ventile le courrier en quatre tas* par rapport à *Luc ventile le courrier dans les services*).

Presque tous les critères font référence à un des arguments syntaxiques de la construction de base à travers une numérotation. Les arguments syntaxiques sont le sujet et les compléments essentiels. Nous devons donc savoir comment gérer la numérotation des arguments. Ils sont supposés être numérotés à partir de 0 et conformément à l'ordre (ou à un des ordres possibles) des compléments dans la construction. Le sujet porte donc le numéro 0. De plus, on place les compléments directs avant les compléments indirects lorsque cet ordre est acceptable (*Luc formule ses réflexions à Marie* plutôt que *Luc formule à Marie ses réflexions*). De même, on place un complément de lieu source avant un complément de lieu de destination lorsque l'ordre est acceptable (*Max expédie les colis de Gap à Dax* plutôt que *Max expédie les colis à Gap de Dax*). Ces principes laissent parfois le choix entre plusieurs numérotations. Dans ce cas, l'arbre de classement tente de prévoir toutes les numérotations recevables, quitte à inverser deux numéros et recommencer. Si l'on reprend l'exemple précédent de la table 39, où l'on est passé par le nœud non (N2 = : V0-inf W), on doit également vérifier la formule booléenne non (N1 =: V0-inf W) et non (Prép2 # <E>+comme). Supposons que l'on souhaite accepter la propriété N1 = : V0-inf W, cela signifie que l'on doit inverser la numérotation et ne pas passer par le nœud non (N2 = : V0-inf W) mais N2 = : V0-inf W, ce qui nous amène à la table 2T (*Max gravit la colline prévenir ses amis*).

Les prépositions éventuelles introduisant des compléments essentiels sont indicées par le numéro correspondant. Les groupes nominaux ou propositions constituant les arguments syntaxiques sont symbolisés par N indicé par le numéro. Ainsi, dans *Luc formule ses réflexions à Marie*, le symbole N1 représente *ses réflexions*, Prép2 symbolise *à* et N2 symbolise *Marie*.

7.2. *Notation des critères*

Les critères utilisés dans l'arbre correspondent à des propriétés syntaxiques et sémantiques de l'entrée à classer. La plupart prennent la forme d'un des intitulés décrits dans la documentation des propriétés (cf. 2.2), par exemple N1 =: Qu Pind qui indique la possibilité d'une complétive objet à l'indicatif en position N1. D'autres sont des formules booléennes sur de tels intitulés (cf. section 6) : par exemple, (N1 =: Qu Pind) ou (N1 =: Qu Psubj) indique la possibilité d'une complétive objet à l'indicatif ou au subjonctif. Les quelques intitulés qui ne sont pas explicitement documentés utilisent les mêmes notations que les autres. Il faut appliquer successivement entre 2 et 14 critères pour déterminer à quelle classe appartient une entrée. Chaque étape propose un choix entre plusieurs critères exclusifs les uns des autres. Lorsqu'il y en a deux, ils sont souvent la négation logique l'un de l'autre. Par exemple, le critère associé à (N1 =: Qu Pind) ou (N1 =: Qu Psubj) est sa négation non (N1 =: Qu Pind) et non (N1 =: Qu Psubj).

Beaucoup de critères sont des critères distributionnels dont nous devons connaître la signification : ils contiennent le symbole =: et indiquent une valeur que peut prendre un des éléments de la construction de base (ou parfois plusieurs valeurs). Ainsi, N0 =: N-hum indique que le sujet N0 peut prendre comme valeur un groupe nominal N-hum dénotant une entité non humaine. Un tel critère n'indique pas une valeur exclusive : si le sujet peut aussi prendre d'autres valeurs, cela n'empêche pas que le critère donne un résultat positif. La seule exception à cette convention est le symbole Npl obl qui désigne un groupe nominal obligatoirement pluriel ou à sens collectif. Le symbole <*E*> représente l'absence de forme explicite, par exemple l'absence de préposition. Le symbole # se lit « différent de » et précède une ou plusieurs valeurs que l'élément ne doit pas prendre. Ainsi, Prép2 # à indique que la préposition doit avoir une valeur autre que *à*. Le critère Prép2 # <E> indique que la préposition doit avoir une forme explicite. Le critère non (Prép2 # <E>) indique qu'aucune préposition explicite ne peut apparaître en cette position.

Tous les critères sémantiques (ou rôles sémantiques d'arguments) sont exclusifs les uns des autres par nature : N1 bénéficiaire indique que l'objet N1 est interprété comme bénéficiaire du référent N2 (*Max goinfre Bob de gâteaux*), alors que N1 détrimentaire indique que l'objet N1 est interprété comme perdant le référent N2 (*Max a possédé Luc de 100 euros*). De même pour lieu source, lieu de destination, lieu statique et lieu de passage ; apparition et disparition, etc. De plus, on donne la priorité à un complément de lieu de destination par rapport à un complément bénéficiaire (*Max expédie la lettre à Luc* par rapport à *Max expédie la lettre à Gap*). De même, on donne la priorité à un complément de lieu source par rapport à un complément détrimentaire.

Un même critère peut généralement être formulé de plusieurs façons, (cf. 3.2) nous devons savoir quelle variante de formulation choisir : par exemple, dans le contexte des verbes transitifs directs à deux arguments, la possibilité d'une complétive objet à l'indicatif peut être notée N1 =: Qu Pind ou N0 V Qu Pind. La formulation choisie dans l'arbre de classement n'est pas toujours la même que celle choisie dans les intitulés

des propriétés décrites dans les tables. Le critère de notation principal reste le même que celui utilisé pour définir les propriétés définitoires, c'est-à-dire que nous notons les arguments N0, N1, etc. sans indication sur leur distribution et nous ajoutons ensuite des propriétés ditributionnelles, telles que N1 =: Qu P, afin que chacun des arguments soit numéroté et que l'on puisse s'y référer dans les autres propriétés. Principalement, les notations sont les mêmes que celles des propriétés définitoires codées dans la table des classes. Nous pouvons par exemple citer le fait d'intégrer la préposition dans la construction de base lorsqu'une seule est possible, comme vu en 4.2. En revanche, la notation marquant l'absence de préposition est utilisée alors qu'elle n'existe pas dans la table des classes. Par exemple, pour la table 9, sa construction de base est codée N0 V N1 N2 dans la table des classes, alors qu'elle figure sous cette notation dans l'arbre : N0 V Prép N1 Prép N2, avec Prép1 =: <E> et Prép2 =: <E>.

7.3. *Résultat du classement*

À l'issue de l'application des critères, l'arbre indique la classe[26] à laquelle appartient l'entrée. Dans certains cas, il indique en outre une ou plusieurs propriétés additionnelles que doit posséder l'entrée et qui n'ont pas été vérifiées dans les critères qui ont mené à cette branche. Par exemple, la branche menant à la classe 39 passe uniquement par 4 nœuds de l'arbre, comme vu précédemment :
3 arguments → Prép1 =: <E> → Prép2 =: <E> → non (N2 =: V0-inf W).
Mais elle indique comme propriétés additionnelles :
non (N1 =: V0-inf W) et non (Prép2 # <E>+comme).
En d'autres termes, les auteurs du Lexique-Grammaire ont constaté lors de leur travail que toutes les entrées qui vérifient à la fois les 4 propriétés ci-dessus vérifient également ces propriétés additionnelles.

8. Conclusion

Ce travail a permis de formaliser la classification existante pour les verbes distributionnels à l'aide de formules logiques. Ces formules contiennent aussi bien des propriétés définitoires vraies que fausses pour toutes les entrées d'une table, ainsi que des formules booléennes contenant plusieurs propriétés dont au moins une est vraie pour chaque entrée d'une table. Ce sont ces formules logiques qui ont permis de confirmer que l'ensemble des classes étaient disjointes, en justifiant ainsi leur existance. Elles ont permis également de représenter la classification sous forme d'arbre de décision, afin d'aider à trouver la classe associée à toute nouvelle entrée. Ceci permet d'assurer la maintenance du lexique et d'en garantir sa future alimentation, avec des critères précis à appliquer.

Après avoir proposé une définition formelle de la classification des verbes distributionnels, nous envisageons de réaliser ce travail pour l'ensemble des tables des

26. La notation **_part** indique qu'il existe plusieurs chemins menant à cette même classe.

autres catégories, à savoir les noms prédicatifs, les expressions figées et les adverbes. Cela peut également être envisagé pour les tables d'autres langues, telles que le grec (Ioannidou et Tolone, 2011). Ce travail montre que la classification des tables du Lexique-Grammaire n'est pas triviale mais repose cependant sur des critères formels.

Remerciements

Je tiens à remercier Éric Laporte et Christian Leclère pour leur constante amélioration des tables et l'aide qu'ils m'ont apportée pour la réalisation de ce travail.

9. Bibliographie

Boons J.-P., Guillet A., Leclère C., La structure des phrases simples en français : Classes de constructions transitives, Technical report, LADL, CNRS, Paris 7, 1976a.

Boons J.-P., Guillet A., Leclère C., *La structure des phrases simples en français : Constructions intransitives*, Droz, Genève, Suisse, 1976b.

Boullier P., Sagot B., « Analyse syntaxique profonde à grande échelle : SXLFG », *Traitement Automatique des Langues (T.A.L.)*, vol. 46, n° 2, p. 65-89, 2005.

Constant M., Tolone E., « A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables », *in* M. D. Gioia (ed.), *Actes du 27^{e} Colloque international sur le lexique et la grammaire (L'Aquila, 10-13 septembre 2008), Seconde partie*, vol. 1 of *Lingue d'Europa e del Mediterraneo, Grammatica comparata*, Aracne, Rome, Italie, p. 79-193, 2010. ISBN 978-88-548-3166-7.

Gardent C., Guillaume B., Perrier G., Falk I., « Extracting subcategorisation information from Maurice Gross' Grammar Lexicon », *Archives of Control Sciences*, vol. 15, n° 3, p. 253-264, 2005. In Memoriam Maurice Gross. Special issue on Human Language Technologies as a challenge for Computer Science and Linguistics. Part I. (2nd Language and Technology Conference).

Gardent C., Guillaume B., Perrier G., Falk I., « Extraction d'information de sous-catégorisation à partir des tables du LADL », *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN'06)*, Louvain, Belgique, 2006.

Giry-Schneider J., Balibar-Mrabti A., Classes de noms construits avec avoir, Technical report, LADL, Université Paris 7, 1993.

Gross M., *Méthodes en syntaxe : Régimes des constructions complétives*, Hermann, Paris, France, 1975.

Gross M., *Grammaire transformationnelle du français : Syntaxe du verbe*, vol. 1, Cantilène, Paris, France, 1986.

Guillet A., Leclère C., *La structure des phrases simples en français : Les constructions transitives locatives*, Droz, Genève, Suisse, 1992.

Hathout N., Namer F., « Automatic Construction and Validation of French Large Lexical Resources : Reuse of Verb Theoretical Linguistic Descriptions », *Proceedings of the 1st Language Resources and Evaluation Conference (LREC'98)*, Grenade, Espagne, 1998.

Ioannidou K., Tolone E., « Construction du lexique *LGLex* à partir des tables du Lexique-Grammaire des verbes du grec moderne », *Actes du 30e Colloque Lexique et Grammaire (LGC'11)*, Nicosie, Chypre, 2011.

Laporte E., « Le Lexique-Grammaire est-il exploitable pour le traitement des langues ? », *Cahiers du CENTAL*, vol. 6, p. 207-218, 2010. Mélanges en hommage à Christian Leclère.

Meunier A., Nominalisations d'adjectifs par verbes supports, PhD thesis, LADL, Université Paris 7, France, 1981.

Paumier S., « Some remarks on the application of a Lexicon-Grammar », *Lingvisticæ Investigationes*, vol. 24, n° 2, p. 245-256, 2001.

Paumier S., De la reconnaissance de formes linguistiques à l'analyse syntaxique, PhD thesis, Université Paris-Est Marne-la-Vallée, France, 2003.

Thomasset F., de La Clergerie E., « Comment obtenir plus des Méta-Grammaires », *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN'05)*, Dourdan, France, 2005.

Tolone E., « Les tables du Lexique-Grammaire au format TAL », *Actes de MajecSTIC 2009*, Avignon, France, 2009. (8 p.).

Tolone E., Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français, PhD thesis, LIGM, Université Paris-Est, France, 2011. (340 p.).

Tolone E., Sagot B., « Using Lexicon-Grammar tables for French verbs in a large-coverage parser », *in* Z. Vetulani (ed.), *Human Language Technology. Challenges for Computer Science and Linguistics. 4th Language and Technology Conference, LTC 2009, Poznań, Poland, November 6-8, 2009, Revised Selected Papers*, vol. 6562 of *Lecture Notes in Artificial Intelligence (LNAI)*, Springer Verlag, p. 183-191, 2011.

Tolone E., Sagot B., de La Clergerie E., « Evaluating and improving syntactic lexica by plugging them within a parser », *Proceedings of the 8th Language Resources and Evaluation Conference (LREC'12)*, Istanbul, Turquie, 2012.

Tolone E., Voyatzi S., Leclère C., « Constructions définitoires des tables du Lexique-Grammaire », *in* L. Popović, C. Krstev, D. Vitas, G. Pavlović-Lažetić, I. Obradović (eds), *Actes du 29e Colloque Lexique et Grammaire (LGC'10)*, Belgrade, Serbie, p. 321-331, 2010.

van den Eynde K., Mertens P., *Le dictionnaire de valence* DICOVALENCE *: manuel d'utilisation*. 2006, http ://bach.arts.kuleuven.be/dicovalence/manuel_061117.pdf.

Vivès R., Avoir, prendre, perdre : constructions à verbe support et extensions aspectuelles, PhD thesis, LADL, Université Paris 7, France, 1983.

A. Annexe - Notations du Lexique-Grammaire

De manière générale, les notations utilisées sont celles de Gross (1986).

Les constructions syntaxiques sont représentées par des suites de symboles telles que N0 V N1 Prép N2, dénotant une suite « sujet - verbe - objet direct - complément prépositionnel », comme par exemple dans la phrase *Paul débat cette question avec Luc*.

N désigne un argument syntaxique, c'est-à-dire le sujet ou un complément essentiel.

Les chiffres à droite des N indiquent leur placement de gauche à droite dans la construction de base :

– N0 : sujet ;

– N1 : premier complément ;

– N2 : deuxième complément, etc.

La notation Ni est utilisée pour désigner le sujet à l'intérieur d'une complétive, comme dans Qu Psubj =: Qu Ni Vsubj W = (Ni) (de Vi-inf W), qui indique qu'une complétive au subjonctif introduite par *que* et de la forme Ni Vsubj W peut être remplacée par un constituant Ni suivi d'une infinitive Vi-inf W, introduite par la préposition *de* (*Paul empêche que Pierre vienne = Paul empêche Pierre de venir*).
Les chiffres à droite des autres symboles, tels que Adj, Det, Prép, Loc, C, etc. indiquent ce même placement. Par exemple, Prép1 désigne la préposition du premier complément, même si la préposition peut ne pas être numérotée dans les constructions, comme c'est le cas pour les verbes (par exemple, dans N0 V Prép N1).

De plus, cette numérotation peut être utilisée pour faire référence à un argument syntaxique de la phrase. Par exemple, le pronom *lui-même* peut être noté lui1-même, ce qui indique que le pronom *lui* est coréférent à l'objet N1 (*Le froid a recroquevillé la plante sur elle-même*).
N peut également représenter un substantif ou un groupe nominal lorsqu'un trait sémantique apparaît à droite du N, ou du chiffre (sauf Nnr qui peut désigner une complétive ou une infinitive). Dans la mesure où l'on se focalise très peu sur les déterminants, les adjectifs et les relatives, cette ambiguïté, loin de présenter des inconvénients, permet de représenter simultanément tout un groupe nominal, ainsi que le substantif tête de ce groupe nominal.
Voici quelques exemples de traits sémantiques figurant dans les constructions :

– N0hum : sujet pris dans la classe des substantifs humains, par exemple, **(L'ennemi+Luc)** *quitte la ville* ;

– N1pl obl : premier complément obligatoirement au pluriel, par exemple, *La bouteille a éclaté en* **mille morceaux** ;

– N2pc : deuxième complément pris dans la classe des substantifs parties du corps, par exemple, *Paul joint le pouce avec* **l'index**.

Ces mêmes traits sémantiques peuvent faire l'objet à eux seuls d'une propriété distributionnelle écrite sous la forme N0 =: Nhum, N1 =: Npl obl ou N2 =: Npc. Pour cette dernière, on peut indiquer la coréférence avec un substantif de la même phrase en ajoutant un chiffre à droite du trait sémantique pc. Par exemple, C1 =: Npc0 (C1 désignant le substantif tête du premier complément figé dans une expression figée) est employé pour C1pc de N0, c'est-à-dire C1pc portant obligatoirement sur N0 (par exemple, *Max a la* **tête** *ailleurs*).

Les parenthèses contenant plusieurs éléments séparés par le signe + indiquent un choix possible entre ceux-ci ; la lettre E désigne l'élément vide. Ainsi :
N0 V (E+N1) : *Jean lit (E+un livre).*

correspond aux deux structures suivantes :
N0 V : *Jean lit.*
N0 V N1 : *Jean lit un livre.*

Un signe + entourant deux chiffres sans parenthèses désigne les arguments concernés, comme par exemple dans N0 V N1 + 2, qui représente une phrase avec un objet direct interprété par métonymie comme les deux objets (*Léa a boutonné un pan de sa robe avec l'autre = Léa a boutonné sa robe*).

Les parenthèses ne contenant pas de signe + permettent de délimiter un argument, comme par exemple dans N0 V (N1 de N1c) = N0 V (N1c) (Prép N1) (*Luc stimule la curiosité de Marie = Luc stimule Marie dans sa curiosité*).

Les deux notations =: et = se différencient par le fait que la première signifie « se spécifie ou se développe en » : elle précise une distribution possible d'un ou plusieurs éléments d'une construction ; alors que le = signifie « est transformationnellement lié à » : il suppose toujours l'existence d'une nouvelle construction (représentée à droite du signe) par rapport à une déjà connue (représentée à gauche). Par exemple, à N1 = Ppv =: le signifie que l'argument à N1 peut être pronominalisé en *le* (*Paul apprend à lire = Paul l'apprend*).

Les crochets représentent une transformation, telle que [extrap] pour l'extraposition, ou [passif par] (respectivement, [passif de]) pour le passif introduit par la préposition *par* (respectivement, *de*).

La notation => désigne une implication, telle que dans impératif => subj, qui implique que la complétive soit au subjonctif lorsque la phrase exprime un ordre (*Max ordonne que Paul vienne*), ou (Nég, interro) => subj, où la négation ou l'interrogation peuvent entraîner la mise au subjonctif de la subordonnée (*Je crois qu'il viendra / Je ne crois pas qu'il vienne / Crois-tu qu'il vienne ?*).

La notation # signifie « différent de », par exemple Loc # de désigne une préposition locative différente de *de*.

La notation 'P' désigne un discours direct, comme dans N0 V à N2 : 'P', où c'est la complétive N1 qui peut prendre la forme d'un discours direct (*Luc répond à Léa qu'il va au cinéma = Paul répond à Marie : « Je vais au cinéma »*), ou 'P', V N0 à N2, où de plus, elle apparaît en tête de phrase (*« Je vais au cinéma », répond Luc à Léa*).

La notation <ENT> représente tous les mots faisant partie de l'entrée et la notation <OPT>, ceux faisant partie d'entrées associées[27].

Traditionnellement (Gross, 1986), les chiffres sont soit en indice pour numéroter les arguments syntaxiques, soit en exposant pour noter la coréférence, ce qui facilite l'interprétation des intitulés. Par exemple, la construction N0 V Loc N1 V0-inf W est notée N_0 V Loc N_1 V^0-inf W. Nous n'avons pas retenu cette distinction dans les

27. Pour les tables verbales, cette notation est employée pour désigner la colonne contenant l'exemple. De plus, <ENGLISH> représente la traduction du verbe en anglais.

intitulés de propriétés des tables, car il n'existe aucun cas où la distinction entre deux intitulés repose uniquement sur la distinction indice/exposant. De plus, l'interprétation des intitulés repose maintenant sur une documentation précise (cf. 2.2).

Les symboles utilisés sont :

– Adj : adjectif ; peut être suivi de permut obl pour indiquer que l'adjectif doit être obligatoirement permuté avec le nom ;

– Adj-ment : adverbe dérivé d'un adjectif, auquel on a ajouté *-ment* ;

– Adj-n : nom morphologiquement associé à un adjectif ;

– Adv : adverbe ; le rôle sémantique de l'adverbe peut être spécifié : Advm pour adverbe de manière, Advp pour adverbe de prix, Advl pour adverbe de lieu, Advt pour adverbe de temps, Advfut pour adverbe de temps futur, Advtd pour adverbe de temps duratif, etc. ;

– AdvPhrase : adverbe de phrase ;

– autre suivi d'un autre symbole (par exemple, Loc1) : autres valeurs lexicales possibles de ce symbole, en plus de celles représentées dans les propriétés binaires ;

– Aux : auxiliaire ;

– C : substantif figé inclus dans la structure d'un argument figé tel que par exemple, Det1 C1 Adj1 dans la construction N0 faire Det1 C1 Adj1 à N2, à contraster avec les arguments libres, ici N0 et N2 ;

– combien ? : complément précisant une quantité ou une mesure intéressant le procès, et souvent à déterminant numéral ;

– Conj : conjonction ; la nature de la conjonction peut être spécifiée : ConjC pour conjonction de coordination et ConjS pour conjonction de subordination ;

– Det : déterminant (simple ou accompagné d'un modifieur, par exemple *un certain*) ; la nature du déterminant peut être spécifiée : Det1 =: déf pour déterminant défini, Det1 = : indéf pour déterminant indéfini, Dnum pour un déterminant numéral ;

– Det N : déterminant et prédicat nominal ;

– Detc : déterminant du complément de nom Nc ;

– dé-V : verbe dérivé de V par un préfixe négatif ;

– du : article partitif (*du*+*de la*) ;

– E ou <E> : absence ou effacement d'un élément ; représente l'élément neutre de la concaténation et sert à marquer la séquence vide (préposition zéro, déterminant zéro, etc.) ;

– GN : groupe nominal ;

– le : article défini (*le*+*la*+*l'*) ;

– Loc : préposition locative, c'est-à-dire introduisant un complément de lieu (*dans*, *sur*, *à*, etc.) ;

– Modif : tout modifieur (relative, adjectif, complément de nom, épithète, etc.) d'un groupe nominal ; un déterminant suivi d'un modifieur (avec la notation Det-Modif)

représente un constituant discontinu formé par le déterminant et le modifieur obligatoire ;

– N : substantif ou groupe nominal, ou argument syntaxique (sujet ou complément essentiel), comme vu précédemment :

- traits sémantiques possibles (attachés au N) : hum (entité humaine), -hum (entité qui n'est pas une personne ni un animal linguistiquement assimilé à une personne), pc (partie du corps d'une personne), pc obl (obligatoirement une partie du corps, ou, par métonymie, une personne), plur (pluriel), pl obl (pluriel obligatoire ou collectif), abs (entité abstraite), conc (objet concret), nr (substantif dénotant une personne, un objet concret, une entité abstraite, une complétive ou une infinitive), pr (nom propre), monnaie (nom de monnaie, une somme d'argent), mes (nom d'unité de mesure), esprit (esprit d'une personne), idée, texte, mot, chemin (situation statique dans laquelle une personne ou une chose peut effectuer un trajet sur ce chemin), coup, trou, couche (couche d'une substance concrète), zone, transport (moyen de transport), instrument, point, trace (trace ou marque), déformation, mal (maladie), psy (psychologique), nc (non contraint),

- rôles thématiques possibles (séparés de N par un espace) : lieu source (lieu source du référent d'un autre argument), lieu de destination (lieu de destination du référent d'un autre argument), nv-dest (nouvelle destination), mouvement (objet ou lieu en mouvement), lieu du passage (lieu par lequel passe le référent du sujet), lieu du procès (lieu où se déroule le procès), apparition (apparaissant ou étant créé au cours du procès), disparition (disparaissant au cours du procès), bénéficiaire (bénéficiaire du référent d'un autre argument), détrimentaire (détrimentaire du référent d'un autre argument), matériau (matériau utilisé dans le procès), attache (système d'attache interprété comme un instrument), résultat, actif (personne interprétée comme active), neutre (la phrase dénote un événement datable), statique (la phrase dénote une situation statique), métaphore (la phrase a un sens métaphorique), scénique (locatif) ;

– Nc : complément de nom ;

– Neg : adverbe de négation, ou pronom clitique *ne* figé avec le verbe V dans Ppv =: Neg ;

– P : phrase ou proposition ;

– Ppv : pronom clitique ou particule préverbale (*me+m'+te+t'+se+s'+le+la+l'+les+lui+nous+vous+leur+en+y*) ; il peut être obligatoirement figé avec le verbe V si figé est mentionné, comme par exemple dans Ppv =: en figé ;

– Poss : déterminant possessif (*mon+ton+son+ma+ta+sa+mes+tes+ses+notre+votre+leur+nos+vos+leurs*) ; un chiffre peut indiquer à quel argument le déterminant possessif est coréférent, par exemple Poss0 est coréférent au sujet N0 ;

– Prép : préposition ; Prép-adv désigne sa modification sous une forme adverbiale ;

– Qu P : complétive sans distinction de contenu, introduite par le pronom *que* ; le mode de la complétive peut être spécifié : Qu Pind pour une complétive à l'indicatif et Qu Psubj pour une complétive au subjonctif ; le *ce* de la complétive peut être indiqué, mais également la locution du type *le fait que* introduisant la complétive notée

le fait Qu P ;

– Tc : temps (éventuellement de l'adverbe) faisant partie de l'infinitive ;

– thèmeN1 : exemple prototypique de nom qui peut occuper la position de l'objet N1 ;

– tout : déterminant indéfini dérivé de *tout* (*tout+tous+toute+toutes*) ;

– Tp : temps (éventuellement de l'adverbe) faisant partie de la principale ;

– trajet : complément locatif introduit par la préposition *sur* ou *le long de* et interprété comme un lieu de passage ;

– un : article indéfini (*un+une*) ;

– V : verbe, défini morphologiquement ;

– V-able, V-ateur, V-eur, ou V-eux : adjectif déverbal lié à V avec un suffixe *-able*, *-ateur*, *-eur*, ou *-eux* ;

– V-adj : adjectif déverbal lié à V ;

– V-ant : adjectif déverbal lié à V avec un suffixe *-ant* ou *-ent* (par exemple, dans N0 être V-ant : *Paul sourit = Paul est souriant*), ou verbe au participe présent dans N1 = (N) (V-ant W) (*J'ai repéré que Paul travaille = J'ai repéré Paul travaillant*) ;

– V-inf W : verbe à l'infinitif, suivi de toute suite de compléments, y compris vide ; le sujet des infinitives peut être spécifié par un chiffre, par exemple :
N0 V V0-inf W : *Jean veut manger cela*
V1-inf W V N1 : *Venir ici ennuie Marie*
V2-inf W V N1 à N2 : *Faire ceci donne du mal à Paul*
Le sujet peut également être coréférent à un complément de nom de l'objet N1 par exemple et noté N0 =: V1c-inf W (*Se présenter aux élections a germé dans la tête de Paul*), ou encore être coréférent avec le sujet de la complétive objet N1 et noté N0 =: Vi-inf W (*Être trop gros empêche Luc de passer dans le couloir*) ;

– V-n : substantif de la même famille morphologique que V, -n étant un suffixe nominalisateur ; parfois le suffixe, noté Sfx, est précisé, par exemple dans Sfx = -ment ;

– Vc : verbe faisant partie de l'infinitive ;

– Vconv : verbe support converse ;

– Vop : verbe opérateur, généralement causatif ; la phrase *Paul fait boire Marie* est analysé par application de l'opérateur *Paul fait* à la phrase *Marie boit* ;

– Vsup : verbe support ;

– Vpp : verbe au participe passé ;

– W : suite quelconque, éventuellement nulle, de compléments ; cette notation peut indiquer la conservation des autres compléments éventuels dans une construction.

Les autres symboles sont des valeurs lexicales de verbes, prépositions, pronoms, conjonctions, adverbes ou modifieurs.

B. Annexe - Arbre de classement des verbes

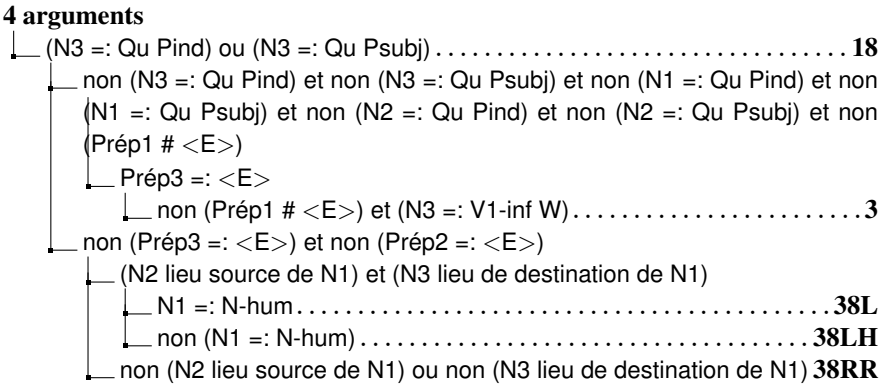

Figure 2. *Arbre de classement des verbes (1)*

3 arguments

- Prép1 =: <E>
  - Prép2 =: <E>
    - N2 =: V0-inf W **2T**
    - non (N2 =: V0-inf W)
      - non (N1 =: V0-inf W) et non (Prép2 # <E>+comme) **39**
  - non (Prép2 =: <E>)
    - (N2 =: Qu Pind) ou (N2 =: Qu Psubj) ou (N2 =: V1-inf W) ou (N2 =: V0-inf W)
      - (N1 =: Qu Pind) ou (N1 =: Qu Psubj) . . . . . **16_part**
      - non (N1 =: Qu Pind) et non (N1 =: Qu Psubj) et non (N1 =: V0-inf W)
        - (Prép2 # à) et (Prép2 # de) . . . . . **16_part**
        - non (Prép2 # à) **11**
        - non (Prép2 # de) **13**
    - non (N2 =: Qu Pind) et non (N2 =: Qu Psubj) et non (N2 =: V1-inf W) et non (N2 =: V0-inf W)
      - (N1 =: Qu Pind) ou (N1 =: Qu Psubj) ou (N1 =: si P ou si P)
        - Prép2 # à . . . . . **10_part**
        - non (Prép2 # à)
          - N2 =: Nhum **9**
          - non (N2 =: Nhum) . . . . . **10_part**
      - non (N1 =: Qu Pind) et non (N1 =: Qu Psubj) et non (N1 =: si P ou si P) et non (N1 =: V0-inf W) . . . . . **voir figure 4**
- non(Prép1 =: <E>)
  - (N2 =: Qu Pind) ou (N2 =: Qu Psubj) ou (N2 =: V1-inf W)
    - (N1 =: Qu Pind) ou (N1 =: Qu Psubj)
      - non (Prép2 =: <E>) . . . . . **16_part**
    - non (N1 =: Qu Pind) et non (N1 =: Qu Psubj)
      - (Prép1 # avec+pour) ou (Prép2 # à)
        - non (Prép2 =: <E>) . . . . . **16_part**
      - non (Prép1 # avec+pour) et non (Prép2 # à) . . . . . **14_part**
  - non (N2 =: Qu Pind) et non (N2 =: Qu Psubj) et non (N2 =: V1-inf W) et non (N2 =: si P ou si P)
    - (N1 =: Qu Pind) ou (N1 =: Qu Psubj)
      - non (Prép1 # à)
        - non (Prép2 # avec+pour) . . . . . **14_part**
      - non (Prép1 # de)
        - non (Prép2 =: <E>) **15**
    - non (N1 =: Qu Pind) et non (N1 =: Qu Psubj) et non (N1 =: V2-inf W)
      - (N2 =: V0-inf W)
        - non (Prép2 # <E>) **2**
      - non (N2 =: V0-inf W)
        - (N1 lieu source de N0) et (N2 lieu de destination de N0)
          - non (N1 =: V0-inf W) et non (N0 V N1 et N2) **35L**
        - non (N1 lieu source de N0) ou non (N2 lieu de destination de N0)
          - non (Prép2 =: <E>) **35RR**

Figure 3. *Arbre de classement des verbes (2)*

3 arguments *(suite, cf. figure 3)*

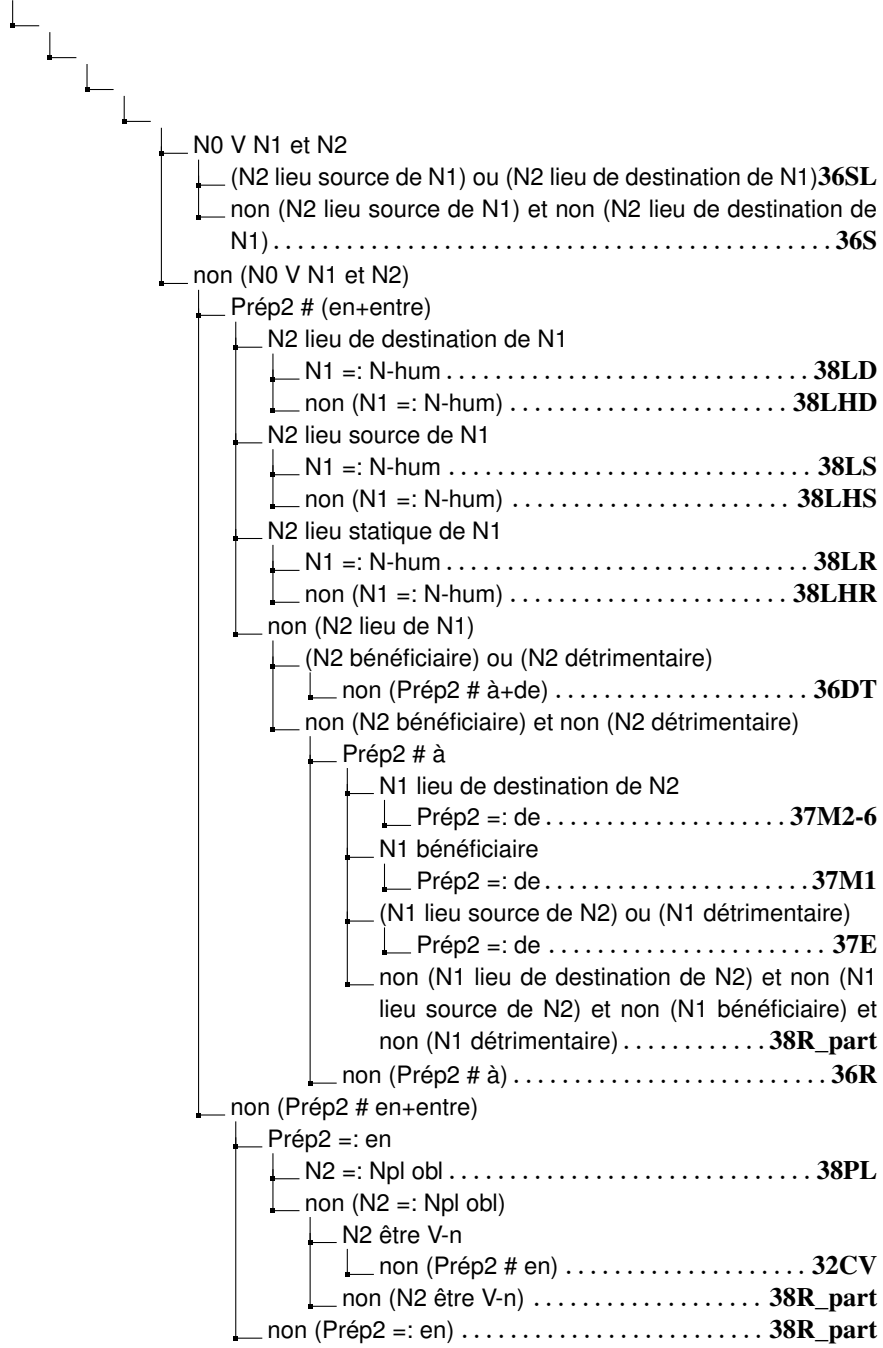

Figure 4. *Arbre de classement des verbes (3)*

2 arguments
- Prép1 # <E>
  - (N1 =: Qu Pind) ou (N1 =: Qu Psubj)
    - (Prép1 # à) et (Prép1 # de)
      - non (N1 = où) et non (N1 lieu de N0) et non (N0 et N1 V) . **35R_part**
    - non (Prép1 # à) **7**
    - non (Prép1 # de) **8**
  - non (N1 =: Qu Pind) et non (N1 =: Qu Psubj)
    - N1 =: V0-inf W
      - non (N1 = :si P ou si P) et non (N1 =: combien P) . . . . . **1_part**
    - non (N1 =: V0-inf W)
      - (N0 =: Qu P) et [extrap]
        - non (N0 rendre N1 V-adj) et non (N0 rendre N1 plus V-adj)**5_part**
      - non (N0 =: Qu P) ou non [extrap]
        - N1 = où
          - N0 et N1 V
            - Prép1 =: avec . . . . . **35S_part**
          - non (N0 et N1 V)
            - N1 lieu de destination de N0 . . . . . **35LD**
            - (N1 lieu source de N0) ou (N1 lieu statique de N0)
              - N1 V de N0 . . . . . **34L0**
              - non (N1 V de N0)
                - N1 lieu source de N0 . . . . . **35LS**
                - N1 lieu statique de N0 . . . . . **35ST**
            - non (N1 lieu de N0)
              - non (N1 V de N0) **35LR**
        - non (N1 = où) et non (N1 lieu de N0)
          - Prép1 # à+<E>
            - N0 et N1 V
              - Prép1 =: avec . . . . . **35S_part**
            - non (N0 et N1 V) . . . . . **35R_part**
          - non (Prép1 # à) **33**
- non (Prép1 # <E>)
  - (N1 =: Qu Pind) ou (N1 =: Qu Psubj) ou (N1 = :si P ou si P) ou (N1 =: combien P)
    - (N1 =: Qu Psubj) et (Qu Psubj =: Qu Ni Vsubj W = (Ni) (de Vi-inf W)) **12**
    - non (Qu Psubj =: Qu Ni Vsubj W = (Ni) (de Vi-inf W))
      - N0 =: V1-inf W
        - (N0 =: Qu P) et (N1 =: Nhum) et non (N1 =: Nconc) et non (N1 =: V0-inf W) et non [extrap] . . . . . **4_part**
      - non (N0 =: V1-inf W) **6**
  - non (N1 =: Qu Pind) et non (N1 =: Qu Psubj) et non (N1 = :si P ou si P) et non (N1 =: combien P) . . . . . **voir figure 6**

Figure 5. *Arbre de classement des verbes (4)*

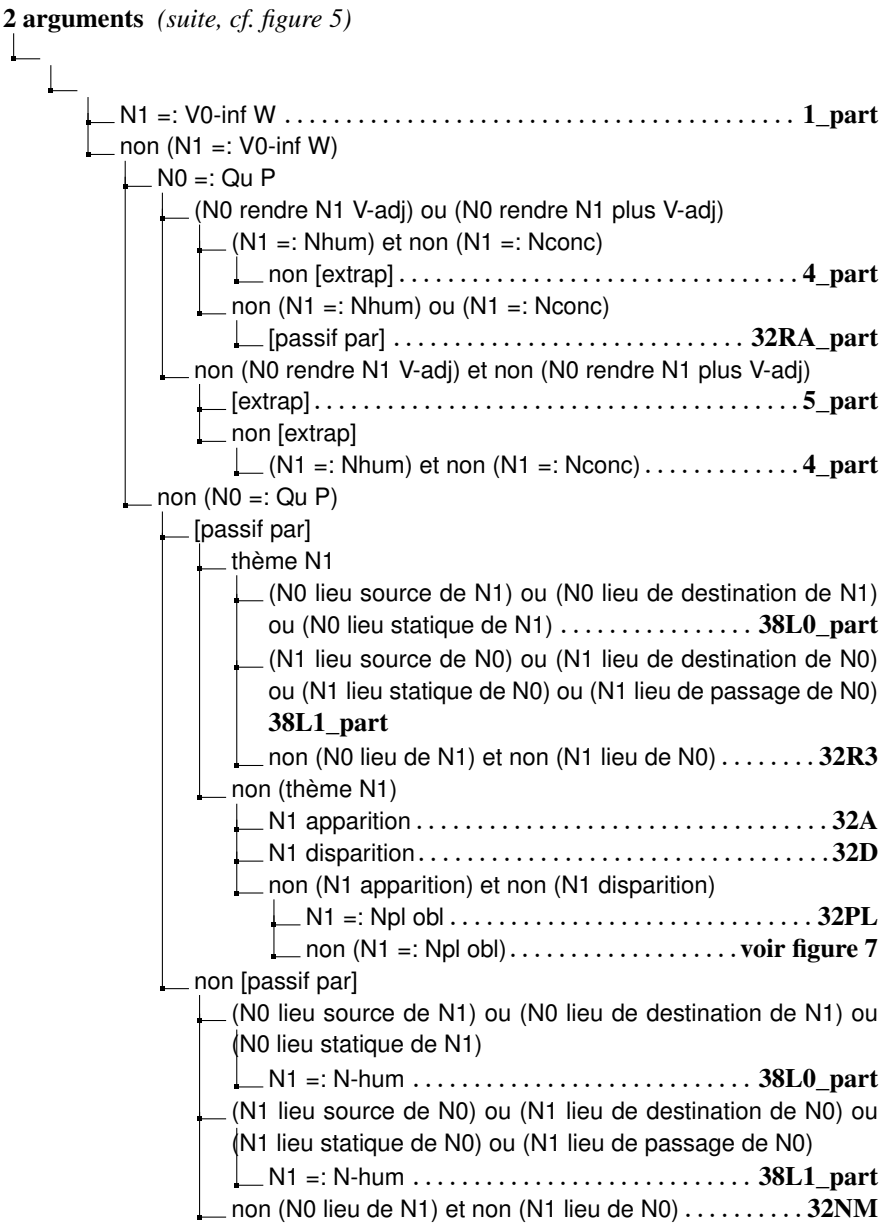

Figure 6. *Arbre de classement des verbes (5)*

2 arguments *(suite, cf. figure 6)*

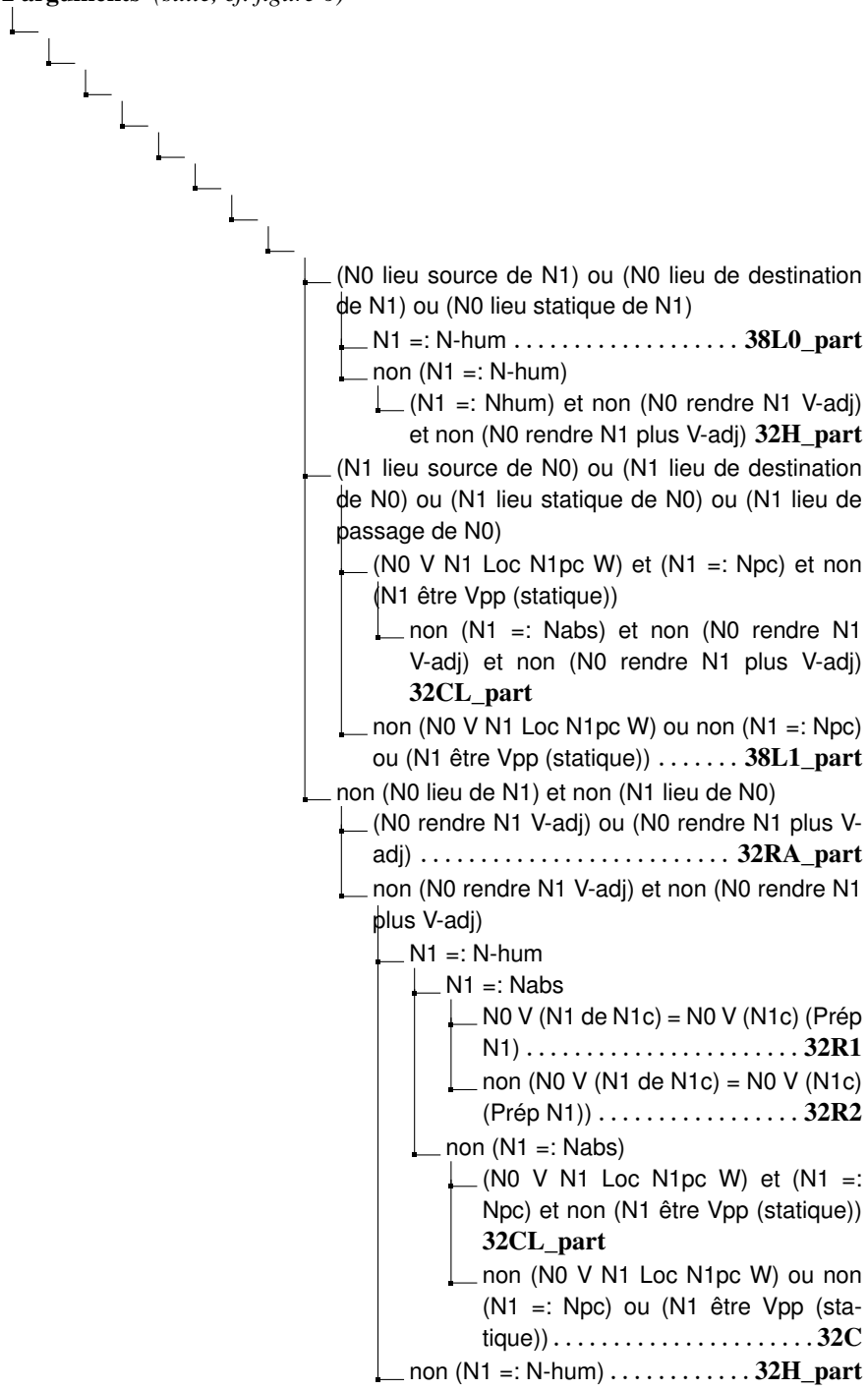

Figure 7. *Arbre de classement des verbes (6)*

1 argument
- N0 =: N-hum **31R**
- non (N0 =: N-hum) et N0 =: Nhum **31H**

Figure 8. *Arbre de classement des verbes (7)*