

La constitution d'un lexique de complémentation verbale du français André VALLI Delic-Université de Provence

Nous sommes en train de construire une base de données décrivant la complémentation verbale du français qui pourra servir dans le traitement automatique des langues.¹

Cette base de données réunira bientôt une liste des 975 entrées des verbes d'emploi le plus fréquent². A terme, elle réunira les entrées de plus de 6000 verbes, ce qui en fera une base de données plus importante que le lexique *Proton*.³

Cette base de données, qui exploite de façon manuelle les données des tables du Lexique-Grammaire de Maurice Gross⁴ présente une alternative intéressante à l'entreprise de chercheurs comme ceux du LORIA qui développent le projet de construction d'un lexique de verbes dédié au TAL grâce à une méthode de traduction du lexique-grammaire dans un format approprié aux systèmes de traitement automatique des langues⁵.

Les entrées verbales du Lexique de valences sont élaborées à partir de la consultation des dictionnaires de langue française de référence (*Le Grand Robert*, *Le Trésor de la Langue Française (TLF)* dans leurs versions électronique) et des données du lexique-grammaire de M. Gross et J.-P. Boons, A. Guillet et C. Leclère.⁶

Pour désigner les compléments figurant dans ces entrées verbales nous avons utilisé les notations disponibles dans la liste de compléments verbaux établie par Morris Salkoff dans sa grammaire en chaîne du français⁷.

Cette liste comporte des compléments signalés dans les tables du lexique-grammaire, dans l'espace « Tables » sur le site de visualisation des tables de l'Institut Gaspard Monge et ils sont illustrés par des exemples proposés dans l'espace « Phrases » du même site. C'est le

¹ M. Salkoff and A. Valli. 2005. A dictionary of French verbal complementation. In Proceedings of 2nd Language and Technology Conference. Human Language and Technologies as a Challenge for Computer Science and Linguistics. In memory of M. Gross and A. Zampolli, Poznan, Poland.

² Il s'agit des 975 verbes répertoriés parmi les 5.000 mots les plus fréquents du français dans l'étude de Juilland, Brodin, Davidovitch, *Frequency Dictionary of French Words* (Mouton, The Hague, 1970),

³ K. van den Eynde & P. Mertens (Université de Leuven), « Le Lexique syntaxique Proton: une base de données de la valence verbale en français », *Journée d'étude de l'ATALA : Interface lexique-grammaire et lexiques syntaxiques et sémantiques*, Organisé par S. Kahane (Modyco, UMR 7114), ENST, Paris, 12 mars 2005.

⁴ M. Gross. 1975. *Méthodes en syntaxe*. Hermann, Paris et J.P. Boons, A. Guillet, et Ch. Leclère. 1976. *La structure des phrases simples en français. Constructions transitives*. Droz. Genève- Paris., J.P. Boons, A. Guillet, et Ch. Leclère 1992. *La structure des phrases simples en français. Constructions transitives locatives*. Droz. Genève-Paris.

⁵ C. Gardent, B. Guillaume, I. Falk, G. Perrier, « Le lexique-grammaire de M. Gross et le traitement automatique des langues », *Journée d'étude de l'ATALA : Interface lexique-grammaire et lexiques syntaxiques et sémantiques*, Organisé par S. Kahane (Modyco, UMR 7114), ENST, Paris, 12 mars 2005

⁶ Nous remercions Christian Leclère qui a mis à notre disposition un classeur Excel qui donne l'inventaire des verbes étudiés avec mention des tables dans lesquelles ils sont recensés, et d'un exemple pour chacune des tables et Eric Laporte qui a accepté de mettre à notre disposition l'intégralité des Tables du Lexique-Grammaire sous un format Excel, dans le cadre d'une licence LGPL-LR.

⁷ Y. Geerbrant, J.M. Marandin, and M. Salkoff. 1976. Les sous-classes verbales pour un lexique automatique du français. In *Rapport du Laboratoire d'Automatique Documentaire et Linguistique, Université ParisVII*, Paris. M. Salkoff. 1979. *Analyse syntaxique du français. Grammaire en chaîne*. John Benjamins Amsterdam.

cas du complément [N₁ comme N₂/Adj], donné comme complément dans la table 38R pour le verbe *regarder* : *Il regarde Léa comme (une déesse+ un esprit)* .

Cette liste comporte également des compléments identifiés dans les tables du lexique-grammaire, par exemple [N_{sujet} V_{ant} COMP], à la table 6. Mais pour le verbe *regarder* on n'en trouve pas d'exemple illustré dans l'espace « Phrases ». C'est dans le Web que nous avons trouvé un exemple attesté de cette structure de complément : *Il suffisait de la regarder courant sur la Place St-Jacques...*(G)

On y trouve enfin des compléments non signalés dans les tables, sauf erreur de notre part. C'est le cas du complément [N_{sujet} comme V_{ant} COMP], même si M. Gross en signale l'existence, pour un nombre réduit de verbes ;⁸ notre lexique en retient deux attestations, une dans le Web, l'autre dans le *TLF* :

Je regarde cette déclaration comme prédisant une catastrophe (G)

Je regarde cette affaire comme étant .. mon affaire personnelle (TLF)

En construisant ces entrées, nous avons exploité aussi le corpus du Web. Nous en avons obtenu deux avantages. En premier lieu, ce corpus nous a permis de dégager des attestations des constructions de compléments phrastiques complexes décrits dans les tables du lexique-grammaire (les tables 4-19) qui ne figurent pas souvent dans les dictionnaires. Ainsi, pour le verbe *mériter*, la table 16 signale le complément [*que P à ce que P*], pour lequel il nous faudrait « inventer » une phrase. Nous avons trouvé dans le Web l'exemple suivant, qui offre une attestation de la structure tout à fait satisfaisante : *On mesure que la proposition d'écriture touche juste à ce que les textes qui viennent ne sont pas prévisibles* (G)

D'autre part, on connaît l'intérêt manifesté dans le domaine TAL pour la problématique du rattachement des groupes prépositionnels⁹. Le recours au corpus que représente le Web nous a permis d'élargir notre recherche aussi loin que possible afin d'identifier toutes les structures PN qui manifestent des propriétés de complément verbal, et cela pour chaque verbe étudié. Le résultat obtenu dépasse largement les données disponibles aussi bien dans les tables du lexique-grammaire que dans les dictionnaires classiques.

Enfin, nous avons commencé à réaliser la transcription des fichiers Word contenant les entrées verbales du lexique pour un environnement XML, selon le standard en train d'être développé actuellement¹⁰. De cette manière, seront rendus lisibles les patrons de compléments retenus. L'utilisation de XML permettra aussi d'introduire dans les entrées verbales toutes les informations concernant les sous-catégorisations des noms (essentiellement), qui sont nécessaires à l'exécution de l'ensemble des restrictions lexicales et grammaticales (des tests de contexte). Celles-ci limitent la possibilité d'occurrence d'un complément donné dans tel ou tel contexte.

⁸ *Méthodes en syntaxe 1975* : 67

⁹ M. Volk. 2001. Exploiting the WWW as a corpus to resolve PP-attachment. In *Proceedings of Conference on Corpus Linguistics* pages 601–606, Lancaster.

¹⁰ « Lexical Markup Framework » (LFM) item de travail de l'ISO TC 37/SC4, accessible en ligne sur <http://www.tagmatica.fr>