

Graphes paramétrés et lexique-grammaire

Éric Laporte
Institut Gaspard-Monge (IGM)
Université de Marne-la-Vallée
5, bd Descartes
77454 Marne-la-Vallée CEDEX 2

Le projet de représentation formelle syntaxique et sémantique du lexique et de la grammaire du français à l'IGM présente plusieurs spécificités.

1. Préexistence d'un lexique syntaxique

La première est historique. Elle repose sur un ensemble préexistant de données lexicales, syntaxiques et sémantiques, le lexique-grammaire du français, élaboré sous la direction de Maurice Gross de 1968 à 2001 (Gross, 1994). Les premières tentatives d'analyse syntaxique de phrases du français à l'aide d'une portion significative du lexique-grammaire datent du début des années 1990 (Roche, 1994). Pendant plus de vingt ans, la préparation des tables du lexique-grammaire est donc restée un travail essentiellement descriptif, qui était considéré comme un préliminaire à la réalisation effective ou même à la conception de structures de données et d'algorithmes pour l'analyse syntaxique. Tous les autres formalismes d'analyse syntaxique des années 1970 et 1980 ont été conçus sans prendre en compte des données lexicales formelles de plus que quelques dizaines d'entrées, considérées comme des exemples suffisants. Cette différence de stratégie, outre peut-être une opposition "sociologique" entre linguistes et ingénieurs, reflétait probablement des vues divergentes sur les dépendances mutuelles entre les contenus des lexiques syntaxiques et la structure des grammaires syntaxiques. Maurice Gross jugeait prématuré de fixer un formalisme syntaxique avant d'avoir une idée précise de la structure, de la forme et de la taille du lexique syntaxique avec lequel il aurait à fonctionner. Par exemple, y aurait-il beaucoup d'entrées lexicales de formes figées et à quoi ressembleraient-elles ? et la même question se posait pour les constructions à verbe support et nom prédicatif. Les auteurs de formalismes grammaticaux, quant à eux, repoussaient dans le futur le début de la construction d'un lexique, qu'il serait toujours temps de concevoir et de réaliser de telle sorte qu'il soit compatible avec les choix qu'ils faisaient alors en matière de représentation de la grammaire. Les points de vue ont commencé à converger dans les années 1990 avec, d'une part, la vogue de la lexicalisation des formalismes grammaticaux, et d'autre part, au CERIL¹, les essais d'exploitation des données du lexique-grammaire pour l'analyse syntaxique par transducteurs finis.

Le lexique-grammaire du français et de quelques autres langues, résultat de ces événements, présente une large couverture lexicale. Pour le français on compte 15 000 entrées de verbes simples ; les effectifs de chacun des autres fragments (noms prédicatifs, phrases figées, adverbes figés) se comptent en dizaines de milliers d'entrées. Il ne s'agit pas d'un lexique exhaustif, et ce encore moins dans les autres langues représentées, pour lesquelles il existe des fragments du même ordre de grandeur mais moins nombreux.

Les distinctions entre emplois sont particulièrement appréciées par les utilisateurs du lexique-grammaire. Par exemple, une expression figée comme *remonter la pente*, ou un sens figuré lexicalisé comme dans *Luc remonte sa montre*, sont décrits dans des entrées indépendantes de celle consacrée au sens premier (*Luc remonte la valise dans le grenier*). De telles distinctions d'emplois sont absentes des autres ressources à visées syntaxiques (ex. Penn Treebank), on ne les retrouve que dans les réseaux sémantiques.

¹ Centre d'études et de recherches en informatique linguistique, Évry, Université Paris 7.

2. La forme tabulaire et la forme grammaire

La forme tabulaire du lexique-grammaire telle qu'elle est décrite dans la littérature (Gross 1994) est conçue pour la construction par les lexicographes, la lecture, la maintenance manuelle. L'exploitation pour le traitement automatique des langues nécessite une forme un peu différente, que nous appellerons une "forme grammaire". Une compilation de celle-là vers celle-ci permet d'assurer la maintenance de la forme grammaire. De ce point de vue, la forme tabulaire est comparable au code source d'un programme, et la forme grammaire au code exécutable : il est déraisonnable d'imaginer la maintenance manuelle directe de dizaines de milliers d'entrées codées dans un formalisme de structures de traits.

L'avantage de la forme tabulaire est que les données sont présentées d'une façon dense, faisant apparaître sur une page ou un écran, en clair, plusieurs dizaines d'éléments lexicaux décrits, et d'une manière également assez lisible, plusieurs dizaines de propriétés syntaxiques et sémantiques codées. Cette densité de l'information est essentielle à la construction de données lexicologiques à partir de l'introspection. Les entrées lexicales sont repérées par des exemples de phrases, indispensables en cas de mots ambigus. Cette forme tabulaire était gérée sous la forme de table Excel au cours des dernières années, et nous passons maintenant à un format XML avec éditeur spécialisé.

Sous leur simple forme tabulaire, donc sans représentation structurée des propriétés syntaxiques et sémantiques qui apparaissent en colonnes, les données sont a priori compatibles avec les algorithmes d'apprentissage automatique à partir d'exemples. Cependant, leur utilisation pour une analyse syntaxique plus précise nécessite de les compiler sous une forme grammaire, qui peut être l'un des nombreux formalismes grammaticaux en usage.

On peut citer plusieurs expériences passées de génération de formes grammaire : Roche 1994 (transducteurs finis), Hathout & Namer 1998 (PATR-II, HPSG et TAG), Paumier 2001 (RTN), et d'autres expériences sont en cours. La difficulté d'une telle entreprise dépend bien sûr de la complexité du formalisme cible. Il s'agit aussi d'effectuer des approximations si l'on désire incorporer dans la grammaire des propriétés qui ne sont pas explicitement codées dans le lexique-grammaire, par exemple l'effacement du complément d'objet second dans une phrase passive à verbe distributionnel. En dehors de ces difficultés prévisibles et inévitables, ces expériences n'ont pas rencontré d'obstacles particuliers et démontré la faisabilité du principe.

Il existe bien sûr une nécessité supplémentaire pour l'exploitation du lexique-grammaire en TAL : un logiciel d'analyse syntaxique. Toute ressource linguistique nécessite du logiciel pour son application, mais l'analyse syntaxique (dite "profonde") a peut-être plus de complexité inhérente.

Les utilisations passées du lexique-grammaire sont peu nombreuses si l'on prend en compte qu'un fragment significatif existait déjà il y a 30 ans. L'explication "sociologique" que nous avons déjà évoquée ci-dessus vient à l'esprit : il est difficile d'associer au lexique-grammaire des problèmes à la fois purement informatiques et indépendants du contenu linguistique, ce qui a certainement contribué à décourager les informaticiens purs. Par ailleurs, le lexique-grammaire tend à mettre en évidence la complexité de la langue : diversité des constructions syntaxiques, irrégularité de la correspondance sens/syntaxe... ce qui peut également décourager des amateurs potentiels. En fait, la plupart des utilisateurs se restreignent à un sous-ensemble des entrées (seulement les verbes locatifs, par exemple) ou des propriétés, par exemple Bourigault & Frérot 2004 (uniquement l'appartenance de verbes distributionnels à certaines tables) ou Danlos 2005 (une sélection de propriétés liées aux constructions impersonnelles).

Étant donné les potentialités d'exploitation des tables du lexique-grammaire, l'IGM a entrepris de les rendre accessibles au public sous format électronique avec l'infrastructure

juridique (licence LGPL-LR), technique (visualisation, format XML et éditeur XML spécialisés) et scientifique (documentation des propriétés) nécessaire. Elles ont été utilisées dans plusieurs travaux de recherche récents et en cours. Les deux tiers des tables de verbes simples sont ainsi désormais librement accessibles sur le web. Comme dans le cas de nombreuses autres ressources linguistiques, la mise à disposition n'est pas totale. Cette politique, également pratiquée par le Linguistic Data Consortium à l'Université de Pennsylvanie, et d'autres producteurs de ressources, s'appuie sur plusieurs raisons : d'une part, la nécessité d'un contact entre l'utilisateur et le gestionnaire de la ressource favorise une relation scientifique entre eux qui est utile à la maintenance de la ressource ; d'autre part, des ressources créées à partir de fonds publics nationaux peuvent avoir des bénéficiaires internationaux, notamment les ressources réalisées en France sur d'autres langues que le français ; enfin, les ressources linguistiques constituent un marché dont les acteurs ne sont pas prêts à accepter le principe de l'échange automatiquement gratuit (le catalogue d'ELDA/ELRA fournit une documentation sur certains types de ressources).

3. Une forme grammaire simple : les graphes paramétrés

Une autre spécificité du lexique-grammaire est la simplicité du formalisme syntaxique sous-jacent (cf. Harris, 1976). Le formalisme utilisé ne comporte pratiquement pas de structures profondes (quelques "phrases théoriques" pourraient cependant y être comparées). Les structures de surface sont simples : les constituants utilisés sont la phrase et le groupe nominal, et il n'y a pas de distinction entre phrase et groupe verbal, ni entre groupe nominal et groupe prépositionnel. Le formalisme est neutre, par exemple, en ce qui concerne le choix entre grammaires de constituants et grammaires de dépendance. Mis à part les traits sémantiques comme humain, concret, abstrait, qui servent d'approximation à la sélection des arguments des prédicats, les traits ne sont pas formalisés (le genre et le nombre, par exemple, interviennent assez peu dans la formulation des propriétés syntaxiques des éléments lexicaux). Cette simplicité est un parti pris volontaire, elle vise à ne pas ajouter à la complexité inhérente de l'interface lexique/grammaire une complexité artificielle supplémentaire.

Les formalismes grammaticaux qui ont été interfacés avec le lexique-grammaire au LADL, au CERIL et à l'IGM sont essentiellement les réseaux de transitions récurrents (RTN) et les transducteurs finis. Ces deux choix sont naturellement cohérents avec les deux partis pris que nous avons évoqués ci-dessus : d'une part, la lisibilité des données linguistiques, d'autre part la simplicité formelle.

La lisibilité des données est essentielle dans notre modèle, qui valorise la construction et la maintenance manuelles (ou assistées d'aides automatiques mais essentiellement fondées sur l'introspection) non seulement des lexiques mais aussi des grammaires (par opposition aux modèles qui recherchent dans l'apprentissage automatique à partir d'exemples une neutralité des résultats vis-à-vis des erreurs que commettent les auteurs humains). Les RTN structurent la grammaire en graphes petits, lisibles, dans lesquels les parties communes sont souvent mises en facteur.

Les RTN tels qu'ils ont été utilisés par exemple par Paumier (2001) sont compatibles avec la simplicité formelle du formalisme syntaxique sous-jacent. En particulier, la structuration du RTN en graphes ne coïncide pas nécessairement avec les limites des constituants syntaxiques comme les phrases ou les groupes nominaux. Ces limites de constituants peuvent, au moins dans certaines des étapes de la construction de la grammaire, ne pas être marquées du tout. La structuration du RTN en graphes reste ainsi un outil dans la main du concepteur de la grammaire, et cet outil est libre pour un des problèmes essentiels de la grammaire: la lexicalisation.

La lexicalisation de la grammaire représentée sous forme de RTN est automatisée, les informations lexicales sur la syntaxe spécifique aux éléments lexicaux étant recherchées dans

les tables de lexique-grammaire. La préparation manuelle de cette lexicalisation consiste à marquer dans les graphes les éléments (mots grammaticaux, acceptabilités de constructions...) à déterminer en fonction de la valeur d'entrées lexicales, éléments qui devront donc être modifiés lors de la lexicalisation. Lors de ce marquage, des paramètres sont insérés dans les graphes, d'où leur nom de graphes paramétrés. Le système Unitex (logiciel libre) produit des versions lexicalisées des graphes paramétrés en remplaçant les paramètres par des valeurs déterminées par un élément lexical présent dans le graphe, l'élément lexicalisateur. Les valeurs des paramètres dépendent soit de la sous-catégorie syntaxique à laquelle appartient l'élément lexicalisateur, soit de la valeur lexicale précise de cet élément. Dans les deux cas, les valeurs sont trouvées dans des tables du lexique-grammaire.

Ce formalisme a ses limites, qui ne pourront probablement être dépassées qu'au fur et à mesure de son utilisation. La présentation par affiche de Blanc, Constant & Sastre dans cette même Journée présente une extension récente à des RTN avec unification. On peut aussi faire les quelques remarques suivantes, qui ouvrent la voie à d'autres extensions éventuelles.

Le formalisme des graphes paramétrés ne permet de lexicaliser un graphe que par rapport à un seul élément lexicalisateur, ce qui oblige à structurer la grammaire en fonction de la lexicalisation. Cette obligation n'a pas été ressentie comme une difficulté jusqu'à maintenant, dans la mesure où, comme on l'a vu, la structuration du RTN en graphes n'est pas imposée théoriquement, mais reste un outil dans la main de l'auteur de la grammaire. De plus, la phrase simple composée d'une forme prédicative et d'un ensemble d'arguments se prête tout à fait à cette structuration, l'élément lexicalisateur étant l'entrée lexicale qui constitue la forme prédicative. Cependant, certaines constructions mettront probablement en jeu une lexicalisation par rapport à plusieurs éléments lexicaux : la forme prédicative et un déterminant, la forme prédicative et un substantif, etc.

On peut également remarquer que le modèle actuel permet de lexicaliser le contenu d'un graphe paramétré, mais pas le nom qui identifie le graphe lui-même. Or c'est ce nom ou cet identifiant qui insère une construction syntaxique dans une autre, et cette opération peut dépendre d'un élément lexical présent dans la construction syntaxique insérée (par exemple, une relative dont le pivot est un nom obligatoirement à sémantique humaine peut être inséré dans un groupe nominal dont la tête est à sémantique humaine).

4. Évaluation

Plusieurs expériences d'évaluation du contenu du lexique-grammaire peuvent être citées. Il n'existe que des évaluations approximatives et partielles, faute d'un corpus étiqueté suffisamment étendu et qui contiendrait des informations aussi précises que celles du lexique-grammaire, comme les distinctions d'emplois. Hathout & Namer 1998 prennent comme référence un corpus de textes pour l'évaluation de la couverture lexicale (70 %) et de la pertinence des analyses présentées par l'analyseur syntaxique (51 à 78 %). La référence pour l'évaluation peut également être un dictionnaire, comme le TLFi dans l'expérience de Jacquey & Naels présentée dans cette même Journée.

Quoi qu'il en soit, le lexique-grammaire reflète une importante connaissance accumulée sur le lexique de formes prédicatives du français et sur les propriétés syntaxiques dépendant des entrées lexicales. Il aborde également plusieurs autres langues européennes et non européennes. Peu d'entreprises systématiques de description syntaxique indépendante d'un corpus de textes lui sont comparables. À ce titre, sa préservation et son évolution future concernent les équipes qui s'intéressent à la lexicalisation.

Références

- Bourigault, Didier, Cécile Frérot. 2004. « Ambiguïté de rattachement prépositionnel : introduction de ressources exogènes de sous-catégorisation dans un analyseur syntaxique de corpus endogène », In *TALN 2004, XIe Conférence sur le Traitement Automatique des Langues Naturelles, Fès, Maroc*.
- Danlos, Laurence. 2005. "Automatic recognition of French expletive pronoun occurrences", *Proceedings of Ninth International Symposium on Social Communication, Santiago de Cuba*.
- Gross, Maurice. 1994. "Constructing Lexicon-grammars". In *Computational Approaches to the Lexicon*, Atkins and Zampolli (eds.), Oxford Univ. Press, pp. 213-263.
- Harris, Zellig. 1976. *Notes du cours de syntaxe*. Trad. M. Gross, Paris: Le Seuil.
- Namer, Fiammetta, Nabil Hathout. 1998. "Automatic Construction and Validation of French Large Lexical Resources. Reuse of Verb Theoretical Linguistic Descriptions", *Proceedings of the First International Conference on Language Resources and Evaluation*, Grenade.
- Paumier, Sébastien. 2001. "Some remarks on the application of a lexicon-grammar", *Linguisticae Investigationes* 24:2, Amsterdam/Philadelphia, John Benjamins, pp. 245-256.
- Roche, Emmanuel. 1994. Two Parsing Algorithms by Means of Finite State Transducers. *COLING 1994*, pp. 431-435.