

Le lexique-grammaire de M. Gross et le traitement automatique des langues

Claire Gardent, Bruno Guillaume, Ingrid Falk, Guy Perrier

LORIA & ATILF, Nancy

{Claire.Gardent, Bruno.Guillaume, Ingrid.Falk, Guy.Perrier}@loria.fr

Résumé

Nous proposons les grandes lignes d'une méthode de traduction du lexique-grammaire de Maurice Gross dans un format approprié aux systèmes de traitement automatique des langues.

1. Le lexique-grammaire de Maurice Gross

Une large part des travaux en syntaxe se concentre sur l'identification et la formalisation de règles générales s'appliquant à une classe étendue de mots. Typiquement, les règles de transformation de Chomsky décrivent des relations systématiques entre les diverses structures syntaxiques. Mais, comme Chomsky lui-même le remarquait (Cho65), ces généralisations sont sujettes à de fortes contraintes lexicales. Étant donné un mot particulier, la question se pose de savoir si une généralisation donnée s'applique à ce mot. En d'autres termes, la description complète de la syntaxe d'une langue implique non seulement l'identification de règles générales mais aussi la détermination de quel mot exige, autorise ou interdit l'application de quelle règle. C'est ce qu'a cherché à faire pour le français Maurice Gross à travers le lexique-grammaire (Gro75).

Le lexique-grammaire de Maurice Gross est une description systématique des propriétés syntaxiques et sémantiques des foncteurs syntaxiques du français, c'est-à-dire les verbes, les noms prédictifs et les adjectifs. Il est organisé en groupes de tables, chaque groupe étant associé à une catégorie syntaxique donnée (verbes pleins, verbes supports, noms, etc.). Au sein d'un groupe, une table correspond à une ou deux constructions syntaxiques particulières et rassemble tous les mots qui entrent dans cette ou ces constructions. Par exemple, la table 1 des verbes contient tous les verbes qui admettent en plus d'un sujet un complément infinitif mais pas un complément qui soit une complétive finie ou un nom. Une table est divisée en lignes selon les mots qu'elle contient et en colonnes selon les propriétés syntaxiques ou sémantiques qui s'appliquent à ces mots et leurs arguments. À l'intersection d'une ligne et d'une colonne, un signe + ou - indique que la propriété indiquée en tête de la colonne s'applique positivement ou négativement au mot placé en tête de la ligne. Cette propriété est soit un ajout d'information sur le mot ou un de ses arguments, soit une transformation du cadre de sous-catégorisation de base associée à la table.

Actuellement, le lexique-grammaire est surtout développé pour les verbes et les locutions prédictives.

Pour ce qu'on appelle les verbes simples, 5000 d'entre eux ont été écrits sur un total de 15000 en usage (Gro75; BGL76a; BGL76b). En outre, 25 000 locutions prédictives ont été décrites de même que 20 000 locutions construites avec *être* ou *avoir* (Gro89).

2. Lexiques électroniques et traitement automatique des langues

La robustesse et la précision des systèmes TAL reposent entre autres sur l'acquisition des connaissances linguistiques et extra-linguistiques pertinentes. Dans cette section, nous identifions deux types de connaissances présentes dans le lexique-grammaire qui sont particulièrement pertinentes pour le TAL : l'information sur la sous-catégorisation et l'information sur les alternances.

2.1. sous-catégorisation

Le lexique-grammaire contient des informations détaillées et exhaustives sur la sous-catégorisation des foncteurs syntaxiques, c'est-à-dire sur le nombre et le type de leurs arguments. Ainsi par exemple, les tables du LADL incluent pour chaque emploi verbal :

- un ou plusieurs cadres de sous-catégorisation sous la forme d'une liste d'arguments,
- pour chaque cadre de sous-catégorisation, une information morpho-sémantico-syntaxique sur les arguments et sur le verbe qui inclura par exemple :
 - pour le verbe, une information sur son type (verbe normal, verbe U de Harris), sur l'auxiliaire utilisée pour construire les temps composés (être ou avoir), sur les contraintes de concordance des temps portant sur les arguments verbaux, etc ;
 - pour les arguments nominaux, une information sur le caractère humain ou non du référent, sur le nombre, sur la pronominalisation, sur les déterminants permis, etc ;
 - pour les arguments prépositionnels, une information sur le type (e.g., locatif) et sur la valeur de la préposition utilisée ;
 - pour les arguments phrastiques, une information sur le mode (indicatif, infinitif, subjonctif), sur la structure de contrôle (par le sujet, l'objet), sur les verbes admis, etc.

Or l'information sur la sous-catégorisation est une composante essentielle d'un système de TAL. Plus spécifiquement, (Gar05) récapitule un certain nombre de résultats récents montrant qu'une information fine et exhaustive sur la sous-catégorisation permet d'une part,

Nous remercions Eric Laporte et l'équipe d'informatique linguistique de l'IGM d'avoir mis à notre disposition sous un format électronique certaines tables du LADL. Nous remercions aussi la Région Lorraine et les établissements publics qui financent le Contrat de Plan Etat Région dans lequel s'inscrit le travail de recherche présenté dans ce papier.

d'accroître la précision et la couverture des analyseurs syntaxiques et d'autre part, d'améliorer le traitement sémantique utilisée par exemple dans certains systèmes de questions/réponses. Ainsi par exemple, (BC93) montre que la moitié des erreurs d'analyse résultent d'information manquante ou erronée sur la sous-catégorisation tandis que (CF04) démontre que pour un domaine donné, l'utilisation d'une grammaire enrichie par une information de sous-catégorisation détaillée permet d'améliorer le taux de succès de l'analyse syntaxique de 15% par rapport à des évaluations faisant intervenir la même grammaire avec une information de sous-catégorisation moins riche. En outre, (HLP⁺00) présente des résultats indiquant que l'information de sous-catégorisation est un facteur clé pour obtenir des traductions automatiques de bonne qualité tandis que (JMdr04) montre que l'extraction de relations syntaxiques permet une augmentation substantielle du nombre de questions répondues par rapport à la simple utilisation de chablon de surface.

2.2. Alternances

Un autre type d'information contenue dans les tables du LADL qui est pertinente pour les systèmes TAL est l'information sur les alternances¹ c'est-à-dire, sur les effacements et mouvements que les arguments d'un foncteur syntaxique peuvent subir. Par exemple, dans le lexique grammaire un verbe peut être spécifié comme acceptant/rejetant les alternances suivantes :

- passif - *Le chat mange la souris/La souris est mangée par le chat*
- réciproque - *Luc flirte avec Léa/Luc et Léa flirtent*
- alternance locative - *Les fautes pullulent dans ce texte/Ce texte pullule de fautes*
- alternance source - *Un paradoxe résulte de cette situation/De cette situation résulte un paradoxe*
- extraposition du sujet - *Un malheur arrive à Paul/Il arrive un malheur à Paul*
- impersonnel avec extraposition du sujet - *On a consenti à ce qu'elle vienne/Il a été consenti à ce qu'elle vienne*
- forme inchoative - *Jean sonne la cloche/La cloche sonne*
- construction à verbe support - *Jean crie/Jean pousse un cri*
- alternance d'ascension possesseur/partie du corps - *Jean imite l'attitude de Marie/Jean imite Marie dans son attitude*

C'est précisément sur la base de ce type d'information que Beth Levin a développé son système de classification sémantique pour les verbes anglais (Lev93). L'intuition sous-jacente y est que les propriétés syntaxiques d'un verbe reflètent ses propriétés sémantiques. La méthodologie employée consiste à identifier pour chaque verbe l'ensemble des alternances qu'il accepte/rejette puis de déterminer des classes sur la base de cette information : les verbes qui acceptent/rejettent le même ensemble d'alternances forment une classe.

¹Il est usuel de distinguer dans la littérature entre alternances et redistributions, les redistributions étant vues comme des opérations plus générales que les alternances. Cette distinction n'étant pas fondamentale dans le contexte de cet article, nous ne la ferons pas et engloberons sous le terme "alternance" à la fois les alternances et les redistributions.

Parce qu'il assied la classification des verbes sur une base théorique et empirique solide, le travail de Beth Levin a eu un impact majeur dans le monde de la sémantique computationnelle. Il est à la base en particulier de VerbNet (KDP00), un lexique électronique répertoriant une information syntaxique et sémantique détaillée pour 2500 verbes anglais.

3. Les lexiques électroniques existant

Bien qu'il soit aujourd'hui clair que le lexique est une composante essentielle des systèmes TAL, les ressources disponibles sont rares.

Pour l'anglais, COMLEX Syntax (MGM94) contient une information détaillée sur la sous-catégorisation de 38 000 mots dont 6 000 verbes tandis que VerbNet décrit 4 000 sens verbaux à partir de 191 classes sémantiques et 52 cadres syntaxiques.

Pour le français, plusieurs lexiques électroniques sont disponibles mais la plupart concernent la morphologie plutôt que la syntaxe². Ainsi le lexique LEFFF (Lexique des Formes Fléchies du Français) contient 5 000 verbes et 200 000 formes fléchies mais l'information associée est purement flexionnelle (CSL04). De même, le lexique Litote³ définit 300 000 formes fléchies pour 6000 verbes en extension comme en intension. De façon similaire, Morfalou⁴, MulText (IV94) et ABU (Ass) se limitent à associer aux mots une information essentiellement morphosyntaxique. Quant aux alternances, (SD99) décrit les alternances des verbes français mais ne traite que de 1 000 verbes.

4. La construction d'un lexique des verbes dédié au TAL à partir du lexique-grammaire

Comme nous l'avons vu dans la section 2., le lexique-grammaire de Gross contient une information détaillée et exhaustive à la fois sur la sous-catégorisation et les alternances. De plus, il a été numérisé par le Laboratoire d'Automatique Documentaire et Linguistique (LADL) et il est maintenant partiellement disponible sous une licence LGPL-LR. Tout ceci facilite la constitution d'une ressource lexicale appropriée au TAL. Pour y parvenir, il est cependant nécessaire de changer la structure et le format de l'information :

- toute l'information relative à un verbe doit être collectée à travers les différentes tables et regroupée dans une ou plusieurs entrées lexicales attachées à ce verbe ;
- les structures de données et les catégories linguistiques doivent être compatibles avec la pratique habituelle en linguistique informatique ;
- le format des données doit être compatible avec les standards existants ou en cours de définition pour les

²Nous mettons ici de côté les dictionnaires électroniques tels que le TLFi (Trésor de la langue française informatisée) qui sont des dictionnaires traditionnels et donc à destination humaine plutôt que computationnelle.

³<http://www.loria.fr/equipes/calligramme/litote/>

⁴<http://lorelei.loria.fr/morfalou/>

données linguistiques (norme ISO/TC 37/SC 4⁵).

Dans la suite, nous nous concentrerons sur le premier point, c'est-à-dire le regroupement de toute l'information se rattachant à un verbe donné dans une ou plusieurs entrées lexicales.

4.1. La procédure générale

Si l'on veut disposer en fin de traitement d'un lexique praticable le plus largement possible par des utilisateurs variés, linguistes ou informaticiens, il est important de représenter l'information lexicale d'une manière qui soit relativement neutre tant par rapport à la théorie linguistique que par rapport aux applications TAL. Une façon standard de le faire consiste à écrire le contenu d'une entrée lexicale à l'aide de structures de traits récursives, c'est-à-dire d'ensembles de couples (attribut-valeur) où les valeurs peuvent être des atomes, des disjonctions d'atomes, des négations de disjonctions ou elles-mêmes des structures de traits.

Ensuite, la manière de procéder consiste à considérer les tables les unes après les autres et à convertir le contenu de chaque table en un ensemble d'entrées lexicales, chaque entrée associant un usage particulier d'un verbe à l'ensemble des propriétés linguistiques que lui attribue la table. Plus précisément, une procédure générale de conversion peut être décrite de la façon suivante pour une table T quelconque :

1. pour chaque ligne L de la table T, créer une entrée lexicale associant le verbe V concerné par la ligne L au cadre de sous-catégorisation de base associée à T ;
2. enrichir chaque entrée lexicale créée à l'étape 1 en utilisant le contenu des colonnes de la table T.

C'est ce travail que nous avons commencé à entreprendre et dont nous rendons compte ici. Pour illustrer notre propos, nous prendrons l'exemple de la table 1 et dans la table 1 la ligne correspondant au verbe *entraîner*.

La première étape consiste à créer une nouvelle entrée lexicale pour *entraîner* avec comme contenu le cadre de sous-catégorisation de base associée à la table 1. Un cadre de sous-catégorisation est défini par une structure de traits fermée, l'un des traits v correspondant au verbe et les autres à ses arguments a_0, a_1, \dots, a_n . La valeur de chacun des traits est elle-même une structure de traits mais cette fois ouverte, c'est-à-dire qu'elle peut être enrichie de nouveaux traits. Par exemple, le cadre de sous-catégorisation de base associée à la table 1 se présente comme suit, sachant que la valeur u du trait $v\text{-type}$ renvoie à la classe U des verbes de Harris et *contrôleur* au contrôleur du complément infinitif, ici le sujet.

$$\left[\begin{array}{l} a_0 = [\], \\ v = [\text{cat}=v, v\text{-type}=u], \\ a_1 = [\text{cat}=p, \text{mode}=\text{inf}, \text{contrôleur}=a_0] \end{array} \right]$$

L'entrée lexicale qui a été créée pour le verbe *entraîner* comme contenu la structure de traits ci-dessus.

La seconde étape consiste à parcourir les colonnes de la table 1 de gauche à droite. Le traitement de chaque colonne consiste soit à enrichir des entrées existantes, soit à

créer de nouvelles entrées. Ainsi, par exemple, le traitement de la table 1 pour le verbe *entraîner* enrichit son cadre de sous-catégorisation de base de la façon suivante⁶ :

$$\left[\begin{array}{l} a_0 = [\text{hum} = -, \text{nc} = +], \\ v = \left[\begin{array}{l} \text{particule_post} = \text{là}, \\ \text{cat} = v, \\ v\text{-type}=u, \\ \text{concTemps} = -, \\ \text{statique} = +, \\ \text{prep} = [\text{à}], \\ \text{aux} = [\text{être}] \end{array} \right], \\ a_1 = \left[\begin{array}{l} \text{vc} = [\text{pouvoir}, \text{savoir}, \text{devoir}], \\ \text{tc} = [\text{passé}, \text{présent}, \text{futur}], \\ \text{cliticisable} = +, \\ \text{cat} = p, \\ \text{mode} = [\text{inf}, \text{ind}, \text{subj}], \\ \text{contrôleur} = a_0, \\ \text{optionnel} = + \end{array} \right] \end{array} \right]$$

D'autres colonnes de la table 1 indiquent qu'une transformation donnée est applicable au cadre de sous-catégorisation de base. Ainsi, pour *entraîner*, sont créées les entrées lexicales suivantes qui correspondent au remplacement du complément à l'infinitif par un complément humain qui est soit introduit par une préposition, soit un objet direct, ce qu'exprime le trait cat de a_1 :

$$\left[\begin{array}{l} a_0 = [\text{hum} = -, \text{nc} = +], \\ v = \left[\begin{array}{l} \text{particule_post} = \text{là}, \\ \text{cat} = v, \\ v\text{-type}=u, \\ \text{concTemps} = -, \\ \text{statique} = +, \\ \text{prep} = [\text{à}], \\ \text{aux} = [\text{être}] \end{array} \right], \\ a_1 = [\text{cat} = \text{sp}, \text{hum} = +] \end{array} \right]$$

$$\left[\begin{array}{l} a_0 = [\text{hum} = -, \text{nc} = +], \\ v = \left[\begin{array}{l} \text{particule_post} = \text{là}, \\ \text{cat} = v, \\ v\text{-type}=u, \\ \text{concTemps} = -, \\ \text{statique} = +, \\ \text{prep} = [\text{à}], \\ \text{aux} = [\text{être}] \end{array} \right], \\ a_1 = [\text{cat} = \text{n}, \text{hum} = +] \end{array} \right]$$

La traduction du lexique-grammaire en un ensemble d'entrées lexicales se fait en deux étapes. La première étape se fait manuellement par l'étude des entêtes des colonnes : on cherche à déterminer l'effet d'une colonne (enrichissement d'une entrée lexicale existante ou création d'une nouvelle entrée) et à formaliser cet effet par le biais d'une procédure. Comme on retrouve certaines entêtes d'une table à l'autre, on peut bien entendu réutiliser les procédures créées pour des tables précédentes. La deuxième étape consiste à répéter l'application des procédures ainsi créées à l'ensemble des verbes et à l'ensemble des colonnes d'une table donnée.

⁶ nc signifie sémantiquement non-contraint c'est-à-dire référant soit à un humain, un objet concret ou abstrait, soit à un événement.

4.2. Quelques questions cruciales

Le choix des traits Pour être utilisable largement, une ressource doit se conformer à l'usage informatique et linguistique général. Linguistiquement, les noms de traits et les catégories utilisées doivent avoir un sens pour un auditoire le plus large possible. A cette fin, nous comptons avoir recours aux catalogues proposés par Multext, EAGLES, et plus récemment par le standard ISO (TC37/SC4) "Lexical Markup Framework". Ce dernier en particulier fournit un modèle de représentation de données de haut niveau et garantit un maximum d'interopérabilité avec des applications informatiques diverses.

Le cadre et éventuellement l'argument de référence pour la procédure associée à une colonne Lorsqu'une procédure associée à une colonne consiste à ajouter un nouveau cadre sous-catégorisation, se pose la question de savoir à partir de quel cadre celui-ci est créée : est-ce le cadre de sous-catégorisation initial de la table ou un cadre précédemment ajouté ? Les entêtes des colonnes sont structurées de façon arborescente, ce qui aide à résoudre ce problème mais malheureusement dans la version électronique des tables dont nous disposons cette structuration est absente. Cette absence est aussi gênante lorsque la procédure associée à une colonne consiste à enrichir un cadre existant par enrichissement d'un trait de sa tête ou de ses arguments. Dans ce cas en effet, il faut en plus retrouver l'argument qui va être enrichi.

L'interdépendance et le traitement conjoint de plusieurs colonnes d'une même table Dans le lexique-grammaire, on observe que souvent plusieurs colonnes d'une même table sont interdépendantes. Aussi, plutôt que de définir des procédures séparées par colonne, il semble plus judicieux de définir une procédure qui permette de traiter simultanément des colonnes interdépendantes. Souvent, cette opportunité se présente quand plusieurs colonnes contribuent à la mise à jour du même trait. Dans la table 1, on le rencontre par exemple pour les traits *aux*, *tc* et *hum* qui renvoient respectivement à l'auxiliaire de temps qui s'emploie avec le verbe, au temps de la complétive et au caractère humain ou pas de l'argument a_1 . Pour ne prendre que *tc*, trois colonnes de la table 1, $Tc = : passé$, $Tc = : présent$ et $Tc = : futur$ contribuent à sa valeur, ce qui justifie leur traitement simultané.

5. Conclusion

Nous ne présentons ici qu'un début de travail qui doit se poursuivre par le traitement de toutes les tables disponibles. Ensuite, il s'agit d'abstraire des données lexicales produites des principes généraux sous-jacent au lexique-grammaire de Gross, qui vont permettre notamment de factoriser le lexique produit. Nous projetons d'utiliser le lexique résultant pour ancrer des grammaires électroniques telles que les grammaires d'arbres adjoints ou les grammaires d'interaction (GP05). Il sera intéressant aussi de comparer les résultats et la méthode utilisée avec la traduction partielle des tables du LADL dans un lexique intermédiaire menée par (HN98). Ce lexique est écrit en PATR-II et il a été utilisé ensuite pour engendrer un lexique HPSG et un lexique TAG.

6. Références

- Dictionnaire des mots communs. Conservatoire National des Arts et Métiers.
- E. Briscoe and J. Carroll. Generalised probabilistic Ir parsing for unification-based grammars. *Computational Linguistics*, 1993.
- J.-P. Boons, A. Guillet, and C. Leclère. *La structure des phrases simples en français. I : Constructions intransitives*. Droz, Genève, 1976.
- J.-P. Boons, A. Guillet, and C. Leclère. *La structure des phrases simples en français. ii : Classes de constructions transitives*. Technical report, Univ. Paris 7, 1976.
- J. Carroll and A. Fang. The automatic acquisition of verb subcategorisations and their impact on the performance of an hpsg parser. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP)*, pages 107–114, Sanya City, China, 2004.
- N. Chomsky. *Aspects of the theory of syntax*. The MIT Press, 1965.
- L. Clément, B. Sagot, and B. Lang. Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of LREC'04*, Lisbonne, 2004.
- C. Gardent. Maurice Gross Grammar-Lexicon and Natural Language Processing. 2005. submitted.
- B. Guillaume and G. Perrier. Interface lexique-grammaire via des structures de traits. In *Journée d'étude de l'ATALA (Interface lexique-grammaire)*, 2005. soumis.
- M. Gross. *Méthodes en syntaxe*. Hermann, 1975.
- G. Gross. *Les constructions converses du français*. Droz, Genève, 1989.
- C. Han, B. Lavoie, M. Palmer, O. Rambow, R. Kittredge, T. Korelsky, and N. Kim. Handling structural divergences and recovering dropped arguments in a korean/english machine translation system. In *Proceedings of the Association for Machine Translation in the Americas*, pages 40–53, Berlin/New York, 2000. Springer Verlag.
- N. Hathout and F. Namer. Automatic construction and validation of french large lexical resources : Reuse of verb theoretical linguistic descriptions. In *First International Conference on Language Resources and Evaluation, Granada, Spain*, pages 627–636, 1998.
- N. Ide and J. Veronis. Multext : Multilingual text tools and corpora. In *Proceedings of COLING 94*, Kyoto, 1994.
- V. Jijkoun, J. Mur, and M. de Rijke. Information extraction for question answering : Improving recall through syntactic patterns. In *COLING-2004*, 2004.
- K. Kipper, H. Trang Dang, and M. Palmer. Class based construction of a verb lexicon. In *Proceedings of AAAI-2000*, Austin TX, 2000.
- B. Levin. *English verb classes and alternations : a preliminary investigation*. Chicago University Press, 1993.
- C. Macleod, R. Grishman, and A. Meyers. Complex syntax : Building a computational lexicon. In *Proceedings of COLING '94*, pages 268–272, 1994.
- P. Saint-Dizier. Alternation and verb semantic classes for french : Analysis and class formation. In *Predicative forms in natural language and in lexical knowledge bases*. Kluwer Academic Publishers, 1999.