Extraction de règles de grammaire à partir de treebanks : développement d'un outil et premiers résultats

Santiago Herrera, Sylvain Kahane, Bruno Guillaume

Extraction de règles de grammaire à partir de treebanks : développement d'un outil et premiers résultats

Santiago Herrera[1] (en collaboration avec Sylvain Kahane[1] et Bruno Guillaume[2])
(1) MoDyCo, Université Paris Nanterre, France
(2) Loria, Nancy, France
santiago.herrerayanez@parisnanterre.fr, sylvain@kahane.fr,
bruno.guillaume@loria.fr

RÉSUMÉ
Ce travail présente une méthode et un outil d'extraction et d'exploration automatique de motifs statistiquement significatifs, de potentielles règles de grammaire, à partir de corpus arborés.

ABSTRACT
Grammar rules extraction from treebanks : a tool and some first results

This work presents a method and a tool for automatic extraction and exploration of statistically significant patterns and potential grammar rules from treebanks.

MOTS-CLÉS : Extraction de grammaire, règles de grammaire, treebank.

KEYWORDS: Grammar extraction, grammar rules, treebank.

1 Introduction

Construire la grammaire d'une langue et repérer l'ensemble de ses règles est une tâche aussi fondamentale pour l'étude de la langue et pour le développement d'autres ressources langagières que coûteuse. Plusieurs domaines, notamment la typologie linguistique (Dryer & Haspelmath, 2013), la linguistique formelle (Bender *et al.*, 2014; Howell *et al.*, 2017) et le traitement automatique des langues (Ponti *et al.*, 2019; Chaudhary *et al.*, 2020), cherchent depuis longtemps à systématiser les contraintes des langues, de façon plus ou moins automatique, afin de rendre compte de leurs propriétés structurelles. La plupart de ces approches s'appuient sur des corpus annotées à travers lesquels il est possible d'exploiter les propriétés statistiques du langage et se concentrent soit sur des règles générales, soit sur l'ensemble des structures possibles d'une langue. Pour notre part, nous cherchons à rendre compte des règles dans un espace local, en les considérant comme une série de contraintes qui déclenchent une sur-représentativité d'un motif donné spécifique dans un contexte précis.

Ce travail présente une méthode et une première version d'un outil d'extraction et d'exploration automatique de motifs statistiquement significatifs et de potentielles règles de grammaire, à partir de corpus annotés en syntaxe ou treebanks en dépendance. On vise plus précisément à l'extraction de règles interprétables et pondérables selon leurs propriétés quantitatives et statistiques. Inspirés par des travaux de la lexicométrie (Lafon, 1980) et de la linguistique du corpus (Evert, 2005; Pecina, 2010), on tire profit des tests et des mesures statistiques qui permettent le repérage de motifs, leur classement et la comparaison des motifs significatifs entre différents corpus. Ce système d'extraction

est implémenté à travers un outil [1] accessible qui combine ces méthodes statistiques avec l'utilisation de GREW-MATCH (Guillaume, 2021), afin d'interroger et d'explorer les arbres syntaxiques.

Le travail présenté ici est issu de Herrera (2022) et il a été réalisé dans le cadre du projet ANR Autogramm (Kahane, 2021), dont un des objectifs est de construire conjointement des treebanks et des grammaires pour des langues peu décrites. Notre outil cherche justement à faciliter le développement de grammaires basées sur (peu) des données avec le but d'aider le linguiste dans la description d'une langue.

2 Hypothèses

Une règle de grammaire est une contrainte d'une langue qui montre une régularité relative dans le système de cette langue. Dans un corpus, cette régularité se traduit par la répétition élevée et non aléatoire d'un motif, lequel est représenté dans une proportion inattendue par rapport à un sous-ensemble pertinent de motifs. Une règle explicite aussi les conditions qui, dans un contexte donné, déclenchent un motif particulier de façon statistiquement élevée.

Nous définissons donc une règle de grammaire à partir de trois éléments :

$$M \Rightarrow M(X) \,|\, C$$

(a) Définition générale d'une règle de grammaire.

$$M1 \Rightarrow M1\&M2 \,|\, M3$$

(b) Définition opérationnelle à partir de motifs d'une règle de grammaire.

FIGURE 1 – Formalisation d'une règle de grammaire.

Partant d'un motif M donné ou d'un espace de recherche spécifique, on cherche les conditions C qui favorisent de façon statistiquement significative les occurrences d'un phénomène linguistique X dans M. Autrement dit, on veut extraire et mettre en évidence quelles sont les variables ou motifs C qui montrent une dépendance statistique, positive en termes d'occurrences, avec X dans M. Cette formalisation s'inspire des règles de correspondance de la Théorie Sens-Texte (Mel'čuk, 1988), bien que l'on cherche à représenter des relations de dépendance entre variables et non pas une correspondance entre différents niveaux de représentation. Les règles sur lesquelles nous travaillons sont plus similaires aux règles de bonne formation ou de linéarisation. L'utilisation de la notion de contrainte trouve également un écho dans les travaux de Grammaires des Propriétés (Blache, 2001), où l'accent est mis sur les contraintes ou *propriétés* qui règlent la combinaison des unités et la formation des phrases. Néanmoins, nos règles sont motivées et limitées par l'information contenue dans le corpus.

Nous supposons que les conditions ou prédicteurs qu'on peut prendre en compte sont présents dans les contextes immédiats possibles du motif M considéré. Dans un arbre de dépendance, c'est la tête syntaxique d'une unité qui détermine dans la plupart de cas les propriétés linguistiques de cette unité. Mais toute autre relation, co-dépendance ou information linguistique encodée dans la phrase ou dans le treebank sont exploitables. Pour extraire de règles, nous élaborons une hypothèse de « localité » laquelle on suppose que les prédicteurs se trouvent dans le contexte immédiat du motif M. Notre espace de recherche (voir Figure 2) se limite donc aux différents traits des nœuds et des dépendances de M, ainsi que des nœuds directement connectés à M.

1. Voir https://github.com/santiagohy/grammar-rules-extraction

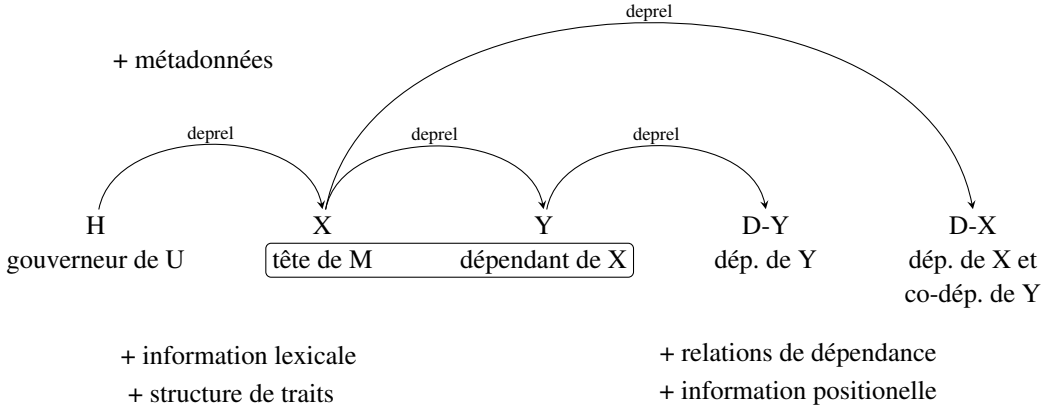

FIGURE 2 – Espace de recherche où X et Y forment, à titre d'exemple, le motif M. La structure de traits inclue les informations encodées dans les nœuds, notamment les traits morphologiques, les lemmes, etc.

3 Méthode d'extraction

Afin de rendre opérationnelle nos hypothèses, on peut formaliser l'extraction d'une règle à partir de trois motifs (Kahane, 2021; Kahane *et al.*, 2021). On cherche les motifs M3 qui, étant donné un motif M1, favorisent significativement les occurrences de M1&M2. Par exemple (voir Table 1), dans l'ensemble des noms modifiés par des adjectifs (M1), le fait que l'adjectif soit un numéral ordinal (M3) favorise significativement l'antéposition de l'adjectif (M1&M2) [2].

	Formalisation	**Exemple**
M1	Espace de recherche de départ	X->Y ; X[upos=NOUN] ; Y[upos=ADJ]
M2	Variable dépendante	Y << X
M3	Variable(s) explicative(s)	Y[NumType=Ord]

TABLE 1 – Formalisation de l'extraction avec l'exemple simple de l'antéposition de l'adjectif en français. On utilise le langage de requête GREW et les étiquettes morphosyntaxiques et syntaxiques UD/SUD. Dans l'exemple, X->Y indique une relation de dépendance orientée entre les nœuds X et Y, et Y << X signale la position relative entre ces deux nœuds dans la phrase, Y étant avant X. Les traits de chaque nœud, avec leurs valeurs, sont explicités entre crochets.

Pour déterminer si un motif M3 est statistiquement significatif, on calcule la probabilité d'obtenir la distribution observée, ou une distribution plus extrême, sous l'hypothèse nulle selon laquelle les motifs M2 et M3, dans le contexte M1, seraient indépendants. Si la probabilité est inférieure à la valeur critique fixée (p-value < 0.01), on rejette l'hypothèse nulle et on considère le motif significatif. Le logarithme de cette probabilité est utilisé comme valeur de significativité.

2. Pour un étude plus approfondie sur la position de l'adjectif et l'ordre de mots en français, voir Thuilier (2012).

Nous choisissons d'utiliser le test exact de Fisher pour calculer la significativité d'un motif à partir de ses occurrences car il nous permet de travailler avec des échantillons de petite taille et il est donc plus adapté au travail avec des langues pour lesquelles il y a peu de ressources (voir Lafon (1980) pour une utilisation de la méthode en lexicométrie). La tâche que nous réalisons se différencie des travaux connexes qui cherchent à évaluer la contribution de la combinaison linéaire d'un ensemble de variables binaires et numérique, dans la prédiction d'un phénomène linguistique (Bresnan *et al.*, 2007; Thuilier, 2012). Nous cherchons, en revanche, à extraire les meilleures prédicteurs, facilement interprétables, à partir d'un vaste ensemble composé, dans la plupart des cas, de variables catégorielles non binaires.

Il s'agit d'une méthode non-déterministe, où les motifs extraits peuvent être des motifs non disjoints. Nous privilégions donc le classement de ces motifs. C'est la raison pour laquelle nous utilisons aussi d'autres mesures complémentaires pour décrire et classer les motifs extraits, notamment la taille d'effet et les proportions des occurrences trouvées sur la totalité des motifs M1&M2 et sur la totalité des motifs M1&M3. La première de ces deux proportions peut s'interpréter comme la couverture de la règle sur le motif et le phénomène linguistique que l'on veut expliquer. La deuxième représente la précision de la règle ou combien de motifs M3 sont effectivement concernés par elle.

Motif	Significativité	OR	Proportion sur M1&M2	Proportion sur M1&M3
Y[NumType=Mod]	89	3.77	18.63%	92.22%

TABLE 2 – Résultat pour l'exemple sur le treebank SUD_French-Sequoia version 2.10.On prend le logarithme arrondi de la p-valeur comme valeur de la significativité. La valeur du OR est plus précisément le logarithme du rapport des chances ou, en anglais, *Odds Ratio*.

En reprenant l'exemple de la place de l'adjectif épithète, sur l'ensemble des traits de l'adjectif (nœud Y), que l'adjectif soit un numéral ordinal est très significatif du fait qu'il se place avant le nom modifié. On obtient donc un nombre d'adjectifs placés avant le nom très inattendue. Plus précisément, avoir obtenu 89 comme valeur de significativité signifie qu'il existe une probabilité très basse, d'autour de 10^{-89}, d'observer un nombre égale ou supérieur à la valeur observé de numéraux ordinaux placés avant leur nom, sous l'hypothèse d'indépendance. Ce motif significatif conforme, donc, une règle de linéarisation de l'adjectif par rapport au nom, qui exprime que les adjectifs dépendants d'un nom s'inversent très largement si ce sont des numéraux ordinaux. Cette règle explique que 18.63% des adjectifs antéposés sont des ordinaux et que plus de 92% des ordinaux se placent avant le nom qu'ils modifient.

4 Résultats et perspectives

La méthode nous permet d'obtenir les résultats que nous attendions. La plupart des règles construites sont des règles de grammaires (règles d'ordre, d'accord, de régime, etc.), bien qu'on extrait aussi des propriétés du corpus et des séries non pertinentes de motifs, d'un point de vue informatif. Ce dernier groupe de résultats est généralement constitué de sous-motifs d'autres motifs plus significatifs ou de motifs faisant partie d'une règle qui ne peut pas être expliquée par les informations linguistiques disponibles dans le corpus.

L'outil qui a été développé permet, d'autre part, d'interroger les règles d'un corpus arborés dans deux directions : de façon ascendante ou descendante. Dans le premier cas, il est possible de l'utiliser pour

explorer un treebank et pour découvrir les règles qui émergent d'un corpus précis, quelle que soit sa taille. Cela permet de mettre en parallèle l'annotation d'un corpus et la construction d'une grammaire, un des objectifs du projet ANR Autogramm, sachant que ces deux tâches peuvent mutuellement se compléter.

Dans le deuxième cas, rien n'empêche d'utiliser cet outil pour valider les règles d'une théorie linguistique précise de façon quantitative et statistique à partir d'un corpus donné. Ainsi, une règle théorique composée de conditions ou de contraintes, comme on peut en trouver dans la Théorie Sens-Texte ou dans le cadre des Grammaires de Propriétés, sera décrite selon sa significativité statistique dans le corpus et à partir des autres valeurs expliquées. Cela favorise le dialogue entre les théories linguistiques et les données empiriques, mettant dans la mesure du possible en relief la possible distance entre les deux et en hiérarchisant les règles selon leur importance dans le corpus. Le travail à partir de corpus de différente nature représente aussi l'occasion de valider ou de mettre à jour certaines affirmations théoriques en fonction du type ou de la variété de la langue étudiée.

L'extraction de règles, comme nous l'avons dit, dépend de l'information contenue dans l'espace de recherche et dans le corpus arborés. Nous travaillons pour l'instant avec les treebanks UD/SUD tels qu'ils se présentent, ce qui permet essentiellement d'extraire des règles syntaxiques de surface (ordre des mots et accord). Une prochaine étape consistera à travailler sur l'enrichissement automatique de treebanks, notamment sur le raffinage de certains traits pour obtenir des traits plus élémentaires et, nous espérons, plus informatifs.

Au moment de l'écriture de ce document, nous explorons d'autres méthodes statistiques et d'autres mesures d'association qui nous permettront d'extraire des motifs significatifs de façon plus efficace et d'avoir un outil plus performant. En parallèle au développement de l'outil, nous travaillions sur des méthodes d'évaluation quantitatives, notamment l'évaluation de la couverture de l'ensemble des règles extraites sur un treebank.

Remerciements

Les auteurs remercient les relecteurs pour leurs commentaires. Ce travail a été réalisé dans le cadre du projet ANR Autogramm (ANR-21-CE38-0017).

Références

BENDER E. M., CROWGEY J., GOODMAN M. W. & XIA F. (2014). Learning grammar specifications from IGT : A case study of chintang. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, p. 43–53, Baltimore, Maryland, USA : Association for Computational Linguistics. DOI : 10.3115/v1/W14-2206.

BLACHE P. (2001). *Les grammaires de propriétés : Des contraintes pour le Traitement Automatique des Langues Naturelles*. Hermés.

BRESNAN J., CUENI A., NIKITINA T. & BAAYEN H. (2007). Predicting the dative alternation. *Cognitive Foundations of Interpretation*, p. 69–94.

CHAUDHARY A., ANASTASOPOULOS A., PRATAPA A., MORTENSEN D. R., SHEIKH Z., TSVETKOV Y. & NEUBIG G. (2020). Automatic extraction of rules governing morphological agree-

ment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 5212–5236, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.422.

DRYER M. S. & HASPELMATH M., Éds. (2013). *WALS Online*. Leipzig : Max Planck Institute for Evolutionary Anthropology.

EVERT S. (2005). *The Statistics of Word Cooccurrences : Word Pairs and Collocations*. Thèse de doctorat, University of Stuttgart.

GUILLAUME B. (2021). Graph matching and graph rewriting : GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : System Demonstrations*, p. 168–175, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-demos.21.

HERRERA S. (2022). Extraction automatique de règles de grammaire à partir de treebanks. Mémoire de master, Master pluriTAL, Université Sorbonne Nouvelle, Université Paris Nanterre et l'INALCO.

HOWELL K., BENDER E. M., LOCKWOOD M., XIA F. & ZAMARAEVA O. (2017). Inferring case systems from IGT : Enriching the enrichment. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, p. 67–75, Honolulu : Association for Computational Linguistics. DOI : 10.18653/v1/W17-0110.

KAHANE S. (2021). Autogramm : Induction of descriptive grammars from annotated corpora. *Soumission à l'Agence National de la Recherche*.

KAHANE S., GUILLAUME B., GERDES K., CARON B. & LOISEAU S. (2021). Le projet ANR Autogramm et l'extraction automatique de grammaires. llustration par la négation. In *3e Journées GdR LIFT, Linguistique Informatique Formelle et de Terrain*, Grenoble, France.

LAFON P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, **1**(1), 127–165. DOI : 10.3406/mots.1980.1008.

MEL'ČUK I. (1988). *Dependency Syntax : Theory and Practice*. State University of New York Press.

PECINA P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, **44**(1/2), 137–158.

PONTI E. M., O'HORAN H., BERZAK Y., VULIĆ I., REICHART R., POIBEAU T., SHUTOVA E. & KORHONEN A. (2019). Modeling language variation and universals : A survey on typological linguistics for natural language processing. *Computational Linguistics*, **45**(3), 559–601. DOI : 10.1162/coli_a_00357.

THUILIER J. (2012). *Contraintes préférentielles et ordre des mots en français*. Thèses, Université Paris-Diderot - Paris VII.