Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Autogramm : développement simultané de treebanks et de grammaires à partir de corpus

Sylvain Kahane[1] Santiago Herrera[1] Bruno Guillaume[2] Kim Gerdes[3]
(1) Modyco, Université Paris Nanterre, CNRS
(2) Sémagramme, Loria, Inria Nancy - Grand Est, Université de Lorraine
(3) LISN, Université Paris-Saclay, CNRS
sylvain@kahane.fr, s.herrera@parisnanterre.fr,
bruno.guillaume@inria.fr, kim@gerdes

RÉSUMÉ

Ce projet de recherche vise à créer de nouveaux treebanks en dépendance pour des langues sous-dotées, en unifiant autant que possible leur développement avec celui de grammaires descriptives quantitatives. Nous présenterons notre chaîne de traitement et de développement de treebanks et nous discuterons du type de grammaire que nous voulons extraire. Enfin, nous examinerons l'utilisation de ces ressources en typologie quantitative.

ABSTRACT

Autogramm : Simultaneous development of treebanks and corpus-driven grammars

This research project aims to create new dependency treebanks for low-resource languages, unifying as far as possible their development with that of quantitative descriptive grammars. We will present our processing pipeline and discuss the type of grammar we want to extract. Finally, we will examine the use of these resources in quantitative typology.

MOTS-CLÉS : Autogramm, treebanks, extraction de grammaires, typologie quantitative.

KEYWORDS: Autogramm, treebanks, grammar extraction, quantitative typology.

1 Introduction

Les études linguistiques comparatives nécessitent des corpus de haute qualité qui représentent au mieux la variation et la diversité des langues du monde, avec des annotations suffisamment riches pour en extraire des descriptions grammaticales, et suffisamment comparables pour permettre des études contrastives et typologiques.

Le projet de recherche *Autogramm* [1] ici présenté vise à répondre à ces besoins, du moins en partie, en créant de nouveaux treebanks syntaxiques pour plus de 15 langues sous-dotées et en unifiant autant que possible leur développement avec celui de grammaires descriptives quantitatives. La production de corpus annotés et de grammaires descriptives est certainement complémentaire et leur développement simultané pourrait permettre de réduire le temps de travail et d'améliorer la qualité des

1. Autogramm est un projet financé par l'Agence Nationale de la Recherche (ANR-21-CE38-0017). Pour accéder à la liste complète des participants, y compris les linguistes et les langues sur lesquelles ils travaillent, ainsi que les outils développés, voir https://autogramm.github.io/.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 6 : projets, pages 37–42.

deux ressources. De plus, les grammaires basées sur des corpus encodent facilement des informations quantitatives, permettant, par exemple, la hiérarchisation des observations grammaticales et leur comparaison.

Plus précisément, nous présenterons notre chaîne de développement de treebanks en dépendances, en utilisant les schémas d'annotation *Universal Dependency* (UD) (Nivre *et al.*, 2020; de Marneffe *et al.*, 2021) et *Surface Syntactic* UD (SUD) (Gerdes *et al.*, 2018, 2019a). Ensuite, nous discuterons du type de grammaire que nous voulons extraire à partir de treebanks et nous présenterons les prochaines étapes de notre travail qui tend vers les études comparatives et vers une typologie quantitative.

2 Chaîne de traitement et nouvelles ressources

Le projet rassemble une équipe diversifiée, comprenant des linguistes de terrain spécialisés en langues peu décrites et des experts en annotation de corpus. Une chaîne de traitement a été mise en place pour que chacun puisse contribuer au développement de ces ressources pour chacune des langues étudiées.

En général, le processus commence par la transformation des gloses interlinéaires (IGT), souvent utilisées par les linguistes de terrain, en un pre-treebank, sans perdre les informations qu'elles contiennent (la segmentation, les traits morpho-syntaxiques, les gloses, etc.). Cela implique un travail avec le linguiste qui consiste en sélectionner et normaliser les informations pertinentes. L'annotation syntaxique peut alors se faire au niveau des mots ou des morphes (e.g. Kahane *et al.*, 2021). Nous utilisons l'outil d'annotation en ligne `ArboratorGrew`, qui propose un système de bootstrapping syntaxique (Guibon *et al.*, 2020; Peng *et al.*, 2022) : l'analyseur syntaxique peut être entraîné avec le travail déjà effectué pour annoter automatiquement le reste du corpus, autant de fois que nécessaire. En parallèle, nous construisons des grammaires pour chacune des langues (voir section 3).

Plusieurs treebanks sont actuellement en cours de développement pour les langues suivantes : Amdo Tibetan (sino-tibetain), dialectes arabes (marocain, égyptien, tunisien ; sémitique), bambara (mandingue), breton (celte), gbaya (oubanguien), haïtien (créole), hausa (chadique), salar (turc), sungwadia (austronésien), tuwari (papoue), vietnamien (austroasiatique), yali (papoue), ye'kwana (caribe), etc. Un treebank pour le Beja (Kahane *et al.*, 2021) et un pour le Zaar (Caron, 2015) ont déjà été développées en utilisant une approche similaire et publiées dans la base de données d'UD.

3 Extraction de grammaires à partir de corpus

Il existe un grand nombre des travaux utilisant différents formalismes et différentes stratégies pour extraire de la manière la plus automatique possible les grammaires et les propriétés typologiques des corpus annotés. La plupart des méthodes produisent des grammaires formelles à partir de ressources linguistiques, telles que les IGT, en utilisant des connaissances grammaticales externes et élaborées à la main (voir Bender *et al.*, 2002; Zamaraeva *et al.*, 2022; Howell & Bender, 2022). Ces grammaires ne contiennent généralement pas d'informations quantitatives, bien que le fait de disposer de telles données permet d'obtenir une description fine de la langue étudiée et de classer les descriptions extraites en fonction de leur importance au sein d'un corpus. D'autres systèmes d'extraction parviennent à encoder des informations quantitatives (e.g. Blache *et al.*, 2016), mais le nombre de règles extraites restent élevé et la forme des règles est limitée. En outre, certaines propriétés

ne sont encodées qu'au niveau de leurs constructions. Par exemple, ces grammaires indiqueront si chaque construction a une tête en position finale, mais pas si la langue étudiée est une langue à tête finale.

Notre objectif est d'extraire des descriptions grammaticales facilement interprétables par n'importe quel utilisateur et, puisque la tâche est réalisée à partir de données annotées, nous cherchons à associer à chacune d'entre elles des informations quantitatives. Cependant, contrairement aux grammaires analysées, nous visons à classer les observations grammaticales en fonction de leur importance dans un corpus et à obtenir des grammaires de différentes tailles en fonction de la manière dont les règles extraites sont classées et regroupées. Les données quantitatives peuvent également être utilisées pour mettre en évidence les relations qui existent entre les différentes propriétés afin d'expliquer certains phénomènes et d'en découvrir d'autres qui seraient autrement passés inaperçus (voir Bresnan *et al.* (2007) dans une étude classique sur l'alternance dative ; plus récemment, Chaudhary *et al.* (2020) et Chaudhary (2022) ont extrait des règles d'accord, d'ordre, ainsi que des cas à l'aide de treebanks). Dans ce but, nous nous concentrons sur la fréquence des phénomènes observés et sur d'autres mesures continues (Levshina, 2019 et Gerdes *et al.*, 2019b).

Nous travaillons actuellement sur différents systèmes d'extraction de grammaires. Dans le cadre du projet, nous avons déjà développé une première méthode qui permet d'extraire avec succès des motifs grammaticaux à partir de treebanks et de les classer en fonction de leur importance statistique dans le corpus (Herrera *et al.*, 2022). Plus précisément, nous calculons la probabilité d'obtenir la distribution observée de certains motifs apparentés à partir d'une hypothèse d'indépendance. Plus cette probabilité est élevée, plus le motif est significatif.

Enfin, en interagissant avec les descriptions et les grammaires extraites, le linguiste travaillant sur une langue peu décrites, sera en mesure de vérifier si les caractéristiques (spécifiques de la langue) choisies pour annoter le corpus représentent bien la grammaire de la langue en question.

4 Typologie quantitative

Les descriptions grammaticales, du moins en ce qui concerne les propriétés universelles, sont exprimées à l'aide du même jeu d'étiquettes et du même formalisme en dépendance. Cela signifie que les grammaires que nous extrayons sont des grammaires comparables qui permettent des analyses comparatives d'une même observation entre différentes langues et différents corpus. De cette façon, nous pouvons déterminer précisément ce qui est particulier à une langue par rapport à d'autres, sans limiter l'étude typologique à une liste préétablie d'observations, qui peuvent pour certaines (familles de) langues se révéler impertinentes.

Lorsque l'on travaille avec des informations quantitatives, on peut également comparer des observations en termes de valeurs continues plutôt que de valeurs discrètes. Contrairement aux bases de données importantes et fondamentales (WALS (Dryer & Haspelmath, 2013), APiCS (Michaelis *et al.*, 2013), ValPal (Hartmann *et al.*, 2013), entre autres), il sera possible d'expliciter dans quelle mesure une caractéristique typologique est présente dans un corpu spécifique et dans quelle mesure elle diffère de celle d'autres langues. Ce faisant, nous travaillons dans le cadre de la typologie quantitative (cf. Cysouw, 2005), en suivant une nouvelle perspective d'étude encore à explorer (Futrell *et al.*, 2015; Gerdes *et al.*, 2021).

Nous explorons des méthodes d'échantillonnage et de comparaison pour trouver des similitudes et

des différences entre les corpus en utilisant les observations extraites, tout en cherchant de nouvelles métriques autres que la fréquence. Nous construirons une base de données typologique contenant les observations quantitatives collectées. Il est à noter que ces méthodes pourraient également être utilisées pour détecter d'autres variations dans la langue, telles que les variations sociolinguistiques et diachroniques.

5 Obstacles et perspectives

Les différents objectifs du projet se heurtent évidement à plusieurs obstacles, dont les suivants : l'explosion combinatoire des variables lorsque l'on tente d'extraire des modèles grammaticaux des banques d'arbres, des résultats incongrus dus à différentes interprétations du schéma d'annotation et aux particularités du corpus, et des échantillons linguistiques déséquilibrés qui empêchent les études typologiques cohérentes. Une partie du projet consiste à étudier ces problèmes afin de contribuer à la linguistique théorique et à la documentation linguistique.

Références

BENDER E. M., FLICKINGER D. & OEPEN S. (2002). The grammar matrix : An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *COLING-02 : Grammar Engineering and Evaluation*.

BLACHE P., RAUZY S. & MONTCHEUIL G. (2016). MarsaGram : an excursion in the forests of parsing trees. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 2336–2342, Portorož, Slovenia : European Language Resources Association (ELRA).

BRESNAN J., CUENI A., NIKITINA T. & BAAYEN R. (2007). *Predicting the Dative Alternation*, In *Cognitive foundations of interpretation*, p. 69–94. KNAW : Amsterdam.

CARON B. (2015). Zaar grammatical sketch. In ASDT, Éd., *Corpus-based Studies of Lesser-described Languages. The CorpAfroAs corpus of spoken AfroAsiatic languages*. Amsterdam-Philadelphia : John Benjamins.

CHAUDHARY A. (2022). *Automatic Extraction and Application of Language Descriptions for Under-Resourced Languages*. Thèse de doctorat, Carnegie Mellon University. DOI : 10.1184/R1/21708035.v1.

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

CYSOUW M. (2005). Quantitative methods in typology (quantitative methoden in der typologie). In R. KÖHLER, G. ALTMANN & R. G. PIOTROWSKI, Éds., *Quantitative Linguistik / Quantitative Linguistics - Ein internationales Handbuch / An International Handbook*, p. 554–557. DeGruyter.

DE MARNEFFE M.-C., MANNING C. D., NIVRE J. & ZEMAN D. (2021). Universal Dependencies. *Computational Linguistics*, **47**(2), 255–308. DOI : 10.1162/coli_a_00402.

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

FUTRELL R., MAHOWALD K. & GIBSON E. (2015). Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, p. 91–100, Uppsala, Sweden : Uppsala University, Uppsala, Sweden.

GERDES K., GUILLAUME B., KAHANE S. & PERRIER G. (2018). SUD or surface-syntactic Universal Dependencies : An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, p. 66–74, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-6008.

GERDES K., GUILLAUME B., KAHANE S. & PERRIER G. (2019a). Improving surface-syntactic Universal Dependencies (SUD) : MWEs and deep syntactic features. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, p. 126–132, Paris, France : Association for Computational Linguistics. DOI : 10.18653/v1/W19-7814.

GERDES K., KAHANE S. & CHEN X. (2019b). Rediscovering greenberg's word order universals in UD. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, p. 124–131, Paris, France : Association for Computational Linguistics. DOI : 10.18653/v1/W19-8015.

GERDES K., KAHANE S. & CHEN X. (2021). Typometrics : From implicational to quantitative universals in word order typology. *Glossa : a journal of general linguistics*, **6**(1). DOI : 10.5334/gjgl.764.

GUIBON G., COURTIN M., GERDES K. & GUILLAUME B. (2020). When collaborative treebank curation meets graph grammars. In *Proceedings of The 12th Language Resources and Evaluation Conference*, p. 5293–5302, Marseille, France : European Language Resources Association.

HARTMANN I., HASPELMATH M. & TAYLOR B., Éds. (2013). *The Valency Patterns Leipzig online database*. Leipzig : Max Planck Institute for Evolutionary Anthropology.

HERRERA S., KAHANE S. & GUILLAUME B. (2022). Extraction de règles de grammaire à partir de treebanks : développement d'un outil et premiers résultats. In L. BECERRA, B. FAVRE, C. GARDENT & Y. PARMENTIER, Éds., *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, p. 93–98, Marseille, France : CNRS. HAL : hal-03846825.

HOWELL K. & BENDER E. (2022). Building analyses from syntactic inference in local languages : An hpsg grammar inference system. *Northern European Journal of Language Technology*, **8**. DOI : 10.3384/nejlt.2000-1533.2022.4017.

KAHANE S., VANHOVE M., ZIANE R. & GUILLAUME B. (2021). A morph-based and a word-based treebank for Beja. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, p. 48–60, Sofia, Bulgaria : Association for Computational Linguistics.

LEVSHINA N. (2019). Token-based typology and word order entropy : A study based on universal dependencies. *Linguistic Typology*, **23**(3), 533–572. DOI : doi :10.1515/lingty-2019-0025.

MICHAELIS S. M., MAURER P., HASPELMATH M. & HUBER M., Éds. (2013). *APiCS Online*. Leipzig : Max Planck Institute for Evolutionary Anthropology.

NIVRE J., DE MARNEFFE M.-C., GINTER F., HAJIČ J., MANNING C. D., PYYSALO S., SCHUSTER S., TYERS F. & ZEMAN D. (2020). Universal Dependencies v2 : An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 4034–4043, Marseille, France : European Language Resources Association.

PENG Z., GERDES K. & GUILLER K. (2022). Pull your treebank up by its own bootstraps. In L. BECERRA, B. FAVRE, C. GARDENT & Y. PARMENTIER, Éds., *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, p. 139–153, Marseille, France : CNRS. HAL : hal-03846834.

ZAMARAEVA O., CURTIS C., EMERSON G., FOKKENS A., GOODMAN M., HOWELL K., TRIMBLE T. & BENDER E. M. (2022). 20 years of the grammar matrix : cross-linguistic hypothesis testing of increasingly complex interactions. *Journal of Language Modelling*, **10**(1), 49–137. DOI : 10.15398/jlm.v10i1.292.