Autogramm: Simultaneous development of treebanks and corpus-driven grammars

Santiago Herrera[†] **Kim Gerdes**[*] **Bruno Guillaume**[‡] **Sylvain Kahane**[†]
[†]Modyco, Paris Nanterre University, CNRS
[*]LISN, Paris-Saclay University, CNRS
[‡]Sémagramme, Loria, Inria Nancy - Grand East, University of Lorraine, CNRS
s.herrera@parisnanterre.fr, kim@gerdes.fr
bruno.guillaume@inria.fr, sylvain@kahane.fr

Relevant UniDive working groups: WG1, WG4

1 Introduction

Cross-linguistic studies require high-quality corpora that best represent the variation and diversity of the world's languages, with annotations rich enough to extract descriptive grammars from them, and sufficiently comparable to allow contrastive and typological studies.

This paper presents part of the ongoing research project Autogramm, which aims to address these problems, at least in part, by creating new treebanks for low-resource languages and by unifying as much as possible the development of syntactic treebanks and descriptive grammars.[1]

Usually, the process of developing a descriptive grammar has been seen as subsequent to the process of building annotated corpora. But, on the one hand, the task of linguistic annotation presupposes some, probably unsystematized, knowledge of the grammar of the language in which one is working. On the other hand, the annotation process, the choice of labels to be used, and the data itself force one to re-examine some of this grammatical knowledge. From this perspective, the production of annotated corpora and descriptive grammars are complementary and their simultaneous development could help to reduce working time and improve the quality of both resources. Moreover, corpus-driven grammars can easily incorporate quantitative information, allowing, for example, the hierarchization of grammatical observations and their comparison across languages and corpora.

In the next sections, we will introduce our processing pipeline and the languages on which we intended to create dependency treebanks, following the universal schema of Universal Dependencies (UD) (Nivre et al., 2016, 2020; de Marneffe et al., 2021) and Surface Syntactic UD (SUD) (Gerdes et al., 2018, 2019a). We will then discuss the type grammar we want to extract from treebanks and present our next steps towards cross-linguistic studies.

2 Processing pipeline and new resources

The project brings together a heterogeneous team to develop treebanks and grammars, including field linguists with expertise in poorly described languages and specialists in annotated corpora. A processing pipeline has been set up to deal with the difficult task of developing these resources for each of the languages under study (see sketch in Figure 1).

In general, the process starts by transforming interlinear glosses (IGTs), often used by field linguists, into a pre-treebank, without losing the information they contain (segmentation, morphosyntactic features, glosses, etc.). This involves working with the linguist to select and normalise the relevant information. The syntactic annotation could be then done at the level of words or morphs (e.g. Kahane et al., 2021). We use the available online annotation tool ArboratorGrew, which offers a system of syntactic bootstrapping (Guibon et al., 2020; Peng et al., 2022): the parser can be trained with the work already done to automatically annotate the rest of the corpus, as many times as needed. In parallel, we build grammars for each of the languages (see section 3).

Treebanks are currently being developed for the following languages: Amdo Tibetan (Sino-Tibetan), Arabic dialects (Moroccan, Egyptian, Tunisian; Semitic), Bambara (Manding), Breton (Indo-European), Gbaya (Ubanguian), Haitian (Creole), Hausa (Chadic), Salar (Turkic), Sungwadia (Austronesian), Tuwari (Papua), Vietnamese (Austrasiatic), Yali (Papua), Ye'kwana (Carib), etc.

[1] Autogramm is funded by the *Agence nationale de la recherche* (ANR-21-CE38-0017). The complete list of participants, including the linguists and the languages they are working on, as well as the tools being developed, can be found in https://autogramm.github.io/.

A treebank for Beja (Kahane et al., 2021) and one for Zaar (Caron, 2015) have already been developed using a similar approach and published in the UD database.

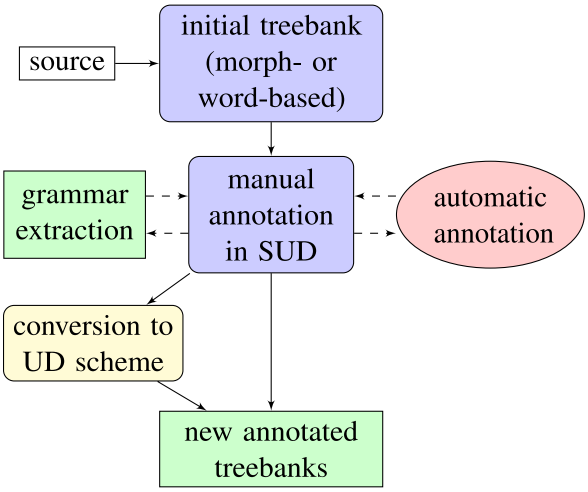

Figure 1: Sketch of the processing pipeline. The sources used to create the initial or pre-treebank are usually IGTs, but could be of any other type.

The SUD scheme was chosen as the annotation scheme because it has the same flexibility as the UD scheme, but it is more suitable for extracting surface morpho-syntactic rules, especially word order regularities. Since it can also be converted to the UD format, we ended up with two different treebanks, each with its own strengths.

3 Extraction of corpus-driven grammars

There is a large body of work that uses different linguistic formalisms and combine different strategies to infer or extract grammars and typological properties from annotated corpora as automatically as possible.

Most of the methods produce large formal grammars from linguistic resources, such as IGTs, using external and hand-engineered grammatical knowledge (e.g. Bender et al., 2002; Zamaraeva et al., 2022; Howell and Bender, 2022). These grammars usually do not contain quantitative information, although having such data allows one to have a fine-grained description of the language under study and to rank the extracted descriptions according to their importance within a corpus. Other extraction systems successfully encode quantitative information (e.g., Blache et al., 2016), but the number of extracted rules is still high and the form of the rules is limited. In addition, certain properties are only encoded at the level of the structures. For example, these grammars will indicate whether each construction has a head in the final position, but not whether a language is a head-final language.

As mentioned previously, our goal is also to extract grammar descriptions from treebanks for a variety of languages. These descriptions must be easily interpretable by any linguist or user and, since the task is done using annotated data, each one of them must be supported by quantitative information. However, in contrast to the analysed grammars, we aim to rank the grammatical observations according to their relevance in a corpus and to be able to obtain grammars of different sizes depending on the way the extracted rules are ranked and clustered. The quantitative data can be also used to compute the interactions between linguistic features in order to explain some phenomena and discover others that would otherwise go unnoticed (see Bresnan et al. (2007) for classic study on dative alternation; more recently, Chaudhary et al. (2020) and Chaudhary (2022) extract agreement, word order and case marking non-ranked rules using SUD treebanks). For this purpose, we focus on the frequency of the observed phenomena and on other continuous metrics and features (e.g. Levshina, 2019 and Gerdes et al., 2019b).

We are currently working on different grammar extraction systems. In the context of the project, we have already developed a first method that successfully extracts grammatical patterns from treebanks and ranks them according to their statistical significance in the corpus (Herrera et al., 2022). More precisely, we compute the probability of getting the observed distribution of some related patterns under an independence hypothesis. The more extreme this probability, the more significant the pattern.

Finally, by interacting with the extracted descriptions and grammars, the linguist working on the language will be able to check whether the (language-specific) features chosen to annotate the corpus are a good representation of the grammar of the language.[2] At the same time, we could also uncover previously unknown regularities in the corpus to support the annotation.

[2] See regular discussions on the UD GitHub repository about adding new universal features to enrich the annotation (e.g. Add features for interrogative, exclamative and comparative clauses? #877)

4 Quantitative and inductive typology

The grammar descriptions, at least as far as the universal features are concerned, are expressed using the same tagset and dependency formalism. This means that the grammars we extract are comparable grammars that allow cross-linguistic comparisons of the same observation between different languages and corpora. In this way, we can determine precisely what is peculiar about a language in comparison with other languages, and not limit the typological study to a pre-established list of observations, which may not be very relevant for certain (families of) languages.

When working with quantitative information, we can also compare observations in terms of continuous values rather than discrete values. Unlike important and fundamental databases (WALS (Dryer and Haspelmath, 2013), APiCS (Michaelis et al., 2013), ValPal (Hartmann et al., 2013), among others), it will be possible to make explicit the extent to which a typological feature occurs in specific corpora and the degree to which it differs from other languages. In doing so, we work within the framework of quantitative typology (cf. Cysouw, 2005), following a new perspective that is still little explored (Futrell et al., 2015; Gerdes et al., 2021).

More precisely, we are exploring sampling and comparison methods to find similarities and differences between corpora using the extracted observations, while looking for new metrics other than frequency. We will build a typological database containing the collected quantitative observations. It is noted that these methods could also be used to detect other variations in language, such as sociolinguistic and diachronic variations.

5 Obstacles and perspectives

The various aims of the project undoubtedly face several obstacles, among which we note: the combinatorial explosion of variables when trying to extract grammatical patterns from treebanks, incongruent results due to different interpretations of the annotation scheme and due to corpus peculiarities, and unbalanced language samples that prevent consistent typological studies. Part of the project is to investigate these problems in order to contribute to theoretical linguistics and language documentation.

References

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 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*.

Philippe Blache, Stéphane Rauzy, and Grégoire Montcheuil. 2016. MarsaGram: an excursion in the forests of parsing trees. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2336–2342, Portorož, Slovenia. European Language Resources Association (ELRA).

J. Bresnan, A. Cueni, T. Nikitina, and R.H. Baayen. 2007. *Predicting the Dative Alternation*, page 69–94. KNAW, Amsterdam.

Bernard Caron. 2015. Zaar grammatical sketch. In asdt, editor, *Corpus-based Studies of Lesser-described Languages. The CorpAfroAs corpus of spoken AfroAsiatic languages*. John Benjamins, Amsterdam-Philadelphia.

Aditi Chaudhary. 2022. *Automatic Extraction and Application of Language Descriptions for Under-Resourced Languages*. Ph.D. thesis, Carnegie Mellon University.

Aditi Chaudhary, Antonios Anastasopoulos, Adithya Pratapa, David R. Mortensen, Zaid Sheikh, Yulia Tsvetkov, and Graham Neubig. 2020. Automatic extraction of rules governing morphological agreement. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5212–5236, Online. Association for Computational Linguistics.

Michael Cysouw. 2005. Quantitative methods in typology (quantitative methoden in der typologie). In Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski, editors, *Quantitative Linguistik / Quantitative Linguistics - Ein internationales Handbuch / An International Handbook*, pages 554–557. De-Gruyter.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 91–100, Uppsala, Sweden. Uppsala University, Uppsala, Sweden.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 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)*, pages 66–74, Brussels, Belgium. Association for Computational Linguistics.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 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)*, pages 126–132, Paris, France. Association for Computational Linguistics.

Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2019b. Rediscovering greenberg's word order universals in UD. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 124–131, Paris, France. Association for Computational Linguistics.

Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2021. Typometrics: From implicational to quantitative universals in word order typology. *Glossa: a journal of general linguistics*, 6(1).

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5293–5302, Marseille, France. European Language Resources Association.

Iren Hartmann, Martin Haspelmath, and Bradley Taylor, editors. 2013. *The Valency Patterns Leipzig online database*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Santiago Herrera, Sylvain Kahane, and Bruno Guillaume. 2022. Extraction de règles de grammaire à partir de treebanks : développement d'un outil et premiers résultats. In *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, pages 93–98, Marseille, France. CNRS.

Kristen Howell and Emily Bender. 2022. Building analyses from syntactic inference in local languages: An hpsg grammar inference system. *Northern European Journal of Language Technology*, 8.

Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 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)*, pages 48–60, Sofia, Bulgaria. Association for Computational Linguistics.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*, 23(3):533–572.

Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors. 2013. *APiCS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Ziqian Peng, Kim Gerdes, and Kirian Guiller. 2022. Pull your treebank up by its own bootstraps. In *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, pages 139–153, Marseille, France. CNRS.

Olga Zamaraeva, Chris Curtis, Guy Emerson, Antske Fokkens, Michael Goodman, Kristen Howell, T.J. Trimble, and Emily M. Bender. 2022. 20 years of the grammar matrix: cross-linguistic hypothesis testing of increasingly complex interactions. *Journal of Language Modelling*, 10(1):49–137.