Building Quantitative Contrastive Grammars from Syntactic Treebanks

Santiago Herrera[1], Ioana-Madalina Silai[1], Caio Corro[2], Bruno Guillaume[3], Sylvain Kahane [1,4]

[1]*Modyco, Université Paris Nanterre, CNRS*
[2]*Sorbonne Université, CNRS, ISIR*
[3]*LORIA, Inria, Université de Lorraine*
[4]*Institut Universitaire de France*

Keywords: contrastive grammar, quantitative grammar, grammar extraction, sparse logistic regression

Our goal is to develop corpus-driven contrastive grammars, that is, to identify the most salient patterns of a given language with respect to another language or to a set of languages, and more generally, for a given corpus with respect to another corpus (for a survey in contrastive studies, see Hasselgård, 2020). To achieve this, we use treebanks, and we compare them to induce their specific characteristics. We obtain a set of linguistic rules or tendencies ordered by importance.

Growing interest in quantitative linguistics has led to the adoption of continuous features inferred from corpora (cf. Levshina, 2019; Baylor et al., 2024) and the extraction of fine-grained patterns directly from linguistic data to answer complex linguistic questions (Chaudhary et al., 2020, 2022; Herrera et al., 2024). For example, corpora are used not only to assess whether the dominant order for a language is subject-verb (SV), but also to quantitatively capture the cases where the order is inverted. Extracting quantitative patterns and rules is particularly important in the case of comparison between language varieties, where most differences are due to variation in frequency (Blanche-Benveniste, 1997).

We built two different experimental scenarios to capture inter-corpus variation:

- Comparison between a written French corpus (GSD: Guillaume et al., 2019) and spoken French corpus (ParisStories: Kahane et al., 2021).
- Comparison between three Romance language: French (GSD), Spanish (AnCora: Taulé et al., 2008) and Romanian (RTT: Barbu Mititelu, 2018).

For each scenario, we use gold treebanks (manually annotated or corrected) to ensure the linguistic quality of our results. We employ sample techniques to ensure the comparability of the treebanks. Whenever possible, we harmonize the annotations between the treebanks.

We rely on the Universal Dependencies (UD) treebank collection (de Marneffe et al., 2021), conceived to maximize language comparability. Some of the treebanks used are from the Surface Syntactic UD (SUD) (Gerdes et al., 2018, 2022) collection. We use the converted UD versions of them which are notably uniform due to the conversion.

We follow the work of Herrera et al. (2024), inspired in turn by previous works (Chaudhary et al., 2020, 2022) which proposes a general formalization of a corpus-driven syntactic rule and

uses sparse logistic regressions to extract and rank quantitative syntactic and morpho-syntactic patterns from treebanks. In these works, rule extraction is defined as a classification task, where the linguistic phenomenon of interest (e.g. agreement, word order, case marking) is the target to predict.

We adapt the latter approach to extract the most salient syntactic patterns characterizing a treebank among a set of treebanks. Given two patterns S (scope) and Q (question), we look for patterns P (predictors) that trigger Q better than S does. For instance, if we want to know what characterizes the French verbs with respect to Romanian verbs, S will be a pattern selecting all verbs of the sample and Q the language/treebank we want to predict. We look for contexts P that favor one language/treebank over the other (such as French verbs having more subjects because Romanian is a pro-drop language).

As a result, we obtain (1) patterns that are unique to one treebank and (2) patterns that are present in both treebanks, but which are more salient in one of them. The extracted patterns are of linguistic interest and capture linguistic variation between the two treebanks. They also reveal problems due to annotation variation, providing a way to capture the difference in annotation.

References

Barbu Mititelu, V. (2018). Modern Syntactic Analysis of Romanian, pages 67–78. Editura Universității Alexandru Ioan Cuza, Iaşi.

Baylor, E., Ploeger, E., and Bjerva, J. (2024). Multilingual gradient word-order typology from Universal Dependencies. In Graham, Y. and Purver, M., editors, Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pages 42–49, St. Julian's, Malta. Association for Computational Linguistics.

Blanche-Benveniste, C. (1997). La notion de variation syntaxique dans la langue parlée. Langue française, 115(1):19–29. Included in a thematic issue: La variation en syntaxe.

Chaudhary, A., Anastasopoulos, A., Pratapa, A., Mortensen, D. R., Sheikh, Z., Tsvetkov, Y., and Neubig, G. (2020). Automatic extraction of rules governing morphological agreement. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5212–5236, Online. Association for Computational Linguistics.

Chaudhary, A., Sheikh, Z., Mortensen, D. R., Anastasopoulos, A., and Neubig, G. (2022). Autolex: An automatic framework for linguistic exploration.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. Computational Linguistics, 47(2):255–308.

Gerdes, K., Guillaume, B., Kahane, S., and Perrier, G. (2018). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In Lynn, T. and Schuster, S., editors, Universal Dependencies Workshop 2018, Brussels, Belgium.

Gerdes, K., Guillaume, B., Kahane, S., and Perrier, G. (2022). Starting a new treebank? Go SUD! Theoretical and practical benefits of the Surface-Syntactic distributional approach. In SyntaxFest Depling 2021 - 6th International Conference on Dependency Linguistics, pages 35–46, Sofia, Bulgaria. Association for Computational Linguistics.

Guillaume, B., de Marneffe, M.-C., and Perrier, G. (2019). Conversion et améliorations de corpus du français annotés en Universal Dependencies. Revue TAL : traitement automatique des langues, 60(2):71–95.

Hasselgård, H. (2020). Corpus-based contrastive studies: Beginnings, developments, and directions. Languages in Contrast, 20(2):184–208.

Herrera, S., Corro, C., and Kahane, S. (2024). Sparse logistic regression with high-order features for automatic grammar rule extraction from treebanks. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 15114–15125, Torino, Italia. ELRA and ICCL.

Kahane, S., Caron, B., Strickland, E., and Gerdes, K. (2021). Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. In Dakota, D., Evang, K., and Kübler, S., editors, Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021), pages 35–47, 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.

Taulé, M., Martí, M. A., and Recasens, M. (2008). AnCora: Multilevel annotated corpora for Catalan and Spanish. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).