# Uncovering Typological Context-Sensitive Features

**Chiara Alzetta**[•◇], **Felice Dell'Orletta**[◇], **Simonetta Montemagni**[◇], **Giulia Venturi**[◇]
[•]DIBRIS, Università degli Studi di Genova, Italy
chiara.alzetta@edu.unige.it
[◇]Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR) - ItaliaNLP Lab
{name.surname}@ilc.cnr.it

**Introduction.** Over the last two decades, linguistic typology went through a renewing phase that brought to the development of what is called "Distributional Typology" (Bickel et al., 2015) aimed at uncovering diversities among languages more that at determining what is possible, as opposed to impossible, in human language. In this respect, both typological linguistic databases, such as WALS (Dryer and Haspelmath, 2013), and corpora, such as linguistically annotated treebanks, see among others Liu (2010); Futrell et al. (2015); Gulordava and Merlo (2015); Sharma et al. (2019), represent highly important information sources. Thus, defining methods to automatically infer typological features from these sources become a hot topic. This created a virtuos circle between the Natural Language Processing (NLP) and the linguistic typology communities where the first can contribute to the investigation of the issues opened by Distributional Typology and linguistic typology can support new solutions towards the development of robust and multilingually applicable NLP technologies. We contribute to the debate by showing how the methodology for typological feature identification in multilingual treebanks that we proposed in the 2019 edition of SIGTYP workshop (Alzetta et al., 2019a) and discussed in Alzetta et al. (2019b), can be exploited to measure the variability of the linguistic context in which the features occur and how the uncovered typological information can be used in a dependency parsing evaluation scenario.

**Method and Data.** Our methodology relies on the LISCA algorithm (Dell'Orletta et al., 2013) which creates a Statistical Model (SM) collecting statistics about a wide set of linguistically-motivated features from an automatically parsed reference corpus and it uses the SM to assign a score to each Dependency Relation (DR) instances contained in a target corpus. The output consists in a list of all DRs instances contained in the target corpus ranked by decreasing score. The LISCA score, computed by taking into account both local and global features, can be understood as a context-sensitive and frequency-based measure reflecting the degree of similarity of the "linguistic environments" in which a given DR instance occurs in the reference and target corpus. In other words, the score encodes the probability to observe a DR instance occurring in a specific context on the basis of the statistical model constructed starting from the reference corpus: higher LISCA scores identify DR instances that are prototypical with respect to the statistics acquired from the reference corpus, lower scores identify less common or even atypical DR instances of the target corpus.

As detailed in Alzetta et al. (2019b,a), 40 million tokens extracted from Wikipedia and parsed by the UDPipe pipeline (Straka et al., 2016) were used as reference corpus, and multilingual UD gold test sets as target corpora since they guarantee a consistent annotation formalism and cross-language comparability (Nivre et al., 2017). Thus, by comparing the rankings obtained for each language, it is possible to capture and measure similarities and differences across languages: the higher the position in the ranking, the more prototypical for that language the context where the DR instance occur. Differently from other approaches to infer typological features, this methodology allows observing a gradual transition from typical to atypical linguistic contexts rather than providing, e.g. only dominant features.

**Results.** Due to space constraints, we focus here on the results obtained for *i)* word order, computed as DR direction, and *ii)* DR length, computed as the linear distance from a dependent to its syntactic head (see Figure 1). We consider the distribution across two language-specific LISCA rankings, i.e. Italian and English, of the instances of a single UD DR: nominal subject (`nsubj`). The ranked list has

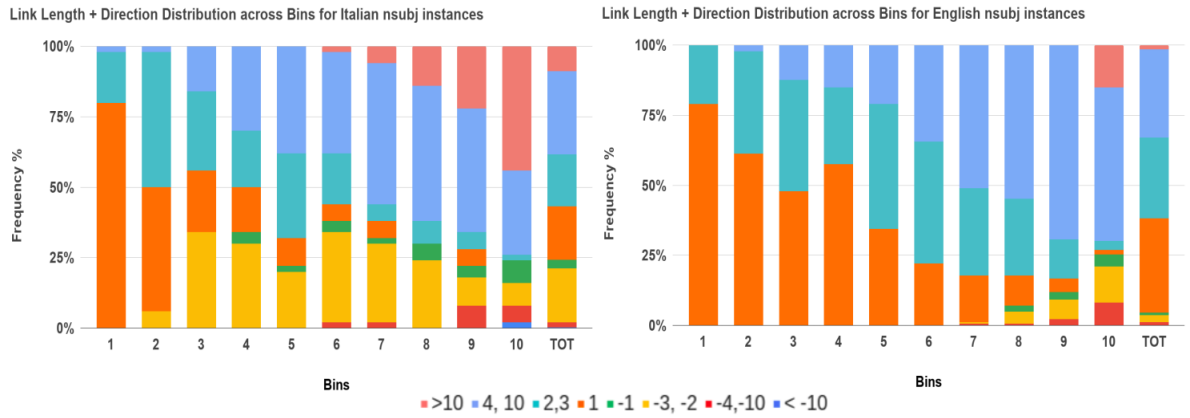


Figure 1: Relative frequency distribution across the LISCA bins and in the total set of Italian and English `nsubj` relations with respect to their link length and direction. Positive values identify right-headed relations, while negative values correspond to left-headed subjects.

been divided into 10 groups of equal size ("bins"), with the first bins containing `nsubj` instances presenting a high LISCA score and, conversely, the last bins with DR instances characterized by lower LISCA scores; while the last `TOT` column reports the flat distribution over the entire target corpus. The evidence emerging from the `TOT` column is not surprising if we consider the typological properties of the two languages: Italian nominal subjects resulted to be characterised by a higher word-order flexibility with respect to English ones. On the contrary, the distribution of the considered features across the LISCA bins provides a rich and articulated picture proving that considering typological features as discrete characteristics oversimplifies the description of language properties. For both languages, shorter right-headed (i.e. Subj-Verb order) links predictably concentrate in the first bins and, vice-versa, longer relations possibly following a "marked" order mainly occur in the bottom part of the ranking. For Italian, left-headed subjects appear from the first half of the ranking, with very few instances of $>$ 10-token long links all occurring in the last bin. For English, left-headed subjects concentrate in the last three bins and dependency length seems to be the main feature at play.

**Applications.** The proposed methodology can be reliably used for example in a multilingual dependency parsing evaluation scenario to built test suites that include typologically-relevant constructions which are more challenging for a parser (i.e. difficult-to-parse), rather than relation types or whole sentences (Naseem et al., 2012; Täckström et al., 2013; Scholivet et al., 2019). Preliminary results in this direction were achieved correlating the ranking positions obtained ordering UD DR types of the UD treebank test sets on the basis of *i)* the LISCA score assigned to each automatically produced DR type, *ii)* the LISCA score assigned to each gold DR type, *iii)* the *F-score*, computed for each DR type parsed with UDPipe by taking into account both head assignment and dependency label and *iv)* the *average length* of each DR type in the gold test set which represents a feature playing a key role for what concerns parsing accuracy (McDonald and Nivre, 2007). As shown in Figure 2, the correlation values between the LISCA scores and the F-score are significantly higher: this suggests that high positions in the LISCA ranking correspond to easier-to-parse DR types, as opposed to those relations ranked lower in the list which are more difficult-to-parse. In addition, the high correlation between the two LISCA scores testifies that LISCA produces a reliable ranking of dependency relations even without using gold data. This paves the way to create training sets with an homogeneous complexity degree (either easy or difficult) in a completely automatic way.

| | | LISCA Gold | LISCA Parsed | Lenght | F-score |
|---|---|---|---|---|---|
| IT | LISCA Gold | 1 | 0.99 | 0.61 | 0.87 |
| | LISCA Parsed | | 1 | 0.62 | 0.88 |
| | Lenght | | | 1 | 0.49 |
| | F-score | | | | 1 |
| EN | LISCA Gold | 1 | 0.99 | 0.64 | 0.74 |
| | LISCA Parsed | | 1 | 0.63 | 0.74 |
| | Lenght | | | 1 | 0.42 |
| | F-score | | | | 1 |

Figure 2: Ranking Correlation.

## References


Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2019a. Dissecting treebanks to uncover typological trends. a multilingual comparative approach. In *Proceedings of the ACL Workshop "Typology for Polyglot NLP"*.

Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2019b. Inferring quantitative typological trends from multilingual treebanks. a case study. *Lingue e linguaggio*, 18(2):209–242.

Balthasar Bickel, Bernd Heine, and Heiko Narrog. 2015. Distributional typology: Statistical inquiries into the dynamics of linguistic diversity.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2013. Linguistically-driven selection of correct arcs for dependency parsing. *Computaciòn y Sistemas*, 2:125–136.

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.

Kristina Gulordava and Paola Merlo. 2015. Structural and lexical factors in adjective placement in complex noun phrases across romance languages. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 247–257, Beijing, China. Association for Computational Linguistics.

Haitao Liu. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120:1567–1578.

Ryan McDonald and Joakim Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 122–131, Prague, Czech Republic. Association for Computational Linguistics.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–637, Jeju Island, Korea. Association for Computational Linguistics.

Joakim Nivre, Lars Ahrenberg Zeljko Agic, et al. 2017. Universal dependencies 2.0. lindat/clarin digital library at the institute of formal and applied linguistics, charles university, prague.

Manon Scholivet, Franck Dary, Alexis Nasr, Benoit Favre, and Carlos Ramisch. 2019. Typological features for multilingual delexicalised dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3919–3930, Minneapolis, Minnesota. Association for Computational Linguistics.

Kartik Sharma, Kaivalya Swami, Aditya Shete, and Samar Husain. 2019. Can Greenbergian universals be induced from language networks? In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 25–37, Paris, France. Association for Computational Linguistics.

M. Straka, J. Hajic, and J. Strakova. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1061–1071, Atlanta, Georgia. Association for Computational Linguistics.