Atypical or Underrepresented? A Pilot Study on Small Treebanks

Akshay Aggarwal[1] **and Chiara Alzetta**[2]
1. Twilio, Prague, Czechia
2. Istituto di Linguistica Computazionale “A.Zampolli”, CNR, Pisa - ItaliaNLP Lab
aaggarwal@twilio.com, chiara.alzetta@ilc.cnr.it

Abstract

We illustrate an approach for multilingual treebanks explorations by introducing a novel adaptation to small treebanks of a methodology for identifying cross-lingual quantitative trends in the distribution of dependency relations. By relying on the principles of cross-validation, we reduce the amount of data required to execute the method, paving the way to expanding its use to low-resources languages. We validated the approach on 8 small treebanks, each containing less than 100,000 tokens and representing typologically different languages. We also show preliminary but promising evidence on the use of the proposed methodology for treebank expansion.

1 Introduction and Motivation

Linguistically-annotated language resources like treebanks are fundamental for developing reliable models to train and test tools used to address Natural Language Processing (NLP) tasks acquiring linguistic evidence from corpora. Concerning the latter, researchers frequently rely on multilingual or parallel resources in contrastive studies to quantify the similarities and differences between languages (Jiang and Liu, 2018). Over the past few years, the Universal Dependencies (UD) initiative[1] (Zeman et al., 2021) has further encouraged such studies. UD defines a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages (Nivre, 2015; de Marneffe et al., 2021), and, at present, the project includes about 200 treebanks representing over 100 languages. The consistent annotation of linguistic phenomena under a shared representation and across different languages makes UD treebanks exceptionally well suited for quantitative comparison of languages (see, for example, Croft et al. (2017), Berdicevskis et al. (2018), Vylomova et al. (2020) and among our works, Alzetta et al. (2019a) and Alzetta et al. (2020a)).

Despite their great relevance for linguistic investigations, large treebanks are available for only a tiny fraction of the world’s languages (Vania et al., 2019). Even within the UD project, around 60% of the treebanks can be considered small, i.e. containing less than 100,000 tokens. Treebank size, in fact, is generally identified as the bottleneck for obtaining high-quality representative models of language use to be employed in downstream NLP applications. In general terms, larger datasets allow for better generalisations of language constructions, leading to better performances of systems trained using such data (Zeman et al., 2018). In fact, *ad-hoc* strategies are generally needed when dealing with low-resourced languages (Hedderich et al., 2021).

This paper illustrates a novel workflow specifically designed to adapt an existing methodology for treebank exploration to small treebanks. The base method, extensively described by Alzetta et al. (2020b), relies on an unsupervised algorithm called *LISCA* (*LInguistically–driven Selection of Correct Arcs*) (Dell’Orletta et al., 2013). LISCA has been successfully employed in past works for performing quantitative cross-lingual analyses (Alzetta et al., 2019a; Alzetta et al., 2019b; Alzetta et al., 2020a) and error detection on UD treebanks (Alzetta et al., 2017). The algorithm works in two main steps. First, it acquires evidence about language use from the distributions of phenomena in annotated sentences. The algorithm then uses such evidence to distinguish *typical* from *atypical constructions* in an unseen set of sentences. The

[1] https://universaldependencies.org

typicality of a construction is determined with respect to the examples observed in a corpus used as a reference, and is encoded with a score. This score, in fact, reflects the probability of observing a dependency occurring in a given context (both sentence-level and corpus-level) on the basis of the constructions sharing common properties reported in the reference corpus. Hence, from our point of view, typicality and frequency are tightly related concepts, as non-standard constructions are also usually less frequent in natural language use.

As such, the LISCA methodology relies on large sets of automatically parsed sentences to collect the statistics about phenomena distributions: even if the data contains parsing errors[2], the corpus size guarantees the collected statistics reflect the actual language use. However, such an approach can be employed only for analysing languages for which large amounts of data are available, or at least for which the parser outputs are generally considered reliable. To overcome such a limit, Aggarwal (2020) suggested that if the statistics are acquired from gold annotations (such as treebanks), the algorithm could collect the statistics from fewer data since these resources are assumed to be error-free.

We implemented this proposal by adapting the original LISCA workflow as detailed in Section 2. Our variation to the original methodology is inspired by the k-fold approach commonly used for performing systems' cross-validation: according to this approach, a dataset is split into sub-sets of equal size, iteratively used for training and/or evaluating a system. We employ a similar strategy for evaluating the typicality of the dependency relations in each treebank split, acquiring the statistics from the sentences contained in the other splits rather than from an external reference corpus. This small but substantial change in the method workflow allows us to apply the LISCA algorithm to small treebanks, which is particularly relevant in the case of analyses performed on low-resource languages.

We tested the methodology in a case study, reported in Section 3, involving 8 languages represented using UD treebanks. Our goal is to test if our method can support linguistic investigations for exploring and quantifying similarities and differences between typologically different languages. To this aim, we first validate the adaptation to the original LISCA approach proposed here in Section 3.1. Then, we exemplify how the obtained results can be employed for linguistic investigations in Section 3.2. To improve the cross–linguistic comparability of the analysis, we relied on Parallel UD (PUD) treebanks: a collection of parallel treebanks developed for the CoNLL–2017 Shared Task on multilingual parsing (Zeman et al., 2017) and linguistically annotated under the UD representation. Being parallel, PUDs are particularly well suited for carrying out multilingual studies since they contain only 1,000 sentences manually translated from English into the other languages, representing a perfect testbed for our approach.

Before concluding the paper in Section 5, we report the results of preliminary investigations to explore whether our approach could also be employed for automatically identifying underrepresented phenomena in treebanks. Søgaard (2020) and Anderson et al. (2021) argue that some treebanks cover only a restricted sample of the structures commonly used in a language, leaving out less common phenomena. This *leakiness* might affect the performances of NLP systems even more than the system architecture. Thus, treebanks should be expanded not only to improve their representativeness but also to obtain more truthful performances of systems trained using them. Section 4 investigates if our methodology can contribute to this issue by exploring its application in automatic treebank expansion.

The **contributions** of the paper can be listed as: *(i)* a novel approach specifically designed for carrying out multilingual investigations on small treebanks; *(ii)* a case study involving eight typologically different languages to test the methodology; and *(iii)* a novel formula, introduced in Section 3.2, to measure the distance between dependents and their syntactic head which improves the cross-lingual comparability of treebanks with respect to such property.

2 Approach

The method presented in this paper relies on a methodology for treebank exploration based on the unsupervised algorithm LISCA (Dell'Orletta et al., 2013), which we adapted to expand its usage for small treebanks, namely containing less than

[2] An assumption when producing automatically parsed data is that most of the errors made by a parser are consistent. As we showed in (Alzetta et al., 2017), the LISCA-based method allows to spot these errors types in annotations.

100,000 tokens.

As mentioned earlier, LISCA can be employed to quantify the typicality of each dependency relation (hereafter *deprel*)[3] of a linguistically annotated corpus with respect to a large set of examples taken as reference (Alzetta et al., 2020b). To achieve this goal, the algorithm first collects statistics about linguistically motivated properties of *deprels* extracted from a corpus of automatically parsed sentences (called *reference corpus*) to create a statistical model (SM). Then, the algorithm calculates a typicality score for each *deprel* appearing in a *test corpus* relying on the SM while also considering its linguistic context to assess the relevance of the *dependency label* used for marking the *dependency* in the given context. When interpreting the assigned LISCA score, a *deprel* marked by LISCA as highly typical was possibly frequently observed in similar contexts also in the reference corpus. In contrast, an atypical *deprel* could be characterised by certain properties which make it somehow distant from the other instances of *dependency* marked with the same *label* in the reference corpus.

In essence, LISCA computes the score for a given *deprel* taking into account local properties (e.g., dependency length and direction) of each *deprel* in the test corpus as well as the linguistic context where it is located (e.g., distance form root, leaves and number of siblings), comparing them both against the properties and contexts of all *dependencies* annotated with the same *dependency label* in the reference corpus. For this reason, the reference corpus has generally corresponded to a large corpus of around 40M tokens: the corpus size allows accounting for a more comprehensive set of examples of linguistic constructions while also compensating for possible parser errors.

Workflow. For this study, we implemented the adaptation of the LISCA workflow proposed by Aggarwal (2020). Inspired by the k-fold validation approach, we modified the original approach as follows:
1) Split a treebank into k portions of equal size ($k = 4$ for this work), each containing the same number of sentences;
2) Use LISCA to acquire the statistics (encoded in the SM) about the distribution of linguistic phenomena from a reference corpus obtained by merging $k-1$ portions of the previously split treebank;
3) Use the obtained SM to compute the typicality score of the *deprels* appearing in the remaining treebank portion (i.e., the one not included in the reference corpus);
4) Repeat steps 2 and 3 until all k portions are analysed;
5) Merge the analysed portions and order the *deprels* by decreasing LISCA score to have a unique ranking of all the *deprels* in the treebank.

The ordered ranking of *deprels* can be explored to investigate which linguistic constructions, represented by means of the *deprels*, were marked as typical or atypical, characterised by higher and lower scores, respectively.

2.1 Data and Languages

We tested our method on a selection of Parallel UD (PUD) treebanks (Zeman et al., 2017), each containing 1,000 sentences. In order to encompass different language families and genera[4], we carried out the case study on the following eight languages: **Arabic** (AR; Afro-Asiatic, Semitic), **Czech** (CZ; Indo-European, Slavic), **English** (EN; Indo-European, Germanic), **Hindi** (HI; Indo-European, Indic), **Finnish** (FI; Uralic, Finnic), **Indonesian** (ID; Austronesian, Malayo-Sumbawan), **Italian** (IT; Indo-European, Romance) and **Thai** (TH; Tai-Kadai, Kam-Tai).

3 Results

3.1 Validating the Approach

We report the results of an analysis to verify whether the adapted and original LISCA-based methods return comparable results. To this aim, we compared the LISCA ranking of PUD deprels obtained using the original algorithm workflow, which employs a large reference corpus to build the language SM, and the novel workflow defined above, which acquires the statistics from the treebank itself. We carried out this analysis for Italian and English PUD treebanks. We manually verified in previous studies that the original approach applied to those languages allows capturing elements of linguistic and parsing complexity

[3] Given a deprel $A \xrightarrow{nsubj} B$, we refer to $A \rightarrow B$ as the *dependency*, with $nsubj$ as the *dependency label*.

[4] The language family and genus, reported between parenthesis as (ISO language code, family, genus), are acquired from the World Atlas of Language Structures (WALS, available online `https://wals.info/languoid`) (Dryer and Haspelmath, 2013).

distinguishing between typical and atypical constructions along with the produced ranking of *deprels* (Alzetta et al., 2019a; Alzetta et al., 2020b).

We compared the *deprel* rankings obtained using the two methodology workflows in terms of Spearman correlation, which returns a rank correlation coefficient indicating a statistical dependence between the rankings of two observed variables. The analysis showed a strong and significant correlation between the rankings produced relying on the two workflows in both languages. Specifically, we obtained a Spearman correlation coefficient of 0.95 ($p < 0.5$) for Italian and English.

Such high correlations confirm that gold corpora, although small, can be used to acquire relevant statistics about language use. Manually revised data might be limited in size. However, their annotations are also generally correct in the case of rare phenomena, which a parser could wrongly annotate due to their low frequency in the data. While large reference corpora compensate for the possibly wrong parses assigned to rare constructions with their size, small reference corpora shall compensate with consistency and correctness. Hence, we could say that using gold data for building the SM allows reducing the number of examples for acquiring language statistics. We notice a difference between the two rankings only when focusing on the bottom part, where we find *deprels* with the lowest scores. While the original method produces only a tiny number of *deprels* with LISCA score equal to 0, which we usually excluded from the analyses, we observe many more of them in the ranking produced with our workflow adaptation. LISCA score zero is assigned to those dependencies never observed in the reference corpus; thus, their typicality is extremely low. It is not surprising that smaller reference corpora produce a higher number of these cases, given their limited coverage. However, the high correlation coefficient reported above suggests that such *deprels* are still interesting from a linguistic perspective. They correspond to rare constructions in the language, obtaining a score slightly higher than zero in the case of a larger reference corpus but are still placed in the lower positions of the ranking.

$$LL_{adjusted} = \begin{cases} \frac{LL_{raw} \cdot exp\left(1 - \frac{TrbAvgSentLen}{SentLength}\right)}{SentLength} & \text{if } \frac{SentLength}{TrbAvgSentLen} < 0.5 \\ \frac{LL_{raw}}{SentLength} & \text{if } \frac{SentLength}{TrbAvgSentLen} \in [0.5, 1.25] \text{ AND} \\ & LL_{raw} < \lfloor TrbAvgSentLen \rfloor \\ min\left(1, \frac{LL_{raw}}{TrbAvgSentLen}\right) & \text{otherwise} \end{cases}$$

Figure 1: $LinkLengthAdjusted$ formula for normalising *deprel length* in multilingual comparisons. **Note:** $\lfloor \cdot \rfloor$ denotes floor function, while $[a, b]$ denotes closed interval over a and b.

3.2 Rankings Exploration

This subsection exemplifies how the ranking of *deprels* obtained with our adapted approach can be employed in linguistic analyses to identify similarities and differences between languages. For this case study, we focused on a specific property of *deprels*, namely the *length* of the dependency link. The length of a *deprel*, measured as the linear distance in terms of intervening tokens between a word and its syntactic head, is a property frequently explored in linguistically annotated corpora. It is highly related to processing complexity in all languages (Demberg and Keller, 2008; Temperley, 2007; Futrell et al., 2015; Yu et al., 2019). For example, McDonald and Nivre (2011) observed that parsers tend to make more mistakes on longer sentences and longer dependencies. Such complexity makes this property particularly interesting from a multilingual perspective, especially when dealing with parallel corpora, as in our case study.

We inspected the ranking of *deprels* to monitor the LISCA score associated with *deprels* of different lengths and their distribution along the ranking of each language. To facilitate the rankings exploration and comparison, we split each ranking into three portions of equal size, referred to as *top*, *middle* and *bottom*, where *top* contains deprels obtaining the highest scores (more typical). In contrast, the *bottom* contains the *deprels* with the lowest scores (atypical).

In order to allow a proper multilingual comparison of the distribution of *deprel lengths* along with the rankings, we defined the novel measure called **Adjusted Link Length** ($LL_{adjusted}$, cf. Figure 1). The measure, inspired by Brevity Penalty used in BLEU score (Papineni et al., 2002), is designed to compute the length of *deprels* involving content words as dependant while simultaneously improving cross-language comparability as the length of

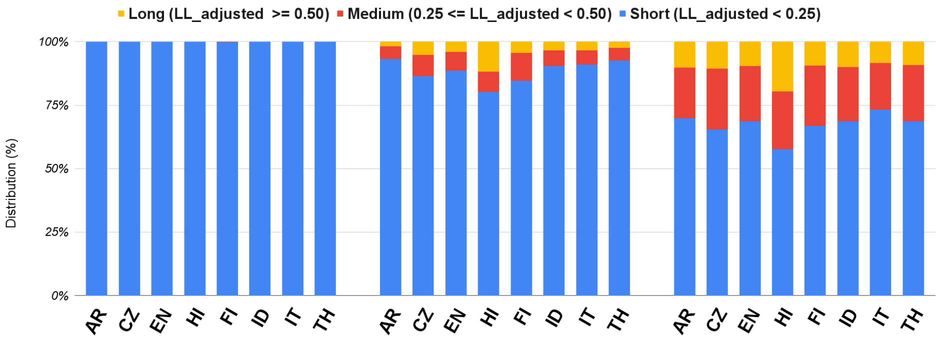

Figure 2: Distribution of Adjusted Link Length on content words across LISCA Rankings.

a *deprel* is measured keeping in mind the overall length of the sentence where it is located and the average sentence length in the treebank. This way, instead of comparing absolute length values, we can observe the tendency of languages towards producing longer or shorter *deprels*.

In $LL_{adjusted}$, we operationally compute the length of *deprels* as a function of *a)* the average sentence length in the treebank ($TrbAvgSentLen$), *b)* the length of the sentence where the *deprel* appears ($SentLength$), and *c)* the distance, in tokens, between the dependent and its syntactic head (LL_{raw}). The formula's values of 0.5 and 1.25 were determined empirically to account for unusually short and long sentences, respectively, in the treebank. Thus, the resulting value associated with each *deprel* denotes it as 'long', 'medium' or 'short' with respect to the average *deprel length* computed in the treebank. Note that, although our analysis focuses on content words, function words are still accounted for when computing the LISCA score as they might be part of the context of content words.

Figure 2 displays the distribution of *deprels* of different lengths (computed using $LL_{adjusted}$) along the portions of the treebank ranking of each language. The distributions show that longer *deprels* are given a lower plausibility score by LISCA in all languages. Interestingly, the length distributions are pretty similar across different languages except for Hindi. Such difference could be due to the typical word order of constituents of the considered languages. Hindi, in fact, is the only language of our set where the order of the main constituents is of the type S(ubject)O(bject)V(erb)[5], and the dominant word order of a language has been shown to influence the dependency length across major dependency types by Yadav et al. (2020).

It should be noted that such difference between languages could also be observed computing the length of dependency relations straightforwardly on PUD treebanks: the average linear link length computed on Hindi PUD is 6.54, for Thai PUD, the language showing shorter relations, is 2.67, while the remaining languages show a value ranging between 3.1 and 3.5. However, our methodology allows us to combine multiple properties simultaneously into a score, thus isolating in different portions of the rankings the *deprels* that show an atypical value for a given property but could be still considered quite typical for the language based on their context. As proof, observe that long and medium *deprels* in Hindi tend to appear earlier in the ranking than in other languages: 19.73% of deprels located in the middle bin are covered by medium and long *deprels*, suggesting that longer *deprels* are more common in Hindi. On the contrary, only 7% of *deprels* of the middle bin are long in Thai, pointing to their atypicality in the language.

The above results show the methodology's effectiveness for exploring tendencies and peculiarities of languages in multilingual studies. However, small samples like PUD treebanks are usually not suited for analysing infrequent phenomena (Taherdoost, 2016). Hence, one might wonder if we are actually capturing the atypicality of linguistic constructions, or instead, we are biased by phenomena underrepresented in the treebank. In the following Section, we will explore whether low LISCA scores might be associated with infrequent linguistic phenomena due to under-representation in the data used to build the SM.

[5] All the other languages are S(ubject)V(erb)O(bject) languages.

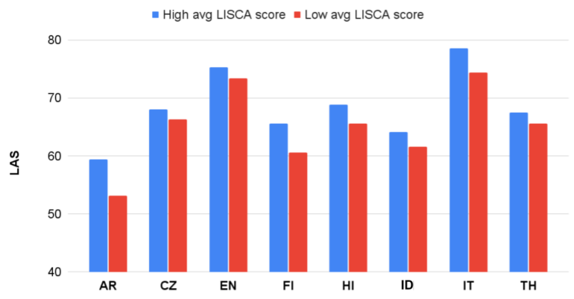

Figure 3: Parsing accuracy (LAS) on sentences having high and low LISCA scores.

4 Towards Treebank Expansion

Our analyses started from the premise that PUD treebanks are error-free. Therefore we can look at the rankings as containing correctly annotated examples of language use. However, the approach employed in this study does not exclude the scenario that a deprel might obtain a low LISCA score because of a lack of similar constructions in the treebank. We explored this idea both at *deprel* and sentence level, as described below.

Concerning the *deprel*–level analysis, we tested the accuracy of a parser for *deprels* in the three portions of the LISCA rankings. To this aim, we parsed each PUD treebanks using UDPipe (Straka et al., 2016), relying on the k-fold approach used to train LISCA: we split each PUD into 4 portions of 250 sentences each, trained UDPipe with $\frac{3}{4}$ of the portions and parsed the remaining portion. Then, we checked if *deprels* were parsed accurately. Again, we excluded function words from this analysis to improve cross-language comparability and avoid biased results as function words are usually more accurately parsed than content words. We observed that wrongly parsed deprels mainly concentrate in the bottom bins for all languages based on the obtained results. This suggests that there might be a relationship between low LISCA scores and underrepresented phenomena.

For the sentence-level analysis, we computed the LISCA score for each sentence in all PUD treebanks as the arithmetic mean of the scores of the individual *deprels* belonging to the sentence to get a sentence–level LISCA score. In the analysis, we explored whether sentences with low average LISCA scores are also more difficult to parse than those with higher average LISCA scores. Having computed the sentence–level LISCA scores, we collected two test sets of 100 sentences each by grouping sentences showing the highest and lowest LISCA scores. Then, we trained UDPipe using the remaining 800 sentences of PUD. The performances of UDPipe on the test sets are reported in terms of Labelled Attachment Score (LAS).

The results of this experiment are reported in Figure 3. We observe that the test sets composed of sentences characterised by the highest scores are more accurately parsed than the lower-score sets for all the languages involved. Differences between languages in terms of overall Label Attachment Score (LAS) and between the two subgroups of sentences will be further investigated in future work. Such results complement the *deprel*-level analysis: they suggest that the methodology could isolate difficult-to-parse sentences, and not only *deprels*, that could be employed to expand treebanks.

Treebank expansion is extremely valuable for low-resourced languages and small resources in general as it allows to include unseen examples to treebanks. Our results suggest that the sentence suites collected by grouping sentences characterised by the lowest LISCA scores contain difficult-to-parse constructions, possibly underrepresented in PUD, that should be included in the treebank to improve its representativeness.

5 Conclusion

We proposed a novel workflow to adapt an existing approach for treebank exploration to small treebanks and low-resourced languages. Results of our analyses showed the effectiveness of the methodology in multiple scenarios. First, the adapted method allows obtaining reliable results on par with the original method workflow when performing linguistic explorations of the treebanks. Secondly, the results also show the potential of the method for automatically identifying underrepresented constructions in treebanks. The latter result paves the way for the automatic identification of cases required to expand the treebanks, which we plan to further investigate in future work.

Acknowledgments

We would like to sincerely thank the anonymous reviewers for their helpful comments.

References

Akshay Aggarwal. 2020. Consistency of Linguistic Annotation. Master's thesis, Univerzita Karlova (ÚFAL), Prague, Czechia, September. Thesis Supervisor Zeman, Daniel.

Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2017. Dangerous Relations in Dependency Treebanks. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 201–210, Prague, Czech Republic.

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

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

Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni, Petya Osenova, Kiril Simov, and Giulia Venturi. 2020a. Quantitative Linguistic Investigations across Universal Dependencies treebanks. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it)*, Bologna (online), Italy, March.

Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2020b. Linguistically-driven Selection of Difficult-to-Parse Dependency Structures. *IJCoL. Italian Journal of Computational Linguistics*, 6(6-2):37–60.

Mark Anderson, Anders Søgaard, and Carlos Gómez-Rodríguez. 2021. Replicating and Extending "Because Their Treebanks Leak": Graph Isomorphism, Covariants, and Parser Performance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1090–1098.

Aleksandrs Berdicevskis, Çağrı Çöltekin, Katharina Ehret, Kilu von Prince, Daniel Ross, Bill Thompson, Chunxiao Yan, Vera Demberg, Gary Lupyan, Taraka Rama, et al. 2018. Using Universal Dependencies in cross-linguistic complexity research. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 8–17.

William Croft, Dawn Nordquist, Katherine Looney, and Michael Regan. 2017. Linguistic Typology meets Universal Dependencies. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, CEUR Workshop Proceedings, pages 63–75.

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

Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2013. Linguistically-driven Selection of Correct Arcs for Dependency Parsing. *Computación y Sistemas*, 17(2):125–136.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

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. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online, June. Association for Computational Linguistics.

Jingyang Jiang and Haitao Liu. 2018. *Quantitative Analysis of Dependency Structures*, volume 72. Walter de Gruyter GmbH & Co KG.

Ryan McDonald and Joakim Nivre. 2011. Analyzing and integrating dependency parsers. *Computational Linguistics*, 37(1):197–230.

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *International conference on intelligent text processing and computational linguistics*, pages 3–16. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Anders Søgaard. 2020. Some Languages Seem Easier to Parse Because Their Treebanks Leak. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2765–2770.

Milan Straka, Jan Hajic, and Jana Straková. 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 2016)*, pages 4290–4297.

Hamed Taherdoost. 2016. Sampling methods in research methodology; how to choose a sampling technique for research. *How to Choose a Sampling Technique for Research (April 10, 2016)*.

David Temperley. 2007. Minimization of dependency length in written English. *Cognition*, 105(2):300–333.

Clara Vania, Yova Kementchedjhieva, Anders Søgaard, and Adam Lopez. 2019. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1105–1116.

Ekaterina Vylomova, Edoardo M Ponti, Eitan Grossman, Arya D McCarthy, Yevgeni Berzak, Haim Dubossarsky, Ivan Vulić, Roi Reichart, Anna Korhonen, and Ryan Cotterell. 2020. Proceedings of the Second Workshop on Computational Research in Linguistic Typology. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*.

Himanshu Yadav, Ashwini Vaidya, Vishakha Shukla, and Samar Husain. 2020. Word Order Typology Interacts With Linguistic Complexity: A Cross-Linguistic Corpus Study. *Cognitive science*, 44(4):e12822.

Xiang Yu, Agnieszka Falenska, and Jonas Kuhn. 2019. Dependency length minimization vs. word order constraints: an empirical study on 55 treebanks. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 89–97.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada, August. Association for Computational Linguistics.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium, October. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, Bilge Nas Arıcan, H͡órunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Deniz Baran Aslan, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Lauren Cassidy, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Mihaela Cristescu, Philemon. Daniel, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-

Rae Han, Muhammad Yudistira Hanifmuti, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Eva Huber, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Kaoru Ito, Tomáš Jelínek, Apoorva Jha, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Krister Lindén, Nikola Ljubešić, Olga Loginova, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Dmitry Sichinava, Janine Siewert, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Rachele Sprugnoli, Steinh͡ór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, Hanzhi Zhu, Anna Zhuravleva, and Rayan Ziane. 2021. Universal dependencies 2.8.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.