What quantifying word order freedom can tell us about dependency corpora

Maja Buljan
University of Oslo / Language Technologies Group
majabu@ifi.uio.no

Abstract

Building upon existing work on word order freedom and syntactic annotation, this paper investigates whether we can differentiate between findings that reveal inherent properties of natural languages and their syntax, and features dependent on annotations used in computing the measures. An existing quantifiable and linguistically interpretable measure of word order freedom in language is applied to take a closer look at the robustness of the basic measure (word order entropy) to variations in dependency corpora used in the analysis. Measures are compared at three levels of generality, applied to corpora annotated according to the Universal Dependencies v1 and v2 annotation guidelines, selecting 31 languages for analysis. Preliminary results show that certain measures, such as subject-object relation order freedom, are sensitive to slight changes in annotation guidelines, while simpler measures are more robust, highlighting aspects of these metrics that should be taken into consideration when using dependency corpora for linguistic analysis and generalisation.

1 Introduction

With the breadth of existing resources and research into developing dependency treebanks, cross-linguistic research has expanded to large-scale comparative work, formalising and computing quantifiable properties of natural language. The use of morphological and syntactic annotations, to name a few, has enabled typological research to move from type-based—treating languages as individual data points with a categorical value—to token-based—making generalisations and comparative analyses by using corpora to observe linguistic units in language use and express their behaviour using aggregate measures (Levshina, 2019).

In this work, the focus is on word order freedom, a property of natural language syntax, extensively covered in previous work that makes use of dependency treebanks (Liu, 2010; Futrell et al., 2015; Naranjo and Becker, 2018). The main point of interest is word order freedom expressed by the measure of Word Order Entropy (WOE), as defined by Futrell et al. (2015).

The cited work expands on methodological issues, aiming to find a balance between linguistic interpretability, robustness independent of corpus size, and cross-lingual applicability. The defined measure also enables quantitative verification of hypotheses on the relation between case marking and word order freedom (Kiparsky, 1997); word order freedom and patterns across languages with respect to head direction; and the positions of subject and object in the main clause (Greenberg et al., 1963).

However, in applying this measure to different corpus domains and sources, several issues arise and require further addressing—mainly, when expressing word order freedom with measures based on dependency annotations, does the measure reveal more about the language itself, or the annotation used as a layer between the raw text and the computable data? Further, and in line with the question raised in the original study, is this measure consistent across corpus sizes, and different text samples?

These questions are investigated through a replication of the methodology on the same set of languages covered by the original study (with minor exceptions). The aim is to compare two generations of Universal Dependency annotation styles (Nivre et al., 2016b, 2020), using the latest releases of Universal Dependencies v1 (Nivre et al., 2016a) and v2 (Zeman et al., 2021). The analysis is focused on three levels—(1) comparing scores obtained over the full corpus with multiple random samples, to verify whether the measure is robust to sample size; (2) comparing scores across two versions of annotation guidelines in the same style, to test whether

Proceedings of the Seventh International Conference on Dependency Linguistics (Depling, GURT/SyntaxFest 2023), pages 54 - 67
March 9-12, 2023

the measure remains consistent through alterations in annotation guidelines and treebank development; and (3) comparing this replication study to the original findings, partially overlapping in corpora, to further verify consistency.

Section 2 gives a brief summary of the key methodological points of Futrell et al. (2015) (further also referred to as "the original study"). Section 3 highlights the specifics of the experimental setup. Results and findings are presented in Section 4, and Section 5 concludes the paper.

2 Background

Futrell et al. (2015) define *word order freedom* as "the extent to which the same word or constituent in the same form can appear in multiple positions while retaining the same propositional meaning and preserving grammaticality." The cited study aims to employ dependency treebanks in computing quantitative properties of natural language syntax—specifically, word order freedom—and develop linguistically interpretable measures.

The degree of word order freedom is quantified through the unordered dependency graph of a sentence, using conditional entropy:

$$H(X|C) = \sum_{c \in C} p_C(c) \sum_{x \in X} p_{x|c}(x|c) \log p_{X|C}(x|c) \tag{1}$$

where X is the dependent variable, conditioned on C, the conditioning variable. Since directly measuring the conditional entropy of sequences of words is intractable, the authors decide on three entropy measures over partial information about dependency trees, considering three parameters: (1) estimating H from joint counts of X and C (further discussed in 3.2); (2) information contained in X; and (3) information contained in C. The goal is to balance the need to avoid data sparsity against the preference to retain linguistic interpretability.

To avoid the issue of sparsity, entropy is computed only on local subtrees—consisting of a head and its immediate dependents. To avoid issues with misrepresented variability in certain word order phenomena, this means preferring annotation styles with content-head dependency. This requirement is satisfied in Universal Dependencies annotations.

Futrell et al. (2015) introduce three measures of word order entropy (WOE):

Relation Order Entropy (ROE) Conditioning on the unordered local subtree structure (C being the set of dependency relations and part-of-speech (POS) tags of constituents), the dependent variable X is the linear order of relation types expressed in the local subtree.

Subject-Object Relation Order Entropy (SOE) Assuming that ROE will result in some data sparsity issues despite limiting the search to local subtrees, SOE narrows the criteria to local subtrees containing relations of type *nsubj* and *dobj* (UDv1) or *obj* (UDv2), conditioned on the POS of these dependents.

Head Direction Entropy (HDE) The most narrowly defined of these measures, HDE is conditioned only on a dependent and its head, for all relation types; the dependent variable denotes whether the head is to the left or right of the dependent.

3 Experimental setup

This study follows the methodology of Futrell et al. (2015) as closely as possible, with three exceptions: omitting three languages from the original study due to data limitations, adjusting entropy estimation due to technical limitations, and performing computations over multiple random subcorpora samples to perform a more robust evaluation of the effects of sampling and data sparsity. The experimental setup is further detailed in subsequent paragraphs.

3.1 Treebank matching

In order to compare WOE scores between UDv1 and UDv2 annotations of the same text, it is first necessary to consolidate the available treebanks across the 34 languages of the original study. The aim is to retain the maximum number of sentences with both UDv1- and UDv2-style annotations.

The last release of UDv1 is used: version 1.4 (Nivre et al., 2016a); and the latest release of UDv2 at the time when the experiments were carried out: version 2.8 (Zeman et al., 2021).

Two of the languages featured in the original study—Bengali and Telugu—do not have a UDv1 release; the original study used HamleDT annotations (Zeman et al., 2012). For this reason, they cannot be featured in the analysis, so the total number of languages is reduced to 32, with a total of 52 available treebanks.

	UD1 vs. UD2		
	<	=	>
no. of treebanks	17	24	11

Table 1: Breakdown of available treebanks and their UD1 vs. UD2 coverage, by treebank count, for 32 languages featured in the original study.

Despite the continuous growth of both the number of languages featured in UD, as well as the respective treebanks (Nivre et al., 2020), the data is limited to the intersection of treebanks (or, in certain cases, individual sentences) between UDv1.4 and UDv2.8. Table 1 breaks down the treebank coverage between releases for the 32 languages group. The majority of treebanks either have an exact match between the two releases, or UDv2 expands the treebanks featured in UDv1, in terms of sentence count. For a fifth of the cases, there is a reduction in the number of sentences going from the UDv1 to the UDv2 version of the treebank.

To ensure a truly "parallel" corpus of UDv1 and UDv2 annotations, those treebank sentences that do not feature in either of the two latest releases need to be removed. Given that the releases followed no set sentence identifier standard before UDv2.0, this means resorting to heuristic matching methods.

The heuristic matching raised unexpected challenges in equating sentences that a human reader would consider superficially identical. Most of these challenges stemmed from increased annotator experience and refined annotation guidelines—resulting in, e.g., altered dependency relations between constituents, and different annotation conventions for multi-word expressions and complex names—or were the result of updated tokenisation, lemmatisation, and treatment of abbreviations. Due to this, the features taken into consideration in the matching process were wordform and lemma comparisons, POS tags and dependency relations, and the Levensthein distance of sentence surface forms.

During the matching process, Japanese was also removed from the pool of languages, due to a negligibly small (roughly 200) number of sentences identified as matches in the only treebank featured both in the UDv1.4 and UDv2.8 release.

Finally, Figure 1 visualises the total size of the annotated corpora[1] per language, from the smallest treebank (Tamil, 600 sentences) to the largest

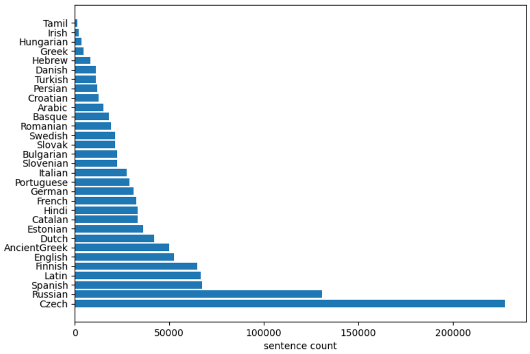

Figure 1: Total corpus size in number of sentences.

collection of treebanks (Czech, 113 682 sentences).

Due to the large variation in corpus sizes, and in line with Futrell et al. (2015), the experiments are performed both on the full corpora for each language, and on 10 randomly sampled subcorpora of 1000 sentences for each language. Note that, while the 1000 sentences are picked randomly, the samples are matched between the UDv1 and the UDv2 versions of the corpus—maintaining the "same sentence, two annotations" setup.

3.2 Entropy estimation

Apart from the equally sized subcorpora, Futrell et al. (2015) address the issue of sample size by applying the bootstrap entropy estimator of DeDeo et al. (2013), arguing that entropy is otherwise underestimated. However, due to backward compatibility issues with the implementation of the bootstrap estimator in the original study, this study resorts to using the naive estimator (Cover et al., 1991), assuming that the analysis performed is not sensitive to the order of magnitude of absolute entropy scores, as its internal consistency allows for forming and comparing rankings between languages. This is further discussed in Section 4.3.

3.3 Variables

In line with the approach of Futrell et al. (2015), conditional entropy is computed on local subtrees: a head and its immediate dependents. The conditioning variable is the unordered set of dependency relations between the head and its dependent(s), and the POS tags of all constituents.

In the case of relation order entropy, the dependent variable is the linear order of relation types in the subtree. For subject-object entropy, the dependent variable is the linear order of the subject and

[1]Detailed statistics are given in Appendix A.

object in subtrees whose predicate head has both a subject and an object in its dependents. Finally, head direction entropy is computed over all head-and-dependent pairs, where the dependent variable notes whether the head is to the left or right of its dependent.

4 Analysis

The aim of the analysis is threefold: (1) comparing the scores obtained on the full corpora against the random samples, to evaluate the effects of sampling and data sparsity, as well as comparing the random samples to estimate variance; (2) comparing UDv1 scores to UDv2 scores, to evaluate the effect of annotation; and (3) comparing the results of the original study to the rankings obtained on UDv1 and UDv2.

4.1 Full corpus vs. random sample

Figures 2 through 4 present the entropy estimates over the full corpora[2] and randomly sampled subcorpora, for UDv1 and UDv2, over the three metrics described in Section 2.

In the case of Relation Order Entropy (Figure 2), there is a clear difference between the full-corpus entropy estimates and the random-sample scores, which would also affect the rankings of the featured languages on a scale from "least-" to "most word order freedom", if the WOE score was used as the main quantifying metric. As mentioned in Section 3.2, Futrell et al. (2015) argue that the entropy estimator plays a role in under- or overestimating the entropy score, considering data sparsity and the long-tailed frequency distribution of words in natural language. However, with the naive estimator, this difference between the full corpus and the 1000-sentence samples is not nearly as striking for the other two metrics, SOE (Figure 3) and HDE (Figure 4); nor do the full-corpus rankings correlate, at a glance, with the corpus sizes shown in Figure 1. An observed explanation for this discrepancy is the fact that ROE—the least narrowly defined metric—allows for an explosion in the number of possible values for the conditioning variable when computing over the full corpus, compared to the relatively limited set of values available in the subcorpora.

Subject-Object Relation Order Entropy (Figure 3) shows less of a discrepancy between full-corpus entropy and that of subcorpora, in line with the SOE metric being more limited in the number and type of constituents forming the values for the conditioning variable. However, there is more of a variance between the entropy scores of different subcorpora (represented with red dots in the figures) than seen with the other two metrics. Furthermore, the different subcorpora scores again have the potential to dramatically alter the rankings. In the case of a relatively narrow definition of word-order metric, where the dependent variable values are permutations of (subject, object, predicate) paired with POS tags, this brings into question the reliability of random samples to give an accurate WOE score according to which languages may consistently be compared as more or less rigid in word order freedom.

Finally, Head Direction Entropy (Figure 4) demonstrates the highest (visual) match between full-corpus and subcorpora scores. Intuitively, this is in line with expectations, considering the narrow definition of HDE and the binary value of the dependent variable—a small random sample will likely have a similar distribution to the full corpus.

The figures alone imply that random samples may be less reliable than full-corpus scores if the WOE metric is less narrowly defined. However, in an attempt to not rely on visualisations alone, these differences are also quantified by calculating the Kendall rank correlation coefficient between rankings obtained from the full-corpus entropy scores, and those based on random-sample scores. Table 2 presents these coefficients, comparing the UDv1 and UDv2 computations, as well as the rankings from the original study for comparison.

The correlation between random samples and full-corpus scores expressed in Kendall τ (Table 2, top) is rather low—and in most cases not significant. The only metric that shows a weak correlation is HDE. Table 3 presents the correlation score between WOE rankings and rankings according to corpus size. No correlation is found between corpus size and WOE ranking, which seems to support the decision to use naive entropy estimations to formulate rankings.

4.2 UDv1 vs. UDv2

Figures 2 through 4 also allow for comparison between scores and rankings computed over the UDv1 and UDv2 annotations.

Figure 2, ROE, apart from a shift in rankings,

[2]Note that, for all metrics, entropy estimates for the full Tamil corpus match all random samples—as the full corpus comprises 600 sentences in total.

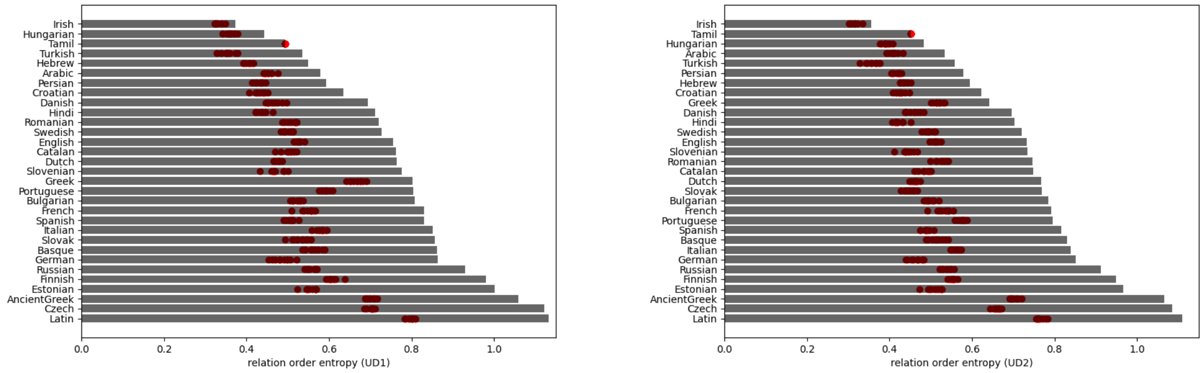

Figure 2: ROE; UDv1 (left) vs. UDv2 (right). The bar represents the relation order entropy estimated from the full corpora; the red dots represent entropies estimated from ten random samples of 1000-sentence subcorpora. Languages are ranked according to the full-corpus ROE estimate.

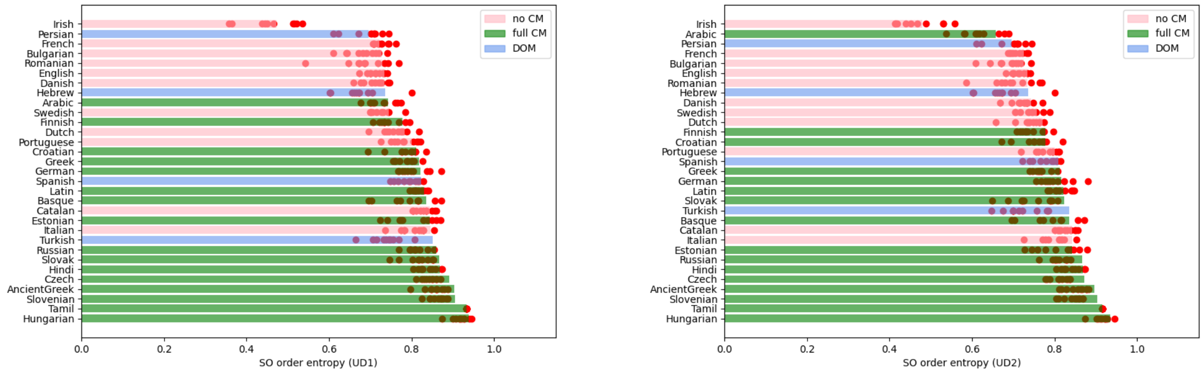

Figure 3: SOE; UDv1 (left) vs. UDv2 (right). The bar represents the relation order entropy estimated from the full corpora; the red dots represent entropies estimated from ten random samples of 1000-sentence subcorpora. Languages are ranked according to the full-corpus SOE estimate. Bars are coloured in line with Futrell et al. (2015), denoting the nominative-accusative case marking system type: “full” means fully present case marking; “DOM” means Differential Object Marking (Aissen, 2003).

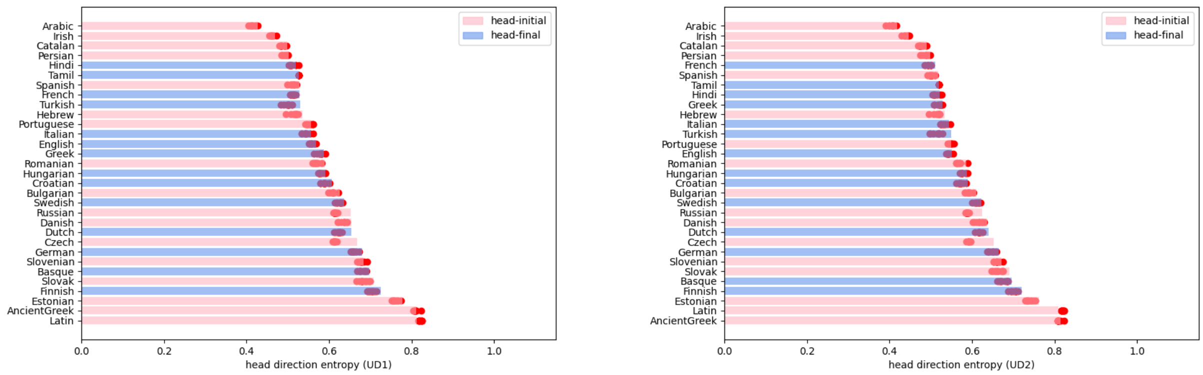

Figure 4: HDE; UDv1 (left) vs. UDv2 (right). The bar represents the relation order entropy estimated from the full corpora; the red dots represent entropies estimated from ten random samples of 1000-sentence subcorpora. Languages are ranked according to the full-corpus HDE estimate.

	ori	UDv1	UDv2
ROE	.161 $_{p=0.210}$	.165 $_{p=0.259}$	.098 $_{p=0.484}$
SOE	.449 $_{p=0.001}$	.068 $_{p=0.584}$	.187 $_{p=0.215}$
HDE	.372 $_{p=0.003}$	.297 $_{p=0.071}$	.200 $_{p=0.176}$

Table 2: Kendall τ entropy estimate rank correlation (averaged in the case of UDv1 and UDv2), comparing full corpus vs. random sample rankings. "ori" denotes rank correlation between full corpus and random sample rankings for data from the original study—note that these scores are based on rankings obtained from visualisations (as absolute entropy estimates were not available), and using only a single data point for each language's random samples.

	full	sample	
ROE	.027 $_{p=0.839}$	.027 $_{p=0.839}$	UDv1
	-0.07 $_{p=0.566}$	.006 $_{p=0.973}$	UDv2
SOE	.002 $_{p=1.0}$	.088 $_{p=0.499}$	UDv1
	.062 $_{p=0.636}$	.118 $_{p=0.361}$	UDv2
HDE	-0.17 $_{p=0.164}$	-0.16 $_{p=0.198}$	UDv1
	-0.01 $_{p=0.919}$	-0.18 $_{p=0.144}$	UDv2

Table 3: Kendall τ scores for WOE vs. corpus size rankings.

also shows different discrepancies between full-corpus scores and random-sample scores for particular languages, as well as different "outliers" in this sense.

The differences are even more notable in the case of SOE (Figure 3). Futrell et al. (2015) make observations on word order freedom implying the presence of case marking, as in the highest-scoring third of languages according to Figure 3. However, certain outliers demonstrate different behaviour between annotation versions. While superficial changes in labelling, e.g., direct objects and passive subjects from UDv1 to UDv2 are accounted for in the computing process, these results imply a non-negligible effect of annotation guidelines or annotator choices on results quantifying word order freedom. In fact, looking into differences between the "parallel" UD corpora reveals nearly universal discrepancies in the number of annotated *nsubj* and *(d)obj* relations, resulting in the more severely affected languages changing their relative positions in the rankings.

As in the previous section, HDE (Figure 4) is the most consistent between annotation versions, with the same group of head-initial languages ranking most- and least-rigid with respect to word order, and variations in rank mostly being pairwise switching. This again confirms the most narrowly-defined measure to be the most robust.

	full	sample
ROE	.105 $_{p=0.417}$	.273 $_{p=0.089}$
SOE	.088 $_{p=0.499}$	.110 $_{p=0.465}$
HDE	.397 $_{p=0.001}$	.380 $_{p=0.013}$

Table 4: Kendall τ entropy estimate rank correlation, comparing UDv1 vs. UDv2 rankings, for full corpus scores and random samples.

	full	sample
ROE	.225 $_{p=0.076}$	.051 $_{p=0.525}$
SOE	.075 $_{p=0.566}$	.052 $_{p=0.612}$
HDE	.075 $_{p=0.566}$	.025 $_{p=0.555}$

Table 5: Kendall τ entropy estimate rank correlation, original study vs. newly obtained rankings; UDv1 only.

Again, Table 4, top shows an attempt to quantify the differences between UDv1 and UDv2 scores through the Kendall τ of rankings. Again, the scores are mostly insignificant, with HDE being the least unstable measure across annotation versions.

4.3 Comparing across studies

Finally, WOE rankings obtained on UDv1 data are compared[3] with those retrieved from the Futrell et al. (2015) study. Rank correlations, again expressed in Kendall τ only, are given in Table 5.

No correlation is found between the rankings obtained on random samples for any of the metrics. Further work is needed to determine how much this is influenced by differences in the corpus content and annotations, or possibly different methods of entropy estimation—especially in the case of ROE, the only notable outlier in this case.

5 Conclusion

This paper has taken a deeper look into an existing methodology of quantifying word order freedom in dependency corpora. The study attempted to determine whether this methodology and measure allows for draw reliable conclusions about word order freedom, or whether it depends to a relevant extent on the underlying dependency annotations—both in terms of annotation guidelines, and in the quality of annotation depending on annotator experience and consistency. The study identified diffi-

[3]In the interest of space, visual comparisons between the scores provided in the original study and those obtained through these computations are not included in the main body of this work; however, they are available in Appendix C.

culties in finding a definition of measure that would be robust enough to avoid noise and misrepresentation, yet fine-grained enough to give meaningful linguistic insight. The analysis shows that changes in annotation styles can alter the results of estimates and change the comparative presentation of word order freedom across languages. Furthermore, it has shown that the observed measures may be susceptible to differences between samples, and that random sampling as defined by this methodology is selectively unreliable, depending on measure complexity. In conclusion, there is merit in cross-testing treebank-based metrics on different versions of treebanks, considering changes in annotation guidelines or even annotator teams, as well as on random subsamples of treebanks. Future work may also investigate the optimal size for these samples—currently fixed on an arbitrary count.

Building on existing work on Universal Dependencies, the question that next arises concerns what potential levels of complexity using Enhanced Universal Dependencies would introduce to this method of quantifying word order freedom. Future work may also focus on similar comparisons between manually annotated (gold-standard) and automatically generated dependency annotations, as well as possible differences between domains (e.g., newswire vs. literary text; written vs. spoken corpora), as well as across different annotation styles.

Acknowledgements

I would like to thank Stephan Oepen, Lilja Øvrelid, Joakim Nivre, and Paola Merlo for valuable discussions and comments on this work. I also thank the anonymous reviewers for their comments.

References

Judith Aissen. 2003. Differential object marking: Iconicity vs. economy. *Natural Language & Linguistic Theory*, 21(3):435–483.

Thomas M Cover, Joy A Thomas, et al. 1991. Entropy, relative entropy and mutual information. *Elements of information theory*, 2(1):12–13.

Simon DeDeo, Robert XD Hawkins, Sara Klingenstein, and Tim Hitchcock. 2013. Bootstrap methods for the empirical study of decision-making and information flows in social systems. *Entropy*, 15(6):2246–2276.

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.

Joseph H Greenberg et al. 1963. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2:73–113.

Paul Kiparsky. 1997. The rise of positional licensing. *Parameters of morphosyntactic change*, 460:494.

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

Haitao Liu. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578.

Matías Guzmán Naranjo and Laura Becker. 2018. Quantitative word order typology with ud. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), December 13–14, 2018, Oslo University, Norway*, 155, pages 91–104. Linköping University Electronic Press.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Riyaz Ahmad Bhat, Eckhard Bick, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Fabricio Chalub, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Claudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Jan Hajič, Linh Hà Mỹ, Dag Haug, Barbora Hladká, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Jessica Kenney, Natalia Kotsyba, Simon Krek, Veronika Laippala, Lucia Lam, Phng Lê Hồng, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Keiko Sophie Mori, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Hanna Nurmi, Petya Osenova, Robert Östling, Lilja Øvrelid, Valeria Paiva, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Baiba Saulīte, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Carolyn Spadine, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Lars Wallin, Jing Xian Wang, Jonathan North Washington, Mats Wirén, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2016a. Universal dependencies 1.4. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016b. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

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. *arXiv preprint arXiv:2004.10643*.

Daniel Zeman, David Marecek, Martin Popel, Loganathan Ramasamy, Jan Stepánek, Zdenek Zabokrtskỳ, and Jan Hajic. 2012. Hamledt: To parse or not to parse? In *LREC*, pages 2735–2741.

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, Ĥórunn 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, Phng 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, Lng 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.

A Corpus statistics

Table 6: Comprehensive list of corpus statistics; sentence count, subtree count, number of subtrees with noun subject and direct object, total count of noun subjects, *nsubj* of which passive, total count of direct objects; per language, and per annotation guidelines version, sorted by total corpus size (ascending).

TBs	UDv	sen	st	has(ns,do)	mult(ns)	nsubj	(o. w. passive)	(d)obj
Tamil	1	600	3901	205	1	665	1	12705
	2	600	3937	167	0	664	1	10492
Irish	1	1010	8762	298	1	1600	0	35252
	2	1010	9052	307	1	1562	0	34987
Hungarian	1	1800	16213	849	0	2614	0	44215
	2	1800	16252	850	1	2621	0	44298
Greek	1	2302	20713	1139	6	3299	0	68570
	2	2302	20106	1011	0	2499	711	62047
Hebrew	1	4198	36479	896	8	5447	0	67334
	2	4198	37177	896	8	5447	0	67371
Danish	1	5509	33106	3257	129	8402	683	110374
	2	5509	33943	3282	95	9085	0	110304
Turkish	1	5619	23750	1027	14	3588	0	58166
	2	5619	23440	976	14	3730	0	54963
Persian	1	5997	62226	1786	22	8861	149	128609
	2	5997	63611	1786	22	8861	149	128609
Croatian	1	6283	51595	2500	2	7798	818	128826
	2	6283	52995	3194	20	9944	0	137521
Arabic	1	7651	123462	7865	35	15732	562	1101114
	2	7651	128242	5246	448	17815	774	494711
Basque	1	8993	45923	2473	4	8716	0	102881
	2	8993	46946	2473	4	8716	0	102881
Romanian	1	9519	83019	3180	7	10178	1857	182848
	2	9519	84178	3183	0	10090	1928	177917
Swedish	1	10589	59962	5564	16	28792	3756	180440
	2	10589	61477	5871	4	29880	3888	182691
Slovak	1	10601	36869	2884	0	7120	220	80395
	2	10601	37791	2003	0	7121	220	57701
Bulgarian	1	11137	56582	3721	1	10209	1240	109351
	2	11137	57622	3354	0	10066	1434	99099
Slovenian	1	11168	56792	2745	0	17496	0	160994
	2	11168	58212	2747	0	17494	0	160187
Italian	1	13779	100170	4458	1	12401	2280	297825
	2	13779	101065	4478	2	12425	2275	296198
Portuguese	1	14400	106352	6431	8	33456	1416	305249
	2	14400	108011	6270	1	31196	3230	338361
German	1	15590	95538	6699	9	17346	3191	176865
	2	15590	97725	6468	10	17412	3192	171913
French	1	16334	136590	9666	24	21005	2716	423183
	2	16334	141126	7232	0	19689	3114	359869
Hindi	1	16611	134715	9020	8	21023	562	410484
	2	16611	128192	9021	8	21023	562	410484
Catalan	1	16677	187178	16818	223	27523	0	1405814
	2	16677	192623	16500	74	27431	25	1408426

(cont. on next page)

TBs	UDv	sen	st	has(ns,do)	mult(ns)	nsubj	(o. w. passive)	(d)obj
Estonian	1	18009	81927	5277	0	20099	0	181768
	2	18009	83159	5226	0	20201	0	181212
Dutch	1	20906	104414	6809	20	40866	0	309076
	2	20906	101450	6838	11	41118	5802	170403
AncientGreek	1	24929	126503	7193	27	42958	4578	428714
	2	24929	125374	6646	15	42610	3788	402153
English	1	26298	142986	12266	37	111537	7005	475591
	2	26298	145774	12320	26	111255	7245	482468
Finnish	1	32302	122859	7206	11	60748	0	256977
	2	32302	125952	7237	12	61190	0	257568
Latin	1	33309	172925	11014	30	96978	29253	503707
	2	33309	176146	7583	29	101202	24639	359377
Spanish	1	33693	346221	20607	205	45537	1182	1803269
	2	33693	355407	17591	30	45460	1234	1604446
Russian	1	65378	438671	14965	4	166572	11406	699931
	2	65378	451072	15224	2	150972	16170	709338
Czech	1	113682	761586	48833	8	334719	34563	2278743
	2	113682	780840	33216	3	334953	34563	1482098

B Corpus statistics, visualised

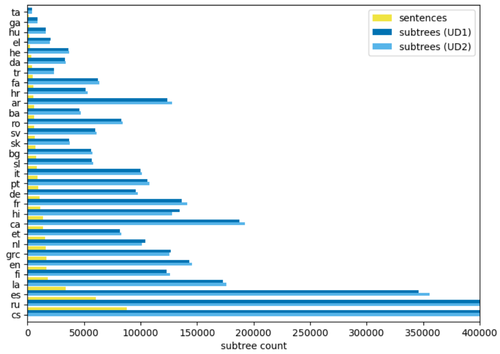

Figure 5: Number of subtrees, per language, across annotation gudeline versions.

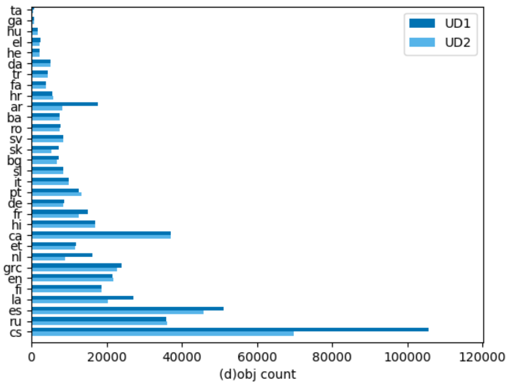

Figure 6: Number of (D)OBJ relation heads, per language, across annotation guideline versions.

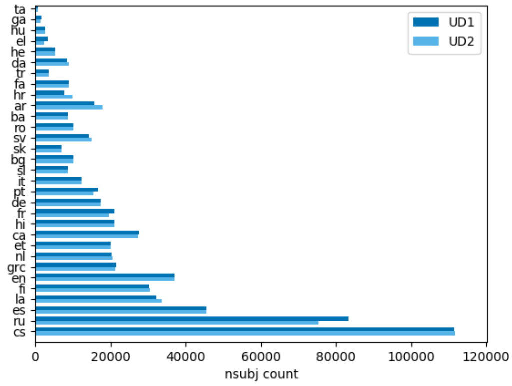

Figure 7: Number of NSUBJ relation heads, per language, across annotation guideline versions.

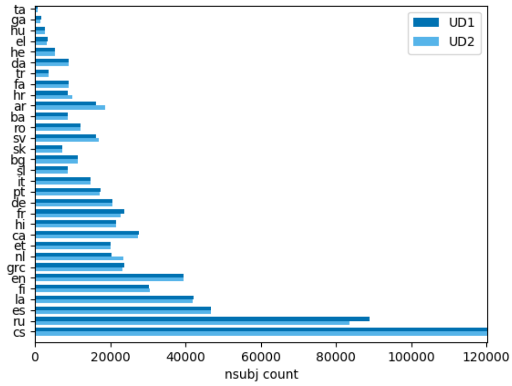

Figure 8: Number of NSUBJ relation heads, incl. variations of PASS, per language, across annotation guideline versions.

C Additional comparisons

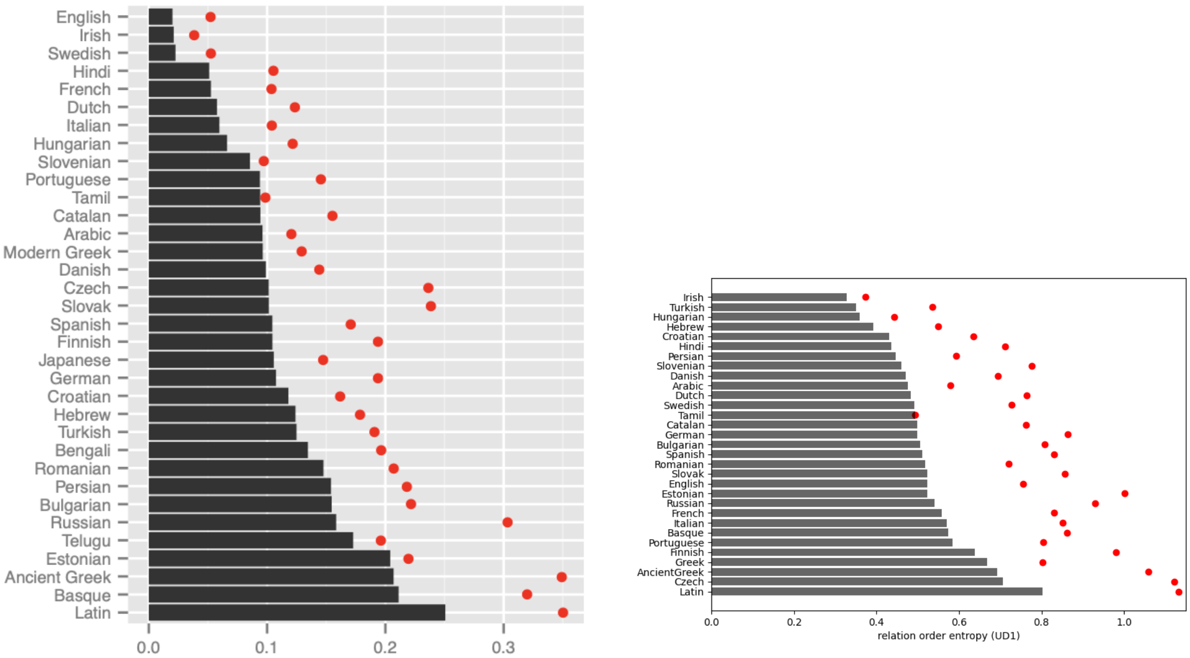

Figure 9: ROE; original study vs. UD1 rerun (random sample vs. full treebank)

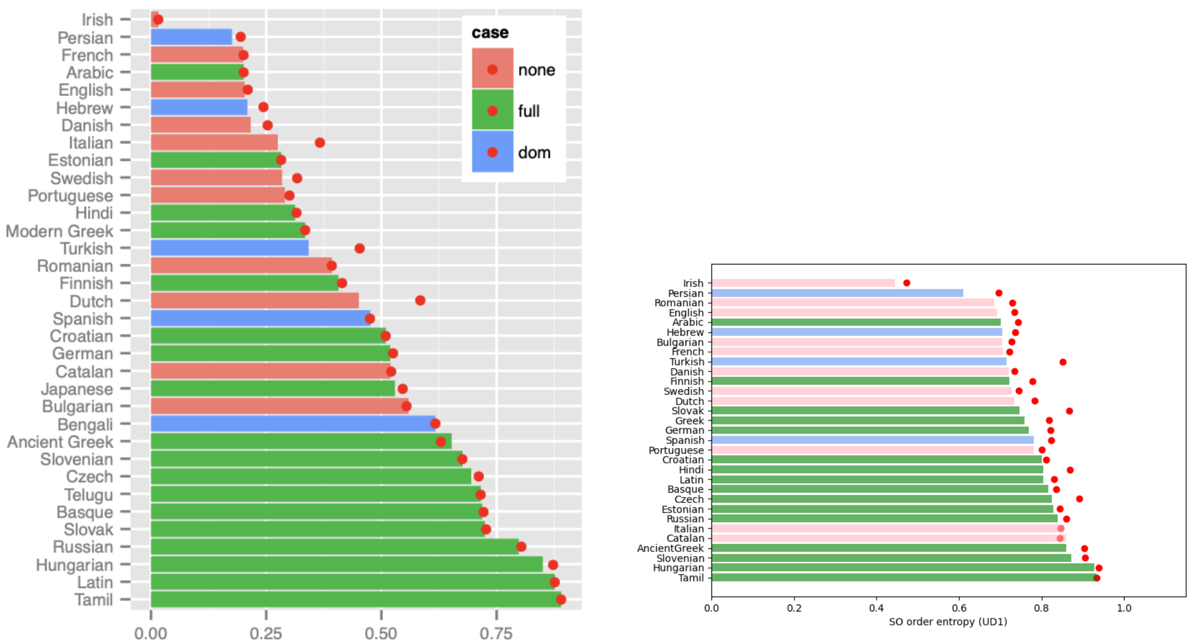

Figure 10: ROE; original study vs. UD1 rerun (random sample vs. full treebank)

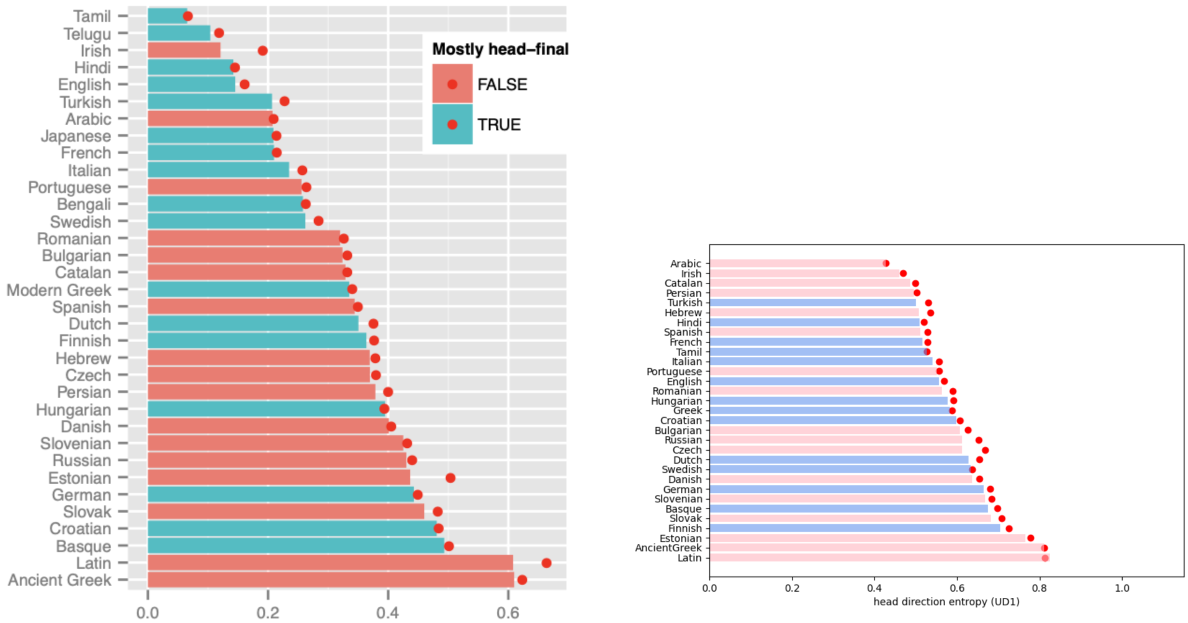

Figure 11: ROE; original study vs. UD1 rerun (random sample vs. full treebank)