Tweaking UD annotations to investigate the placement of determiners, quantifiers and numerals in the noun phrase

Luigi Talamo
Language Science and Technology, Saarland University
luigi.talamo@uni-saarland.de

Abstract

We describe a methodology to extract with finer accuracy word order patterns from texts automatically annotated with Universal Dependency (UD) trained parsers. We use the methodology to quantify the word order entropy of determiners, quantifiers and numerals in ten Indo-European languages, using UD-parsed texts from a parallel corpus of prosaic texts. Our results suggest that the combinations of different UD annotation layers, such as UD Relations, Universal Parts of Speech and lemma, and the introduction of language-specific lists of closed-category lemmata has the two-fold effect of improving the quality of analysis and unveiling hidden areas of variability in word order patterns.

1 Introduction

Most of the work on word order variation using Universal Dependencies (UD: de Marneffe et al., 2021) is based on curated dependency treebanks, with only a few works using dependency corpora derived from raw texts. Although the accuracy rate of NLP systems trained on UD models is reportedly very high (Hajič and Zeman, 2017; Zeman and Hajič, 2018; Straka et al., 2019; Qi et al., 2020), a certain level of noise i.e., erroneous annotations is in fact present when working with automatically annotated texts (Levshina et al., to appear; Talamo and Verkerk, to appear); furthermore, different layers of UD annotations such as Universal Parts of Speech (UPOS) and UD Relations are not always used consistently across languages, often resulting in the cross-linguistic comparison of different categories.

We discuss a methodology to tweak the UD annotations in order to achieve a better representation of word order entropy; the methodology is exemplified on three categories that are particularly difficult to analyze with automatic methods and from a cross-linguistic perspective. Determiners, quantifiers and numerals are often treated in descriptive grammars as heterogeneous categories; the lexical category of determiners includes articles and demonstratives, while the category of quantifiers often includes elements of other closed categories, such as pronouns and gradation markers, and sometimes members of open categories, such as adjectives and adverbs; finally, numerals are often not restricted to cardinal, ordinal and distributive numbers, but overlap with quantifiers.

This heterogeneity is reflected by the UD implementation of these categories, both at the Relation and the UPOS annotation layer. Numerals are treated as a separate category and represented at the syntactic level by the `nummod` UD Relation and at word category level by the NUM UPOS; by contrast, the UD framework conflates articles, demonstratives and quantifiers into one UPOS tag (DET) and into one UD Relation (`det`), resulting in the 'Determiners & Quantifiers' macro-category. At the language-specific level several individual POS tags and UD Relation subtypes are used; for instance, in Slavic languages quantifiers get two specific subtypes, `det:numgov` and `det:numposs`, and morpho-syntactic features of numerals can be specified using additional UD Relation subtypes.

Our methodology combines two layers of UD annotations, UPOS and UD Relations, with manually-compiled and language-specific lists of lemmata. We test the methodology against a parallel corpus of fiction texts and their translations in 10 Indo-European languages; given their particular genre, these texts are quite challenging for parsers that are mostly trained on non-fiction data such as Wiki and News. Following previous studies, we employ here Shannon's entropy as a metric for word order variation.

2 Related work

Since its inception in 2015 (Nivre et al., 2015), UD has been widely used in corpus-based studies on word order variability. However, as earlier

Proceedings of the 4th Workshop on Computational Typology and Multilingual NLP (SIGTYP 2022), pages 36 - 41

mentioned, corpora used in previous studies are "dependency corpora of the HamleDT 2.0 and Universal Dependencies 1.00" (Futrell et al., 2015), "the Universal Dependencies Treebank version 2.2" (Naranjo and Becker, 2018), "a selection of 55 treebanks from Universal Dependencies v2.4" (Yu et al., 2019), "Surface-Syntactic Universal Dependencies (SUD) [treebanks]" (Gerdes et al., 2019) and the "Universal Dependencies project, release 2.1" (Futrell et al., 2020). UD Treebanks can be considered de-facto gold standards, as large parts of them are manually compiled or at least semi-automatically checked for wrong annotations, allowing scholars to work with high quality of data. However, as UD Treebanks wildly vary across languages with respect to size and text genres (Levshina et al., to appear), results from most of the previous works are biased against these factors. Exceptions are represented by works using the LISCA parse assessment algorithm (Dell'Orletta et al., 2013), whose models have been trained on UD-parsed Wikipedia corpora (Alzetta et al., 2018) and tested on the so-called 'reference corpora', which consist "of a monolingual corpus of texts from the news and Wikipedia domains [...] morpho-syntactically annotated and dependency parsed by the UDPipe pipeline trained on the Universal Dependency treebanks, version 2.2" (Alzetta et al., 2020); automatically annotated texts are also partially employed in Levshina (2019), who uses eleven UD-parsed corpora from the Leipzig Corpora Collection for one of her case-studies on word order entropy. Finally, Talamo and Verkerk (to appear) is a study on word order variation in the nominal phrase and is entirely based on parallel texts that are parsed by the UDPipe pipeline trained on UD treebanks v.~2.5. To the best of our knowledge, Levshina (2019) and Talamo and Verkerk (to appear) are the only studies on word order variation combining UD Relations with other annotation layers; although her methodology is not fully disclosed, Levshina (2019) applies the UPOS annotation layer to the head in her first case study, where word order variability is taken from a syntactic perspective, and the wordform annotation layer to the dependent in her second case-study, where word order variability is investigated with respect to the lexically specific level; Talamo and Verkerk (to appear) take a step further and operationalize this methodology by introducing several combinations of UD Relation and UPOS annotation layer to restrict either the head, the dependent or both, and introducing language-specific list of lemmata to match modifiers at the lexical level.

3 Data and Methods

3.1 Corpus

We use the Parallel Corpus of Indo-European Prose and more (CIEP+: Talamo and Verkerk, to appear), which features prosaic texts and their translations in more than 30 languages; we select a sample of 10 Indo-European languages, belonging to the following branches: Balto-Slavic (Lithuanian, Polish), Celtic (Irish), Germanic (Danish, Dutch and German), Greek (Modern Greek), Romance (French, Portuguese and Spanish). All languages feature 18 books (approximately 120K of sentences for each language), with the exception of Irish that features 5 books (approximately 13K of sentences).

The corpus has been parsed using Stanford Stanza[1] with pre-trained UD Models[2] for the 10 languages. The resulting CoNLL-U files are processed with a Python script using the pyconll library[3]. The script extracts the occurrences of the specific UD Relations (see below) and determines the relative position of head and dependent; for each occurrence, we collect the following annotation fields for both the head and the dependent: UD Relations, UPOS tag and lemma.

Scripts and dataset, with the exception of the parsed corpus containing copyrighted texts, are available in the Supplementary Material[4].

3.2 Tweaking the UD annotations

Working with a dependency grammar, the most important annotation layer is represented by the UD Relations, which identifies the head and the dependent within the phrase; for our case study, we deal with various dependents (Table 1) and one type of head, the noun. This basic layer of annotation is then combined with other layers of annotation. By formulating the categories in Table 1 as comparative concepts (Haspelmath, 2010), we seek to integrate cross-linguistic definitions with the different layers of annotation (see 'Comparative Concepts

[1] Version 1.3.0 https://stanfordnlp.github.io/stanza/

[2] Version 2.8 https://stanfordnlp.github.io/stanza/available_models.html

[3] https://pyconll.readthedocs.io/en/stable/

[4] https://doi.org/10.5281/zenodo.6580701

Category	UPOS	UD Relation
nominal head	NOUN, PROPN	-
article	DET	`det`
demonstrative	DET *PRON*	`det`
quantifier	DET *ADJ ADV PRON*	`det det:nummod det:numgov`
numeral	NUM	`nummod nummod:entity nummod:gov nummod:flat`

Table 1: Values used to capture the categories of nominal heads, articles, demonstratives, quantifiers and numerals. Non-specific values are given in italics.

and Universal Dependencies' in the Supplementary Material).

Elaborating on Talamo and Verkerk (to appear), we propose three combinations, with each combination building on top of the previous one. The first combination, `rel`, uses the specific UD relation for the dependent, thus corresponding to most of the approaches taken in previous works; the second combination, `rel+upos`, adds the UPOS layers, using specific UPOS tags for the nominal head, and specific and non-specific UPOS tags for the dependent; the third combination, `rel+pos+lemma`, introduces language-specific list of lemmata for the dependent.

With 'specific' UPOS tags we refer to the values that are described by the UD Annotation Guidelines[5] as relevant for the investigated categories; we additionally add non-specific UPOS tags, which are based on the consultation of descriptive grammars and on the comparison of UD-parsed texts across the ten languages of the sample. See Table 1 for a detailed list of these values. Finally, language-specific lists of lemmata are aimed to capture the three components of the 'Determiners & Quantifiers' macro-category, namely, articles, demonstratives and quantifiers; we use these lists of lemmata, which are compiled using descriptive grammars and with the aid of native speakers, in intersective queries (positive match) on the lemma field for articles, demonstratives and quantifiers, and in non-intersective queries (negative match) on the lemma field for numerals.

3.3 Metrics

Word order variability is assessed using Shannon's entropy:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

where the upper bound of summation, n, is set to 2, indicating that there are two possible word order patterns i.e., the modifier is either prenominal or postnominal, and P represents the probability of the two order patterns; the resulting value of entropy is given in bits and ranges from 0 (only one of the two word order patterns is attested) to 1 (both word order patterns are attested with the same probability). For instance, we find in the Danish part of CIEP+ 5089 occurences of postnominal `nummod` (and subtypes) and 1392 of prenominal `nummod` (and subtypes), with a resulting entropy of .75.

4 Results

As shown in the left panel of Figure 1, entropy values captured by the second combination, `rel+upos`, are significantly lower than values captured by the `rel` combination. This is particularly clear for numerals, which display a substantial drop in entropy in half of the languages; furthermore, in some languages the entropy of numerals is reduced to near-zero values (German: .02, French and Irish: .04, Polish: .06). As for determiners & quantifiers, the introduction of the UPOS layer is overall less significant; a significant reduction of entropy is observed only for three languages (Greek, Lithuanian and Spanish). This is partly due to the low value of entropy already captured by the `rel` combination, but it also reflects the biunivocal relation between the DET UPOS tag `det` and the UD Relation and its subtypes; by contrast, the `nummod` UD relation and its subtypes are not in biunivocal relation with the NUM UPOS tag, since, as already mentioned above, this UD Relation is often used with quantifiers as well. The third combination, which adds language-specific lists of lemmata to the UPOS and UD Relation layers and is plotted in the right panel of Figure 1, allows us to zoom in on the entropy of determiners & quantifiers, disentangling the category into articles, demonstratives

[5] https://universaldependencies.org/guidelines.html

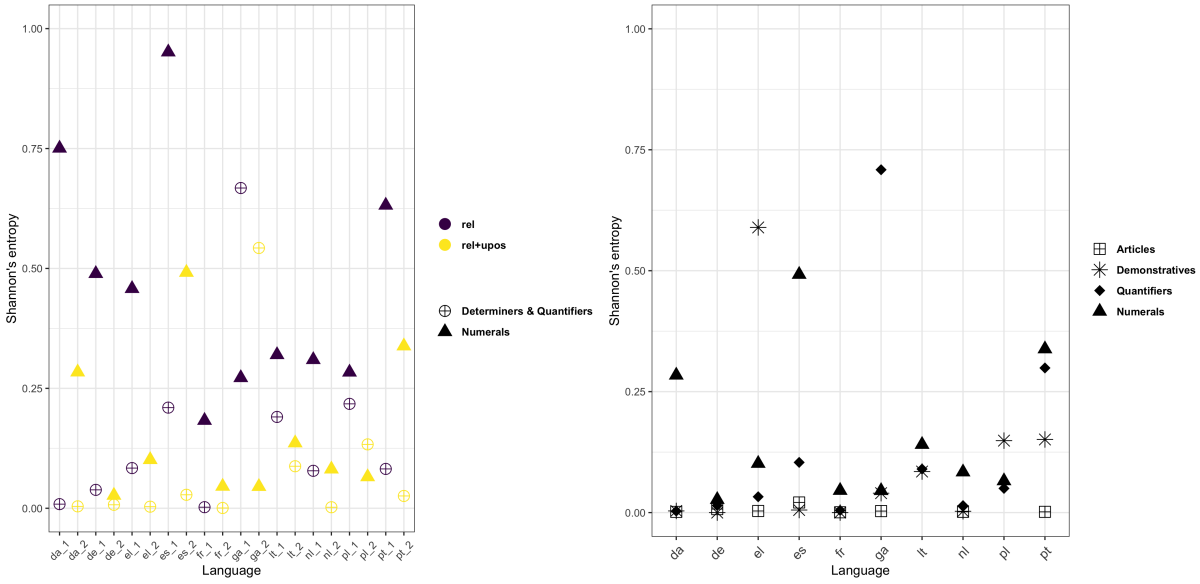

Figure 1: Left: entropy values for the categories of determiners & quantifiers and numerals, as captured by the `rel` combination and the `rel+upos` combination. Right: entropy values for the categories of articles, demonstratives, quantifiers and numerals, as captured by `rel+pos+lemma` combination.

and quantifiers; although the entropy of determiners & quantifiers is extremely low for all languages of the sample, two languages have moderate values of entropy for demonstratives or quantifiers. More specifically, Irish has the highest value of entropy for quantifiers (.71), while Greek the highest value for demonstratives (.59). According to Stenson (2020, 189-192), the position of quantifiers in Irish is lexically determined; most Irish quantifiers precede the noun, while few follow it; postnominal quantifiers include the high-frequency lexeme *uilig* 'all', which explains the high entropy of Irish quantifiers. As for Greek, the high value of entropy for demonstratives can be accounted on a pragmatic and semantic basis, as postnominal demonstratives have an emphatic reading (Lascaratou, 1998, 164). Furthermore, low-to-moderate values of entropy are observed for quantifiers in Portuguese (.30) and Spanish (.10) and for demonstratives in Portuguese (.15). As for numerals, we use language-specific lists of quantifiers as negative matches against the lemma field; this approach is however of little use, as entropy values captured by the third combination are the same of the second combination. Thanks to the introduction of language-specific list of lemmata, the third combination is suitable for closed categories such as articles, demonstratives and quantifiers, while the second combination is already effective for capturing open categories such as numerals.

5 Conclusion

We have discussed a methodology to extract with better accuracy word order patterns from CoNLL-U files obtained from the automatic parsing of raw texts; the methodology, which exploits different UD annotation layers and language-specific list of lemmata, is exemplified on the heterogenous lexical categories of determiners and numerals, whose word order patterns are analyzed in a parallel corpus of 10 Indo-European languages. We have shown that the methodology is able to correct some of the errors introduced by the automatic parsing and inconsistent use of UPOS tags and UD Relations, thus improving the quality of the analysis, as shown with the category of numerals. Furthermore, the methodology sheds light on areas of variability, which were previously hidden by the UD lumping of articles, demonstratives and quantifiers into a unitary category. Given the very high frequency of articles, whose variability is close to zero, this unitary category displays very low values of entropy across languages; once this unitary category is split into its three components, some languages show moderate-to-high levels of entropy with respect to demonstratives (Greek) and quantifiers (Irish and Portuguese).

References

Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni, Petya Osenova, Kiril Simov, and Giulia Venturi. 2020. Quantitative linguistic investigations across universal dependencies treebanks. In *CLiC-it*.

Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2018. Universal Dependencies and quantitative typological trends. a case study on word order. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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*, 17:125 – 136.

Richard Futrell, Roger P. Levy, and Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language*, 96(2):371–412.

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

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

Jan Hajič and Dan Zeman, editors. 2017. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, Vancouver, Canada.

Martin Haspelmath. 2010. Comparative concepts and descriptive categories in cross-linguistic studies. *Language*, 86(4).

Chryssoula Lascaratou. 1998. Basic characteristics of modern greek word order. In Anna Siewierska, editor, *Constituent Order in the Language of Europe*, pages 151–171. Mouton de Gruyter, Berlin.

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

Natalia Levshina, Savithry Namboodiripad, Alex Kramer, Annemarie Verkerk, Luigi Talamo, Marc Tang, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Gabriela Garrido Rodriguez, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, Natalya Stoynova, Liu Ying, and Sasha Wilmoth. to appear. Why we need a gradient approach to word order. *Linguistics*.

Matias-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)*, pages 91–104, Oslo, Norway. Linköping Electronic Conference Proceedings 155:10.

Joakim Nivre, Željko Agić, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Cristina Bosco, Sam Bowman, Giuseppe G. A. Celano, Miriam Connor, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Daniel Galbraith, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Berta Gonzales, Bruno Guillaume, Jan Hajič, Dag Haug, Radu Ion, Elena Irimia, Anders Johannsen, Hiroshi Kanayama, Jenna Kanerva, Simon Krek, Veronika Laippala, Alessandro Lenci, Nikola Ljubešić, Teresa Lynn, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Shunsuke Mori, Hanna Nurmi, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Prokopis Prokopidis, Sampo Pyysalo, Loganathan Ramasamy, Rudolf Rosa, Shadi Saleh, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Kiril Simov, Aaron Smith, Jan Štěpánek, Alane Suhr, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Sumire Uematsu, Larraitz Uria, Viktor Varga, Veronika Vincze, Zdeněk Žabokrtský, Daniel Zeman, and Hanzhi Zhu. 2015. Universal dependencies 1.2. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Nancy Stenson. 2020. *Modern Irish: A Comprehensive Grammar*. Comprehensive grammars. London & New York: Routledge.

Milan Straka, Jana Straková, and Jan Hajic. 2019. Evaluating contextualized embeddings on 54 languages in POS tagging, lemmatization and dependency parsing. *CoRR*, abs/1908.07448.

Luigi Talamo and Annemarie Verkerk. to appear. A new methodology for an old problem: a corpus-based typology of adnominal word order in European languages. *Italian Journal of Linguistics*.

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, Paris, France. Association for Computational Linguistics.

Daniel Zeman and Jan Hajič, editors. 2018. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, Brussels, Belgium.