# A Computational Method for Analyzing Syntactic Profiles: The Case of the ELEXIS-WSD Parallel Sense-Annotated Corpus

**Jaka Čibej**
Centre for Language Resources and Technologies
Faculty of Computer and Information Science
Faculty of Arts
University of Ljubljana
jaka.cibej@ff.uni-lj.si

## Abstract

In the paper, we present an approach to comparing corpora annotated with dependency relations. The method relies on the compilation of syntactic profiles – numeric vectors representing the relative frequencies of different syntactic (sub)trees extracted automatically with the *STARK 3.0* open-access dependency tree extraction tool. We perform the extraction on the *ELEXIS-WSD Parallel Sense-Annotated Corpus*, which has recently been published as version 1.2 with UD dependency relation annotations for 10 European languages. The corpus provides an additional resource for contrastive studies in quantitative syntax. In addition to presenting the corpus and conducting some proof-of-concept analyses, we discuss several other potential uses and improvements to the proposed approach.

## 1 Introduction

The proliferation of corpus resources annotated with dependency relations in the last decade (such as *Universal Dependencies Treebanks*; de Marneffe et al., 2021) has facilitated automatic syntactic analyses with different computational approaches. However, the field of quantitative syntax analysis is arguably still discovering its full potential, and methods that have been ubiquitous in other (sub)fields of computational linguistics are still to be implemented in quantitative syntax studies. The same is true for language resources, with new corpora being developed every year but not included in syntactic studies. The growing interest of the research community in quantitative syntax studies is emphasized by studies focusing on the benefits of quantitative methods (e.g. Gibson et al., 2012), as a counterweight to the prevalent methods of obtaining a judgment of the acceptability of a sentence pair by a handful of participants (Gibson and Fedorenko, 2010). In addition, data extraction for quantitative analyses has been facilitated by recently developed tools specialized for syntactic features (Krsnik et al., 2024; Krsnik and Dobrovoljc, 2025; Yang and Liu, 2025).

The goal of this paper is to make a contribution to the growing toolbox of quantitative syntax methods by (a) presenting a new approach to comparing syntactically annotated corpora with the use of syntactic profiles (numerical vectors of quantitative syntactic features; see Section 4), and (b) introducing the *ELEXIS-WSD Parallel Sense-Annotated Corpus 1.2* (see Section 3), a new multilayered and multilingual parallel corpus that can be used for syntactic analyses.

The paper is structured as follows: we first provide a brief overview of related work in analyses and tools for syntactically annotated (parallel) corpora (Section 2). We then describe the latest version of the ELEXIS-WSD Parallel Sense-Annotated Corpus (Section 3) and the method for extracting syntactic profiles from its subcorpora as well as individual sentences (Section 4). We analyze the corpus-level and sentence-level syntactic profiles (Section 5) with statistical tests to determine the most statistically significant differences in distributions of syntactic structures across different languages. In Section 6, we focus on the analysis of individual syntactic structures. We conclude the paper (Section 7) with several suggestions for future improvements to the method.

## 2 Related Work

Many studies in quantitative syntax so far have focused on a restricted set of specific syntactic phenomena (see e.g. van Craenenbroeck et al., 2019 for a study of word order in verb clusters in 186 Dutch dialects; Poppek et al., 2021 for an analysis of differences between regular transitive and experiencer-object verbs in German; or Niu et al., 2021 for an analysis of the properties of rare constructions such as it-clefts and topicalization in



*Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025)*, pages 63–71

| Language | Tokens |
|---|---|
| Bulgarian | 33,978 |
| Danish | 33,012 |
| English | 34,497 |
| Spanish | 37,822 |
| Estonian | 26,378 |
| Hungarian | 29,851 |
| Dutch | 35,543 |
| Italian | 41,609 |
| Portuguese | 41,136 |
| Slovene | 31,233 |
| **Total** | **345,059** |

Table 1: Number of tokens in subcorpora of ELEXIS-WSD 1.2.

English) or tests of pre-determined language universals (Choi et al., 2021). Instead of focusing on a specific syntactic phenomenon, our approach is designed in more bottom-up manner (see Section 4).

The study most similar to our approach was conducted by Klyshinsky and Karpik, 2019, who extracted syntactic profiles from the Universal Dependencies corpora by focusing on co-occurrences of words and syntactic relations, then cross-comparing the most frequent pairs to obtain similarity/correlation scores between languages. However, the method only provided results on the level of individual languages and their subcorpora, and the syntactic profiles used were limited to a limited set of the most frequent tuples. We build on this approach and focus not only on subcorpora, but individual sentences. In addition, we do not deconstruct syntactic (sub)trees into tuples of relations and focus on a much larger set of complete syntactic (sub)trees as features extracted from the *ELEXIS-WSD Parallel Sense-Annotated Corpus* (see Section 3).

## 3 Corpus

The ELEXIS-WSD Parallel Sense-Annotated Corpus (Martelli et al., 2021) is a dataset that in its current version (1.2; Čibej et al., 2025) consists of subcorpora containing the same 2,024 sentences in 10 European languages: Bulgarian, Danish, English, Spanish, Estonian, Hungarian, Italian, Portuguese, and Slovene. An example of a sentence and some of its parallel equivalents is shown in Table 2. The size of the corpus in tokens is shown in Table 1.

The corpus was primarily designed within the ELEXIS project[1] as a word-sense disambiguation dataset in which the content words (verbs, nouns, adjectives, and adverbs) in each subcorpus are annotated with their corresponding senses from an accompanying sense inventory (a collection of lexemes and their sense divisions with definitions).

The sentences were extracted from WikiMatrix (Schwenk et al., 2021), a collection of parallel sentences from Wikipedia, and selected according to several mostly semantic criteria (e.g., the number of semantically ambiguous words). Missing translations into other languages were automatically translated and manually validated by native speakers. The final versions were tokenized, lemmatized and morphosyntactically tagged using UDPipe (Straka et al., 2016; Straka, 2018).[2] These annotation layers were also manually validated, and the corpus is available in the CoNLL-U format under a Creative Commons BY-SA 4.0 license.

Within the context of the UniDive COST Action (*Universality, Diversity and Idiosyncrasy in Language Technology*; Savary et al., 2024), which at the time of writing this paper is still underway, the ELEXIS-WSD corpus is being extended with new languages on the one hand, and new annotation layers on the other. This includes Universal Dependencies parsing annotations (Tiberius et al., 2024), which were absent in previous versions. For the Slovene and Estonian subcorpora, the annotations have already been manually validated. For the other languages, the dependency relations were added using the UDPipe 2.15 models.[3] The performance of the models on gold tokenization is shown in Table 3.[4] All models achieve relatively high F1 scores, with the Hungarian model being the least accurate. The majority of automatic syntactic annotations in the corpus are thus expected to be correct. The corpus, although somewhat small in size and not entirely manually validated, should thus be sufficient for our proof-of-concept experiment on comparing syntactic profiles of corpora.

Version 1.2 is the first version that makes ELEXIS-WSD suitable as an additional resource

[1] *European Lexicographic Infrastructure* (ELEXIS): https://project.elex.is/

[2] UDPipe: https://lindat.mff.cuni.cz/services/udpipe/

[3] For Dutch, the validation is still ongoing at the time of writing this paper, so only automatic annotations have been included in version 1.2.

[4] A more detailed overview of model performance is available at: https://ufal.mff.cuni.cz/udpipe/2/models

| Sentence ID | Text |
|---|---|
| en.4 | More than 7,000 people visited the film's premiere in Damascus. |
| es.4 | A la presentación del documental en Damasco asistieron más de 7000 personas. |
| et.4 | Rohkem kui 7000 inimest külastas Damaskuses filmi esilinastust. |
| nl.4 | Meer dan 7.000 mensen bezochten de première van de film in Damascus. |

Table 2: Examples of parallel sentences for English, Spanish, Estonian, and Dutch from ELEXIS-WSD 1.2.

| Model | UAS | LAS | MLAS | BLEX |
|---|---|---|---|---|
| Bulgarian (bulgarian-btb-ud-2.15-241121) | 95.31 | 92.57 | 86.55 | 87.25 |
| Danish (danish-ddt-ud-2.15-241121) | 89.97 | 87.93 | 80.65 | 82.80 |
| Dutch (dutch-alpino-ud-2.15-241121) | 94.92 | 92.86 | 86.60 | 83.78 |
| English (english-ewt-ud-2.15-241121) | 93.42 | 91.52 | 85.10 | 86.21 |
| Hungarian (hungarian-szeged-ud-2.15-241121) | 88.70 | 85.08 | 75.20 | 78.33 |
| Italian (italian-isdt-ud-2.15-241121) | 95.08 | 93.39 | 87.08 | 88.14 |
| Portuguese (portuguese-bosque-ud-2.15-241121) | 93.46 | 91.08 | 81.78 | 85.74 |
| Spanish (spanish-ancora-ud-2.15-241121) | 94.00 | 92.35 | 87.30 | 88.85 |

Table 3: F1 scores of UDPipe models used to annotate ELEXIS-WSD 1.2.

for contrastive cross-lingual syntactic analyses. Because it is a parallel corpus, the included sentences are directly comparable in terms of content and genre. In the following sections, we perform several statistical comparisons to demonstrate the uses of our method for insights into syntactic differences between languages.

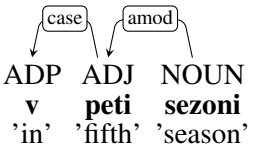


Figure 1: An example of a syntactic tree extracted from the Slovene subcorpus and corresponding to the structure *ADP <case ADJ <amod NOUN*.

## 4 Extraction of Syntactic Profiles

We prepare the data for statistical analysis by extracting syntactic profiles of individual subcorpora as well as individual sentences from ELEXIS-WSD. We define a syntactic profile of a unit as a numerical vector of relative frequencies of various syntactic features extracted from the unit. In this paper, we focus on features representing the relative frequencies of different syntactic trees and subtrees in different units. We extract the frequencies using *STARK 3.0* (Krsnik et al., 2024), an open-access dependency-tree extraction tool available under the Apache 2.0 license. STARK takes a CoNLL-U file with syntactic annotations as input and, based on several customizable parameters, outputs a frequency list of syntactic structures (trees) represented with the simple *dep_search* query language.[5] An example is shown in Figure 1.

Depending on the settings, the frequency list contains absolute and relative frequencies of syntactic structures (normalized by the number of tokens in the extracted unit per million).

Instead of feeding entire subcorpora to STARK, we first split the files into individual sentences and performed the extraction[6] on each sentence individually. From each sentence, we extracted complete syntactic (sub)trees encompassing the head and all its (in)direct dependants, as well as the order of the dependants. A sample of extracted (sub)trees is shown in Table 4.

After extracting syntactic (sub)trees from all sentences, we removed the structures occurring less than 3 times throughout the entire corpus and ended up with a set of 2,582 distinct (sub)trees. These were used as features for the numerical vectors representing the syntactic profile of each sentence. For each sentence $s$, its syntactic profile is compiled by

[5] A more detailed overview of the *dep_search* query language is available at: https://orodja.cjvt.si/drevesnik/help/en/

[6] We used STARK 3.0 (commit 'bed75dc' on GitHub): https://github.com/clarinsi/STARK. The following parameters were used: size="2-10000", processing_size=None, complete="yes", labeled="yes", fixed="yes", node_type="upos", example="yes", detailed_results_file="(path to file with detailed results)". The rest of the parameters (apart from the obligatory 'input', 'output', and 'config_file') were set to None.

concatenating the relative frequencies (within $s$) of each tree $t$ from the set of $n$ distinct (sub)trees: $s = [f_r(t_1), f_r(t_2), f_r(t_3), ..., f_r(t_n)]$. In our case, this generated a 20,240 x 2,582 matrix that was used for statistical comparisons (see Section 5). An additional 10 x 2,582 matrix of syntactic profiles was compiled for individual subcorpora, consisting of the means of relative frequencies of each syntactic tree.

## 5 Global Feature Analysis

### 5.1 Syntactic Profiles of Subcorpora

We first performed an analysis to compare the syntactic profiles of the individual subcorpora. Due to the limited size of the corpus, we first observed whether a bird's-eye view of the extracted corpus vectors revealed any expected differences and similarities between languages in order to confirm that it was sensible to continue with sentence-level comparisons. If the differences between corpus-level syntactic profiles had been completely random, further analyses on sentence-levels.

We performed multiple instances of $k$-means clustering[7] on the syntactic profiles of subcorpora and calculated the silhouette score[8]) to determine the optimal $k$, i.e. the most sensible division of groups by similarity between syntactic profiles. The silhouette scores for different cluster numbers are shown in Table 5.

The optimal number of clusters (4) divides the languages in the following manner: Cluster 1 – Hungarian; Cluster 2 – English, Dutch, Spanish, Italian, Portuguese; Cluster 3 – Bulgarian, Slovene, Danish; Cluster 4 – Estonian. We visualized the syntactic profiles using multidimensional scaling (MDS)[9] (see Figure 2). With some exceptions (like Danish being clustered with Bulgarian and Slovene despite its proximity to English and Dutch; and English and Dutch being grouped together with the Romance languages), the division is largely expected and follows the distinction between Romance, Germanic, and Slavic languages, with Hungarian and Estonian as separate clusters.

The differences and similarities between languages are more accurately represented with cosine similarity scores ($sim$) calculated based on the subcorpora's syntactic profiles (see Figure 3). The highest similarity can be observed between the three Romance languages ($0.94 \leq sim \leq 0.96$) and between Dutch and English ($sim = 0.93$). In terms of the distribution of syntactic structures, Danish indeed seems to be more similar to Bulgarian ($sim = 0.83$) and Slovene ($sim = 0.76$) than to Dutch ($sim = 0.73$). This outcome is not entirely intuitive and warrants further research and a more detailed comparison of syntactic (sub)trees. When interpreting the results, it should also be taken into account that most of the subcorpora were parsed automatically, so the comparison of distributions of syntactic structures should be conducted a second time once the data has been manually validated, or cross-referenced with results from comparisons between relevant UD treebanks. This is beyond the scope of this paper, but we focus on a number of differences between corpora in terms of specific syntactic (sub)trees in the following sections.

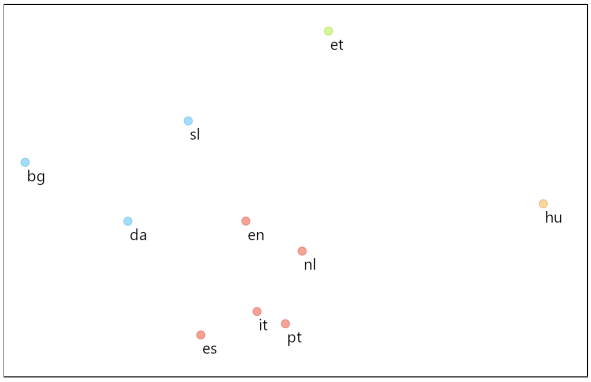


Figure 2: MDS Visualization of the Syntactic Profiles of the ELEXIS-WSD 1.2 Subcorpora.

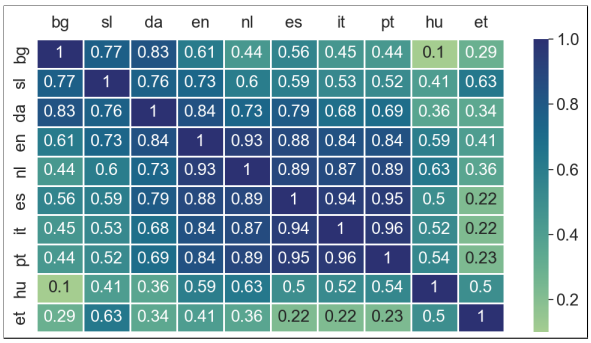


Figure 3: Matrix of cosine similarities between the syntactic profiles of individual ELEXIS-WSD subcorpora.

[7] $k$-means clustering was performed using the Scikit-Learn Python package (Pedregosa et al., 2011).

[8] The silhouette score was calculated taking into account the Euclidean distance using the *scikit-learn* Python package: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

[9] MDS was performed using Orange Data Mining v3.38.0 (Demšar et al., 2013).

| Tree | Order | Nodes | Head | Example |
|---|---|---|---|---|
| (DET <det NOUN >case PART) <nmod NOUN | ABCD | 4 | NOUN | the film's premiere |
| ADJ >fixed ADP | AB | 2 | ADJ | More than |
| ADP <case PROPN | AB | 2 | PROPN | in Damascus |
| DET <det NOUN >case PART | ABC | 3 | NOUN | the film's |
| ((ADJ >fixed ADP) <advmod NUM) <nummod NOUN | ABCD | 4 | NOUN | More than 7,000 people |
| (ADJ >fixed ADP) <advmod NUM | ABC | 3 | NUM | More than 7,000 |

Table 4: A sample of syntactic (sub)trees and their frequencies extracted from the en.4 English sentence using STARK 3.0; all have an $f_a$ = 1 and $f_r$ = 83,333.3.

| Clusters | Silhouette Score |
|---|---|
| 2 | 0.246 |
| 3 | 0.278 |
| **4** | **0.294** |
| 5 | 0.236 |
| 6 | 0.172 |
| 7 | 0.163 |
| 8 | 0.123 |
| 9 | 0.058 |

Table 5: Silhouette scores for different numbers of clusters in $k$-means clustering of ELEXIS-WSD subcorpora.

### 5.2 Syntactic Profiles of Sentences

To delve deeper into the syntactic differences between corpora, we performed the Kruskal–Wallis H test[10] (Kruskal and Wallis, 1952) ($k = 10$, $n = 20,240$) to determine statistically significant differences in the distribution of the 2,582 extracted syntactic (sub)trees. For 1,712 (sub)trees, the difference in distribution is statistically significant ($p \leq 0.05$), but only 756 (29%) pass the Bonferroni correction[11] (at $p \leq 1.936e-05$). The results of the test with the highest effect sizes[12] are shown in Table 6.

Some of the outcomes are expected, as several of the top 10 syntactic (sub)trees with the highest differences in distribution point out the more directly obvious differences between languages. For instance, several of the syntactic (sub)trees contain determiners, which are much less frequent in Slovene and Bulgarian compared to English, Dutch, and the three Romance languages. Although the overall results are promising and show that more detailed comparisons of syntactic tree distributions should be made, we limit our analysis to a handful of the most statistically significant differences due to space limitations. We describe them in the following sections.

## 6 Statistical Analysis of Selected Features

To determine in which specific languages the differences in frequencies of a given syntactic tree are statistically significant, we performed a series of pair-wise Mann–Whitney U tests (Mann and Whitney, 1947) with Bonferroni correction (at $p < 0.001$).[13] The effect sizes were measured with the rank-biserial correlation coefficient ($r$) (Cureton, 1956).[14]

### 6.1 ADJ <amod NOUN – AB

The structure *ADJ <amod NOUN – AB* refers to a noun modified by an adjective on the left (e.g. *immediate fame*). The results of the test confirm that the syntactic structure is notably less frequent in Spanish, Italian, and Portuguese compared to the other languages, with more significant differences when comparing to Estonian, Hungarian, Bulgarian, and Slovene. The most noticeable difference is between Estonian and Portuguese ($k = 2$, $n = 4,048$, $n_1 = n_2 = 2,024$, $U_1 = 2,623,642.5$, $p \leq 0.0001$, $r = 0.28$). This is an expected outcome; the Romance languages

[10] We opted for the non-parametric Kruskal-Wallis H test because of the non-normal distributions for the vast majority of extracted syntactic (sub)trees. A statistically significant result reveals that at least one of the groups that are being compared stochastically dominates at least one other group. The differences are then further inspected with additional statistical tests (see Section 6).

[11] Due to the limited size of the corpus, we opted for the more conservative Bonferroni correction method as opposed to other less restrictive methods (e.g., Holm-Bonferroni method or the Benjamini-Hochberg prodecure).

[12] Effect size was calculated as $\eta^2 = (H - k + 1)/(n - k)$, as reported in (Tomczak and Tomczak, 2014). The $\eta^2$ effect size ranges from 0 to 1, and multiplied by 100% indicates the percentage of variance in the dependent variable explained by the independent variable.

[13] Again, we opted for the non-parametric Mann-Whitney U test because the distribution of relative frequencies is not normal for the majority of syntactic (sub)trees.

[14] The rank-biserial correlation coefficient is a value between $-1$ and $+1$, with a value of zero indicating no relationship.

| Tree and Node Order | $f_a$ | $H$ | $p$ | $\eta^2$ |
|---|---|---|---|---|
| ADP <case DET <det NOUN – ABC | 3,191 | 2,461.36 | $p \leq 0.0001$ | 0.121 |
| DET <det NOUN – AB | 5,063 | 2,060.96 | $p \leq 0.0001$ | 0.101 |
| ADP <case NUM <amod NOUN – ABC | 243 | 2,008.85 | $p \leq 0.0001$ | 0.099 |
| ADJ <amod NOUN – AB | 2,225 | 1,984.65 | $p \leq 0.0001$ | 0.098 |
| ADP <case NOUN – AB | 4,996 | 1,888.27 | $p \leq 0.0001$ | 0.093 |
| PROPN <nmod NOUN – AB | 427 | 1,867.59 | $p \leq 0.0001$ | 0.092 |
| NOUN <nmod NOUN – AB | 323 | 1,671.90 | $p \leq 0.0001$ | 0,082 |
| ADP <case DET <det NUM – ABC | 185 | 1,530.56 | $p \leq 0.0001$ | 0.075 |
| ADP <case ADJ <amod NOUN – ABC | 1,366 | 1,471.42 | $p \leq 0.0001$ | 0.072 |
| DET <det ADJ <amod NOUN – ABC | 1,109 | 1,430.00 | $p \leq 0.0001$ | 0.070 |

Table 6: Top 10 syntactic (sub)trees with the most significant differences in distributions according to the Kruskal-Wallis H test.

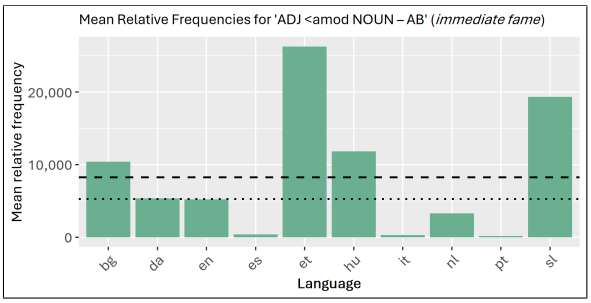


Figure 4: Mean relative frequencies (per million words) for the structure *ADJ <amod NOUN – AB*.

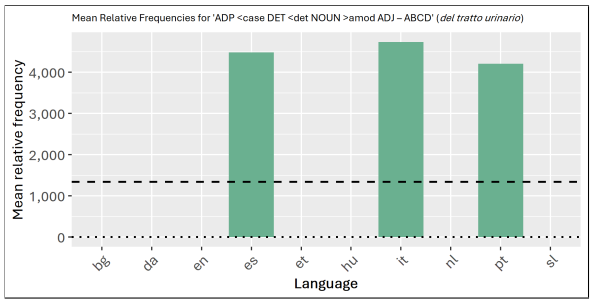


Figure 5: Mean relative frequencies (per million words) for the structure *ADP <case DET <det NOUN >amod ADJ – ABCD*.

usually modify their nouns with an adjective on the right and typically also include a determiner (see Section 6.2 for an analysis of a similar structure). A barplot of mean frequencies is shown in Figure 4, with the dashed line representing the global mean and dotted line the global median value.

### 6.2 ADP <case DET <det NOUN >amod ADJ – ABCD

On the other hand, the structure *ADP <case DET <det NOUN >amod ADJ – ABCD* (e.g. *del (di + il) tratto urinario* 'of the urinary tract' in Italian), which contains a noun modified by an adjective to the right, is much more typical of the Romance languages and is in fact completely absent in the rest (see Figure 5). The most statistically significant and largest difference is between Italian and Dutch ($k = 2$, $n = 4,048$, $n_1 = n_2 = 2,024$, $U_1 = 2,222,352.0$, $p \leq 0.0001$, $r = 0.085$); the same difference can also be observed between Italian and Slovene, while similar differences are confirmed for pairs that include other Romance languages, such as Spanish-Estonian ($U_1 = 2,213,244.0$, $p \leq 0.0001$, $r = 0.085$) and Portuguese-Slovene ($U_1 = 2,203,124.0$, $p \leq 0.0001$, $r = 0.076$).

### 6.3 NOUN >nummod NUM – AB

The structure *NOUN >nummod NUM – AB* (e.g. *junija 2014* 'in June of 2024' in Slovene; *juunis 2014* in Estonian) seems to be much more frequent in Slovene compared to the other languages in the corpus (see Figure 6). The difference is confirmed by the pair-wise Mann–Whitney U tests, which find statistically significant differences between Slovene and all other languages, with the highest difference between Slovene and Portuguese/Italian/Spanish/Hungarian/Danish on the one hand and Slovene on the other (for all these comparisons: $k = 2$, $n = 4,048$, $n_1 = n_2 = 2,024$, $U_1 = 1,891,428.0$, $p \leq 0.0001$, $r = 0.077$). Statistically significant differences can also be found between Estonian and e.g. Hungarian/Italian/Portuguese/Dutch, but the effect sizes are smaller ($r = 0.015$).

## 7 Conclusion and Future Work

We have presented the latest version of the ELEXIS-WSD parallel corpus, which also contains UD dependency relations and can be used

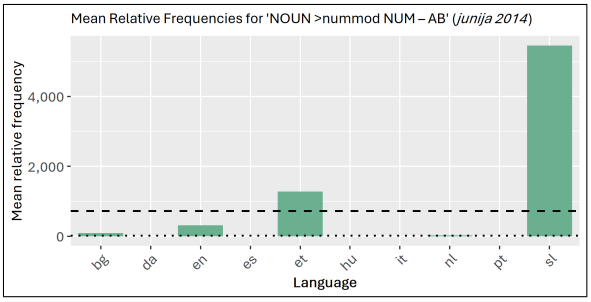


Figure 6: Mean relative frequencies (per million words) for the structure *NOUN >nummod NUM – AB*.

as an additional resource for studies in quantitative syntax alongside the many existing UD treebanks, including parallel UD treebanks.[15] We have also presented a method to observe the differences in the distribution of syntactic (sub)trees between corpora by using the STARK 3.0 dependency-tree extraction tool. While we showcased the method on a parallel corpus, it can also be used to compare e.g. two corpora in the same language (e.g. a spoken and a written corpus; a learner vs. a general corpus) to determine the most salient differences in syntactic structures. In addition to contrastive syntactic comparisons, the method could also provide a basis for several other uses. First, by generating quantified syntactic profiles of sentences in a corpus, groups of syntactically similar sentences can be extracted by exporting clusters with high cosine similarity scores compared to a reference sentence. Second, the method could be used to compare whether (and to what degree) a sampled corpus is syntactically representative of the whole. On the other hand, the method can help extract syntactically diverse samples to ensure as many syntactic structures are included as possible.

However, there are potential challenges with the scalability of the method. In this paper, we have limited the extraction of syntactic profiles to only complete syntactic (sub)trees. Extracting all parts of syntactic (sub)trees would help provide a more accurate profile, but would also be much more computationally expensive. During our tests, extracting partial and full syntactic profiles resulted in approximately 2kB vs. 10MB of data per sentence, respectively. More tests are required to compare which (additional) features are best at representing the syntactic characteristics of the remaining links not extracted when focusing solely on complete syntactic (sub)trees.

In the future, we intend to publish new versions of the ELEXIS-WSD corpus within the UniDive COST Action. On the one hand, the corpus will be extended with subcorpora for new languages, and the dependency relation annotations for more of the existing corpora will be manually validated. The corpora will eventually also contain several other annotations that can be cross-compared with syntax, such as sense-, named entity-, and multiword expression annotations.

Once the corpus is fully manually annotated, the parallel alignment of sentences will allow for an even more direct comparison of syntactic structures. Exporting co-occurrences of syntactic (sub)trees between equivalent sentences from different languages will enable us to observe the most frequently or typically co-occurring structures (by calculating association measures such as pointwise mutual information (Church and Hanks, 1990)).

The next step should also involve extending the method of extracting syntactic profiles by including e.g. combinations of syntactic structures and additional quantitative features, such as direction, frequency, and depth of individual dependency relations (or combinations thereof), which have been shown to be effective at representing certain aspects of syntactic complexity (see e.g. Terčon, 2024) and can be easily extracted with recently developed tools such as *ComparaTree* (Terčon and Dobrovoljc, 2025) and *QuanSyn* (Yang and Liu, 2025). These options will be explored in future studies, in which the method will also be tested on other corpora.

## Limitations

The research in this paper has required no ethical considerations. In terms of limitations, it should be noted that many texts from the corpus were missing from Wikimatrix and were machine-translated, then corrected manually at a later stage. The translations are thus not entirely manual, and syntactic structures have been influenced by the decisions of the machine translation systems used for different languages. In addition, all texts were translated from English, so some English influence can also be expected. Most of the subcorpora were parsed automatically (only Slovene and Estonian have been manually validated so far), so the corpus cannot be considered a gold-standard dataset

[15] See e.g. Polish-PUD: https://github.com/UniversalDependencies/UD_Polish-PUD; and English-PUD https://github.com/UniversalDependencies/UD_English-PUD.

in terms of dependency relations. All sentences are taken from WikiPedia, so the corpus is also biased in terms of genre. Lastly, in this paper, syntactic profiles only take into account distributions of syntactic (sub)trees, while many other syntactic features could be taken into account as well to better represent the wide range of syntactic characteristics present in all subcorpora.

## Acknowledgments

The work presented in the paper was supported by the COST Action CA21167 – *Universality, Diversity and Idiosyncrasy in Language Technology* (UniDive). The author also acknowledges the financial support from the Slovenian Research and Innovation Agency (research core funding No. P6-0411 – *Language Resources and Technologies for Slovene*) and thanks the anonymous reviewers for their constructive comments.

## References

Hee-Soo Choi, Bruno Guillaume, and Karën Fort. 2021. Corpus-based language universals analysis using universal dependencies. In *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, page 33–44.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Jaka Čibej, Simon Krek, Carole Tiberius, Federico Martelli, Roberto Navigli, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, and 23 others. 2025. Parallel sense-annotated corpus ELEXIS-WSD 1.2. Slovenian language resource repository CLARIN.SI.

Edward E. Cureton. 1956. Rank-biserial correlation. *Psychometrika*, 21(3):287–290.

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

Janez Demšar, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, Miha Štajdohar, Lan Umek, Lan Žagar, Jure Žbontar, Marinka Žitnik, and Blaž Zupan. 2013. Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, 14:2349–2353.

Edward Gibson and Evelina Fedorenko. 2010. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, pages 1–37.

Edward Gibson, Steven T. Piantandosi, and Evelina Fedorenko. 2012. Quantitative methods in syntax/semantics research: A response to sprouse and almeida (2012). *Language and Cognitive Processes*, pages 1–12.

Edward S. Klyshinsky and O.V. Karpik. 2019. Quantitative evaluation of syntax similarity. *Mathematica Montisnigri, Vol XLVI*, pages 123–132.

Luka Krsnik and Kaja Dobrovoljc. 2025. Stark: A toolkit for dependency (sub)tree extraction and analysis. In *SyntaxFest 2025*, Ljubljana, Slovenia.

Luka Krsnik, Kaja Dobrovoljc, and Marko Robnik-Šikonja. 2024. Dependency tree extraction tool STARK 3.0. Slovenian language resource repository CLARIN.SI.

William H. Kruskal and W. Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621.

H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60.

Federico Martelli, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña Ruiz, José Luis Sancho Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, Simon László, and Tina Munda. 2021. Designing the elexis parallel sense-annotated dataset in 10 european languages. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference*, pages 377–395.

Ruochen Niu, Yaqin Wang, and Haitao Liu. 2021. The properties of rare and complex syntactic constructions in english: A corpus-based comparative study. In *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, page 74–83.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Johanna M. Poppek, Simon Masloch, Amelie Robrecht, and Tibor Kiss. 2021. A quantitative approach towards german experiencer-object verbs. In *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, page 84–91.

Agata Savary, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz, and Alina Wróblewska. 2024. UniDive: A COST action on universality, diversity and idiosyncrasy in language technology. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Underresourced Languages @ LREC-COLING 2024*, pages 372–382, Torino, Italia. ELRA and ICCL.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka, Jan Hajič, 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'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Luka Terčon and Kaja Dobrovoljc. 2025. Comparatree: A multi-level comparative treebank analysis tool. In *SyntaxFest 2025*, Ljubljana, Slovenia.

Luka Terčon. 2024. Uporaba šestih mer skladenjske kompleksnosti za primerjavo jezika v govornem in pisnem korpusu. In *Proceedings of the Language Technologies and Digital Humanities Conference 2024*, page 668–686.

Carole Tiberius, Jaka Čibej, Jelena Kallas, Kertu Saul, Kadri Muischnek, and Simon Krek Krek. 2024. Ud syntax for the elexis-wsd parallel sense-annotated corpus: A pilot study. In *UniDive 2nd General Meeting (Naples, Italy)*, Naples, Italy.

Maciej Tomczak and Ewa Tomczak. 2014. The need to report effect size estimates revisited. an overview of some recommended measures of effect size. *Trends in Sport Sciences*, 1(21):19–25.

Jeroen van Craenenbroeck, Marjo van Koppen, and Antal van den Bosch. 2019. A quantitative-theoretical analysis of syntactic microvariation: Word order in dutch verb clusters. *Language 95, no. 2*, pages 333–370.

Mu Yang and Haitao Liu. 2025. Quansyn: A package for quantitative syntax analysis. *Journal of Quantitative Linguistics*, 32(2):1–18.