Quasy 2025

Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025)

Proceedings

August 29, 2025

The Quasy organizers gratefully acknowledge the support from the following sponsors.

Mestna občina
Ljubljana

Organized by

As part of SyntaxFest 2025

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-293-0

Introduction

Quantitative Syntax (QUASY) is a workshop series dedicated to advancing quantitative, statistical, and computational methods in syntactic research. The workshops bring together researchers to discuss and explore quantitative, statistical, and computational methods in syntax research, responding to the growing need for linguistic meetings that focus on empirical data-driven approaches to syntactic theory. Since the first QUASY workshop in 2019 in Paris, the series has brought together researchers working at the intersection of syntax, corpus linguistics, computational methods, and related empirical approaches. This year's workshop, the third QUASY 2025, is held as part of SyntaxFest 2025 in Ljubljana, Slovenia, which brings together five related but independent events:

- 18th International Conference on Parsing Technologies (IWPT 2025)
- 8th Universal Dependencies Workshop (UDW 2025)
- 8th International Conference on Dependency Linguistics (DepLing 2025)
- 23rd Workshop on Treebanks and Linguistic Theories (TLT 2025)
- 3rd Workshop on Quantitative Syntax (QUASY 2025)

In addition, a pre-conference workshop organized by the COST Action CA21167 – Universality, Diversity and Idiosyncrasy in Language Technology (UniDive) was held prior to the main event, with dedicated sessions on the 1st UniDive Shared Task on Morphosyntactic Parsing and the 2nd Workshop on Universal Dependencies for Turkic Languages.

SyntaxFest 2025 continues the tradition of SyntaxFest 2019 (Paris, France), SyntaxFest 2021 (Sofia, Bulgaria), and GURT/SyntaxFest 2023 (Washington DC, USA) in bringing together multiple events that share a common interest in using corpora and treebanks for empirically validating syntactic theories, studying syntax from quantitative and theoretical points of view, and training machine learning models for natural language processing. Much of this research is increasingly multilingual and cross-lingual and requires continued systematic analysis from various theoretical, applied, and practical perspectives. By co-locating these workshops under a shared umbrella, SyntaxFest fosters dialogue between overlapping research communities and supports innovation at the intersection of linguistics and language technology.

As in previous editions, all five workshops at SyntaxFest 2025 shared a common submission and reviewing process, with a unified timeline, identical submission formats, and a shared program committee. During submission, authors could indicate one or more preferred venues, but the final assignment of papers was determined by the collective program chairs, composed of the individual workshop chairs, based on thematic alignment. All accepted submissions were peer-reviewed by at least three reviewers from the shared program committee.

In total, SyntaxFest 2025 received 94 submissions, of which 73 (78%) were accepted for presentation. The final program included a total of 47 long papers, 21 short papers, and 5 non-archival contributions, distributed across the five workshops: 5 papers were presented at IWPT (2 long, 3 short); 20 at UDW (14 long, 5 short, 1 non-archival); 16 at DepLing (12 long, 2 short, 2 non-archival); 18 at TLT (10 long, 7 short, 1 non-archival); and 14 at QUASY (9 long, 4 short, 1 non-archival).

Our sincere thanks go to everyone who made this event possible. We thank all authors for their submissions and the reviewers for their time and thoughtful feedback, which contributed to a diverse and high-quality program. Special thanks go to the local organizing team at the University of Ljubljana and the Slovene Language Technologies Society for hosting the event, and to the sponsors for their generous support. Finally, we gratefully acknowledge ACL SIGPARSE for endorsing the event and the ACL Anthology for publishing the proceedings.

Kenji Sagae, Stephan Oepen (IWPT 2025 Chairs)
Gosse Bomma, Çağrı Çöltekin (UDW 2025 Chairs)

Eva Hajičová, Sylvain Kahane (DepLing 2025 Chairs)
Heike Zinsmeister, Sarah Jablotschkin, Sandra Kübler (TLT 2025 Chairs)
Xinying Chen, Yaqin Wang (QUASY 2025 Chairs)
Kaja Dobrovoljc (SyntaxFest 2025 Organization Chair)

Ljubljana, August 2025

Organizing Committee

TLT Chairs

Heike Zinsmeister, University of Hamburg
Sarah Jablotschkin, University of Hamburg
Sandra Kübler, Indiana University

DepLing Chairs

Eva Hajičová, Charles University, Prague
Sylvain Kahane, Université Paris Nanterre

UDW Chairs

Gosse Bomma, University of Groningen
Çağrı Çöltekin, University of Tübingen

IWPT Chairs

Kenji Sagae, University of California, Davis
Stephan Oepen, University of Oslo

QUASY Chairs

Xinying Chen, University of Ostrava
Yaqin Wang, Guangdong University of Foreign Studies

Publication Chair

Sarah Jablotschkin, University of Hamburg

Local SyntaxFest 2025 Organizing Committee

Kaja Dobrovoljc, University of Ljubljana, SDJT
Špela Arhar Holdt, University of Ljubljana
Luka Terčon, University of Ljubljana
Marko Robnik-Šikonja, University of Ljubljana
Matej Klemen, University of Ljubljana
Sara Kos, University of Ljubljana
Timotej Knez, University of Ljubljana, SDJT
Tinca Lukan, University of Ljubljana

Special Thanks for designing the SyntaxFest 2025 logo to

Kim Gerdes, Université Paris-Saclay

Program Committee

Shared Program Committee

V.S.D.S.Mahesh Akavarapu, Eberhard-Karls-Universität Tübingen
Leonel Figueiredo de Alencar, Federal University of Ceará (UFC)
Patricia Amaral, Indiana University
Giuseppe Attardi, University of Pisa
John Bauer, Stanford University
David Beck, University of Alberta
Laura Becker, Albert-Ludwigs-Universität Freiburg
Aleksandrs Berdicevskis, Gothenburg University
Ann Bies, University of Pennsylvania
Igor Boguslavsky, Universidad Politécnica de Madrid
Bernd Bohnet, Google
Cristina Bosco, University of Turin
Gosse Bouma, University of Groningen
Miriam Butt, Universität Konstanz
G. A. Celano, Universität Leipzig
Heng Chen, Guangdong University of Foreign Studies
Xinying Chen, University of Ostrava
Jinho D. Choi, Emory University
Çağrı Çöltekin, University of Tuebingen
Daniel Dakota, Leidos
Stefania Degaetano-Ortlieb, Universität des Saarlandes
Kaja Dobrovoljc, University of Ljubljana
Jakub Dotlacil, Utrecht University
Gülşen Eryiğit, Istanbul Technical University
Kilian Evang, Heinrich Heine University Düsseldorf
Pegah Faghiri, CNRS
Ramon Ferrer-i-Cancho, Universidad Politécnica de Cataluna
Marcos Garcia, Universidade de Santiago de Compostela
Kim Gerdes, Université Paris-Saclay
Loïc Grobol, Université Paris Nanterre
Bruno Guillaume, INRIA
Carlos Gómez-Rodríguez, Universidade da Coruña
Eva Hajicova, Charles University
Dag Trygve Truslew Haug, University of Oslo
Santiago Herrera, University of Paris Nanterre
Richard Hudson, University College London
Maarten Janssen, Charles University Prague
Jingyang Jiang, Zhejiang University
Mayank Jobanputra, Universität des Saarlandes
Sylvain Kahane, Université Paris Nanterre
Václava Kettnerová, Charles University Prague
Sandra Kübler, Indiana University
Guy Lapalme, University of Montreal
François Lareau, Université de Montréal
Miryam de Lhoneux, KU Leuven
Zoey Liu, University of Florida

Teresa Lynn, Dublin City University
Jan Macutek, Slovak Academy of Sciences
Robert Malouf, San Diego State University
Marie-Catherine de Marneffe, UCLouvain
Nicolas Mazziotta, Université de Liège
Alexander Mehler, Johann Wolfgang Goethe Universität Frankfurt am Main
Maitrey Mehta, University of Utah
Wolfgang Menzel, Universität Hamburg
Marie Mikulová, Charles University
Aleksandra Miletić, University of Helsinki
Jasmina Milićević, Dalhousie University
Simon Mille, Dublin City University
Yusuke Miyao, The University of Tokyo
Noor Abo Mokh, Indiana University
Simonetta Montemagni, Institute for Computational Linguistics "A. Zampolli" (ILC-CNR)
Jiří Mírovský, Charles University Prague
Kaili Müürisep, Institute of computer science, University of Tartu
Anna Nedoluzhko, Charles University Prague
Ruochen Niu, Beijing Language and Culture University
Joakim Nivre, Uppsala University
Stephan Oepen, University of Oslo
Timothy John Osborne, Zhejiang University
Petya Osenova, Sofia University "St. Kliment Ohridski"
Agnieszka Patejuk, Polish Academy of Sciences
Lucie Poláková, Charles University Prague
Prokopis Prokopidis, Athena Research Center
Mathilde Regnault, Universität Stuttgart
Kateřina Rysová, University of South Bohemia
Magdaléna Rysová, Charles University Prague
Tanja Samardzic, University of Zurich
Giuseppe Samo, Beijing Language and Culture University
Haruko Sanada, Rissho University
Nathan Schneider, Georgetown University
Djamé Seddah, Sorbonne University
Anastasia Shimorina, Orange
Maria Simi, University of Pisa
Achim Stein, University of Stuttgart
Daniel G. Swanson, Indiana University
Luka Terčon, Faculty of Arts, University of Ljubljana
Giulia Venturi, Institute for Computational Linguistics "A. Zampolli" (ILC-CNR)
Veronika Vincze, University of Szeged
Yaqin Wang, Guangdong University of Foreign Studies
Pan Xiaxing, Huaqiao University
Chunshan Xu, Anhui Jianzhu University
Nianwen Xue, Brandeis University
Jianwei Yan, Zhejiang University
Zdenek Zabokrtsky, Faculty of Mathematics and Physics, Charles University Prague
Eva Zehentner, University of Zurich
Amir Zeldes, Georgetown University
Daniel Zeman, Charles University Prague
Šárka Zikánová, Charles University Prague

Heike Zinsmeister, Universität Hamburg

Keynote

The rhetorical and pragmatic functions of syntactically complex structures in academic and second language writing

Xiaofei Lu
The Pennsylvania State University

Abstract: Previous studies of linguistic complexity in academic and second language (L2) writing has often focused on quantitative differences across different writer groups and/or longitudinal changes over time, without systematic attention to the rhetorical or pragmatic functions that complex forms are used to convey. This talk argues for the importance of and delineates the scope of the function dimension of linguistic complexity analysis in L2 writing research, reviews the methods and findings of emerging efforts on this dimension, and discusses how future L2 writing research could attend to this dimension.

Bio: Xiaofei Lu is the George C. and Jane G. Greer Professor of Applied Linguistics and Asian Studies at The Pennsylvania State University. His research has long centered on computational and quantitative analyses of linguistic complexity in reading materials, second language production, and academic writing. His current work explores mappings between linguistic forms and rhetorical/pragmatic functions in language production and sense-aware measurements of linguistic complexity that account for the specific meanings of polysemous linguistic forms in context. He has published over 90 peer-reviewed articles in leading journals, including Applied Linguistics, Behavior Research Methods, Computer Assisted Language Learning, Language Learning, Studies in Second Language Acquisition, TESOL Quarterly, and The Modern Language Journal. He received the 2023 Ken Hyland Best Paper Award from the Journal of English for Academic Purposes. His latest book, Corpus Linguistics and Second Language Acquisition: Perspective, Issues, and Findings, was published by Routledge in 2023.

Non-Archival Abstract

Syntactic Complexity and News Credibility in Czech Media

Miroslav Kubát, Xinying Chen, Michaela Nogolová and Michal Místecký
University of Ostrava

This study examines how syntactic complexity varies across news articles differing in credibility, using a Czech-language corpus annotated with five credibility levels: credible, partially credible, misleading, manipulative, and unclassifiable. We apply a dependency parsing pipeline and compute five syntactic metrics measuring features such as sentence length, clause density, and hierarchical depth. Results show that manipulative texts are structurally the most complex, while misleading and unclassifiable texts are simpler and more fragmented. Credible texts display balanced complexity consistent with journalistic norms. These findings highlight the role of syntax in shaping rhetorical strategies and contribute to the linguistic understanding of news credibility.

Table of Contents

Subject-Verb Agreement Alternations in Spanish Pseudopartitive Constructions: A Corpus Study

Marina Cerebrinsky
Bar-Ilan University - marinacere@hotmail.com

Abstract

Pseudopartitive constructions, following the format N1-*of*-N2 (such as *a group of students*), are known to feature alternations in their subject-verb agreement patterns, either with the N1 or the N2. Through a Spanish corpus analysis, this study investigates the possibility of a correlation between the choice of N1/N2 as an agreement trigger and the semantic type of the N1, as well as the animacy status of the N2. Although a positive correlation was found for N1 semantic type, no statistically significant results emerged for N2 animacy.

1 Introduction

The present paper deals with subject-verb agreement alternations in pseudopartitive constructions in Spanish. Pseudopartitive constructions—such as *a group of students*—are structures of the form N1-*of*-N2, where a bare noun phrase (N2; *students*) is quantified or measured by a noun functioning as N1 (*group*), typically denoting quantity, collectivity, or containment (Milner, 1978; Schwarzschild, 2006). Although this description is based on English, analogous structures exist cross-linguistically, including in German (Grestenberger, 2015), Romanian (Cornilescu, 2009), Finnish (Huumo, 2018), Hebrew (Shatil, 2015), and Spanish (Demonte and Pérez Jiménez, 2015). These constructions contrast with partitive structures, where the N1 selects a subset out of a specific set, which is preceded by a determiner (Chierchia, 1998; Zamparelli, 2008).

At a first glance, pseudopartitive constructions may seem indistinguishable from other (det)-N1-*of*-N2 genitive structures—such as *a portrait of children*—since both follow the same sequence of word classes. However, one key distinction is that pseudopartitive constructions are known to feature two different verb agreement patterns: 1. with the N1; or 2. with the embedded N2 (Foppolo et al., 2023). See (1) for an example in which subject-verb agreement is possible either with the N1 (*group*) or the N2 (*researchers*).

(1) a. A group of researchers is analyzing this phenomenon.
b. A group of researchers are analyzing this phenomenon.

The present study investigates the possibility of a correlation between the choice for either N1 or N2 subject-verb agreement and the semantic type of the N1; and between the choice for either N1 or N2 subject-verb agreement and N2 animacy. It does so through the analysis of 1,200 occurrences of pseudopartitive subject-verb agreement in subject position, present in the Spanish-language esTenTen18 corpus (Kilgarriff and Renau, 2013), available on Sketch Engine (Kilgarriff et al., 2014). Even though previous studies have been conducted to test for a correlation between N1/N2 choice for agreement and the semantic type of the N1 (Demonte and Pérez Jiménez, 2017; Foppolo et al., 2023), no studies have analyzed the role of N2 features in pseudopartitive agreement.

As stated previously, the possibility of dual agreement sets pseudopartitive constructions apart from other superficially similar N1-*of*-N2 sequences. An example showing the impossibility of dual agreement in other seemingly equivalent English N1-*of*-N2 constructions is provided in (2) below. Even though native speakers might produce a sentence like that of (2b), these

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 1–8

instances are considered errors in the psycholinguistic literature, deeming them as cases of agreement attraction (Foppolo et al., 2023).

(2) a. The portrait of children looks beautiful.
b. *The portrait of children look beautiful.

(Pseudo)partitive dual agreement has been explained in terms of structural ambiguity (Selkirk, 1977; Pesetsky, 1982; Franks, 1994), and, more recently, in terms of feature behavior across constituents (Wechsler and Zlatić, 2003; Danon, 2013). Dual agreement has also been attested for (pseudo)partitive constructions in Spanish (Demonte and Pérez Jiménez, 2017), Hebrew (Danon, 2011), Italian (Foppolo et al., 2023), Greek (Stavrou, 2003), and other languages. See (3) for an example in Spanish.

(3) a. Un grupo de investigadores está analizando este fenómeno.
A group.SG of researchers.PL is.SG analyzing this phenomenon.
"A group of researchers is analyzing this phenomenon."

b. Un grupo de investigadores están analizando este fenómeno.
A group.SG of researchers.PL are.PL analyzing this phenomenon.
"A group of researchers are analyzing this phenomenon."

Quantitative studies in the literature have suggested a correlation between subject-verb agreement and the semantic properties of the N1 in pseudopartitive constructions. In a series of three experiments in Italian—including acceptability judgments, production tasks, and eye-tracking—Foppolo et al. (2023) observed that N2 agreement was more likely when the N1 was interpreted primarily as a unit of measurement. For instance, quantifier N1s, which tend to facilitate the sole interpretation of measuring the N2 (e.g., *a lot of students*), tend to display a more balanced distribution in agreement patterns. On the other hand, N1s that have an independent referential meaning (e.g., *a box of chocolates,* in which *box* could denote a unit of measure and an actual cardboard box) strongly favor N1 agreement. Based on their findings, the authors proposed a semantic hierarchy reflecting how easily each semantic type of N1 supports a measure reading: containers allow it the least, followed by collectives, while quantifiers allow it the most.

Demonte and Pérez Jiménez (2015) conducted a corpus-based study of Spanish and observed that certain semantic types of N1s, which they termed "collective numeral nouns"—including expressions like *un centenar de* 'hundreds of' as well as "non-numerical items" like *un montón de* 'a lot of'—tend to facilitate agreement with the embedded noun (N2). Moreover, the authors stated that what they termed "multiplying numeral nouns"—such as *el doble de* 'double of'— tend to facilitate N2 agreement. In contrast, N1s categorized by them as "group nouns" (e.g., *un grupo de* 'a group of,' *una pila de* 'a pile of') and "fixed measure nouns" (e.g., *un kilo de* 'a kilo of') were found to favor agreement with the N1. Their data also showed that constructions headed by "container nouns" exclusively triggered N1 agreement. The authors further investigated whether subject-verb agreement was influenced by the type of determiner preceding the N1, or by the presence of adjectives modifying either noun. In both cases, they found no significant correlation.

The present study builds on prior research by adopting the three-way semantic categorization of N1s—container, collective, and quantifier—proposed by Foppolo et al. (2023), who used it to investigate agreement patterns in Italian pseudopartitive constructions. While this categorization was originally developed within a psycholinguistic framework, applying it to corpus data represents an innovative methodological extension. It allows for the comparison of findings across studies that use distinct methodologies while preserving theoretical consistency. In contrast, previous corpus-based research on Spanish pseudopartitives (such as Demonte and Pérez Jiménez, 2015) employed more fine-grained categorizations of N1s, which, while descriptively rich, pose challenges for cross-linguistic and cross-methodological replicability. By working with a smaller set of broader categories, the present study promotes comparability across languages and approaches. Moreover, corpus linguistics offers the advantage of enabling researchers to efficiently analyze hundreds or thousands of naturally occurring instances of the phenomenon in question.

Although prior studies have explored how N1 semantics may influence agreement patterns, no research to date has systematically examined whether N2 animacy plays a role in agreement alternations within pseudopartitive constructions—or, more broadly, whether any characteristics of the N2 can serve as predictors. Given that animacy is a well-established semantic feature influencing grammatical behavior across languages (Özsoy, 2009; Bresnan and Hay, 2008; Gámez and Vasilyeva, 2015; Bayanati and Toivonen, 2019; Rosenbach, 2008) and has been shown to affect language processing (Vihman and Nelson, 2019; Branigan, Pickering and Tanaka, 2008), it constitutes a strong starting point for investigating whether N2 features impact pseudopartitive agreement. The present study thus introduces a novel dimension by examining whether N2 animacy contributes to subject-verb agreement patterns in Spanish pseudopartitives.

2 Methodology

2.1 Materials

Part of the TenTen corpus family (Suchomel, 2020), esTenTen18 comprises approximately 16.9 billion words sourced from internet texts (Sketch Engine, 2025). It includes a broad range of materials representing both Peninsular and Latin American Spanish varieties with a wide variety of registers. esTenTen18 is tagged morphologically by FreeLing (Padró and Stanilovsky, 2012). Every word in the corpus is tagged based on its part-of-speech and, furthermore, on its morphological features.

2.2 Corpus Annotation Procedure

The N1s analyzed in this study fall into three semantic categories: container, collective, and quantifier. Within each semantic group, four N1s were selected, and for each one, 100 occurrences of subject-verb agreement in subject-position pseudopartitive constructions were annotated.

To account for irrelevant or incomplete results often returned by corpus queries, the first 200 randomized hits per noun were downloaded. Annotation proceeded until 100 valid subject-position tokens were obtained, with the remainder discarded to ensure equal representation across N1s. An example search is shown in (4), with (5) illustrating a specific query for the noun *porcentaje* 'percentage.' The list of N1s used to represent each semantic category was assembled based on the author's intuition as a native speaker of Spanish, with the aim of capturing nouns that are most frequently used in pseudopartitive constructions. As this is an exploratory study, no corpus-based or frequency-driven selection criteria were applied; however, future work will employ a more rigorous and systematic approach to N1 selection. A complete list of the selected N1s by semantic category appears in Table 1.

(4) determiner + (any number of optional adjectives) + the N1 analyzed + (any number of optional adjectives) + the word *de* ("of") + (any number of optional adjectives) + a random N2 (plural forms only) + (any number of optional adjectives) + a random verb

(5) <s> [tag="D.*"] [tag="A.*"]* [word="porcentaje"] [tag="A.*"]* [word="de"] [tag="A.*"]* [tag="N..P.*"] [tag="A.*"]* [tag= "V.*"]

Category	N1 analyzed
Container Nouns	*Bolsa* 'bag' *Caja* 'box' *Paquete* 'package'/'pack' *Puñado* 'handful'
Collective Nouns	*Grupo* 'group' *Equipo* 'team' *Pila* 'pile' *Conjunto* 'set'
Quantifier Nouns	*Montón* 'lot' *Número* 'number' *Par* 'pair' *Porcentaje* 'percentage'

Table 1. N1s analyzed per semantic category

Each occurrence was annotated in an Excel spreadsheet for two key parameters: agreement (N1 or N2) and the animacy of the N2. For agreement, the annotation was either "N1" or "N2," with no ambiguity expected, as the N1 was always singular and the N2 was forced to be always plural. Given that Spanish verbs overtly mark number, the source of agreement can be identified with confidence. For N2 animacy, one of four categories was assigned: human, animal, collective, or inanimate. The "collective" category applies to entities such as groups, organizations or institutions. For example, in the

sentence *a group of hotels were built*, the noun *hotels* would be categorized as inanimate, referring to physical structures. In contrast, in *a group of hotels offers significant discounts*, *hotels* would be considered collective, as it denotes an organization acting as an agent offering the discounts. In the event of an ambiguous sentence, "inanimate" was used as the default label. Once all occurrences were annotated, descriptive and inferential statistical analyses were conducted in R (R Core Team, 2021) to identify patterns and test for statistical significance. A total of 1,200 annotated pseudopartitive constructions were analyzed, with 400 occurrences for each N1 semantic type.

3 Results and Implications

N1 semantic type was associated with clear differences in agreement patterns. The majority of constructions with collective and container N1s strongly favored agreement with the N1 (84.2% and 87%, respectively), while constructions with quantifier N1s displayed a more balanced distribution, with 54.5% N1 agreement and 45.5% N2 agreement. These results are summarized in Table 2, with a visual representation provided in Fig. 1.

N1 semantic type	N1 agr	N2 agr	Total	N1 agr (percentage)	N2 agr (percentage)
collective	337	63	400	84.2	15.8
container	348	52	400	87.0	13.0
quantifier	218	182	400	54.5	45.5

Table 2: Agreement by N1 semantic type

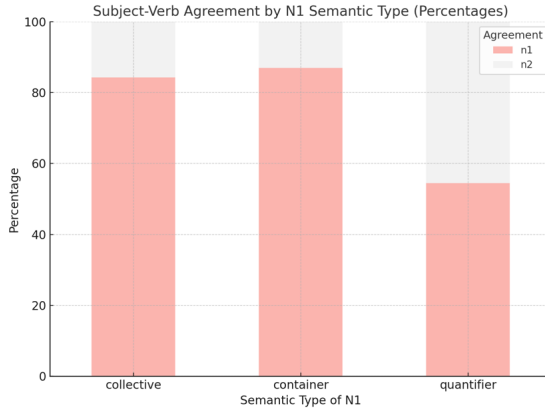


Figure 1: Bar chart. Agreement by N1 semantic type

With respect to N2 animacy, inanimate N2s showed the strongest tendency toward N1 agreement, with only 20.3% of cases exhibiting N2 agreement. Human and animal N2s behaved similarly to each other, showing N2 agreement in roughly 30% of cases (30.4% and 35.0%, respectively). The strongest tendency toward N2 agreement was observed with collective N2s, which displayed an almost even split between N1 and N2 agreement (48.7% vs. 51.3%). Full descriptive counts and percentages are presented in Table 3 and illustrated in Fig. 2.

N2 animacy	N1 agr	N2 agr	Total	N1 agr (percentage)	N2 agr (percentage)
animal	13	7	20	65.0	35.0
collective	19	20	39	48.7	51.3
human	263	115	378	69.6	30.4
inanimate	608	155	763	79.7	20.3

Table 3: Agreement by N2 animacy

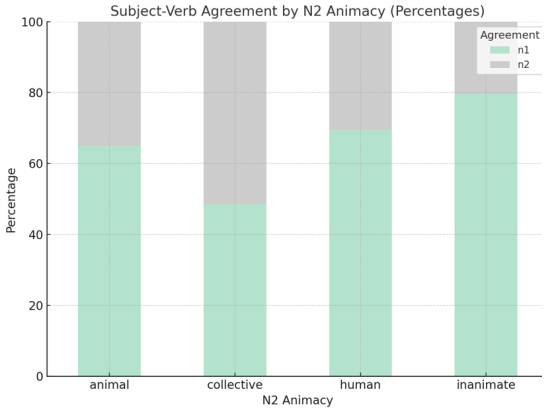


Figure 2: Bar chart. Agreement by N2 animacy

Two chi-square tests of independence were conducted to examine whether N1 semantic type and N2 animacy were associated with subject-verb agreement choice. A chi-square test of independence assesses whether two categorical variables are associated by comparing the actual frequencies observed in the data to the frequencies that would be expected if the variables were statistically independent. In the present study's case, it tests whether the distribution of subject-verb agreement (N1 vs N2) depends on the semantic type of the N1 or the animacy of the N2. If the observed frequencies differ substantially from what would be expected under the

assumption of no relationship, the test produces a large chi-square value and a small p-value, indicating a significant association between the variables.

The first test revealed a strong and statistically significant association between the N1 semantic type variable and agreement, χ^2 (2, N = 1200) = 139.52, p < .001. This indicates that the semantic type of N1 used in the pseudopartitive construction significantly influenced whether the verb agreed with the N1 or N2. This effect appears to be largely driven by the higher rate of N2 agreement observed with quantifier N1s, compared to container and collective N1s (see Table 2). The second chi-square test found a weaker, but still statistically significant, association between the N2 animacy variable and agreement, $\chi^2(3, N = 1200) = 30.46$, p < .001. This suggests that the animacy status of the N2 (whether it referred to a human, animal, collective, or inanimate entity) had some influence on agreement patterns, though not as strong as the effect of N1 semantic type. This pattern appears to be driven in part by the relatively high N2 agreement rates observed for collective N2s (see Table 3).

To investigate which specific conditions influenced agreement patterns, a post-hoc binary logistic regression was conducted. While the earlier chi-square tests showed that both N1 semantic type and N2 animacy were associated with agreement, they could not identify which conditions within each variable were driving the effect. Logistic regression addresses this limitation by estimating the contribution of each condition to the likelihood of N2 agreement, while controlling for the other variable. This allows for testing whether particular conditions—quantifier, container, or collective N1s, and human, animal, collective, or inanimate N2s—significantly increase the probability of N2 agreement when other factors are held constant. In this study, the model predicted whether agreement occurred with the embedded noun (N2) or the head noun (N1), based on the values of the two variables.

The overall model was statistically significant, $\chi^2(5, N = 1200) = 154.4$, p < .001, indicating that the variables helped explain variation in agreement patterns. Among the individual conditions, only quantifier N1s had a statistically significant effect. Compared to constructions with collective N1s, those with quantifier N1s were substantially more likely to show N2 agreement. The coefficient for quantifier N1s was 1.513 (p < .001), corresponding to an odds ratio of 4.54, calculated by exponentiating the coefficient. This means that, all else being equal, constructions headed by quantifier N1s were more than four times as likely to display N2 agreement. Container N1s did not differ significantly from collectives.

With respect to N2 animacy, none of the categories reached statistical significance in this model. However, the effect for collective N2s approached significance, showing a somewhat higher likelihood of N2 agreement than inanimate N2s, though this difference did not meet the conventional threshold for significance.

Predictor	Estimate (β)	Std. Error	z value	p value	Significance
(Intercept)	-1.606	0.5152	-3.117	0.00183	**
semantic type: container	-0.1464	0.2134	-0.686	0.49256	
semantic type: quantifier	1.513	0.1739	8.699	<2e-16	***
n2 animacy: collective	1.0466	0.6099	1.716	0.08617	†
n2 animacy: human	0.1474	0.5116	0.288	0.77332	
n2 animacy: inanimate	-0.3194	0.5064	-0.631	0.52821	

Table 4: Logistic regression results predicting the likelihood of N2 agreement

An additional model including the interaction between N1 semantic type and N2 animacy failed to converge meaningfully due to quasi-complete separation. Several N1-N2 combinations (e.g., container N1s paired with animal N2s) had zero or near-zero cases of N2 agreement, resulting in inflated standard errors and uninterpretable coefficients. As such, only the main effects model is reported.

The descriptive and inferential statistics replicated what was found by Foppolo et al. (2023), in the sense that quantifier N1s facilitated a more balanced distribution of N1/N2 agreement.

Therefore, the present study's results provide more evidence that semantic characteristics can accurately predict agreement patterns. Moreover, Foppolo et al. (2023) predicted and confirmed a gradient in agreement preferences based on the semantic type of the N1, with N2 (plural) agreement becoming increasingly acceptable from containers to collectives to quantifiers, reflecting the increasing accessibility of a measure construal. The present study replicated this gradient in Spanish to some extent: both descriptive statistics and logistic regression results show that quantifier N1s favored N2 agreement the most, followed by collectives, while containers showed the strongest preference for N1 agreement. However, collective N1s and container N1s feature only a three percent difference in descriptive statistics, and this small difference was not statistically significant in the inferential model. Hence, caution should be taken in this regard.

Although the descriptive statistics suggested a potential correlation between N2 animacy and subject-verb agreement patterns, the inferential analyses did not support this relationship. No statistically significant link was found between N2 animacy and the choice of N1 or N2 agreement. The present findings do not show a direct effect of animacy in the N2 and choice of N1/N2 subject-verb agreement in Spanish pseudopartitives. However, the role of animacy in this domain should not be ruled out at this point, with further studies needed to fully explore its potential influence.

4 Conclusion

In sum, this study contributes to the understanding of subject-verb agreement variation in Spanish pseudopartitive constructions by applying a corpus-based methodology informed by a replicable three-way categorization of N1s proposed by Foppolo et al. (2023). This approach facilitated cross-linguistic comparison and revealed that quantifier N1s significantly increased the likelihood of N2 agreement, a result supported by both descriptive statistics and logistic regression. While the overall order of agreement preferences—containers showing the least N2 agreement, collectives in the middle, and quantifiers the most—mirrored the gradient observed in Foppolo et al. (2023), the minimal difference between container and collective N1s (just three percent) and the lack of a statistically significant difference between them suggest that the gradient is only partially replicated.

The study also explored N2 animacy as a novel predictor, but found no statistically significant correlation with agreement patterns. However, the question of whether certain characteristics of the N2 can predict N1/N2 agreement in Spanish pseudopartitve structures is not exhausted. Further studies should be conducted, analyzing other possible characteristics of the N2 that might facilitate one type of agreement over the other.

These findings reinforce the importance of N1 semantics in agreement variation and highlight the value of combining psycholinguistically informed frameworks with corpus-based methods. At the same time, they point to several directions for future research. As an exploratory study, this analysis focused on a limited set of twelve N1 nouns selected to represent three broad semantic categories. While this approach enabled clear comparisons across N1 types, it does not capture the full range of variation found in Spanish pseudopartitive constructions. Expanding the dataset and incorporating more ambiguous or marginal cases would allow for a more comprehensive understanding of the phenomenon.

To support this broader coverage, future work will adopt a more systematic approach to N1 selection, potentially drawing on corpus-based frequency data or cross-linguistic comparability measures. In parallel, greater attention should be given to the role of embedded noun features. While N1 semantics were categorized with reference to psycholinguistic literature, no equivalent framework was applied to N2 animacy. Adopting psycholinguistically grounded categories for N2s may help clarify their contribution to agreement patterns and support more robust cross-linguistic comparisons.

Acknowledgments

This research was supported by the ISRAEL SCIENCE FOUNDATION (grant No. 1861/23).

References

Bayanati, S. and Toivonen, I. (2019). Humans, Animals, Things and Animacy. *Open*

Linguistics 5(1), 156–170. DOI: https://doi.org/10.1515/opli-2019-0010

Branigan, H. P., Pickering, M. J. and Tanaka, M. (2008). Contributions of Animacy to Grammatical Function Assignment and Word Order during Production. *Lingua* 118(2), 172–189. https://doi.org/10.1016/j.lingua.2007.02.003

Bresnan, J. and Hay, J. (2008). Gradient Grammar: An Effect of Animacy on the Syntax of *give* in New Zealand and American English. *Lingua* 118(2), 245–259. https://doi.org/10.1016/j.lingua.2007.02.007

Chierchia, G. (1998). Partitives, reference to kinds and semantic variation. In A. Lawson (Ed.), *Proceedings of Semantics And Linguistic Theory (SALT) VII* (pp. 73–98). CLC Publications. DOI: https://doi.org/10.3765/salt.v7i0.2792

Chomsky, N. (2000). Minimalist inquiries: The framework. In Martin, Roger, David Michaels, and Juan Uriagereka (Eds.), Step by Step: Essays in Minimalist Syntax in Honor of Howard Lasnik, pp. 89–155. Cambridge, Mass: MIT Press.

Chomsky, N. (2001). Derivation by phase. In Kenstowicz, Michael (Ed.), *Ken Hale: A Life in Language*, 36, 1-52. The MIT Press. DOI: https://doi.org/10.7551/mitpress/4056.001.0001

Cornilescu, A. (2009). Measure Phrases and the Syntax of Romanian Nouns and Adjectives. *Bucharest Working Papers in Linguistics*, 1, 35–67. Available at: https://www.ceeol.com/search/article-detail?id=235512

Danon, G. (2011). Agreement with quantified nominals: implications for feature theory. In O. Bonami & P. Cabredo Hofherr (Eds.), Empirical Issues in Syntax and Semantics, 8, 75-95. Available at: http://www.cssp.cnrs.fr/eiss8/

Danon, G. (2013). Agreement alternations with quantified nominals in Modern Hebrew. *Journal of Linguistics*, 49, 55–92.

Demonte, V. and Perez-Jimenez, I. (2015). Construcciones partitivas y pseudopartitivas en español: concordancia híbrida y variación en la interficie sintaxis-semántica. In Hernández, E. Butragueño, P. M. (Eds.), *Variación y diversidad lingüística: Hacia una teoría convergente*. El Colegio de México, 15-98. Available at: https://www.researchgate.net/publication/376782936_Demonte_Perez Jimenez_2015_Construcciones_partitivas_y_pseudopartitivas_en_espanol_concordancia_hibrida_y_variacion_en_la_interficie_sintaxis-semantica

Foppolo et al. (2023). A Group of Researchers Are Testing Pseudopartitives in Italian: Notional Number is not the Key to the Facts. *Glossa Psycholinguistics*, 2(1). Available at: https://escholarship.org/uc/item/18g1c99t

Franks, S. (1994). Parametric properties of numeral phrases in Slavic. *Natural Language and Linguistic Theory*, 12, 597–674.

Gámez, P. B. and Vasilyeva, M. (2015). Exploring Interactions between Semantic and Syntactic Processes: The Role of Animacy in Syntactic Priming. *Journal of Experimental Child Psychology* 138, 15–30. https://doi.org/10.1016/j.jecp.2015.04.009

Grestenberger, L. (2015). Number marking in German measure phrases and the structure of pseudo-partitives. *The Journal of Comparative Germanic Linguistics*, 18, 93-138. https://doi.org/10.1007/s10828-015-9074-1

Huumo, T. (2018). The partitive A: On uses of the Finnish partitive subject in transitive clauses. In A. Ilja Seržant & Alena Witzlack-Makarevich (Eds.), *Diachrony of differential argument marking* (pp. 383–411). Language Science Press. DOI:10.5281/zenodo.1228271

Kilgariff, A. et al. (2014). The Sketch Engine: Ten Years On. *Lexicography*, 1(1), 7-36. Available at: https://journal.equinoxpub.com/lexi/article/view/17961

Kilgarriff, A. and Renau, I. (2013). esTenTen, a Vast Web Corpus of Peninsular and American Spanish. *Procedia – Social and Behavioral Sciences*, 95, 12-19. DOI:10.1016/j.sbspro.2013.10.617

Milner, J. C. (1978). *De la syntax à l'interprétation*. Editions de Seuil.

Padró, L. and Stanilovsky, E. (2012). *FreeLing 3.0: Towards Wider Multilinguality.* Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul, Turkey.

Pesetsky, D. (1982). *Paths and Categories* (PhD dissertation), MIT.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rosenbach, Anette. (2008). Animacy and grammatical variation – Findings from English genitive variation. *Lingua* 118(2). 151–171. DOI: https://doi.org/10.1016/j.lingua.2007.02.002

Schwarzschild, R. (2006). The role of dimensions in the syntax of noun phrases. *Syntax*, 9, 67-110.

Selkirk, E. (1977). Some remarks on noun phrase structure. In P. Culicover, T. Wasow & A. Akmajian (Eds.), Formal syntax (pp. 285–316). Academic Press.

Shatil, N. (2015). The Nature and Diachrony of Hebrew Quality Pseudo-Partitives: Are They a Calque from the Contact Languages? *Journal of Jewish Languages*, 3(1-2), 301-308. https://doi.org/10.1163/22134638-12340046

Sketch Engine. (n.d.). *Timestamped Spanish corpus*. Retrieved February 22, 2025a, from https://www.sketchengine.eu/jozef-stefan-institute-newsfeed-corpus/#toggle-id-1

Stavrou, M. (2003). Semi-Lexical Nouns, Classifiers, and the Interpretation(s) of the Pseudopartitive Construction. In Martine Coene and Yves D'hulst (Eds.), *From NP to DP. Volume 1: The Syntax and Semantics of Noun Phrases*. John Benjamins Publishing Company, 329-353.

Suchomel, V. (2020). Better Web Corpora For Corpus Linguistics And NLP. Doctoral thesis. Masaryk University, Faculty of Informatics, Brno. Available at: https://is.muni.cz/th/u4rmz/.

Özsoy, A. (2009). Argument structure, animacy, syntax and semantics of passivization in Turkish: A corpus-based approach. In Y. Kawaguchi, M. Minegishi & J. Durand (Eds.), *Corpus Analysis and Variation in Linguistics* (pp. 259-279). John Benjamins Publishing Company. https://doi.org/10.1075/tufs.1.16ozs

Vihman, V. A. and Nelson, D. (2019). Effects of Animacy in Grammar and Cognition: Introduction to Special Issue. *Open Linguistics* 5(1). 260–267. DOI: https://doi.org/10.1515/opli-2019-0015

Wechsler, S. and Zlatić, L. (2003). *The Many Faces of Agreement*. CSLI Publications.

Zamparelli, R. (2008). Dei ex machina: A note on plural/mass indefinite determiners. *Studia Linguistica*, *62*(3), 301–327. DOI: https://doi.org/10.1111/j.1467-9582.2008.00149.x

Degree centrality as a measure of robustness of dependency structures of the sentences in a large-scale learner corpus of English

Masanori Oya
School of Global Japanese Studies,
Meiji University
masanori_oya2019@meiji.ac.jp

Abstract

This paper examines the differences in the robustness of syntactic dependency structures in written English produced by learners of varying proficiency levels and by native English speakers. The robustness of these dependency structures is represented by their degree centralities, and corpus-based investigation revealed that learners with higher proficiency levels tend to produce sentences with lower degree centralities. This means that they produce more robust, and more embedded sentences. It is also revealed that the sentences produced by native speakers of English tend to produce more embedded sentences than non-native speakers.

1 Introduction

The aim of this paper is to examine the differences in the robustness of syntactic dependency structures in written English produced by learners of varying proficiency levels and by native English speakers.

Structural properties of sentences have been explored in the field of second language acquisition (SLA) using a variety of metrics such as word per sentence or type-token ratio with the cover term of *sentence complexity* (e.g., Bardovi-Harlig 1992, Brown 1973, Ellis and Yuan 2005, Hunt 1965, Michel et al. 2007, Norris and Ortega 2009, Ortega 2003, Robinson 2007, Scarborough 1990, Scott 1988, Skehan and Foster, 2005, Wolf-Quintero et al. 1998). The basic tenet behind them is that the proficiency levels of learners can be represented by these metrics. In other words, it is expected that these metrics increase in proportion to the advancement of learners' proficiency levels. For example, Wolfe-Quintero et al. (1998) pointed out that depth of clauses in the sentences produced by leaners of English increases in proportion to their proficiency levels, hence depth of clauses can function as a measure of sentence complexity.

Sentence complexity should not be regarded as a single independent variable, but as a dependent variable that can be represented by multiple variables (depth of clauses is one of them). In this context, it is essential to address these variables related to sentence complexity individually rather than treating them collectively and indiscriminately. By focusing on each factor in turn, we can understand the structural characteristics of the sentences produced by speakers/writers with certain attributes (e.g., native/non-native, beginners/intermediate/advanced, non-native with different backgrounds) more objectively.

This paper introduces the robustness of dependency structures as one of these variables related to sentence complexity. Specifically, I adopt *degree centrality* of the dependency structure of a sentence as a metric to measure its robustness. By modeling the dependency structures of English sentences in a corpus—organized by learners' proficiency levels—as graphs, I compute their degree centralities and investigate whether the distribution of these values reflects the learners' proficiency levels.

This paper is organized as follows: Section 2 summarizes the idea of dependency structures as graphs and their degree centralities, and explains the relationship between degree centralities of dependency structures and their robustness. In Section 3, previous studies are briefly reviewed to point out their drawbacks. Section 4 describes this study of degree centralities of English sentences in a large-scale learner corpus, which is followed by discussions in Section 5. Section 6 concludes this paper.

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 9–16

2 Dependency trees as graphs

In network analysis, a graph consists of a collection of nodes and a set of edges linking these nodes (Freeman, 1978; Wasserman & Faust, 1994). In this context, the degree of a node is determined by the number of edges connected to it. Previous research (Oya, 2010, 2013, and 2014) has posited that the dependency tree (or structure) of a sentence can be conceptualized as a graph. More specifically, within a dependency tree, words function as nodes, while their dependency relationships are represented as edges, and the degree of a word is the number of other words depending on it and the word which it depends on. For example, an English sentence "I have written this article" has the dependency structure in the format of *Universal Dependencies* (de Marneffe et al. 2021, Zeman et al. 2017) in Figure 1. The degree of the word "written" is four, because it depends on "root," and three words depend on it.

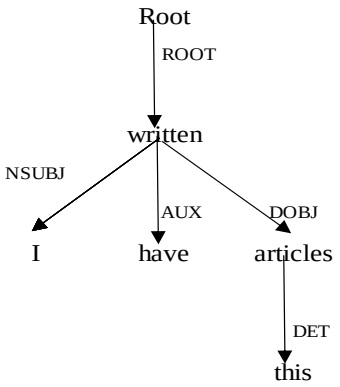


Figure 1: The dependency structure of the sentence "I have written this article."

Graph theory establishes a variety of metrics that quantify the structural characteristics of graphs. If the dependency tree of a sentence is considered a specific type of graph, these metrics can be utilized to analyze its structural properties systematically. This approach enables a more objective and scientifically rigorous examination of its structure, as opposed to relying solely on intuitive interpretations. Based on this premise, Oya (2010) applied *degree centrality* (Freeman 1978; Wasserman & Faust 1994) as a metric to assess the complexity of dependency trees in English sentences (yet the use of the word "complexity" is rather problematic; discussed later).

Degree centrality is a type of index that indicates the significance of a given node within a specific graph. The degree centrality of a graph C_D which contains g nodes is calculated by the following formula (Freeman 1978, Wasserman & Faust 1994):

$$C_D = \frac{\sum_{i=1}^{g}[C_D(n^*) - C_D(n_i)]}{max\sum_{i=1}^{g}[C_D(n^*) - C_D(n_i)]} \quad (1)$$

$C_D(n^*)$ is the largest degree in the given graph, and $C_D(n_i)$ is the degree of a node. The enumerator represents the sum of the largest degree minus the degrees of all the other nodes. The denominator represents the maximal possible sum of the largest degree minus the degrees of all the other nodes. For a graph which contains g nodes, the largest possible degree of its node is g-1.

In principle, degree centrality ranges from 0 to 1.

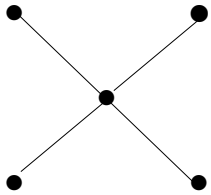


Figure 2: A star graph.

If a graph has a degree centrality of 1, this signifies that a single node within the graph is connected to all other nodes, forming a *star graph*, as illustrated in Figure 2.

The largest degree in the star graph in Figure 2 is 4, which is the largest possible degree of a graph which contains 5 nodes. The degree of all the other nodes is 1. The enumerator and the denominator are the same, as indicated by the following formula, hence the degree centrality of the star graph is 1:

$$C_D = \frac{(4-1)+(4-1)+(4-4)+(4-1)+(4-1)}{(4-1)+(4-1)+(4-4)+(4-1)+(4-1)} = 1 \quad (2)$$

If the graph representing the dependency tree of a sentence has a degree centrality of one, this indicates that a single word serves as the dependency head for all other words in the sentence. In other words, the dependency structure of the sentence is entirely flat.

Degree centrality values decrease as the structure of graphs becomes more linear, meaning that no single node holds greater significance than the others. In Figure 3, the five nodes are arranged in a linear sequence (forming a *line graph*).

The largest degree in the line graph in Figure 3 is 2, and 3 of its nodes have that degree. The other 2 nodes have the degree 1. Hence, the enumerator is 2. The denominator is the same as that of the star graph in Figure 2. Therefore, as indicated by the formula (3), the degree centrality of the line graph in Figure 3 is approximately 0.1667:

$$C_D = \frac{(2-1)+(2-2)+(2-2)+(2-2)+(2-1)}{(4-1)+(4-1)+(4-4)+(4-1)+(4-1)} = \frac{2}{12} \approx 0.1667 \quad (3)$$

The degree centrality of the example sentence "I have written this article" is calculated as follows: its dependency structure contains 6 nodes (including Root). Its largest observed degree is 4 (with *written*). One node has the degree 2 (with *article*), and all the other nodes have the degree one. The largest possible degree of a graph with 6 nodes is 5. Therefore, as indicated by the formula (4), the degree centrality of the dependency structure of the example sentence is 0.7:

$$C_D = \frac{(4-1)+(4-4)+(4-2)+(4-1)+(4-1)+(4-1)}{(5-1)+(5-1)+(5-5)+(5-1)+(5-1)+(5-1)} = \frac{14}{20} = 0.7 \quad (4)$$

If the graph representing the dependency tree of a sentence exhibits a low degree centrality, this indicates that one word depends on another, which in turn depends on yet another, and so forth, resulting in a more embedded dependency structure.

The degree centrality of a network (and a dependency structure of a sentence) is concerned with its robustness. Sentences with larger degree centralities (with flatter dependency structure) contain a certain core word (or words) on which many of the other words in the same sentence depend. If the core word is deleted or overlooked, then the whole structure falls apart into unrelated words, and fails to be interpreted appropriately. On the other hand, sentences with lower degree centralities (with more embedded dependency structure) have no such core, or more than one core, and therefore, even if one of the words is deleted, there will be some fragments of structure which can be interpreted, though not completely. Hence, sentences with lower degree centralities are more robust than those with higher degree centralities.

It should be noted that the robustness of dependency structures as discussed here differs conceptually from syntactic robustness.

3 Previous studies

Figure 3: A line graph.

Some previous studies assume that the degree centrality values of sentences within a corpus can function as an indicator of their syntactic complexity, in which the word "complexity" is used as something represented by degree centrality. Oya (2010) observed that the degree centrality values of English essays written by Japanese learners tend to be higher than those of academic journal abstracts, suggesting that the former exhibit flatter and less embedded syntactic structures compared to the latter. Oya (2013) conducted corpus-based research of degree centrality as a syntactic complexity measure. He revealed that sentences in different genres show different distributions of degree centralities, more specifically, sentences in fictions tend to have higher degree centralities than those in journals, meaning that the former have flatter syntactic structure than the latter. Oya (2014) applied the idea of using degree centrality as a syntactic complexity measure into Japanese, based on an English-Japanese small-scale parallel corpus, and it is found that Japanese sentences tend to have higher degree centralities than their English translations, meaning that Japanese sentences are flatter than their English translations.

The previous studies on the degree centralities of sentences contain the following two drawbacks: First, it is assumed that the degree centrality of a sentence can be used as a measure of its complexity without explicit explanation on why it can be. It is certain that structures with lower degree centralities are more robust, and it is found that the robustness is one of the characteristics of complex systems (e.g., Artime, Grassia, De Domenico et al. 2024), yet it is not certain that more robust sentences are more complex. These previous

studies should have used the word “robustness” instead of “complexity” of syntactic structure. Second, there has been no study of degree centrality as a measure of structural robustness of English sentences generated by learners of English as a second language (L2) at different proficiency levels, let alone comparing and contrasting the degree centralities of English sentences generated by non-native speakers of English (non-ENS) and those generated by native speakers of English (ENS). In this context, this study is the first attempt to examine whether the degree centralities of sentences generated by non-ENS in different proficiency levels show distributions which are different across these different proficiency levels, and from those by ENS. If any difference between them is found, that will give us a new insight into the difference between non-ENS and ENS in terms of the robustness of the sentences they generate, based on the theoretical background of graph theory.

4 This study

The research question of this study is as follows:
(1) Do degree centralities of the sentences generated by non-ENS at different proficiency levels show different distributions across these levels?
(2) Do degree centralities of the sentences generated by non-ENS show distributions which are different from those generated by ENS?

4.1 Data

The production data examined in this study are the written essay section of the International Corpus Network of Asian Learners of English (ICNALE; Ishikawa 2013, 2023), a learner corpus of English, production data of English from college-level students with a variety of backgrounds across Asia (China, Hong Kong, Indonesia, Japan, Korea, Pakistan, the Philippines, Singapore/Malaysia, Taiwan, and Thailand), along with production data from native speakers of English. In the written essay section of ICNALE, the topics of the essays are as follows:

Topic A: College students need to have a part-time job.

Topic B: Smoking should be completely banned at all restaurants in the country.

A notable characteristic of the ICNALE is its systematic classification of production data based on learners’ proficiency levels, as defined by the Common European Framework of Reference for Languages (CEFR). These levels include A2, B1_1 (B1 low), B1_2 (B1 high), and B2+. In the written essay section of the ICNALE, each learner is assigned a proficiency level according to their scores on various English proficiency tests, and their essays are then categorized accordingly within these CEFR levels.

4.2 Procedure

The average degree centrality of the sentences of the essays of Topic A and Topic B in each CEFR category in the ICNALE is calculated by a Python script which was coded by the 1st author, then these average degree centralities are compared across these CEFR categories. Also, the distributions of degree centralities of individual sentences are compared across these CEFR categories, in terms of the percentage of the degree centralities falling within particular subranges of the interval of 0.1. Since degree centralities fall within the range from 0 to 1, it is further divided into those subranges. If it is found that the degree centralities of sentences produced by learners who are categorized into one particular CEFR category, say, B1_1, fall within a certain subrange, such as that from 0.2 to 0.3, significantly more often than other subranges, then it indicates the structural characteristics of the sentences produced by learners of that CEFR category, suggesting that learners who belong to the CEFR category B1_1 tend to produce sentences whose degree centralities often fall within that subrange.

This procedure was conducted for each Topic individually, so that we can examine whether there is any difference of degree centralities due to the difference of topics: Topic A is related to college life, and therefore it must be more familiar to the learners than Topic B, which is related to one of the social issues. As Oya (2013) pointed out that sentences of different genres show different distributions of degree centralities, it is expected that the difference of topics in the ICNALE would result in different distributions of degree centralities.

4.3 Results

	N	ADC	SD
A2	7287	0.36	0.19
B1_1	14369	0.34	0.17
B1_2	12967	0.31	0.17
B2	6244	0.3	0.15
ENS	1779	0.23	0.14

Table 1: The average degree centralities of the sentences in the essays about Topic A (part-time job). ADC: average degree centralities

	N	ADC	SD
A2	7460	0.38	0.2
B1_1	14678	0.37	0.18
B1_2	13440	0.34	0.17
B2	3249	0.33	0.17
ENS	1981	0.27	0.16

Table 2: The average degree centralities of the sentences in the essays about Topic B (Ban on smoking). ADC: average degree centralities

Table 1 shows the average degree centralities of the sentences in the essays about Topic A (part-time job), and Table 2 shows those in the essays about Topic B (smoking ban on local restaurants):

In both groups, the average degree centralities decrease from A2 at the largest among them to the ENS at the lowest.

A one-way between subjects ANOVA was conducted to compare the average degree centralities across the categories for each topic group. For Topic A, there was a significant effect of categories on average degree centralities at the $p<.01$ level [F(4, 4264) = 276.87]. Post hoc comparisons using the Tukey HSD test indicated that the mean scores for A2 (M = 0.36, SD = 0.19), B1_1 (M = 0.34, SD = 0.17), B1_2 (M = 0.31, SD = 0.17), B2 (M = 0.3, SD = 0.15) and ENS (M = 0.23, SD = 0.14) are all different from each other. For Topic B, there was also a significant effect of categories on average degree centralities at the $p<.01$ level [F(4, 4080) = 211.32]. Post hoc comparisons using the Tukey HSD test indicated that the mean scores for A2 (M = 0.33, SD = 0.2), B1_1 (M = 0.37, SD = 0.18), B1_2 (M = 0.34, SD = 0.17), B2 (M = 0.33, SD = 0.17) and ENS (M = 0.27, SD = 0.16) are all different from each other, except for the pair of B1_2 and B2. These results suggest that degree centralities of sentences on average decrease in negative proportion to the proficiency of the learners, and yet they are still larger than those produced by ENSs. The scenario can be summarized roughly as follows: Leaners at lower proficiency levels write English sentences which contain flatter structure (with larger degree centralities), and as their proficiency level gets higher, they come to produce sentences with more embedded structure (with smaller degree centralities).

It is also interesting to note that the average degree centrality of Topic A is smaller than that of Topic B regardless of the CEFR category. This may

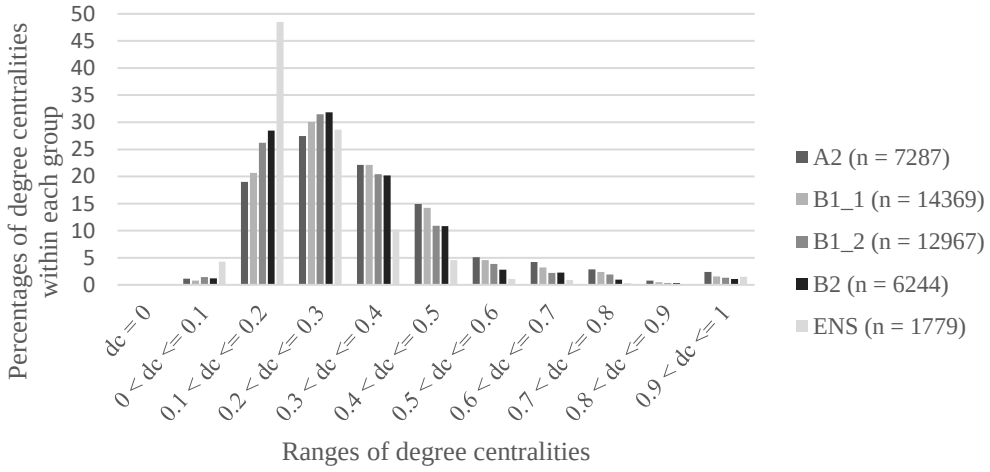


Figure 3: Percentages of degree centralities of the sentences in each group

(Topic A: Part-time job)

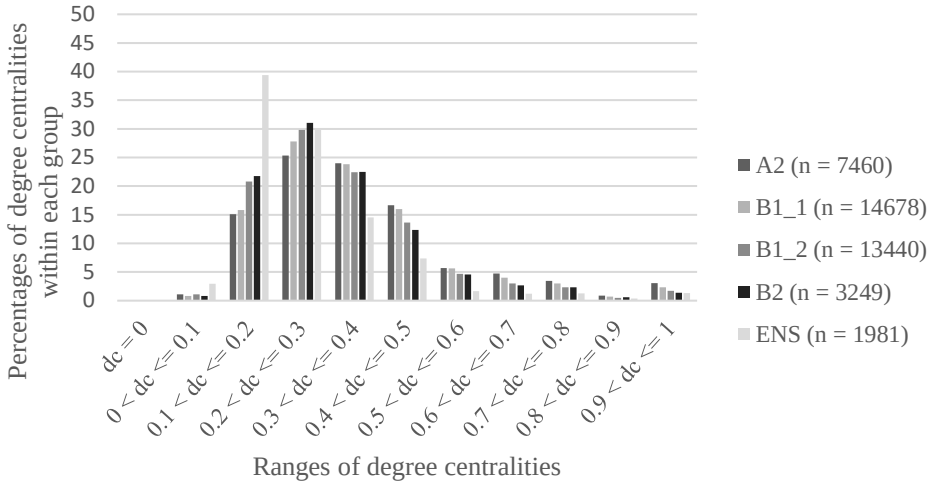


Figure 4: Percentages of degree centralities of the sentences in each group

(Topic B: Ban on smoking)

suggest that the topic of an essay influences the degree centralities of its sentences. As far as Topic A and Topic B in the ICNALE are concerned, Topic A is related to their own life and therefore they may find it easier to describe their own experiences and opinions, adding into their sentences more phrases describing the details, resulting in more embedded sentences with smaller degree centralities. Topic B, on the other hand, needs to establish their own argument on a topic which they may be less familiar with, hence their description stays simple and this can be reflected on less embedded sentences.

Figure 3 and 4 show the distributions of degree centralities of individual sentences in all the CEFR categories, represented by the percentages of degree centralities which fall within each subrange. About 50% of the sentences the degree centralities of the category ENS for Topic A (part-time job) fall within the range of degree centralities which are 0.2 or smaller and larger than 0.1, while in the same range the degree centralities of the sentences generated by non-ENS groups are less than 30% of those in each group. Similar distributions of degree centralities can be found for Topic B (Ban on smoking); About 40% of the sentences the degree centralities of the category ENS for this group fall within the range of degree centralities which are 0.2 or smaller and larger than 0.1, while in the same range the degree centralities of the sentences generated by non-ENS groups are less than 25% of those in each group.

5. Discussion

The results described above seem to answer the research questions positively: First, in the ICNALE, average degree centralities of the sentences generated by non-ENSs at different proficiency levels show different distributions across these levels, decreasing from lower levels to higher levels. Second, average degree centralities of the sentences generated by non-ENS show distributions which are different from those generated by ENSs. These findings must be put into the context of linguistic description along with the explanation on why it is the case. There are at least three questions to be addressed: (1) Why do non-ENSs at lower proficiency levels tend to produce sentences with flatter structure? (2) Why do they come to produce more embedded sentences as they become more proficient? And (3) Why do ENSs produce more embedded sentences than non-ENSs? If degree centralities of sentences (or their robustness) on average decrease as the proficiency of learners gets higher, it can be explained that it gets more robust than before. These issues need to be addressed in future research, for better understanding of degree centralities of sentence structure, from the viewpoint of (1) investigating the relationship between degree centralities and other sentence complexity measures, such as type-token ratio, word per sentence, and (2) formulating the theory which explains how and why the robustness of sentences increases along with the development of learners' proficiency.

An anonymous reviewer noted that these issues have been addressed in previous research from the perspective of sentence length. Ferrer-i-Cancho and Gómez-Rodríguez (2019) argue that shorter sentences often conflict with the principle of dependency distance minimization (DDM) (cf. Liu, 2008; Futrell et al., 2015). According to DDM,

language users tend to prefer shorter dependency distances, as longer distances increase the cognitive burden on working memory. Ferrer-i-Cancho and Gómez-Rodríguez (2019) suggest that the apparent violation of DDM in short sentences arises from their characteristically flat, star-like dependency structures. A plausible scenario can thus be outlined: beginner-level language learners, who typically produce shorter sentences, tend to generate structures with higher degree centralities. As their proficiency improves, they begin to construct longer, more syntactically complex sentences, which in turn exhibit less star-like configurations—potentially as a strategy to mitigate the cognitive load imposed by long dependency distances within such structures. This hypothesis warrants further empirical investigation in future research.

6. Conclusion

This paper examined the differences in the robustness of syntactic dependency structures in written English produced by learners of varying proficiency levels and by native English speakers. The robustness of these dependency structures is represented by their degree centralities, and corpus-based investigation revealed that learners with higher proficiency levels tend to produce sentences with lower degree centralities, meaning they produce more robust, and more embedded sentences, yet the sentences produced by native speakers of English tend to produce more embedded sentences than non-native speakers. The results of this study lead us to further exploration of degree centralities of dependency structures as a measure of their robustness.

References

Oriol Artime, Marco Grassia, Manlio De Domenico, James P. Gleeson, Hernán A. Makse, Giuseppe Mangioni, Matjaž Perc and Filippo Radicchi. 2024. Robustness and resilience of complex networks. *National Review of Physics* 6, 114–131. https://doi.org/10.1038/s42254-023-00676-y

Kathleen Bardovi-Harlig. 1992. A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly*, 26, 390-395.

Roger Brown. 1973. *A First Language: The Early Stages*. Harvard University Press.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2).

Rod Ellis and Fanguan Yuan. 2005. 'The effects of careful within-task planning on oral and written task performance' in R. Ellis (ed.): *Planning and Task Performance in a Second Language*, 167-192. John Benjamins.

Ramon Ferrer-i-Cancho and Carlos Gómez-Rodríguez. 2019. Anti dependency distance minimization in short sequences. A graph theoretic approach. *Journal of Quantitative Linguistics*, 28(1), 50–76. https://doi.org/10.1080/09296174.2019.1645547

Linton Freeman. 1978. Centrality in social networks. *Social Networks* vol.1, 215-239.

Richard Futrell, Kyle Mahowald, and Edward Gibson. Large-scale evidence of dependency length minimization in 37 languages. *The Proceedings of the National Academy of Sciences (PNAS)* 112 (33), 10336-10341. https://doi.org/10.1073/pnas.1502134112

Kellog W. Hunt. 1965. Grammatical structures written at three grade levels. *NCTE Research Report* No. 3. Champaign, IL, USA: NCTE.

Shin-ichiro Ishikawa. 2013. The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner corpus studies in Asia and the world*, 1, 91-118.

Shin-ichiro Ishikawa. 2023. *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*. Routledge.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2), 159-191. http://dx.doi.org/10.17791/jcs.2008.9.2.159

Marije C. Michel, Folkert Kuiken and Ineke Vedder. 2007. The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching*, 45, 241-259.

John M. Norris and Lourdes Ortega. 2009. Towards an organic approach to investigating CAF in instructed SLA: The case of complexity, *Applied Linguistics* 30(4), 555-578, https://doi.org/10.1093/applin/amp044

Lourdes Ortega. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics* 24(4), 492–518, https://doi.org/10.1093/applin/24.4.492

Masanori Oya. 2010. Directed acyclic graph representation of grammatical knowledge and its application for calculating sentence complexity. *Proceedings of the 15th International Conference of Pan-Pacific Association of Applied Linguistics*, 393-400.

Masanori Oya. 2013. Degree centralities, closeness centralities, and dependency distances of different genres of texts. *Selected Papers from the 17th Conference of Pan-Pacific Applied Linguistics*, 42-53.

Masanori Oya. 2014. *A Study of Syntactic Typed-Dependency Trees for English and Japanese and Graph-centrality Measures*. Ph.D. dissertation, Waseda University.

Peter Robinson. 2007. Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics*, 45, 237-257.

Hollis S. Scarborough 1990. Index of productive syntax. *Applied Psycholinguistics* 11, 1-22.

Cherry M. Scott. 1988. Spoken and written syntax. In M. Nippold (ed.): *Later Language Development: Ages Nine through Nineteen*. Little, Brown.

Peter Skehan and Pauline Foster. 2005. Strategic and on-line planning: The influence of surprise information and task time on second language performance. In R. Ellis (ed.): *Planning and Task Performance in a Second Language*. John Benjamins.

Stanley Wasserman and Katherine Faust. 1994. *Social Network Analysis*. Cambridge: Cambridge University Press.

Kate Wolfe-Quintero, Shunji Inagaki, and Hae-Yong Kim. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*. University of Hawaii Press.

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gökırmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.C., Sanguinetti, M., Simi, M., Kanayama, H., de Paiva, V., Droganova, K., Alonso, H.M., Çöltekin, Ç., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H.F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R., and Li, J. (2017). CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 1–19, Vancouver.

Application of Existing Readability Methods to the Ukrainian Language: A Comprehensive Study

Serhii D. Prykhodchenko*, Oksana Yu. Prykhodchenko[x]
∗ Computer Systems Software Department, Dnipro University of Technology, Dnipro, Ukraine
prykhodchenko.s.d@nmu.one
x Computer Systems Software Department, Dnipro University of Technology, Dnipro, Ukraine
prykhodchenko.o.yu@nmu.one

Abstract

The Ukrainian language currently lacks a well-developed framework for assessing text readability. This study addresses this gap by focusing on three key contributions. First, we present the creation of UkrTB, a Ukrainian-language corpus of texts categorized by reader age. Second, we conduct a statistical analysis of the corpus, evaluating key linguistic features such as sentence length, word complexity, and part-of-speech distribution. Third, we systematically assess the applicability of existing readability formulas, including Flesch, Flesch-Kincaid, Matskovskii, Pisarek, and Solnyshkina et al., to Ukrainian texts. Our findings indicate that readability models developed for English and other Slavic languages exhibit significant limitations when applied to Ukrainian. While some methods demonstrate partial correlation with expected readability levels, others produce inconsistent results, underscoring the need for a specialized readability metric tailored to Ukrainian. This work lays the foundation for further research in Ukrainian readability assessment and the development of language-specific models.

Keywords: Readability, Ukrainian language, Natural Language Processing, Corpus Linguistics, Text Complexity

1 Introduction

Research on quantitative readability assessment began in the 1940s with Flesch's early work (Flesch, 1948), leading to the development of various readability metrics. However, these methods were primarily designed for English and did not account for structural differences in other languages. English, as an analytical language, relies on a fixed word order and auxiliary words to convey meaning, whereas Slavic languages, including Ukrainian, use a more synthetic structure characterized by extensive inflection, case systems, and grammatical gender. These differences complicate the direct application of English-based readability models to Ukrainian, as they fail to capture the complexity introduced by its flexible word order and morphological variation. Ukrainian is characterised by rich morphology (cases, long word forms), syllabic structure (complex syllable divisions, less vowel reduction) and syntax (long sentences, free word order). Polish has consonant clusters (ex. ‘szczegółowy’). At the same time, Russian is characterised by vowel reduction - this does not coincide with Ukrainian, where vowels are fuller. Formula coefficients must be calibrated on the Ukrainian corpus of texts and tested on native speakers, otherwise they give incorrect results. Since the 1960s, readability studies have been conducted for some Slavic languages, particularly Polish and Russian, leading to the development of language-specific models. However, Ukrainian remains underexplored in this context, despite having a substantial number of native speakers and a distinct grammatical system. The absence of a dedicated readability assessment framework for Ukrainian presents a challenge for text classification, educational content adaptation, and NLP applications. Existing readability formulas, whether developed for English or other Slavic languages, may not be directly transferable to Ukrainian due to its unique linguistic features.

To address this gap, this study makes three key contributions. First, we construct a Ukrainian-language text corpus categorized by readability levels to provide a foundation for further research. Second, we conduct a statistical analysis of the corpus, evaluating linguistic features such as sentence structure, word complexity, and part-of-speech distribution. Third, we systematically

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 17–25

assess the applicability of existing readability formulas, including those developed for English, Polish, and Russian, to determine their effectiveness for Ukrainian. Our findings will help establish whether current models can be adapted or if a new methodology is required to develop an accurate readability assessment framework for the Ukrainian language.

2 Related work

To consider the task of determining the applicability of existing methods for calculating readability to Ukrainian-language texts, it is necessary to solve several successive tasks:

1. Create a text corpus suitable for testing the research hypotheses and, in the long term, for further research.

2. Determine existing methods for determining readability that use the desired parameters for determining readability created for Ukrainian and other languages.

3. Apply previously determined methods, evaluate their accuracy for the Ukrainian language, and draw conclusions for further research.

Prior research has explored readability assessment across multiple domains. Studies on text complexity for second-language learners have focused (Xia et al., 2016) on CEFR-graded English datasets, addressing the challenge of limited annotated data. These works have adapted models trained on native corpora, leveraging domain adaptation and self-training techniques to improve performance. High accuracy and strong correlation coefficients suggest that similar approaches could be explored for the Ukrainian language.Lexical richness has also been shown to correlate with perceived quality in ESL learners' oral narratives (Lu, 2012), suggesting that vocabulary diversity is an important dimension of readability and language proficiency assessment.

Efforts to standardize readability assessment in educational materials have also been undertaken, particularly in the context of Ukrainian textbooks. Research from Lviv Polytechnic (Krychkovska et al., 2014) investigated text complexity using the "Chitanka" program, analyzing the accessibility of scientific content. The study emphasized the importance of systematically increasing textual complexity in educational materials, advocating for a more structured approach to readability in Ukrainian academia.

Recent advancements in multilingual readability estimation highlight the potential of cross-lingual transfer learning. The paper "An Open Multilingual System for Scoring Readability of Wikipedia" (Trokhymovych et al., 2024) presents a novel approach to assessing the readability of Wikipedia articles across multiple languages. It introduced a model trained on Wikipedia articles in 14 languages, demonstrating significant improvements over previous benchmarks. By aligning document pairs across different languages, the model effectively assessed text complexity even in languages with limited resources, providing a strong foundation for future Ukrainian readability models.

The UKRMED corpus (Cherednichenko et al., 2020), focuses on Ukrainian medical texts, including clinical protocols and forums, categorized by complexity. Its creation involved preprocessing, tokenization, and statistical analysis, with crowdsourcing to improve quality. Studies using Pymorphy2 showed that frequency-based methods are insufficient, highlighting the need for advanced linguistic features and lexical resources—relevant for broader readability assessment.

Next studies (Cherednichenko et al., 2018) highlight the diverse methodologies for readability evaluation and corpus development. While significant progress has been made in multilingual and domain-specific readability assessment, Ukrainian remains underexplored. By leveraging insights from existing research, this study aims to construct a robust Ukrainian readability framework, integrating statistical, linguistic, and computational approaches. They used the Pymorphy2 morphological analyzer and stopword lists for preprocessing. Their results underscored the need for specialized lexical resources and ontologies to simplify medical texts for better comprehension by non-specialist readers effectively.

The article (Cherednichenko and Kanishcheva. 2021) examined the application of readability formulas to Ukrainian medical texts using the UKRMED corpus, which categorizes texts into three levels of complexity: simple, moderate, and complex. The results indicated that while existing formulas produce similar rankings, medical texts remain inherently difficult to understand. The study

suggests that further refinement of readability metrics and detailed text markup could improve accessibility for both non-native speakers and individuals with varying educational backgrounds.

The paper (Vajjala and Meurers, 2012) investigated the impact of second-language acquisition (SLA) research on readability classification. By integrating SLA-based complexity measures with traditional readability metrics, the researchers improved classification accuracy. Their model, which combined lexical and syntactic features, outperformed conventional methods, achieving over 93% accuracy in predicting text difficulty across different grade levels. This work demonstrates the potential of interdisciplinary approaches in enhancing automated readability assessment, particularly in educational and language-learning contexts.

For Russian texts, (Solnyshkina, 2018) proposed a modified readability formula incorporating syntactic, lexical, and frequency features. Tested on Russian school textbooks, it showed improved accuracy over previous models and highlighted the potential of regression-based approaches for genre-specific readability assessment.

Similarly, research (Broda et al., 2014) on Polish texts evaluated multiple readability assessment methods, including traditional formulas such as Gunning-Fog and Pisarek's method. In addition, the study introduced novel approaches, such as distributional lexical similarity and an automated Taylor test based on statistical language modelling. The authors developed an online tool for readability evaluation, showing a strong correlation between various readability indices and effective classification of text complexity levels.

Corpus parameterization has also been a focus in readability studies (Starko and Cheylytko, 2013), as highlighted in research on optimizing corpus balance and representativity. A hybrid methodology combining statistical models, expert evaluations, and adaptive monitoring was proposed to address the proportionality of text types in linguistic corpora. This study emphasizes the importance of dynamic, data-driven strategies in corpus design, ensuring that robust and well-structured datasets support readability research.

Adapting readability assessment methods for the German language has seen significant advancements through various innovative approaches. Recent research has focused on integrating machine learning and natural language processing techniques to enhance the accuracy and applicability of these methods. An online service has been developed that utilizes five statistical methods and two machine learning models, including BERT, to evaluate German text readability at the sentence level (Pickelmann, 2023). This tool is particularly beneficial in educational contexts, helping to assess the suitability of teaching materials for different grade levels.

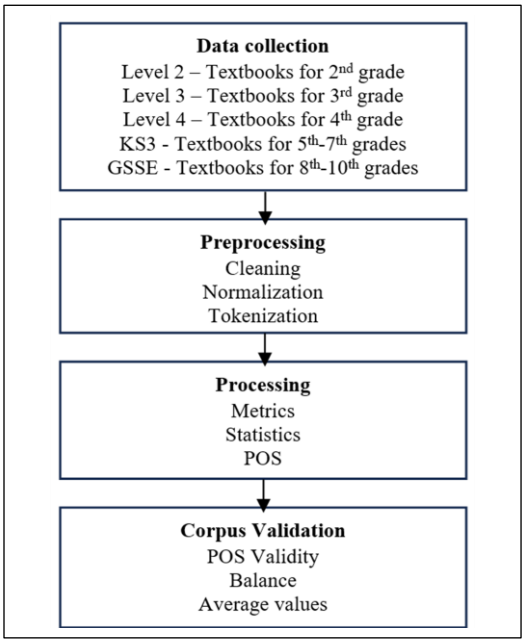


Figure 1: Algorithm of Text corpus creation

Recent research in German has demonstrated the use of neural models (e.g., BERT-based) for sentence-level readability prediction, showing potential for future adaptation to low-resource languages.(Blaneck et al., 2022), (Mohtaj et al., 2022).

3 Methods

This study follows a structured approach to evaluate the applicability of existing readability metrics to Ukrainian texts. The first step involved constructing UkrTB, a corpus of 750 texts extracted from Ukrainian educational materials, categorized into five readability levels corresponding to different school grades. The texts were preprocessed through cleaning, normalization, and tokenization. Part-of-speech tagging was performed using pymorphy3, leveraging OpenCorpora dictionaries to ensure consistency in morphological analysis. We

computed statistical features — including sentence length, word length, and syllable count — to quantitatively evaluate text readability.

To assess readability, we applied six established formulas: Flesch Readability Ease (Flesch, 1948), Flesch-Kincaid Grade Level (Kincaid et al., 1975), Pisarek's (Pisarek, 1969) linear and nonlinear models, Matskovskii's readability index (Matskovskii, 1976), and Solnyshkina et al.'s model (Solnyshkina, 2018). These metrics were chosen based on their relevance to English and Slavic languages. Each formula was systematically applied to the UkrTB corpus, producing readability scores for each text. The evaluation of these methods involved a correlation analysis using Spearman and Pearson coefficients to examine the relationship between formula scores and the intended educational levels. The correlation between computed readability scores and the predefined grade levels was quantified using both Pearson's and Spearman's coefficients. These statistical measures allow for evaluating the degree of linear and monotonic relationships, respectively. A distribution analysis was also conducted to assess intra-class variability, determining whether readability scores were consistent within the same grade level.

4 Collection of the dataset

Creating a text corpus for a source dataset is challenging, as the source data strongly influences the processing results. Currently, there are no standard techniques for creating text corpora in linguistics. Our method for collecting a corpus of texts has similarities with the WeeBit corpus, which involved Weekly Reader texts ranked by reader age (Vajjala and Meurers, 2012). We propose our own method for collecting texts to create a training and general corpus based on the work of (Cherednichenko and Kanishcheva. 2021)(Vajjala and Meurers, 2012). The main requirement for the corpus is to provide Ukrainian text data ranked by readers' ages to study language problems.

The corpus of Ukrainian-language texts we are creating, which we call UkrTB - Ukrainian Textbooks[1], was originally conceived as a corpus without saved tags and attributes. The first reason is the availability of modern tagging and attribution techniques, including the rapid creation of keyword lists using artificial intelligence elements. The second reason for not tagging is the perceived versatility of the corpus, which, as research progresses, is planned to be expanded to the size of WeeBit, i.e. about 6000 texts, divided into appropriate ages of potential readers, where texts range from folk tales to excerpts from history or physics textbooks.

This study proposes the following process (Fig. 1) based on the work (Cherednichenko et al., 2020) to create a text corpus ranging by readers' ages in Ukrainian. As data sources, texts from the electronic library formed by the Ministry of Education of Ukraine, which contains all textbooks and auxiliary literature recommended by the Ministry of Education from the first to the 11th grade, are considered. These texts include both folk tales and samples of literature from the 11th to 21st centuries, as well as scientific explanations of, for example, physics or chemistry,

It should be explained that there is no single textbook for the discipline studied in Ukraine; an author may submit their textbook, which meets the specified criteria, to the Ministry of Education, get approval, and then their textbook will be included in the list of recommended textbooks. Thus, on the recommended textbooks website, we can find 5 history textbooks for 7th grade and 6 Ukrainian literature textbooks for 8th grade, which may differ in the selection of materials and writing style. Nevertheless, the Ministry's recommendation confirms that these texts fulfil the criteria set out for textbooks, which means that a pupil of the relevant grade will be able to read the material.

The corpus design and validation processes were inspired by the works of Vasyl Starko (Starko and Cheylytko, 2013), who is one of the creators of the Brown Corpus of the Ukrainian language. Obviously, his works are based on the ideas of creating the Brown Corpus of English (Standard Corpus of Present-Day Edited American English, or shortened Brown Corpus). At the same time, we will note that the exact date of the creation of the text in this work is of little importance to us in comparison with the indication of the age audience of the reader. Thus, according to our criteria, a folk tale recorded more than a hundred years ago but present in a modern textbook for the 3rd grade is suitable for inclusion in the corpus. The average number of words per page of a textbook depends

[1] github.com/prykhodchenkosd/ukrtb

on the format, font and text density. Still, it averages around 250-300 words for a standard A4 page or a textbook without many illustrations and formulas. Textbooks with a dense layout (e.g., high school textbooks) may have up to 400 words, and those with many graphics or examples (e.g., elementary school textbooks) may have about 200 words. To create this corpus, we used an average article length of 2 pages, i.e., an average of about 500 words on the topic of one lesson.

The data are collected by dividing the texts by the reader's age group. We identify five levels of text complexity: Level2, Level3, Level4, KS3, and GCSE, similar to the age categories of the WeeBit corpus (Vajjala and Meurers, 2012) KS3 (Key Stage 3, UK educational stage for ages 11–14) and GCSE (General Certificate of Secondary Education, typically for ages 14–16). In the first stage, we collected equal amounts of data for each group using semantically similar sources, prioritizing textbooks from the same authors or editors. Texts were then cleaned, encoded uniformly, and processed using tokenization, normalization, and statistical analysisAt the processing stage, the program calculates statistical indicators of the text and metrics, such as the number of letters in a sentence, the number of words in a sentence, the average number of letters in a word, the average number of words in a sentence, the average number of syllables in a word, etc. The primary metrics and statistical indicators are given in Table 1. In further studies of readability formulas, each level will appear under a specific number: 1 – Level2, 2 – Level3, 3 – Level4, 4 – KS3, and 5 – GCSE. This coding is present in Fig. 2-6 in the form of marks on the abscissa axis.

To evaluate statistical information about parts of speech, the work used automatic POS tagging, carried out using the pymorphy3 [2] project. The original versions 1 and 2 of the pymorphy project can currently be considered abandonware; the development and support of the latest versions of Python is provided by its fork, pymorphy3. This project is based on the dictionaries of the OpenCorpora project, which are the basis for a significant number of scientific works in the field of text corpora processing (Korobov, 2015), (D. Kalugin-Balashov. 2023), (Tmienova and Sus, 2019), which allows us to talk about the quality and reliability of the assessments of this solution. The program using pymorphy3 produces a significant array of data on each of the words of the text, including the POS decision, which is issued in an abbreviated form[3].

Grade **Level**	Age in Years	Num. of articles	Avg. number of sentences per article	Avg. syllables in a word	Avg. words in a sentence
Level 2	7-8	150	24,17	2,27	10,6542
Level 3	8-9	150	31,16	2,34236	10,28945
Level 4	9-10	150	41,35	2,15672	13,10985
KS3	11-14	150	37,82	2,34837	12,9896
GCSE	14-16	150	27,91	2,51437	15,7587

Table 1: Primary metrics and statistical indicators

As a result of processing the corpus texts, categorizations of words by parts of speech were obtained for each text passage that makes up the corpus. The results are presented in Table 2.

5 Applicability of existing readability methods to Ukrainian texts

Classic works on calculating the readability coefficient are the works of Flesch, who defined a general readability formula and, later, Kincaid derived the grade-level readability formula, which determined the readability index depending on the level of education of the reader. Flesch's works were aimed at studying the English-language texts; they also served as a starting point for several similar studies of later times, which resulted in the methods for determining readability ARI, SMOG, and several others, which were also focused on English-language texts. Attempts to use these methods for other languages, even of the related Germanic language group, usually concluded with the need to create a method entirely adapted to the corresponding language (Pickelmann et al., 2023), (Blaneck et al., 2022), (Mohtaj et al., 2022).

For the languages of the Slavic group, such studies began to be conducted in the 1960s-70s, gaining recognition for such languages as Russian and Polish, for which studies of readability

[2] https://pypi.org/project/pymorphy3/

[3] https://github.com/no-plagiarism/pymorphy3/blob/master/docs/user/grammemes.rst

	Level 2	Level 3	Level 4	KS3	GCSE
Total	26276	36132	53548	59498	55109
NOUN	9957	13622	18357	23409	23623
ADJF	2115	3516	3915	6444	7880
VERB	4721	6274	9599	8996	6700
NPRO	2389	3285	5682	5219	3651
ADVB	1253	1640	2588	2688	2153
PRCL	1606	2064	4030	3329	2525
CONJ	2514	3278	5775	5016	4355
PREP	1479	2081	3005	3684	3617
GRND	99	171	252	355	270
NUMR	143	201	345	358	335

Table 2: Number of part-of-speech in the text corpus

assessment methods and software products based on these methods are still relevant. At the same time, evaluations of other reading methods for other Slavic languages, including Ukrainian, have either not been conducted at all or have not received sufficient publicity and recognition.

The Ukrainian language has the most remarkable lexical similarity with the Belarusian language and, to a lesser extent, with Russian and Polish. Studies of readability assessment methods for the Belarusian language have also not been conducted, so in this paper, we will try to apply methods applicable to Russian and Polish, as well as the Flesch and Flesch-Kincaid methods, which have become classics, for assessing the readability of Ukrainian-language texts collected in the UkrTB corpus. On the way to this part of the study, we manually checked the automatically performed morphological partitioning of the corpus presented in Table 2.

Flesch's work assumed the calculation of the readability coefficient based on such parameters as total words, sequences, and syllables. Its consideration is still encountered today when considering readability coefficients for languages different from English. Formula (1), proposed by Flesch, was calculated for the texts of the UkrTB corpus, resulting in the graph presented in Fig. 2. This graph shows each class's maximum, average, and minimum values.

$$F = 206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - -84.6\left(\frac{total\ syllables}{total\ words}\right) \quad (1)$$

In 1975, Kincaid at al. improved Flesch's readability method (Kincaid et al., 1975) by including a proposed division by educational level, thus calibrating the readability coefficient by the level of the proposed reader. This method relies on the same parameters as in the original work - total words, sequences, and syllables, and Formula 2 is a mathematical description of this method. In this study, all corpus texts were also processed using this method, resulting in the graphs presented in Fig. 3, which also determined the maximum, minimum, and average coefficients for each class of readers. Although Figure 3 shows an upward trend in Flesch-Kincaid scores with increasing grade level, the within-group variance is substantial. Furthermore, the score ranges overlap significantly between neighboring levels, making it difficult to reliably distinguish readability tiers based on these scores alone. This explains the assessment of low correspondence, despite the overall trend.

$$FK = 0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\left(\frac{total\ syllables}{total\ words}\right) - -15.59 \quad (2)$$

Later, linguists developed readability methods for Slavic languages, as English-based formulas proved unsuitable. Matskovskii and Pisarek's formulas were chosen due to their relevance to Slavic structures. Pisarek's Polish method, adapted for inflected languages with complex morphology, is a strong candidate for Ukrainian. Matskovskii's formula, designed for Russian, incorporates syntactic and lexical complexity, offering insights into how readability metrics perform in another East Slavic language. Given Ukrainian's linguistic position between Polish and Russian, testing these models provides a comparative framework for assessing their applicability and identifying necessary modifications for a more accurate Ukrainian readability metric.

The author of the readability coefficient for the Polish language is prof. Pisarek (Pisarek, 1969) proposed calculating the linear (3) and nonlinear

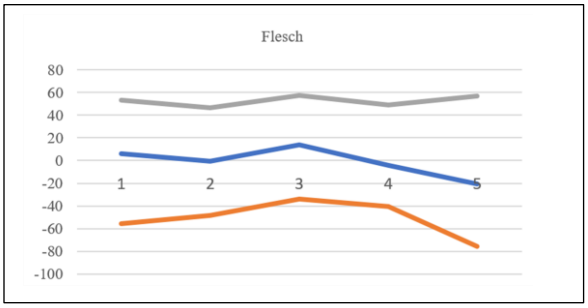


Figure 2: Classes of UkrTB by Flesch readability ease

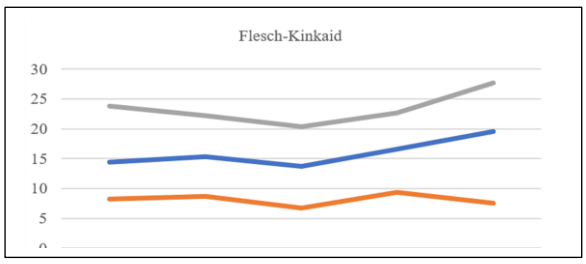


Figure 3: Classes of UkrTB by Flesch-Kinkaid readability ease

(4) dependencies for the readability coefficient of the Polish language based on ASL (average number of words per sentence) and PCW (percentage of complex words). His work in this field is still used to analyze Polish-language texts (Broda et al., 2014). In our implementation, complex words (PCW) were defined as those exceeding three syllables and not belonging to a predefined list of stopwords or functional parts of speech. This operationalization was chosen to approximate the lexical difficulty as perceived by native readers, following practices in Polish readability research.

$$P_l = \frac{1}{3}ASL \cdot \frac{1}{3}PCW + 1 \quad (3)$$

$$P_{nl} = \frac{1}{2}\sqrt{ASL^2 + PCW^2} \quad (4)$$

This study applied these techniques to UkrTB corpus. The results can be seen in Figure 4.

Some methods have been proposed for the Russian language, which differs significantly from each other in the time of creation and the number and composition of the studied text features. In this study, we considered the method (5) (Matskovskii, 1976), and based on such parameters as average sentence length and X_3 is the percentage of words of more than 3 syllables in the text.

$$M = 0.62ASL + 0.123X_3 + 0.051 \quad (5)$$

We also considered a relatively recent work by M. Solnyshkina and co-authors (Solnyshkina, 2018), based on the analysis of ASW (average number of syllables per word), ASL (average number of words per sentence), UNAV (relation between the number of unique words in text: (number of unique Adjectives + number of unique Nouns)/(number of unique Verbs)) and NAV (relation between the number of words in text: (type-token ratio of Adjectives + type-token ratio of Nouns)/(type-token ratio of Verbs)).

$$S = -0.124ASL + 0.018\,ASW - 0.007\,UNAV + \\ + 0.007\,NAV - 0.003\,ASL^2 + 0.184\,ASL\,ASW + \\ + 0.097\,ASL\,UNAV - 0.158\,ASL\,NAV + \\ + 0.09\,ASW^2 + 0.091\,ASW\,UNAV + \\ + 0.023\,ASW\,NAV - 0.157\,UNAV^2 - \\ - 0.079\,UNAV\,NAV + 0.058\,NAV^2 \quad (6)$$

The results of applying formulas 5 and 6 to the corpus texts can be seen in Fig. 5.

The overall figure comparing the applications of the above-described methods can be seen in Fig. 6. To facilitate the visual comparison of readability scores across methods with different scales and directionalities, we applied min-max normalization to each set of scores before plotting. The normalized scores are shown on a unified vertical axis, where higher values indicate higher perceived text complexity. This allows for approximate comparison of method behavior across grade levels.

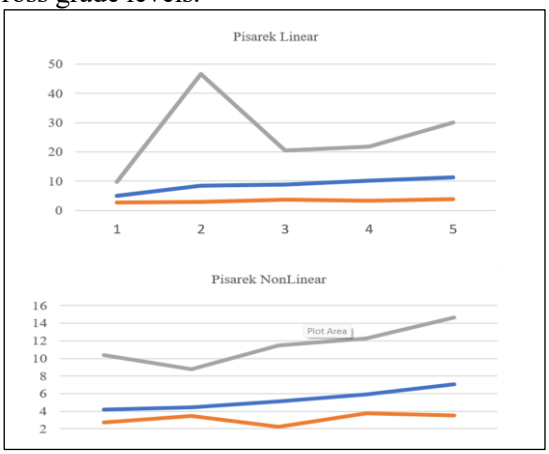


Figure 4: Classes of UkrTB by Pisarek linear and non-linear readability ease

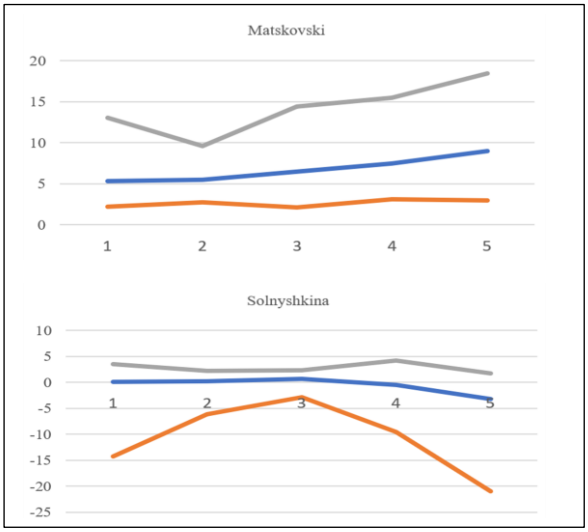


Figure 5: Classes of UkrTB by Matskovskii and Solnyshkina readability ease

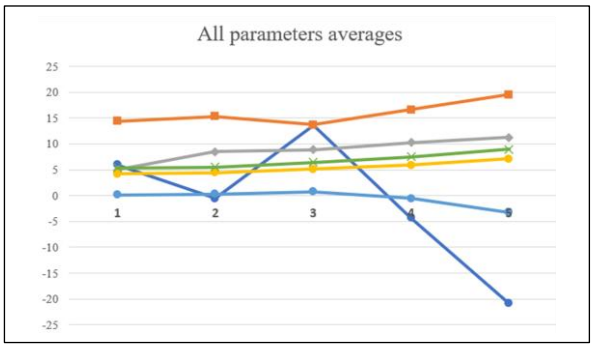


Figure 6: Classes of UkrTB by methods 1-6 of readability ease

The examination of the above-described methods applied to the texts of the UkrTB corpus showed the following results:

1. The Flesch and Flesch-Kincaid methods showed a low correspondence of the calculated readability coefficients to the expected change in the level of text complexity from low to high, which was initially assumed due to the initial calculation of the applicability of these methods for the English language.

2. Pisarek's methods, on average, showed a proportional increase in the expected complexity of texts. Still, the difference is not obvious when considering the range of values within classes, which is especially important for the linear model.

3. Among the considered models of the Russian language applied to the texts of the UkrTB corpus, Solnyshkina's model showed an average inversely proportional nonlinear dependence, with an initial directly proportional one, and Matskovskii's model - a directly proportional one. However, as in the case of Pisarek's models, the results of both models within a class show a wide range of values, which does not allow us to talk about sufficient accuracy in determining the level of readability for the Ukrainian language.

6 Conclusions

As a result of the research work, two main results were obtained, the first of which is the creation of a Ukrainian-language corpus of texts divided by the assumed levels of education into 5 classes.

The second result of this work is a series of tests of the applicability of existing methods for determining readability coefficients to corpus texts written in Ukrainian.

To summarize, some methods for determining readability in average results show dependencies between the level of readability and the assumed level of education. In some cases, this dependence is the opposite of what was declared by the creators of the method. Nevertheless, the spread of values obtained as a result of processing is such that in the presence of two texts of different classes, false identifications of the readability level are possible not only with the nearest class but - through a class, two, and in some cases - through three, i.e., at the other end of the readability level and the assumed level of education. Consequently, the methods proposed earlier for languages other than Ukrainian can be applied to Ukrainian texts only in limited cases, using additional coefficients, and for large average samples, which makes their real application difficult. For instance, one text designed for Level 2 (age 7–8), consisting of short sentences and simple vocabulary, was assessed by the Solnyshkina model as equivalent in complexity to GCSE-level texts. This highlights how non-adapted formula parameters may misinterpret language features that are not penalized in Ukrainian, such as long noun phrases or complex morphologies.

In general, this study shows that readability formulas such as Flesch, Flesch-Kincaid (for English), Pisarek (for Polish) and Matskowski and Solnyskina (for Russian) are not suitable for Ukrainian because of linguistic and statistical differences. Formulas like Flesch count the length of sentences and words, but their coefficients are customised for English. Readers' cultural expectations are also important: complex constructions in Ukrainian are perceived naturally, but existing formulas “penalise” them.

The Ukrainian language needs a new formula with empirical data and adapted parameters that consider specificity. Thus, as a result of these studies, it can be concluded that additional research is necessary to determine the dependence of the parameters of Ukrainian-language texts on the expected level of education and thus determine a specialized methodology for determining the readability coefficient for the Ukrainian language.

References

P. G. Blaneck, T. Bornheim, N. Grieger, and S. Bialonski. 2022. Automatic Readability Assessment of German Sentences with Transformer Ensembles. https://doi.org/10.48550/arXiv.2209.04299.

B. Broda, B. Nitoń, W. Gruszczyński, and M. Ogrodniczuk. 2014. Measuring Readability of Polish Texts: Baseline Experiments. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 573–580, Reykjavik, Iceland. European Language Resources Association (ELRA).

O. Cherednichenko and O. Kanishcheva. 2021. Readability Evaluation for Ukrainian Medicine Corpus (UKRMED). COLINS, pages 402–412.

O. Cherednichenko, O. Kanishcheva, and N. Babkova. 2018. Complex term identification for Ukrainian medical texts. CEUR Workshop Proceedings, 2255:146–154.

O. Cherednichenko, O. Kanishcheva, and O. Yakovleva. 2020. Collection and Processing of a Medical Corpus in Ukrainian. Computational Linguistics and Intelligent Systems, COLINS, CEUR Workshop Proceedings, 2604:272–282.

Rudolf Flesch. 1948. A new readability yardstick. Journal of Applied Psychology, 32(3):221–233. https://doi.org/10.1037/h0057532. PMID 18867058.

D. Kalugin-Balashov. 2023. Advancing Full-Text Search Lemmatization Techniques with Paradigm Retrieval from OpenCorpora. https://doi.org/10.48550/arXiv.2305.10848.

J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel. Research Branch Report 8–75. Chief of Naval Technical Training: Naval Air Station Memphis.

M. Korobov. 2015. Morphological Analyzer and Generator for Russian and Ukrainian Languages. In: M. Khachay, N. Konstantinova, A. Panchenko, D. Ignatov, and V. Labunets (Eds.), Analysis of Images, Social Networks and Texts. AIST 2015, Communications in Computer and Information Science, vol 542. Springer, Cham. https://doi.org/10.1007/978-3-319-26123-2_31.

Lu, Xiaofei. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. The Modern Language Journal. 96. 190-208. 10.2307/41684069.

M. S. Matskovskii. 1976. Problems of Readability of Printed Material In Semantic Perception of a Speech Message in Mass Communication, pages 126–142. Nauka, Moscow.

S. Mohtaj, B. Naderi, S. Möller, F. Maschhur, C. Wu, and M. Reinhard. 2022. A Transfer Learning Based Model for Text Readability Assessment in German. https://doi.org/10.48550/arXiv.2207.06265.

A. M. Krychkovska, Zh. P. Parashchyn, O. V. Shved, I. I. Gubytska, L. D. Bolibrukh, V. P. Novikov . Application of information technologies for standardization of the methodology of creating educational literature // Bulletin of the Lviv Polytechnic National University. Informatization of a higher educational institution. -2014. - No. 803. - S. 81-85.

F. Pickelmann, M. Färber, and A. Jatowt. 2023. Ablesbarkeitsmesser: A System for Assessing the Readability of German Text. https://doi.org/10.1007/978-3-031-28241-6_28.

W. Pisarek. 1969. Jak mierzyć zrozumiałość tekstu. Zeszyty Prasoznawcze, (4):35–48.

M. Solnyshkina, V. Ivanov, and V. Solovyev. 2018. Readability Formula for Russian Texts: A Modified Version. In 17th Mexican International Conference on Artificial Intelligence, MICAI 2018, Guadalajara, Mexico, October 22–27, 2018, Proceedings, Part II. https://doi.org/10.1007/978-3-030-04497-8_11.

V. Starko and N. Cheylytko. 2013. Parameterization of the corpus as a way to increase its representativeness and balance. Ukrainian Linguistics, 43:87–94.

N. Tmienova and B. Sus. 2019. System of Intellectual Ukrainian Language Processing. CEUR Workshop Proceedings, 2577:199–209.

M. Trokhymovych, I. Sen, and M. Gerlach. 2024. An Open Multilingual System for Scoring Readability of Wikipedia. https://doi.org/10.48550/arXiv.2406.01835.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pages 163–173. ACL.

Mingqiang Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text Readability Assessment for Second Language Learners. Pages 12–22. https://doi.org/10.18653/v1/W16-0502.

Extraction of Contrastive Rules from Syntactic Treebanks: A Case Study in Romance Languages

Santiago Herrera[1], Ioana-Madalina Silai[1], Bruno Guillaume[2], Sylvain Kahane[1,3]
[1]Modyco, Université Paris Nanterre, CNRS
[2]Université de Lorraine, CNRS, Inria, LORIA
[3]IUF - Institut Universitaire de France
{s.herrera, 43016143, skahane}@parisnanterre.fr, bruno.guillaume@loria.fr

Abstract

In this paper, we develop a data-driven contrastive framework to extract common and distinctive linguistic descriptions from syntactic treebanks. The extracted contrastive rules are defined by a statistically significant difference in frequency and precision, and classified as common and distinctive rules across the set of treebanks. We illustrate our method by working on object word order using Universal Dependencies (UD) treebanks in 6 Romance languages: Brazilian Portuguese, Catalan, French, Italian, Romanian and Spanish. We discuss the limitations faced due to inconsistent annotation and the feasibility of conducting contrastive studies using the UD collection.

1 Introduction

Cross-lingual corpus-based studies normally focus on finding common and distinctive structural features or tendencies among languages, language families, or typological balanced samples. Word order tendencies and their correlation with other language formal properties are an example of typological high-level descriptions. However, one might be also interested in comparing fine-grained patterns that explain the variation or the similarity between the compared corpora. This is a common goal in translation, second language teaching, textual genre research, and more generally, in corpus-based contrastive linguistics.

Comparing languages becomes more challenging the closer the languages are to each other. For example, syntactic objects vary considerably among Romance languages, even though they also exhibit some shared properties. Nominal objects often follow their verb, and personal pronominal objects tend to be in preverbal position. Both word order rules are common and dominant (in terms of frequency) in all Romance languages.

However, personal pronouns are enclitics of infinitives (or gerunds where they exist) in languages such as Spanish, Catalan or Italian, among others, while this is not possible in French. And even if the exact syntactic configuration exists in all languages, the relative frequency may vary. In addition, fine-grained differences within a language family are not always shared by the same group of languages. As mentioned, Spanish and French do not have the same word order between infinitive verbs and personal pronominal objects, but they do when the verb is an imperative (see Figure 1).

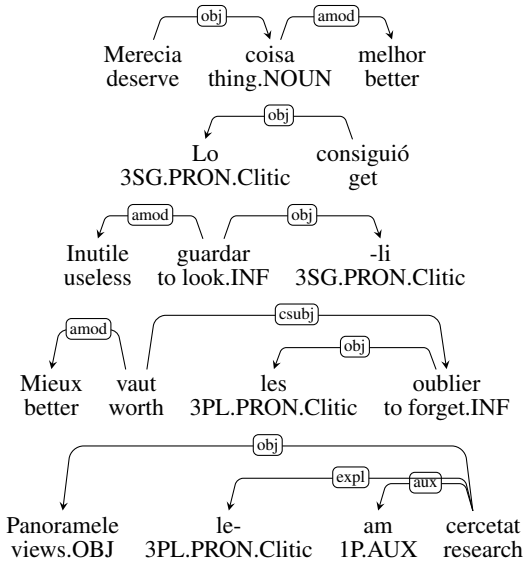


Figure 1: Different nominal and pronominal object word orders in Portuguese, Spanish, Italian, French and Romanian.

From this perspective, corpus-based contrastive analysis should be able to identify absolute differences across a set of languages, such as unique linguistic features, and be flexible enough to capture significant differences in frequency for fine-grained phenomena. Moreover, a contrastive approach should focus not only on distinctive patterns, but also on common ones. This approach allows a more detailed analysis of languages and corpora, focusing on variation or similarity in a specific lin-

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 26–38

guistic feature rather than on language profiles.

To address this issue, we develop a data-driven framework to extract fine-grained common and distinctive linguistic descriptions from syntactic treebanks. The extracted contrastive rules are defined by a statistically significant difference in precision and classified as common and distinctive rules across the set of treebanks. For each distinctive rule, we regroup languages having the same behavior. To test our method, we analyse the word order of objects in six Romance languages using Universal Dependencies (UD) treebanks. Object word order, especially object clitic order, varies considerably across Romance languages (Roberts, 2016). The present study is limited to this specific phenomenon. Further studies are left for future exploration.

Our approach follows and contributes to the core principles of most recent descriptive grammar extraction works. We seek descriptive but systematic descriptions of linguistic phenomena (Chaudhary et al., 2020, 2022), where extracted rules are overlapping rules, more or less fine-grained, and associated to quantitative data (Herrera et al., 2024a,b). The rules should be few and easy to interpret.

The contributions of this study can be summarised as follows:

- We adapt and extend current grammar extraction methods to extract concise and systematic contrastive descriptions for a set of corpora.
- We take a linguistically and statistically motivated approach to understanding common and distinctive patterns.
- We discuss the limitations of the use of universality based datasets (like UD) for automatic contrastive grammar description.
- We provide the UD community (de Marneffe et al., 2021) with a different perspective on annotation at the language family level.

2 Related Works

Corpus-based approaches to cross-linguistic analysis are widespread in both typological and contrastive linguistics, although in recent years they have become much more prominent in the latter. In either case, multilingual corpora have offered the possibility to capture and compare quantitative and gradient properties on a larger scale (Levshina, 2022).

Corpus-based Typological Studies

Typological studies based on parallel and comparable treebanks have mainly focused on high-level structural properties, typically word order or linguistic complexity. For example, Choi et al. (2021) explore the main word order categories in UD treebanks using a rewriting graph tool, and Gerdes et al. (2021) study quantitative word order and implicational universals on the same corpora, generalising to a larger variety of word order patterns. Similar studies have also been conducted using massive parallel corpora in more languages (Östling, 2015).

A common way to compare languages is to cluster them using different syntactic representations to see if phylogenetic groups are reconstructed according to typological databases (Alves et al., 2023). To the best of our knowledge, there are no studies using UD treebanks that cluster languages according to fine-grained syntactic patterns as we do in this paper.

Corpus-based Contrastive Studies

Contrastive linguistics lies somewhere between single-corpus studies and multi-corpus comparative studies, where the interest is in capturing fine-grained similarities and differences of internal linguistic properties in a small set of languages (Hasselgård, 2020). Contrastive approaches include frequency and statistical modeling (Gries et al., 2020) and studies based on information theory (see for example Alves (2025) for comparison between two Portuguese varieties). Normally, contrastive studies favour the use of parallel corpora (Nikolaev et al., 2020; Alves et al., 2023) because of the easy alignment between constructions from different sources.

Our work is closer to the quantitative syntax tradition (Bresnan et al. 2007; see also Thuilier 2012 for an example in a Romance language), where the main goal is to discover how some selected factors explain specific syntactic phenomena, such as the dative alternation or adjective order preferences. Such approaches have also been examined from a comparative perspective (e.g. Bresnan and Ford, 2010). In our case, we work with more than two languages and our approach consists of automatically identifying the predictive factors.

Descriptive Grammar Extraction

This paper builds on recent work on descriptive grammar extraction, where the main goal is to pro-

duce quantitative fine-grained grammar rules using traditional machine learning (ML) techniques. Chaudhary et al. (2020, 2022) formalise the task as a classification problem using decision trees to extract agreement, word order, and case marking rules across all UD treebanks. Herrera et al. (2024a,b) use a sparse linear classifier for agreement and word order for a few languages to obtain more expressive rules. They do this in single-corpus and one-to-one contrastive scenarios, but not for a collection of languages.

An advantage of (regularized) linear models is their ability to extract overlapping rules, which reflects how grammatical rules often behave. In contrast, decision trees create disjoint partitions of the data. In addition, their rule structure is highly sensitive to hyperparameters, such as tree depth. Tuning these hyperparameters can be challenging when the goal is descriptive insight rather than predictive accuracy. For instance, limiting a tree's depth to improve interpretability may result in uninformative residual nodes that use complex negative conditions to account for the remaining examples.

3 Task definition

Our main goal is to identify common and distinctive rules given a set of treebanks, focusing on object word order. Following Herrera et al.'s (2024a) formalisation, a quantitative rule posits a predictive relationship within the data, where the presence of a pattern P, identified within a starting sample S, increases the likelihood, by an α extent, of a phenomenon of interest Q.

$$S \implies (P \overset{\alpha\,\%}{\Longrightarrow} Q)$$

For example, to extract object word order rules for all objects in the sample, we look for linguistic patterns that favour the right or the left position with respect to its governor:

$$S : {}^{\text{OBJECT}}_{\text{RELATION}} \implies (P \overset{\alpha\,\%}{\Longrightarrow} Q : {}^{\text{OBJECT}}_{\text{POSITION}})$$

Given this formalisation, we consider a common grammatical rule to be evenly distributed across all languages. A distinctive rule, on the other hand, is unevenly distributed, though it may be shared equally by a subset of languages. If a corpus-based rule is quantitative in nature and captures gradient phenomena, a contrastive rule additionally posits whether the distribution of the predictive factor P is uniformly distributed across subsamples, in our case across languages.

To explore word order, we extract contrastive rules for the following four questions: (1) word order between two nodes connected by a dependency, (2) object word order, (3) pronominal object word order, and (4) nominal object word order. Analyzing object word order and its subtypes in two steps might be considered unnecessary, as the rules emerge when examining general object word order when the regularization parameter of the linear model is low enough. However, extracting rules for each type of object allows us to examine them in a more precise subspace. Furthermore, it is reasonable to explore pronouns in more detail as they are selected as good predictors of object word order. Overall, this multistep process yields fewer, more significant rules on average.

Before discussing the methods in detail, we first present the data, its preprocessing, and the sampling strategies employed.

4 Data

This study focuses on Romance languages for three main reasons. First, they are well represented in the UD treebank collection, providing us with a robust dataset. Second, we are experts and have native proficiency or possess a comprehensive understanding in all of the chosen languages. Finally, the use of these languages as a test case is advantageous due to their closeness. We hypothesize that the successful detection of differences and similarities between closely related languages will serve as an indication of the applicability of our method.

When multiple treebanks were available in the UD collection for a single language, the largest one in terms of tokens was selected, except in cases where annotation quality or methodology warranted a different choice. In Table 1 we provide an overview of each treebank, including its size, annotation process, and text sources where available.

4.1 Data Processing and Sampling

In all our experiments, we use the listed treebanks and ensure cross-linguistic comparability by selecting the same number of relevant syntactic patterns for each language. For broad pattern types, e.g. all dependencies, we collect approximately 10% of the matches for the smallest treebank to ensure a balanced sample size across languages. For more

UD Treebank	Tokens \ Sentences	Genres	Annotation Process
Brazilian Portuguese (Porttinari) (Duran et al., 2023)	168k \ 8k	News	Morphology: automatically tagged (reviewed); Syntax: manual annotation
Catalan (AnCora) (Taulé et al., 2008)	553k\ 16k	News	Morphology: automatically tagged (reviewed); Syntax: manual annotation
French (GSD) (Guillaume et al., 2019)	300k \ 16k	News, blog, reviews, wiki	Morphology: automatically tagged (reviewed); Syntax: converted from non-UD (corrected)
Italian (ISDT) (Bosco et al., 2014)	298k \ 14k	News, legal, wiki	Originally manually annotated, converted from non-UD
Romanian (RRT) (Barbu Mititelu and Irimia, 2016)	218k \ 9k	News, legal, fiction, academic, etc.	Morphology: automatically tagged (reviewed); Syntax: converted from non-UD (corrected)
Spanish (AnCora) (Taulé et al., 2008)	568k \ 17k	News	Morphology: automatically tagged (reviewed); Syntax: manual annotation

Table 1: Overview of selected UD treebanks used in this study.

specific target patterns (for example all object dependencies), we set the match count to the minimum available across all languages (approximately 7260 matches in this case) to ensure that each treebank was equally represented.

To compile these datasets, we randomly sampled sentences from each treebank until the desired number of matches was reached. To improve consistency across samples, we applied an interquartile range (IQR) filter based on sentence length, where length was defined as the total number of nouns, verbs and adjectives in the sentence. Sentences whose length fell outside the IQR-adjusted bounds were excluded and replaced with others from the original treebank falling within those bounds.

In order to limit the amount of noise in our final results, we removed punctuation due to lack of consistency across UD, and any enhanced dependencies as not all treebanks contained them.

5 Methodology

Our method can be divided into three separate steps. First, we extract and rank the most overall salient patterns for a given linguistic phenomenon across the set of treebanks using a linear classifier. Secondly, in order to identify common and distinctive patterns, we assess if the distributions of selected patterns are statistically different from a proportionally expected distribution. Finally, for each distinctive pattern we cluster languages to find those that share the same behaviour.

5.1 Rule Extraction Method

In order to achieve our first objective of extracting a small set of important features, we employed Herrera et al.'s (2024a) method of automatically extracting and ranking fine-grained grammatical rules from the combined treebanks. We train a series of sparse logistic regression models on features of all nodes within a defined search space. The extraction task is framed as a classification problem, where the goal is to predict the likelihood of a linguistic phenomenon occurring based on its associated features. The linear model is trained using the **negative log-likelihood loss** and **L1-norm** regularization to force sparsity, which makes the model easier to interpret and the selected features less redundant.

For any given node, the search space includes the node itself, its parent and children. We consider only the universal features of UD, parts of speech (UPOS) and morpho-syntactic features (FEATS), in order to minimise noise in the decision process and make samples more comparable.

The core logic of this approach for the word order of objects is as follows: given all the object dependencies in the treebank the goal is to identify patterns P, like being a object pronoun, that better predict its position with respect to its governor. The scope S and the target question Q are manually defined, reflecting the linguistic phenomena of interest, while patterns P are selected by the ML model. The model outputs a binary indication for each feature, indicating whether it is a reliable indicator of Q (i.e., postverbal position) or $\neg Q$ (i.e., preverbal position).

Patterns selected by the classifier are not relevant to a single language, but rather to the general classification task. Importantly, the selected patterns are ranked via the regularization path determined by

the series of trained models. In each run, the regularization parameter is decreased to allow more features to be activated, providing a ranking of importance inherent to the model. For more details, refer to the cited paper.

Herrera et al. (2024a) proposed several descriptive measures to describe a corpus-based grammar rule. One of them is **precision**, defined as the probability of Q happening, given that P has already happened (e.g. out of all the objects (S), the number of objects placed after their governor (Q) or before it ($\neg Q$) that are pronouns (P)). For our purpose, we have applied this metric to all occurrences of a rule in each language as follows:

Rule	**Precision for Treebank** t
$S \implies (P \overset{\alpha\,\%}{\Longrightarrow} Q)$	$\frac{\#_t(S \wedge P \wedge Q)}{\#_t(S \wedge P)}$
$S \implies (P \overset{\alpha\,\%}{\Longrightarrow} \neg Q)$	$\frac{\#_t(S \wedge P \wedge \neg Q)}{\#_t(S \wedge P)}$

Since we are interested in how languages differ from each other, we compute the **Coefficient of Variation (CV)** to measure the dispersion over the precision scores of treebanks t:

$$\text{CV} = \frac{\text{standard deviation}(\{\text{prec}_t\})}{\text{mean}(\{\text{prec}_t\})} \quad \text{for } t \in T$$

The CV measures the spread of the sample standard deviation relative to the mean of the precision scores. Language subsets with higher CV values exhibit more diversity, whereas those with low CV values are more similar. The extracted rules can be explored and ranked not only by their precision or the predictive importance given by the linear model, but also by their dispersion.

5.2 Evaluating Distribution Proportionality

Selected rules are relevant to a given linguistic phenomenon. In practice, we capture general properties of our sample, such as the fact that the nominal object follows the verb or that being a prepositional phrase does not favour being an object. It is still unclear whether the selected rules are common or distinctive properties across languages. The aforementioned measures also do not account for the cross-lingual behaviour of each pattern. For instance, the CV indicates how the patterns P are spread across languages, but it does not reveal the significance of that spread or which languages are driving it. Inspired by Chaudhary et al. (2020), we perform a statistical test to evaluate whether this difference is significant enough to conclude that the rule applies differently across languages.

To asses whether a selected pattern P is common or distinctive, we evaluate if there is a statistically significant difference between the observed distribution and a uniform expected distribution across languages of occurrences of P (which satisfies condition Q). The expected distribution is based on the assumption that the probability of having pattern Q given P is the same across all languages (e.g., the probability of being a preverbal object when the object is a pronoun is the same for all languages). To test this, we formulate two hypotheses:

Null hypothesis: The occurrences of patterns P satisfying Q are uniformly distributed across languages, meaning each language has an equal probability of exhibiting the pattern P satisfying Q. The pattern is a **common pattern** across our sample.

Alternative hypothesis: The occurrences of patterns P satisfying Q are not uniformly distributed across languages, meaning certain languages show a higher or lower relative frequency of the pattern. The pattern is a **distinctive pattern** across our sample.

We employed a conservative significance level by applying the Bonferroni correction. We divide the base alpha level (p-value < 0.01) by the number of statistical tests performed, which is equivalent to the number of rules selected by the model.[1] We also report Cramér's V effect size. If the null hypothesis is rejected, we consider it as **distinctive rule**, otherwise we consider it a **common rule**. Rules with patterns that are not present in all languages are also considered distinctive patterns but no statistical test is computed. We test our hypothesis using χ^2 goodness-of-fit test between the expected and observed distribution, as follows:

$$\chi^2 = \sum_{t \in T} \frac{(O_t - E_t)^2}{E_t}$$

where O the observed counts and E the expected frequency under the null hypothesis. The expected values are computed as:

[1] A more exploratory approach should consider a lower significance level and using a weaker regularization parameter.

$$E_t = \#(P \wedge Q) \cdot \frac{\#_t P}{\#P}$$

This is equivalent to computing expected values that follow the same precision distribution.

Although the goodness-of-fit test separates common from distinctive patterns, it does not specify individual language behaviour relative to a selected pattern under the null hypothesis. Therefore, in order to provide a better description, we compute normalised residuals between observed and expected frequencies to identify which languages are driving the deviation and in which direction (more or less frequent than expected):

$$r_t = \frac{O_t - E_t}{\sqrt{E_t}}$$

Large residuals indicate a significant difference between the observed and expected counts. Residuals close to zero indicate that the observed counts are similar to the expected counts under the null hypothesis. More specifically, a $|r_t| > 2.58$ is highly significant.

It is important to note that statistically significant findings reflect substantial frequency differences between treebanks. These differences may arise from either genuine linguistic variation or systematic annotation discrepancies. Our conservative approach, while excluding variations of lower significance, might still capture major systematic differences, including potential annotation artifacts.

5.3 Language Clustering by Pattern

Distinctive rules apply differently across treebanks. To automatically regroup treebanks with similar behaviour, we cluster their precision scores for each rule. We employ a hierarchical and incremental approach using Euclidean distance and the Ward variance minimization algorithm to group languages that together have low variance. Early merges represent highly similar languages, while later merges, occurring at higher levels of the dendrogram, involve increasingly dissimilar groups that contribute more substantially to the overall variance.

6 Results

We present the raw results without postselection, even though some rules may be redundant. For the object order excluding the Catalan treebank (refer to Section 6.1 for the reason), we extracted 69 potential grammar rules or tendencies. Of these, 39 are common, 14 are distinctive, 4 have low-frequency occurrences, and 12 are present only in one treebank. For an overview of shared and distinctive rules after the clustering process refer to Appendix B for all languages.[2]

We evaluate the sparse linear models with the selected features on the treebank test sets, and they generalise well (refer to Appendix C). However, such an evaluation provides only limited insight into the extracted rules, as we are not interested in classification scores but in the selected features. Since it is not feasible to qualitatively evaluate all the extracted rules, we explore a few relevant rules to illustrate how to interpret them and what the limits are.

6.1 Object as a Word Order Cue

Before looking at the word order of objects, a good starting point is to check whether being an object is associated with word order in general. In other words, we examine whether and to what extent word order is predictable from being an object. To do this, we trained a classifier to select the most important features that predict word order for all pairs of nodes with a dependency relation. Selected patterns could be, as it was mentioned before, global properties of word order given our entire sample, or properties of sub-samples/treebanks.

Among the selected patterns, being an object is a salient pattern for general word order, but not, as expected, the most important one. Three rules concern the object order, all favouring the right position (see Table 2). The second pattern, involving nominal objects, is a common one in our sample and is part of a highly precise rule, where 99% of nominal objects are to the right of their governor (example 1 in Figure 1). The first and third patterns concern the objects in general and those governed by verbs. While both patterns are highly correlated, as UD objects should be governed by verbs, they show a significant difference in the distribution of precision scores per language.

A closer look reveals that the Catalan treebank is the big outlier, showing a much lower probability of objects to the right of their governors than other treebanks. This difference is an annotation artifact. Reflexive clitics are incorrectly labeled as objects in reflexive passive clauses, when they are dative oblique complements, and when they are part of a pronominal verb. This artificially multiplies the

[2]The results and code are available at https://github.com/s-herrera/contrastive-grex-syntaxfest-2025.

n rule	pattern P	rule precision	predicted order	λ	CV	type
32	X-[obj]->O	83%	XO	0.007	0.117	distinctive
33	X-[obj]->O; O[upos=NOUN]	99%	XO	0.007	0.004	common
129	X-[obj]->O; X[upos=VERB]	83%	XO	0.001	0.119	distinctive

Table 2: Word order rules for the object considering all pairs of nodes connected by a dependency. The pattern P is expressed in GREW (Guillaume, 2021) format. X corresponds to an undefined node, while O is the head of the object. λ is the value of the regularization parameter at the moment of the feature activation. For rule precision, CV, significance, see subsections 5.1 and 5.2. Results included the Catalan treebank.

number of objects, making comparisons based on pronominal object counts unrealistic.

As previously mentioned, a significant difference can reflect a real syntactic difference or a systematic difference in annotation. In this case, it is the latter. In the following, we exclude the Catalan treebank, ensuring that the rules involving pronominal objects analyzed are reliable.

In any case, this provides an overview of the order of objects, with the post-governor order being preferred. However, it is still unclear whether different languages have different strategies for ordering objects. To explore this question, we will focus on directly investigating which factors favor the order of objects.

6.2 Order of Pronominal Objects

The rule A.2 of Table 3 is the first common rule selected by the linear model (the second overall) for pronominal object order. It captures that, among all treebanks, 86% of pronominal objects of finite verbs are preverbal (example 2 in Figure 1). In other words, in all the considered Romance treebanks, the object pronouns are frequently placed to the left of finite verbs, although this is not the only order, and object position may not follow the same strategies in all languages. As stated, our significance threshold is highly conservative, and while we consider this a common pattern, the dispersion is not null. For example, Brazilian Portuguese has more postverbal object pronouns with finite verbs, because, among other reasons, finite verbs allow right object pronouns.

On the contrary, the most important rule (A.1) for our model shows a higher dispersion. It captures that 75% of object pronouns tend to be to the left of the verb and this is a good predictor of word order. However, CV is relatively high, and the observed deviation relative to the expected proportional distribution is statistically significant. It is important to note that this rule includes all types of pronouns. Rule A.10 restricts pronouns to

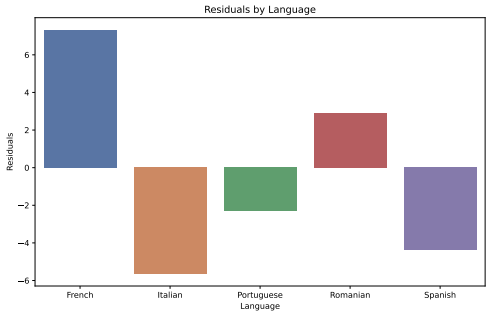


Figure 2: Residuals per language for rule A.10 of Table 3 on personal pronoun-object order (O[PronType=Prs]).

personal pronouns (*PronType=Prs*), a pattern that also favours the preverbal order. The difference in the expected distribution is also significant. If we examine the residuals for this rule in Figure 2, we see that there are more preverbal personal object pronouns in French, and Romanian, and much fewer in Italian, Brazilian Portuguese, and Spanish than expected under the null hypothesis. Being an object pronoun is then not a uniform predictor of word order in our sample, but it does discriminate well between two groups of languages where the behaviour with regards to this rule is different.

In order to further examine this difference in behaviour across the treebanks, we focus on the word order of pronominal object. The most predictive rule (B.1) states that being governed by an infinitive verb (*VerbForm=Inf*) favours the postverbal position (example 3 in Figure 1). This contrasts with the aforementioned rule involving finite verbs, listed here in the third position, where the behaviour is fairly uniform across the treebanks. Figure 3 shows that, in the case of object pronouns governed by an infinitive verb, there are clearly two clusters of languages: Italian, Portuguese and Spanish where the object pronoun immediately follows the infinitive verb in the majority of cases, and the rest for which the opposite is true (example 4 in Figure 1)

The model also extracts salient rules for less fre-

	pattern P	rule precision	predicted order	λ	CV	significance
(A) OBJECT ORDER						
1	O[upos=PRON]	75%	OV	0.1	0.19	distinctive
2	O[upos=PRON]; V[VerbForm=Fin]	86%	OV	0.026	0.09	common
10	O[PronType=Prs]	77%	OV	0.006	0.23	distinctive
(B) ORDER OF PRONOMINAL OBJECTS						
1	V[VerbForm=Inf]	56%	VO	0.7	0.76	distinctive
4	V[VerbForm=Fin]; O[PronType=Prs]	95%	OV	0.06	0.03	common
6	O[PronType=Rel]	98%	OV	0.05	0.55	common*
48	V[VerbForm=Ger]	74%	VO	0.005	0.64	common*
61	V[Mood=Imp]	85%	VO	0.003	0.17	low freq
95	O[PronType=Int]	88%	OV	0.001	0.93	low freq*
(C) ORDER OF NOMINAL OBJECTS						
1	with {Vchild[PronType=Prs]}	2.6%	OV	0.002	0.86	distinctive
5	with {V-[expl]->Vchild}	21%	OV	0.001	2.1	low freq*

Table 3: Word order rules concerning (A) object dependencies, (B) pronominal objects and (C) nominal objects. Refer to Table 2 for columns descriptions. The *with* clause in rules C.1 and C.5 should be interpreted as indicating the existence of at least another dependent that satisfies the specified condition. χ^2 test is not calculated for patterns with low frequency. *Patterns are selected by the model but are not shared by all languages.

quent phenomena. Rule B.6 indicates that the order of relative object pronouns does not vary significantly across treebanks and are almost always in a preverbal position. Rule B.95 indicates that being a interrogative pronoun favors the preverbal position. Romanian is not taken into account in these two cases because it uses a different label. Rules B.48 and B.61 favour the post-verbal position when the verb is a gerund form or is in the imperative mood, respectively. The expected frequency of these rules is low, and therefore the assumptions for computing the χ^2 statistic are not met.

Overall, we identify the main patterns of clitic object variation. Some findings challenge established knowledge. Brazilian Portuguese, for example, has a higher frequency of enclitic pronouns with infinitives than proclitics, resembling Spanish and Italian (c.f., Roberts, 2016, p. 791). However, the rules' limited expressivity, including the absence of negative conditions, prevents capturing phenomena such as clitic climbing with modal and aspectual verbs, as well as person-case constraints (Roberts, 2016, p. 789).

6.3 Order of Nominal Objects

When focusing on nominal objects, fewer rules emerge compared to other scopes. Rules indicating a post verbal position of the object have very low precision, confirming the postverbal dominant position (99%). Most of them are less reliable and difficult to understand. However, they hide some regularities and syntactic tendencies. We focus on the first rule (C.1), labeled as distinctive, which concerns nominal objects whose verbal governor has at least one more personal pronoun as a dependent. In French, the rule captures preposed nominal objects in direct interrogatives, where the verb bears a clitic. In Romanian and in Spanish, it captures clitic doubling phenomena which is not obligatory but strongly preferred when the nominal object is preposed. In addition, in Spanish, we find several passive subjects annotated wrongly

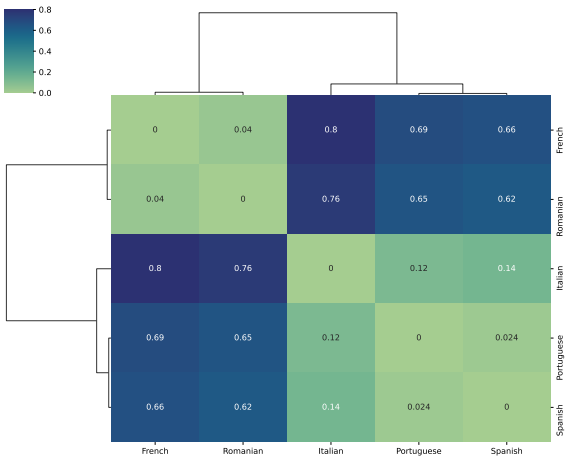


Figure 3: Clustermap of distances between precision distributions of each language for the rule B.1 of the Table 3 on pronominal object of infinitive verbs.

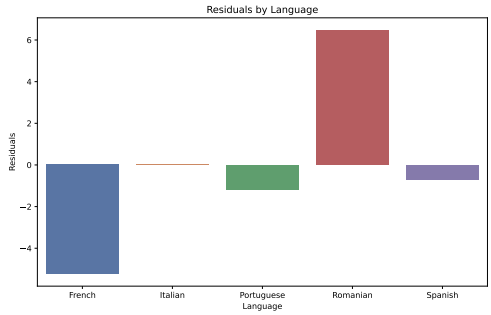


Figure 4: Residuals for rule C.1 of Table 3 on verbs having another pronoun dependant besides the object.

as objects. In Italian too, most occurrences are mediopassive constructions in *si* where the deep object promoted as a subject has been wrongly analyzed as an object (*il vero relax non si improvvisa* 'true relaxation cannot be improvised'). Residuals in Figure 4 show that Romanian in particular has more occurrences of the C.1 pattern than what is expected under the null hypothesis, while the opposite is the case for French. This is partly explained by the absence of the double clitic and the smaller number of preposed objects in French, as well as a seemingly productive strategy in Romanian. However, such low-frequency phenomena are probably influenced by the sample and its genres. This makes it impossible to draw a final conclusion about these patterns.

6.4 Annotation Inconsistencies and Error Detection

We mentioned that some results are affected by annotation inconsistencies as rules with extreme dispersion values reflect most of the time annotation inconsistencies. Our model uses these patterns in order to isolate anomalies in the sample, as they are extremely precise. This is the case, for example, for the feature *Emph=No*, which is used only in the French treebank to distinguish between emphatic and non-emphatic pronouns. It is also the case of the use of *PrepCase=Npr* in the Catalan and Spanish treebanks (both developed by the same team), to indicate that it is not a pronoun that changes form before a preposition. It should be noted that the other studied Romance languages also have this property, but do not use the feature. Sometimes the difference is not whether a label is used or not, but how it is used: the Italian treebank is the only one where the post-posted object clitics of infinitive verbs are annotated with the feature *Clitic=Yes*, even though this is a characteristic shared by all studied languages, except for French.

Cases like missing features in one or more treebanks are extreme, but annotation inconsistencies also arise from different annotation strategies. This is the case of rule C.5 of the table concerning clitic doubling. For instance, Romanian uses the relation `expl` (example 5 in Figure 1), while Catalan and Spanish use `obj`. In practice, the UD guidelines do not encourage the doubling of the *obj* dependency.

This illustrates how our method additionally captures regular error annotation, sharing thus conceptual ground with other error detection approaches. Related approaches use ML models trained on existing annotations to highlight inconsistencies between predicted and observed labels (Aquino et al., 2025) or compare annotations to predefined grammatical rules (Oepen et al., 2004). Hybrid systems combine both strategies (Agrawal et al., 2013; Ambati et al., 2011), while others identify consistent sequences in order to extract reliable patterns before extracting anomalies (Dickinson, 2015).

Our approach can be reframe as an error detection method for harmonizing annotations across corpora. First, we use ML techniques to extract salient syntactic patterns, treating those unique to one corpus as potential inconsistencies. Subsequently, statistical tests are used to analyze variations in shared patterns, which can signal either genuine linguistic differences or annotation discrepancies. Interpretation depends on corpus similarity: in closely related treebanks, variations more likely indicate annotation errors, while for distant corpora, linguistic knowledge is required to determine the cause.

7 Takeaways

We present a comprehensive framework to extract contrastive grammar rules and tendencies from syntactic treebanks. It allows us to induce a concise set of grammar rules that reflect statistical differences between closely related languages. A more exploratory and less conservative approach is possible by adjusting a few hyperparameters. Indirectly, the method can be used to find annotation inconsistencies across treebanks. Experiments also show the limitations of doing automatic grammar extraction and linguistic analysis with universal collections. For this reason, we encourage UD contributors maintaining related language treebanks to work together to harmonise annotation choices.

Limitations

Our sample presents three potential limitations. Firstly, our Romance sample does not cover all the language family diversity. Additionally, we only focus on object clitics, leaving out locative or genitive clitics. More critically, the heterogeneity of genres present within the employed treebanks introduces a confounding variable. Weak statistical trends may be attributable to variations or properties inherent to specific genres, rather than solely reflecting inherent linguistic characteristics. Third, the sample suffers from annotation inconsistencies and errors, introducing some noise in our results. Finally, concerning our methodology, it is important to emphasize that extracted grammar rules should be interpreted as having a predictive or directional nature, and not as causal factors.

Acknowledgments

This work is supported by the Université Paris Nanterre and the French National Research Project Autogramm (ANR-21-CE38-0017). It has benefited from discussions within CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology). Finally, we would like to thank Guillaume Bonfante for his valuable insights and the three anonymous reviewers for their constructive feedback.

References

Bhasha Agrawal, Rahul Agarwal, Samar Husain, and Dipti M Sharma. 2013. An automatic approach to treebank error detection using a dependency parser. In *Computational Linguistics and Intelligent Text Processing*, Lecture notes in computer science, pages 294–303. Springer Berlin Heidelberg, Berlin, Heidelberg.

Diego Alves. 2025. Information theory and linguistic variation: A study of Brazilian and European Portuguese. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 9–19, Abu Dhabi, UAE. Association for Computational Linguistics.

Diego Alves, Božo Bekavac, Daniel Zeman, and Marko Tadić. 2023. Analysis of corpus-based word-order typological methods. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 36–46, Washington, D.C. Association for Computational Linguistics.

Bharat Ram Ambati, Rahul Agarwal, Mridul Gupta, Samar Husain, and Dipti Misra Sharma. 2011. Error detection for treebank validation. In *Proceedings of the 9th Workshop on Asian Language Resources*, pages 23–30, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Angelina A. Aquino, Lester James V. Miranda, and Elsie Marie T. Or. 2025. The ud-newscrawl treebank: Reflections and challenges from a large-scale tagalog syntactic annotation project. *Preprint*, arXiv:2505.20428.

Verginica Barbu Mititelu and Elena Irimia. 2016. Linguistic data retrievable from a treebank. In *Proceedings of the Second International Conference on Computational Linguistics in Bulgaria (CLIB 2016)*, pages 19–27, Sofia, Bulgaria. Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

Q. Bertrand, Q. Klopfenstein, P.-A. Bannier, G. Gidel, and M. Massias. 2022. Beyond l1: Faster and better sparse models with skglm. In *NeurIPS*.

Cristina Bosco, Felice Dell'Orletta, and Simonetta Montemagni. 2014. The evalita 2014 dependency parsing task. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014 9-11 December 2014, Pisa*. pisa university press.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. In G. Bouma, I. Krämer, and J. Zwarts, editors, *Cognitive Foundations of Interpretation*, pages 69–94. Royal Netherlands Academy of Arts and Sciences, Amsterdam.

Joan Bresnan and Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in american and australian varieties of english. *Language*, 86(1):168–213.

Aditi Chaudhary, Antonios Anastasopoulos, Adithya Pratapa, David R. Mortensen, Zaid Sheikh, Yulia Tsvetkov, and Graham Neubig. 2020. Automatic extraction of rules governing morphological agreement. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5212–5236, Online. Association for Computational Linguistics.

Aditi Chaudhary, Zaid Sheikh, David R Mortensen, Antonios Anastasopoulos, and Graham Neubig. 2022. Autolex: An automatic framework for linguistic exploration. *Preprint*, arXiv:2203.13901.

Hee-Soo Choi, Bruno Guillaume, Karën Fort, and Guy Perrier. 2021. Investigating dominant word order on Universal Dependencies with graph rewriting. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 281–290, Held Online. INCOMA Ltd.

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

Markus Dickinson. 2015. Detection of annotation errors in corpora. *Language and Linguistics Compass*, 9(3):119–138.

Magali Duran, Lucelene Lopes, Maria das Graças Nunes, and Thiago Pardo. 2023. The dawn of the porttinari multigenre treebank: Introducing its journalistic portion. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 115–124, Porto Alegre, RS, Brasil. SBC.

Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2021. Typometrics: From Implicational to Quantitative Universals in Word Order Typology. *Glossa: a journal of general linguistics (2021-...)*, 6(1):17.

Stefan Th. Gries, Marlies Jansegers, and Viola G. Miglio. 2020. Quantitative methods for corpus-based contrastive linguistics. In Renata Enghels, Bart Defrancq, and Marlies Jansegers, editors, *New Approaches to Contrastive Linguistics*, pages 53–84. De Gruyter.

Bruno Guillaume. 2021. Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In *EACL 2021 - 16th conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Kiev/Online, Ukraine.

Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. 2019. Conversion et améliorations de corpus du français annotés en Universal Dependencies. *Revue TAL : traitement automatique des langues*, 60(2):71–95.

Hilde Hasselgård. 2020. Corpus-based contrastive studies: Beginnings, developments and directions. *Languages in Contrast*, 20(2):184–208.

Santiago Herrera, Caio Corro, and Sylvain Kahane. 2024a. Sparse logistic regression with high-order features for automatic grammar rule extraction from treebanks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15114–15125, Torino, Italia. ELRA and ICCL.

Santiago Herrera, Ioana-Madalina Silai, Bruno Guillaume, and Sylvain Kahane. 2024b. Building quantitative contrastive grammars from syntactic treebanks. *Langues & Langages à la croisée des Disciplines (LLcD)*.

Natalia Levshina. 2022. Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology*, 26(1):129–160.

Badr Moufad, Pierre-Antoine Bannier, Quentin Bertrand, Quentin Klopfenstein, and Mathurin Massias. 2023. skglm: improving scikit-learn for regularized generalized linear models.

Dmitry Nikolaev, Ofir Arviv, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Maria Saeboe, and Omri Abend. 2020. Fine-grained analysis of cross-linguistic syntactic divergences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1159–1176, Online. Association for Computational Linguistics.

Stephan Oepen, Dan Flickinger, and Francis Bond. 2004. Towards holistic grammar engineering and testing. *Beyond shallow analyses–formalisms and statistical modelling for deep analysis (workshop at the first international joint conference on natural language processing (IJCNLP-04)*.

Robert Östling. 2015. Word order typology through multilingual word alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 205–211, Beijing, China. Association for Computational Linguistics.

Ian Roberts. 2016. Object clitics. In *The Oxford Guide to the Romance Languages*. Oxford University Press.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Juliette Thuilier. 2012. *Contraintes préférentielles et ordre des mots en français*. Ph.D. thesis, Université Paris-Diderot - Paris VII, Paris, France.

A Sparse Logistic Regression Hyperparameters

We use skglm (Bertrand et al., 2022; Moufad et al., 2023) Sparse Logistic Regression implementation. We use default hyperparameters, except for the regularization parameter, which we vary from 0.1 to 0.001 in 100 steps.

B Distinctive Rule Distributions

To analyze the interrelationships between languages according to their distinctive rules, we present two series of matrices. The first series consists of co-occurrence matrices (Figure 5), which quantify how often languages fall into the same cluster for a given rule. The second series, composed of difference matrices (Figure 6), shows the inverse: how frequently languages are separated into different clusters. The matrices reveal the nuanced relationships across treebanks. For example, in the case of pronominal object word order (28 rules in total), Spanish and French only co-cluster in eight rules and are separated in 20, primarily due to their differing behavior in infinitive and gerund constructions. Conversely, Romanian and French are often grouped together. This is not due to a strong resemblance between the two languages, but rather because they are both systematically different from the more homogeneous Italo-Iberian group.

(a) Object WO (14)

	French	Italian	Portuguese	Romanian	Spanish
French	-				
Italian	2	-			
Portuguese	0	11	-		
Romanian	11	4	2	-	
Spanish	10	6	3	10	-

(b) Pronominal Object WO (28)

	French	Italian	Portuguese	Romanian	Spanish
French	-				
Italian	3	-			
Portuguese	0	24	-		
Romanian	25	3	0	-	
Spanish	8	22	20	8	-

(c) Nominal Object WO (2)

	French	Italian	Portuguese	Romanian	Spanish
French	-				
Italian	2	-			
Portuguese	2	2	-		
Romanian	0	0	0	-	
Spanish	2	2	2	0	-

Figure 5: Co-occurrence matrices of shared rules for object, pronominal object, and nominal object word order across treebanks. Languages are ordered alphabetically.

(a) Object WO (14)

	French	Italian	Portuguese	Romanian	Spanish
French	-				
Italian	12	-			
Portuguese	13	3	-		
Romanian	3	10	12	-	
Spanish	4	8	11	4	-

(b) Pronominal Object WO (28)

	French	Italian	Portuguese	Romanian	Spanish
French	-				
Italian	24	-			
Portuguese	28	4	-		
Romanian	3	25	28	-	
Spanish	20	6	8	20	-

(c) Nominal Object WO (2)

	French	Italian	Portuguese	Romanian	Spanish
French	-				
Italian	0	-			
Portuguese	0	0	-		
Romanian	2	2	2	-	
Spanish	0	0	0	2	-

Figure 6: Difference matrices of distinctive rules for object, pronominal object, and nominal object word order across treebanks. Languages are ordered alphabetically.

C Model Evaluation

Table 4, on the next page, shows the evaluation scores for the selected rules on the training and test sets. Sparse linear models generalize well across all test sets. Performance scores for the nominal object order model were excluded due to extreme class imbalance. Macro-averaged measures are reported. The simplicity of the task makes the evaluation scores relatively uninformative.

Dataset	Accuracy	Precision	Recall	F1-score	Majority-class Baseline
GENERAL WORD ORDER					0.57
Train	0.93	0.93	0.94	0.93	
Test	0.93	0.93	0.93	0.93	
OBJECT ORDER					0.87
Train	0.98	0.97	0.95	0.96	
Test	0.98	0.97	0.95	0.96	
ORDER OF PRONOMINAL OBJECTS					0.72
Train	0.98	0.97	0.97	0.97	
Test	0.97	0.97	0.96	0.96	

Table 4: Scores on train and test (25%), with the selected features, excluding the Catalan treebank.

A Quantitative Study of Syntactic Complexity across Genres: Dependency Distance in English and Chinese

Yaqin Wang
Center for Linguistics and Applied Linguistics
Guangdong University of Foreign Studies
yqinw688@gmail.com

Abstract

This study investigates syntactic complexity in fiction and news genres by analyzing mean dependency distances (MDD) across controlled sentence lengths in English and Chinese corpora. Results show that English fiction exhibits greater MDD than news, while Chinese fiction shows the reverse. More complex syntactic structures, i.e., complex coordination structures, are found in English fiction texts than in news writing. In contrast, Chinese news writing relies more on nominal modification and prepositional phrases that create long-distance dependencies than fiction texts. These findings show deviations from uniform correlations between genre formality and syntactic complexity across languages.

1 Introduction

Syntactic complexity of genres has been given explicit attention in the field of quantitative syntax (Biber & Conrad, 2019). Measurements, such as dependency distance and sentence length, have been employed to compare genres' syntactic difficulty (Oya, 2011; Wang & Liu, 2017, 2022; Chen & Kubát, 2024). Dependency distance, based on the concept of dependency grammar which describes the asymmetric syntactic relationship between two words concerned, refers to the linear distance between the governor and the dependent (Liu, 2008; Liu et al., 2017).

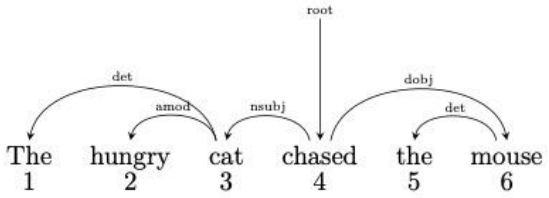


Figure 1: Dependency analysis of an example sentence.

For example, in the pair (*cat*, *chased*) in Figure 1, *chased* is the governor and *cat* the dependent, the type shown as the label above the arc is *nsubj* (nominal subject). The dependency distance of this pair is $|3 - 4| = 1$. The notion of dependency distance signals the difficulty level of a given sentence. The greater the dependency distance, the more difficult the syntactic analysis of a sentence is (Gibson, 1998; Liu et al., 2017).

Recent advances in quantitative dependency syntax (Futrell et al., 2020; Yadav et al., 2022) have established more complicated frameworks for cross-linguistic analysis. These developments raise further questions about how dependency distance interacts with genre characteristics. Formal genres, such as informative texts, are presumably expected to have greater syntactic complexity, shown as the longer sentences and greater dependency distances. However, as investigated by Wang & Liu (2017), when sentence lengths are controlled, imaginative texts show greater dependency distances, which indicates more complex syntactic structures than do informative texts. They did not further conduct the statistical analysis of the descriptive statistics though, thus leaving questions regarding statistical significance unresolved. More recently, Chen & Kubát (2024) found that short stories in Czech National Corpus have greater dependency distances than newspapers, to some extent challenging the assumption that more formal texts

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 39–46

necessarily contain more syntactically difficult sentences. Wang (2020) found that Chinese news' dependency distances are greater than fiction, albeit without controlling sentence lengths. These findings suggest that the relationship between genre formality and syntactic complexity warrants deeper investigation.

Thus, the present study intends to explore mean dependency distances of genres within controlled sentence lengths, with corpora from both English and Chinese. Research questions are as follows:

1) Do fiction genres demonstrate greater syntactic complexity than news genres when sentence length is controlled, and is this pattern language-specific or a cross-linguistic phenomenon?

2) Which dependency types, taking into account of both mean dependency distances and frequencies, contribute most to the observed patterns across genres and languages?

2 Methods and Materials

Texts in the English corpus were collected from FLOB corpus. FLOB (Freiburg-LOB Corpus of British English) is a British written English corpus, mainly including 15 genres. This paper extracted three types of news and three types of fiction genres, specifically including press reportage (A), press editorials (B), press review (C), general fiction (K), mystery and detective fiction (L), and humor (R). Ten texts are collected from each genre (except for humor fiction, which only has 9 texts in the corpus), resulting in a total of 59 English texts.

To ease English-Chinese register comparison, the Chinese corpus in this paper uses the Lancaster Corpus of Mandarin Chinese (LCMC), which was designed as a Chinese counterpart to the FLOB corpus following similar sampling principles and genre categories. Consistent with the English selection, this paper selected six Chinese genres—A, B, C, K, L, and R—comprising three news registers and three fiction registers, for a total of 59 Chinese texts.

According to the previous experience (Wang & Liu, 2017), the sentences from each text were grouped by length ranges from 1-5 words, 6-10 words, etc., as shown in Table 1. To ensure reliable statistical comparison, the present study primarily focuses on ranges up to 36-40 words, where each range contains at least 10 sentences in each genre and 200 sentences in all genres within one language. Detailed sentence counts for each genre after the data extraction are provided in Appendix A due to space limitations.

Length Range of English genres	Sent. Count	Length Range of Chinese genres	Sent. Count
1-5	936	1-5	375
6-10	1574	6-10	826
11-15	1288	11-15	1000
16-20	952	16-20	882
21-25	777	21-25	756
26-30	558	26-30	510
31-35	397	31-35	311
36-40	278	36-40	232
41-45	202	41-45	150
46-50	109	46-50	99
51-55	61	51-55	77
56-60	41	56-60	37
61-65	24	61-65	20
66-70	12	66-70	17
71-75	11	71-75	7
76-80	4	76-80	9
81-85	5	81-85	6
86-90	2	86-90	7
91-95	1	91-95	2
96-100	1	96-100	3
126-130	1	101-105	1
136-140	1	111-115	1

Table 1: Sentence count in English and Chinese genres.

Following the data extraction process, the Stanford parser (De Marneffe & Manning, 2008) (version 3.9) was employed to output the typed-dependency relations. Due to the relatively low accuracy rate of Stanford parser for Chinese, the current research, manually modified the parsed Chinese treebank.

In measuring the dependency distance of a sentence and of a sample, i.e., a large corpus, Liu et al., (2009) propose several methods. Let $W_1 \ldots W_i \ldots W_n$ be a word string. For any dependency relation between the words W_a and W_b ($1 \leq a \leq b \leq n$), suppose W_a is a governor and W_b is its dependent. The dependency distance (hereafter

referred to as *DD*) between them is defined as their difference, i.e., $|a - b|$. The mean dependency distance (hereafter referred to as MDD) of an entire sentence can be calculated as the average of dependency distances. For instance, MDD of the example sentence in Figure 1 is (2+1+1+2+1) / 5 = 1.4. In the current research, to analyze sentences of similar lengths across different genres, the author calculated the MDD for groups of sentences (similar to a small corpus) within specific length ranges. The MDD of a corpus can be defined as:

$$MDD = \frac{1}{n-s}\sum_{i=1}^{n-s}|DD_i| \quad (1)$$

Here n is the total number of words, s is the total number of sentences and DD_i is the *DD* of the i-th syntactic link of the sample.

When the dependency type was investigated, two features, i.e., their dependency distances and frequencies were considered. The present study quantified those two factors by calculating the relative contribution of MDDs of each dependency type. Here is the formula for calculating relative contribution of each dependency type:

$$\textbf{\textit{Raw Contribution}} = Proportion \times MDD_{type} \quad (2)$$

where *Proportion* is the dependency type's frequency relative to all dependencies in the corpus, and MDD_{type} is the mean dependency distance for that specific dependency type.

$$\textbf{\textit{Relative Contribution (\%)}} = (Raw\ Contribution\ /\ Total\ MDD_{corpus}) \times 100 \quad (3)$$

where $Total\ MDD_{corpus}$ is the overall mean dependency distance across all dependency types in the corpus. This normalizes contributions as percentages of overall MDD, allowing for direct comparison across different dependency types and corpora.

3 Results and Discussion

3.1 MDD distribution

The distribution of MDDs across different sentence length ranges in English and Chinese genres is displayed in Appendix A and Figure 2 and Figure 3.

Figure 2a and Figure 3a show the average MDD deviation from the overall MDD. Figure 2b and Figure 3b shows deviation of each genre's MDD from the overall average across all genres, with red showing positive deviation (higher than average MDD) and blue showing negative deviation (lower than average MDD).

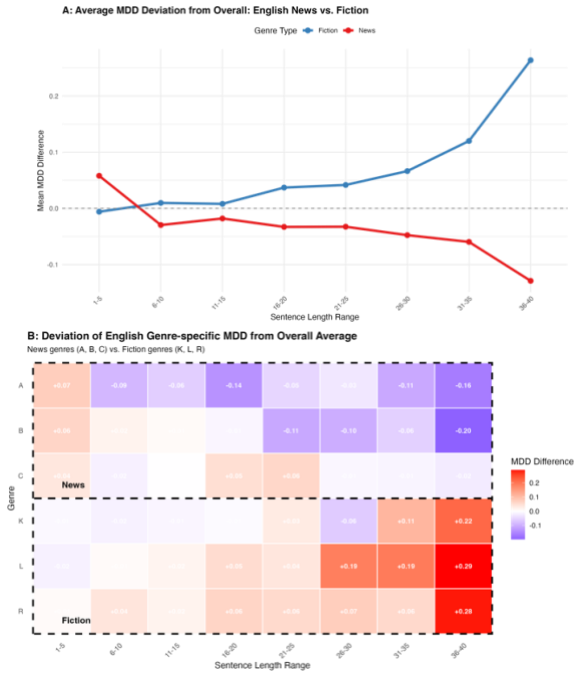


Figure 2: MDD distribution in English fiction and news

As shown in Figure 2, the gap between English fiction and news becomes greater as sentence length increases, with the maximum difference occurring in the 36–40 word range. Fiction (represented by genres K, L, and R) starts slightly below average for the shortest sentences but then rises above average as sentence length increases. For instance, genre R (humor) exhibits the most obvious positive deviation in the 36-40 word ranges. News (as shown by genres A, B, C) begins above average for very short sentences (1-5 words) but quickly falls below average and continues to decrease relative to the overall mean. Genre A (press reportage) showing the most consistent negative deviation across almost all length ranges.

As shown in Figure 3 and Appendix A, Chinese fiction and news texts display a different pattern from English genres. Chinese news genres (A, B, C) generally maintain above-average MDD across nearly all sentence length ranges, with slightly negative deviations in genre A (press reportage). Chinese fiction genres (K, L, R) generally display

below-average MDD values. This pattern is reversed from what is shown in English, where fiction texts trend toward higher-than-average MDD in longer sentences.

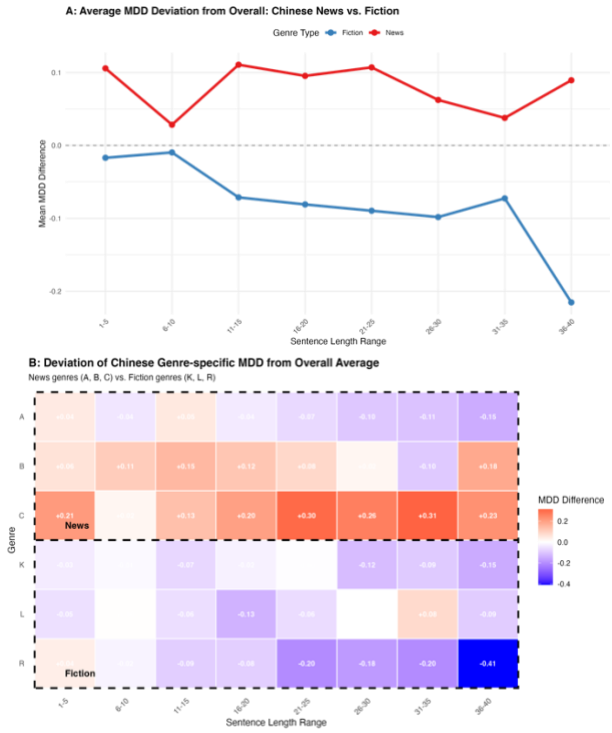


Figure 3: MDD distribution in Chinese fiction and news

A linear model was then fitted to examine the relationship between MDD, sentence length, genre, and language[1]. The model result of predicting MDD from the interaction of language, i.e., English and Chinese, genre type, i.e., news or fiction, and sentence length range, i.e., the average value of the sentence length range, shows an excellent fit with an R^2 value of 0.958 (F (7,88) = 289.03, $p < 0.001$).

Above all, there is a marginal significant main effect of language ($p = 0.06$), indicating differences in MDD between English and Chinese texts are marginally significant. Further, sentence length shows a highly significant main effect ($p < 0.001$), confirming that MDD consistently increases with sentence length across languages and genres. This confirms the previous finding that dependency distances are more optimized in English than in Chinese (Ferrer-I-Cancho et al., 2022).

The two-way interaction between genre and length range is not significant ($p = 0.31$), indicating that the relationship between sentence length and MDD is consistent across different genres when not considering language differences. However, the three-way interaction between language, genre type, and sentence length range is significant (F (1,88) = 6.83, $p = 0.01$). This interaction confirms the observation that the relationship between genre and sentence length differs significantly between English and Chinese.

The model coefficients support this interpretation as well. For Chinese fiction (the reference level), the coefficient for sentence length range was 0.060 ($p < 0.001$), indicating a strong positive relationship between sentence length and MDD. The interaction between language and length is also significant ($p = 0.02$), with a negative coefficient (-0.009), suggesting that sentence length has a different effect on MDD in English compared to Chinese. Further, the interaction between language, genre type, and sentence length range had a significant negative coefficient (-0.014, $p = 0.01$), demonstrating that the genre difference in how MDD changes with increasing sentence length is significantly different between languages.

3.2 Distribution of dependency types

To understand the underlying causes why these MDD differences emerge in Chinese and English, a detailed analysis of dependency types in longer sentences (ranges 21-25, 26-30, 31-35, and 36-40 words) was conducted. This analysis involved calculating both the MDDs and frequencies of various dependency types across these sentence lengths.

[1] The current research chose a standard linear model rather than a mixed-effects model based on a comparative analysis of both approaches. The random effects in the initial mixed-effects model (with random intercepts for genre nested within language, i.e., K, L, R, A, B, C) accounted for only 7.48% of the total variance, with a random intercept variance of 0.001486 compared to a residual variance of 0.018384. The improvement in explanatory power was minimal, with the R^2 increasing only from 0.9587 (fixed effects only) to 0.9618 (with random effects) – a mere 0.32% improvement. These statistics indicate that the simpler linear model provides equivalent explanatory power. Besides, due to small sample size in each genre, i.e., eight observations, as a random effect, the present study chose the linear model to avoid potential overfitting and to ensure more reliable parameter estimates.

Genre	Dep. Type	Prop.	MDD	Rel. Contribution
News	*prep*	0.125	2.66	12.14
News	*pobj*	0.119	2.24	9.74
News	*nsubj*	0.081	3.09	9.07
News	*conj*	0.036	6.96	9.02
News	*det*	0.106	1.67	6.43
News	*cc*	0.034	5.08	6.39
News	*ccomp*	0.016	7.95	4.65
Fiction	*conj*	0.042	8.18	12.04
Fiction	*prep*	0.106	2.48	9.18
Fiction	*cc*	0.042	6	8.76
Fiction	*nsubj*	0.108	2.13	8.07
Fiction	*pobj*	0.102	2.06	7.32
Fiction	*ccomp*	0.022	7.28	5.48
Fiction	*det*	0.099	1.53	5.29

Table 2 Top seven contributing dependency types in English genres.

Genre	Dep. Type	Prop.	MDD	Rel. Contribution
News	*conj*	0.088	8.62	21.38
News	*dobj*	0.088	3.77	9.38
News	*prep*	0.046	7.12	9.24
News	*nsubj*	0.089	3.27	8.25
News	*advmod*	0.088 6	2.78	6.95
News	*nn*	0.128	1.6	5.77
News	*ccomp*	0.03	5.5	4.66
Fiction	*conj*	0.104	10.07	31.61
Fiction	*nsubj*	0.122	2.68	9.87
Fiction	*advmod*	0.110	2.34	7.77
Fiction	*dobj*	0.085	2.84	7.27
Fiction	*ccomp*	0.052	4.42	6.87
Fiction	*prep*	0.039	4.96	5.84
Fiction	*nn*	0.055	1.49	2.47

Table 3 Top seven contributing dependency types in English genres.

Table 2 and Table 3 present the seven dependency types that contribute most significantly to the overall MDD in news and fiction genres for Chinese and English, respectively. In English, fiction texts show more complex coordination structures than news, with conjunctions (*conj*) having greater MDD (8.18 vs. 6.96) and higher frequency (0.042 vs 0.036) and coordinating conjunctions (*cc*) showing both higher frequency (0.042 vs. 0.034) and MDD (6.00 vs. 5.08). These patterns, to some degree, explain why English fiction demonstrates greater overall MDD than news.

In Chinese, the pattern differs. While conjunctions in fiction have higher MDD (10.07 vs. 8.62) and frequency (0.104 vs. 0.088) than in news, Chinese news writing shows greater complexity in other structures. Prepositional phrases (*prep*) in news have greater MDD (7.12 vs. 4.96) and higher frequency (0.046 vs. 0.039). Direct objects (*dobj*) show similar frequency across genres but higher MDD in news (3.77 vs. 2.84). News writing also employs extensive nominal modification (*nn*: 0.128 vs 0.055), which partly contributes to information density. These patterns likely contribute to Chinese news having higher overall MDD than fiction.

3.3 Discussion

The present study reveals a clear cross-linguistic divergence in the relationship between genre and syntactic complexity, as measured by mean dependency distance (MDD). In English, fiction genres demonstrate higher MDD values than news genres when sentence length is controlled. In contrast, Chinese shows the opposite pattern: news genres exhibit greater MDD values than fiction. These findings deviate from the assumption that syntactic complexity correlates with genre formality across languages. The English findings align with Wang & Liu's (2017) observation that when sentence lengths are controlled, imaginative texts show greater dependency distances, which indicates more complex syntactic structures than do informative texts. Furthermore, the present study provides the statistical significance test that Wang & Liu (2017) acknowledged was missing from their work, demonstrating that these differences are indeed statistically significant ($p < 0.01$). These findings also complement Chen & Kubát's (2024) work on Czech, which found that short stories in Czech National Corpus have greater dependency distances than newspapers. This supports that more formal texts do not necessarily contain more syntactically difficult sentences across all languages.

In English, the higher MDD in fiction genres is to some degree driven by the prevalence of complex coordination. Specifically, dependency types such as *conj* and *cc* have greater distances and higher frequencies in fiction than in news, contributing the most to overall MDD. This suggests that English fiction frequently employs longer-range syntactic dependencies, likely reflecting narrative style and the use of more elaborated clause structures. This distinction is consistent with prior observations that formal English writing often emphasizes phrasal over clausal elaboration (Biber & Gray, 2010).

In Chinese, the pattern reverses: news genres produce greater MDD values than fiction. This is shown by the syntactic features of Chinese news writing, which often include extensive nominal modification and prepositional phrases that introduce long-distance dependencies. For example, *prep*, *dobj*, and *nn* dependencies in Chinese news texts tend to have higher MDD values and occur more frequently compared to fiction texts. These constructions allow news writers to convey dense informational content, resulting in longer syntactic dependencies.

While the analysis demonstrates clear genre effects, it should also be noted that sentence length remains the dominant factor affecting dependency distance, as shown by the linear model results. This finding is consistent with previous studies (Ferrer-I-Cancho & Liu, 2014; Jiang & Liu, 2015), which have reported the significant relationship between sentence length and dependency distance. The genre and language effects operate as modulating factors on this main relationship—altering the rate of MDD increase as sentences grow longer rather than overriding the basic length-distance correlation.

These findings contribute to our understanding of how dependency distance minimization principles (Liu, 2008; Futrell et al., 2015; Liu et al., 2017) operate under different genre contexts and typological constraints. While minimizing dependency distance may reduce cognitive processing load (Gibson, 1998; Hawkins, 2004), the results suggest that this principle is not applied uniformly across genres or languages. Instead, it is adapted in ways that reflect the communicative goals and syntactic norms of each genre-language combination.

4 Conclusion

Overall, the current research shows that metrics such as dependency distance can capture genre-sensitive patterns of syntactic complexity, but their interpretation is better understood in the context of both cross-linguistic variation and genre-specific conventions. These insights may shed a new light on genre analysis and highlight the value of quantitative syntax in uncovering structural differences across languages. Future work could broaden this investigation by combining a wider range of genres and additional typologically diverse languages to further assess the cross-linguistic and cross-genre consistency of these findings.

References

Douglas Biber and Susan Conrad. 2019. *Register, genre, and style*. Cambridge University Press, Cambridge, UK.

Douglas Biber and Bethany Gray. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, 9(1):2-20. https://doi.org/10.1016/j.jeap.2010.01.001.

Xinying Chen and Miroslav Kubát. 2024. Quantifying Syntactic Complexity in Czech Texts: An Analysis of Mean Dependency Distance and Average Sentence Length Across Genres. *Journal of Quantitative Linguistics*, 31(3):260-273. https://doi.org/10.1080/09296174.2024.2370459.

Marie-Catherine De Marneffe and Christopher D. Manning. 2008. *Stanford Typed Dependencies Manual*, pp.pages 338–345. Technical report, Stanford University.

Ramon Ferrer-i-Cancho and Haitao Liu. 2014. The risks of mixing dependency lengths from sequences of different length. *Glottotheory*, 5(2):143-155. https://doi.org/10.1515/glot-2014-0014.

Ramon Ferrer-i-Cancho, Carlos Gómez-Rodríguez, Juan L. Esteban, and Lluís Alemany-Puig. 2022. Optimality of syntactic dependency distances. *Physical Review E*, 105(1):014308. https://doi.org/10.1103/PhysRevE.105.014308.

Richard Futrell, Roger P. Levy, and Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language*, 96(2):371-412. https://doi.org/10.1353/lan.2020.0024.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33):10336-10341. http://dx.doi.org/10.1073/pnas.1502134112.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68 (1),):1—76. http://dx.doi.org/10.1016/S0010-0277(98)00034-1.

John A. Hawkins, 2004. *Efficiency and Complexity in Grammars*. Oxford University Press, Oxford, UK.

Jingyang Jiang and Haitao Liu. 2015. The effects of sentence length on dependency distance, dependency direction and the implications-based on a parallel English-Chinese dependency treebank. *Language Sciences*, 50 :93-104. https://doi.org/10.1016/j.langsci.2015.04.002.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159-191. http://dx.doi.org/10.17791/jcs.2008.9.2.159.

Haitao Liu, Richard Hudson, and Zhiwei Feng. 2009. Using a Chinese Treebank to measure dependency distance. *Corpus Linguistics and Linguistic Theory*, 5(2):161-174. http://doi.org/10.1515/CLLT.2009.007.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of life reviews*, 21:171-193. https://doi.org/10.1016/j.plrev.2017.03.002.

Masanori Oya. 2011. Syntactic dependency distance as sentence complexity measure. In: *Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics*, pages 313–316.

Yaqin Wang. 2020. *Quantitative Syntactic Features of Genres from Multi-perspectives* (in Chinese). Zhejiang University dissertation, Hangzhou.

Yaqin Wang and Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Sciences*, 59: 135-147. http://dx.doi.org/10.1016/j.langsci.2016.09.006.

Yaqin Wang and Haitao Liu. 2022. Creativity complicates tweets: a quantitative lens on syntactic characteristics of twitter. *Digital Scholarship in the Humanities*, 37(1):264-279. https://doi.org/10.1093/llc/fqab028.

Himanshu Yadav, Shubham Mittal, and Samar Husain. 2022. A reappraisal of dependency length minimization as a linguistic universal. *Open Mind*, 6:147-168. https://doi.org/10.1162/opmi_a_00060.

Appendix A MDDs in different sentence length range across genres in English and Chinese

English	SL range	Sentence Count	MDD
All genres	1-5	936	1.412
All genres	6-10	1574	1.818
All genres	11-15	1288	2.152
All genres	16-20	952	2.393
All genres	21-25	777	2.581
All genres	26-30	558	2.736
All genres	31-35	397	2.906
All genres	36-40	278	3.064
A	1-5	12	1.486
A	6-10	74	1.728
A	11-15	123	2.091
A	16-20	163	2.257
A	21-25	163	2.531
A	26-30	120	2.706
A	31-35	92	2.798
A	36-40	74	2.9
B	1-5	70	1.475
B	6-10	138	1.837
B	11-15	186	2.16
B	16-20	151	2.381
B	21-25	129	2.472
B	26-30	113	2.634
B	31-35	83	2.848
B	36-40	51	2.866
C	1-5	29	1.449
C	6-10	68	1.8
C	11-15	102	2.151
C	16-20	106	2.442
C	21-25	129	2.642
C	26-30	98	2.725
C	31-35	91	2.893
C	36-40	62	3.039
K	1-5	324	1.404
K	6-10	534	1.801
K	11-15	303	2.14

K	16-20	164	2.387
K	21-25	121	2.612
K	26-30	63	2.672
K	31-35	43	3.016
K	36-40	26	3.281
L	1-5	336	1.395
L	6-10	456	1.825
L	11-15	355	2.169
L	16-20	216	2.446
L	21-25	127	2.619
L	26-30	74	2.924
L	31-35	47	3.099
L	36-40	34	3.354
R	1-5	165	1.419
R	6-10	304	1.857
R	11-15	219	2.171
R	16-20	152	2.457
R	21-25	108	2.637
R	26-30	90	2.811
R	31-35	41	2.963
R	36-40	31	3.348
Chinese	SL range	Sentence count	MDD
All genres	1-5	375	1.507
All genres	6-10	826	2.105
All genres	11-15	1000	2.519
All genres	16-20	882	2.847
All genres	21-25	756	3.15
All genres	26-30	510	3.403
All genres	31-35	311	3.613
All genres	36-40	232	3.87
A	1-5	35	1.551
A	6-10	125	2.061
A	11-15	169	2.567
A	16-20	164	2.811
A	21-25	119	3.084
A	26-30	98	3.306
A	31-35	50	3.507
A	36-40	48	3.725
B	1-5	10	1.568
B	6-10	81	2.214
B	11-15	129	2.669
B	16-20	151	2.969
B	21-25	145	3.234
B	26-30	85	3.423
B	31-35	55	3.518
B	36-40	40	4.053
C	1-5	10	1.719
C	6-10	61	2.125
C	11-15	98	2.653
C	16-20	104	3.047
C	21-25	93	3.453
C	26-30	107	3.667
C	31-35	60	3.927
C	36-40	59	4.1
K	1-5	116	1.473
K	6-10	183	2.094
K	11-15	212	2.449
K	16-20	170	2.822
K	21-25	136	3.146
K	26-30	74	3.286
K	31-35	59	3.52
K	36-40	31	3.722
L	1-5	90	1.454
L	6-10	164	2.108
L	11-15	156	2.461
L	16-20	124	2.712
L	21-25	137	3.089
L	26-30	74	3.403
L	31-35	52	3.692
L	36-40	38	3.781
R	1-5	114	1.543
R	6-10	212	2.084
R	11-15	236	2.433
R	16-20	169	2.764
R	21-25	126	2.946
R	26-30	72	3.225
R	31-35	35	3.409
R	36-40	16	3.461

Syntactic Complexity in L2 Reading: A Comparison of Adapted and Original Czech Texts

Žaneta Stiborská and Michaela Nogolová and Xinying Chen and Miroslav Kubát
University of Ostrava
Correspondence: zaneta.stiborska@osu.cz, michaela.nogolova@osu.cz

Abstract

This corpus-based study explores the syntactic complexity of adapted Czech texts designed for learners of Czech as a second language (L2). It investigates how syntactic complexity varies according to learner proficiency levels (A2, B1, B2) as defined by the Common European Framework of Reference for Languages (CEFR) and how these adapted texts differ from their original versions. Quantitative analyses using metrics such as average sentence length (ASL), average clause length (ACL), mean dependency distance (MDD), and mean hierarchical distance (MHD) demonstrate clear systematic simplifications in adapted texts at lower proficiency levels. At A2 and B1 levels, adapted texts were found to be significantly less syntactically complex compared to their original counterparts. However, these differences diminished notably at the B2 proficiency level, indicating a gradual alignment of adapted texts with native-level syntactic complexity as learner proficiency increased. These results underscore the importance of careful syntactic calibration in creating educational materials for language learners, highlighting implications for curriculum design, instructional methodologies, and materials development. The findings offer valuable insights for language educators and textbook authors aiming to optimize reading materials to support language acquisition effectively.

1 Introduction

Reading comprehension is a fundamental component of second language acquisition. Adapted texts are commonly utilized in language education to support learners toward reading authentic, native-level materials. These texts are simplified versions of original works, specifically designed to align with learners' proficiency levels and support reading development. Adaptations typically aim to reduce lexical and syntactic complexity, making the text more accessible while preserving the original storyline and stylistic value (Goodman and Freeman, 2018; Crossley et al., 2012).

While significant research has examined lexical simplification and vocabulary development in adapted texts (Bahrainian et al., 2024; Štajner et al., 2022; Truică et al., 2023; Bingel et al., 2018) less attention has been dedicated to the syntactic complexity of their texts and how it is systematically modified according to learner proficiency. For example, Jin et al. (2020) examined syntactic complexity of English foreign language teaching materials for various grade levels in China, using eight syntactic measures. Their findings demonstrated significant differences in syntactic complexity between texts adapted for different proficiency levels. Additional studies have empirically compared simplified and authentic texts, revealing substantial linguistic differences – including lexical variation, syntactic structure, and textual cohesion (e.g. Crossley et al., 2007; Goodman and Freeman, 2018; Davison and Kantor, 1982; Carrell, 1987; Allen, 2009). Tools such as Coh-Metrix have been used to show that simplified texts typically contain fewer complex connectives, simpler syntactic constructions, greater lexical repetition, and enhanced local coherence.

In the Czech research context, adapted texts are most commonly studied in relation to reading literacy. Reading literacy is widely recognized as a crucial competency, enabling learners to understand, interpret, and engage with textual information effectively. However, Czech academia in this area has largely relied on qualitative approaches (e.g. Slavík, 2003; Vondrová et al., 2022), and quantitative methodologies remain underdeveloped. Existing Czech quantitative studies have primarily focused on assessing textual complexity and readability in textbooks (e.g. Pluskal, 1996; Průcha, 1998; Greger, 1999), often analyzing limited textual excerpts. In contrast, research on foreign languages increasingly adopts computational and

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 47–55

large-scale approaches to text analysis (Rupp et al., 2001; Graesser et al., 2011; Benjamin, 2012; Rafatbakhsh and Ahmadi, 2023), highlighting the need for similarly robust methods in Czech language education research.

The present study addresses this gap by investigating the syntactic complexity of adapted Czech literary texts designed for learners of Czech as a second language (L2) across the proficiency levels A2, B1, and B2, as defined by the Common European Framework of Reference for Languages (Council of Europe, 2001, CERF). Using quantitative syntactic metrics, this research compares adapted texts to their original versions, assessing the degree of syntactic simplification employed and how it correlates with learner proficiency.

2 Language Material

The corpus under analysis consists of ten adapted literary texts designed specifically for L2 learners of Czech. These texts represent the complete set of officially published adapted prose publications in Czech for this target group. Each book is aimed at a specific proficiency level according to the CERF, ranging from A2 to B2. These adaptations facilitate reading comprehension and linguistic acquisition by systematically adjusting the syntactic and lexical complexity to match learners' language proficiency. Our study examines how these adapted texts differ in syntactic complexity and compares them with their original Czech counterparts.

All adapted texts belong to the series *Adaptovaná česká próza* "Adapted Czech prose", published by Akropolis, a publishing house specializing in materials for L2 instruction in Czech.

At the A2 proficiency level, the corpus includes three books – *Brněnské legendy* "Brno Legends" (Trchová, 2017), *O pejskovi a kočičce* "A Doggie and a Pussycat" (Čapek et al., 2019) and *Pohádky* "Fairytales" (Holá, 2013).

At the B1 proficiency level, the corpus contains five adapted texts – *Povídky malostranské* "Prague Tales" (Neruda and Holá, 2012), *Pražské legendy* "Prague Legends" (Holá, 2011), *První láska a jiné povídky* "First love and other stories" (Šabach et al., 2014), *Staré pověsti české a moravské* "Old Czech and Moravian tales" (Holá, 2012) and *Báječná léta pod psa* "Bliss Was It in Bohemia" (Viewegh and Šichová, 2021).

Finally, at the B2 proficiency level, two adapted books were analyzed – *Košík plný milenců a jiné povídky* "A Basket Full of Lovers and Other Stories" (Pawlowská et al., 2015) and *Povídky z jedné kapsy a Povídky z druhé kapsy* "Tales from Two Pockets" (Čapek and Korková, 2010).

The motivation behind the adaptation of literary texts into simplified versions for learners of Czech as a second language stems from a combination of pedagogical, cultural, and practical considerations. According to a recent analysis of author interviews (Šimková, 2019), many of the adaptors were driven by a lack of suitable reading materials for non-native speakers on the Czech market. Their primary aim was to provide accessible texts that would both support language acquisition and introduce learners to Czech literature and cultural heritage. Furthermore, several adaptors emphasized the practical utility of these texts in classroom contexts and found inspiration in similar foreign editions of simplified literature.

The process of selecting titles for adaptation was influenced by a variety of factors, including personal interest, availability of copyright, timelessness, literary value, thematic appeal, and the potential for further language work. While some adaptors favored canonical Czech works, others focused on folklore, legends, or even translations of global literary texts. Importantly, there was no consensus on the primary nature of the adapted text: most authors viewed the pedagogical and artistic dimensions as equally important, striving to preserve aesthetic quality while facilitating language learning. This duality is reflected in the content, structure, and linguistic features of the texts, suggesting a deliberate attempt to bridge literary authenticity with didactic function (Šimková, 2019).

Notably, all the adaptors were also experienced educators, and most identified primarily as teachers in the adaptation process, often combining this role with that of editor, co-author, or translator. Their decisions regarding linguistic simplification – whether in vocabulary, syntax, or discourse structure – were typically informed by their teaching practice, CEFR guidelines, and existing teaching materials. Although no strict methodology was applied, the adaptations exhibit a high degree of consistency in using exercises, audio recordings, and illustrations, all aimed at supporting comprehensive language development. These findings demonstrate a nuanced, intuitive approach to text simplification that balances linguistic accessibility with cultural and literary integrity (Šimková, 2019).

This Czech approach closely mirrors the for-

eigner approach in L2 text simplification. As discussed by Crossley et al. (2011, 2012) or Allen (2009), the majority of adapted materials for L2 learners worldwide are created through an intuitive approach, in which the adaptor's experience as a teacher, language learner, or material developer plays a central role. Instead of relying on fixed word lists or formal grammatical constraints, adaptors make subjective judgments about what language structures are appropriate for learners at specific proficiency levels. These intuitively simplified texts aim to enhance readability and comprehensibility while maintaining the narrative coherence and stylistic essence of the original work.

In contrast, structural approaches – which rely on pre-defined lexical and grammatical frameworks or readability formulas – are less commonly applied and have been criticized for failing to account for deeper cognitive and discourse – level processing (Davison and Kantor, 1982; Carrell, 1987). Even in graded reader schemes that use controlled language, the goal remains similar: to reduce cognitive load and support language development through extensive reading. Overall, both Czech and international findings suggest that successful text adaptation depends less on rigid simplification rules and more on pedagogical sensitivity, linguistic intuition, and a balanced integration of aesthetic and instructional goals.

In addition to the adapted texts, the corpus also includes parts of three original Czech literary works: *O pejskovi a kočičce* "A Doggie and a Pussycat" (Čapek, 2018a), *Povídky malostranské* "Prague Tales" (Neruda, 2011), *Povídky z jedné kapsy* "Tales from First Pocket" (Čapek, 2018c) and *Povídky z druhé kapsy* "Tales from Second Pockets" (Čapek, 2018b). These texts correspond to the adapted versions used at the A2, B1, and B2 proficiency levels, respectively. Therefore, the built corpus enables a direct comparison between adapted and original texts across all three proficiency levels, providing insight into the syntactic modifications applied during the adaptation process. Table 1 provides an overview of the number of texts, sentences, and tokens analyzed in this study.

material	text_n	sentence_n	word_n
A2	32	1041	8887
B1	46	2215	21307
B2	19	1070	12761
Adapted book A2	10	464	3627
Adapted book B1	6	354	3015
Adapted book B2	9	670	7513
Original book A2	10	676	9868
Original book B1	6	685	8664
Original book B2	9	1085	14032

Table 1: Number of texts, sentences and tokens.

3 Methodology

Each text (individual chapter) was analyzed using UDPipe 2.0 (Straka, 2018) with Universal Dependencies (UD) 2.15 models (Zeman et al., 2019), a well-established framework for syntactic parsing. Subsequently, the dependency trees were converted into the Surface Syntactic Universal Dependencies (SUD) scheme (Gerdes et al., 2018), which more emphasizes distributional criteria and syntactic relations rather than the content-focused approach of UD.

To maintain data consistency, only sentences that met the following criteria were included in the analysis: (i) they contained a predicate (a finite verb or auxiliary) as the sentence root and (ii) they did not include abbreviations, numerical digits, or special characters. In this paper, four syntactic indices were used to examine syntactic complexity: average sentence length (ASL), average clause length (ACL), mean dependency distance (MDD), and mean hierarchical distance (MHD).

ASL was measured using two methods: (i) the ratio of total words to sentences (words per sentence) and (ii) the ratio of total clauses to sentences (clauses per sentence). The first metric represents overall sentence length, while the second reflects clause density within sentences.

ACL was determined as the ratio of total words to total clauses, serving as an indicator of clause complexity.

MDD, based on Liu (2008), measures syntactic complexity by calculating the average dependency distance (DD) between syntactically related words throughout the text, excluding punctuation and root nodes. The DD of a word corresponds to the absolute difference between its position in the sentence (captured by id of each word) and the position of its syntactic parent. MDD was computed by dividing the sum of all DDs in the text by the total number of dependent words (i.e., total words minus the number of sentences), as shown in Formula 1:

$$\text{MDD} = \frac{\sum_{i=1}^{n-s} |DD_i|}{n-s} \quad (1)$$

where n represents the total token count, s denotes the number of included sentences, and DDi is the dependency distance of the i-th token.

MHD, introduced by Jing and Liu (2015), was calculated using the same approach as MDD, but instead of dependency distances, hierarchical distances (HDs) were used. The HD of a word indicates the number of dependency edges separating it from the sentence root. MHD offers a broader structural perspective, illustrating the extent to which syntactic elements are embedded within a sentence. Let us illustrate the computation process of the four syntactic complexity indexes using sentence 1 from the adapted version of *O pejskovi a kočičce*.

1. Mám v kuchni myši a nechci je tam mít.
   "I have mice in the kitchen, and I do not want to have them there."

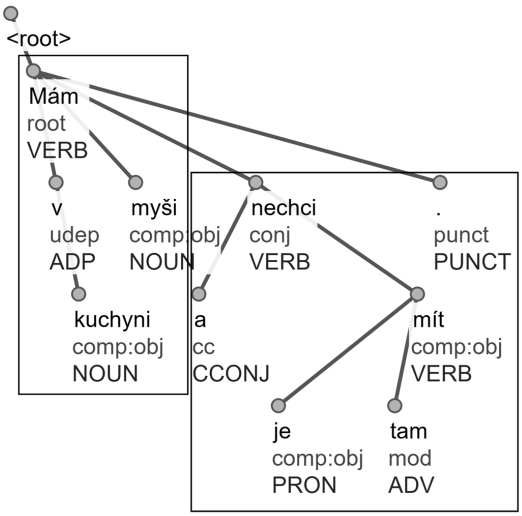


Figure 1: The dependency tree of sentence 1.

Figure 1 presents the dependency tree of sentence 1, rectangles capture individual clauses. The sentence consists of 9 words (excluding punctuation) and contains 2 clauses, as it includes 2 predicates: mám "have" and nechci "do not want to". The ASL in terms of words is 9, calculated as 9 words divided by 1 sentence ($9/1 = 9$). The ASL in terms of clauses is 2, 2 clauses divided by 1 sentence ($2/1 = 2$). The ACL is 4.5, calculated as 9 words divided by 2 clauses ($9/2 = 4.5$). The MDD is 2.125, based on the sum of all dependency distances in the sentence: $\frac{1+1+3+1+5+2+1+3}{8} = \frac{17}{8} = 2.125$ The MHD is 1.875, calculated from the sum of all hierarchical distances: $\frac{1+2+1+1+2+2+3+3}{8} = \frac{15}{8} = 1.875$. To evaluate statistical significance, comparisons were made between the following groups:

1. Adapted texts across proficiency levels A2, B1 and B2.

2. Adapted texts in contrast with their original versions.

Prior to statistical testing, the normality of each dataset was assessed using the Shapiro-Wilk test (Shapiro and Wilk, 1965). If normality was violated in any group, the Mann-Whitney U test (Mann and Whitney, 1947) was applied as a non-parametric alternative. In cases where both groups followed a normal distribution, an independent samples t-test was performed.

4 Results

4.1 Adapated texts

The results reveal that adapted texts across proficiency levels (A2, B1 and B2) demonstrated a clear trend of increasing syntactic complexity aligned with higher proficiency. Sentence length, measured in words, shows a clear upward trend from A2 to B2 levels (see Figure 2 and Table 2), with statistically significant differences between all levels ($p < 0.05$).

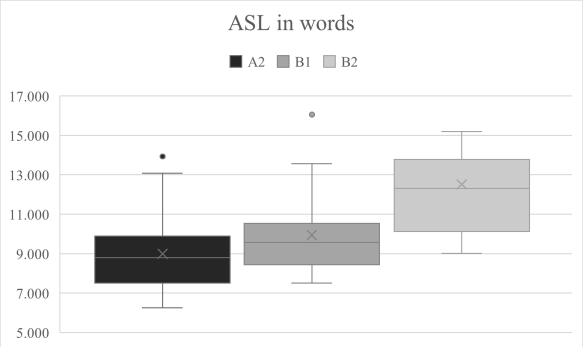


Figure 2: The average sentence length in words.

When measuring ASL in terms of clauses, we observe an overall increasing trend from level A2 to B2. However, the use of clauses within a sentence remains relatively similar between levels A2 and B1 (see Figure 3 and Table 2).

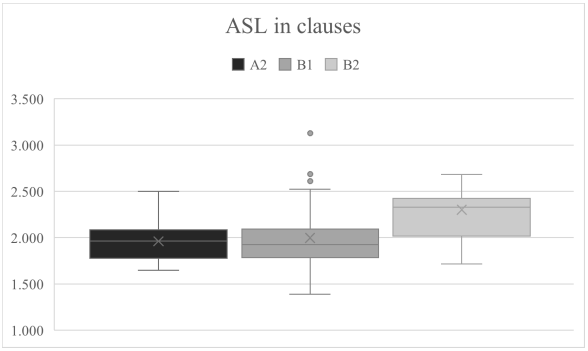


Figure 3: The average sentence length in clauses.

Statistical analysis confirms significant differences only between A2 and B2 ($p < 0.001$), and between B1 and B2 ($p = 0.001$).

level	ASL in words	sd	ASL in clauses	sd
A2	8.991	1.794	1.961	0.202
B1	9.946	2.160	1.996	0.334
B2	12.512	2.882	2.302	0.416

Table 2: Means and standard deviations (sd) of ASL in words, ASL in clauses

The average clause length shows a clear upward trend from A2 to B2 (see Figure 4 and Table 3), with statistically significant differences across all proficiency levels ($p < 0.05$). This increase mirrors the pattern observed in sentence length measured in words, indicating that syntactic complexity grows not only through the expansion of sentence structure but also through the internal development of individual clauses.

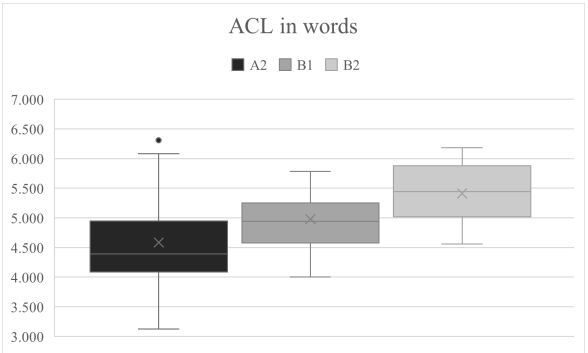


Figure 4: The average clause length.

The alignment between sentence and clause length trends suggests that as learners progress, they are gradually exposed to more elaborated syntactic constructions, both at the interclausal and intraclusal levels. This supports the notion of controlled complexity progression in adapted texts, designed to match learners' growing proficiency and prepare them for authentic language use.

level	ACL	sd
A2	4.583	0.750
B1	4.978	0.530
B2	5.412	0.477

Table 3: Means and standard deviations (sd) of ACL

Both MDD and MHD display a clear upward trajectory from A2 to B2, with consistent and statistically supported increases ($p < 0.05$) across the three proficiency levels (see Figures 5 and 6 and Table 4).

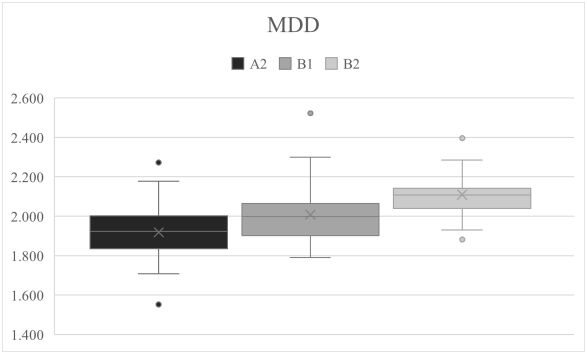


Figure 5: Mean dependency distance.

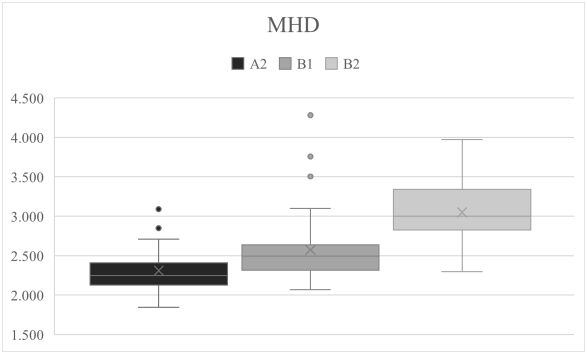


Figure 6: Mean hierarchical distance.

These findings indicate that B2 texts involve greater syntactic depth and longer dependencies, signaling a more complex and layered sentence structure. The rising values of these two measures suggest a shift toward deeper hierarchical embedding and longer syntactic spans, which corresponds to preparing advanced learners for exposure to more complex, native-level language.

level	MDD	sd	MHD	sd
A2	1.919	0.143	2.309	0.301
B1	2.008	0.147	2.573	0.426
B2	2.109	0.120	3.049	0.459

Table 4: Means and standard deviations (sd) of MDD and MHD.

Overall, the results indicate that syntactic complexity in adapted Czech texts increases progressively across proficiency levels, with B2 texts exhibiting the most sophisticated structures. This supports the assumption that higher-level adapted texts are designed to approximate native-level syntax more closely, regarding dependency distance and hierarchical depth. The findings confirm a systematic calibration of syntactic features according to learner proficiency, aligning with pedagogical goals of gradually preparing learners for authentic reading experiences.

4.2 Comparison of adapted and original texts

Regarding ASL measured in words (see Figure 7 and Table 5), statistically significant differences ($p < 0.05$) were found between the adapted and original versions at A2 and B1 levels, while no significant difference was observed at B2 ($p = 0.103$).

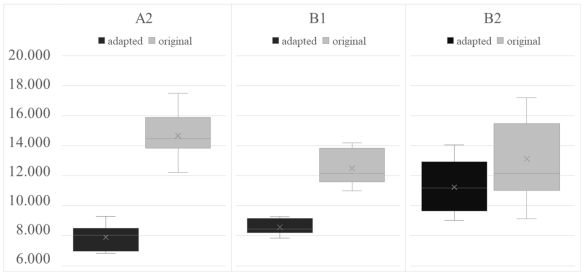


Figure 7: ASL in words of original and adapted texts.

	adapted		original	
level	mean	sd	mean	sd
A2	7.907	0.849	14.669	1.568
B1	8.571	0.531	12.494	1.220
B2	11.244	1.809	13.118	2.705

Table 5: Means and standard deviations (sd) of ASL in words for adapted and original texts.

In contrast, ASL measured in clauses (see Figure 8 and Table 6) showed statistically significant differences across all three pairs ($p \leq 0.05$).

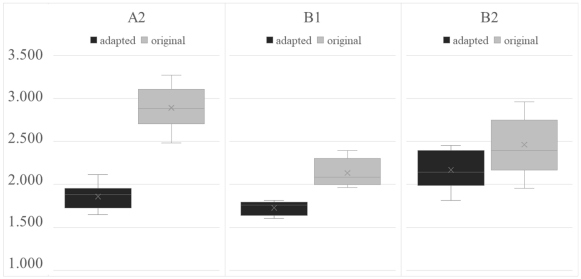


Figure 8: ASL in clauses of original and adapted texts.

	adapted		original	
level	mean	sd	mean	sd
A2	1.858	0.145	2.890	0.256
B1	1.730	0.082	2.133	0.165
B2	2.168	0.231	2.462	0.344

Table 6: Means and standard deviations (sd) of ASL in clauses for adapted and original texts.

Although the adapted B2 text is similar in sentence length (measured in words), it still displays a notably simpler syntactic structure compared to the original. The consistent differences in clause-based measures suggest that adaptations maintain reduced syntactic complexity, even when overall sentence length appears comparable.

In line with the findings on sentence length, further syntactic measures confirmed significant differences between the adapted and original versions of the A2 and B1 texts, whereas the B2 pair showed no statistically significant divergence.

ACL was significantly lower in the A2 and B1 ($p < 0.05$) adapted versions indicating a tendency toward structurally simpler clauses at lower proficiency levels. However, no significant difference ($p = 0.597$) was observed in text adapted for B2 level, suggesting that the clause structure in this adaptation remains relatively close to the original (see Figure 9 and Table 7).

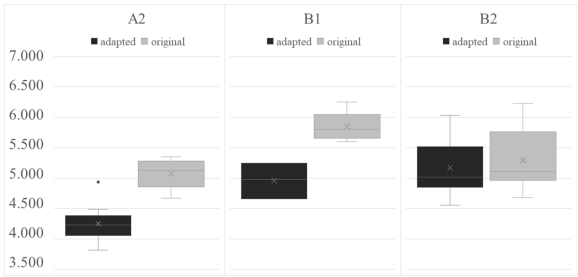


Figure 9: ACL of original and adapted texts.

	adapted		original	
level	**mean**	**sd**	**mean**	**sd**
A2	4.254	0.305	5.072	0.240
B1	4.960	0.300	5.852	0.240
B2	5.171	0.456	5.294	0.511

Table 7: Means and standard deviations (sd) of ACL for adapted and original texts.

The analysis of structural dependency measures revealed a consistent pattern across both MDD and MHD (see Table 8 and Figure 10 and Table 9 and Figure 11). Significant differences were identified in the A2 and B1 text pairs, reflecting reduced linear distance and syntactic embedding in the adapted versions. At the B2 level, however, these differences were no longer statistically significant, suggesting convergence between adapted and original texts regarding syntactic complexity.

	adapted		original	
level	**mean**	**sd**	**mean**	**sd**
A2	1.870	0.083	2.364	0.088
B1	1.962	0.061	2.225	0.096
B2	2.065	0.086	2.169	0.131

Table 8: Means and standard deviations (sd) of MDD for adapted and original texts.

	adapted		original	
level	**mean**	**sd**	**mean**	**sd**
A2	2.168	0.172	3.071	0.219
B1	2.386	0.194	2.910	0.130
B2	2.721	0.332	2.951	0.416

Table 9: Means and standard deviations (sd) of MHD for adapted and original texts.

Taken together, these results highlight a consistent pattern – syntactic simplification in adapted texts is most pronounced at the lower proficiency levels, both in terms of clause structure and dependency complexity. At the B2 level, the adapted texts retain much of the syntactic sophistication of the original works. This suggests that while simplification is a key strategy in materials for beginning and intermediate learners, advanced-level adaptations aim to approximate native-level structures more closely, supporting learners' transition to authentic reading.

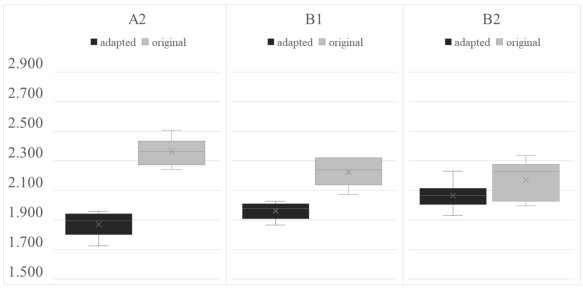


Figure 10: MDD of original and adapted texts.

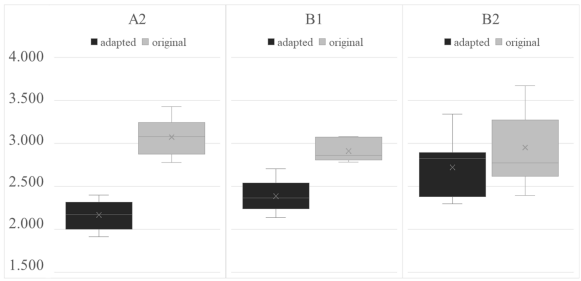


Figure 11: MHD of original and adapted texts.

5 Conclusion

This study investigated the syntactic complexity of adapted Czech literary texts across CEFR proficiency levels (A2–B2) and compared a subset of these texts with their original, non-adapted versions. Using a set of quantitative syntactic measures – including average sentence and clause length, mean dependency distance (MDD) and mean hierarchical distance (MHD) – we observed a clear proficiency-aligned increase in syntactic complexity within adapted texts. Texts intended for B2 learners exhibited significantly higher levels of structural complexity than those aimed at A2 and B1, suggesting deliberate calibration in the design of adapted materials.

The comparative analysis between adapted and original texts further revealed that syntactic simplification is most pronounced at lower proficiency levels. While the A2 and B1 adaptations showed statistically significant reductions in syntactic complexity across all core measures, the B2 adaptation did not differ significantly from its original counterpart. This suggests that adapted texts at more advanced levels tend to retain authentic syntactic structures, thus providing learners with exposure to language that approximates native-level usage.

Together, these findings highlight the role of syntactic adaptation in facilitating reading comprehension and language acquisition, particularly

at the earlier stages of L2 development. At the same time, they confirm that advanced learners are increasingly challenged with structurally complex input – a necessary step in the transition toward full linguistic competence. Future research could extend these findings by incorporating lexical, semantic or discourse-level features and by examining a broader range of genres and adaptation practices.

Beyond highlighting the increasing syntactic complexity across proficiency levels, our findings may serve as a foundation for developing a more systematic methodology for syntactic adaptation of texts. Such a framework could support authors and educators in producing level-appropriate reading materials, especially for L2 learners, by offering evidence-based guidelines for adjusting sentence length, clause structure, and syntactic depth. While current adaptations are largely guided by intuition or pedagogical experience, our data suggest the potential for a more standardized approach to aligning textual complexity with CEFR proficiency levels. This could lead to more effective and transparent practices in textbook design and literary adaptation for language education.

Acknowledgments

The research is supported by Grant SGS04/FF/2025, University of Ostrava.

References

David Allen. 2009. A study of the role of relative clauses in the simplification of news texts for learners of english. *System*, 37(4):585–599.

Seyed Ali Bahrainian, Jonathan Dou, and Carsten Eickhoff. 2024. Text simplification via adaptive teaching. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6574–6584, Bangkok, Thailand. Association for Computational Linguistics.

Rebekah George Benjamin. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24:63–88.

Joachim Bingel, Gustavo Paetzold, and Anders Søgaard. 2018. Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 245–258.

Patricia L Carrell. 1987. Readability in esl. *Reading in a Foreign Language*, 4(1):21–40.

Council for Cultural Co-operation Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Scott A Crossley, David Allen, and Danielle S McNamara. 2012. Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research*, 16(1):89–108.

Scott A Crossley, David B Allen, and Danielle S McNamara. 2011. Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a foreign language*, 23(1):84–101.

Scott A Crossley, Max M Louwerse, Philip M McCarthy, and Danielle S McNamara. 2007. A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, 91(1):15–30.

Alice Davison and Robert N Kantor. 1982. On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading research quarterly*, pages 187–209.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. Sud or surface-syntactic universal dependencies: An annotation scheme near-isomorphic to ud. In *Universal dependencies workshop 2018*.

Kenneth S Goodman and David Freeman. 2018. What's simple in simplified language? In *Simplification: Theory and application*, pages 69–76. Singapore: SEAMEO Regional Language Center.

Arthur C Graesser, Danielle S McNamara, and Jonna M Kulikowich. 2011. Coh-metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, 40(5):223–234.

David Greger. 1999. Obtížnost textů učebnic českého jazyka pro 2. ročník zŠ. *Pedagogická orientace*, 9(2):96–99.

Lída Holá. 2011. *Pražské legendy*. Akropolis, Praha.

Lída Holá. 2012. *Staré pověsti české a moravské*. Akropolis, Praha.

Lída Holá. 2013. *Pohádky*. Akropolis, Praha.

Tan Jin, Xiaofei Lu, and Jing Ni. 2020. Syntactic complexity in adapted teaching materials: Differences among grade levels and implications for benchmarking. *The Modern Language Journal*, 104(1):192–208.

Yingqi Jing and Haitao Liu. 2015. Mean hierarchical distance augmenting mean dependency distance. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 161–170, Uppsala, Sweden. Uppsala University, Uppsala, Sweden.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Jan Neruda. 2011. *Povídky Malostranské*. Městská knihovna v Praze, Praha.

Jan Neruda and Lída Holá. 2012. *Povídky malostranské*. Akropolis, Praha.

Halina Pawlowská, Silvie Převrátilová, and Petra Bulejčíková. 2015. *Košík plný milenců a jiné povídky*. Akropolis, Praha.

Miroslav Pluskal. 1996. Zdokonalení metody pro měření obtížnosti didaktických textů. *Pedagogika*, 46(1):62–76.

Jan Průcha. 1998. *Učebnice: Teorie a analýzy edukačního média*. Paido.

Elaheh Rafatbakhsh and Alireza Ahmadi. 2023. Predicting the difficulty of efl reading comprehension tests based on linguistic indices. *Asian-Pacific Journal of Second and Foreign Language Education*, 8(1):41.

Andre A Rupp, Paula Garcia, and Joan Jamieson. 2001. Combining multiple regression and cart to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing*, 1(3-4):185–216.

Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.

J Slavík. 2003. Lesk a bída oborových didaktik. pedagogika. *Pedagogika*, 53(2):137–140.

Sanja Štajner, Daniel Ferrés, Matthew Shardlow, Kai North, Marcos Zampieri, and Horacio Saggion. 2022. Lexical simplification benchmarks for english, portuguese, and spanish. *Frontiers in artificial intelligence*, 5:991242.

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.

Martina Trchová. 2017. *Brněnské legendy*. Akropolis, Praha.

Ciprian-Octavian Truică, Andrei-Ionuţ Stan, and Elena-Simona Apostol. 2023. Simplex: a lexical text simplification architecture. *Neural Computing and Applications*, 35(8):6265–6280.

Michal Viewegh and Kateřina Šichová. 2021. *Báječná léta pod psa*. Akropolis, Praha.

Naďa Vondrová, Martina Šmejkalová, and Irena Smetáčková. 2022. Zadání slovních úloh jako podklad pro rozvoj čtení s porozuměním a dovednosti slovní úlohy řešit. *Pedagogika*, 72(1).

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, and 326 others. 2019. Universal dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics ÚFAL, Faculty of Mathematics and Physics, Charles University.

Josef Čapek. 2018a. *Povídaní o pejskovi a kočičce*. Městská knihovna v Praze, Praha.

Josef Čapek, Silvie Převrátilová, and Petra Bulejčíková. 2019. *O pejskovi a kočičce*. Akropolis, Praha.

Karel Čapek. 2018b. *Povídky z druhé kapsy*. Městská knihovna v Praze, Praha.

Karel Čapek. 2018c. *Povídky z jedné kapsy*. Městská knihovna v Praze, Praha.

Karel Čapek and Radomila Korková. 2010. *Povídky z jedné a Povídky z druhé kapsy*. ASA, Praha.

Petr Šabach, Silvie Převrátilová, and Petra Bulejčíková. 2014. *První láska a jiné povídky*. Akropolis, Praha.

Věra Šimková. 2019. Specifika literární komunikace u jedinců s odlišným mateřským jazykem v českém prostředí. Master's thesis, Masaryk University.

Modeling the Law of Abbreviation in Classical, Modern, and ChatGPT-Generated Chinese: A Power-Law Analysis of Structural Economy

Jianwei Yan[1], **Heng Chen**[2]

[1] Department of Linguistics, Zhejiang University;
[2] Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies.
jwyan@zju.edu.cn; chenheng@gdufs.edu.cn

Abstract

This study investigates the Law of Abbreviation—the inverse relationship between word length and frequency—across Classical, Modern, and ChatGPT-generated Chinese. Using a tri-partite parallel corpus and a power-law model $y = a * x^{-b}$, we analyze the relationship between word length and the average usage frequency of words within a given word length category to assess structural economy. Results confirm consistent Zipfian distribution across all text types, with high R^2 values indicating strong model fit. However, the parameter b varies significantly: Classical Chinese shows the steepest decline, suggesting strong pressure for brevity; Modern Chinese exhibits a moderated pattern; ChatGPT-generated texts display the weakest pressure, prioritizing fluency over compression. These differences reflect evolving communicative priorities and reveal that while AI models can mimic statistical distributions, they underrepresent deeper structural pressures found in natural language evolution. This study offers insights into lexical optimization and the parameter b offers a useful metric for comparing structural efficiency across modalities. Implications are discussed in relation to language modeling, cognitive economy, and the evolution of linguistic structure.

1 Introduction

One of the most enduring empirical patterns in quantitative linguistics is the inverse relationship between word length and word frequency, commonly known as the Law of Abbreviation. Zipf (1935: 25) famously hypothesized that "the magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences," and also asserted that "the larger a word is in length, the less likely it is to be used" (Zipf, 1935: 22). This observation has given rise to a long-standing discussion regarding the appropriate directionality of modeling: should word length be treated as a function of frequency, or should frequency be modeled as a function of word length (Strauss et al., 2007: 277)?

Researchers who model word length as a function of frequency often draw on Zipf's (1949) "principle of least effort," which suggests that frequently used linguistic forms tend to be shorter for communicative efficiency (e.g., Köhler, 1986). For instance, Breiter (1994) found that higher-frequency words tend to be shorter based on frequency dictionaries. Wang (2014) used corpus data and confirmed a negative correlation between frequency and length in Chinese, consistent with a power-law distribution. Moreover, Bentz and Ferrer-i-Cancho (2016), in a large-scale cross-linguistic study covering 1,262 texts in 986 languages, found robust negative correlations between frequency and word length, attributing the universal pattern to fundamental principles of information processing: "Words that are used more frequently tend to be shorter" (p. 1).

In contrast, other scholars have argued for the reverse modeling direction, treating word frequency as a function of word length. This approach aligns with Zipf's (1935: 22) assertion that longer words are inherently less frequent,

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 56–62

reflecting structural constraints on usage. It has been adopted by Miller et al. (1958), Chen et al. (2015), Linders and Louwerse (2023), and Li and Lei (2025), etc. For example, Linders and Louwerse (2023) demonstrated that the Law of Abbreviation holds in natural spoken dialogues, extending prior findings based on written corpora. Li and Lei (2025) validated the law in Chinese texts across four genres, revealing that while the inverse relationship persists, genre-specific factors and character polysemy may modulate the strength of the effect.

While both approaches may be mathematically equivalent under parameter transformation (Strauss et al., 2007: 279-280), they embody distinct theoretical assumptions, and entail different perspectives on how form and usage interact. Modeling word length as a function of frequency suggests that linguistic structure is shaped by language use—frequent forms tend to become shorter over time. Conversely, modeling frequency as a function of word length assumes that structural features of language constrain how often a form is used, with shorter or simpler forms being more cognitively efficient and therefore more likely to recur.

Notably, most prior studies have focused on languages using alphabetic writing systems, particularly those using Latin scripts such as English, Dutch, or German. Research on Chinese remains limited, with only a handful of studies examining the length–frequency relationship (e.g., Breiter, 1994; Wang, 2014; Chen et al., 2015; Li and Lei, 2025). Moreover, few studies utilize parallel corpora or consider how emerging forms of language generation—such as AI-generated text—may reflect or deviate from natural linguistic patterns. This study addresses both gaps by focusing on Chinese and incorporating AI-generated texts as a comparison.

In this study, we adopt the modeling perspective that treats frequency as a function of word length, based on three key considerations. First, this direction emphasizes the structural constraints that word form imposes on usage, aligning with the linguistic principle that shorter forms are more cognitively efficient and thus more likely to recur. Second, in diachronic and cross-system comparisons, word length is more stable than frequency, making it a more reliable independent variable; and treating frequency as a response variable enables us to assess whether different language production systems—including large language models—adhere to the same efficiency principles observed in human language. Finally, this approach is empirically grounded in Chen et al. (2015), who modeled the length–frequency relationship in Chinese using the power-law function and demonstrated that the parameter b captures the rate at which average usage frequency decreases with increasing word length, reflecting the evolutionary dynamics of the Chinese lexicon: A larger value of b indicates a steeper decline in frequency as length increases, signaling a stronger pressure for efficiency and simplification.

Specifically, we apply the power-law function to the corpus comprises three parallel versions: (1) Classical Chinese texts, (2) their Modern Chinese equivalents, and (3) their Modern Chinese translations generated by ChatGPT from the same Classical Chinese input. This study aims to address three research questions:

Question 1: Does the inverse relationship between word length and word frequency hold consistently across Classical Chinese, Modern Chinese, and ChatGPT-generated Chinese?

Question 2: How do the fitted power-law parameters, namely the parameter b, differ across these text types, and what do they reveal about structural pressures toward lexical economy?

Question 3: To what extent does ChatGPT-generated language replicate or diverge from the natural patterns of lexical economy observed in natural languages?

This diachronic and cross-modal design allows us to examine how structural features such as word length influence language usage across ancient, modern, and AI-generated language.

2 Material and Method

2.1 Material

The data for this study were drawn from the Classical-Modern Chinese parallel corpus,[1] which provides sentence-aligned pairs of Classical Chinese texts and their Modern Chinese

[1] https://github.com/NiuTrans/Classical-Modern

equivalents. From this corpus, we randomly extracted ten Classical Chinese texts, each with 200 sentences, and their sentence-aligned Modern Chinese equivalents, yielding ten parallel pairs and a total of 2,000 aligned sentence pairs.

To generate the AI-translated dataset, we prompted ChatGPT-4o using batches of 100 Classical Chinese sentences with the following instruction (original in Chinese, with an academic English translation provided below):

Chinese prompt: "以下是 100 个文言文句子，请将这些句子翻译为流畅、自然的现代汉语。翻译时不必拘泥于文言原文的句式结构，重点在于准确传达原意，使现代读者易于理解。请仅输出现代汉语译文，保持语义准确、语言通顺，避免逐字直译。"

English translation: "The following are 100 Classical Chinese sentences. Please translate them into fluent and idiomatic Modern Chinese. You are not required to adhere strictly to the syntactic structures of the source text; instead, prioritize conveying the intended meaning clearly and naturally for a contemporary readership. Provide only the translated Modern Chinese sentences. Ensure semantic fidelity and linguistic fluency, and avoid literal, word-for-word translation."

All ChatGPT translations were generated in separate sessions, each using the same prompt. This procedure yielded ChatGPT-generated translations for each Classical Chinese sentence, resulting in aligned triplets for every source sentence: (1) the Classical Chinese, (2) the Modern Chinese version, and (3) the ChatGPT-generated Modern Chinese version. Each text type comprised 10 files, with 200 sentences per file.

Then, text segmentation was conducted using language-specific tools. The Classical Chinese texts were segmented with *udkanbun* (Yasuoka, 2019),[2] a syntactic parser based on Universal Dependencies and specifically designed for Classical Chinese (漢文/文言文). For both human and ChatGPT-generated Modern Chinese texts, segmentation was performed with *stanza* 1.10.1 (Qi et al., 2020),[3] a Python-based NLP toolkit supporting multiple languages including Chinese. The segmentations were manually checked to ensure accuracy. An overview of token and type counts for the three versions across the ten files is presented in Table 1.

Table 1: Word counts of the 30 text files.

Classical		Modern		ChatGPT	
token	**type**	**token**	**type**	**token**	**type**
3,234	1,163	3,468	1,929	2,860	1,681
3,181	1,160	3,325	1,899	2,784	1,724
3,172	1,184	3,356	1,852	2,797	1,673
3,202	1,185	3,466	1,940	2,953	1,726
3,062	1,112	3,173	1,788	2,749	1,647
3,137	1,146	3,417	1,860	2,919	1,763
3,111	1,170	3,265	1,853	2,913	1,708
3,031	1,136	3,265	1,844	2,665	1,653
3,331	1,231	3,535	1,993	2,913	1,721
3,309	1,222	3,774	2,061	2,899	1,739

2.2 Method

There are various approaches to evaluating the relationship between word length and word frequency, including non-parametric methods, linear mixed-effects regression models, and power-law formulations. For example, Bentz and Ferrer-i-Cancho (2016) employed a non-parametric approach using Kendall's τ, avoiding any specific functional form. Li and Lei (2025) predicted log-transformed frequency from character length using linear mixed-effects models. In addition, numerous equations describing the relationship between word length and frequency (or frequency rank) have been theoretically developed and employed in empirical studies (Ferrer-i-Cancho, 2025).

Informed by prior studies and for consistency with diachronic studies on Chinese, the present study adopts the modeling approach of Chen et al. (2015). Specifically, we fit the data to a power-law function of the form: $y = a * x^{-b}$, where x is word length (in characters), y is the mean word ratio (MWR), calculated as the token count divided by the type count for each word length class, indicating the average usage frequency of words within a given word length. Parameters a and b are estimated from the data.

[2] https://github.com/KoichiYasuoka/UD-Kanbun

[3] https://github.com/stanfordnlp/stanza

For each segmented text file, we recorded: (1) Tokens—the total number of word occurrences of a given length; (2) Types—the number of unique words of that length; and (3) MWR—the ratio of tokens to types. For example, in the segmented Classical Chinese sentence "余 / 啖 / 林檎 / 一 / 枚 / ， / 梨 / 二 / 枚 / ， / 山胡桃 / 五 / 枚 (I ate one apple, two pears, and five hickory nuts), words of one character appear 9 times (tokens, 余, 啖, 一, 枚, 梨, 二, 枚, 五, 枚) across 7 unique items (types, 余, 啖, 一, 枚, 梨, 二, 五), yielding an MWR of 1.29. Two-character (林檎) and three-character (山胡桃) words each occur once, with 1 token and 1 type, resulting in an MWR of 1.00.

We then fit a power-law model to the data, and computed the coefficient of determination (R^2) to assess the goodness of fit. The parameter *b* serves as an index of structural economy in the lexicon and is used to trace diachronic trends and production modality effects on the length–frequency relationship.

3 Results and Discussion

3.1 Regularity of the Inverse Length–Frequency Relationship

To empirically test the universality of the inverse relationship between word length and frequency in Chinese, we fitted power-law functions to all 30 texts. Table 2 reports the goodness of fit, and Appendix A presents the fitted curves and observed data for each text. [4]

Traditionally, an R^2 value greater than 0.9 (Mačutek and Wimmer, 2013: 233) or 0.8 (Eom, 2006: 121) is considered satisfactory. As shown in Table 2 and Appendix A, all three text types demonstrate excellent model fit, with most texts achieving R^2 exceeding 0.9, indicating that the power-law relationship holds robustly. The results consistently support the hypothesis of structural economy: as word length increases, average frequency usage sharply declines. For Classical Chinese, the R^2 values range from 0.8532 to 0.9921 (M = 0.9655, SD = 0.0426). For Modern Chinese, R^2 values range from 0.8288 to 0.9533 (M = 0.9195, SD = 0.0373). For ChatGPT-generated Chinese, the results are similarly robust, with R^2 values between 0.8355 and 0.9550 (M = 0.9210, SD = 0.0342).

These findings empirically confirm that Zipf's Law of Abbreviation is consistently observed across all three modalities, demonstrating the robustness of the inverse length–frequency relationship in Chinese, providing evidence that this relationship reflects a pervasive regularity of lexical systems. These results echo the cross-linguistic patterns reported by Bentz and Ferrer-i-Cancho (2016), and further suggest that even large language models like ChatGPT reproduce this statistical regularity—possibly as a byproduct of optimizing communicative efficiency.

Table 2: Power-law modeling results of word length-frequency distributions in 30 text files.

Classical		Modern		ChatGPT	
R^2	*b*	R^2	*b*	R^2	*b*
0.9779	1.1008	0.8288	0.7749	0.8355	0.7173
0.9358	0.8996	0.9194	0.9122	0.9411	0.9566
0.9829	1.0290	0.9524	1.0296	0.9164	0.8451
0.9873	1.0235	0.9237	0.9085	0.9207	0.8930
0.9921	1.0254	0.9243	0.9059	0.9550	0.9437
0.9808	0.9130	0.9179	0.9406	0.9165	0.8159
0.9716	1.0593	0.8911	0.8288	0.9509	1.0032
0.9863	1.0102	0.9533	1.0560	0.9105	0.7463
0.9872	1.0174	0.9524	1.0959	0.9453	0.9716
0.8532	0.8205	0.9318	0.9948	0.9178	0.8231

3.2 Variation in the Power-Law Parameter *b* and Lexical Economy

As demonstrated by Chen et al. (2015), the parameter *b* in the power-law model reflects the rate at which average word frequency decreases with increasing word length, thus serving as a quantitative index of lexical economy. A higher *b*-value indicates a steeper decline, suggesting a stronger systemic preference for brevity through more frequent use of shorter words.

As shown in Table 2 and Appendix A, the average *b*-values across the three text types reveal meaningful distinctions in lexical economy. Classical Chinese exhibits the highest *b*-value,

[4] Although the consistently high R^2 values and visualizations confirm the Zipfian patterns across the three text types, a modeling-related limitation remains—namely, the small number of word length categories available for power-law fitting, particularly in some Classical Chinese texts. Future research may adopt alternative models and complementary approaches to triangulate the results and enhance generalizability.

demonstrating the steepest inverse relationship (M = 0.9899, SD = 0.0850), with b-values ranging from 0.8205 to 1.1008. This indicates the strongest structural pressure for brevity, consistent with the compactness and density characteristic of traditional literary forms. Modern Chinese shows a slightly lower mean (M = 0.9447, SD = 0.1006; range: 0.7749 to 1.0959), suggesting a somewhat weaker pressure for brevity and greater tolerance for longer word forms. In contrast, ChatGPT-generated Chinese presents the smallest b-value (M = 0.8716, SD = 0.0977), with a range from 0.7173 to 1.0032. This comparatively flatter decline implies that the model prioritizes fluency and plausibility over compression, consistent with its generative objective to optimize for readability rather than structural economy.

These findings are supported by lexical diversity patterns (see Table 1). Classical Chinese has the highest token/type ratio (M = 2.7136, SD = 0.0394), suggesting low lexical diversity and high repetition. Modern Chinese follows with a moderate ratio (M = 1.7897, SD = 0.0292), while ChatGPT-generated texts show the lowest ratio (M = 1.6701, SD = 0.0350), indicating greater lexical variety and less repetition. These patterns align with the b-value trends: the steeper slope in Classical Chinese reflects tighter lexical economy, whereas the flatter slope in ChatGPT texts suggests weaker brevity constraints.

To statistically assess these differences, a Kruskal–Wallis H test was conducted on the b-values, revealing a significant difference among the three text types, H (2) = 6.68, p = 0.036. Follow-up Mann–Whitney U tests showed a significant difference between Classical Chinese and ChatGPT-generated texts (p = 0.009, *Cliff's delta* = 0.70), while comparisons between Classical and Modern (p = 0.385, *delta* = 0.24) and between Modern and ChatGPT (p = 0.162, *delta* = 0.38) were not statistically significant, though they yielded small-to-moderate effect sizes. Figure 1 visualizes these distributional differences.

Overall, the results reveal a progressive attenuation in the pressure for lexical economy across both historical and generative dimensions. While all three modalities conform to Zipfian scaling, the magnitude of b highlights systematic variation in structural optimization: Classical Chinese reflects strong efficiency-driven constraints, Modern Chinese represents a moderated form of such pressure, and ChatGPT-generated language prioritizes fluency, coherence, and accessibility. This diachronic and modality-based divergence likely mirrors evolving communicative priorities. Notably, while ChatGPT reproduces surface-level statistical regularities, it may not fully internalize the deeper structural pressures that govern naturally evolved human language.

The parameter b, therefore, serves not only as an indicator of Zipfian adherence but also as a sensitive metric for comparing structural economy across modalities.

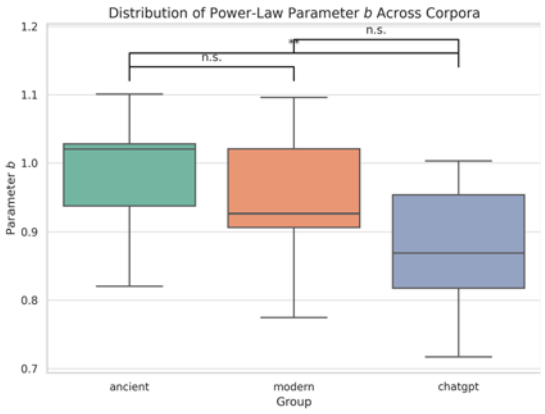


Figure 1: Distribution of power-law b values across text types.

3.3 Structural Divergence in Lexical Economy between AI-Generated and Human Texts

Although ChatGPT outputs exhibit high R^2 values—indicating conformity to Zipfian distributions—they consistently yield lower b-values than both Classical and Modern Chinese. This suggests that while the model captures the general pattern of frequency decline with word length, the intensity of this relationship is weaker.

This divergence implies that ChatGPT approximates, but may do not fully internalize, the structural constraints underlying natural lexical distributions. Human-authored texts—especially Classical Chinese—reflect strong pressures for brevity and efficiency. In contrast, ChatGPT-generated language appears driven more by statistical plausibility than by structural economy, associating shorter words with higher frequency without being governed by communicative constraints.

The contrast with Classical Chinese is particularly notable. Its historical evolution favored compression and information density—qualities not explicitly optimized in neural models. Instead, ChatGPT is trained to maximize coherence and fluency based on probabilistic exposure, often resulting in less disciplined lexical structures despite surface Zipfian regularity.

Taken together, these findings affirm the generalizability of Zipf's Law of Abbreviation across modalities, while also revealing graded differences in lexical economy.

4 Conclusion

This study examined whether the inverse relationship between word length and frequency—commonly known as the Law of Abbreviation—holds across three language modalities: Classical Chinese, Modern Chinese, and ChatGPT-generated Chinese. Using a power-law model, we analyzed the relationship between word length and average usage frequency measured by MWR across three text types. All texts showed strong Zipfian patterns, with high R^2 values indicating good model fit.

However, significant differences emerged in the fitted *b* parameter, which reflects lexical economy. Classical Chinese exhibited the largest *b*-values, indicating the strongest preference for short, frequent words. Modern Chinese showed moderate brevity pressure, while ChatGPT-generated texts had the smallest *b*-values, suggesting weaker structural constraints. A Kruskal–Wallis H test confirmed significant group differences, and post hoc analysis found a significant contrast between Classical Chinese and ChatGPT, while differences involving Modern Chinese were not significant.

These findings suggest that while large language models like ChatGPT can replicate surface-level Zipfian distributions, they do not fully reproduce the deeper efficiency pressures observed in human-authored language, particularly in highly compressed systems like Classical Chinese. The *b* parameter thus serves as a useful indicator of structural economy across production modalities.

There are some limitations for this study. First, the corpus was limited to 2,000 sentences per text type. Although balanced and systematically sampled, the dataset may not capture the full lexical variability of each modality. Second, Chinese segmentations relied on the NLP tools which, despite manual check, may overlook certain morphological subtleties. Third, ChatGPT outputs were generated using a fixed prompt under a single condition, which may have limited stylistic variation. Repeating the generation process under varied prompts and conditions may offer greater lexical and stylistic diversity.

Future research could extend the corpus to include larger and more genre-diverse datasets, compare across different LLMs (e.g., DeepSeek, Claude, Gemini, etc.), and incorporate additional complexity metrics such as syntactic depth, semantic density, or information-theoretic entropy. Longitudinal tracking of AI-generated texts across training iterations may also reveal whether structural economy emerges or erodes as model architectures evolve.

Acknowledgments

This work was partly supported by the National Social Science Foundation of China (Grant No. 24CYY064) and the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities in China (Grant No. 22JJD740018). The authors would like to sincerely thank the three anonymous reviewers for their valuable comments and constructive suggestions.

References

Christian Bentz and Ramon Ferrer-i-Cancho. 2016. Zipf's law of abbreviation as a language universal. In Christian Bentz, Gerhard Jäger, and Igor Yanovich (eds.), *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. University of Tübingen, Tübingen, pages 1–4. https://doi.org/10.15496/publikation-10057.

Maria A. Breiter. 1994. Length of Chinese words in relation to their other systemic features. *Journal of Quantitative Linguistics*, 1(3):224–231. https://doi.org/10.1080/09296179408590020.

Heng Chen, Junying Liang, and Haitao Liu. 2015. How does word length evolve in written Chinese? *PLOS ONE*, 10(9): e0138567. https://doi.org/10.1371/journal.pone.0138567.

Jee-Hyeon Eom. 2006. *Rhythmus im Akzent: Zur Modellierung der Akzentverteilung als einer Grundlage des Sprachrhythmus im Russischen*. Otto Sagner, München.

Ramon Ferrer-i-Cancho. 2025 On the class of coding optimality of human languages and the origins of Zipf's law. 1–7. https://doi.org/10.48550/arXiv.2505.20015.

Reinhard Köhler. 1986. *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Brockmeyer, Bochum.

Zhuolun Li and Lei Lei. 2025. Deciphering cross-genre dynamics: Testing the Law of Abbreviation and the Meaning-Frequency Law in Chinese across genres. *Lingua*, 320:1–18. https://doi.org/10.1016/j.lingua.2025.103934.

Guido M. Linders and Max M. Louwerse. 2023. Zipf's law revisited: Spoken dialog, linguistic units, parameters, and the principle of least effort. *Psychonomic Bulletin & Review*, 30:77–101. https://doi.org/10.3758/s13423-022-02142-9.

Ján Mačutek and Gejza Wimmer. 2013. Evaluating Goodness-of-Fit of Discrete Distribution Models in Quantitative Linguistics. *Journal of Quantitative Linguistics*, 20 (3):227–240. https://doi.org/10.1080/09296174.2013.799912.

George A. Miller, Edwin B. Newman, and Elizabeth A. Friedman. 1958. Length-frequency statistics for written English. *Information and Control*, 1(4):370–389. https://doi.org/10.1016/S0019-9958(58)90229-8.

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*, pages 101–108, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.14.

Udo Strauss, Peter Grzybek, and Gabriel Altmann. 2007. Word length and word frequency. In Peter Grzybek (ed.), *Contributions to the Science of Text and Language,* pages 277–294. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-4068-9_13.

Lu Wang. 2014. Synergetic studies on some properties of lexical structures in Chinese. *Journal of Quantitative Linguistics*, 21(2):177–197. https://doi.org/10.1080/09296174.2014.882186.

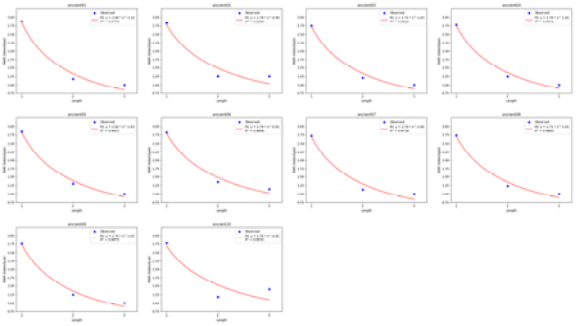
Koichi Yasuoka. 2019. Universal Dependencies Treebank of the Four Books in Classical Chinese. In *Proceedings of the 10th International Conference of Digital Archives and Digital Humanities (DADH2019),* pages 20–28.

George Kingsley Zipf. 1935. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin, Boston.

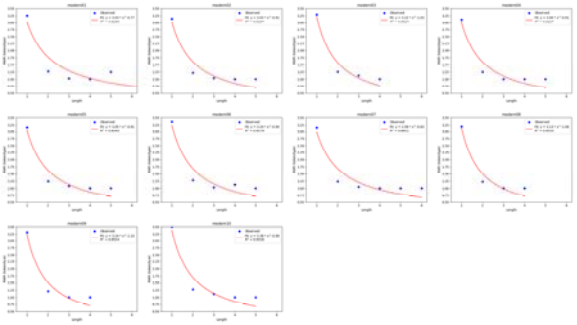
George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Hafner, New York.

Appendix A

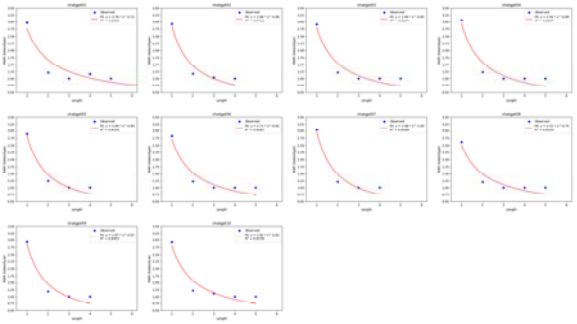
Power-law fitting results for the Classical, Modern and ChatGPT-Generated Chinese texts.



(a) Power-law fitting results for the Classical Chinese texts.



(b) Power-law fitting results for the Modern Chinese texts.



(c) Power-law fitting results for the ChatGPT-Generated Chinese texts.

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.

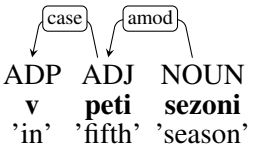


Figure 1: An example of a syntactic tree extracted from the Slovene subcorpus and corresponding to the structure *ADP <case ADJ <amod NOUN*.

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.

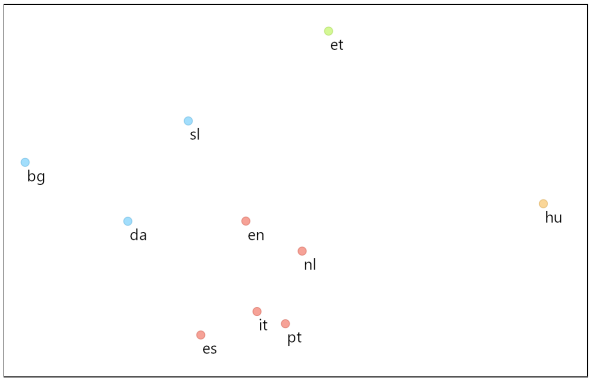


Figure 2: MDS Visualization of the Syntactic Profiles of the ELEXIS-WSD 1.2 Subcorpora.

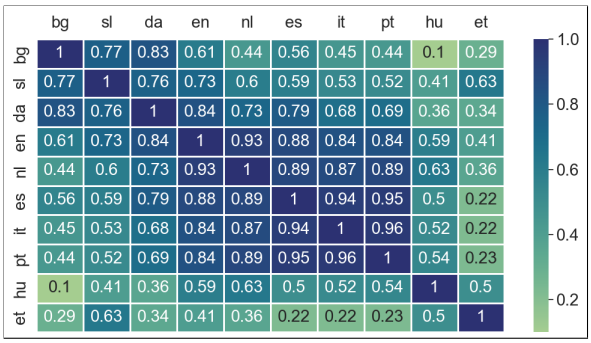


Figure 3: Matrix of cosine similarities between the syntactic profiles of individual ELEXIS-WSD subcorpora.

[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 η^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	η^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.

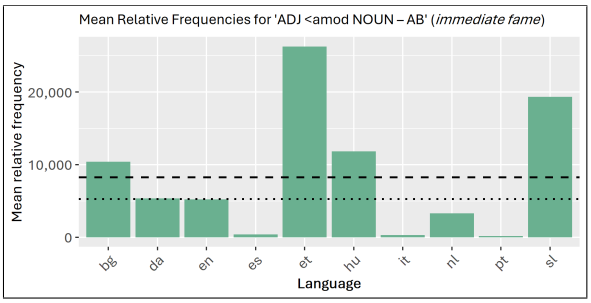


Figure 4: Mean relative frequencies (per million words) for the structure *ADJ <amod NOUN – AB*.

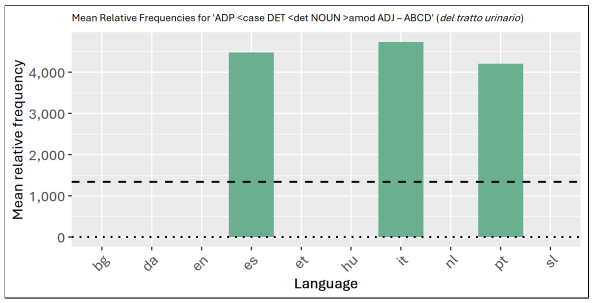


Figure 5: Mean relative frequencies (per million words) for the structure *ADP <case DET <det NOUN >amod ADJ – ABCD*.

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

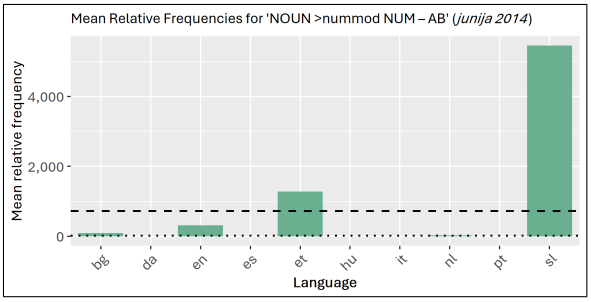


Figure 6: Mean relative frequencies (per million words) for the structure *NOUN >nummod NUM – AB*.

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.

The Interplay of Noun Phrase Complexity and Modification Type in Scientific Writing

Isabell Landwehr
Department of Language Science and Technology
Saarland University
isabell.landwehr@uni-saarland.de

Abstract

We investigate the interplay of noun phrase (NP) complexity and modification type, namely the choice between pre- and postmodification, using a corpus-based approach. Our dataset is the Royal Society Corpus (RSC; Fischer et al., 2020), a diachronic corpus of English scientific writing. We find that the number of dependents, length of the head noun and distance to the head noun's own syntactic head (typically the main verb) affect the likelihood of pre- vs. postmodification: NPs with more dependents are more likely to be premodified, NPs with a longer head noun and a head noun closer to its own head are more likely to be postmodified. In addition, we find an effect of syntactic role and definiteness as well as time: The likelihood of premodification over postmodification increases with time and subject NPs as well as indefinite NPs are more likely to be premodified than NPs in other syntactic roles or definite NPs.

1 Introduction

Language use has been argued to be shaped by optimization constraints (e.g. Levshina, 2022), such as minimizing dependency length between syntactic heads and dependents (e.g. Gibson, 1998; Gibson et al., 2000). This has also been posited for the register of English scientific writing (Degaetano-Ortlieb and Teich, 2022), in which complex noun phrases (e.g. NPs consisting of a head noun modified by several dependents) are a key feature (Halliday, 1988). English allows both premodification (e.g., in the form of nouns or adjectives, see Example 1) and post-modification (e.g. prepositional phrases or finite and non-finite relative clauses, see Example 2) in the noun phrase. The two types of modification may also occur at the same time (Example 3). [1]

(1) However, in this case we may proceed to calculate **the total plasma velocity** directly [...]. *(RSC, rsta_1996_0136)*

(2) But when **velocity relative to aether** was finally abandoned [...]. *(RSC, rsbm_1942_0016)*

(3) So far we have calculated **the flow velocity normal to the field lines** [...]. *(RSC, rsta_1996_0136)*

Previous studies on scientific writing have observed a diachronic shift from postmodification to premodification of the noun phrase (Degaetano-Ortlieb, 2021). Premodification results in more compressed structures than postmodification, which is particularly the case for nouns premodified by other nouns, i.e. compounds: Not only does a compound like *the plasma velocity* contain fewer words compared to a prepositional phrase (*the velocity of plasma*) or a relative clause (*the velocity which plasma possesses*), it also makes the semantic relationship between modifier and head implicit: The relationship between *plasma* and *velocity* could theoretically be interpreted as *plasma* ***has*** *velocity* (similarly to *eye color*), *velocity* ***consists of*** *plasma* (similarly to *stone pillar*) or *velocity* ***found in*** *plasma* (similarly to *forest animal*). [2] Selecting the correct semantic relation out of several competing ones is a crucial task in compound processing and high entropy of possible relations increases processing difficulty (Benjamin and Schmidtke, 2023). Moreover, the internal embedding structure of compounds may be ambiguous as well. A three-constituent compound such as *energy*

[1] Examples are taken from the Royal Society Corpus (RSC; Fischer et al., 2020).

[2] A detailed discussion and annotation scheme of semantic relations between constituents can be found in Gagné and Shoben (1997) or Ó Séaghdha (2007), among others.

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 72–82

flow velocity could refer to the *velocity of the energy flow* or the *flow velocity of energy*. This means that, while premodification streamlines linguistic structures, it also adds a new level of complexity. In addition, the choice of modification also influences other features of linguistic complexity like dependency length (see Section 2.2).

The present study aims to investigate how the increased complexity introduced by highly compressed structures interacts with other aspects of NP complexity. Premodification has become very common in scientific writing, but does this hold equally for all types of NPs, regardless of their complexity (e.g. in terms of number of dependents)? We assume that language users, in general but particularly in scientific writing, aim to maintain communicative efficiency (Levshina, 2022; Degaetano-Ortlieb and Teich, 2022), for instance by avoiding excessive complexity. Given processing constraints, we investigate the hypothesis that more complex NPs (e.g. with larger numbers of dependents) tend to be postmodified rather than premodified. Taking a corpus-based approach, we utilize Universal Dependencies annotation (de Marneffe et al., 2021) to consider different dependency-based complexity features. In our statistical analysis, we find that several complexity features influence modification type: In contrast to our original hypothesis, we find that a higher dependency number is associated with a greater likelihood of premodification. Features like larger distance to the verbal head and greater head noun length, on the other hand, are all associated with a greater likelihood of postmodification. Discourse status and syntactic role affect modification type as well, with indefinite NPs and subject NPs being more likely to be premodified. We also observe an effect of time in line with previous studies, namely an increased likelihood of premodification in later years.

This paper is structured as follows: Section 2 introduces relevant previous work on scientific writing and linguistic complexity, taking both psycholinguistic and theoretical approaches into account. It also motivates the selection of complexity features included in the analysis, while Section 3 describes our dataset and the preprocessing steps. Section 4 presents the statistical analysis, with a discussion in subsection 4.4. Section 5 on limitations and possible future research wraps up the paper.

2 Background and Rationale

2.1 Complex Noun Phrases in Scientific Writing

We conceptualize the general writing process in a similar way as described by Flower and Hayes (1981) and Hayes and Flower (1987): A writer first plans what they are going to write about (e.g. about the concept of *plasma velocity*). They then translate their ideas into syntactic structure, generating a sentence. During this step, NPs are used to encode the main concepts (e.g. *velocity*), possible modifiers further elaborate these core concepts (e.g. *plasma* and *total*). If modifiers are included, a choice between premodification, postmodification or a combination of both needs to be made at this stage. In a final step, the writer revises and edits the produced text. These three main steps can overlap and be repeated recursively. In scientific writing, complex NPs fulfill a central role for encoding concepts: Nominalization is a key feature of this register and NPs frequently describe very technical and specialized concepts (Halliday, 1988; Banks, 2008). Historically, there has been a development from an emphasis on clausal structures to an emphasis on phrasal structures, allowing information to be conveyed in a more compressed way (Biber and Gray, 2011). This is particularly exemplified by the case of compounds, which are informationally denser than their prepositional counterparts, and which have increased in frequency over time (Degaetano-Ortlieb, 2021). In this way, scientific English writing evolved to be optimized for written communication among experts (Degaetano-Ortlieb and Teich, 2022).

Writing a formal text such as a scientific article is also an audience-directed process (Hayes and Flower, 1987): The writer aims to accommodate the needs of the potential reader(s) and to make the text understandable. This means that possible processing demands on the reader need to be considered as well.

2.2 Processing Complex Structures

Previous studies have analyzed linguistic complexity from different perspectives. The processing cost associated with complex structures and the influence of complexity features on constituent order have been of particular interest.

A frequently investigated feature of complexity is dependency length, which describes the distance between a syntactic head and its dependent(s).

Greater distance has been associated with increased processing cost: According to Dependency Locality Theory (DLT; Gibson, 1998; Gibson et al., 2000), greater distance means that the prediction about upcoming material needs to be kept in memory for a longer time. This increases the cost for maintaining the prediction and for finally integrating it into the mental syntactic representation. Dependency length has therefore been proposed to measure processing difficulty (Liu, 2008). Support for DLT comes from various studies: Gibson (1998) showed its ability to account for different complexity phenomena, such as the processing of subject- versus object-extracted relative clauses. Liu (2008) analyzed dependency distance in a corpus study covering 20 languages and found a trend towards minimization of average dependency distance. Demberg and Keller (2008) found that DLT successfully predicts the reading times for nouns, while Temperley (2007) tested its predictions from a production perspective. Accordingly, the principle of Dependency Length Minimization (DLM) has been proposed, which posits that language users aim to place syntactic heads and dependents in proximity to each other (Futrell, 2019) and is often regarded as a linguistic universal (Liu et al., 2017). Similarly, dependency locality has been associated to information locality, e.g. by Levshina (2022), who argues that language users aim to minimize dependency length in order to maintain communicative efficiency.

Another complexity feature is the length of syntactic constituents, which has been found to affect constituent order: Behaghel (1909) already observed that long, complex phrases tend to occur at the ends of clauses (called *end-weight* in other studies, see e.g. Eitelmann (2016)). Discourse status needs to be considered as well: *Given* information tends to precede *New* information (Gundel et al., 1988; Prince, 1992). Arnold et al. (2000) found that heavy and new NPs tend to be postponed in the sentence, giving the speaker more time to plan the utterance and easing memory load on the listener. Syntactic role has also been considered when investigating dependency length, constituent length and discourse status: Temperley (2007), for instance, found that in written English, direct objects tend to be longer than subjects, and that postmodifying adverbial clauses tend to be longer than premodifying adverbial clauses.

In addition, word length itself has also been shown to affect processing (Baddeley et al., 1975; Jalbert et al., 2011; Guitard et al., 2018): Shorter words are recalled better than longer words, indicating a higher load on working memory associated with longer words.

Focusing specifically on the effect of NP structure on language understanding, an experimental study by Mota and Igoa (2017) compared simple NPs, which consisted of a series of coordinate NPs, and complex NPs, which contained embedded prepositional phrases. They found that language comprehenders were sensitive to the NP complexity, but only in the case of subject NPs.

2.3 Rationale

We investigate how different features of NP complexity interact with modification type. We selected the features based on previous literature (see Sections 2.1 and 2.2) and chose to consider all of them in order to limit possible confounding effects and improve the validity of our results. Modification type may be influenced by the overall **dependency length**, the distance between the head noun of the NP to its own verbal head. If many tokens already intervene between an NP and its head, the choice between pre- and postmodification can further increase this distance, depending on the **syntactic role** of the NP: A subject NP's distance to the head is increased by postmodifiers, an object or oblique's distance to the head is increased by premodification. In Example 2, for instance, the distance between *velocity*, the subject NP's head noun and *abandoned*, its verbal head, is 6 steps. Without the subject NP's postmodification (*relative to aether*), the distance would be only 4 steps. Similarly, in Example 1, the distance between the direct object *velocity* and its verbal head *calculate* is 4 steps, which would be only two steps without the premodifiers *total* and *plasma*. Following the principle of Dependency Length Minimization, we therefore predict that for subjects, premodification is preferred, while objects and obliques display a preference for postmodification.

The **number of dependents** affects the distance to the head as well, again depending on the syntactic role: We expect subjects with a large number of modifiers to show a preference for premodification, while objects and obliques are expected to display a preference for postmodification. We also predict that a greater **length of the head noun** decreases the likelihood of pre-modification: Larger structures increase memory load, in which case post-modification as the less complex modification

type may be preferred to reduce overall complexity. Moreover, discourse status needs to be considered, for which we use **definiteness** as a proxy. We consider discourse-new information to be more complex than given information: In order to limit excess complexity, we therefore predict new NPs (here: NPs without a definite determiner) to show a smaller likelihood of premodification, the more compressed and complex alternative, than given NPs (NPs with a definite determiner). Examples 1 and 2 would fulfill this expectation: The discourse-given NP *the total plasma velocity* in Example 1 is premodified, while the discourse-new NP *velocity relative to aether* in Example 2 is postmodified.

Finally, we expect to see an effect of **time**: Due to the diachronic development of scientific English towards an efficient register with more and more compressed structures (Degaetano-Ortlieb and Teich, 2022), the likelihood of premodification should increase with the progression of time.

3 Dataset

We use the Royal Society Corpus (RSC; Fischer et al., 2020; Menzel et al., 2021), a diachronic corpus of English scientific writing. The full version 6 contains the *Philosophical Transactions and Proceedings of the Royal Society of London* from 1665 to 1995, with over 290 million tokens in more than 47,000 documents. The corpus was built in accordance to FAIR principles (Wilkinson et al., 2016), preprocessed using standard tools (Baron and Rayson, 2008; Schmid, 1995) and annotated with meta-data (Menzel et al., 2021). These include author, year of publication, text type (e.g. article, lecture, report, obituary), primary topic and journal (e.g. Series A - Mathematics and Physics, Series B - Biology).

We use version 6.0.3 which was parsed with the Python package *stanza* (Qi et al., 2020) and contains Universal Dependencies annotations (de Marneffe et al., 2021; Nivre et al., 2017). This version ("good sentences version") contains fewer tokens than the full version 6, since ungrammatical sentences or sentences which might be problematic for parsing (e.g. appendices, image titles, foreign language sentences) were not considered. Table 1 shows the composition of this corpus version over time.

Years	# Texts	# Tokens
1665-1699	1,312	2,194,828
1700-1749	1,674	2,895,445
1750-1799	1,806	5,037,372
1800-1849	2,709	7,001,970
1850-1899	5,502	12,923,443
1900-1949	6,879	21,014,576
1950-1996	20,413	78,142,577

Table 1: Composition of the Royal Society Corpus (version 6.0.3) over time.

4 Statistical Analysis

4.1 Preprocessing

For this study, we sampled 3,805 documents from the corpus. We used stratified sampling by publication year, meaning that the proportion of documents per year of the whole corpus was maintained in the sample.

Using a script written in Python (Van Rossum and Drake, 2009), version 3.10.15, we extracted the noun phrases from these documents. We consider only NPs headed by a common noun and only top-level NPs, i.e. NPs which are not embedded in other NPs. For these, we extracted the following linguistic features: head noun, number of dependencies of the head noun, syntactic role of the head noun, number of modifiers and modification type (premodification, postmodification, both). We also extracted the following metadata features of the noun or its context: text ID, sentence ID, head noun ID, publication year, author(s), text type and journal. This procedure resulted in information about 1,986,592 NPs.

For the statistical analysis, we filtered the data: First, we considered only NPs which possessed at least one modifier and were either pre- or post-modified, but not both. Determiners were not counted as modifiers, but as dependents of the head noun. We removed outliers which were most likely the result of parsing errors: NPs with more than 20 dependents, NPs with a distance to the head greater than 25, NPs with a head noun consisting of more than 20 characters. We also only focused on some syntactic roles since many roles were not attested frequently enough in our data (e.g. indirect objects, roots of a sentence). We considered nominal subjects, direct objects and oblique arguments. For the nominal subjects and the obliques, the various sub-categories (e.g. *oblique agent*) were subsumed into the overall category (e.g. *oblique*). We only

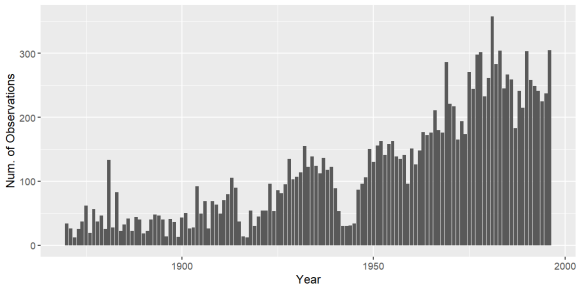


Figure 1: Number of observations per year.

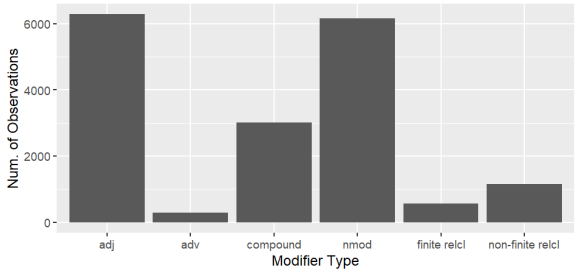


Figure 2: Number of observations of each modifier type.

considered the text type *article* in order to focus on scientific writing proper and to exclude non-scientific text types, such as obituaries and biographies. In addition, scientific text types other than *article* (e.g. lectures, speeches) were not strongly represented in the sample and contributed much fewer observations.

This filtering still resulted in 746,817 NPs, so we again applied stratified sampling by publication year, resulting in N = 14,934 observations to be included in the statistical analysis. The final dataset included 7,353 postmodified NPs and 7,581 premodified NPs. 7,843 NPs had no definite determiner, while 7,091 possessed a definite determiner. Most NPs (6,055) were oblique arguments, while 5,924 NPs acted as nominal subjects and 2,955 NPs as direct objects in their sentences. Most observations (5,605) stemmed from the journal *Proceedings of the Royal Society, Series A*, which encompasses the disciplines of mathematics, physics and engineering. An overview of the temporal distribution of our observations is given in Figure 1, with publication years ranging from from 1870 to 1996. The different types of modification in our data sample are shown in Figure 2.

4.2 Regression Model

We fit a mixed-effects logistic regression model in the statistical programming language R, version 4.4.2 (R Core Team, 2024) and using the library *glmmTMB* (Brooks et al., 2017). We chose regression modeling due to the large number of theoretically-motivated predictor variables which we took into account here: Mixed-effects regression modeling allows us to consider all of the variables and is appropriate given the hierarchical nature of corpus data and the resulting dependencies among observations (several observations from the same journal, author etc.).

Our dependent variable was modification type, with the levels *premodification* and *postmodification*. As predictor variables, we included year, distance to syntactic head, length of the head noun (in characters), discourse status (operationalized as the presence of a definite determiner for the status *Given*), sentence length (in number of words) and an interaction of dependency number and syntactic role. We also tested an interaction of dependency length and syntactic role, however, this model did not converge. The variable year was centered for ease of interpretation with regards to the intercept, all other numerical variables were centered and scaled. The factor variables were treatment-coded, with *postmodification* as the baseline for modification type and *nominal subject* as the baseline for syntactic role. To account for within-group variability, we included random intercepts for journal, author and noun, as well as a by-noun random slope for number of dependencies. Testing the variables for multicollinearity using the library *performance* (Lüdecke et al., 2021) revealed only mild correlation between the variables (variance inflation factors < 5). Model diagnostics (e.g. inspection of residuals) were performed with the package *DHARMa* (Hartig, 2024) and showed no overly problematic trends.

4.3 Results

The full model summary (Table 2) is included in Appendix A.

We found significant effects of year (p <0.001, z = 7.99, Figure 3), number of dependencies (p <0.001, z = 8.69, Figure 4), distance to syntactic head (p <0.001, z = -28.48, Figure 5), noun length (p <0.001, z = -6.67, Figure 6) and sentence length (p <0.001, z = -6.51, Figure 7): Over time, the likelihood of premodification over postmodification increases. The likelihood of premodification also increases for NPs with more dependencies. However, it decreases with greater distance to the head and with longer nouns and sentences.

We also found a significant effect of definiteness

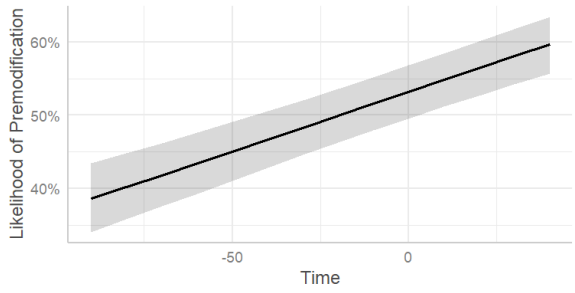


Figure 3: Effect of time on premodification likelihood: NPs in later years are more likely to be premodified.

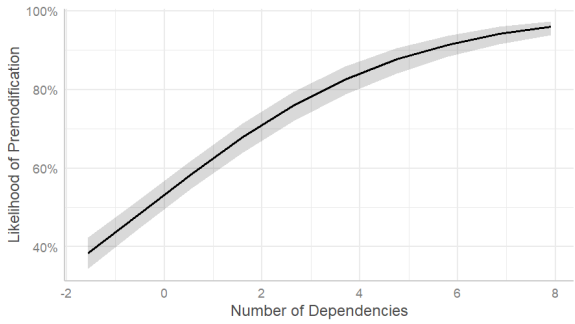


Figure 4: Effect of dependency number on premodification likelihood: NPs with more dependents are more likely to be premodified.

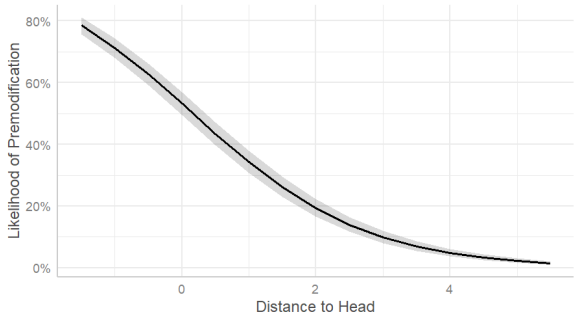


Figure 5: Effect of distance to head on premodification likelihood: NPs with greater distance to their head are less likely to be premodified.

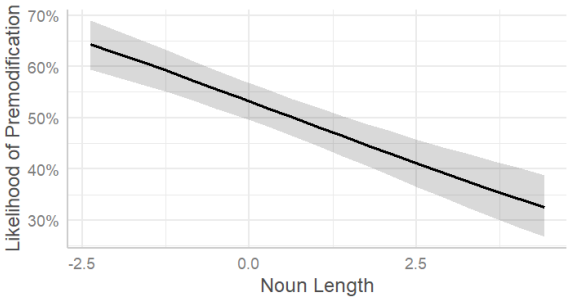


Figure 6: Effect of head noun length on premodification likelihood: NPs with a longer head noun are less likely to be premodified.

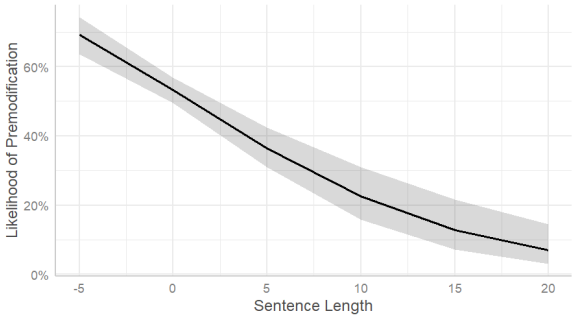


Figure 7: Effect of sentence length on premodification likelihood: NPs in longer sentences are less likely to be premodified.

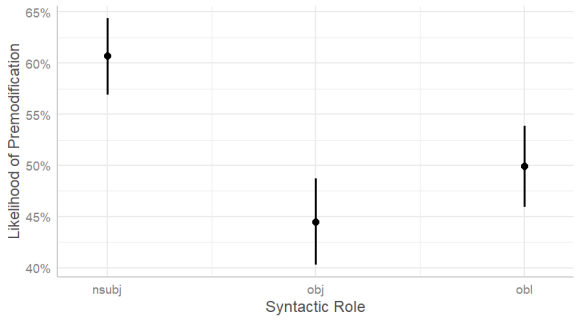


Figure 8: Effect of syntactic role on premodification likelihood: Subjects are the most likely to be premodified, followed by obliques and then direct objects.

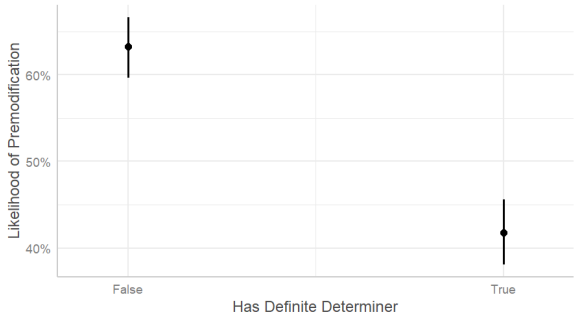


Figure 9: Effect of definiteness on premodification likelihood: Indefinite NPs are more likely to be premodified.

(p <0.001, z = -19.20, Figure 9) and of syntactic role (Figure 8): NPs without a definite determiner (i.e. NPs with either an indefinite or no determiner) had a higher chance of being premodified than noun phrases with a definite determiner. Compared to nominal subjects, NPs acting as direct objects (p <0.001, z = -9.94) or obliques (p <0.001, z = -7.65) had a lower chance of being premodified, with direct objects being the most unlikely to have premodification.

The interaction of dependency number and syntactic role, however, was not significant.

4.4 Discussion and Conclusion

Some of our predictions could be confirmed by the statistical analysis: In line with the principle of DLM, we observed a strong preference for subjects to be premodified, while direct objects and obliques were more likely to be postmodified (such as the object NP in Example 4). Moreover, a greater length of the head noun decreased the likelihood of premodification (consider the NP in Example 5 with a 7-syllable head noun). This supports the hypothesis that, in the face of higher memory load, language users opt for a less compressed modification type in order to ease processing difficulty. High memory load may also be an explanation for the observation that NPs in longer sentences are less likely to be premodified.

(4) Studies of these pterosaurs have revealed **a number of general features with regard to patterns of bone ossification** [...]. *(RSC, rspb_1996_0008)*

(5) **Reproducibility of results**, greater methodological awareness, and more rigorous assessment of hypothesis robustness are identified as additional issues [...]. *(RSC, rspb_1996_0205)*

Contrary to our expectations, however, definite NPs were less likely to be premodified than indefinite NPs. Our expectation may not have been confirmed because the notion of givenness was insufficiently operationalized by definiteness.

A larger distance to the verbal head was generally associated with a decreased likelihood of premodification. DLM can explain this for objects and obliques: In their case, premodification further increases the distance to the verbal head and should therefore be avoided. For subject NPs, a larger distance to their head in combination with postmodification (see Example 6) might be the result of an attempt to avoid the increased compression of premodified structures when memory load is high: If many dependents need to be integrated in the NP and stored in memory, writers might aim to avoid additional processing strain by selecting a less complex modification type.

(6) **Adoption of cladistic methods by students of archosaurs** has clearly been a slow and gradual process. *(RSC, rspb_1996_0205)*

It is interesting that the influence of dependency number was not modulated by syntactic role (non-significant interaction): Contrary to our expectations, NPs with more dependents were generally more likely to be premodified and not only in the case of subjects. This may be due to the influence of predictability: Some constituents of complex NPs might actually be very commonly used together and have a high transitional probability between the constituents. Encountering the first

element(s) of such an NP might lead the reader to correctly predict the whole structure. Over time, this facilitating effect for comprehension may lead to a preference in production. This may in particular be the case for compounds, highly compressed structures, which are derived from a process between syntax and morphology. Compound processing has been shown to be influenced by various factors such as constituent frequency, compound frequency, compound word length, compound family size or semantic transparency (Baayen et al., 2010; Schmidtke et al., 2021). Some factors actually have a facilitating effect on processing, so that compounds with high-frequency constituents and high semantic transparency are processed faster than compounds with low-frequency constituents and low semantic transparency. These effects might counteract and outweigh factors decreasing processing speed. A technical term consisting of an NP with several nominal modifiers, such as *heat shock cognate protein*, might be considered complex when judging merely from its syntactic structure: However, since *heat shock* and *protein* as well as *cognate* and *protein* co-occur frequently in biochemical texts and are established terms, this syntactic complexity might be outweighed by lexical frequency effects.

This study gives further support to the principle of Dependency Length Minimization and shows that it is also relevant for the choice between pre- and postmodification. It also supports the hypothesis that premodification might indeed be adding complexity to an NP, since it is dispreferred in the case of longer dependencies or for NPs with longer head nouns. However, this analysis also highlights that other factors, such as predictability and co-occurrence patterns, need to be considered as well when investigating optimization mechanisms in language use. Overall, this analysis supports the theory that optimization plays an important role in the evolution of scientific writing (Degaetano-Ortlieb and Teich, 2022). Considering the key role of NPs in scientific language, the results highlight the fact that these optimization pressures also act on the NP-level.

From a diachronic perspective, our study shows that the likelihood of premodification increases over time, even when controlling for other variables influencing the choice. This points towards a conventionalization trend within scientific writing: As register-specific norms become established over time, more compressed NPs are preferred, possibly outweighing competing constraints.

5 Limitations and Outlook

A major limitation of this analysis is the way discourse-given and discourse-new NPs were identified: While givenness and definiteness are correlated in English, they are not identical (Gundel et al., 1988): An NP with a demonstrative is usually given, an NP with a definite article, on the other hand, does not necessarily have to be given, only uniquely identifiable (Gundel et al., 1993). An investigation with more refined discourse annotation might lead to clearer insights on the influence of this factor.

Moreover, since our focus was on NPs headed by a common noun, other possible heads of NPs, like pronouns and proper names, were not included in this analysis. Future investigations should consider them as well in order to investigate if the results presented here are generalizable to pronouns and proper names. Special consideration should also be given to compounds, since the relationship between head and modifier(s) of a compound are presumably stronger than between head and phrasal modifiers.

NPs with both pre- and postmodification were also not considered here. It might be interesting to look at them in future research: Which dependents are added before and which after the head noun? Length and internal structure of modifiers are also a factor of interest, since modifiers may themselves contain heads with dependents. Investigating these aspects more closely may shed more light on the internal order of NP constituents and on the question whether the same principles apply here as for the clause level. It may also illustrate in more detail how competing pressures interact with each other in the process of language optimization.

Acknowledgements

The author would like to thank Elke Teich and Stefania Degaetano-Ortlieb as well as three anonymous reviewers for their insightful remarks on a previous version of this paper. This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102 Information Density and Linguistic Encoding.

References

Jennifer E. Arnold, Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. Newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1):28–55.

Harald Baayen, Victor Kuperman, and Raymond Bertram. 2010. Frequency effects in compound processing. In *Cross-disciplinary Issues in Compounding*, pages 257–270. John Benjamins Publishing Company.

Alan D. Baddeley, Neil Thomson, and Mary Buchanan. 1975. Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 14(6):575–589.

David Banks. 2008. *The Development of Scientific Writing: Linguistic Features and Historical Context*. University of Toronto Press.

Alistair Baron and Paul Rayson. 2008. Vard2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*.

Otto Behaghel. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*, 25:110.

Shaina Benjamin and Daniel Schmidtke. 2023. Conceptual combination during novel and existing compound word reading in context: A self-paced reading study. *Memory & Cognition*, 51(5):1170–1197.

Douglas Biber and Bethany Gray. 2011. Grammatical change in the noun phrase: The influence of written language use. *English Language & Linguistics*, 15(2):223–250.

Mollie E. Brooks, Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Maechler, and Benjamin M. Bolker. 2017. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400.

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

Stefania Degaetano-Ortlieb. 2021. Measuring informativity: The rise of compounds as informationally dense structures in 20th-century scientific English. In *Corpus-based Approaches to Register Variation*, pages 291–312. John Benjamins Publishing Company.

Stefania Degaetano-Ortlieb and Elke Teich. 2022. Toward an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory*, 18(1):175–207.

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

Matthias Eitelmann. 2016. Support for end-weight as a determinant of linguistic variation and change. *English Language & Linguistics*, 20(3):395–420.

Stefan Fischer, Jörg Knappen, Katrin Menzel, and Elke Teich. 2020. The Royal Society Corpus 6.0: Providing 300+ years of scientific writing for humanistic study. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 794–802. European Language Resources Association.

Linda S. Flower and John R. Hayes. 1981. A cognitive process theory of writing. *College Composition & Communication*, 32(4):365–387.

Richard Futrell. 2019. Information-theoretic locality properties of natural language. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 2–15.

Christina L. Gagné and Edward J. Shoben. 1997. Influence of thematic relations on the comprehension of modifier–noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(1):71.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson et al. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, Language, Brain*, 2000:95–126.

Dominic Guitard, Andrew J. Gabel, Jean Saint-Aubin, Aimée M. Surprenant, and Ian Neath. 2018. Word length, set size, and lexical factors: Re-examining what causes the word length effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(11):1824.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1988. On the generation and interpretation of demonstrative expressions. In *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307.

Michael A. K. Halliday. 1988. On the language of physical science. *Registers of written English: Situational factors and linguistic features*, 162:177.

Florian Hartig. 2024. *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.4.7.

John R. Hayes and Linda S. Flower. 1987. On the structure of the writing process. *Topics in Language Disorders*, 7(4):19–30.

Annie Jalbert, Ian Neath, Tamra J. Bireta, and Aimée M. Surprenant. 2011. When does length cause the word length effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(2):338.

Natalia Levshina. 2022. *Communicative Efficiency*. Cambridge University Press.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193.

Daniel Lüdecke, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. 2021. performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60):3139.

Katrin Menzel, Jörg Knappen, and Elke Teich. 2021. Generating linguistically relevant metadata for the Royal Society Corpus. *Research in Corpus Linguistics*, 9(1):1–18.

Sergio Mota and José Manuel Igoa. 2017. Parsing complex noun phrases: Effects of hierarchical structure and sentence position on memory load. *The Spanish Journal of Psychology*, 20:E37.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Diarmuid Ó Séaghdha. 2007. Designing and evaluating a semantic annotation scheme for compound nouns. In *Proc. Corpus Linguistics*, pages 1–17.

Ellen F. Prince. 1992. The ZPG letter: Subjects, definiteness, and information-status. In *Discourse Description: Diverse linguistic analyses of a fund-raising text*, pages 295–326. John Benjamins Publishing Company.

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*, pages 101–108, Online. Association for Computational Linguistics.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

Daniel Schmidtke, Julie A. Van Dyke, and Victor Kuperman. 2021. Complex: An eye-movement database of compound word reading in english. *Behavior Research Methods*, 53:59–77.

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

Guido Van Rossum and Fred L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):1–9.

A Appendix: Regression Model Summary

	Est.	**SE**	***z***	***p***
Intercept	0.85	8.48e-02	10.03	<0.001
Year	6.55e-03	8.20e-04	7.99	<0.001
Dependency Number	3.77e-01	4.34e-02	8.69	<0.001
Distance to (Verbal) Head	-7.80e-01	2.74e-02	-28.48	<0.001
Head Noun Length	-1.95e-01	2.92e-02	-6.72	<0.001
Syntactic Role *direct object*	-6.57e-01	6.61e-02	-9.94	<0.001
Syntactic Role *oblique*	-4.39e-01	5.74e-02	-7.65	<0.001
Definiteness	-8.73e-01	4.55e-02	-19.20	<0.001
Sentence Length	-1.37e-01	2.10e-02	-6.51	<0.001
Dep. Num * Synt. Role *direct obj.*	5.65e-02	6.79e-02	6.86e-01	0.493
Dep. Num * Synt. Role *oblique*	-7.69e-04	5.54e-02	-1.40e-02	0.989

Table 2: Regression model summary.

Predictability Effects of Spanish-English Code-Switching: A Directionality and Part of Speech Analysis

Josh Higdon
University of Florida
j.higdon@ufl.edu

Valeria Pagliai
University of Florida
vpagliai@ufl.edu

Zoey Liu
University of Florida
liu.ying@ufl.edu

Abstract

Research on code-switching (CS), the spontaneous alternation between two or more languages within a discourse, remains relatively new and often limited by the use of elicited production tasks, with some exceptions leveraging naturalistic corpora. This study analyzes the effects of language directionality and part-of-speech (POS) tags on Spanish-English CS production between corpus modalities and speech communities. We use data from two spoken corpora: Miami Bangor Corpus (MBC; N = 261,711) and Spanish in Texas Corpus (STC; N = 416,784), as well as the written LinCE Corpus (N=278,093). Bootstrap analyses indicate that Spanish serves as the matrix language (i.e., the most used) for MBC and LinCE, while English is for STC. Logistic regression analyses show that the particle-coordinating conjunction combination was the strongest POS predictor of a CS. The results suggest that corpus modality and the speech community affect matrix language proportions and that both previously attested and unseen POS combinations modulate the production of Spanish-English CS.

1 Introduction

Code-switching (CS), or the fluid switching between languages in bilingual speech or text (Poplack, 1980), is ubiquitous in bilingual communities. This language alternation is known to be structured yet spontaneous (Myers-Scotton, 1993), as it is believed that two (or more) languages are active in the speaker's brain (Van Hell et al., 2015). CS is categorized as inter-sentential (e.g., "No saldré hoy [*I am not going out today*]. I'm too tired.") or intra-sentential (e.g., "She bought una casa en Florida [*a house in Florida*]"). Since languages are not used with the same frequency, the most dominant is considered the matrix language, strongly influencing the morphosyntactic structure of the discourse (Myers-Scotton, 1993).

Previous research has identified several characteristics of CS. For example, the presence of cognates seems to trigger a language switch (Kootstra et al., 2020), and the POS combination of determiner-noun has been noted as a frequent point of switch (Balam et al., 2020). Specifically in the case of Spanish-English CS, there is evidence suggesting that Spanish often serves as the matrix language in CS (Carter et al., 2010), and, regardless of this matrix, there is a consistent pattern where a Spanish determiner is followed by an English noun (e.g. *la house*) (Toribio, 2023). However, these studies mainly rely on (small) sets of controlled stimuli from psycholinguistic experimentation (cf. Soto et al. (2018); Winata et al. (2023)), lacking analysis of naturalistic data (e.g., corpora), meaning that it is unclear whether the mentioned patterns are present in naturalistic CS speech.

Our work addresses these limitations. Leveraging three Spanish-English CS corpora covering both spoken and written modalities, we ask: (1) To what degree is CS production constrained by the language directionality (i.e., Spanish-English vs. English-Spanish) of the CS? (2) To what degree can the POS combination of a pair of words predict whether a CS occurs between those words? (3) If directionality and POS effects are found, are these findings modulated by corpus modality (oral vs. written) and speech community?

Given that Spanish tends to occur as the matrix language in Spanish-English CS (Carter et al., 2010), we predict that Spanish-English CS occurs at higher proportions than in the opposite direction. We also predict that particular POSs will be able to predict the occurrence of CS, specifically with determiners (DETs, among other tags) occurring as the first word in a bigram containing a CS (Eichler et al., 2012; Balam et al., 2020) and conjunctions (CONJ) occurring as the second word in a bigram containing a CS (Soto et al., 2018).

It is difficult to hypothesize the answers to the

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 83–89

third question, since there is a lack of literature on Spanish-English CS in the written domain, and also that the effect of speech community on CS has been generally neglected (Chan, 2009; Couto et al., 2021). Regarding the effect of corpus modality on CS production, it might not be surprising to see differences in matrix language selection between oral and written corpora, taking into account various differences (including register variations) between modalities (Rabinovich et al., 2019). In one of the few studies examining the effects of speech community on Spanish-English CS, Blokzijl et al. (2017) compared determiner-noun (DET-N) CS production in the Miami Bangor Corpus (MBC) to a corpus of interviews conducted in Nicaragua. The authors found significantly higher rates of Spanish-English DET-N CSs in the MBC, while English-Spanish DET-N rates were higher in the Nicaragua Corpus (Blokzijl et al., 2017). As such, we conjecture that differences in regional norms may manifest in directionality effects. For example, Spanish-English CS may occur at higher rates in the MBC than the Spanish-English proportions of the Spanish Texas Corpus (STC), since Spanish has been documented as the matrix language in Spanish-English CS in the MBC (Carter et al., 2010), but not in STC.

2 Related Work

Recent experiments have used computational analyses of linguistic corpora to uncover Spanish-English CS trends (Winata et al., 2023), but they face limitations in two respects. First, most experiments focus on modeling distributional trends in CS (e.g., the preference for the *estar* + English gerund switch (Tsoukala et al., 2019)) or improving a model's ability to diagnose the presence of a CS (Iliescu et al., 2021) through semi-supervised language identification methods that leverage monolingual data and models like Viterbi decoding. Second, much of the corpus-based research on Spanish-English CS is limited to one corpus, such as the MBC (Deuchar et al., 2009) or the STC (Bullock and Toribio, 2024), raising the question of whether the trends discovered would be representative of those across speech modalities and communities. Furthermore, given that non-computational analyses have found regional differences in DET-N CS patterns, it stands to reason that these differences would appear in CS behavior (Blokzijl et al., 2017). However, this analysis did not account for regional differences in other structural properties of Spanish-English CS (e.g., other POS pairings or trends in matrix language usage).

Closer to this research, through a cross-linguistic analysis of Spanish-English and Mandarin-English corpora, Chi and Bell (2024) highlight POS tags as strong predictors of CS and expand this idea by concluding that this predictive strength decreases the farther a word is from the CS point. While their study underscores the value of POS based approaches, their Spanish-English analysis was also limited to the MBC. We extend overall previous research by comparing CS patterns in three corpora of different modalities, identifying a range of POS tag combinations that predict the presence or absence of CS, and testing prior assumptions about Spanish as the default matrix language. We highlight the role of modality and speech community as key factors shaping CS behavior.

3 Experiments

3.1 Datasets and preprocessing

In contrast to prior corpus-based studies on Spanish-English CS, which typically used only one of these corpora, we selected three corpora that enabled us to include both spoken and written data across different speaking communities.

Miami Bangor Corpus The Miami Bangor Corpus (MBC) (Deuchar et al., 2009), previously used in studies of bilingual speech (Soto et al., 2018; Soto and Hirschberg, 2017), consists of transcriptions of conversations from 86 speakers (33 male, 53 female) based in Miami, Florida. The mean speaker age was 33 years old, and 91.6% reported having at least a college education. On average, speakers acquired Spanish between 2 and 4 years old and English between 4 years old and primary school age. The MBC corpus comprises of a total of 242,475 words, of which 63% is in English, 34% in Spanish, and the rest is undetermined.

Spanish in Texas Corpus The Spanish in Texas Corpus (STC) (Bullock and Toribio, 2024) consists of transcribed interviews and conversations with speakers based in different cities across Texas. It contains approximately 500,000 words from 96 speakers (36 male and 60 female) whose mean age was 39.1 years upon corpus creation. Although information about age of acquisition was not directly provided, 78.1% of speakers reported speaking primarily Spanish with their parents. Additionally, 92.7% of speakers reported having received at least

a high school education. As such, we highlight the similarities between the MBC and the STC in terms of speaker backgrounds. However, in terms of the linguistic makeup of the corpus, the STC is 96% Spanish and 4% English, making it the corpus with the most pronounced language imbalance (and the highest proportion of Spanish) in this study.

LinCE Corpus Unlike the previous corpora, which consist of spoken data, the LinCE Corpus (LinCE) (Aguilar et al., 2020) is the only one based on written texts retrieved from X (formerly Twitter). No background information regarding the speakers of the LinCe corpus is publicly available. The LinCe corpus contains 390,953 words, of which approximately 33% are in Spanish, 64% in English, and the remaining portion is undetermined.

Although the corpora contain POS information, this is not directly comparable across all three due to differences in coding schema. To ensure consistency and scalability, we automatically annotated the POS tags using tools aligned with the Universal Dependencies (UD) annotation scheme (de Marneffe et al., 2021), instead of the tags provided within the data. This decision was also motivated by our observations of several inconsistencies in the existing annotations. We did retain the language tags present in all three corpora and based the subsequent analysis on them.

We used Stanza (Qi et al., 2020) to parse every sentences with the English model. Each word was then processed based on the language tag they had in the original data. Spanish words were reanalyzed with spaCy's es_core_news_sm model, which is trained on the AnCora corpus (Taulé et al., 2008), while unrecognized English words (POS tagged as "X") were reanalyzed with the spaCy model en_core_web_sm to resolve this. We applied spaCy's models at the token level to ensure accurate annotation of individual words within mixed-language utterances.

To investigate the effects of POS production on CS, we analyzed the relationship between the occurrence (or lack thereof) of CS and POS tag bigrams. This was motivated by Soto et al. (2018) who also examined the roles of POS tags on CS production. After filtering punctuation at the sentence level for each corpus, we extracted POS bigrams sequences and included their corresponding language tags to diagnose the presence of a CS. If the two words in a bigram belonged to different languages, we categorized it as a CS.

3.2 Statistical analyses

To examine language directionality effects on CS production, we subjected each corpus to bootstrapping analysis (Efron and Tibshirani, 1994). We calculated the number of bigrams in which the language of the first word was Spanish and that of the second word was English, indicating a Spanish-English CS. The reverse was true for calculating the number of English-Spanish CSs. We divided the count of each CS type (Spanish-English, English-Spanish) by the number of total CSs to calculate its proportions for both CS types. We conducted 10,000 iterations of this process to derive 95% confidence intervals (CIs). Mean proportions and CIs were calculated using a) exclusively the POS bigrams labeled as CSs, and b) all three corpora combined to understand the quantity of each CS type relative to the entire corpus.

We used a logistic regression model (Cox, 1958) to analyse POS effects on CS production across data. Soto et al. (2018) found that a variety of tags, such as determiners (DET), nouns (NOUN), pronouns (PRON), subordinating conjunctions (SCONJ), tend to be associated with CS occurrence; however, their analyses were derived from descriptive statistics with Chi-squared tests. In contrast, our usage of logistic regression models enables us to reliably assess whether a certain combination of POS tags can predict the presence of a CS. In detail, for each POS bigram, which was used as the fixed effect, a binary variable was created to indicate whether a CS was ('1') or was not ('0') present; this variable was the dependent variable. To control for the potential effect of the individual corpus on CS production, we included the corpus as a fixed effect, with the MBC corpus as the reference level (see the formula below).

$$\text{CS\_OCCURRENCE} \sim \text{POS BIGRAM} + \text{CORPUS}$$

All regression models were fit using the glm function in R version 4.3.3 (R Core Team 2018).

4 Results

4.1 Bootstrapping analysis

The results of the bootstrapping analyses can be seen in Table 1. Reliable differences were found for each corpus; these were consistent even when looking at the entirety of MBC, STC and LinCe combined. For the MBC and LinCe corpora, the mean proportion of Spanish-English CS was no-

Bootstrap estimation	Corpus	ES-EN Proportion	EN-ES Proportion
CS tokens exclusively	MBC	0.55 [0.54, 0.57]	0.44 [0.42, 0.46]
	LinCE	0.55 [0.54, 0.57]	0.44 [0.42, 0.45]
	Texas Spanish	0.47 [0.44, 0.48]	0.52 [0.51, 0.53]
Entire corpus	MBC	0.00980 [0.00939, 0.0120]	0.00784 [0.00747, 0.00822]
	LinCE	0.0100427 [0.0096, 0.01045]	0.00797 [0.00763, 0.00834]
	Texas Spanish	0.0129 [0.0125, 0.0132]	0.0143 [0.0139, 0.0146]

Table 1: Proportions of Spanish-English and English-Spanish CSs, measured from a) exclusively the POS bigrams labeled as CSs, and b) all three corpora combined.

POS tag before CS	Coef.	*N*
INTJ	0.13	2,815
PROPN	0.08	1,305
NOUN	0.06	2,960
PART	0.04	60
ADJ	0.02	1,062
VERB	0.009	2,338
NUM	0.005	139
DET	0.004	1,843
ADP	0.0007	1,636
ADV	-0.0002	1,349
AUX	-0.003	822
CCONJ	-0.005	749
SCONJ	-0.01	643
PRON	-0.01	912

Table 2: Part of Speech (POS) tags of words *preceding* a CS with their coefficient estimates and counts, ordered from the most predictive to the least.

POS tag after CS	Coef.	*N*
INTJ	0.12	1,767
PROPN	0.06	1,370
PART	0.04	185
CCONJ	0.024	1,168
SCONJ	0.021	982
ADJ	0.016	1,364
NOUN	0.012	2,838
ADV	0.006	1,509
ADP	0.003	1,527
AUX	-5.20E-07	753
PRON	-0.0005	2,561
DET	-0.001	906
VERB	-0.003	1,535
NUM	-0.01	168

Table 3: Part of Speech (POS) tags of words *following* a CS with their coefficient estimates and counts, ordered from the most predictive to the least.

tably higher, while the mean proportion of English-Spanish CS was higher in STC. STC's higher proportion of English-Spanish CS goes against the prediction that Spanish-English CS proportions would be higher than English-Spanish CS proportions. Moreover, it is unexpected given that previous literature (Carter et al., 2010; Couto et al., 2021) has demonstrated that Spanish tends to be the matrix language in Spanish-English CS.

4.2 Logistic regression analysis

Of 195 POS combinations detected across the three corpora, 150 POS combinations were statistically significant predictors of CS. Estimate values from the model were consulted to determine whether a given POS combination predicted that a CS would occur (POSitive estimate) or not (negative estimate). 60 POS combinations predicted the occurrence of a CS, while 90 combinations predicted that a CS would not occur. We calculated mean estimate values across all statistically significant POS combinations to derive the values found in Table 2 and Table 3.

As presented in Table 2, among all POS tag bigrams that indicate the occurrence of a CS, interjections (INTJ) and proper nouns (PROPN) were the most predictive of a CS occurring (please see Example 1 in Table 4 for an example of a CS where INTJ precedes the CS). However, when compared to all bigrams predicting that a CS would not occur, the subordinating conjunction (SCONJ) and pronoun (PRON) tags were the least effective at predicting a CS would occur. DET, which was predicted to be the first word in a CS pair due to the prevalence of the DET-N CS (e.g., Valdés Kroff et al. (2018); Balam et al. (2020)), had the second-smallest estimate value of the tags that were predictive of CS. See Example 2 (Table 4) for an example of a DET-N CS.

Similarly, the estimate values provided in Table 3 indicate that INTJ, PROPN, and PART tags being the second tag in a bigram effectively diagnose that a CS has occurred. See Example 3 in Table 4 for an example of a CS containing a PART tag. The auxiliary (AUX), PRON, and DET tags (among others) effectively determine that a CS does *not* occur in the same context. Interestingly, we found that SCONJ and CCONJ happen regularly as the second, but not the first, word in a CS bigram, contrasting with Soto et al. (2018), who found that both tags occur at significantly higher rates as the first word in a bigram containing a CS. Examples

Number	Corpus	Tag preceding CS with Lang ID	Tag following CS with Lang ID	Example
1	LinCe	INTJ (EN)	AUX (ES)	**Like**, hay que vacilar (*like, you have to stagger*)
2	STC	DET (ES)	NOUN (EN)	... para el **deadline** (*for the deadline*)
3	MBC	CCONJ (ES)	PART (EN)	**Pero** to test for lead (*but to test for lead*)
4	LinCe	ADJ (EN)	SCONJ (ES)	... how **funny que** I'm sitting there (*how funny that I'm sitting there*)

Table 4: Specific examples of code-switching contexts from the three different corpora.

of some of the pairs mentioned can be found in Table 4.

5 Discussion and Limitation

After analyzing three Spanish-English spoken and written corpora, we expanded previously known patterns of directionality and POS in CS. Bootstrapping showed a tendency in Spanish-English switches throughout the MBC and LinCE corpora, while the opposite was found in the STC. The logistic regression analysis concluded a robust number of POS combinations that predicted CS: tags like INTJ and PROPN were commonly present during switches, while PRON and SCONJ were not.

This paper contributes to the field by highlighting the complex role that POS plays in Spanish-English CS. Similarly to Soto et al. (2018), we found that a variety of tags predicts the event that precedes (INTJ, PROPN, NOUN and PART, among others) or follows (INTJ, PROPN, PART, CCONJ and SCONJ, among others) a given CS. Phrases like Example 1 in Table 4 illustrate how frequent INTJs are more characteristic of naturalistic language use than of elicited stimuli, which may explain why this POS had not been previously identified as statistically significant.

The fact that Soto et al. (2018) did not find predictive validity of the SCONJ and CCONJ tags following a CS but our experiment did, can be explained by a difference in the corpora analyzed. Soto et al. (2018) examined solely a portion of the MBC, while our work covered two additional corpora, besides the entirety of MBC; the CS behavior in Soto et al. (2018)'s analyses may be qualitatively and quantitatively different from the data used in this study. Moreover, a switch occurring before a CCONJ and SCONJ could be expected since these POS tags often introduce new clauses or sentence-like units, creating opportunities for a language switch (as seen in Example 4, Table 4).

To our knowledge, this is the first research to empirically establish a higher proportion of English-Spanish CS relative to Spanish-English CS with the written STC corpus. This finding may be motivated by pragmatic differences in CS across written and oral corpora. However, it could also be the case that speech community and corpus modality combine to affect distributional trends of Spanish-English and English-Spanish CS. Further investigation of this would require more written Spanish corpora, which is lacking at the moment; however, we hope that future work can mitigate this gap.

Despite the implications of this study, it is not without limitations. First, only three corpora were analyzed due to the lack of more publicly available data. We look forward to expanding our experiments to additional speech communities in the future. Second, we only examined Spanish-English CS, two languages considered relatively high-resource; we hope that our analyses can be extended to different language pairings, particularly those that involve understudied languages, to probe the generalizability of our findings.

One final limitation of our analyses is that we used POS tags and not dependency relations to investigate Spanish-English CS. Our decision, which is line with with previous CS research (i.e. Soto et al. (2018), Balam et al. (2020), inter alia), was motivated by the fact that POS tags are (albeit a somewhat simplistic) way to gain information about structural characteristics of Spanish-English CS. For example, our analyses provided novel insight into the predictability of INTJ, NOUN, and PART occurring before a CS and INTJ, PART, CCONJ, and SCONJ occurring after a CS. However, we recognize the value of including dependency relations in our analyzes, given that in German-English CS, two adjacent words have been reported to be more likely to be in the same language if they are more directly related in terms of dependency relations (Eppler, 2010). We expect that future analyses consider both POS tags and dependency relations to further enrich the field's understanding of Spanish-English CS.

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Osmer Balam, María del Carmen Parafita Couto, and Hans Stadthagen-González. 2020. Bilingual verbs in three Spanish/English code-switching communities. *International Journal of Bilingualism*, 24(5-6):952–967.

Jeffrey Blokzijl, Margaret Deuchar, and M Carmen Parafita Couto. 2017. Determiner asymmetry in mixed nominal constructions: The role of grammatical factors in data from Miami and Nicaragua. *Languages*, 2(4):20.

Barbara E. Bullock and Almeida Jacqueline Toribio. 2024. Spanish in Texas Corpus.

Diana Carter, Peredur Davies, Ma Carmen Parafita Couto, and Margaret Deuchar. 2010. A corpus-based analysis of codeswitching patterns in bilingual communities. *Revista Española de Lingüística*.

Brian Hok Shing Chan. 2009. *Code-switching between typologically distinct languages*. Cambridge University Press.

Jie Chi and Peter Bell. 2024. Analyzing the role of part-of-speech in code-switching: A corpus-based study. In *Findings of the Association for Computational Linguistics*, page 1712–1721. Association for Computational Linguistics. The 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 ; Conference date: 17-03-2024 Through 22-03-2024.

M Carmen Parafita Couto, Miriam Greidanus Romeli, and Kate Bellamy. 2021. Code-switching at the interface between language, culture, and cognition. *Lapurdum*.

David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232.

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

Margaret Deuchar, Maria del Carmen Parafita Couto, Peredur Davies, Kevin Donnelly, Fraibet Aveledo, Diana Carter, Marika Fusser, Jon Herring, Lowri Jones, Siân Lloyd-Williams, Myfyr Prys, Elen Robert, and Jonathan Stammers. 2009. Bangortalk Corpora. Accessed April 21, 2025.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. Chapman and Hall/CRC.

Nadine Eichler, Malin Hager, and Natascha Müller. 2012. Code-switching within determiner phrases in bilingual children: French, Italian, Spanish and German. *Zeitschrift für französische Sprache und Literatur*, 122(3):227–258.

Eva Eppler. 2010. *Emigranto:: The syntax of German-English code-switching*. Wilhelm Braumüller Universitäts-Verlagsbuchhandlung.

Dana-Maria Iliescu, Rasmus Grand, Rob van der Goot, and Sara Qirko. 2021. Much gracias: Semi-supervised code-switch detection for Spanish-English: How far can we get? In *Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, page 65. Association for Computational Linguistics.

Gerrit Jan Kootstra, Ton Dijkstra, and Janet G. van Hell. 2020. Interactive alignment and lexical triggering of code-switching in bilingual dialogue. *Frontiers in Psychology*, Volume 11 - 2020.

Carol Myers-Scotton. 1993. *Social motivations for codeswitching: Evidence from Africa*. Oxford University Press.

Shana Poplack. 1980. Sometimes I'll start a sentence in Spanish y termino en español: Toward a typology of code-switching. *Linguistics*, 18(7-8):581–618.

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*, pages 101–108, Online. Association for Computational Linguistics.

Ella Rabinovich, Masih Sultani, and Suzanne Stevenson. 2019. CodeSwitch-Reddit: Exploration of written multilingual discourse in online discussion forums. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, page 446, Hong Kong, China. Association for Computational Linguistics.

Victor Soto, Nishmar Cestero, and Julia Hirschberg. 2018. The role of cognate words, POS tags and entrainment in code-switching. In *Interspeech*, pages 1938–1942.

Victor Soto and Julia Hirschberg. 2017. Crowdsourcing universal Part-of-Speech tags for code-switching. *CoRR*, abs/1703.08537.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Almeida Jacqueline Toribio. 2023. 'Doing' Romance Linguistics. *Isogloss. Open Journal of Romance Linguistics*, 9(2):1–14.

Chara Tsoukala, Stefan L Frank, APJ van den Bosch, J Valdés Kroff, and Mirjam Broersma. 2019. Simulating Spanish-English code-switching: El modelo está generating code-switches. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics.

Jorge R Valdés Kroff, Rosa E Guzzardo Tamargo, and Paola E Dussias. 2018. Experimental contributions of eye-tracking to the understanding of comprehension processes while hearing and reading code-switches. *Linguistic Approaches to Bilingualism*, 8(1):98–133.

Janet G Van Hell, Kaitlyn Litcofsky, and Caitlin Y Ting. 2015. *Intra-sentential code-switching: Cognitive and neural approaches*. Cambridge University Press.

Genta Indra Winata, Alham Fikri Aji, Zheng-Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. *Preprint*, arXiv:2212.09660.

On the Flatness, Non-linearity, and Branching Direction of Natural Language and Random Constituency Trees: Analyzing Structural Variation within and across Languages

Taiga Ishii and Yusuke Miyao
The University of Tokyo
{taigarana,yusuke}@is.s.u-tokyo.ac.jp

Abstract

Natural languages exhibit remarkable diversity in their syntactic structures. Previous research has investigated the cross-lingual differences in local structural features such as word order or dependency relations. However, considering structural variation within individual language, it remains unclear how such features influence the variation in the overall constituency tree structure and hence the structural variation across languages. To this end, we focus on the shape of constituency trees, analyzing the cross-lingual overlap in the distributions of flatness, non-linearity, and branching direction. While acknowledging that the findings may be influenced by the potential annotation idiosyncrasies across treebanks, the experiments quantitatively suggest that flatness and branching direction vary significantly across languages. As for non-linearity, the cross-lingual difference was relatively small, and the distributions tend to skew towards linear structures. Furthermore, comparison with randomly generated trees suggests that while phrase category and frequency information is crucial for reproducing the branching direction found in natural languages, non-linearity can be replicated reasonably well even without such information.

1 Introduction

Uncovering the universals and variations in syntactic structures across natural languages is a central challenge in computational linguistics and natural language processing. In the context of linguistic typology, differences and universal properties among languages have been extensively discussed from perspectives such as word order (Dryer, 1992; Östling, 2015; Baylor et al., 2024; Alves et al., 2023), dependency relations (Blache et al., 2016; Chen and Gerdes, 2017), morphology (Cotterell et al., 2019; Bentz et al., 2016; Bjerva and Augenstein, 2018), and phonology (Cotterell and Eisner, 2017; Bjerva and Augenstein, 2018).

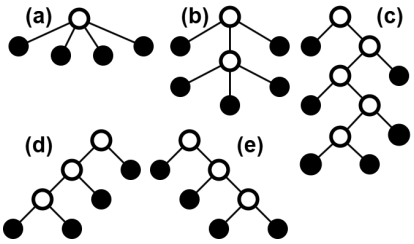


Figure 1: Example constituency trees: (a) fully flat, (b, c) fully non-linear, (d, e) fully left/right-branching

However, capturing structural variation within language remains a challenge for existing typological studies (Ponti et al., 2019). When considering within-language variations, it is not trivial how local structural features, such as word order and dependency relations, influence the variation in the overall structure of constituency trees and, consequently, relate to cross-lingual structural differences and universals.

To address this issue, we propose an approach based on the shape of constituency trees. Specifically, we quantify three features characterizing tree shape: flatness, non-linearity, and branching direction (Figure 1).[1] We then analyze the distribution of these shape features within each language and their distributional overlap across languages, using treebanks from diverse languages.

A key advantage of our approach, using tree shape features, is its potential to enable comparisons beyond natural languages. While traditional linguistic typology relies mainly on features derived from linguistic theories and often limits its scope to comparisons among natural languages (Dryer and Haspelmath, 2013), our approach allows us to investigate the statistical properties of natural language constituency trees within the broader space of all theoretically possible tree structures.

[1] Examples of English constituency trees are shown in Appendix A.

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 90–104

Tree Shape	Measure	Intuitive Description	Range
Flatness	AR	average number of internal nodes per leaf	$[0, 1]$
Non-linearity	NCE	normalized depth of max center embedding	$[0, 1]$
Branching Direction	CC, EWC, RJ	left-right diff of number of leaves at each node	$[-1, 1]$

Table 1: Overview of the tree shape measures used in this study.

Furthermore, in addition to comparing tree shape distributions across natural languages, we conduct comparisons with randomly generated tree structures. This comparative analysis aims to shed light on the fundamental question of what information (e.g., grammatical category information) is essential for characterizing the structure of natural language trees.

Our analyses are conducted on constituency treebanks from 11 diverse languages. Experimental results quantitatively show that while flatness and branching direction exhibit significant cross-lingual variation with minimal distributional overlap, the distributions of non-linearity are skewed towards the linear region, resulting in a certain degree of overlap across languages. Moreover, comparisons with random trees suggest that category information is crucial for reproducing branching direction, whereas non-linearity can be relatively well replicated even without such information. However, a key limitation is that the treebanks used in the experiments are not harmonized, meaning that the findings may be influenced by differences in the annotation schemes. Disentangling the genuine linguistic differences from potential annotation artifacts is a crucial direction for future work. The implementations of the experiments are available at `https://github.com/mynlp/tree-shape-distribution.git`.

2 Background

Cross-lingual analysis based on the shape features of dependency trees has been actively conducted. For example, (directed) dependency distance is used to investigate typological differences (Chen and Gerdes, 2017; Yadav et al., 2020) and universal tendencies like dependency distance minimization (Futrell et al., 2015; Yu et al., 2019). Previous research has also examined structural properties like clause/center embedding depth across languages (Blasi et al., 2019; Noji and Miyao, 2014) and statistical patterns such as Menzerath's law (Mačutek et al., 2017, 2021; Berdičevskis, 2021), sometimes including comparisons with random trees (Tanaka-Ishii, 2021).

In contrast, cross-lingual analysis based on the shape of constituency trees is relatively sparse compared to those on dependency trees. For example, while there are studies on the relationship between the center embedding depth and human reading time in English (van Schijndel et al., 2015), and a comparison of branching direction in English and Chinese (Zhang et al., 2022), these studies are often limited to a small number of languages. Tanaka-Ishii and Tanaka (2023) conducted an extensive analysis on various natural languages and random trees, but their work is limited to the Strahler number (Strahler, 1957), which measures the lower bounds on memory requirements for processing constituency trees.

This study aims to conduct a systematic comparative analysis specifically for constituency trees, using 3 tree shape features —flatness, non-linearity, and branching direction— across a diverse set of languages.

3 Tree Shape Measures

This section defines the tree shape measures used in this study. Following Chan et al. (2010), we analyze the tree shape of delexicalized constituency trees, where leaves are POS tags and internal nodes represent phrases. While some parsing research assumes only binary trees (Liang et al., 2007; Kim et al., 2019), we consider general n-ary tree structures after removing unary nodes and ignore phrase category labels.

We specifically examine three features: flatness, non-linearity, and branching direction (Table 1). As our interest lies in the overall shape rather than the absolute size of trees, the shape measures are normalized to mitigate the effect of tree size, i.e., the number of leaf nodes, corresponding to sentence length.

3.1 Flatness

Flatness can be interpreted as the degree of nesting within a tree. To quantify flatness, we use the "aspect ratio", a measure adapted from the one

proposed by Chan et al. (2010) as a feature for unsupervised parsing.

For a given unlabeled tree t, the aspect ratio $\mathrm{AR}(t)$ roughly corresponds to the number of internal nodes divided by the number of leaves, and is formally calculated as:

$$\mathrm{AR}(t) \equiv \begin{cases} \frac{|V(t)|-1}{|t|_L-2}, & \text{if } |t|_L > 2 \\ 0, & \text{otherwise} \end{cases}$$

where $|t|_L$ denotes the number of leaves in t, and $|V(t)|$ denotes the number of internal nodes.

This measure is designed such that it equals 0 for a fully flat tree (Figure 1 (a)) and 1 for a fully binary branching tree (Figure 1 (c, d, e)), regardless of the number of leaf nodes $|t|_L$.[2] Note that the original aspect ratio proposed by Chan et al. (2010) was simply $\frac{|V(t)|}{|t|_L}$. Our definition modifies this by subtracting offsets for normalization.

3.2 Non-linearity

Non-linearity is a key concept for discussing whether natural languages are more complex than regular languages (Chomsky, 1956), often captured by center embedding structures in trees. Center embedding also has drawn attention from a cognitive perspective, particularly concerning processing memory load (van Schijndel et al., 2015).

Prior work measured center embedding via the maximum stack depth required by a left-corner parser (van Schijndel et al., 2015; Noji and Miyao, 2014). However, this metric is problematic for purely capturing tree shape, as its inherent left-right asymmetry yields different values for flipped tree structures (Noji, 2016). Furthermore, the Strahler number, employed by Tanaka-Ishii and Tanaka (2023) to measure memory requirement lower bounds, is also not suitable for quantifying non-linearity, because it cannot distinguish between fully center-embedding binary trees (Figure 1 (c)) and fully left/right-branching ones (Figure 1 (d, e)).

We introduce a left-right symmetric center embedding depth measure, calculated via Algorithm 1. Roughly, for a node v, its center embedding depth $\mathrm{CenterEmb}_t(v)$ counts ancestors that are neither the left-most nor right-most child of their respective parent. The overall center embedding depth

Algorithm 1 Function for calculating the center embedding depth of a given node v in tree t.

function $\mathrm{CenterEmb}_t(v)$
  $c \leftarrow 0, \quad nl \leftarrow \mathrm{False}, \quad nr \leftarrow \mathrm{False}$
  while v is not root of t **do**
    if v is not the left-most child in t **then**
      $nl \leftarrow \mathrm{True}$
    if v is not the right-most child in t **then**
      $nr \leftarrow \mathrm{True}$
    if $nl \wedge nr$ **then** ▷ Current node is center embedded
      $c \leftarrow c + 1$
      $nl \leftarrow \mathrm{False}, \quad nr \leftarrow \mathrm{False}$
    $v \leftarrow$ parent of v
  return c

$\mathrm{CE}(t)$ of a tree t is the maximum $\mathrm{CenterEmb}_t$ value among the parents of leaf nodes:[3]

$$\mathrm{CE}(t) \equiv \max_{v\text{: leaf of } t} \mathrm{CenterEmb}_t(\text{parent of } v)$$

To normalize for tree size and capture tree shape purely, we define $\mathrm{NCE}(t)$ as $\mathrm{CE}(t)$ divided by the maximum possible CE value for a tree with $|t|_L$ leaves:

$$\mathrm{NCE}(t) \equiv \begin{cases} \frac{\mathrm{CE}(t)}{\lceil \frac{|t|_L-3}{2} \rceil}, & \text{if } |t|_L > 3 \\ 0, & \text{otherwise} \end{cases}$$

The denominator $\lceil \frac{|t|_L-3}{2} \rceil$ represents the maximum value achieved by fully center-embedding trees (Figure 1 (b, c)). Thus, $\mathrm{NCE}(t)$ approaches 0 for both linear (Figure 1 (d, e)) and flat (Figure 1 (a)) structures, while it approaches 1 for fully center-embedding ones (Figure 1 (b, c)).

3.3 Branching Direction

While local structural features such as word order and dependency direction differ across languages (Dryer and Haspelmath, 2013; Chen and Gerdes, 2017), it is not obvious how these local orderings affect the overall shape, particularly the directional bias, of constituency trees. To this end, we employ three branching direction measures proposed by Ishii and Miyao (2023). These measures are extensions of tree balance indices used in phylogenetics (Heard, 1992; Mooers and Heard, 1997;

[2] Since unary nodes are removed, the range of AR for the trees we analyze is $[0, 1]$. If there were unary nodes, the value could be larger than 1.

[3] Using the parent of the leaf node rather than the leaf node itself is intended to reflect phrase-level embedding, as opposed to word-level embedding. Also, as this is a purely tree shape measure, it does not consider grammatical or semantic constraints, unlike (Wilcox et al., 2019).

Rogers, 1996), adapted to capture left-right asymmetry. They are primarily calculated based on the difference in the number of leaves between the left and right subtrees at each internal node.

To calculate such difference in the number of leaves $h_t(v)$ for a node v in a non-binary tree t, the i-th child of v from the left is weighted by a position-dependent weight $w_v(i)$:

$$h_t(v) \equiv \sum_{i=0}^{|v|_C^t-1} w_v(i) \cdot |t_{v_i}|_L$$

Here, $|v|_C^t$ is the number of child nodes of v, and t_{v_i} denotes the subtree rooted at v_i. The weight $w_v(i)$ is defined as:

$$w_v(i) \equiv \begin{cases} g(i - \frac{|v|_C^t-1}{2}) \cdot \frac{1}{\lfloor |v|_C^t/2 \rfloor}, & \text{if } |v|_C^t > 1 \\ 0, & \text{otherwise} \end{cases}$$

where $g(x) \equiv \text{sign}(x) \cdot \lceil |x| \rceil$ is rounding toward infinity. The weight is symmetric, being close to 0 for central children and -1 or 1 for the outermost children. For example, if $|v|_C^t = 4$, the weights for the children from left to right are $-1, -\frac{1}{2}, \frac{1}{2}, 1$.

The three measures aggregate $h_t(v)$ differently across the tree. The range of all measures is $[-1, 1]$, where values closer to -1 indicate a tree closer to a fully left-branching tree (Figure 1 (d)), and values closer to 1 indicate a tree closer to a fully right-branching tree (Figure 1 (e)).

First, the corrected Colles index (CC) is calculated as:

$$\text{CC}(t) \equiv \frac{2}{(|t|_L - 1)(|t|_L - 2)} \cdot \sum_{v \in V(t)} h_t(v)$$

CC tends to give more weight to the branching bias (h_t) at internal nodes closer to the root. In contrast, the equal weights Colles index (EWC) normalizes the h_t at each internal node by the size of its subtree, aiming to evaluate the contribution of each node to the overall branching direction more evenly. It is calculated as:

$$\text{EWC}(t) \equiv \frac{1}{|t|_L - 2} \cdot \sum_{v \in V(t): |t_v|_L > 2} \frac{h_t(v)}{|t_v|_L - 2}$$

Finally, the Roger's J index (RJ) aggregates branching bias using only the sign of $h_t(v)$, thus evaluating the branching bias at a coarser granularity than EWC:

$$\text{RJ}(t) \equiv \frac{1}{|t|_L - 2} \cdot \sum_{v \in V(t)} \text{sign}(h_t(v))$$

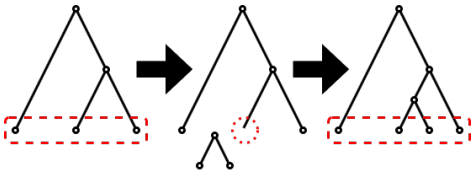


Figure 2: Example of leaf replacement in Yule model.

Algorithm 2 Parameterized Yule model to sample a single non-labeled tree.

Input: w_{len} ▷ Counts of sentence lengths
Input: $w^1_{\text{arity}}, \ldots$ ▷ Counts of arities
Input: $w^1_{\text{pos}}, \ldots$ ▷ Counts of replaced leaf indices
$t \leftarrow \text{SampleCherry}(w^1_{\text{arity}})$ ▷ Initialization
$n_{\text{lim}} \leftarrow \text{Sample}(w_{\text{len}})$ ▷ Sample length limit
while $|t|_L < n_{\text{lim}}$ **do**
  if $\text{sum}(w^{|t|_L}_{\text{pos}}) = 0 \vee \text{sum}(w^{|t|_L}_{\text{arity}}) = 0$ **then**
    Restart from initialization due to lack of statistical infomation
  else
    $i \leftarrow \text{Sample}(w^{|t|_L}_{\text{pos}})$ ▷ Sample leaf index for replacement
    $c \leftarrow \text{SampleCherry}(w^{|t|_L}_{\text{arity}})$ ▷ Sample n-ary cherry
    Replace i-th leaf with cherry c
return t

4 Generating Random Trees

To investigate what statistical information is essential for characterizing the structure of natural language trees, we perform experiments with randomly generated trees. Our methodology is to first extract different levels of statistical information from a given treebank, and then use this information to parameterize random tree models. By comparing the tree shape distributions of the generated trees with those of the original treebank, we can assess the importance of the specific statistical information used by the model.

For this purpose, we employ 6 random tree models based on 2 approaches: the Yule model and Probabilistic Context-Free Grammar (PCFG). Both approaches are hierarchical processes, but they differ primarily in that PCFG utilizes phrase category information, while the Yule model does not.

4.1 Yule Model

The Yule model (Harding, 1971; Yule, 1925; Fischer et al., 2023) is a basic model for generating unlabeled trees by starting from a single leaf node, iteratively replacing a uniformly randomly selected

leaf with a cherry until reaching a target tree size (Figure 2). Typically, a cherry refers to a single internal node tree with two leaves; however, in this study, we consider a general n-ary cherry.

To better capture natural language properties, we parameterize the Yule model using three types of empirical statistics extracted from treebanks: (1) target tree size (stopping criterion) (2) node arity, and (3) leaf replacement position.

Statistics (2) and (3) are estimated as conditional distributions w^k_{arity}, w^k_{pos} dependent on the number of candidate leaves k for the replacement at each step. The process of a Yule model parameterized with these statistics is shown in Algorithm 2. w^k_{arity} and w^k_{pos} are calculated via an inverse Yule process that collapses cherries back into single leaf nodes (Algorithm 3). Since the inverse Yule process for a given tree is generally not unique, we apply the process N times to the treebank.[4]

We compare 4 variants. Yule+arity+pos uses $w^k_{\text{arity}}, w^k_{\text{pos}}$ calculated by Algorithm 3. Yule+arity uses conditional w^k_{arity} with uniform replacement, i.e., $\forall k.w^k_{\text{pos}} = [1, \dots]$. Yule+pos uses conditional w^k_{pos} with corpus-level empirical arity distribution, i.e., $\forall k.w^k_{\text{arity}} = \hat{w}_{\text{arity}}$. Yule uses neither, employing uniform replacement and $\hat{w}_{\text{arity}}$.

4.2 Probabilistic Context-free Grammar

To analyze the role of phrase category information, we also generate random trees using Probabilistic Context-Free Grammars (PCFGs), a standard formalism in parsing (Charniak, 1996; Johnson et al., 2007; Liang et al., 2007). We employ a PCFG with rule probabilities estimated by counts of production rules in the treebank. Since controlling tree size during PCFG generation is non-trivial, we use breadth-first generation, restarting sampling if the number of bottom-most nodes exceeds the maximum tree size, i.e., number of leaves, observed in the original treebank.[5] To isolate the effect of rule frequency information inherent in the PCFG, we additionally evaluate a uniform PCFG (UPCFG) where all production rules for a given nonterminal have uniform probability.

[4]Generation is retried if the empirical distribution for k is unavailable.

[5]Breadth-first generation avoids potential traversal order biases caused by size-based cancellation in depth-first generation.

Algorithm 3 Inverse Yule process to obtain conditional empirical distribution for node arity and leaf replacement positions.

Input: $\mathcal{T}$ ▷ List of trees
Input: N ▷ Number of iteration over given trees
for $n = 1, \dots$ **do** ▷ Initialize the counts of replaced leaf indices and arities when there are n leaves
$w^n_{\text{pos}} \leftarrow [0, \dots], \quad w^n_{\text{arity}} \leftarrow [0, \dots]$
for N times **do**
for $t \in \mathcal{T}$ **do**
$t' \leftarrow t$ ▷ Just copy
while t' is not an n-ary cherry **do**
$l_{\text{cherry}} \leftarrow$ list of root nodes of n-ary cherries in t'
$v \leftarrow \text{UniformSample}(l_{\text{cherry}})$
$a \leftarrow |v|^{t'}_C$
Replace subtree ${t'}_v$ with dummy leaf
$i \leftarrow$ leaf index of replaced dummy leaf
$w^{|t'|_L}_{\text{arity}}[a] \leftarrow w^{|t'|_L}_{\text{arity}}[a] + 1$
$w^{|t'|_L}_{\text{pos}}[i] \leftarrow w^{|t'|_L}_{\text{pos}}[i] + 1$
$w^1_{\text{arity}}[|t'|_L] \leftarrow w^1_{\text{arity}}[|t'|_L] + 1$ ▷ Count the arity of root
return $w^1_{\text{pos}}, \dots, w^1_{\text{arity}}, \dots$

5 Experiments and Discussion

Datasets. In this study, we use treebanks from 11 languages: English (Penn Treebank (Marcus et al., 1993)), Chinese (Chinese Treebank (Palmer et al., 2005)), Japanese (NPCMJ), French, German, Korean, Basque, Hebrew, Hungarian, Polish, and Swedish (SPMRL (Seddah et al., 2013)).[6] It should be noted that these treebanks are not harmonized and thus annotation schemes are not identical. Following Chan et al. (2010), we focus on delexicalized constituency trees. For preprocessing, we apply the following steps to the annotated tree structures in each treebank: (1) remove null elements, (2) strip functional tags from category labels, (3) remove word tokens and treat POS tags as the new leaf nodes, and finally (4) remove unary nonterminals by concatenating their category labels, similar to (Gómez-Rodríguez and Vilares, 2018).[7] Note that we do not remove punctuation.

Furthermore, to analyze the distributions of over-

[6]NPCMJ: NINJAL Parsed Corpus of Modern Japanese (http://NPCMJ.ninjal.ac.jp/).

[7]Further details are described in Appendix B.

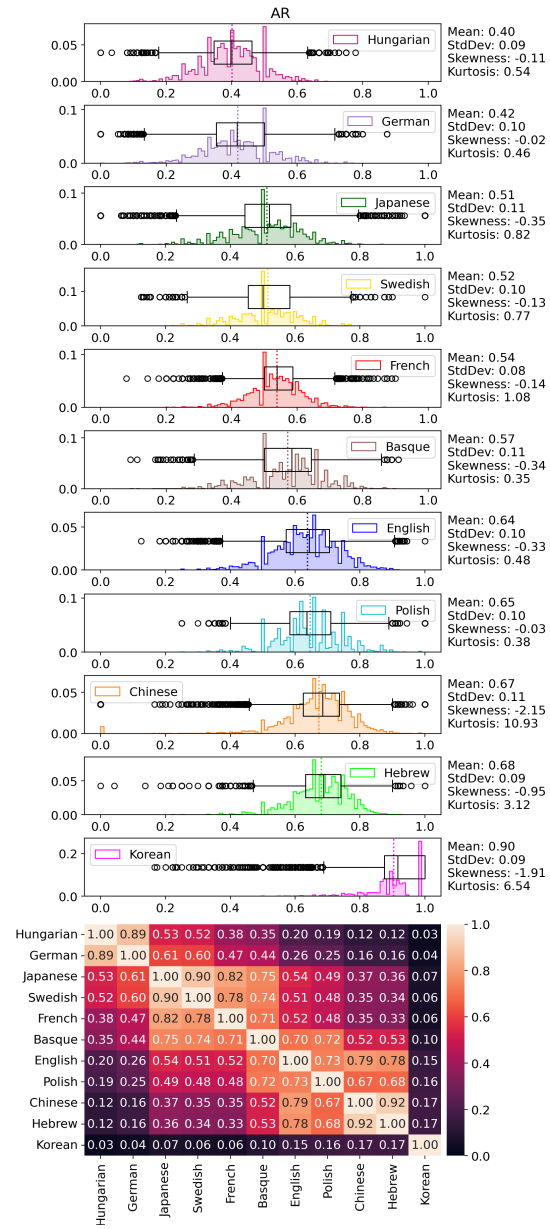


Figure 3: Distributions of AR for 11 natural language treebanks. The values in the heatmap are pairwise HIs.

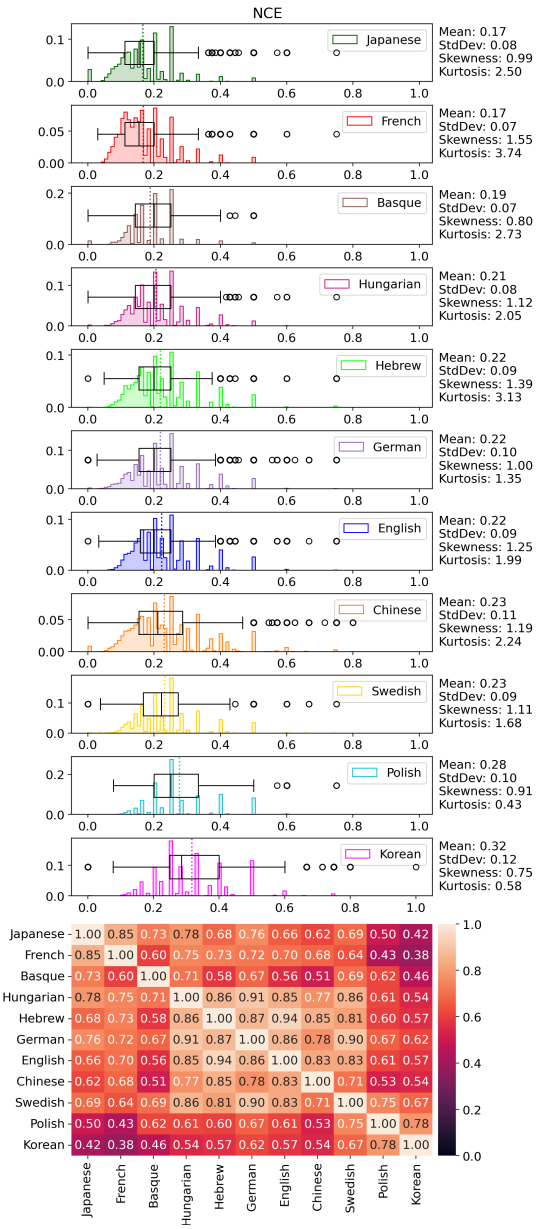


Figure 4: Distributions of NCE for 11 natural language treebanks. The values in the heatmap are pairwise HIs.

all tree shape, we use only sentences with lengths of 10 or more in the experiment. The tree shape distributions for random tree models are also calculated from this subset. This is because shorter sentences have a limited number of possible tree structures, which makes it difficult to analyze cross-lingual differences. Additionally, this length-based filtering provides a simple way to exclude typically short non-sentential fragments, e.g., "(FRAG (PU （) (VV 完) (PU ）))".[8] Table 2 shows the statistics of the preprocessed treebanks, including the number of data points and the mean number of leaf nodes. Note that the values are calculated after applying the length-based filtering. For each model and original dataset, we generate 10000 random trees for analysis.

Evaluation. For all evaluation metrics, we compute distributions as normalized histograms with 100 bins. To quantify cross-lingual differences in tree shape distributions, we use Histogram Intersection (HI). HI measures the proportion of overlap among a set of distributions, yielding a score in $[0, 1]$, where 0 indicates no overlap and 1 indicates identical distributions. This direct measure of overlap makes the score highly intuitive to inter-

[8]The example is from CTB and translates to "(finish)" in English.

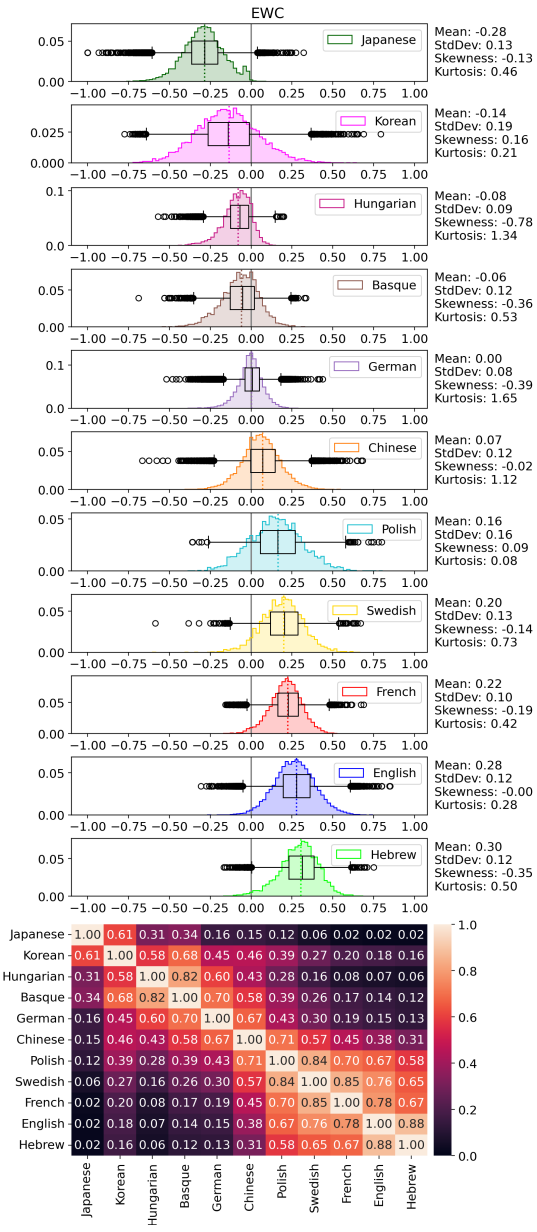


Figure 5: Distributions of EWC for 11 natural language treebanks. The values in the heatmap are pairwise HIs.

pret. Furthermore, unlike pairwise metrics such as Kullback-Leibler divergence, HI has the advantage of being applicable to multiple distributions simultaneously, which allows us to compute a single overall score across all languages.

While HI quantifies the overall overlap, to further understand the characteristics of each distribution, we also analyze its shape using standard deviation, skewness, and kurtosis. Skewness measures how a distribution is biased towards left or right; positive/negative skewness implies large part of the distribution is on the left/right-side. Kurtosis describes the sharpness of a peak of distribution and the weight of its tails compared to a normal distribution, which is defined to have kurtosis of 0; a positive value indicates a more pointed peak and heavier tails.[9]

	#Data	#Leaves
English	35.0K	25.4 ± 10.3
Chinese	14.1K	32.6 ± 18.0
Japanese	47.0K	26.8 ± 18.1
French	13.4K	32.4 ± 16.7
German	30.5K	21.8 ± 9.9
Korean	16.4K	15.3 ± 4.0
Basque	4.9K	15.8 ± 5.1
Hebrew	4.5K	27.6 ± 13.9
Hungarian	6.8K	23.6 ± 11.1
Polish	2.9K	14.4 ± 4.4
Swedish	3.5K	19.0 ± 8.4

Table 2: Statistics of the preprocessed treebanks: the number of data points and the mean ± standard deviation of the number of leaf nodes. The statistics are calculated after applying length-based filtering with a threshold 10.

	AR	NCE	CC	EWC	RJ
HI	0.03	0.33	0.03	0.02	0.02

Table 3: HI across 11 natural language treebanks.

5.1 How Are Natural Language Trees Different?

Table 3 shows the HI across all languages for each tree shape measure. Figure 3, Figure 4, and Figure 5 show the distributions for AR, NCE, EWC in each language, together with box plots within 1.5 IQR and heatmaps of the pairwise HI between languages.[10] Note that languages are sorted by the mean for each measure.

Flatness. As shown in Table 3, the HI is only 3%, indicating that flatness varies considerably across languages. However, languages differ not only in their average flatness but also in the shape of their flatness distributions. While some languages like German and Polish exhibit skewness near 0, others such as Chinese and Korean show values close to −2. Similarly, for kurtosis, Basque has a value around 0.3, whereas Chinese and Korean have much larger values, approximately 6 and 10, respectively.

We speculate two potential factors for the cross-

[9]The presence of outliers can also lead to high kurtosis.
[10]Results for CC and RJ are provided in Appendix C.

	AR	NCE	CC	EWC	RJ
Yule	0.84± 0.01	0.88± 0.02	0.64± 0.06	0.51± 0.08	0.52± 0.08
Yule+arity	0.81± 0.02	0.88± 0.03	0.61± 0.08	0.49± 0.08	0.51± 0.08
Yule+pos	0.84± 0.01	**0.90**± 0.01	0.82± 0.01	0.70± 0.05	0.72± 0.05
Yule+arity+pos	0.81± 0.02	**0.90**± 0.01	0.84± 0.03	0.65± 0.05	0.70± 0.05
UPCFG	0.27± 0.04	0.74± 0.02	0.74± 0.03	0.53± 0.06	0.58± 0.05
PCFG	**0.91**± 0.01	0.88± 0.01	**0.90**± 0.01	**0.92**± 0.01	**0.90**± 0.01

Table 4: Average and standard error of HI between each random model and its original treebank.

lingual differences in flatness. First, differences in annotation schemes across treebanks may play a role. For example, the Japanese and Hungarian treebanks used in this study do not have annotations for VPs (verb phrases) as in PTB due to the relatively free word order (Csendes et al., 2005), and phrases like PP (prepositional phrase) are annotated flatly in German treebank (Brants et al., 2004). Furthermore, we hypothesize that distinctly high AR, i.e., lower flatness, of Korean is due to its tokenization, where multiple word tokens, e.g., compound nouns, are often agglutinated into a single token (Seddah et al., 2013), reducing the number of leaves per nonterminal. This implies that, for any language, the shape of constituency trees calculated based on the number of leaves, can vary depending on the granularity of tokenization, i.e., definition of phrase size.

Non-linearity. From Table 3, we can observe that non-linearity NCE has a 33% overlap across all languages, indicating higher cross-lingual commonality compared to flatness or branching direction. Indeed, the heatmap in Figure 4 shows that pairwise HI values for NCE are generally higher than those for flatness (Figure 3) and branching direction (Figure 5).

Moreover, even Korean, which has the highest average NCE, only reaches 0.32. This suggests that natural language trees are generally quite linear among all possible trees. The skewness ranges from 0.75 to 1.55 across all languages, consistently showing relatively large positive values. This indicates that the distributions are skewed to the left, i.e., towards the more linear region.

Branching Direction. As shown in Table 3, for all branching direction measures CC, EWC, and RJ, the cross-lingual HI is very small, only 2-3%, highlighting significant variation across languages. Furthermore, Figure 5, displaying the distributions and heatmap for EWC, reveals considerable variation in branching direction even within individual languages.

For instance, based on the means, languages such as Japanese, Korean, Hungarian, and Basque tend to be left-branching. However, within each of these languages, right-branching structures are also observed. Similarly, languages such as Hebrew, English, French, Swedish, Polish, and Chinese are right-branching on average, yet they also exhibit left-branching structures internally. It is also interesting to note that the left/right-branching language group based on the mean EWC is mostly the same with that based on the sign of directional dependency distance (Chen and Gerdes, 2017) except Chinese.[11] Conversely, skewness values are close to 0 for all languages, suggesting that the distributions tend to be relatively symmetrical regardless of the language. These results suggest that even when structural variations within individual languages are taken into account, significant structural variation still emerges across languages.

5.2 How Do Random Trees Differ from Natural Language Trees?

Table 4 presents the HI between each random model and the original treebanks, averaged over languages. PCFG performs best overall, achieving nearly 90% overlap on most measures.

Yule models using non-uniform leaf replacement (Yule+pos, Yule+arity+pos) better model non-linearity and branching direction than other Yule variants. However, their lower overlap on EWC, RJ compared to CC suggests that positional information w^k_{pos} for leaf replacement becomes noisy for large k, impacting EWC, RJ that give equal weights to branches near leaves, unlike the root-focused CC. In contrast, PCFG shows consistent strength across CC, EWC, RJ, unlike UPCFG, highlighting the importance of category

[11] However, the order of language itself is not exactly the same as (Chen and Gerdes, 2017).

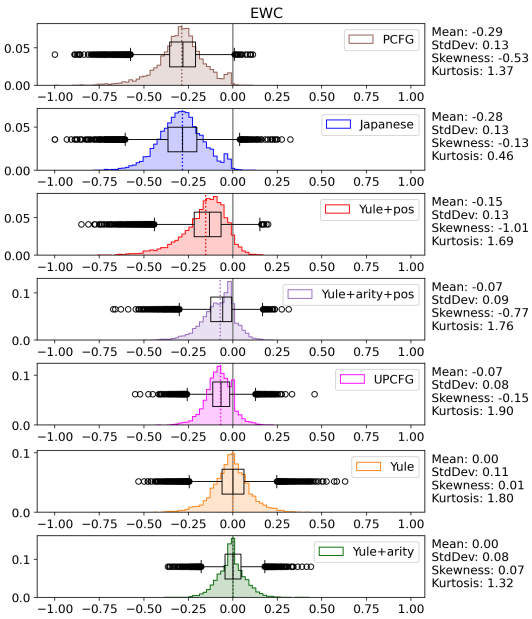


Figure 6: Distributions of EWC for Japanese treebank and its random models.

and frequency information for branching direction. Surprisingly, adding complex arity information (Yule+arity, Yule+arity+pos) degrades performance, suggesting that arity distributions can act as noise when conditioned on the number of leaves for replacement k for models that do not distinguish leaves from nonterminals. The distributions of EWC for Japanese (Figure 6) illustrate these differences; Yule and Yule+arity are mostly centered around 0 while that of the Japanese treebank is around -0.29; Yule+pos seems to capture the left-branching bias to a certain degree, but it is still skewed towards 0.0 compared to PCFG.

Notably, the basic Yule model —which only uses the corpus-level empirical arity distribution and assumes uniform leaf replacement positions— replicates non-linearity NCE well with 88% overlap, suggesting that non-linearity of natural language may be governed by a general mechanism beyond specific grammar or cognitive constraints.

6 Conclusion

We investigated structural variation of constituency trees within and across 11 languages, focusing on flatness, non-linearity, and branching direction. Analysis of the cross-lingual distributional overlap revealed that flatness and branching direction vary significantly across languages, indicating that cross-lingual differences emerge even when considering the structural variation within each language. Meanwhile, the distributions of non-linearity showed smaller cross-lingual difference and tend to skew towards linear trees.

Comparison with 6 random tree models based on the Yule model and PCFG showed that category information, accompanied by frequency statistics, are crucial for reproducing the branching direction patterns in natural language. In contrast, non-linearity was reasonably replicated even by relatively simple Yule models that lack such information, suggesting that non-linearity may be governed by more universal mechanisms independent of fine-grained grammatical details.

While this work focused on the shape of constituency trees, a key future direction is to analyze the joint distribution of overall tree shape features and local structural features, such as word order or dependency relations used in traditional linguistic typology. Such an analysis could lead to a deeper understanding of cross-lingual variations and universality in syntactic structures.

Limitations

As discussed in section 5, since this study analyzes annotated constituency trees, our experimental results can be influenced by the annotation scheme. First, while we included punctuations in the trees, they are sometimes removed in parsing (Li et al., 2020). Given that punctuation annotation methods can also differ across treebanks, investigating the impact of these annotation differences and the presence/absence of punctuation remains a task for future work. Second, as noted in section 5, the category labels annotated in the datasets used for our analysis are not consistent across all languages; for example, VP is not annotated in Japanese and Hungarian. Such difference in annotated phrase categories may also affect the analysis. Third, we did not apply any tokenization to the annotated constituency trees. However, as discussed in section 5, tokenization might affect tree shape.

Acknowledgements

This work was supported by JST SPRING Grant Number JPMJSP2108 and JSPS KAKENHI Grant Number JP24KJ0666 and JP24H00087.

References

Diego Alves, Božo Bekavac, Daniel Zeman, and Marko Tadić. 2023. Analysis of corpus-based word-order typological methods. In *Proceedings of the*

Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023), pages 36–46.

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2024. Multilingual gradient word-order typology from Universal Dependencies. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 42–49.

Christian Bentz, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardzic. 2016. A comparison between morphological complexity measures: Typological data vs. language corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 142–153.

Aleksandrs Berdičevskis. 2021. Successes and failures of Menzerath's law at the syntactic level. In *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pages 17–32.

Johannes Bjerva and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916.

Philippe Blache, Stéphane Rauzy, and Grégoire Montcheuil. 2016. MarsaGram: an excursion in the forests of parsing trees. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2336–2342.

Damian Blasi, Ryan Cotterell, Lawrence Wolf-Sonkin, Sabine Stoll, Balthasar Bickel, and Marco Baroni. 2019. On the distribution of deep clausal embeddings: A large cross-linguistic study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3938–3943, Florence, Italy. Association for Computational Linguistics.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Res. Lang. Comput.*, 2(4):597–620.

Samuel W K Chan, Lawrence Y L Cheung, and Mickey W C Chong. 2010. Tree topological features for unlexicalized parsing. In *Coling 2010: Posters*, pages 117–125, Beijing, China. Coling 2010 Organizing Committee.

Eugene Charniak. 1996. Tree-bank grammars. In *Proceedings of the AAAI Conference on Artificial Intelligence, 13*.

Xinying Chen and Kim Gerdes. 2017. Classifying languages by dependency structure Typologies of delexicalized Universal Dependency treebanks. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 54–63.

N Chomsky. 1956. Three models for the description of language. *IEEE Trans. Inform. Theory*, 2(3):113–124.

Ryan Cotterell and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1182–1192.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Trans. Assoc. Comput. Linguist.*, 7:327–342.

Dóra Csendes, János Csirik, Tibor Gyimóthy, and András Kocsor. 2005. The Szeged treebank. In *Text, Speech and Dialogue*, Lecture notes in computer science, pages 123–131. Springer Berlin Heidelberg, Berlin, Heidelberg.

Matthew S Dryer. 1992. The Greenbergian word order correlations. *Language (Baltim.)*, 68(1):81–138.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo.

Mareike Fischer, Lina Herbst, Sophie Kersting, Annemarie Luise Kühn, and Kristina Wicke. 2023. *Tree balance indices: A comprehensive survey*. Springer International Publishing, Cham.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proc. Natl. Acad. Sci. U. S. A.*, 112(33):10336–10341.

Carlos Gómez-Rodríguez and David Vilares. 2018. Constituent parsing as sequence labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1314–1324, Stroudsburg, PA, USA. Association for Computational Linguistics.

E. F. Harding. 1971. The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Applied Probability*, 3(1):44–77.

Stephen B Heard. 1992. Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. *Evolution*, 46(6):1818–1826.

Taiga Ishii and Yusuke Miyao. 2023. Tree-shape uncertainty for analyzing the inherent branching bias of unsupervised parsing models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 532–547, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mark Johnson, Thomas L Griffiths, and Sharon Goldwater. 2007. Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 139–146.

Yoon Kim, Chris Dyer, and Alexander Rush. 2019. Compound probabilistic context-free grammars for grammar induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2369–2385, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jun Li, Yifan Cao, Jiong Cai, Yong Jiang, and Kewei Tu. 2020. An empirical comparison of unsupervised constituency parsing methods. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3278–3283, Online. Association for Computational Linguistics.

Percy Liang, Slav Petrov, Michael Jordan, and Dan Klein. 2007. The infinite PCFG using hierarchical Dirichlet processes. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 688–697, Prague, Czech Republic. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*.

Ján Mačutek, Radek Čech, and Marine Courtin. 2021. The Menzerath-Altmann law in syntactic structure revisited. In *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pages 65–73.

Ján Mačutek, Radek Čech, and Jiří Milička. 2017. Menzerath-Altmann law in syntactic dependency structure. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 100–107.

Arne O Mooers and Stephen B Heard. 1997. Inferring evolutionary process from phylogenetic tree shape. *The Quarterly Review of Biology*, 72(1):31–54.

Hiroshi Noji. 2016. *Left-corner Methods for Syntactic Modeling with Universal Structural Constraints*. Ph.D. thesis, The Graduate University for Advanced Studies.

Hiroshi Noji and Yusuke Miyao. 2014. Left-corner transitions on dependency parsing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2140–2150, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Martha Palmer, Fu-Dong Chiou, Nianwen Xue, and Tsan-Kuang Lee. 2005. Chinese treebank 5.1 LDC2005T01U01.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Comput. Linguist.*, 45(3):559–601.

James S Rogers. 1996. Central moments and probability distributions of three measures of phylogenetic tree imbalance. *Systematic Biology*, 45(1):99–110.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, and 4 others. 2013. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182.

Arthur N Strahler. 1957. Quantitative analysis of watershed geomorphology. *Eos, Transactions American Geophysical Union*, 38(6):913–920.

Kumiko Tanaka-Ishii. 2021. Menzerath's law in the syntax of languages compared with random sentences. *Entropy*, 23(6):661.

Kumiko Tanaka-Ishii and Akira Tanaka. 2023. Strahler number of natural language sentences in comparison with random trees. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(12).

Marten van Schijndel, Brian Murphy, and William Schuler. 2015. Evidence of syntactic working memory usage in MEG data. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 79–88, Denver, Colorado. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2019. Hierarchical representation in neural language models: Suppression and recovery of expectations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 181–190, Florence, Italy. Association for Computational Linguistics.

Marcin Woliński. 2019. *Automatyczna analiza składnikowa języka polskiego*. Wydawnictwa Uniwersytetu Warszawskiego.

Himanshu Yadav, Ashwini Vaidya, Vishakha Shukla, and Samar Husain. 2020. Word order typology interacts with linguistic complexity: A cross-linguistic corpus study. *Cogn. Sci.*, 44(4):e12822.

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

G. U. Yule. 1925. II.—a mathematical theory of evolution, based on the conclusions of dr. J. C. Willis, F. R. S. *Philos. Trans. R. Soc. Lond.*, 213(402-410):21–87.

Xiaohan Zhang, Shaonan Wang, Nan Lin, and Chengqing Zong. 2022. Is the brain mechanism for hierarchical structure building universal across languages? an fMRI study of Chinese and English. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7852–7861, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Robert Östling. 2015. Word order typology through multilingual word alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 205–211, Stroudsburg, PA, USA. Association for Computational Linguistics.

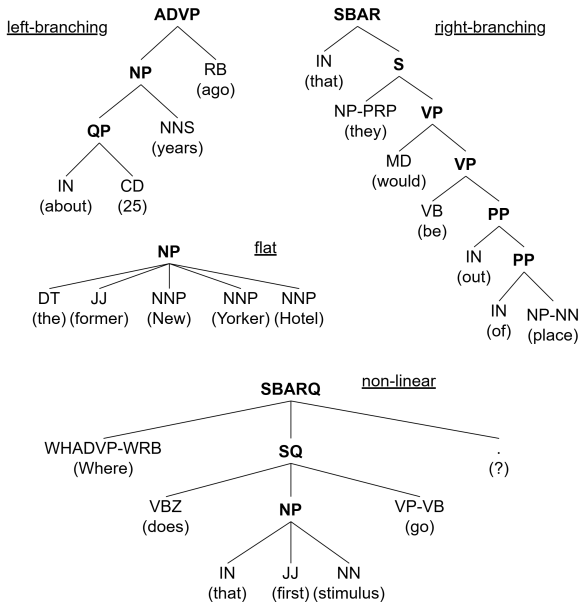


Figure 7: Examples of English constituency trees with fully left/right-branching, flat, and non-linear structures.

A Example Trees in English

Figure 7 shows examples of English constituency trees with fully left/right-branching, flat, and non-linear structures. As it is difficult to find complete sentences that exhibit these specific structures, the example trees in Figure 7 are subtrees extracted from larger constituency trees.

B Setting Details

In this study, we analyze delexicalized constituency trees, treating preterminal nodes, typically POS tags, as leaf nodes. However, the Hebrew and Polish treebanks employ specific annotation conventions that necessitate different preprocessing steps, as detailed below.

First, the Hebrew treebank features two layers of preterminals (Seddah et al., 2013). Therefore, we use the higher preterminal node as the effective leaf node in our analysis.

In the Polish treebank, the lowest-layer nonterminals (i.e., those directly dominating the preterminals) function similarly to preterminals themselves (Woliński, 2019). Unlike the Hebrew data, these lowest-layer nonterminals in Polish are sometimes nested. When these lowest-layer nonterminals are nested, we simply treat the highest ones as leaf nodes.

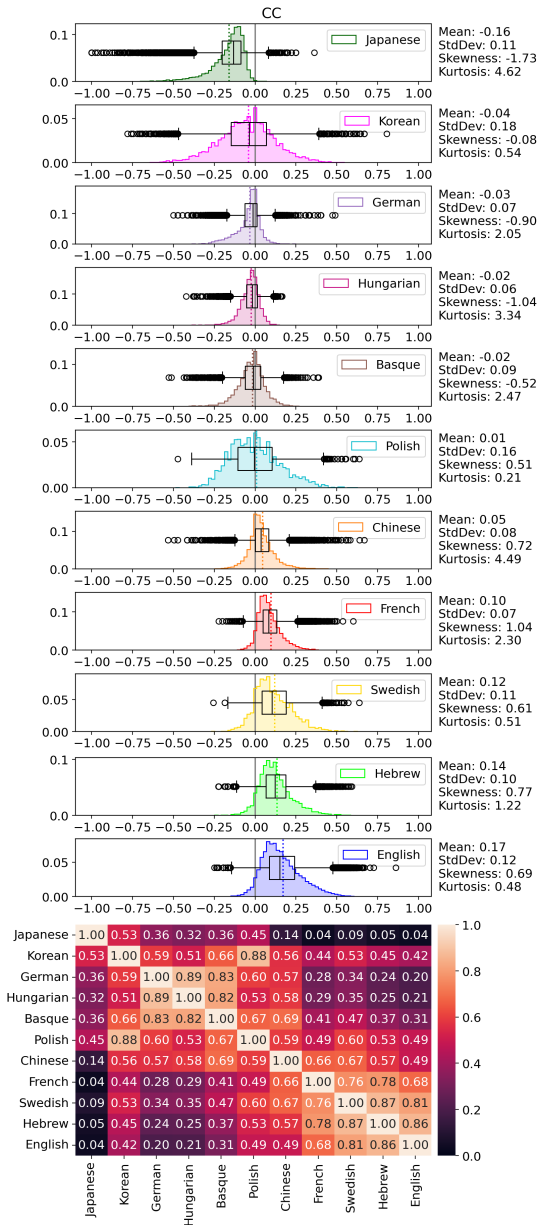


Figure 8: Distributions of CC for 11 natural language treebanks. The values in the heatmap are pairwise HIs.

C Other Results

Figure 8 and Figure 9 show the distributions for CC, RJ in each language, together with box plots within 1.5 IQR and heatmaps of the pairwise HI between languages.

Table 5 shows the HI between each random model and the original treebank for each measures.

AR	English	Chinese	Japanese	French	German	Korean	Basque	Hebrew	Hungarian	Polish	Swedish
Yule	0.93	0.87	0.75	0.78	0.83	**0.88**	0.85	0.80	0.83	0.83	0.84
Yule+arity	**0.94**	0.82	0.70	0.73	0.79	**0.88**	0.83	0.77	0.75	0.82	0.82
Yule+pos	0.93	0.87	0.75	0.78	0.83	**0.88**	0.85	0.80	0.83	0.83	0.84
Yule+arity+pos	**0.94**	0.82	0.70	0.73	0.79	**0.88**	0.83	0.77	0.75	0.82	0.82
UPCFG	0.10	0.21	0.37	0.13	0.55	0.12	0.15	0.20	0.28	0.43	0.46
PCFG	0.91	**0.90**	**0.92**	**0.93**	**0.91**	0.87	**0.89**	**0.89**	**0.95**	**0.89**	**0.92**

NCE	English	Chinese	Japanese	French	German	Korean	Basque	Hebrew	Hungarian	Polish	Swedish
Yule	**0.93**	0.85	0.83	0.87	0.89	0.94	0.64	**0.93**	0.90	0.92	**0.93**
Yule+arity	0.92	0.83	0.89	0.88	0.86	0.94	0.63	**0.93**	0.88	**0.94**	**0.93**
Yule+pos	0.91	**0.92**	0.89	0.83	**0.91**	**0.96**	**0.82**	**0.93**	0.90	0.86	**0.93**
Yule+arity+pos	0.91	0.89	0.90	0.86	0.89	**0.96**	**0.82**	**0.93**	0.89	**0.94**	**0.93**
UPCFG	0.79	0.66	0.81	0.83	0.74	0.60	0.68	0.82	0.83	0.65	0.78
PCFG	0.88	0.86	**0.92**	**0.92**	**0.91**	0.84	**0.82**	0.88	**0.96**	0.85	0.87

CC	English	Chinese	Japanese	French	German	Korean	Basque	Hebrew	Hungarian	Polish	Swedish
Yule	0.37	0.75	0.32	0.44	0.79	0.85	0.81	0.44	0.78	0.87	0.57
Yule+arity	0.30	0.69	0.17	0.33	0.86	0.85	**0.92**	0.33	0.84	**0.92**	0.49
Yule+pos	**0.87**	0.86	0.77	0.82	0.81	0.87	0.74	0.83	0.76	0.78	0.88
Yule+arity+pos	0.83	**0.87**	0.57	0.79	**0.94**	0.86	0.85	0.80	0.90	0.89	0.88
UPCFG	0.57	0.70	0.60	0.76	0.72	0.87	0.87	0.70	0.88	0.72	0.78
PCFG	0.86	0.83	**0.90**	**0.91**	0.91	**0.95**	0.85	**0.93**	**0.96**	0.88	**0.93**

EWC	English	Chinese	Japanese	French	German	Korean	Basque	Hebrew	Hungarian	Polish	Swedish
Yule	0.22	0.73	0.21	0.21	0.93	0.69	0.81	0.19	0.65	0.57	0.37
Yule+arity	0.21	0.71	0.16	0.18	0.91	0.69	0.81	0.17	0.60	0.56	0.35
Yule+pos	0.57	0.88	0.53	0.49	0.87	0.73	0.81	0.47	0.73	**0.92**	0.67
Yule+arity+pos	0.57	0.85	0.32	0.44	0.89	0.73	0.79	0.44	0.63	0.78	0.68
UPCFG	0.26	0.64	0.26	0.30	0.81	0.76	0.65	0.32	0.63	0.73	0.45
PCFG	**0.91**	**0.92**	**0.94**	**0.93**	**0.94**	**0.95**	**0.85**	**0.89**	**0.97**	0.90	**0.93**

RJ	English	Chinese	Japanese	French	German	Korean	Basque	Hebrew	Hungarian	Polish	Swedish
Yule	0.28	0.77	0.17	0.22	0.88	0.72	0.82	0.22	0.69	0.57	0.39
Yule+arity	0.27	0.76	0.13	0.20	0.88	0.71	0.81	0.21	0.68	0.56	0.38
Yule+pos	0.66	**0.93**	0.47	0.50	0.85	0.77	**0.85**	0.54	0.77	**0.90**	0.70
Yule+arity+pos	0.67	0.90	0.28	0.49	0.88	0.77	0.82	0.55	0.72	0.81	0.76
UPCFG	0.38	0.65	0.28	0.40	0.84	0.76	0.70	0.40	0.70	0.71	0.53
PCFG	**0.88**	0.91	**0.93**	**0.90**	**0.95**	**0.93**	**0.85**	**0.86**	**0.95**	0.88	**0.90**

Table 5: HI between each random model and its original treebank.

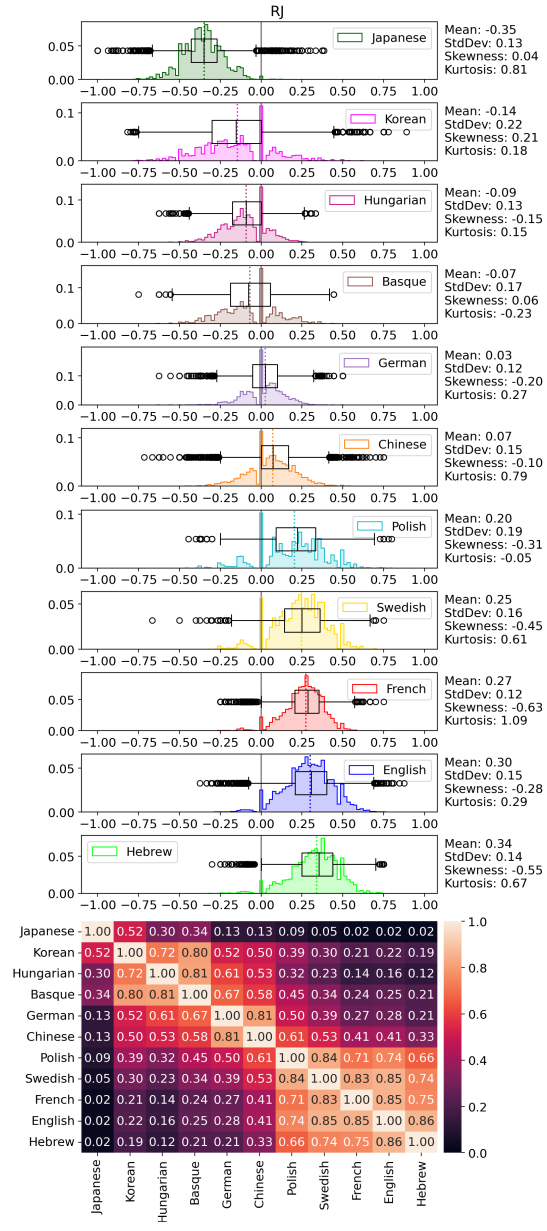


Figure 9: Distributions of RJ for 11 natural language treebanks. The values in the heatmap are pairwise HIs.

First Insights into the Syntax of Slovene Student Writing: A Statistical Analysis of Šolar 3.0 vs. Učbeniki 1.0

Tina Munda
Centre for Language Resources and Technologies,
University of Ljubljana
Večna pot 113, Ljubljana
tina.munda@cjvt.si

Špela Arhar Holdt
Faculty of Arts,
University of Ljubljana
Aškerčeva cesta 2, Ljubljana
spela.arharholdt@ff.uni-lj.si

Abstract

This study investigates the syntactic features of Slovene student writing by comparing essays from the Šolar 3.0 corpus (ages 13–19; primary and secondary school levels) with textbook texts from the Učbeniki 1.0 corpus aligned to the same educational stages. We apply quantitative syntactic analysis at two complementary levels: clause-type frequency (coordination, parataxis, and four types of subordination) and tree-based syntactic complexity measures (number of clauses, clauses per T-unit, and maximum parse-tree depth). Results show that students heavily rely on coordination and specific subordinate clauses (especially object and adverbial), producing more clauses per sentence and per T-unit than textbooks. However, their sentences tend to exhibit flatter syntactic structures, with shallower embedding in primary school and only modest increases in tree depth by secondary school. These findings reveal a divergence between surface-level complexity and hierarchical depth, highlighting developmental trends and instructional targets in written syntactic maturity. We conclude by discussing implications for syntactic development and directions for future research.

1 Introduction

In recent years, the availability of large scale corpora annotated with syntactic relations, improved accuracy of automatic parsing tools, and user-friendly software for corpus-data extraction have opened new opportunities for syntactic research based on automatically processed linguistic data.

This paper investigates the use of syntactic structures in the writing of Slovene primary (age 13–15) and secondary (age 15–19) school students, using the developmental Šolar 3.0 corpus (Arhar Holdt and Kosem, 2024), and compares it to the syntactic patterns found in the corpus of Slovene textbooks—Učbeniki 1.0 (Kosem and Pori, forthcoming).

The Šolar 3.0 corpus comprises 5,485 Slovene texts (1,635,407 words) produced by pupils in grades 6–9 of Slovene primary school (ages 13–15) and by students in Slovene secondary school (ages 15–19). Its composition reflects the variety of classroom genres: essays (58.7%), classroom exercises (15.0%), practical texts (12.6%), and tests (13.7%). Just under one fifth of the texts (19.7%) come from primary students, with the remainder (80.3%) authored by secondary-school students. Most texts (85.4%) were written in Slovene language classes. Although teacher corrections are available for 38.18% of the corpus, our analysis uses the original student productions to capture authentic student syntax. To ensure better comparability in genre and communicative purpose, we restrict our analysis to texts labeled as *esej ali spis* (essay), which constitute the majority text type in Šolar.

The Učbeniki 1.0 corpus consists of 127 Slovene-language textbooks (4,302,857 words) covering grades 1–9 of primary school and all years of secondary school, across 16 subjects. Primary school textbooks make up 71.6% of the corpus, with secondary school titles accounting for 28.4%. For direct comparability with Šolar 3.0, we restrict our analysis to grades 6–9 of primary school and all secondary school textbooks, excluding readers and early grade textbooks. This alignment ensures that both corpora reflect the pedagogical materials and student texts relevant to the same educational stages.

Although the two corpora differ in genre, comparing them is sensible: textbooks—written for these age and comprehension levels—represent the intended or desired level of language competence, while students' texts reveal their actual writing skills. We cannot, however, be sure that textbook authors intentionally or successfully tailored their language to the target age group.

Situated within the domain of literacy develop-

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 105–114

ment, this study aims to provide a clearer picture of how syntactic competence manifests in student writing compared to the standard Slovene patterns promoted through pedagogical materials. The analysis relies on corpus data (Munda et al., 2025a; Munda et al., 2025b) obtained via the Universal Dependencies[1] (UD) framework and represents the first computational syntactic study of student writing in the Slovene context. As a starting point, we ask: To what extent do student texts differ syntactically from textbook texts, and how can these differences be meaningfully interpreted? Can such insights yield pedagogically useful guidance for developing writing competence? To answer these questions, our quantitative analysis proceeds on two complementary levels:

- Clause-level frequency comparison: we compare the raw and normalized frequencies of coordination, parataxis, and four subordination types (subject, object, adverbial, and relative clauses) across Šolar 3.0 and Učbeniki 1.0, separately for primary and secondary school subsets.
- Tree-based syntactic complexity profiling: to capture the deeper structural features of students' sentences, we compute three tree-based syntactic complexity measures—number of clauses, clauses per T-unit, and maximum parse tree depth—derived from UD parses.

Together, these approaches yield a richer account of student syntactic development. The findings offer practical implications for language instruction, highlighting which clause types and complexity dimensions merit explicit teaching, and to what extent students' real world usage aligns with textbook models. This work was conducted within the framework of the *Empirical foundations for digitally-supported development of writing skill*[2] project, which supports applied linguistic research for educational development.

The remainder of this paper is organized as follows. Section 2 provides related work. Section 3 describes data-preparation procedures and the methodology for the quantitative syntactic analyses, including clause-type frequency counts and tree-based syntactic complexity measures. Section 4 reports the results of these analyses for primary and secondary school subsets. Section 5 offers a discussion of our findings. Finally, Section 6 provides our conclusion and outlines avenues for future work.

2 Related Work

Computational syntactic analysis of Slovene corpora has only recently gained significant momentum, primarily due to advancements in syntactic parsing following the Universal Dependencies (UD) framework. This cross-linguistically consistent annotation scheme, having been developing for over the past decade, has enabled more sophisticated analyses of Slovene syntactic structures. The current version of the CLASSLA-Stanza parser (Ljubešić et al., 2024) has achieved remarkably high accuracy rates of 95.54% for UD-relations in Slovene texts, making large-scale automated syntactic analysis increasingly reliable.

Recent studies have begun to leverage UD-annotated corpora for Slovene linguistic research. Dobrovoljc (2024) explored the potential and limitations of UD-relations for analyzing spoken Slovene, highlighting the adaptability of the framework to various contexts. Terčon (2024) demonstrated the value of UD annotations for comparative syntactic analysis by automatically measuring syntactic complexity differences between written and spoken Slovene corpora.

Research on the Šolar developmental corpus has thus far focused primarily on lexical aspects of student writing. Rozman et al. (2018) analyzed collocations occurring within the corpus, while Gantar and Bon (forthcoming) investigated multi-word lexical challenges faced by students in Slovene primary and secondary schools. Arhar Holdt and Rozman (2015) examined the most frequently corrected verbal and pronominal lemmas in the corpus, aiming to develop data-driven instructional materials responsive to actual student needs.

By contrast, syntactic dimensions of student writing remain largely unexplored. This is a notable gap given the importance of syntactic development in educational contexts and the potential for corpus-based findings to inform pedagogical practice. The present study addresses this gap by providing the first quantitative analysis of Slovene student syntax compared to textbook models, leveraging the analytical power of UD annotations to reveal developmental patterns in syntactic complexity and clause usage across educational levels.

Although Slovene-based studies are scarce, re-

[1] https://universaldependencies.org/
[2] https://www.cjvt.si/prop/en/

search on as a first language (L1) writing provides valuable developmental benchmarks and comparative context for assessing syntactic growth across educational levels.

In their studies of adolescent writing in English L1 in the United States, Beers and Nagy (2009); (2011) examined clause density—measured as clauses per T-unit—among other syntactic complexity measures, across various essay types. For Grades 7 and 8 students (ages 12–14), Beers and Nagy (2009) reported a mean of 1.5 clauses per T-unit in narrative essays and 2.0 in persuasive essays. In a broader study of types of school writing, Beers and Nagy (2011) analyzed texts from Grades 3, 5, and 7, finding that Grade 7 (ages 12–13) clause density ranged from 1.21 in descriptive texts to 2.08 in persuasive essays. Since the Šolar corpus essays are not subdivided by types, the average of the four essay types reported by Beers and Nagy (2011)—1.46 clauses per T-unit—offers a useful reference point for interpreting clause density at the primary-school level in the present study. A foundational study by Hunt (1970) further supports these benchmarks. His analysis of 8th-grade (ages 13–14) student writing revealed an average of 1.42 clauses per T-unit, indicating slightly lower syntactic complexity in U.S. school settings compared to the average across essay types reported by Beers and Nagy.

As for the secondary level, Hunt (1970) reported a mean of 1.68 clauses per T-unit for 12th-grade students (ages 17–18), suggesting moderate syntactic growth in late adolescence. This figure provides a relevant benchmark for evaluating clause density in the secondary-school subset of the present study.

These studies also emphasize two important caveats: first, that increased syntactic complexity does not always correlate with higher writing quality; and second, that genre exerts a significant influence on syntactic structures. Different genres give rise to distinct syntactic realizations, depending on their communicative goals. Therefore, syntactic complexity should be interpreted in light of genre conventions rather than as a standalone indicator of writing development.

3 Methodology

3.1 Data Preparation

The syntactic data (Munda et al., 2025a; Munda et al., 2025b) were extracted in advance from both corpora using the STARK tool (Krsnik and Dobrovoljc, 2025), which leverages morphosyntactic and dependency annotations following the UD schema.

To enable direct comparison by school level, educational level metadata were added to each textbook parse file, allowing the pedagogical corpus to be split into primary- and secondary-school subsets. The Šolar corpus already contained this information, so the same split was applied there. In addition, only the syntactic data from texts tagged as *esej ali spis* (essay) were retained from Šolar 3.0, with other text types excluded to reduce structural variability.

After this step, every syntactic structure appears in four subsets:

- Šolar: primary-school student texts (šolar_PS)
- Šolar: secondary-school student texts (šolar_SS)
- Učbeniki: primary-school textbooks (učb_PS)
- Učbeniki: secondary-school textbooks (učb_SS)

These form the basis of all subsequent quantitative analyses; see Table 1 for their size.

Subcorpus	Size (words)
šolar_PS	195,233
šolar_SS	1,075,409
učb_PS	2,039,313
učb_SS	1,252,755

Table 1: Size in words of the four subcorpora of Šolar 3.0 (šolar) and Učbeniki 1.0 (učb) for primary (PS) and secondary (SS) school levels.

3.2 Quantitative Syntactic Analysis

3.2.1 Clause-Type Frequency Analysis

We compare the frequencies of major relations at the clause level—coordination, parataxis, and four subordination types (subject, object, adverbial, and relative clauses)—across the four data subsets. Observed counts of each relation were tabulated, then normalized per 1,000 tokens to adjust for corpus-size differences.

To test whether students and textbooks differ in their use of each structure, we applied the Chi-square test (χ^2) on 2×2 contingency tables (developmental vs. pedagogical), conducted separately for primary and secondary levels. We report $p < 0.0001$ for all comparisons and calculate

the Phi coefficient (ϕ) as an effect-size measure to quantify the magnitude of the association between corpus type and each syntactic structure.

3.2.2 Tree-Based Syntactic Complexity Measures

To capture sentence-level structural complexity, we compute three tree-based syntactic complexity metrics on every sentence in each subset. These measures are modeled after Terčon (2024), who applied them to compare syntactic complexity across written and spoken registers of Slovene. These include:

- NR_OF_CLAUSES: total number of finite and non-finite clauses per sentence.
- CLAUSES_PER_T-UNIT: clause density, calculated as the number of clauses divided by the number of T-units in a sentence, indicating how many subordinate or coordinate clauses are packed into each minimal 'idea' unit.
- MAX_TREE_DEPTH: height of the dependency tree, measuring embedding depth; put simply: the largest number of levels from any word up to the main verb in the sentence's dependency tree, showing how many nested syntactic layers the sentence has.

We excluded the three token-based measures—Mean Dependency Distance, Normalized Dependency Distance and Words per Sentence—that were included in Terčon (2024), because the textbook corpus contains noisy segmentation that distorts token-based calculations. The three selected measures, by contrast, rely on parse-tree structure and are more robust to segmentation errors.

For each metric, we first compute descriptive statistics (mean, standard deviation, median) for Šolar vs. Učbeniki at each school level. To assess group differences, we run Mann–Whitney U tests (non-parametric) and report rank-biserial correlations as effect sizes. Finally, to control the overall error rate across the three comparisons at each level, we apply the Holm–Bonferroni correction to the raw p-values.

4 Results

In this section, we present the results of the quantitative analyses, organized by educational level. We first report findings for the primary-school texts, followed by those from the secondary-school level. At each level, we examine both clause-type frequencies and tree-based syntactic complexity measures.

4.1 Primary-School Level

4.1.1 Clause-Type Frequency Analysis

The quantitative analysis of syntactic structures at the primary-school level (see Table 2) shows clear differences between the developmental corpus (Šolar 3.0) and the textbook corpus (Učbeniki 1.0). Coordination structures (*conj*) occur at a notably higher normalized frequency per 1,000 tokens in the Šolar corpus (31.64) compared to textbooks (11.72). Subordination structures also exhibit higher normalized frequency in student texts (38.41) relative to textbook texts (22.54). Among subordinate clause types, adverbial clause *(advcl*) and object clause (*ccomp*) stand out with notably higher normalized frequencies in student texts (16.59 and 11.63, respectively) compared to textbooks (7.54 and 3.21, respectively). Relative clauses (*acl*), however, show relatively similar frequencies across both corpora (9.29 vs. 11.21). Parataxis (*parataxis*) structures appear slightly less frequently in student texts (13.65) compared to textbooks (14.74).

Chi-square tests confirm these differences are statistically significant (see Table 2). The strongest associations, as indicated by Phi (ϕ), appear for coordination (ϕ=0.0488), object (ϕ=0.0379) and adverbial clauses (ϕ=0.0281). Relative and subject clauses, and parataxis show smaller effect sizes, indicating minimal differences between students and textbooks, despite statistical significance.

4.1.2 Tree-Based Syntactic Complexity Measures

Tree-based syntactic complexity measures reveal additional structural insights into primary-school-level sentence constructions (see Table 3, Figure 1).

The number of clauses per sentence is notably higher in the Šolar corpus (mean=2.22) than in textbooks (mean=1.73). Clause density (number of clauses per T-unit) is slightly higher in student texts (mean=1.45) than in textbooks (mean=1.29). These differences are visually evident in the corresponding boxplots (Figure 1), where student data show a broader range and higher median values for both measures, indicating more frequent use of multi-clause structures and denser clause packaging in learner writing.

Interestingly, the mean maximum tree depth—a measure of syntactic embedding—does not differ

Structure	šolar_PS		učb_PS		Φ
	Ofq	Nfq	Ofq	Nfq	
coordination	6,177	31.64	23,905	11.72	0.0488
subordination	7,499	38.41	45,973	22.54	0.0293
subject c.	177	0.91	1,196	0.59	0.0036
object c.	2,270	11.63	6,544	3.21	0.0379
adverbial c.	3,239	16.59	15,369	7.54	0.0281
relative c.	1,813	9.29	22,864	11.21	0.0052
parataxis	2,665	13.65	30,050	14.74	0.0025
Total	16,341	83.7	99,928	49.00	

Table 2: Observed (Ofq) and normalized (Nfq; per 1,000 tokens) frequencies of syntactic structures in the primary school (PS) subset of Šolar 3.0 and Učbeniki 1.0, along with Phi (Φ) effect sizes for differences in syntactic structure use across both corpora in the primary school (PS) subset. All differences are statistically significant (χ^2, p < 0.0001).

substantially between corpora (4.10 for Šolar vs. 4.28 for textbooks), with nearly identical medians and overlapping interquartile ranges. This suggests that while students produce more clauses, they do not build significantly deeper syntactic structures.

Mann–Whitney U tests confirm statistically significant differences for clause-related metrics (U=1.84×10^9, r_{rb}=–0.363* for number of clauses; U=1.68×10^9, r_{rb}=–0.260* for clauses per T-unit), but not for maximum tree depth (U=1.45×10^9, r_{rb}=0.005). These findings indicate that student essays are structurally more clause-heavy, but not more syntactically embedded than textbooks.

4.2 Secondary-School Level

4.2.1 Clause-Type Frequency Analysis

The syntactic-structure frequencies at the secondary-school level (see Table 4) also exhibit notable differences between the Šolar and textbook corpora. Coordination structures again occur more frequently in the Šolar corpus (30.85 per 1,000 tokens) compared to textbooks (11.16). Similarly, overall subordination structures are used more frequently by students (44.67) than in textbooks (23.00). Among subordinate clause types, adverbial clause (16.23) and object clause (11.99) remain prominently higher in students' texts compared to textbooks (7.17 and 2.94, respectively). Relative clauses have slightly higher frequencies in student texts (14.53) compared to textbooks (12.29), while parataxis structures appear slightly less often in the Šolar corpus (12.37 vs. 14.28 in textbooks).

Chi-square tests confirm the statistical significance of these differences (see Table 4). Coordination (Φ=0.0697), subordination (Φ=0.0605), object clause (Φ=0.0537), and adverbial clauses (Φ=0.0426) show moderate effect sizes, reflecting stronger corpus differences. Relative clause, subject clause, and parataxis have minimal effect sizes despite statistical significance.

Compared to the primary school subset, coordination and subordination frequencies remain similarly elevated in student texts, but effect sizes are notably higher in the secondary-school data—suggesting a stronger divergence in how advanced students structure their writing. Additionally, whereas relative clause frequencies were nearly identical across corpora in the primary school subset, secondary-school students use relative clauses more frequently, though the effect size remains small.

4.2.2 Tree-Based Syntactic Complexity Measures

At the secondary-school level, tree-based syntactic complexity measures show nuanced differences (see Table 5, Figure 2).

Student writing features a markedly higher number of clauses per sentence (mean=2.66 vs. 1.78), indicating a greater tendency to produce multi-clause constructions. In addition, students demonstrate higher clause density, measured as clauses per T-unit (mean=1.62 vs. 1.31), suggesting that more subordinate or coordinate clauses are packed into individual minimal idea units. These findings are supported by the boxplots (Figure 2), which reveal a broader distribution and higher medians for both the number of clauses and clause density in student texts.

Interestingly, mean maximum tree depth is also slightly higher in student texts (4.84 vs. 4.56),

Measure	šolar_PS			učb_PS			U ($\times 10^9$)	r_{rb}
	mean	sd	median	mean	sd	median		
NR_OF_CLAUSES	2.22	1.36	2	1.73	1.36	1	1.84	-0.363*
CLAUSES_PER_T-UNIT	1.45	0.66	1	1.29	0.55	1	1.68	-0.260*
MAX_TREE_DEPTH	4.10	1.43	4	4.28	2.21	4	1.45	0.005

Table 3: Tree-based syntactic complexity measures for the primary school (PS) subset of the Šolar 3.0 (šolar) and Učbeniki 1.0 (učb) corpora. Values are reported as mean, standard deviation (sd), and median per sentence. Mann–Whitney U test statistics (scaled $\times 10^9$) and rank biserial correlations (r_{rb}) are included. All comparisons are statistically significant at $p < 0.0001$ after Holm–Bonferroni correction. Asterisks indicate comparisons that remain significant at $\alpha = 0.05$.

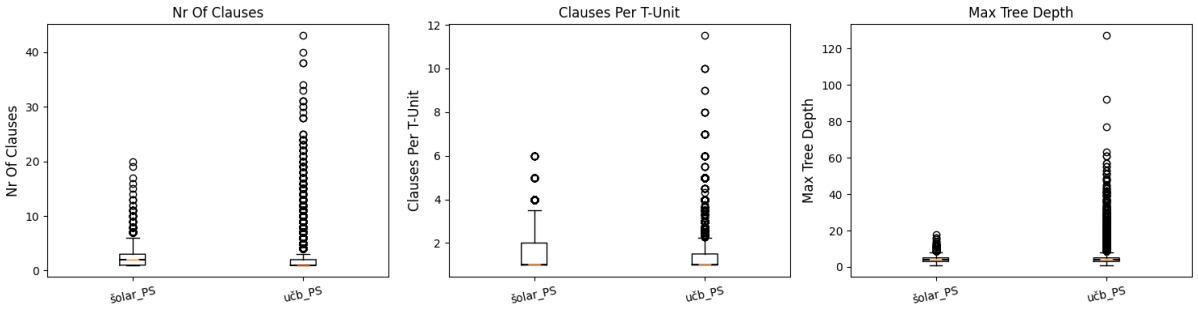


Figure 1: Boxplots of three tree-based syntactic complexity measures—number of clauses, clauses per T-unit, and maximum tree depth—for primary school texts (Šolar_PS vs. Učbeniki_PS). The figure illustrates higher clause density in student writing and a wider range of tree depths and clause counts in textbook data, likely due to segmentation noise.

although the difference is less pronounced and medians remain identical (both 5). However, the broader range and right-skewed distribution in the textbook corpus—reflected in more extreme outliers—suggest occasional structurally deeper sentences, probably due to segmentational noise.

Mann–Whitney U tests confirm statistically significant differences for all three measures ($p < 0.0001$, adjusted using Holm–Bonferroni correction). The effect sizes (rank-biserial correlation, r_{rb}) are strongest for the number of clauses (–0.363) and clauses per T-unit (–0.260), while the difference in maximum tree depth shows a smaller effect (–0.111), consistent with the visual overlap in distributions.

5 Discussion

This study provides novel insights into syntactic patterns characterizing Slovene students' written language in comparison to pedagogical patterns represented in textbooks. By employing a dual-methodological framework—quantitative analyses of clause-type frequencies and tree-based syntactic complexity measures—we uncover distinct patterns in how students utilize syntactic structures at both primary and secondary school levels.

Before interpreting these findings, it is important to consider the genre makeup of each corpus. The Šolar 3.0 corpus used in this study includes only student essays, while the textbook corpus encompasses a broader mix of descriptive, expository, and narrative texts. This genre imbalance partially explains clause chaining in student writing, which may amplify the frequency of coordination and subordination in comparison to textbooks.

5.1 Primary School: Student Texts vs. Textbooks

At the primary school level, students show a clear preference for coordination and subordination strategies. Compared to textbooks, student texts contain more than double the normalized frequencies of coordination (31.64 vs. 11.72) and adverbial clauses (16.59 vs. 7.54), with relative and object clauses also more frequent. These differences are statistically significant across all clause types, with Phi effect sizes ranging from small to moderate (Φ=0.0025–0.0488). This suggests that primary students are already making active use of a range of clause-linking devices in their writing.

Tree-based syntactic complexity measures reveal more nuanced structural patterns. While student sentences contain more clauses per sentence (2.22 vs. 1.73) and show greater clause

Structure	šolar_SS		učb_SS		Φ
	Ofq	Nfq	Ofq	Nfq	
coordination	33,176	30.85	13,983	11.16	0.0697
subordination	48,035	44.67	28,809	23	0.0605
subject c.	2,068	1.92	742	0.59	0.0191
object c.	12,897	11.99	3,684	2.94	0.0537
adverbial c.	17,449	16.23	8,988	7.17	0.0426
relative c.	15,621	14.53	15,395	12.29	0.0097
parataxis	13,307	12.37	17,891	14.28	0.0083
Total	94,518	87.89	60,683	48.44	

Table 4: Observed (Ofq) and normalized (Nfq; per 1,000 tokens) frequencies of syntactic structures in the secondary school (SS) subset of Šolar 3.0 and Učbeniki 1.0, along with Phi (Φ) effect sizes for differences in syntactic structure use across both corpora in the secondary school (SS) subset. All differences are statistically significant (χ^2, $p < 0.0001$).

density as measured by clauses per T-unit (1.45 vs. 1.29), their maximum tree depth is slightly lower (4.10 vs. 4.28). Both clause metrics show significant differences (r_{rb} = –0.363 and –0.260), while tree depth does not. These results indicate that primary-school students tend to produce more linear clause sequences and pack more information per idea unit, but do so using relatively shallow structures—favoring additive linking and minimally embedded subordination over hierarchical nesting.

In terms of cross-linguistic reference points, the mean clause density of 1.45 clauses per T-unit observed in Slovene primary-school essays closely aligns with values reported for English L1 adolescent writing. Beers and Nagy (2011) found that Grade 7 students (ages 12–13) produced between 1.21 and 2.08 clauses per T-unit depending on text type, with an average of 1.46 across four text types. Hunt (1970) similarly reported a mean of 1.42 for 8th-grade students (ages 13–14). These parallels suggest that Slovene students in this age range exhibit comparable syntactic development in terms of clause density, despite language-specific and curricular differences.

5.2 Secondary School: Student Texts vs. Textbooks

In secondary school, the same overall trends persist but become more pronounced. Students continue to use coordination (30.85 vs. 11.16) and subordination (44.67 vs. 23.00) far more frequently than textbooks. This includes object (11.99 vs. 2.94) and adverbial clauses (16.23 vs. 7.17) both of which exhibit moderate effect sizes (Φ=0.0537 and 0.0426). These elevated frequencies suggest that students are increasingly using complex sentence structures to develop arguments and articulate relationships between ideas.

Tree-based syntactic complexity measures reinforce this pattern, showing significantly more clauses per sentence (2.66 vs. 1.78) and higher clause density (1.62 vs. 1.31) in student writing, with large effect sizes (r_{rb}=–0.363 and –0.260). In contrast to the primary level, however, maximum tree depth is also greater in student texts than in textbooks (4.84 vs. 4.56), and while the effect size is smaller (r_{rb}=–0.111), the difference remains statistically significant. These findings point to a developmental progression: secondary-school students not only use more clauses but also begin to imbed them more hierarchically—an indication of growing syntactic proficiency and control.

These clause density findings correspond well with those reported for English L1 writing at this educational level. Hunt (1970) found that 12th-grade students (ages 17–18) produced a mean of 1.68 clauses per T-unit, which aligns closely with the 1.62 observed in Slovene secondary-school essays. This supports the interpretation that clause-per-T-unit density may reflect a broader developmental milestone in adolescent writing, observable across typologically different languages.

5.3 Developmental Trends across School Levels

A comparison of student writing across school levels reveals a developmental trajectory in syntactic maturity. From primary to secondary school, the mean number of clauses per sentence increases (2.22 to 2.66), as does clause density per T-unit (1.45 to 1.62), reflecting a growing tendency to

Measure	šolar_SS			učb_SS			U ($\times 10^9$)	r_{rb}
	mean	sd	median	mean	sd	median		
NR_OF_CLAUSES	2.66	1.67	2	1.78	1.46	1	4.82	-0.363*
CLAUSES_PER_T-UNIT	1.62	0.77	1.5	1.31	0.57	1	4.45	-0.260*
MAX_TREE_DEPTH	4.84	1.72	5	4.56	2.11	4	3.93	-0.111*

Table 5: Tree-based syntactic complexity measures for the secondary school (SS) subset of the Šolar 3.0 (šolar) and Učbeniki 1.0 (učb) corpora. Values are reported as mean, standard deviation (sd), and median per sentence. Mann–Whitney U test statistics (scaled $\times 10^9$) and rank biserial correlations (r_{rb}) are included. All comparisons are statistically significant at $p < 0.0001$ after Holm–Bonferroni correction. Asterisks indicate comparisons that remain significant at $\alpha = 0.05$.

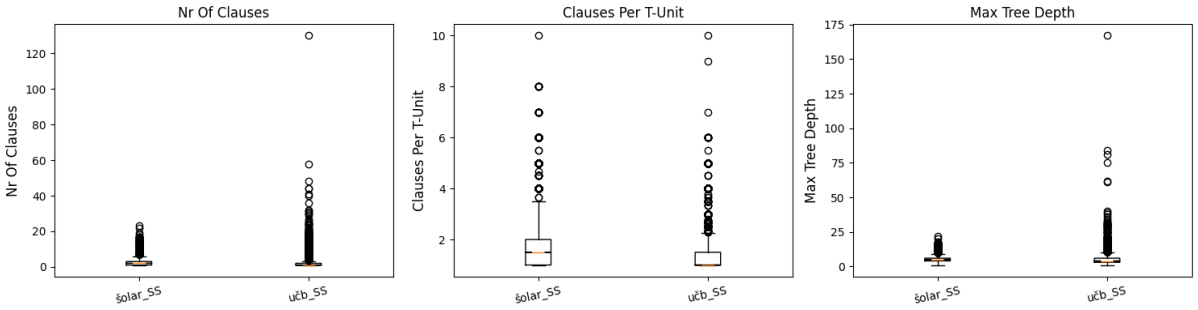


Figure 2: Boxplots of three tree-based syntactic complexity measures—number of clauses, clauses per T-unit, and maximum tree depth—for secondary school texts (šolar_SS vs. učb_SS). The figure illustrates higher clause density in student writing and a wider range of tree depths and clause counts in textbook data, likely due to segmentation noise.

elaborate ideas through clause integration.

At the same time, maximum tree depth rises only modestly (4.10 to 4.84), suggesting that while older students use more clauses, their embedding depth increases more gradually. This indicates that syntactic development may initially proceed through horizontal expansion—adding and linking clauses—before progressing to deeper hierarchical structuring.

Taken together, these developmental trends support the interpretation that Slovene students, as they mature, expand their syntactic repertoire primarily by increasing clause quantity and density, rather than embedding complexity. Instructional efforts might focus on helping students transition from additive, chain-like structures toward more varied and hierarchically integrated syntax.

It is important to interpret these results in light of the reference standard. While in the study, textbooks serve as a syntactic benchmark, they do not represent a neutral writing model. Sentence structures may be intentionally simplified for pedagogical clarity, which can inadvertently limit syntactic exposure and under-challenge learners. The observed disparities thus reflect not only student development but also the editorial and instructional conventions that shape textbook style.

6 Conclusion and Future Work

6.1 Conclusion

This study provides the first large-scale syntactic comparison between Slovene student writing and textbook models, using dependency-parsed data from the Šolar 3.0 and Učbeniki 1.0 corpora. Focusing on essays produced by students in Slovene primary and secondary school, we examined clause-type usage and three tree-based syntactic complexity measures: number of clauses, clauses per T-unit, and maximum tree depth.

The results reveal consistent differences between student writing and textbook texts across educational levels. Slovene students in both primary and secondary school employed significantly more coordination and subordination—particularly object and adverbial clauses—than textbooks, reflecting the clause-chaining tendencies of essayistic expression. At the same time, students produced significantly more clauses per sentence and per T-unit, while maximum tree depth remained comparable or shallower, suggesting a preference for linear rather than deeply embedded syntactic structures.

These findings highlight the developmental nature of student writing: increased clause density signals growing syntactic fluency, yet deeper hier-

archical structuring lags behind. The comparison with textbook models also underscores potential mismatches between pedagogical input and student output, calling for closer alignment between instructional materials and authentic student language.

Ultimately, this study demonstrates the value of corpus-based syntactic analysis for identifying developmental patterns and informing pedagogical practice.

6.2 Future Work

This analysis has already provided valuable insights into student writing, but there remain several directions to deepen and extend our understanding:

Corpus cleaning and improved annotation: The Učbeniki 1.0 corpus currently contains missegmented sentences and non-alphanumeric artifacts from PDF conversion. A thorough cleaning—removing stray characters and correcting sentence boundaries—would enable more accurate automatic UD annotation and allow us to reintroduce token-based syntactic complexity measures (MDD, NDD). These measures could be applied to enrich the overall picture of syntactic development.

Fine-grained conjunction patterns: A detailed examination of conjunction usage—specifically the distribution and frequency of individual conjunctions within each subtype of coordination and subordination—could uncover broader usage patterns and register effects in student writing.

Exploratory and machine-learning analyses: Beyond hypothesis-driven tests, an exploratory, data-driven approach—using clustering or classification techniques—could uncover hidden patterns of clause use, parse-tree configurations, or connective preferences.

Longitudinal developmental studies: Tracking students over time, from primary through secondary and into higher education, would illuminate how syntactic and lexical competencies evolve. Such longitudinal data could reveal critical periods for particular structures or register shifts as students encounter more advanced writing tasks.

Pursuing these avenues will not only validate and refine our current findings but also chart a richer map of syntactic development in educational contexts.

Limitations

Our findings should be interpreted with certain methodological limitations in mind. First, the Učbeniki 1.0 corpus suffers from noise introduced during PDF conversion: many sentences are missegmented or contain stray characters. This inadvertently lead to segmentation errors and parsing mistakes.

Second, all analyses rely on automatic Universal Dependencies annotation. Although UD provides a consistent framework, off-the-shelf parsers can mislabel complex or ungrammatical constructions—especially in student data, where nonstandard usage may further confuse the parser and introduce annotation errors.

Third, there is a text-type mismatch between corpora: Šolar is dominated by essays, whereas the textbook corpus is largely expository and descriptive. We have noted and taken genre effects into account in our interpretation, but residual differences in discourse conventions may still influence our quantitative measures.

Finally, our syntactic analysis does not include the UD *xcomp* (open clausal complement) relation, which in Slovene often corresponds to object clause. Because *xcomp* cannot be reliably distinguished from other non-clausal complements without manual inspection—and it lacks an overt conjunction—these instances are currently unaccounted for in our object-clause counts. Integrating *xcomp* into future analyses would provide a more complete picture of students' use of object clauses.

Acknowledgments

The paper was produced within the project *Empirical foundations for digitally-supported development of writing skills* (No. J7-3159) and the research programme Language Resources and Technologies for Slovene (No. P6-0411) funded by the Slovene Research and Innovation Agency (ARIS).

The authors would also like to thank the anonymous reviewers for their valuable feedback and constructive suggestions.

References

Špela Arhar Holdt and Iztok Kosem. 2024. Šolar, the developmental corpus of slovene. *Language Resources Evaluation*.

Špela Arhar Holdt and Tadeja Rozman. 2015. Možnosti uporabe podatkov iz korpusa Šolar za pripravo

slovarskih priročnikov. In Mojca Smolej, editor, *Slovnica in slovar – aktualni jezikovni opis, del 1*, pages 67–74. Znanstvena založba Filozofske fakultete, Ljubljana.

Scott F. Beers and William E. Nagy. 2009. Syntactic complexity as a predictor of adolescent writing quality: Which measures? which genre? *Reading and Writing*, 22(2):185–200.

Scott F. Beers and William E. Nagy. 2011. Writing development in four genres from grades three to seven: syntactic complexity and genre differentiation. *Reading and Writing*, 24(2):183–202.

Kaja Dobrovoljc. 2024. Uporaba drevesnice sst v raziskavah govorjene slovenščine: prednosti in omejitve. *Jezik in slovstvo*, 69(4):187–209.

Polona Gantar and Mija Bon. forthcoming. Večbesedni leksikalni problemi pri samostojnem tvorjenju besedil v osnovnih in srednjih šolah. *Sodobna pedagogika*.

Kellogg W. Hunt. 1970. Syntactic maturity in schoolchildren and adults. *Monographs of the Society for Research in Child Development*, 35(1):iii–67.

Iztok Kosem and Eva Pori. forthcoming. Prvi koraki do seznama temeljnega besedišča z analizo korpusa slovenskih učbenikov. *Sodobna pedagogika*.

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

Nikola Ljubešić, Luka Terčon, and Kaja Dobrovoljc. 2024. Classla-stanza: The next step for linguistic processing of south slavic languages. In *Proceedings of the Conference on Language Technologies and Digital Humanities (JT-DH 2024)*, Ljubljana, Slovenia.

Tina Munda, Špela Arhar Holdt, Kaja Dobrovoljc, Iztok Kosem, Eva Pori, and Simon Krek. 2025a. Frequency lists of syntactic structures from the učbeniki 1.0 corpus. Slovenian language resource repository CLARIN.SI.

Tina Munda, Špela Arhar Holdt, Kaja Dobrovoljc, Tadeja Rozman, Mojca Stritar Kučuk, Simon Krek, Irena Krapš Vodopivec, Marko Stabej, Eva Pori, Teja Goli, Polona Lavrič, Cyprian Laskowski, Polonca Kocjančič, Bojan Klemenc, Luka Krsnik, and Iztok Kosem. 2025b. Frequency lists of syntactic structures from the Šolar 3.0 corpus. Slovenian language resource repository CLARIN.SI.

Tadeja Rozman, Špela Arhar Holdt, Senja Pollak, and Iztok Kosem. 2018. Kolokacije v korpusu Šolar. *Jezik in slovstvo*, 63(2/3):117–128.

Luka Terčon. 2024. Uporaba šestih mer skladenjske kompleksnosti za primerjavo jezika v govornem in pisnem korpusu. In *Proceedings of the Conference on Language Technologies and Digital Humanities (JT-DH 2024)*, Ljubljana, Slovenia.

Syntactic units and their length distributions: A case study in Czech

Michaela Nogolová[1], Michaela Koščová[2], Ján Mačutek[2], Radek Čech[3],

[1]Department of Czech Language, University of Ostrava
[2]Mathematical Institute, Slovak Academy of Sciences
[3]Department of Czech Language, Masaryk University
nogolovam@gmail.com, koscmichaela@gmail.com, jmacutek@yahoo.com, cechradek@gmail.com

Abstract

This study investigates the length distributions of syntactic units in Czech across multiple hierarchical levels: sentences, independent clauses, clauses, phrases, subphrases, chunks, and words. Using a diverse dataset -– including Universal Dependency treebanks, presidential speeches, the Czech Bible, and random sample from corpora of modern Czech – the analysis examines whether lengths of these syntactic units follow consistent distributional patterns. Length is defined as the number of immediate subunits, and the distributions were modeled using the right-truncated hyper-Poisson distribution. The results demonstrate that this model fits well distributions of length of all abovementioned syntactic units, pointing to a common principle underlying the organization of syntactic structure in Czech.

1 Introduction

The relationship between linguistic units has long been an important topic in quantitative linguistics, particularly through the study of the Menzerath-Altmann law (MAL henceforward, Menzerath, 1954; Altmann, 1980). The law expresses the relationship between the length of a construct and the length of its constituents (parts of the construct); specifically, the longer the construct, the shorter the constituent on average. While the MAL has been extensively validated using words and syllables, its applicability at the syntactic level remains a subject of ongoing investigation (Andres and Benešová, 2012, Sanada, 2016, Berdicevskis, 2021, Mačutek et al., 2017, Mačutek et al., 2021).

Recent study (Nogolová et al., 2025) has explored the MAL across several units in the language unit hierarchy (sentences - independent clauses – clauses – phrases – subphrases – chunks – words – syllables). These investigations suggest that the MAL holds across all these levels, indicating a consistent pattern also with respect to syntactic constructions. It is the first paper that considers several neighbouring language units simultaneously; all previous papers on this topic limited themselves to partial results, mostly only to one triad of units (such as, e.g., Mačutek et al., 2017 focus solely on clause length in phrases and phrase length in words).

Building upon this foundation, the present paper aims to analyze distributions of lengths of syntactic units mentioned in the previous paragraph. Length of a unit is measured in the number of its direct lower neighbours in the language unit hierarchy. Specifically, we will examine length of sentences (measured in the number of independent clauses the sentence contains), of independent clauses (in clauses), of clauses (in phrases), of phrases (ecubphrases), of subphrases (in chunks), of chunks (in words), and of words (in syllables). The goal is to determine whether length distributions of abovementioned syntactic units display similar patterns that might provide insight into the structure and organization of language units.

2 Language material

The language material for this study is sourced from multiple treebanks and corpora, each offering a distinct representation of the Czech language across various genres and time periods.

A significant portion is drawn from the Universal Dependencies 2.13 (Zeman et al., 2023). Specifically, we utilize six Czech dependency treebanks:

1. UD_Czech-CAC is based on the Czech Academic Corpus 2.0 (Vidová Hladká et al., 2008). This treebank encompasses articles from diverse sources, including journalism, administration, and scientific fields.

2. UD_Czech-CLTT originates from the Czech Legal Text Treebank 2.0 (Kríž and Hladká, 2017). It comprises two legal documents on accounting.

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 115–123

3. FicTree (Jelínek, 2017) includes six books of fiction, one book of fiction for children, and one memoir.

4. Czech-PDT UD (Bejček et al., 2013) contains journalistic texts.

5. UD_Czech-Poetry contains samples from 19th-century Czech poetry from the Corpus of Czech Verse (Plecháč and Kolár, 2015).

6. The Czech part of Parallel Universal Dependencies (PUD) consists of 1000 random sentences translated into Czech from English and other languages.[1]

In addition to the UD treebanks mentioned above, we analyze 89 annual speeches delivered by twelve Czechoslovak and Czech presidents. These speeches are the object of research in Kubát et al. (2021).

Our study also incorporates the Czech Ecumenical Translation of the Bible (CET), a contemporary Czech translation undertaken between 1961 and 1979. This translation renders biblical texts into modern Czech while preserving traditional diction and style. We use the 2001 revision of the CET translation.

Finally, we utilize sentences from the SYN2020 corpus (Křen et al., 2020), a comprehensive and balanced collection of contemporary written Czech developed by the Czech National Corpus. Predominantly encompassing texts from 2015 to 2019, the corpus contains 100 million words and is structured into three equally sized segments: fiction, non-fiction, and newspapers/magazines. For this study, a random sample of 50,000 sentences from each segment was used.

The individual parts were merged and treated as a whole, thus encompassing various genres. The language material used is quite heterogeneous. Menzerath (1954) formulated the relation between word length (in syllables) and the mean syllable length (in phonemes) as valid for the vocabulary. Similarly, we suppose that the Menzerath-Altmann law in general is valid for the inventory of units rather than for particular texts or text genres. Of course no dictionary or corpus contains the whole vocabulary, and it is even less realistic to speak about the complete inventory of sentences, clauses, etc., but some measure of heterogeneity reduces the risk of genre-specific syntactic units.

3 Methodology and operationalization

In this section, we present language units we analyze, following the approach from Nogolová et al. (2025).

Those parts of our corpus that do not come from UD treebanks (i.e., the presidential speeches, the Bible translation, and the sentences from the SYN2020 corpus; see Section 2) were processed using UDPipe 2.0 (Straka, 2018), a trainable pipeline that, among other functions, performs dependency parsing. Subsequently, all the annotated texts were converted to the Surface Syntactic Universal Dependencies (SUD) annotation scheme (Gerdes et al., 2018) using Grew software (Guillaume, 2021). The reason is that the SUD annotation reflects more closely dependency syntax based on purely syntactic (rather than semantic or functional) criteria. SUD provides a representation closer to surface-syntactic frameworks such as Meaning-Text Theory (Mel'čuk, 1988), Word Grammar (Hudson, 2010), and the Prague Dependency Treebank (Hajič et al., 2017).

We took sentences as they are determined by the annotation tool. Only sentences satisfying the following three criteria were included in the analysis: (i) they contain a predicate (a finite verb or an auxiliary) as the sentence root; (ii) they do not contain abbreviations, digits, foreign words, words with unknown syntactic functions, special characters, and words assigned the *flat*[2] or *orphan*[3] syntactic function; (iii) they do not contain a chain of more than two coordinated words, with exception of predicates of independent clauses (discussed below).

Sentence length is expressed as the number of independent clauses the sentence contains. The first independent clause comprises the root of the sentence and all words that are directly or indirectly syntactically linked with it if the links are not coordinations with another predicate. If there is another predicate coordinated with the sentence root, it becomes the root of another independent clause within the sentence.

The immediate constituent of an independent clause is a clause, which consists of the predicate

[1] These sentences form a part of the Parallel Universal Dependencies (PUD) treebanks created for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies (http://universaldependencies.org/conll17/).

[2] Personal names, which exhibit a distinct syntactic behavior, are tagged by this relation.

[3] This often indicates ellipsis, which presents analytical challenges.

and all its direct or indirect dependents, excluding other predicates. To illustrate these units[4], Figure 1 presents the dependency tree of the sentence (1) divided into individual independent clauses (on the left) and clauses (on the right).

(1) *Ani u uvědomované motivace nemůžeme mít přehled o všech motivech, které jsou v daném okamžiku ve hře, uvědomujeme si pouze motivy dominantní, převládající.*
"Even with conscious motivation, we cannot be aware of all motives that play a role at a given moment; we are only realising the dominant, predominant motives."

Sentence (1) has two independent clauses, each defined by a main verb: *nemůžeme* "cannot" and *uvědomujeme* "realising". The first independent clause contains two clauses, as it has two verbs: *nemůžeme* "cannot" and *jsou* (translated here as "play"). The second independent clause consists of a single clause, as it contains only one finite verb.

The immediate unit of the clause is a phrase. A phrase is commonly understood as a multi-word constituent where one word, known as the head or root, holds prominence over the others (Osborne, 2019). For our purposes, the initial phrase of a clause includes the predicate and its leftmost dependent, along with all its own dependents. This approach emphasizes the linear progression of the sentence. Subsequent dependents of the predicate are treated as heads of their own phrases, each encompassing the head and all words dependent on it, directly or indirectly, if applicable.

A subphrase is a newly defined unit introduced by Nogolová et al. (2025). It is defined as the longest sequence of dependent words in which each word (except the head) has at most one dependent. Figure 2 provides a detailed view of the first clause in sentence (1). The left side of the figure illustrates the individual phrases. The first phrase includes the predicate and its leftmost dependent, while the second phrase consists of the remaining dependent – functioning as the head of the phrase – along with all of its own dependents. This results in the following division of the first clause:

[*Ani u uvědomované motivace nemůžeme*] [*mít přehled o všech motivech*].

[4]All examples taken from Nogolová et al. (2025).

The right side of the figure shows the individual subphrases. The first subphrase of the first phrase consists of the predicate *nemůžeme* on its own, because its dependent *u* has more than one dependent and cannot be included. The second subphrase is formed by *u* and all its dependents, since each of them has at most one dependent. Hence, the subphrases of the first phrase are [*nemůžeme*], [*Ani u uvědomované motivace*]. The second phrase makes up a single subphrase, as every word in it has at most one dependent.

A subphrase then consists of chunks defined according to the following criteria: (i) all words within the chunk share the same immediate parent (head); (ii) the chunk comprises only one level of dependency; (iii) the words are contiguous in the subphrase; (iv) no dependent word in the chunk has dependents outside of it.[5] Figure 3 presents the first phrase of the first clause of the sentence (1), divided into subphrases (on the left side) and individual chunks (on the right side). The first subphrase contains a single chunk, as it consists of only one word – *nemůžeme*. The second subphrase is divided into two chunks, [*Ani u*] and [*uvědomované motivace*], since only these combinations satisfy all the criteria outlined above.

Table 1 shows the number of sentences, independent clauses, clauses, phrases, subphrases, chunks, and words in language material we use (see Section 2). We include also words, as word length has been studied extensively in the last few decades. We thus can compare length distributions of syntactic units (a relatively new topic) with many previously achieved results for words.

Unit	**Tokens**
sentence	132 159
independent clause	167 132
clause	245 584
phrase	612 167
subphrase	697 244
chunk	1 047 142
word	1 535 506

Table 1: Number of units in the merged corpus.

Some of these units are, admittedly, purely formal, i.e., they are well defined, but, for the time

[5]This definition is taken from Anderson et al. (2019) with one modification; namely, Anderson et al. (2019) define chunks within sentences, while chunks as defined in this paper do not exceed the boundaries between subphrases.

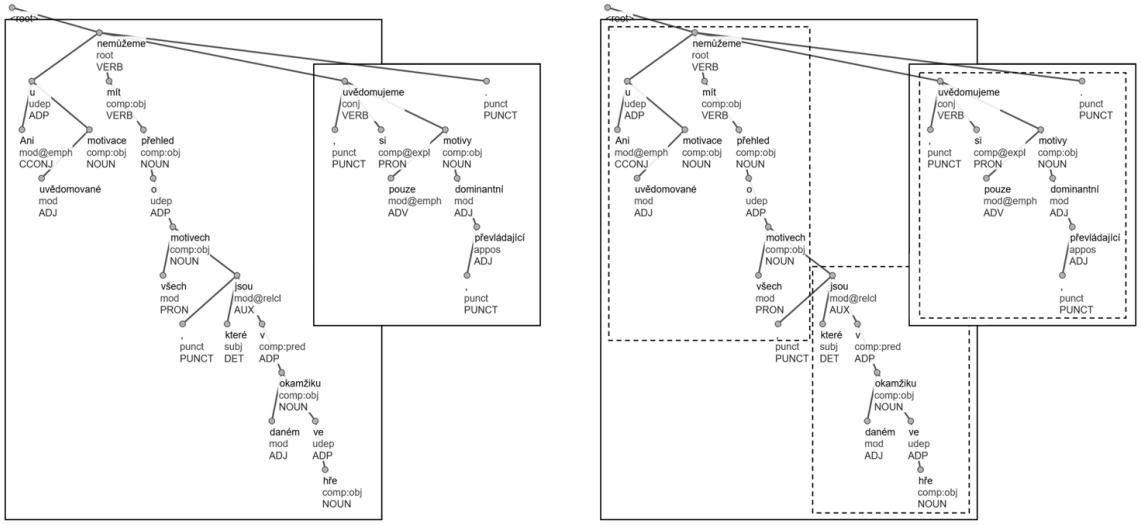


Figure 1: Dependency tree of sentence (1).

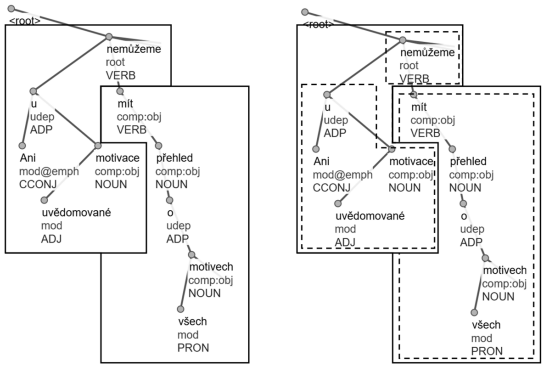


Figure 2: Phrase and subphrase structure of the first clause in sentence (1).

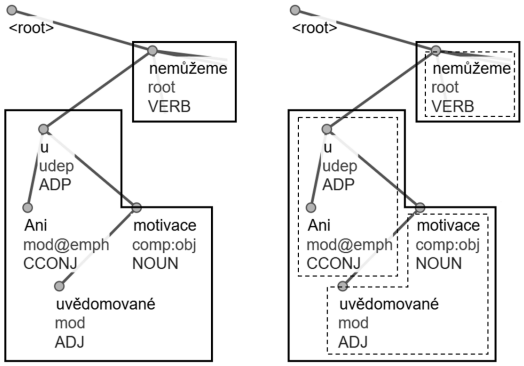


Figure 3: Chunks in both subphrases of the first phrase.

being, they do not have a linguistic interpretation. Nogolová et al. (2025) is a pilot study which shows that the MAL can indeed be modelled across many levels of the linguistic unit hierarchy. These units thus require investigations also from other points of view. This study focuses on modelling of their lengths.

4 Results

The hyper-Poisson distribution (Wimmer and Altmann, 1999, pp. 281-282) is defined as

$$P_x = \frac{a^x}{{}_1F_1(1;a;b)b^{(x)}}, \quad x = 0, 1, \ldots \qquad (1)$$

with $a \geq 0$ and $b > 0$ being its parameters[6]; ${}_1F_1(1;a;b)$ is a hypergeometric function (see e.g. Gasper and Rahman, 1990). However, distribution (1) is defined on the set of all non-negative integers, while our data attain values from a finite set (no length exceeds the value of 13), and, moreover, it attains also the value of 0, while the lowest value of length in our data is 1. Therefore, we use a modification of distribution (1), namely, its right-truncated version shifted to the right by 1, with

$$P_x = C\frac{a^{x-1}}{b^{(x-1)}}, \quad x = 1, \ldots, n, \qquad (2)$$

as a model for length distributions of the units from Section 3. The value of the parameter n is determined as the highest observed length value in the

[6]Symbol $b^{(x)}$ denotes the rising factorial, i.e., $b^{(x)} = \frac{\Gamma(b+x)}{\Gamma(b)}$.

data. For the normalization constant C it holds $C = \left(\sum_{x=1}^{n} \frac{a^{x-1}}{b^{(x-1)}}\right)^{-1}$, i.e., C is not an independent parameter, but its value depends on the values of the parameters $a \geq 0$ and $b > 0$; .

We express the goodness-of-fit of the model in terms of the determination coefficient R^2, with $R^2 \geq 0.9$ indicating a satisfactory fit (see Mačutek and Wimmer, 2013). We created a simple script in statistical software environment R[7] to fit this dataset. Estimated values of the parameters and the resulting determination coefficients are presented in Tables 2 and 3, and visualized in Figure 4.

One can see that the hyper-Poisson distribution fits all the data sufficiently well (all values of the determination coefficient are at least 0.9927). We remind the reader that the hyper-Poisson distribution is one of standard mathematical models for length of linguistic units. The Poisson distribution and its modifications and generalizations (including the hyper-Poisson distribution) and its applications to word length modelling can be found e.g. in Grzybek (2006) and Popescu et al. (2013). Thus, one can say that length of syntactic units (as they are defined in Section 3) behaves in the same way as word length. In addition, the hyper-Poisson distribution is a special case of a very general model of the linguistic theory from Wimmer and Altmann (2005). Therefore, although some of the new units used are waiting for their linguistic interpretation, they at least display a behaviour analogous to the units that are well established.

While the parameters of the hyper-Poisson distribution do not vary systematically between levels, there is a strong hint that their ratios b/a depend on the empirical repeat rate RR defined as $RR = \sum_{k=1}^{n} r_k^2$, with r_k being the relative frequency of length k, see Herdan (1962, pp. 36–40) and Altmann and Lehfeldt (1980, pp. 151–166). The repeat rate is a measure of diversity (or, from the opposite point of view, of uniformity) of the observed distribution. It can attain values from $1/n$ to 1. In the context of this paper, the value of $1/n$ corresponds to the same frequency of all lengths from 1 to n, whereas the value of 1 characterizes the deterministic distribution, i.e., all item have the same length. The values of the ratios b/a and the repeat rates RR for particular levels can be found in Table 4 and Figure 5; they are very strongly correlated, with the Pearson correlation coefficient of 0.973. This highly regular behaviour[8] opens a possibility of the interpretation of parameters in future research.

[7] https://www.r-project.org/

5 Conclusion

Analysis of the lengths of sentences, independent clauses, clauses, phrases, subphrases, chunks, and words revealed that these syntactic units exhibit length distribution patterns comparable to those found in more traditional linguistic units (especially in words). In each case, the length distribution can be modelled by the hyper-Poisson distribution, which has often been used as a mathematical model for (especially, but not only) word length. Moreover, based on Nogolová et al. (2025), these units consistently conform to the Menzerath–Altmann law across all structural levels, from sentences to syllables. Thus, they fit into the framework of interconnected linguistic units in which units and their properties are not isolated, but, rather, they influence each other (Köhler, 2005).

The results are promising, but we are aware of the fact that they are tentative only, as the only language from which we took data is Czech. Several important questions remain. These include the extent to which the findings can be generalized for other languages, as well as the potential influence of text genre and annotation schemes on the observed patterns. Further linguistic research will therefore be essential to determine the generality of the findings.

Acknowledgments

This research was supported by Moravian-Silesian Region, grant programme "Podpora vědy a výzkumu v Moravskoslezském kraji 2023", project "Podpora talentovaných studentů doktorského studia na Ostravské univerzitě VII", project No. 00516/2024/RRC (M. Nogolová), by projects APVV-21-0216 and VEGA 2/0120/24 (M. Koščová), by EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00748 (J. Mačutek), and by project "Lexikon a gramatika češtiny V - 2025 (MUNI/A/1445/2024)" (R. Čech).

[8] While the estimated parameter values depend also on the numerical estimation procedures, the ratio of b/a remains almost constant regardless of the method chosen.

length	**sentence in independent clauses**		**independent clause in clauses**		**clause in phrases**		**phrase in subphrases**	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	103 667	104 238.08	110 722	110 854.54	40 171	39 335.76	550 082	519 882.35
2	23 432	22 289.68	39 948	39 700.70	92 978	97 718.61	45 955	78 425.54
3	4 026	4 544.88	12 309	12 248.36	74 122	69 258.26	11 447	11 785.66
4	763	885.58	3 092	3 318.82	29 518	28 628.04	3 269	1 764.41
5	198	165.22	803	801.67	7 319	8 352.09	951	263.15
6	46	29.57	200	174.69	1 308	1 882.79	301	39.10
7	19	5.08	42	34.67	143	345.82	103	5.79
8	4	0.84	12	6.32	23	53.59	36	0.85
9	2	0.13	3	1.06	1	7.18	14	0.13
10	0	0.02	1	0.17	1	0.85	7	0.02
11	0	< 0.01					2	< 0.01
12	1	< 0.01						
13	1	< 0.01						
a	4.39		2.26		0.99		39.488	
b	20.53		6.21		0.40		261.72	
R^2	0.9998		> 0.9999		0.9954		0.9927	

Table 2: Fitting length frequencies by the hyper-Poisson distribution.

length	**subphrase in chunks**		**chunk in words**		**word in syllables**	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	438 584	444 369.74	591 655	591 539.21	471 587	461 580.80
2	183 484	167 238.90	396 848	392 045.32	487 216	514 703.06
3	48 494	58 448.27	51 432	58 364.50	345 045	333 384.51
4	17 746	19 066.26	6 495	4 894.08	168 721	152 164.10
5	5 936	5 831.10	649	285.64	49 843	53 616.00
6	1 994	1 678.51	58	12.768	10 544	15 384.27
7	658	456.34	5	0.46	2 065	3 723.02
8	215	117.54			362	778.99
9	65	28.76			96	143.56
10	34	6.70			19	23.643
11	5	1.49			4	3.52
12	1	0.32			3	0.48
13	1	0.06			1	0.06
a	4.90		0.19		1.55	
b	13.01		0.29		1.39	
R^2	0.9979		0.9998		0.9970	

Table 3: Fitting length frequencies by the hyper-Poisson distribution.

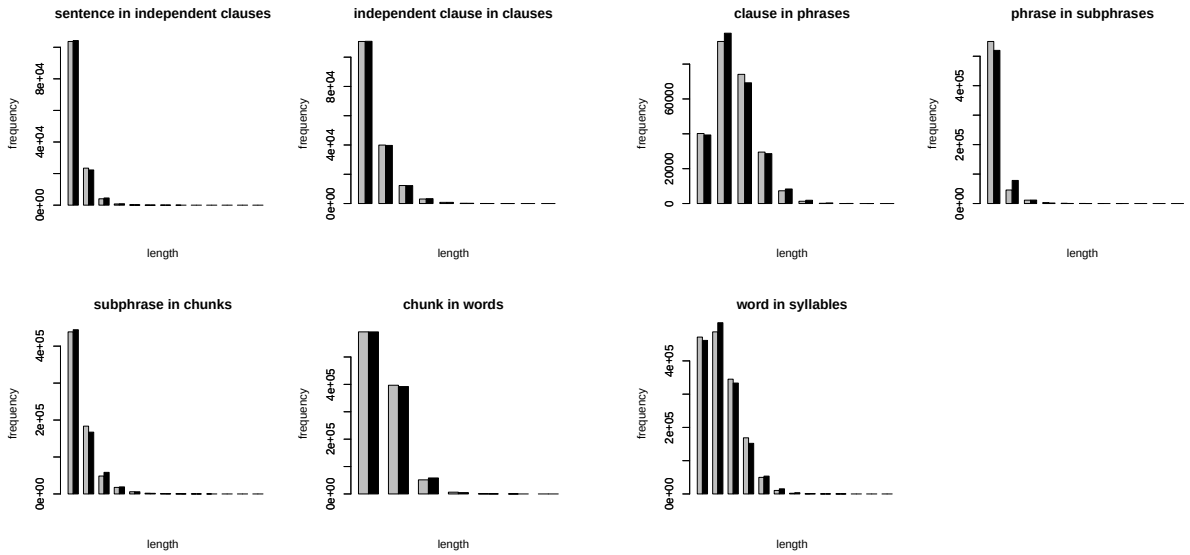


Figure 4: Fitting length frequencies by the hyper-Poisson distribution.

Unit	b/a	RR
sentence	4.677	0.648
independent clause	2.792	0.502
clause	0.403	0.277
phrase	6.629	0.813
subphrase	2.655	0.470
chunk	1.526	0.465
word	0.897	0.259

Table 4: Values of the repeat rate and of the ratio b/a.

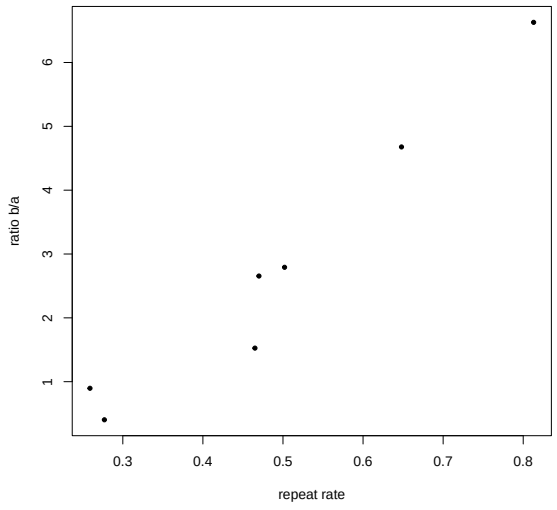


Figure 5: The relationship between the repeat rate and the ratio b/a.

References

Gabriel Altmann. 1980. Prolegomena to Menzerath's law. In Rüdiger Grotjahn (ed.), *Glottometrika 2*, pages 1–10. Brockmeyer, Bochum.

Gabriel Altmann and Werner Lehfeldt. 1980. *Einführung in die Quantitative Phonologie*. Brockmeyer, Bochum.

Mark Anderson, David Vilares, and Carlos Gómez-Rodríguez. 2019. Artificially evolved chunks for morphosyntactic analysis. In Marie Candito, Kilian Evang, Stephan Oepen and Djamé Seddah (eds.), *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 133–143. ACL, Paris.

Jan Andres and Martina Benešová. 2012. Fractal Analysis of Poe's Raven, II. *Journal of Quantitative Linguistics*, 19(4):301–324.

Eduard Bejček et al. 2013. Prague dependency treebank 3.0. *LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics*, http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3.

Aleksandrs Berdicevskis. 2021. Successes and failures of Menzerath's law at the syntactic level. In Radek Čech and Xinying Chen (eds.), *Proceedings of the second workshop on quantitative syntax* (Quasy, SyntaxFest 2021), pages 17–32. ACL, Sofia.

George Gasper and Mizan Rahman. 1990. *Basic Hypergeometric Series*. Cambridge University Press, Cambridge.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or Surface-syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In Marie-Catherine de Marneffe,

Teresa Lynn and Sebastian Schuster (eds.), *Proceedings of the second workshop on universal dependencies* (UDW 2018), pages 66–74. ACL, Brussels.

Peter Grzybek. 2006. History and methodology of word length studies. In Peter Grzybek (ed.), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*, pages 15–90. Springer, Dordrecht.

Bruno Guillaume. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In Dimitra Gkatzia and Djamé Seddah (eds.), *Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: System demonstrations*, pages 168–175. ACL, online.

Jan Hajič, Eva Hajičová, Marie Mikulová, and Jiří Mírovský. 2017. An introduction to word grammar. In Nancy Ide and James Pustejovsky (eds.), *Handbook of Linguistic Annotation*, pages 555–594. Springer, Dordrecht.

Gustav Herdan. 1962. *The Calculus of Linguistic Observations*. Mouton, The Hague.

Richard Hudson. 2010. *An Introduction to Word Grammar*. Cambridge University Press, Cambridge.

Tomáš Jelínek. 2017. FicTree: A manually annotated treebank of Czech fiction. In Jaroslava Hlaváčová (ed.), *ITAT 2017 Proceedings: Information technologies – applications and theory: Conference on theory and practice of information technologies*, pages 181–185. CreateSpace Independent Publishing Platform, Aachen, Technical University & Charleston.

Reinhard Köhler. 2005. Synergetic linguistics. In Reinhard Köhler, Gabriel Altmann and Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook*, pages 760–774. de Gruyter, Berlin.

Vincent Kríž and Barbora Hladká. 2017. Czech Legal Text Treebank 2.0,. In *LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics*. Charles University, Prague. Http://hdl.handle.net/11234/1-2498.

Miroslav Kubát, Ján Mačutek, and Radek Čech. 2021. Communists spoke differently: An analysis of Czechoslovak and Czech annual presidential speeches. *Digital Scholarship in the Humanities*, 36(1):138–152.

Michal Křen et al. 2020. *SYN2020: reprezentativní korpus psané češtiny*. Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague. Http://www.korpus.cz/.

Ján Mačutek and Gejza Wimmer. 2013. Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics*, 20(3):227–240.

Ján Mačutek, Radek Čech, and Marine Courtin. 2021. The Menzerath-Altmann law in syntactic structure revisited: Combining linearity of language with dependency syntax. In Radek Čech and Xinying Chen (eds.), *Proceedings of the second workshop on quantitative syntax (Quasy, SyntaxFest 2021)*, pages 65–73. ACL, Sofia.

Ján Mačutek, Radek Čech, and Jiří Milička. 2017. Menzerath-Altmann law in syntactic dependency structure. In Simonetta Montemagni and Joakim Nivre (eds.), *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 100—-107. Linköping University Electronic Press, Linköping.

Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. SUNY Press, Albany, NY.

Paul Menzerath. 1954. *Die Architektonik des deutschen Wortschatzes*. Dümmler, Bonn.

Michaela Nogolová, Ján Mačutek, and Radek Čech. 2025. The Menzerath-Altmann law: From sentence to phoneme. *Journal of Quantitative Linguistics*, submitted paper.

Timothy Osborne. 2019. *A Dependency Grammar of English: An Introduction and beyond*. Benjamins, Amsterdam.

Petr Plecháč and Robert Kolár. 2015. The corpus of Czech verse. *Studia Metrica et Poetica*, 2(1):107–118.

Ioan-Iovitz Popescu, Sven Naumann, Emmerich Kelih, Andrij Rovenchak, Anja Overbeck, Haruko Sanada, Reginald Smith abnd Panchanan Mohanty, Andrew Wilson, and Gabriel Altmann. 2013. Word length: aspects and languages. In Reinhard Köhler and Gabriel Altmann (eds.), *Issues in Quantitative Linguistics 3*, pages 224–281. RAM-Verlag, Lüdenscheid.

Haruko Sanada. 2016. The Menzerath–Altmann law and sentence structure. *Journal of Quantitative Linguistics*, 23(3):256–277.

Milan Straka. 2018. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning*, pages 197–207. ACL, Stroudsburg, PA.

Barbora Vidová Hladká, Jan Hajič, Jirka Hana, Jaroslava Hlaváčová, Jiří Mírovský, and Jan Raab. 2008. The Czech Academic Corpus 2.0 Guide. *The Prague Bulletin of Mathematical Linguistics*, 89:41–96.

Gejza Wimmer and Gabriel Altmann. 1999. *Thesaurus of univariate discrete probability distributions*. Stamm, Essen.

Gejza Wimmer and Gabriel Altmann. 2005. Unified derivation of some linguistic laws. In Reinhard Köhler, Gabriel Altmann and Rajmund G. Piotrowski (eds.) *Quantitative Linguistics. An International Handbook*, pages 791–807. de Gruyter, Berlin.

Daniel Zeman et al. 2023. Universal Dependencies 2.13. *LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics*, https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-5287.

Do Multilingual Transformers Encode Paninian Grammatical Relations? A Layer-wise Probing Study

Akshit Kumar, Dipti Sharma, Parameswari Krishnamurthy
LTRC, International Institute of Information Technology, Hyderabad, India
akshit.kumar@research.iiit.ac.in
{dipti, param.krishna}@iiit.ac.in

Abstract

Large multilingual transformers such as XLM-RoBERTa achieve impressive performance on diverse NLP benchmarks, but understanding how they internally encode grammatical information remains challenging. This study investigates the encoding of syntactic and morphological information derived from the Paninian grammatical framework—specifically designed for morphologically rich Indian languages—across model layers. Using diagnostic probing, we analyze the hidden representations of frozen XLM-RoBERTa-base, mBERT, and IndicBERT-v2 models across seven Indian languages (Hindi, Kannada, Malayalam, Marathi, Telugu, Urdu, Bengali). Probes are trained to predict Paninian dependency relations (by edge probing) and essential morphosyntactic features (UPOS tags, Vibhakti markers). We find that syntactic structure (dependencies) is primarily encoded in the middle-to-upper-middle layers (layers 6–9), while lexical features peak slightly earlier. Although the general layer-wise trends are shared across models, significant variations in absolute probing performance reflect differences in model capacity, pre-training data, and language-specific characteristics. These findings shed light on how theory-specific grammatical information emerges implicitly within multilingual transformer representations trained largely on unstructured raw text.

1 Introduction

Multilingual pre-trained transformer models such as XLM-RoBERTa (Conneau et al., 2019) have become foundational in natural language processing (NLP), achieving remarkable cross-lingual transfer capabilities. However, their internal representations remain only partially understood. Identifying where and how linguistic structures are encoded across the layers of these transformers improves their interpretability, efficiency, and adaptability across diverse languages and linguistic tasks (Belinkov and Glass, 2019).

A fundamental question is whether multilingual transformer models implicitly encode theory-specific grammatical information despite not being explicitly trained with linguistic annotations grounded in such formalisms.

Much prior work has analyzed transformer representations using probing tasks based mainly on Universal Dependencies (UD) (Hewitt and Liang, 2019; Tenney et al., 2019). While valuable, the UD framework aims for cross-lingual consistency and potentially abstracts away from linguistic phenomena better captured by language-specific or theory-specific grammars. Indian languages (IL), characterized by rich morphology and relatively free word order, are often analyzed using frameworks derived from the classical Paninian grammatical tradition (Bharati et al., 1995). This tradition introduces distinctive linguistic concepts such as Kārakas (syntactico-semantic roles) and emphasizes morphological cues such as Vibhakti markers as fundamental in signaling syntactic structures.

A detailed overview of the specific Paninian annotation scheme (Begum et al., 2007) relevant to this work is provided in Appendix A.

In this work, we carry out the first Paninian analysis of multilingual transformer representations across layers. We probe the hidden layers of XLM-RoBERTa, IndicBERT-v2, and mBERT across seven Indian languages annotated under a shared Paninian annotation scheme (Begum et al., 2007). To measure these linguistic representations, we employ edge probing (Tenney et al., 2019) for syntactic dependency structures and token-classification probes for morphosyntactic features (Universal POS, Vibhakti markers).

Specifically, the paper aims to answer the following research questions clearly:

1. Do multilingual transformer models implicitly

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 124–130

represent Paninian dependency structures and morphosyntactic information?

2. If such representations exist, at which layer do they emerge distinctly?

3. How consistent are these observed patterns across different multilingual transformer architectures and different Indian languages?

Our key findings can be summarized as follows:

1. Paninian dependency structures are indeed implicitly encoded, predominantly emerging in middle layers of multilingual transformers (layers 6–9), in line with prior UD-based studies.

2. Lexical-level morphosyntactic signals (UPOS, Vibhakti markers) peak slightly earlier in the layers, highlighting differentiated storage of structural versus lexical linguistic information.

3. Despite shared general trends, we observe substantial variations across different architectures and languages, reflecting cross-linguistic diversity, data differences, and model-specific factors.

2 Experimental Setup

To investigate and compare the layer-wise encoding of Paninian grammar in different multilingual transformers, we conducted probing experiments using three base models: XLM-RoBERTa-base (Conneau et al., 2019), Multilingual BERT (Devlin et al., 2019) and IndicBERT-v2 (Doddapaneni et al., 2023). For each base model, parameters were kept frozen[1], and we extracted hidden states from all layers (including embeddings; Layer 0 up to Layer 12 for base models). Our analysis covers seven Indian languages: Hindi (hi), Kannada (ka), Malayalam (ml), Marathi (mr), Telugu (te), Urdu (ur), and Bengali (be). The datasets utilize annotations following an extended version of the Paninian dependency annotation scheme proposed by Begum et al. (2007).

Probing Tasks and Models: We designed diagnostic classifiers (Tenney et al., 2019) for three distinct tasks reflecting key aspects of the formalism. Following best practices for probing (Hewitt and Liang, 2019), we aimed for low-capacity probes:

- **Dependency Relations (Edge Probing):** To assess structural syntactic encoding, we trained probes using a biaffine attention/classifier architecture (Dozat and Manning, 2017). This probe incorporates non-linearity through intermediate single-layer MLPs (dimensionality 256, GELU activation) applied to the input hidden states before the biaffine transformations for predicting heads and relations.

- **UPOS Tags:** To capture basic lexical categories, we trained strictly linear classifier probes mapping the hidden state directly to tag logits.

- **Vibhakti Features:** Similarly, we used linear classifier probes to predict grammatical case/postposition markers. Vibhakti labels were extracted from the FEATS column based on observed patterns; non-applicable tokens were ignored during accuracy calculation.

Separate probe models were trained for each task, base model, language, and layer combination.

Data Handling: Input sequences were processed using the respective tokenizer for each base model (XLM-R, mBERT, IndicBERT-v2). Sequences exceeding a tokenized length of 128 tokens were filtered out. Remaining sequences were padded/truncated to 128 tokens. Target labels were aligned to the first sub-word token, with others ignored. Detailed dataset statistics are provided in Appendix B.

Training and Evaluation: Probes were trained independently for 3 epochs using the AdamW optimizer (Loshchilov and Hutter, 2017) with CrossEntropyLoss. Learning rates and batch sizes were tuned per model and task type based on preliminary experiments (Edge Probes: LR=1e-4, BS=64; Token Probes: LR=1.5e-3, BS=256). The model state yielding the best validation performance (LAS for dependencies, Accuracy for features) was selected. Evaluations were performed on held-out validation sets.

3 Results

We evaluated the performance of probes trained on each layer's representation across the three base models (XLM-RoBERTa, mBERT, IndicBERT-v2),

[1]Only probe parameters were trained (Alain and Bengio, 2016; Hewitt and Liang, 2019).

seven languages, and three probing tasks (Dependency Relations, UPOS, Vibhakti). Key findings are presented below.

Dependency Relations (LAS): Figure 1 presents the best Labeled Attachment Score (LAS) achieved by dependency probes trained on each layer (0-12) for the three base models across the seven Indian languages. A consistent trend across all three models and most languages is the distribution of syntactic information across layers. LAS is low at the initial embedding layer (L0), increases through the lower layers, generally peaks in the middle-to-upper-middle layers (typically Layers 7-9), and then declines towards the final layer (L12). This confirms the widely observed phenomenon that structural syntactic information is most saliently represented in the intermediate representations of transformer models (Jawahar et al., 2019; Tenney et al., 2019).

Model-wise, XLM-RoBERTa (Fig. 1a) and mBERT (Fig. 1b) achieve comparable peak performance on higher-resource languages like Hindi (peak LAS 61.6% for XLM-R at L8; 59.1% for mBERT at L7) and Urdu (peak LAS 48.1% for XLM-R; 47.9% for mBERT, both at L7). IndicBERT-v2 (Fig. 1c), despite its Indic pretraining, does not consistently outperform the general multilingual models on this task for all languages. While reaching a high peak for Hindi (62.1% at L6), its LAS scores for several other languages (ka, ml, be, mr, te) are often lower than those achieved by XLM-R or mBERT, particularly in the upper layers. IndicBERT-v2 shows earlier peaks in several languages (e.g., L6 for Hindi compared to L8 and L7 for XLM-R and mBERT respectively).

UPOS Tagging (Accuracy): Figure 2 illustrates the layer-wise accuracy for predicting Universal Part-of-Speech (UPOS) tags. Generally, UPOS tagging accuracy is high, significantly exceeding LAS scores, confirming that basic lexical category information is more readily extractable. Performance tends to peak relatively early compared to LAS, often plateauing across several lower-to-middle layers (e.g., Layers 3-9 for many languages in XLM-R and mBERT) before potentially declining slightly in the final layers.

Model-wise, XLM-RoBERTa and mBERT show strong results, with peak accuracies often reaching 85-90%+ for languages like Hindi and Kannada. IndicBERT-v2 generally achieves somewhat lower peak accuracies on this task compared to the other two models (e.g., peaking around 75% for Hindi).

Vibhakti Feature Prediction (Accuracy): The accuracy for predicting Vibhakti features, a key morphological cue in Paninian grammar, is presented in Figure 3. Overall accuracy for this task is generally higher than LAS but can be lower than UPOS accuracy. Hindi and Urdu consistently yield the highest accuracies, frequently exceeding 80-90% across several layers in XLM-RoBERTa and mBERT, and maintaining high accuracy across almost all layers in IndicBERT-v2.

Compared to the sharper LAS peaks, strong Vibhakti prediction performance often extends across a broader range of middle and sometimes upper layers (e.g., Layers 2-10 for Hindi/Urdu in XLM-R/mBERT). The optimal layers for Vibhakti tend to overlap with or slightly precede the peak layers for dependency relations. Among the models, IndicBERT-v2 demonstrates particularly strong and stable Vibhakti prediction for Hindi and Urdu across nearly all layers.

Cross-Lingual Variation: Performance varies markedly across languages within each model (Figures 1, 2 and 3). While a comprehensive cross-lingual performance comparison is complicated by factors such as differing script representations and varying vocabulary coverage for each language within the pre-trained models, clear tiers of performance emerge. Hindi and Urdu consistently show the strongest LAS results, suggesting the models capture their Paninian structures relatively effectively. Kannada, Malayalam, Bengali, and Marathi form a mid-tier group, with peak LAS typically ranging from ~20% to ~40% depending on the model and language. Telugu consistently exhibits the lowest LAS scores across all models and layers (peak < 10% for XLM-R/mBERT, < 5% for IndicBERT-v2), a finding that strongly correlates with its significantly smaller probing dataset size (Appendix B) and potential underlying data sparsity in the models' pre-training.

4 Analysis and Discussion

Our multi-model, multi-task probing experiments reveal consistent layer-wise patterns for encoding Paninian grammar, alongside notable performance variations.

Layer Specialization for Paninian Grammar: Across all three models, a functional specialization

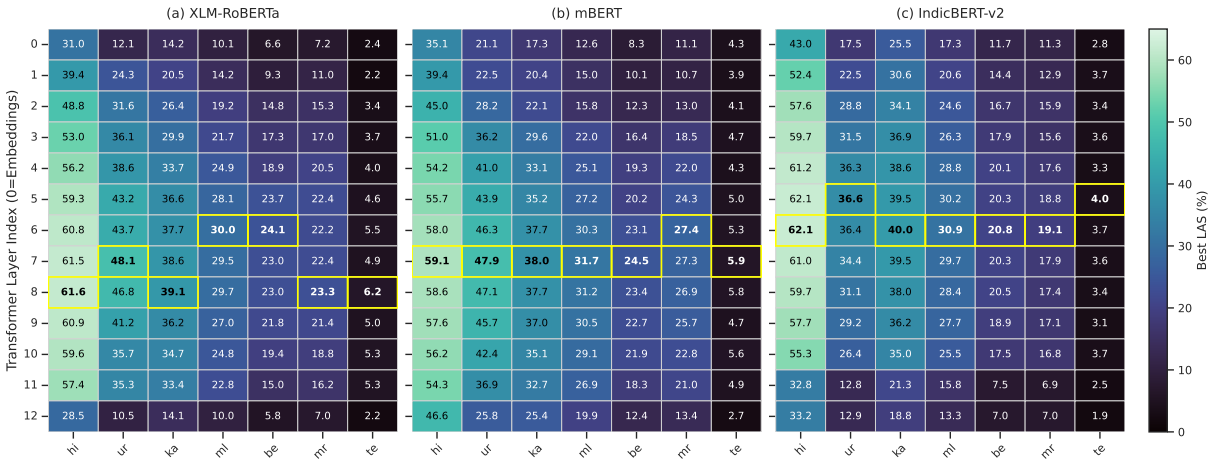


Figure 1: Layer-wise Dependency LAS (%) across Languages for (a) XLM-RoBERTa, (b) mBERT, and (c) IndicBERT-v2. Brighter colors indicate higher LAS. Yellow boxes highlight approximate peak performance regions for selected languages.

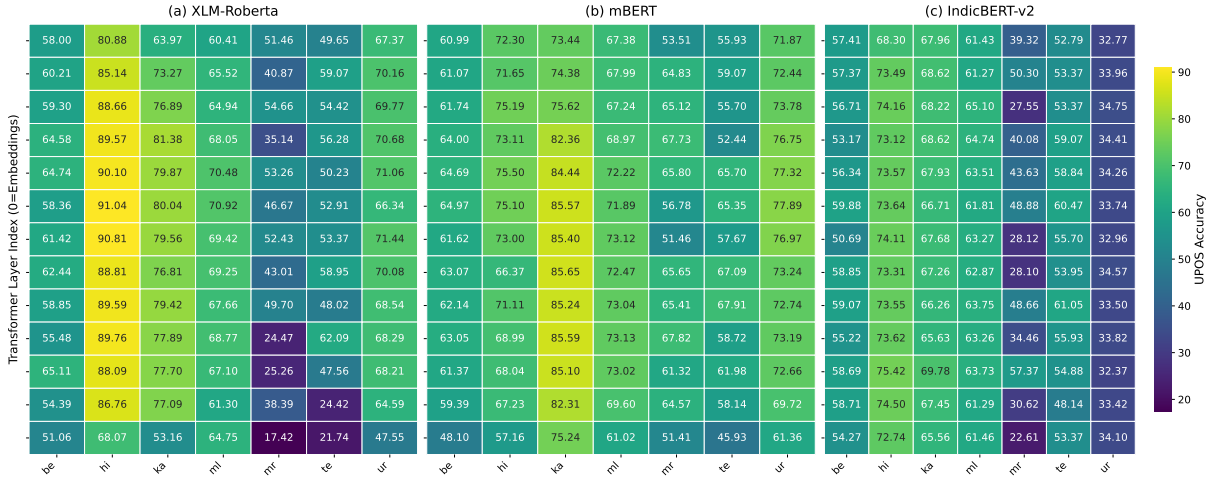


Figure 2: Layer-wise UPOS Accuracy (%) across Languages for (a) XLM-RoBERTa, (b) mBERT, and (c) IndicBERT-v2. Brighter colors indicate higher accuracy.

of layers aligns with prior probing studies on formalisms like Universal Dependencies (UD) (Jawahar et al., 2019; Tenney et al., 2019). Basic lexical information (UPOS tags) becomes accurately predictable in lower-to-middle layers. Morphological features (Vibhakti) also show strong representation across middle layers, their optimal encoding often overlapping with or slightly preceding layers most informative for syntactic structure. Crucially, complex Paninian dependency relations (LAS) consistently peak later, in the upper-middle layers (7-9), suggesting a hierarchical process where models integrate lexical/morphological cues to build syntactic representations. Performance generally degrades in final layers, possibly as representations specialize towards pre-training objectives.

Model Architectures and Pre-training Influence: While layer-wise trends are broadly similar, absolute performance and peak locations vary across models. XLM-RoBERTa and mBERT show comparable LAS capabilities on higher-resource languages like Hindi and Urdu. IndicBERT-v2, despite its Indic-focused pre-training, does not uniformly outperform general multilingual models in LAS for all tested languages, though it achieves strong LAS for Hindi. However, IndicBERT-v2 excels in stable Vibhakti prediction for Hindi/Urdu across most layers, potentially reflecting better morphological encoding due to its specialized training. This nuanced behavior suggests language-family specific pre-training can enhance morphological representation, but generalization to complex syn-

Figure 3: Layer-wise Vibhakti Accuracy (%) across Languages for (a) XLM-RoBERTa, (b) mBERT, and (c) IndicBERT-v2. Brighter colors indicate higher accuracy.

tax across diverse languages within that family remains challenging and interacts with other model properties.

Cross-Lingual Consistency and Variation: Significant cross-lingual variation in probing performance is evident for all tasks. While direct comparison is complicated by differing script representations and vocabulary coverage, Hindi and Urdu consistently yield the strongest results. The markedly lower performance for Telugu correlates with its smaller probing dataset size (Appendix B) and likely reflects underlying challenges from pre-training data sparsity or dataset quality. This underscores that probing performance reflects an interplay between information encoded by the base model and the characteristics of the probe training dataset.

Encoding Paninian-Specific Information: Our results demonstrate that diagnostic probes can successfully extract information pertinent to the Paninian grammatical framework – Kāraka-based dependency relations and Vibhakti features – from the frozen representations of these multilingual transformers. The ability to predict these suggests models implicitly learn representations sensitive to this distinct formalism, primarily consolidating structural knowledge in their middle layers.

5 Conclusion

We presented a layer-wise probing analysis comparing the encoding of Paninian grammatical information within XLM-RoBERTa, mBERT, and IndicBERT-v2 across seven Indian languages, examining dependency relations, UPOS tags, and Vibhakti features. Our findings reveal that Paninian dependency structure generally peaks in the upper-middle transformer layers, following lexical and morphological feature encoding in lower-to-middle layers, consistent with known patterns of linguistic representation.

Substantial cross-lingual and cross-model variations were observed. While IndicBERT-v2 showed strengths in Vibhakti prediction for core Indic languages, it did not uniformly surpass general multilingual models in representing Paninian dependency structures. Performance differences across languages correlate strongly with probing dataset sizes and likely reflect variations in pre-training data. Our results confirm that probing effectively reveals how theory-specific grammatical formalisms are represented within standard multilingual models.

6 Limitations

This study is limited to three base models and seven Indian languages; findings may not generalize broadly. Probe performance reflects both model representations and probe/dataset characteristics, with notable dataset size/quality variations (e.g., Telugu, Appendix B). The Paninian features probed (deprel, Vibhakti) are not exhaustive. Our analysis of frozen representations establishes correlations, not causal model mechanisms. Future work involves expanding model/language scope, probing more fine-grained Paninian features and studying the emergence of these representations.

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Rafiya Begum, Samar Husain, Arun Dhwaj, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. 2007. Dependency annotation scheme for indian languages. In *Proceedings of the Linguistic Annotation Workshop (LAW '07)*, pages 149–156, Prague, Czech Republic. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics (TACL)*, 7:49–72.

Akshar Bharati, Vineet Chaitanya, and Rajeev Sangal. 1995. *Natural Language Processing: A Paninian Perspective*. Prentice-Hall of India, New Delhi.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. *Preprint*, arXiv:2212.05409.

Timothy Dozat and Christopher D Manning. 2017. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations (ICLR)*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

A Background

Our probing analysis leverages the Paninian dependency annotation scheme, specifically developed for accurately representing linguistic phenomena within morphologically rich and relatively free word order ILs (Begum et al., 2007).

The Paninian grammatical tradition, originating from the ancient linguist Pāṇini, describes sentences primarily through modifier-modified dependency structures centered around the verb (Bharati et al., 1995). A crucial aspect of Paninian grammar is the use of Kārakas, specialized syntactico-semantic relations linking verbs and their arguments or modifiers. In the annotation scheme adopted here (Begum et al., 2007), six main Kārakas are identified:

1. *adhikaraṇa* (location)
2. *apādān* (source)
3. *sampradān* (recipient)
4. *karaṇa* (instrument)
5. *karma* (theme, loosely object-like)
6. *karta* (agent, loosely subject-like)

Importantly, Kārakas are not exact equivalents to purely semantic thematic roles (e.g., agent, patient). Instead, they encode a distinctly Paninian syntactico-semantic perspective. Consider for instance the English-like example 'key opened the door'. While semantically an instrument, the 'key' here would be annotated as the *karta* (loosely agent-like) Kāraka in the Paninian tradition (Bharati et al., 1995).

Identification of these Kārakas depends heavily on morphological indicators such as Vibhakti (case-endings, postpositions) and verb-based TAM markers (tense-aspect-modality) within sentences (Begum et al., 2007). The strong, systematic correlation between morphological features and syntactic dependency structures motivates our probing

approach: we probe multilingual transformer models' representations for both the structural Kāraka relations (via dependency links and edge labels) and essential morphological cues (*Vibhaktis*, UPOS tags), analyzing explicitly how these linguistic representations are distributed layer-wise.

A.1 Vibhakti

In the context of Paninian grammar and many modern Indian languages, *Vibhakti* refers to morphological markers, primarily case endings or postpositions, that are attached to nouns or noun phrases. These markers play a crucial role in signaling the grammatical function and syntactico-semantic role of the noun phrase within the sentence.

While often translated loosely as *case*, Vibhakti in the Paninian tradition is intimately linked to the concept of *Kārakas* (described in Appendix A). Specific Vibhaktis are typically associated with signaling specific Kāraka roles (e.g., a particular Vibhakti might commonly mark the *karta* 'agent-like' role, while another marks the *karma* 'theme-like' role, and others mark instrument, location, etc.). However, the mapping is not always one-to-one and can be influenced by other factors like verb semantics and sentence structure.

Essentially, Vibhaktis provide explicit surface cues about the underlying grammatical relationships in the sentence, making them particularly important in languages with relatively flexible word order where syntactic function is not solely determined by position.

B Dataset Statistics

Table 1 provides statistics for the annotated datasets used in our probing experiments. Counts reflect the data *after* filtering sequences longer than 128 tokens but *before* any potential subsetting for development runs. Please note that the Bengali annotated data does not include Vibhakti features.

Language		Sentences		Tokens	
Code	Name	train	val	train	val
be	Bengali	3939	488	99504	12389
hi	Hindi	19855	2478	570772	71333
ka	Kannada	10388	1297	263827	32910
ml	Malayalam	7109	882	178754	21873
mr	Marathi	3608	451	91077	11461
te	Telugu	1514	185	19663	2219
ur	Urdu	4871	607	164129	21606

Table 1: Statistics of the Paninian–annotated datasets used for probing (post-filtering, pre-subsetting).

Author Index