

DepLing 2025

**Eighth International Conference on Dependency Linguistics  
(Depling, SyntaxFest 2025)**

**Proceedings**

August 27-28, 2025

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

VITASIS



UniDive



Ljubljana Tourism



Mestna občina  
Ljubljana



Flanders  
State of the Art



**CJVT** Centre for  
Language Resources  
and Technologies



alpineon))

**AI4DH** CENTRE OF EXCELLENCE IN AI  
FOR DIGITAL HUMANITIES

**Organized by**



**As part of SyntaxFest 2025**



©2025 Association for Computational Linguistics

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](mailto:acl@aclweb.org)

ISBN 979-8-89176-290-9

## Introduction

The Eighth edition of the International Conference on Dependency Linguistics (DepLing) follows a biannual series that started in 2011, in Barcelona and continued in Prague (2013), Uppsala (2015), Pisa (2017), Paris (2019), Sofia (2021), and Washington DC (2023). The series responds to the growing need for linguistic meetings dedicated to approaches in syntax, semantics and the lexicon that are centered around dependency structures as a central linguistic notion. DepLing (2025) took place at SyntaxFest 2025 in Ljubljana, Slovenia, which brought 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 Catalunya  
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 Samardžić, University of Zurich  
Giuseppe Samo, Beijing Language and Culture University  
Haruko Sanada, Ritssho 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

# DepLing Keynote

## Auxiliaries across Languages and Frameworks

**Daniel Zeman**

Charles University, Prague



**Abstract:** In my talk, I will discuss the status of auxiliaries (i.e., auxiliary verbs as well as uninflected non-verbal particles with auxiliary function) in dependency treebanks. The focus will be on two frameworks, Universal Dependencies (UD) and the Prague family of treebanks, rooted in the Functional Generative Description. However, I will occasionally show examples from other treebanks and frameworks, encountered during the HamleDT harmonization effort.

Besides looking at various treatments of auxiliaries in different annotation schemes, I will also discuss the question of delimiting the set of auxiliaries in individual languages (or, more exactly, the set of words that receive the special treatment in the respective annotation schemes). Various grammatical tests may be available, but sometimes the auxiliaries are simply enumerated by traditional school grammar. Moreover, there is a scale of categories ranging from pure grammatical auxiliaries through modals and phase verbs to various semantically bleached verbs that take other verbs as complements, yet their contribution is lexical rather than grammatical and their syntactic behavior shows no anomalies. All these aspects complicate finding a unified definition that would be applicable in a multi-lingual dataset, such as HamleDT or UD.

In the last part of the talk, I will show some examples of contrastive cross-linguistic studies that would benefit from comparably defined auxiliaries. I will show how we encourage comparability in UD using a common database of auxiliaries, and I will argue that it has the potential to become a useful resource of its own.

**Bio:** Daniel Zeman is an associate professor of computational linguistics at the Charles University in Prague. He obtained his PhD (also from Charles University) in 2005 with a dissertation on statistical methods for syntactic parsing of Czech. He then worked on cross-lingual transfer techniques for low-resource languages, and led several projects focused on multilingual NLP and harmonization of linguistic resources, including Interset (for morphological tagsets) and HamleDT (for dependency treebanks). He is one of the founders and leading personalities of the Universal Dependencies initiative, and vice-chair of the COST Action “Universality, Diversity and Idiosyncrasy in Language Technology” (UniDive). His current work extends to harmonized datasets for coreference resolution (CorefUD) and Uniform Meaning Representation (UMR).

# Local SyntaxFest Keynote

## What we learn about syntax when dependencies fail: Experimental insights into syntactic locality constraints

Artur Stepanov  
University of Nova Gorica



**Abstract:** This talk examines a class of syntactic dependencies that cannot be formed: classic island violations (extraction from adjuncts, complex NPs, wh-islands etc.). I survey psycho- and neurolinguistic evidence quantifying the cognitive cost of breaching locality constraints, showing how these findings expose limits on dependency formation that remain invisible in standard treebanks yet are central to real-time sentence processing. I consider implications for parsing, dependency representations, and cross-linguistic variation, with suggestions for incorporating experimental diagnostics into syntactic annotation and parser-evaluation frameworks.

**Bio:** Artur Stepanov is a professor of psycholinguistics at the University of Nova Gorica. His work focuses on the cognitive representation and real-time processing of syntactic dependencies in monolingual and multilingual speakers, exploring how internal grammatical competence maps onto observable linguistic behavior. He combines psycholinguistic experimentation with insights from generative syntax, with particular emphasis on lesser-studied Slavic languages. He is involved in multiple international collaborations on projects related to sentence comprehension and production, the linguistic and cognitive dimensions of multilingualism, and, more recently, the compositional aspects of animal (marine mammal) vocalization sequences.

# Non-Archival Abstract

## Dependency Analysis of Chinese Comparative Sentences

Zexin Liu

Zhejiang University

This paper examines the dependency structures of comparative sentences across various Chinese dialects. In equality comparatives, the comparative result is post-posed (R-back) in all Chinese dialects, which contrasts with English. Although Mandarin also follows the R-back pattern for superiority comparatives, dialects such as Hong Kong Cantonese and Southern Min adopt an R-front type, similar to English. However, Southern Min lacks a comparative marker, while English's comparative marker *than* dominates the standard of comparison. In contrast, the comparative marker in Cantonese does not dominate the standard. Through the calculation of dependency distances and syn-tactic tests, we argue that when the comparative result is preposed, it dominates the standard of comparison. Conversely, when the comparative construction follows an R-back type, the comparative marker dominates the comparative result. This analysis aligns closely with the annotation choices of the Surface-Syntactic Universal Dependencies (SUD), differing significantly from those of the Universal Dependencies (UD) model.



# Non-Archival Abstract

## **A Quantitative Study of Subject-Predicate-Object Word Class Composition in vernacular Chinese Based on Dependency Grammar**

Bingli Liu<sup>1</sup> and Yiyi Zhao<sup>2</sup>

<sup>1</sup>Huaqiao University Quanzhou

<sup>2</sup>Xiamen University

The paper aims at studying the evolution of lexical composition of subject-verb-object sentences in vernacular Chinese. Five corpora are constructed for the Tang and Five Dynasties, Song Dynasty, Yuan and Ming Dynasties, Qing Dynasty, and the present contemporary era which lasts for more than 1,000 years. The syntactic structures of these sentences are labeled, counted, and analyzed based on the theoretical foundation of dependency grammar, with the aim of investigating the evolution of the lexical category composition of the subject-predicate-object in vernacular Chinese over time. The results show that the ratio of nouns and pronouns in each period occupies the majority of the total number of subject lexemes, and the lexical composition of predicates has been very stable since ancient times, with verbal predicates accounting for the vast majority of predicates. Compared with the subject lexical composition, objects are richer and the lexical composition changes more slowly.

## Table of Contents

<i>A Typology of Non-Projective Patterns in Unas's and Teti's Pyramid Texts</i> Roberto A. Diaz Hernandez .....	1
<i>Tracing Syntactic Complexity: Exploring the Evolution of Average Dependency Length Across Three Centuries of Scientific English</i> Marie-Pauline Krielke, Diego Alves and Luigi Talamo .....	13
<i>Modeling Syntactic Dependencies in Southern Dutch Dialects</i> Loic De Langhe, Jasper Degraeuwe, Melissa Farasyn and Veronique Hoste .....	24
<i>Assessing the Agreement Competence of Large Language Models</i> Alba Táboas García and Leo Wanner .....	36
<i>Introducing KIParla Forest: seeds for a UD annotation of interactional syntax</i> Ludovica Pannitto, Eleonora Zucchini, Silvia Ballarè, Cristina Bosco, Caterina Mauri and Manuela Sanguinetti .....	54
<i>Head-initial and head-Final coordinate structures in two annotation schemes of dependency grammar</i> Timothy John Osborne and Chenchen Song .....	74
<i>Genre Variation in Dependency Types: A Two-Level Genre Analysis Using the Czech National Corpus</i> Xinying Chen and Miroslav Kubát .....	84
<i>A morpheme-based treebank for Gbaya, an Ubanguian language of Central Africa</i> Paulette Roulon-Doko, Sylvain Kahane and Bruno Guillaume .....	93
<i>Dative alternations in less-researched syntactic patterns of standard Croatian</i> Matea Andrea Birtić, Siniša Runjaić and Robert Sviben .....	103
<i>Distance and Projectivity as Predictors of Sentence Acceptability in Free Word Order Languages</i> Kirill Chuprinko, Artem Novozhilov and Arthur Stepanov .....	108
<i>UD Annotation of Experience Clauses in Tigrinya</i> Michael Gasser and Nazareth Amlesom Kifle .....	120
<i>A corpus-driven description of OV order in Archaic Chinese</i> Qishen Wu, Santiago Herrera, Pierre Magistry and Sylvain Kahane .....	130
<i>Periphrastic Verb Forms in Universal Dependencies</i> Lenka Krippnerová and Daniel Zeman .....	140
<i>Word Order Variation in Spoken and Written Corpora: A Cross-Linguistic Study of SVO and Alternative Orders</i> Nives Hüll and Kaja Dobrovoljc .....	150

# A Typology of Non-Projective Patterns in Unas’s and Teti’s Pyramid Texts

**Roberto Antonio Díaz Hernández**  
Univeristy of Jaén (radiaz@ujaen.es)

## Abstract

The aim of this paper is to study the use of non-projective structures in Unas’s and Teti’s Pyramid Texts (ca. 2321–2279 BC) annotated in the Egyptian-UJaen treebank. It offers the first typology of non-projective patterns in Old Egyptian, and it discusses the causes for non-projectivity in the Old Egyptian language of Unas’s and Teti’s Pyramid Texts to conclude that non-projectivity is an exceptional phenomenon in these texts.

## 1 Introduction

The Egyptian-UJaen treebank in Universal Dependencies (hereafter UD-EUJA treebank) holds now 21,945 words and 2,192 sentences, most of them from Unas’s and Teti’s Pyramid Texts (ca. 2321–2279 BC) written in Old Egyptian (ca. 2700–2000 BC).<sup>1</sup> It allows the search for any morphosyntactic feature in those texts.

The analysis of non-projective structures in Unas’s and Teti’s Pyramid Texts is intended to shed light on the way they were formed and the morphosyntactic rules that govern their use. This will enable us to develop digital tools for the automatic translation of Egyptian texts.

This paper is divided into the following parts:

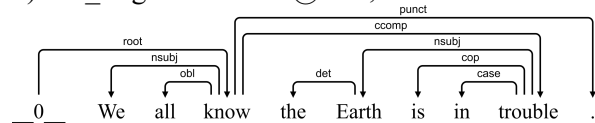
- A conceptual explanation of “projectivity” and “non-projectivity” in dependency grammar (2).
- A critical review of the analysis of “non-projectivity” in Egyptian philology (3).

- A typology of non-projective structures in Unas’s and Teti’s Pyramid Texts according to five patterns (4).
- A discussion of syntactic and pragmatic factors when dealing with non-projective structures in Unas’s and Teti’s Pyramid Texts (5).
- A conclusion (6).

## 2 Concept

“Projectivity” and “non-projectivity” are two key concepts coined in dependency grammar in the 1960s (Lecerf, Ihm, 1960, Hays, 1964, 519 and Marcus, 1965, 181–192). “Projectivity” is used as a label for a *continuous structure* whose dependents are close to their heads in word order (Osborne, 2019, 199 and 203). There are no intersections of connection lines in a dependency tree showing a projectivity structure (Hays, 1964, 519), for example:

1) UD\_English-ParTUT@2.15, id-sent 872:

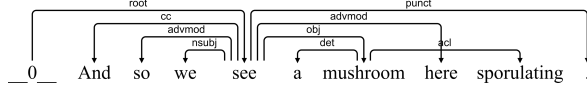


By contrast, non-projectivity refers to structures whose dependents are separated from their heads by one or more words causing a discontinuity. This results in crossing lines in a tree (Osborne, 2019, 213). Although non-projective structures arise from ungrammatical sentences, such as *\*Whose do you like answer?* (Groß and Osborne, 2009, 43), there are grammatically accepted non-projective sentences, for example:

<sup>1</sup> The Pyramid Texts have been edited by Sethe (1908–22) and by Allen (2013). Both works have been used for the annotation of the Pyramid Texts in the UD-EUJA treebank. The present paper is based on the latest

version of the treebank published in UD release 2.16 (May 2025). For a general description of the UD-EUJA treebank see Díaz Hernández and Passarotti 2024.

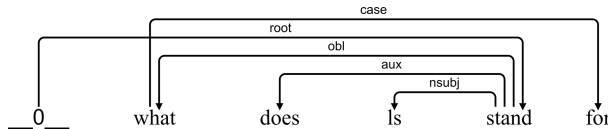
2) UD\_English-ParTUT@2.15, id-sent = 896:



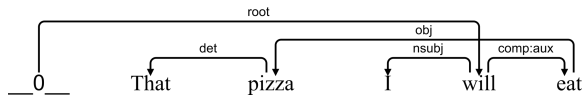
The continuous structure between “mushroom” (*head*) and “sporulating” (*dependent*) is broken by the adverb “here” governed by the verb “see”. This causes a discontinuity in the word order of sentence 2 and an intersection of lines in the tree.

Grammatical non-projective structures follow patterns governed by syntactic rules. However, it should be noted that non-projective patterns may vary according to the dependency theory applied to syntactic analysis. For example, there are three patterns of non-projectivity in English according to the traditional dependency grammar (Osborne, 2019, 204): *wh*-fronting (ex. 3), extraposition (ex. 2), and topicalization (ex. 4).

3) UD\_English-Atis@2.15, id-sent 0033.train; a *wh*-question and preposition at the end of a sentence:

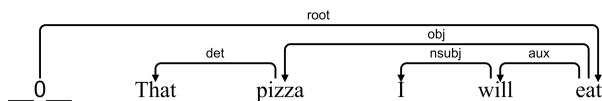


4) Topicalization:



However, according to the Universal Dependencies approach, the root of a verbal sentence is the verb (De Marneffe *et al.*, 2021, 257 and Nivre *et al.*, 2016, 1662), but not an auxiliary so that examples of topicalization such as 4 are considered to be projective structures:

5)



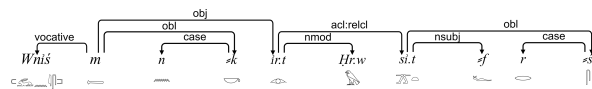
In this paper, only Egyptian structures considered to be non-projective according to the traditional dependency theory and Universal Dependencies have been selected for study. If a sentence is considered non-projective just in terms

of traditional dependency theory, it has been left out.

### 3 Non-projectivity in Egyptian philology

The issue of non-projectivity in Egyptian has been discussed so far only by Landgráfová, who, in her paper on the function of resumptive pronouns in Middle Egyptian, concluded that they are used in order to avoid non-projective structures, especially in relative forms (Landgráfová, 2002, 282). However, syntactic tree diagrams of relative clauses do not support such a conclusion. Given a sentence with the relative form *sl.t* “which went”:

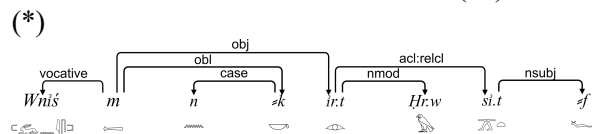
6) UD\_EUJA-150 = PT 31a W



LT:<sup>2</sup> “Unas (*Wniš*), take (*m*) for (*n*) yourself (*sk*) the eye (*ir:t*) of Horus (*Hr:w*) which-went (*sl.t*) he (*šf*) to (*r*) it (*šš*, referring to the eye).”

FT: “Unas, take the eye of Horus for which he went.”

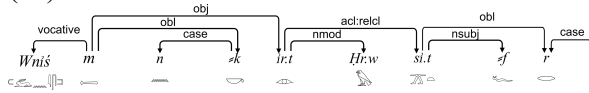
The word order of this sentence is clearly projective (fig. 1, cf. Landgráfová, 2002, 279, no. 35 and 36). It consists of the root *m* “take” governing *Wniš* as its vocative, *n sk* “for yourself” an adverbial phrase, and *ir:t* “eye” its direct object, followed by two modifiers—the name Horus (*Hr:w*) and the relative form *sl.t* “which went”. This relative form governs its own subject (*šf* “he”) and an adverbial phrase consisting of the preposition *r* “for” and the resumptive pronoun *šš*, which semantically refers to the antecedent of the relative form (*ir:t* “eye”), but is syntactically governed by the relative form (*sl.t*). The absence of that resumptive pronoun would not cause any discontinuity in the sentence, nor an intersection of connection lines—either the relative form would lack a prepositional phrase (\*) or the preposition to which it is attached would stand alone (\*\*):



<sup>2</sup> LT stands for “literal translation” and FT for “free translation”. The examples are annotated without glosses due to space limitations. The CoNLL-U file of

the UD-EUJA treebank contains the morphosyntactic annotation for each example: [https://github.com/UniversalDependencies/UD\\_Egyptian-UJaen/tree/master](https://github.com/UniversalDependencies/UD_Egyptian-UJaen/tree/master).

(\*\*)



In any case, the sentence would be syntactically ungrammatical, rather than non-projective. In spite of this, Landgráfová’s pioneering work is inspiring because it invites us to search for real types of non-projective structures in Egyptian texts.

#### 4 Non-projective patterns in Unas’s and Teti’s Pyramid Texts

The UD-EUJA treebank is available in GREW-MATCH for morphosyntactic queries.<sup>3</sup> The search for non-projective structures in the latest version of this treebank using GREW-MATCH yields 31 cases of non-projectivity, of which 17 are found in Unas’s Pyramid Texts and 10 in Teti’s Pyramid Texts<sup>4</sup> (see table in the Appendix). This represents 1.37 % of all sentences from the Unas’s Pyramid Texts and 1.42 % of all sentences from the Teti’s Pyramid Texts in the UD-EUJA treebank. Five types of non-projectivity can be distinguished:

- 1) Extraposition by inserting an adverbial phrase (EUJA-654, 871, 942, 1291, 1313, 1390, 1415, 1511, 1758, 1769, 1835, 2050, 2101, 2119).
- 2) Extraposition by inserting a vocative (EUJA-117, 981, 1406, 1753).
- 3) Extraposition by inserting a verb of utterance (EUJA-1455, 1507, 2038).
- 4) Extraposition of the nisba adjective *n(.t)* in the so-called “indirect genitive” (EUJA-442, 1290, 1294, 1338, 1614).
- 5) Discontinuity due to a dislocated element in a sentence with an emphatic subject (EUJA-385).

##### 4.1 Extraposition by inserting an adverbial phrase

This type is also found in English (see ex. 2, above). It is the most common type of non-projectivity in Unas’s and Teti’s Pyramid Texts. It consists of a noun phrase acting as a head, whose modifier is extraposed by an adverbial phrase linked to the root and inserted between the head

and its modifier. The adverbial phrase (AP) can be an oblique adjunct (obl, ex. 7, 8, 9, 10), an oblique argument (obl:arg, ex. 11 and 13) or a noun in adverbial function (obl, ex. 12). The extraposed modifier (EM) is usually an attribute in the form of an adjective (adj, ex. 7) including a nisba adjective (nadj, ex. 8),<sup>5</sup> a verb conjugated in the Old Semitic suffix conjugation and used in attributive function (OSSC, ex. 9), a participle (part, ex. 10) and a relative form (RF, ex. 11). A noun used in apposition (appos, ex. 12) or in a conjunct relation to its head (conj, ex. 13) can also be extraposed.

7) EUJA-1313 = Pyr. 424a, Unas, AP = obl and EM = adj (cf. fig. 1):

č(t) mṯw ik rr Wniš ‘n.t šf tn ir šk iḫ.t (...)  
 LT: “Saying (č(t)) a speech (mṯw): ‘Unas (Wniš) shall-dart (ik) indeed (rr) this (tn) left (iḫ.t) thumb-nail (‘n.t) of his (šf) against (ir) you (šk) (...)’”

FT: “Recitation: ‘Unas shall dart indeed this left thumb-nail of his against you (...)’”

The prepositional phrase *ir šk* “against you” is inserted between *‘n.t šf tn* “this thumb-nail” and its attribute *iḫ.t* “left”. Violation of the word order is due to the fact that *ir šk* is governed by the verb *ik*.

8) EUJA-2119 = Pyr. 734a–b, Teti, AP = obl and EM = nadj (cf. fig. 2):

(...) trč.t šk n šk im.t mnč(.wi) mw.t ʒš.t  
 LT: “(...) Your (šk) milk (trč.t) (is) for (n) you (šk) which-is-from (im.t) the breasts (mnč(.wi)) of Mother (mw.t), Isis (ʒš.t).”

FT: “(...) Your milk, which comes from the breasts of Mother Isis, is for you.”

The predicate of this adverbial sentence is *n šk* “for you”. The syntactic discontinuity is caused by the placement of the predicate between the subject (*trč.t šk*) and its nisba adjective (*im.t*).

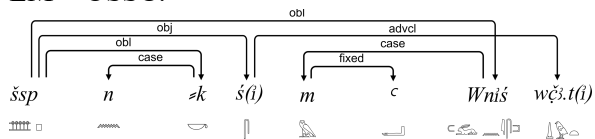
<sup>3</sup>Grewmatch (<https://universal.grew.fr/>) has proven to be an effective digital tool for studying non-projective structures in French, see Perrier, 2021, 41–42.

<sup>4</sup>The sentences EUJA-236, 441 and 729 are from Pepi II’s Pyramid Texts, which are not included in this study. The sentence EUJA-2172 contains an annotation error which causes a non-projectivity structure—the word čw.t.w is used as an apposition of h(n)k.t, rather

than *sp*. This error will be corrected in the next release of the treebank.

<sup>5</sup>In Semitic languages, such as Arabic, “nisba” is used to label an ending added to nouns, and rarely to prepositions and pronouns, to form (relative) adjectives and nouns (see Schulz 2010, 86). The addition of the nisba ending to prepositions to form adjectives and nouns is a common feature in Egyptian.

9) EUJA-1390 = Pyr. 451b–c, Unas, AP = obl and EM = OSSC:<sup>6</sup>



LT: “Accept (*šsp*) for (*n*) yourself (*k*) it (*š(i)*) from-the-hand-of (*m-*) Unas (*Wniš*) being intact (*wčš.t(i)*) (...)”

FT: “Accept it from Unas intact (...)”

The verb *wčš* is conjugated in the 3rd. f. sg. person of the OSSC and used in an attributive function. Its head is the dependent pronoun *š(i)* used as a direct object of the verb *šsp*. The prepositional phrase *m- Wniš* is inserted between the pronoun *š(i)* and *wčš.t(i)* and it causes a non-projective structure, as the prepositional phrase is governed by the root (*šsp*).

10) EUJA-1511 = Pyr. 500a, Unas, AP = obl and EM = part (cf. fig. 3):<sup>7</sup>

*ič k n k Wniš hn' k hn' k nf n k šp.t (...)*

LT: “You (*k*) shall-take (*ič*) Unas (*Wniš*) for (*n*) you (*k*), with (*hn*) you (*k*), with (*hn*) you (*k*), he-who-drives-away (*nf*) storms (*šp.t*) for (*n*) you (*k*) (...)”

FT: “You shall take Unas for you, with you and with you, he who drives away storms for you (...)”

The active present participle *nf* “he who drives away” is governed by the king’s name *Wniš* used as a direct object of the verb *ič*. Non-projectivity is caused by the double insertion of *hn' k* between *Wniš* and its participle, as *hn' k* is a prepositional phrase linked to the root of the sentence (*ič*).

11) EUJA-1758 = Pyr. 599b–c, Teti, AP = obl:arg and EM = RF (cf. fig. 4):

*(...) int š mhn.t tf n.t Mr-n(.i)-h<sup>3</sup> n Tti č<sup>33</sup>.t š nčr(.w) im š (...)*

LT: (...) that-he-may-bring (*int*) that (*tf*) boat (*mhn.t*) belonging-to (*n.t*) Merenkha (*Mr-n(.i)-h<sup>3</sup>*) to (*n*) Teti (*Tti*) which-ferries (*č<sup>33</sup>.t*) he (*š*) the gods (*nčr(.w)*) in (*im*) it (*š*) (...)”

FT: “(...) that he may bring to Teti that boat of Merenkha in which he ferries the gods (...)”

According to the Old Egyptian word order, the oblique argument (obl:arg) consisting of the preposition *n* plus a noun and used as an indirect object should follow the direct object of a sentence (Schenkel, 2012, 68–69). Example 11 shows that

this rule is obeyed even if the oblique argument (*n Tti* “to Teti”) causes a discontinuity between the direct object (*mhn.t* “boat”) and its attribute (*č<sup>33</sup>.t š* “on which he ferries”).

12) EUJA-1769 = Pyr. 606b–d, Teti, AP = obl and EM = appos (cf. fig. 5):

*(...) mr š*i*.t Nw.w ft.t iptw nčr.(w)t hrw š*i*.n š*n* hnt š*i*.t Nb.t-hw.t Ni.t Šrk.t-htw*

LT: “(...) As (*mr*) Nu (*Nw.w*) protected (*š*i*.t*) these (*iptw*) four (*ft.t*) goddesses (*nčr.(w)t*) the day (*hrw*) (that) they (*š*n**) protected (*š*i*.n*) the throne (*hnt*), Isis (*š*i*.t*), Nephthys (*Nb.t-hw.t*), Neith (*Ni.t*), Selqet-hetu (*Šrk.t-htw*).”

FT: “(...) As Nu protected these four goddesses on the day when they protected the throne, namely Isis, Nephthys, Neith, and Selket.”

*Nčr.(w)t* “goddesses” is used after the numeral *ft.t* “four” in apposition according to Old Egyptian grammatical rules (Schenkel, 2012, 121). The syntactic continuity is broken by inserting the noun *hrw* “day” used adverbially between *nčr.(w)t* and the names of the four goddesses Isis, Nephthys, Neith, and Selket.

13) EUJA-2050 = Pyr. 707a–b, Teti, AP = obl:arg and EM = conj (cf. fig. 6):

*in n k trč.t š*i*.t n Tti š*g*b.i Nb(.t)-hw.t (...)*

LT: “Bring (*in*) for (*n*) yourself (*k*) the milk (*trč.t*) of Isis (*š*i*.t*) for (*n*) Teti (*Tti*), the flood (*š*g*b.i*) of Nephthys (*Nb(.t)-hw.t*) (...)”

FT: “Bring to Teti, the milk of Isis, the flood of Nephthys (...)”

The insertion of *n Tti* “for Teti” between two noun phrases in a coordinate relation causes a non-projective structure. However, it is probably an error because *n* + king’s name appears in a projective structure between *n k* and the first noun phrase (*trč.t* “milk”) in the Pyramid Texts witness of Pepi II (Pyr. 707a, N).

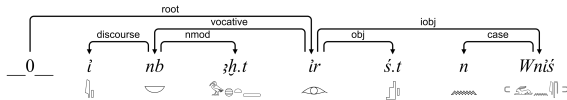
## 4.2 Extraposition by inserting a vocative

A noun phrase, especially a name, can be used as a vocative preceding or following the root of the sentence, for example:

<sup>6</sup> Sim. EUJA-942, 1291, 1415, 1835 and 2101.

<sup>7</sup> Sim. EUJA-654 and 871.

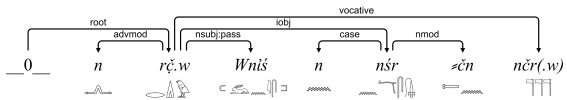
14) EUJA-894 = Pyr. 277a, Unas:



LT: “O (i) lord (nb) of Akhet (ih.t)! Make (ir) a place (s.t) for (n) Unas (Wniš).”

FT: “O lord of Akhet! Make a place for Unas.”

15) EUJA-1049 = Pyr. 323d, Unas:

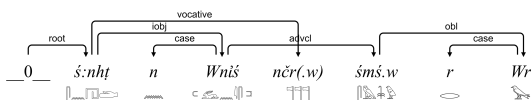


LT: “Unas (Wniš) will-not-be-given (n rč.w) to (n) your (šcn) flame (nšr), (you) gods (nčr(.w)).”

FT: Unas will not be given to your flame, you gods.

A non-projective structure with a vocative occurs when the noun phrase used as a vocative is inserted between a noun governed by the root and its nominal dependent, thus violating the word order, for example:

16) EUJA-981 = Pyr. 306d, Unas:



LT: “Make-salutation (s:nhf), to (n) Unas (Wniš), (you) gods (nčr(.w)), who-is-older (šmš.w) than (r) the Great-One (Wr).”

FT: “Make salutation, you gods, to Unas, who is older than the Great-One.”

The vocative *nčr(.w)* “gods” is governed by the root and it is inserted between *Wniš* as an indirect object (obl:arg) and *šmš.w* which is a participle used as an attribute of *Wniš*.

17) EUJA-1406 = Pyr. 457a–b, Unas (cf. fig. 7):

*s:bʔk r ʔk Wniš m šit ʔk pw sʔb.y sʔb s:wʔb.tw ʔk nčr(.w) im ʔf*

LT: “Make-bright (s:bʔk) indeed (r ʔk) Unas (Wniš) in (m) this (pw) jackal (sʔb.y) lake (šit) of yours (ʔk), (O) Jackal (sʔb), which-cleansed (s:wʔb.tw) you (ʔk) the gods (nčr(.w)) in (im) it (ʔf).”

FT: “Make Unas bright in this jackal lake of yours, O Jackal, in which you cleansed the gods.”

The noun *sʔb* “Jackal” acts as the vocative of the causative verb *s:bʔk* “make bright”. It is inserted between an adverbial phrase (*m šit ʔk pw sʔb.y*) governed by the root and a relative form (*s:wʔb.tw ʔk*) used as an attribute of *šit* “lacke”. This results in

a non-projective structure because *sʔb* breaks the continuity between *šit ʔk* and its relative form.

18) EUJA-117 = Pyr. 22a, Unas (cf. fig. 8):

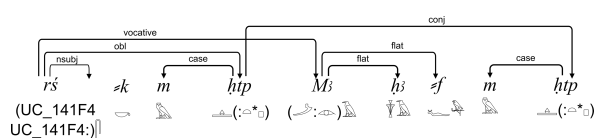
*kbh(.w) ʔk ipn Wsr(.w) kbh(.w) ʔk ipn hʔ Wniš pr.w hr sʔ ʔk (...)*

LT: “These (ipn) libations (kbh(.w)) of yours (ʔk), (O) Osiris (Wsr(.w)), these (ipn) libations (kbh(.w)) of yours (ʔk), O (hʔ) Unas (Wniš), came (pr.w) from (hr) your (ʔk) son (sʔ) (...)”

FT: “These libations of yours, O Osiris, these libations of yours, O Unas, came from your son (...)”

*Wsr(.w)* “Osiris” acts as a vocative of the root (*pr.w*). It follows a noun phrase which is a subject (*kbh(.w) ʔk ipn*) and is repeated after *Wsr(.w)* “Osiris”. A discontinuity of word order happens because the repeated noun phrase is linked in a coordinate relation (conj) to the subject, whereas *Wsr(.w)* “Osiris” is governed by the root. Similarly, in the following example, the vocative (*Mʔ-hʔ-ʔf*) is governed by the root (*rʔ*) and is inserted between two adverbial phrases linked by a coordinate relation (conj):

19) EUJA-1753 = Pyr. 597a, Teti (cf. fig. 9):



LT: “You (ʔk) shall-awake (rʔ) in (m) peace (hʔp), O Mahaef (Mʔ-hʔ-ʔf), in (m) peace (hʔp).”

FT: “You shall awake in peace, O Mahaef, in peace.”

### 4.3 Extraposition by inserting a verb of utterance

If a verb of utterance is inserted into a direct speech text by means of a parataxic relation, and the root of the sentence governs the elements of the direct speech text, there is no syntactic discontinuity:

20) EUJA-512 = Pyr. 147b, Unas:



LT: ““You (kw) (are) distinguished (čn),” said (i.n) they (šn), ‘in (m) your (ʔk) name (rn) belonging-to (n(.i)) god (nčr).”

FT: ““You are distinguished’, said they, ‘in your name of god.””

The root (*čn* “be distinguished”) governs both—the verb of utterance (*l.n ʒsn* “they said”) and the extraposed prepositional phrase in the direct speech sentence (*m rn ʒk n(.i) nčr* “in your name of god”.)

A non-projectivity structure occurs when the second part of the direct speech text is not governed by the root, but rather by a component in the first part of the direct speech text which is separated from its second part by the verb of utterance (see examples 21–23).

21) EUJA-1455 = Pyr. 476a–477b, Unas (cf. fig. 9):

(...) *nfr w(i) ʒ mʒ.w htp w(i) ʒ pt(r) l.n ʒsn in nčr(.w) pr.t r ʒ nčr pn ir p.t (...)*

LT: “(...) ‘How (*w(i)*) lovely (*nfr*) (it is) really (*ʒ*) to see (*mʒ.w*), how (*w(i)*) pleasing (*htp*) (it is) really (*ʒ*) to behold (*pt(r)*)’—said (*l.n*) they (*ʒsn*), namely (*in*) the gods (*nčr(.w)*)—‘that this (*pn*) god (*nčr*) ascends (*pr.t*) indeed (*r ʒ*) to (*ir*) the sky (*p.t*)’ (...)”  
 FT: “(...) ‘How lovely it is really to see and how pleasing it is really to behold’, said they, namely the gods, ‘that this god ascends to the sky’ (...)”

The verb *nfr* “be lovely, nice” governs *l.n ʒsn* “they said”, but not *pr.t* “ascend”, which is an infinitive used in an object clause syntactically linked to *pt(r)* “behold” in the first part of the direct speech text.

22) EUJA-1507 = Pyr. 497b, Unas (cf. fig. 10):  
*wč Wniš r ʒk wč šw wč šw č(t) mṯw sp 4 čṯ(.w) n 4 ipw khʒ.w (...)*

LT: “‘Commend (*wč*) Unas (*Wniš*) indeed (*r ʒk*), commend (*wč*) him (*šw*), commend (*wč*) him (*šw*)’—saying (*č(t)*) a speech (*mṯw*) four (4) times (*sp*) in-succession (*čṯ(.w)*)—‘to (*n*) these (*ipw*) four (4) blustering-winds (*khʒ.w*) (...)’”

FT: “‘Commend Unas, commend him, commend him’—recitation four times in succession—‘to these four blustering winds (...)’”

The verb of utterance is inserted between the oblique argument “to these four blustering winds” (*n 4 ipw khʒ.w*) and its head “commend” (*wč*), thus violating the word order of the sentence in the direct speech text. It can also happen that the direct speech text is introduced by the *č(t) mṯw* formula “saying a speech” and its word order is broken by inserting a ritual remark which completes the text of the *č(t) mṯw* formula, as in the following example:

23) EUJA-2038 = Pyr. 702a, Teti (cf. fig. 11):  
*č(t) mṯw nʒ.w Tti hnʒ ʒk nʒ.w.ti sp 4 čṯ(.w) tp(.i) iʒw.(w)t Wʒč.t*

LT: “Saying (*č(t)*) a speech (*mṯw*): ‘Teti (*Tti*) will-travel (*nʒ.w*) with (*hnʒ*) you (*ʒk*), Traveller (*nʒ.w.ti*)’—4 times (*sp*) in-succession (*čṯ(.w)*)—‘who-is-on (*tp(.i)*) the standards (*iʒw.(w)t*) of Wadjet (*Wʒč.t*).’”

FT: “Recitation: ‘Teti will travel with you, O Traveller’—4 times in succession—‘who is on the standards of Wadjet.’”

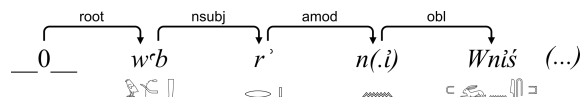
The syntactic relation between *nʒ.w.ti* “Traveller” and its nisba adjective *tp(.i)* “one who is on” is broken by the ritual remark *sp 4 čṯ(.w)* “four times in succession”, which completes the text of the *č(t) mṯw* formula “saying a speech/recitation”

#### 4.4 Extraposition of *n.i* “belonging to/of”

This type of non-projectivity is well known to students of Middle Egyptian when dealing with the so-called “indirect genitive”. However, it has been overlooked that the extraposition of *n.i* in a *pw* sentence is a case of non-projectivity.

According to the dependency approach, the “indirect genitive” consists of the nisba adjective *n.i* “belonging to” used as an adjectival modifier of its head (*amod*) and a noun used as an oblique adjunct of *n.i* (*obl*), for example:

24) EUJA-414 = Pyr. 293a, Unas:

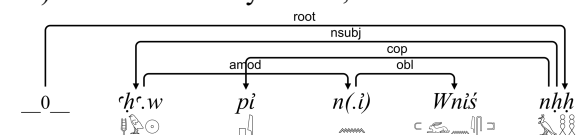


LT: “The mouth (*r*) belonging-to (*n(.i)*) Unas (*Wniš*) (is) pure (*wʒb*) (...)”

FT: “The mouth of Unas is pure (...)”

In a *pw* sentence, non-projectivity occurs when *pw* is used as a copula and inserted between *n(.i)* and its head:

25) EUJA-1290 = Pyr. 412a, Unas:<sup>8</sup>



LT: “The lifetime (*hʒ.w*) is (*pi*) belonging-to (*n(.i)*) Unas (*Wniš*) eternity (*nhh*).”

FT: “Unas’ lifetime is eternity.”

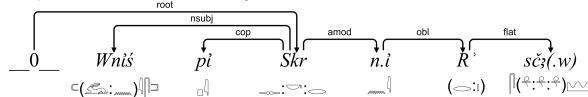
This is a *pw* nominal sentence consisting of *hʒ.w* as its subject and *nhh* as its root. *Pi* is an older variant

<sup>8</sup> Sim. EUJA-442 and 1614.



of *pw* and acts as a copula linking the subject and the root. Its insertion between *h'w* and *n(i)* *Wniš* causes a discontinuity in the word order. It should be noted that there is no discontinuity when *n(i)* follows the root of a *pw* nominal sentence:

26) EUJA-1373 = Pyr. 445b, Unas:

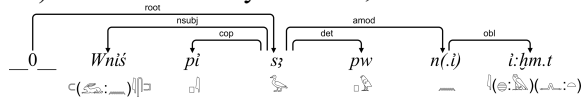


LT: “Unas (*Wniš*) is (*pi*) Sokar (*Skr*) belonging-to (*n.i*) Rostau (*R* -*sčš(.w)*).”

FT: “Unas is Sokar of Rostau.”

The structure is also projective when *pw* / *pi* is used as a demonstrative determiner instead of a copula:

27) EUJA-1544 = Pyr. 515c-d, Unas:



LT: “Unas (*Wniš*) is (*pi*) this (*pw*) son (*s*) belonging-to (*n(i)*) the one-who-is-unknown (*i:hm.t*).”

FT: “Unas is this son of the one who is unknown.”

However, the insertion of an adverbial phrase between a noun after *pw* used as a demonstrative determinative and the nisba adjective *n.i* causes a syntactic discontinuity. This is actually a type of nisba adjective extraposed by an adverbial phrase (cf. ex. 8, above):

28) EUJA-1338 = Pyr. 434e, Unas (cf. fig. 12):

*im šf čt rn šk pw r šk n.i Nm(i) sš Nm(i).t*

LT: “He (*šf*) shall-not (*im*) pronounce (*čt*) this (*pw*) your (*šk*) name (*rn*) against (*r*) you (*šk*) belonging-to (*n.i*) Nemi (*Nm(i)*), son (*sš*) of Nemet (*Nm(i).t*).”

FT: “He shall not pronounce against you this your name of Nemi, son of Nemet.”

Here the noun phrase *rn šk* “your name” is followed by *pw* used as a demonstrative determiner. Non-projectivity is caused by the insertion of the prepositional phrase *r šk* between *rn šk* and the nisba adjective *n.i* because *r šk* is governed by the root of the sentence (*čt*).

The word order in the “indirect genitive” is violated by the insertion of an adverbial phrase between the noun and the nisba adjective *n.i*, even if *pw* is not used as demonstrative determiner of the noun:

29) EUJA-1294 = Pyr. 415a–c, Unas (cf. fig. 13):  
 (...) *hw.(w)t ščn Hr.w hrm.(w) r Wniš m'rk šf r ph šf n(i) bšk n(i) i'n*

LT: “(...) Your (*ščn*) Horus (*Hr.w*) mansions (*hw.(w)t*) are-barred (?) (*hrm.(w)*) to (*r*) Unas (*Wniš*), his (*šf*) bent tail (*m'rk*), belonging-to (*n(i)*) the intestine (*bšk*) belonging-to (*n(i)*) a baboon (*i'n*), (is) at (*r*) his (*šf*) rear (*ph*).”

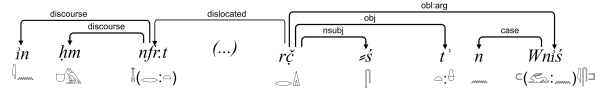
FT: “(...) Your Horus mansions are barred (?) to Unas, his bent tail, of the intestine of a baboon, is at his rear.

Although the meaning of *hrm.(w)* is controversial (Faulkner, 1969, 84 and Sethe, 1936, 176), it acts as the head of this sentence. It is the verb *hrm* conjugated in the 3rd. c. pl. person of the OSSC. Its subject is *hw.(w)t ščn Hr.w* “your Horus mansions” and the prepositional phrase *r Wniš* belongs to its predicate. It is followed by an adverbial sentence consisting of *m'rk šf* as its subject and the prepositional phrase *r ph šf* as its predicate, which is inserted between *m'rk šf* and its nisba adjective *n(i)* causing non-projectivity.

#### 4.5 Discontinuity due to a dislocated element in a sentence with emphatic subject

In Earlier Egyptian, an emphatic subject is introduced by the particle *in* and followed by a participle or a future verb form, for example:

30) EUJA-397 = Pyr. 123d-e, Unas:

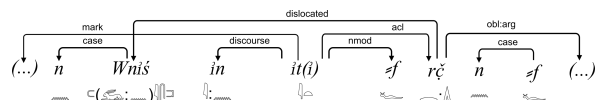


LT: “(It) (is) really (*hm*) the-beautiful-one (*nfr*) (...) she (*šš*) will-give (*rč*) bread (*t*) to (*n*) Unas (*Wniš*) (...)”

FT: “It is really the beautiful one (...) she will give bread to Unas (...)”

The future verb form *rč šš* “she will give” is placed after the emphasised subject and followed by its direct object *t* “bread” and its oblique argument *n* *Wniš* “to Unas”. However, a dislocated element may precede the emphasised subject causing discontinuity in the word order, for example:

31) EUJA-385 = Pyr. 121a, Unas:



LT: “(...) to (*n*) Unas (*Wniš*), (it) (is) his (*šf*) father (*it(i)*) who-gave (*rč*) to (*n*) him (*šf*).”

FT: “(...) to Unas, it is his father who gave to *n* him.”

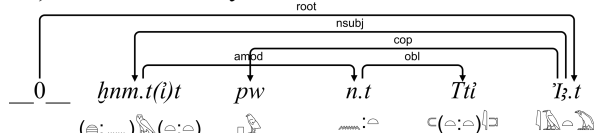
This sentence has two oblique arguments *n Wniš* “to Unas” and *n šf* “to him”. The first of them precedes the emphasised subject *in it(i) šf* “it is his father” causing discontinuity in the word order. It is in a dislocated relation to the participle *rč* “who gave”. The second oblique argument *n šf* “to him” follows the participle according to the Earlier Egyptian word order. The first oblique argument (*n Wniš*) can also be emphasised, because it is referred to by the resumptive pronoun in the second oblique argument. However, this case of non-projectivity may be due to an error involving the addition of *n Wniš* before the emphasised subject (*in it(i) šf*), since no example of this is found in Egyptian grammars.

### 5 Factors for the use of non-projective structures in Old Egyptian

Example 11 showed that non-projectivity may be due to syntactic factors, since the rigid word order of Old Egyptian remains unchanged even when an indirect object expressed by the preposition *n* plus a noun is inserted between the direct object and its modifier. Likewise, the discontinuity of *n.t* “belonging to/of” in a *pw* sentence may be caused by the rigid word order in this type of sentence, in which *pw* usually follows the first noun phrase.

There may also have been pragmatic reasons for using non-projective structures in Egyptian, for in some languages non-projective structures are caused by a difference in discourse function between modifier and head, for example an extraposed attribute acts as focus and its head as theme in Wardaman (Croft, 2022, 163). This may have been another reason for the formation of non-projective structures in Old Egyptian. Given a *pw* sentence such as:

32) EUJA-442 = Pyr. 131d, Teti:



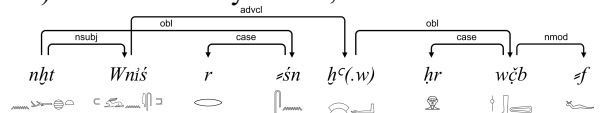
LT: “The nurse (*hnm.t(i)t*) is (*pw*) belonging-of (*n.t*) Teti (*Tti*) Iat (*š.t*).”

FT: “The nurse of Teti is Iat”

The nisba adjectives *n.t* “belonging to (of)” is separated from their head *hnm.t(i)t* “nurse” by *pw* acting as the copula so that the sentence causes a case of non-projectivity (type 4). Since the nisba adjective introduces relevant information about its head, the syntactic discontinuity could be due to the will to emphasise the nisba adjective *n.t*: “The nurse, of Teti, is Iat.”

An attribute may have the same pragmatic function when it is extraposed in a non-projectivity structure (type 1):

33) EUJA-942 = Pyr. 291d, Unas:



LT: “Stronger (*nht*) (is) Unas (*Wniš*), than (*r*) they (*šn*), appearing (*h'(w)*) upon (*hr*) his (*šf*) shore (*wčb*).”

FT: “Unas, who appeared upon his shore, is stronger than they.”

In this sentence there is a contrast similar to that in the *pw* sentences with non-projectivity. *Wniš* is used as a theme and subject, while its attribute *h'(w) hr wčb šf* “appearing upon his shore” introduces relevant information about *Wniš*. The prepositional phrase governed by the root and inserted between *Wniš* and *h'(w)* may be pushed into the background by the extraposition of the attribute, on which the focus is likely to lie (cf. other examples discussed in 4.4., above). Similarly, non-projectivity type 2 may be caused by the intention to highlight an extraposed modifier through the insertion of a vocative. This can be seen in example 19, where *m htp* “in peace” is repeated after the noun used as a vocative *M<sup>3</sup>-h<sup>3</sup>-šf* “Mahaef” to emphasise the action of awaking in peace.

Finally, another factor for non-projectivity in Old Egyptian is the insertion of previously omitted or forgotten information, as occurs when a verb of utterance is inserted in the middle of a direct speech text (type 3, examples 20–23).

### 6 Conclusion

This paper has shown that the formation of grammatically accepted non-projective structures in Old Egyptian is not accidental, but it rather follows patterns governed by syntactic rules. Five types of non-projective patterns have been identified so far in the UD-EUJA treebank. New

types will probably be found during its full development.

Furthermore, this paper has argued for three factors involved in the formation of non-projective structures in Old Egyptian:

- Maintenance of the word order, even if this leads to syntactic discontinuity due to the extraposition of an attribute.
- Emphasis on an extraposed modifier.
- Addition of omitted or forgotten information by the insertion of a verb of utterance in the middle of a direct speech text.

Finally, the scarce presence of non-projective structures in Unas's and Teti's Pyramid Texts—1.37 % and 1.42 % respectively—is probably due to the rigid word order of Old Egyptian language.<sup>9</sup> It contrasts with freer word order languages such as Ancient Greek, which has a higher rate of non-projectivity (15.15%) (Mambrini and Passarotti, 2013, 180, Tab. 3).

## Acknowledgments

The UD-EUJA treebank was created during the CA21167 COST Action UniDive, funded by COST (European Cooperation in Science and Technology).

## References

- James P. Allen. 2013. *A New Concordance of the Pyramid Texts. 6 volumes*. Brown University.
- William Croft. 2022. *Morphosyntax. Constructions of the World's Languages*. Cambridge.
- Roberto A. Díaz Hernández and Marco Carlo Passarotti. 2024. Developing the Egyptian-UJaen Treebank. In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)* Association for Computational Linguistics, pages 1–10. <https://aclanthology.org/2024.tlt-1.1/>.
- Raymond O. Faulkner. 1969. *The Ancient Egyptian Pyramid Texts*. Oxford.
- Thomas Groß, Timothy Osborne. 2009. Toward a Practical Dependency Grammar Theory of Discontinuities. *SKY Journal of Linguistics*, 22:43–90.
- David G. Hays. 1964. Dependency Theory: A Formalism and Some Observations. *Language*, 40(4):511–525.
- Renata Landgráfová. 2002. Resumptive Pronouns in Middle Egyptian – A Means of Avoiding Non-Projective Constructions?. *Lingua Aegyptia* 10: 269–282.
- Yves Lecerf, P. Ihm. 1960 *Éléments pour une grammaire générale des langues projectives. Rapport CETIS* 1: 1–19.
- Francesco Mambrini, Marco Passarotti. 2013. Non-projectivity in the Ancient Greek Dependency Treebank. In *Proceedings of the Second International Conference on Dependency Linguistics*. Prague:177–186.
- Solomon Marcus. 1965. Sur la notion de projectivité. *Mathematical Logic Quarterly*, 11(2):181–192.
- Marie-Catherine de Marneffe (et al.) Universal Dependencies. *Computational Linguistics*, 47(2):255–308.
- Ratna Nirupama, Prakash Mondal. 2022. The Representation of Discontinuity and the Correspondence Principle. In *Pacific Asia Conference on Language, Information and Computation (PACLIC36)*, 20–29.
- Joakim Nivre (et al.) 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Association for Computational Linguistics, pages 1659–1666.
- Timothy Osborne. 2019. *A Dependency Grammar of English. An Introduction and Beyond*. Amsterdam/Philadelphia.
- Wolfgang Schenkel. 2012. *Tübinger Einführung in die klassisch-ägyptische Sprache und Schrift*. Pagina, Tübingen.
- Eckehard Schulz. 2010. *A Student Grammar of Modern Standard Arabic*. Cambridge.
- Kurt Sethe. 1908–1922. *Die altägyptischen Pyramidentexte nach den Papierabdrücken und Photographien des Berliner Museums. 4 volumes*. Heinrich'sche Buchhandlung, Leipzig.
- Kurt Sethe. 1936. *Übersetzung und Kommentar zu den altägyptischen Pyramidentexten II. Band, Spruch 261–365*. Glückstadt.

---

<sup>9</sup> According to Nirupama and Mondal (2022, 20) rigid word order languages tend to have a low rate of syntactic discontinuity.

## A Appendix:

EUJA-	Pyr.	fig.
Category 1, cf. 4.1		
654	202c, Unas	
871	270a–e, Unas	
942	291d, Unas	
1291	412a–413c, Unas	
1313	424a, Unas	1
1390	451b–c, Unas	
1415	460a–b, Unas	
1511	500a, Unas	3
1758	599b–c, Teti	4
1769	606b–d, Teti	5
1835	632b–c, Teti	
2050	707a–b, Teti	6
2101	728a–c, Teti	
2119	734a–b, Teti	2
Category 2, cf. 4.2		
117	22a, Unas	8
981	306d, Unas	
1406	457a–b, Unas	7
1753	597a, Teti	
Category 3, cf. 4.3		
1455	476a–477b, Unas	9
1507	497b, Unas	10
2038	702a, Teti	11
Type 4, cf. 4.4		
442	131d, Teti	
1290	412a, Unas	
1294	415a–c, Unas	13
1338	434e, Unas	12
1614	538c, Teti	
Type 5, cf. 4.5		
385	121a, Unas	

Table: Non-projective structures in Unas's and Teti's Pyramid Texts.

## B Supplementary Material—Syntactic Tree Diagrams

Fig. 1: Ex. 7, EUJA-1313 = Pyr. 424a, Unas

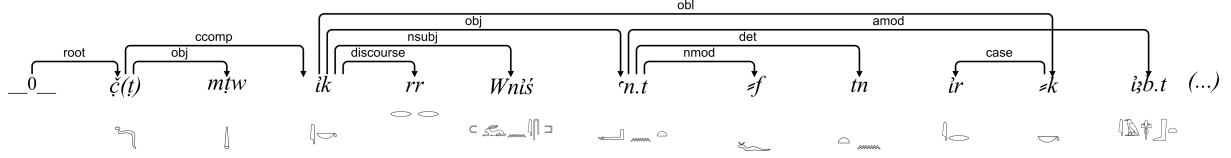


Fig. 2: Ex. 8, EUJA-2119 = Pyr. 734a–b, Teti

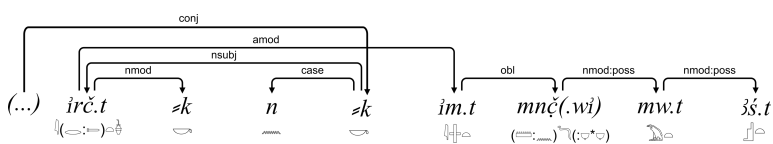


Fig. 3: Ex. 10, EUJA-1511 = Pyr. 500a, Unas

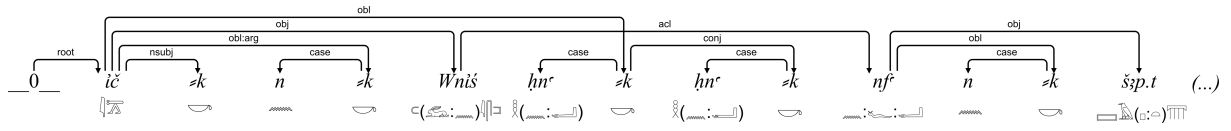


Fig. 4: Ex. 11, EUJA-1758 = Pyr. 599b–c, Teti

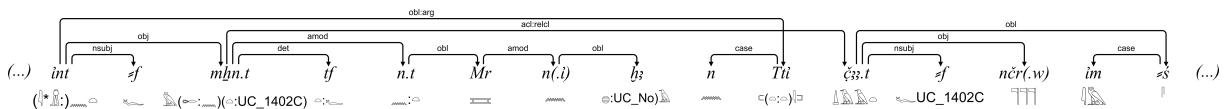


Fig. 5: Ex. 12, EUJA-1769 = Pyr. 606b–d, Teti

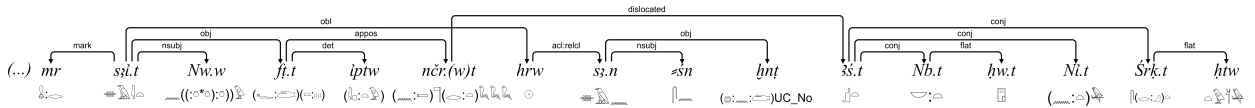


Fig. 6: Ex. 13, EUJA-2050 = Pyr. 707a–b, Teti

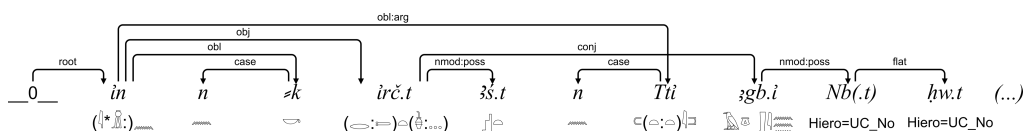


Fig. 7: Ex. 17, EUJA-1406 = Pyr. 457a–b, Unas

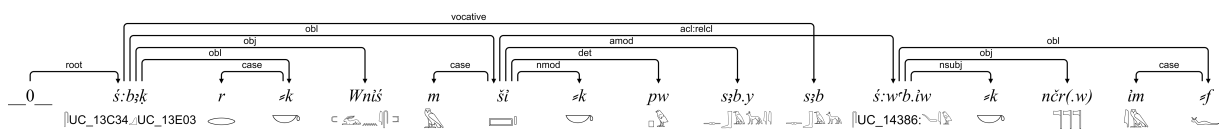


Fig. 8: Ex. 18, EUJA-117 = Pyr. 22a, Unas

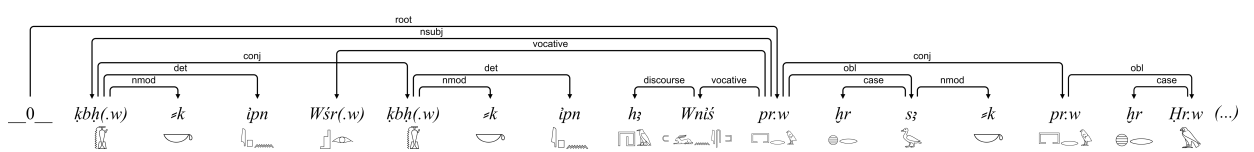


Fig. 9: Ex. 21, EUJA-1455 = Pyr. 476a–477b, Unas

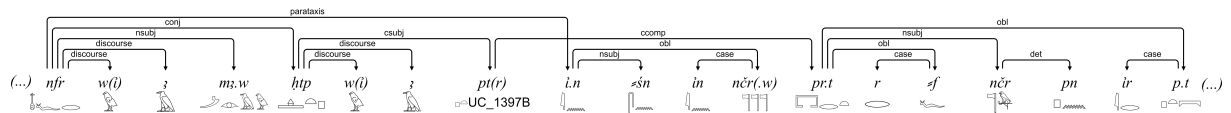


Fig. 10: Ex. 22, EUJA-1507 = Pyr. 497b, Unas

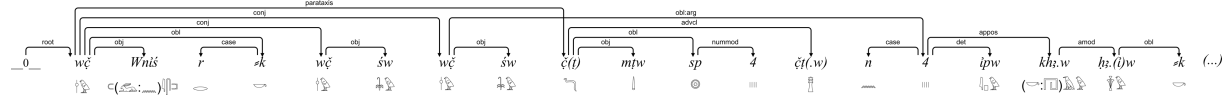


Fig. 11: Ex. 23, EUJA-2038 = Pyr. 702a, Teti

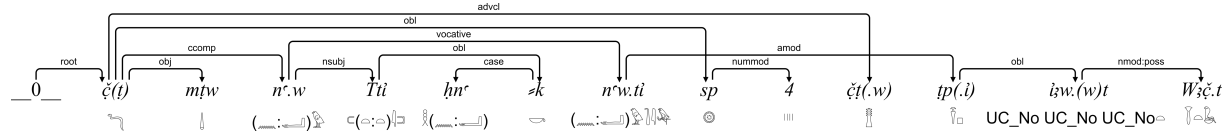


Fig. 12: Ex. 28, EUJA-1338 = Pyr. 434e, Unas

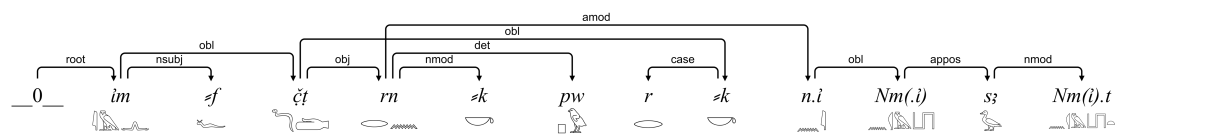
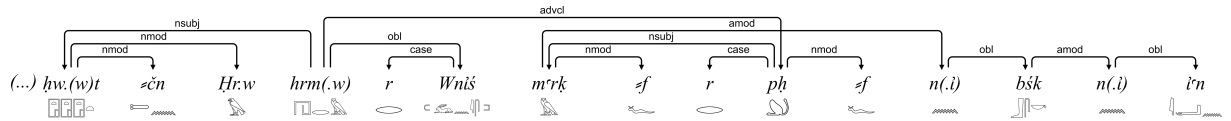


Fig. 13: Ex. 29, EUJA-1294 = Pyr. 415a–c, Unas



# Tracing Syntactic Complexity: Exploring the Evolution of Average Dependency Length Across Three Centuries of Scientific English

**Marie-Pauline Krielke**

Saarland University,  
Germany

mariepauline.krielke@  
uni-saarland.de

**Diego Alves**

Saarland University,  
Germany

diego.alves@  
uni-saarland.de

**Luigi Talamo**

Saarland University,  
Germany

luigi.talamo@  
uni-saarland.de

## Abstract

We present a diachronic analysis of syntactic change in a corpus of 300+ years (1665 - 1996) of scientific English annotated with Universal Dependencies (UD) and Dependency Length (DL). We trace the development of average Dependency Length (aDL) as a measure of syntactic complexity in scientific English between 1665 and 1996 and describe the corpus construction, and report on the UD annotation evaluation. We find that aDL first decreases toward the 19<sup>th</sup> c., but then increases significantly in the 20<sup>th</sup> c. We show that the highly aggregate measure of aDL masks the underlying mechanisms driving shifts in syntactic complexity. A fine-grained analysis of the dependency relations involved in change shows that the increasing use of (multi-word) compounds is a dominant source of long leftward expanded noun phrases, leading to the expansion of syntactic dependencies within and beyond the noun phrase. The results open a new perspective on syntactic complexity, shifting from the sentence to the phrasal level.

## 1 Introduction

Syntactic change in English in the past 300 years was largely of statistical nature, reflecting shifts in usage rather than structural innovation. These shifts were motivated in response to increasingly changing communicative demands from evolving contexts of usage, such as the advent of new genres like scientific English, beginning with the Scientific Revolution and continuing into the present day.

Syntactic change in the evolution of scientific English has been examined from both qualitative and quantitative perspectives. Ranging from descriptive approaches (e.g. Halliday, 1988; Halliday and Martin, 1993) to early quantitative analyses (Biber and Clark, 2002; Biber and Gray, 2010, 2016; Hundt et al., 2012) report a general shift from subordinate constructions to nominal style. More recent approaches, grounded in cognitive and

information-theoretic frameworks, have examined scientific English using measures such as Relative Entropy and Surprisal (e.g. Degaetano-Ortlieb et al., 2016; Bizzoni et al., 2020; Teich et al., 2021), as well as memory-based metrics like Dependency Length (DL, the distance between a syntactic head and its dependent(s)) (Juzek et al., 2020; Krielke, 2024). These studies collectively indicate a long-term trend toward syntactic simplification at the sentence level, characterized by fewer clausal embeddings, increasingly informationally dense noun phrases (NPs), and overall decreasing DL, independent of sentence length.

While these trends are well-documented up to the 19<sup>th</sup> c., little research has focused on syntactic developments in scientific English throughout the 20<sup>th</sup> c. Recent findings suggest a marked tendency toward extremely long nominal phrases in this later period, mainly through increased use of premodification strategies such as compounding (Degaetano-Ortlieb, 2021). This phenomenon leads to a pronounced leftward expansion of the NP. Given that aDL is most optimal when syntactic branching is balanced on both sides of the syntactic head (Temperley and Gildea, 2018), we hypothesize that the increasing reliance on leftward expansion in the 20<sup>th</sup> c. may result in a reversal of the earlier trend, leading to an increase in aDL. This study investigates whether this shift is empirically supported, contributing to a more comprehensive understanding of changes in syntactic complexity in scientific English.

We start by discussing the existing literature on diachronic syntactic change in scientific English and the use of DL as a measure of syntactic complexity (Section 2).

We present our data and methods, i.e. an updated and extended version (+100 years) of the Royal Society Corpus (RSC, Fischer et al., 2020) spanning the years between 1665 and 1996 annotated with Universal Dependencies (UD, de Marneffe et al.,

2021) and DL (Section 3).

Next, we evaluate the dependency annotations using Stanza compared to previous evaluation scores obtained by (Krielke et al., 2022) using UD-pipe (Straka and Straková, 2017). Apart from general improvements in the parsing quality by using a more state-of-the-art parser, we test whether parsing accuracy has additionally improved in the last 100 years due to, among other reasons, a notable decrease in average sentence length (SL) over time (i.e. the shorter a sentence, the better the parsing) considering SL impact on the encountered evaluation results (Section 4).

We then analyze the evolution of DL across three centuries. We measure average aDL per SL per 50-year period to gain a first overview of the development of aDL over time. Against common intuition that DL would further decrease over time (i.e. after 1900), we find an increase in aDL normalized by SL in the 20<sup>th</sup> c. To identify the driving forces behind this upward trend, we analyze the most influential syntactic relations on changes in aDL with SL held stable. The analysis reveals a pivot role of the nominal compounds (Section 5). Section 6 presents a general discussion of the results. We close with conclusions (Section 7) and limitations of this work (Section 8).

## 2 Related Work

Scientific English has been described to have undergone two notable developments: (i) the creation of specialized terminology (Halliday and Martin, 1993; Wang et al., 2023) and (ii) a shift from clausal to phrasal complexity (Biber and Gray, 2016; Alves et al., 2024), i.e. sentences continuously consist of fewer subclauses and rather long and complex noun phrases (e.g. Halliday, 1988; Halliday and Martin, 1993; Biber and Clark, 2002; Biber and Gray, 2011, 2016; Hundt et al., 2012). This process of syntactic reorganization has been attributed to the need for efficiency achieved by linguistic condensation on the one hand and a reaction to increasing shared expert knowledge that makes it possible to use grammatically implicit encodings (e.g. compounds) instead of explicit ones (e.g. relative clauses) (Biber and Gray, 2010).

The preference for dense nominal structures over intricate subordinate structures is associated with a general trend towards lower grammatical complexity, as it is connected to the hypothesis that aDL minimizes diachronically (cf. Juzek et al., 2020;

Krielke, 2024). This assumption is grounded in the Dependency Length Minimization (DLM) Hypothesis, according to which human languages tend to reduce the distance between syntactically related words due to limited working memory capacity and the principle of least effort (Zipf, 1949) and assuming that shorter DL is easier to produce and comprehend (Hawkins, 1994, 2004; Gibson, 2000; Demberg and Keller, 2008).

DLM is widely recognized as a universal property and has been observed across languages (Gildea and Temperley, 2010; Liu et al., 2017; Futrell et al., 2015), across genres (Wang and Liu, 2017), and diachronically (Tily, 2010; Lei and Wen, 2020; Liu et al., 2022). In particular, for scientific English between 1650 and 1900, Juzek et al. (2020) and Krielke (2024) observe a steady reduction in aDL; both papers attribute this trend to a persistent increase in nominal structures, which create rather short dependency relations, and a strong decrease of clausal subordination, which represent rather long dependency relations.

Diachronic research on other genres (e.g., political speeches, Lei and Wen, 2020; Liu et al., 2022) found a similar trend but did not strictly control for SL. Comparing the aDL in scientific and general English, Krielke (2024) in fact showed that aDL is strongly correlated with SL and, when holding SL stable, only scientific texts showed a significant aDL reduction over time. Liu et al. (2022) find general downward trends in their data but looking at specific dependency relations (deprels), they find that nominal relations (attributive adjectives, possessive modifiers, compounds, determiners, etc.) become longer over time while clausal relations become shorter.

Using information theoretic measures, Degaetano-Ortlieb (2021) showed for the 20<sup>th</sup>c. that especially composite terminology undergoes expansion in comparison to other nominal pre- and postmodification patterns; this, on the one hand, seems to point to structural compression but, on the other hand, might indicate a further expansion of the core nominal group leading to increasing distances between dependency relations on average.

## 3 Data and Methods

Our corpus contains texts extracted from the *Philosophical Transactions* and *Proceedings* of the Royal Society of London. It represents an exten-



sion (+76 years) of the open version *The Royal Society Corpus 6.0* (Fischer et al., 2020), covering texts until 1920.

On top of sentence splitting and tokenization using TreeTagger (Schmid, 1994) and spelling normalization, using a trained model of VARD (Baron and Rayson, 2008), we enriched the corpus with two types of token-level annotation layers: (i) UD annotation and (ii) dependency length. Layer (i) was annotated using Stanford Stanza v.1.5 with the English combined model pre-trained on five English treebanks (EWT, GUM, GUMReddit, PUD, and Pronouns) from UD v.2.12<sup>1</sup>, performing the following NLP tasks: lemmatization, parts-of-speech, morphological features tagging, dependency parsing, and named entity recognition. We chose Stanford Stanza because it natively supports UD, can be implemented as a Python library, and performs slightly better on English data with respect to similar tools, i.e., UDPipe and spaCy (for a comparison, see Qi et al., 2020). Layer (ii) consists of four different values calculated at the sentence level and excluding punctuation: dependency length (DL), sentence length (SL), sum of all (absolute) dependency lengths (sumDL), and average dependency length per sentence (aDL). For example, in the sentence displayed in Figure 1 the head of the nominal compound *brachiopod shell* is *shell* (token id = 14) and since *shell* depends on *calcite* (token id = 10), has a DL of  $-4$  i.e.,  $10 - 14$ ; its only modifier, *brachiopod* (token id = 13), has a DL of  $1$  i.e.,  $14 - 13$ ; both compound constituents have an SL of  $14$  - the sentence contains only a punctuation symbol - a sumDL of  $24$  i.e.,  $1 + 1 + 0 + 1 + 2 + 1 + 2 + 2 + 1 + 3 + 3 + 2 + 1 + 4$  and an aDL of  $1.84$  i.e.,  $\frac{1+1+0+1+2+1+2+1+3+3+2+1+4}{13}$ . For our analyses, we bin periods of 50 years, e.g.  $1700 = 1701 - 1750$ ,  $1750 = 1751 - 1800$ .

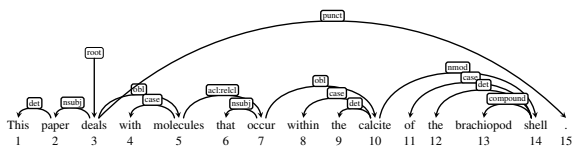


Figure 1: Dependency tree of a sentence with dependency labels and token ids.

<sup>1</sup>[https://stanfordnlp.github.io/stanza/combined\\_models.html](https://stanfordnlp.github.io/stanza/combined_models.html)

Subcorpus	Sentences	Tokens
1665-1700	42,398	2,195,166
1701-1750	57,915	2,860,083
1751-1800	109,771	5,037,846
1801-1850	169,224	7,004,769
1851-1900	352,259	12,935,866
1901-1950	663,561	21,031,436
1951-1996	2,587,326	78,165,925
<b>All</b>	<b>3,982,454</b>	<b>129,231,091</b>

Table 1: Total number of tokens and sentences per 50-year period.

Subcorpus	Sentences	Tokens	Deprel
1665-1700	20	1093	40
1701-1750	20	1024	40
1751-1800	20	1038	40
1801-1850	20	967	40
1851-1900	20	982	38
1901-1950	20	1011	34
1951-1996	20	789	38
<b>All</b>	<b>140</b>	<b>6904</b>	<b>49</b>

Table 2: Evaluation set description in terms of number of sentences, tokens, and dependency relation tags.

## 4 Dependency Parsing Evaluation

To evaluate Stanza’s parsing performance on RSC texts, we carried out a quality check over 140 sentences randomly selected from the parsed corpus (i.e., 20 from each 50-year period<sup>2</sup>). These sentences were manually corrected by a linguist and a student with previous experience with Universal Dependencies; the cases of disagreement were discussed until an agreement was reached. Then, the sentences parsed with Stanza were compared to the manually corrected ones using the DependAble tool (Choi et al., 2015).

Table 2 presents the overall characteristics of the evaluation set manually corrected in terms of the number of tokens and dependency relation labels.

The model used to parse the RSC corpus is the combined English one from the Stanza repository. When combined, these training sets have 53 deprels. Thus, our evaluation set covers the vast majority of the dependency parsing labels excluding *csubj:outer*, *reparandum*, *list*, and *dislocated*.

<sup>2</sup>The parsed sentences along with the manual corrections can be accessed at <https://tinyurl.com/4j5a99dx>.

Subcorpus	UAS	LAS
1665-1700	88.29	85.09
1701-1750	90.33	87.21
1751-1800	91.52	89.02
1801-1850	90.49	87.49
1851-1900	88.80	84.93
1901-1950	94.86	92.88
1951-1996	93.92	90.87
<b>All</b>	91.06	88.11

Table 3: UAS and LAS results of the evaluation set.

#### 4.1 Evaluation Results

Table 3 presents the results concerning unlabeled and labeled attachment scores (UAS and LAS respectively).

We notice a tendency of better parsing scores for more recent texts (i.e., from the 20<sup>th</sup> c.), as one would expect regarding the texts composing the training corpora of the Stanza model. Compared to the similar analysis conducted by (Krielke et al., 2022) where RSC texts from 1665 to 1899 were parsed with UDPipe 1.0 (Straka and Straková, 2017), Stanza seems to provide better parsing scores, an improvement of around 7 points regarding the overall LAS. However, it is important to mention that the number of analyzed sentences concerning the evaluation set is quite small, so further analyses are necessary for a more complete and statistically valid evaluation of the parsed corpus.

As was the case in a previous parsing analysis of the RSC (cf. Krielke et al., 2022), during the manual correction of the new version of the corpus we found that many parsing errors are due to OCR errors and tokenization issues (e.g., "simul taneous" instead of "simultaneous" and "bem" instead of "been"). Furthermore, when equations were part of sentences, they were usually tokenized in random ways and labeled with erroneous deprels.

Besides overall UAS and LAS, the DependAble tool also provides detailed information regarding scores in relation to sentence length (Table 4) and distance between heads and dependents (Tables 5 and 6). These analyses were conducted considering all 140 sentences of the evaluation set.

Sent. Length	≤ 20	≤ 30	≤ 40	≤ 50	≥ 50
No. Sentences	5	16	12	13	21
LAS	91.13	90.49	90.04	87.67	87.26

Table 4: LAS in relation to sentence length.

In terms of SL, we can observe in Table 4 that

Dist.	< -5	-5	-4	-3	-2	-1
LAS	81.16	84.85	86.49	93.92	95.52	95.63

Table 5: LAS by distance when the head is after the dependent.

Dist.	1	2	3	4	5	>5
LAS	77.94	83.09	88.04	80.31	70.34	69.99

Table 6: LAS by distance when the head precedes the dependent.

LAS results tend to deteriorate for longer sentences. Figure 2 shows that SL tends to decrease over time, thus, this may influence the better LAS scores observed in more recent texts.

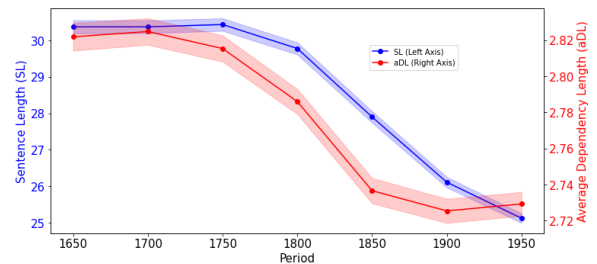


Figure 2: Development of mean SL and mean aDL with 95% CI.

Table 5 shows that the closest the dependent is to the head, the better are LAS results in the cases where the dependent precedes the head in the sentence. However, as shown in Table 6, it is not the case when the dependent follows the head. These counter-intuitive results concerning dependents with a distance of 1 or 2 from their heads can be explained by recurrent OCR and tokenization errors, leading to several occurrences of goes with in the manually annotated evaluation corpus. For example, in one sentence the token *itself* is split into two tokens, *it* and *self*; accordingly, the second token is labeled goes with in the corrected file. However, from distance 3 to >5, the results are as expected, the further the dependent, the lower the LAS.

A qualitative analysis of the dependency parsing evaluation shows that, of the deprels occurring more than 20 times in the evaluation set, the following ones present LAS below 80: parataxis, acl:relcl, goeswith, acl, advcl, and conj. Moreover, we find that obl, cc, and conj tend to have better LAS values in more recent texts.

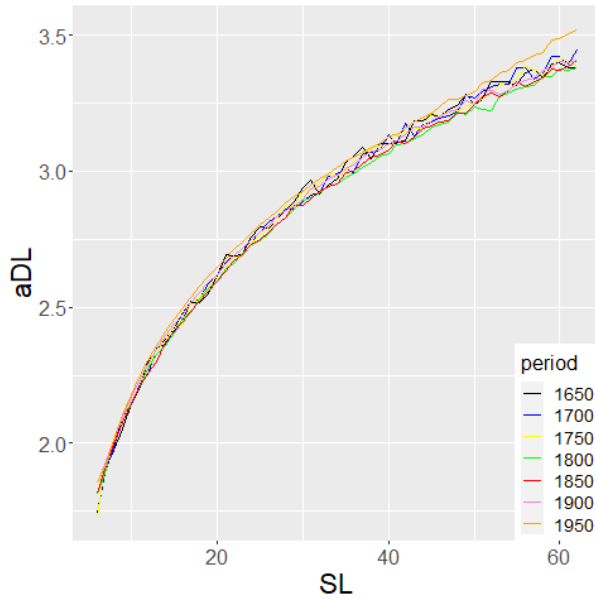


Figure 3: aDL per SL per period.

## 5 Analysis

We conduct a quantitative analysis concentrating on the development of aDL as a measure of syntactic complexity in scientific English texts over the past 300 years. First, our goal is to verify results from earlier studies, especially whether the use of a different parser on the same data influences the outcome of DL calculations. For this comparison, we refer to the results presented by (Juzek et al., 2020) obtained from Stanford Parser (Klein and Manning, 2003). Second, we are interested in the aDL development in the 20<sup>th</sup> c., since earlier studies only cover the years until 1900. Regarding changes in the 20<sup>th</sup> c., we expect to find an increase in aDL due to leftward NP expansion (compare example in Figure 7).

### 5.1 Dependency Length across periods

aDL overall decreases steadily until 1900 together with SL and increases slightly after 1950 (Figure 2). The trends of both measures are moderately correlated (Pearson Correlation Coefficient: 0.5560, p-value < 0.01). Hence, aDL values without controlling for periodically dominant SLs paint a skewed picture. We calculate mean aDL per SL per 50-year period (Figure 4, (a)) and only display SLs occurring  $\leq 250$  times (i.e.,  $SL \geq 6$  and  $SL \leq 62$ ). 1950 (orange) consistently shows the highest aDL per SL among all periods. Only 1650 (black) sporadically shows peaks surpassing the aDL of the latest

period. aDL per SL is lowest in the periods 1800 (green) and 1850 (red).

Since the lines in Figure 3 overlap and are therefore hard to interpret, we conduct independent Two-Sample t-tests to compare the means of aDL per SL between two adjacent periods for each SL and to identify significant differences between our results of aDL per SL per period. Figure 4 shows that aDL in the period 1900 vs. 1950 significantly differs for nearly all SLs. Also, the aDL in the period 1850 vs. 1900 shows significant differences for most (especially the shorter) SLs. Among the earlier periods, only 1750 vs. 1800 shows significant differences for nearly half of all SLs. For the other comparisons, we find only very few significant differences; aDL only seems to differ in these periods sporadically at specific SLs.

The results indicate an SL-independent (for most SLs non-significant) decrease of aDL between the 18<sup>th</sup> c. and the 19<sup>th</sup> c., followed by a highly significant increase throughout the 20<sup>th</sup> c. The development until 1900 is in line with that found by Juzek et al. (2020), who showed that aDL went gradually down between 1650 and 1889. The discovery of an increase in aDL in the 20<sup>th</sup> c. points to a possible impact of increasingly leftward expansion of nominal phrases.

However, this comparison represents a highly aggregated view of the situation. To observe changes in the full range of aDL values over time and still be able to control for SL, we choose a SL that is strongly represented in all periods. For this, we take the arithmetic mean between the observed SLs (6 - 62):  $(6 + 62)/2 = 68/2 = 34$ . We thus use SL 34 for further inspection.

### 5.2 Development of aDL at SL 34

aDL at SL 34 first shows a slight increase between 1650 and 1700, and a downward trend towards the 19<sup>th</sup> c. In the 20<sup>th</sup> c., aDL increases steeply and reaches its highest mean value in the period 1950. Independent Two-Sample t-tests with *period* as the predictor and *aDL* as the response variable show a significant decrease between 1750 and 1800 as well a highly significant increase in aDL in the last three periods, i.e. 1850 vs. 1900 and 1900 vs. 1950 (Table 7). To verify whether the restructuring of the NP (decreasing subordination between the 18<sup>th</sup> and 19<sup>th</sup>c. and increasing premodification in the 20<sup>th</sup>c.) is the driving force behind the observed aDL changes, we further analyze the development of individual deprels.

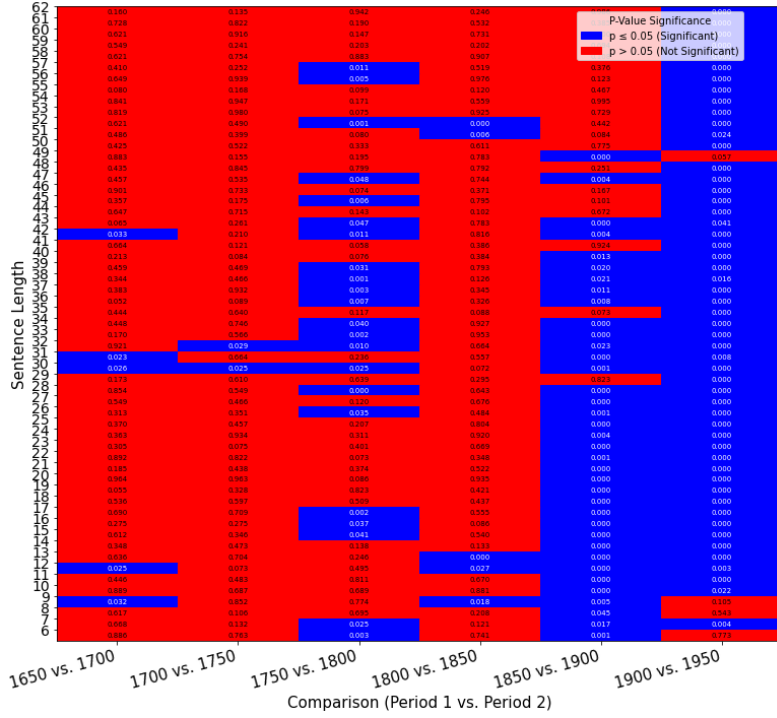


Figure 4: P-values for two-sided t-tests comparing the aDLs per sentence of each SL (6–62) between all adjacent periods.

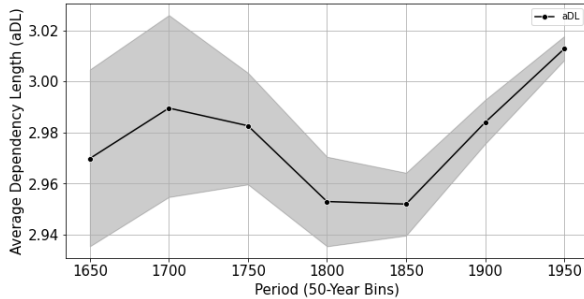


Figure 5: aDL at SL 34 per 50-year period.

Comparison	P-Value	Significant
1650 vs. 1700	0.45	False
1700 vs. 1750	0.74	False
1750 vs. 1800	0.04	True
1800 vs. 1850	0.93	False
1850 vs. 1900	0.00	True
1900 vs. 1950	0.00	True

Table 7: T-Test results for AVD changes in sentence length 34 across time periods

### 5.3 Individual deprels at SL 34

To detect influential deprels for the decrease in aDL between 1750 and 1800, and the increase in the 20<sup>th</sup> c., we take a closer look at two factors influencing the overall aDL (holding SL stable at 34) in a period: changes in aDL per deprel per period (x-axis), and changes in fpm (y-axis) per deprel per period (see Figure 6), formalized as in Equation 1 for  $\Delta fpm$  and Equation 2 for  $\Delta aDL$ .

$$\Delta fpm(\text{deprel}, t_i, t_j) = fpm_{t_i}(\text{deprel}) - fpm_{t_j}(\text{deprel}) \quad (1)$$

$$\Delta aDL(\text{deprel}, t_i, t_j) = aDL_{t_i}(\text{deprel}) - aDL_{t_j}(\text{deprel}) \quad (2)$$

Figures 8a and 8b show deprels occurring  $> 10,000$  times per million tokens and display the frequency and aDL differences of each deprel between two periods; the colors indicate the average aDL of each deprel ranging from short-distance

( $< 3$ ) over mid-distance ( $> 3$ ) to long-distance functions ( $> 6$ ).

#### 5.3.1 Comparison 1750 vs. 1800

Between 1800 and 1750 (Figure 8a), we find a notable increase in two noun phrase (NP) premodifying elements: amod (attributive adjective) and det (determiner), and in two postmodifying elements: case (adposition) and nmod (nominal modifier). This increase in both NP pre- and postmodifiers indicates that during this period, NPs become increasingly loaded with additional information, while showing an equal distribution of pre- and postmodification on both sides of the nominal head, possibly leading to the observed decrease in aDL.

The most substantial frequency decrease is found for nsubj (nominal subject), cc (coordinating con-

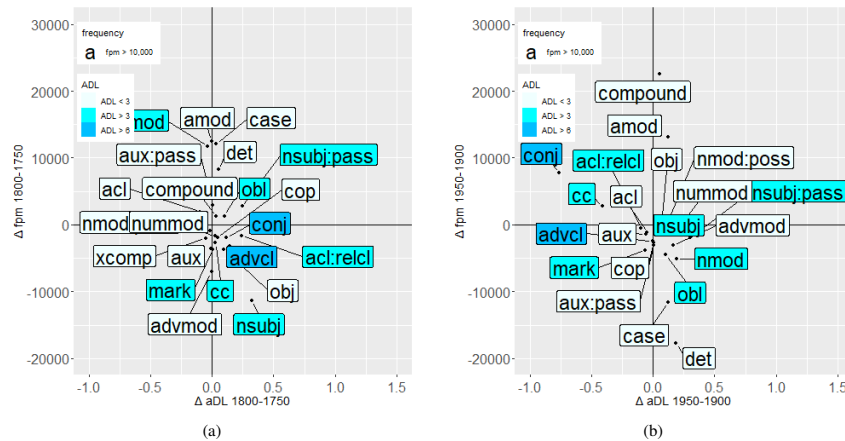


Figure 6: Relative frequency and aDL development of high-frequency deprels between 1750 and 1800 (a) and between 1900 and 1950 (b).

junction), conj (conjunct), indicating a decrease in clausal subordination being an additional factor of Dependency Length Minimization.

In terms of aDL, we can observe that almost all deprels become slightly longer or do not change at all over time. The significant decrease in aDL between 1750 and 1800 could be owed to an equal frequency increase in typically short deprels functioning as NP pre- and postmodifying constituents and creating an equal distribution of information on either side of the nominal head. Additionally, longer deprels typically indicating the presence of a subclause, e.g. mark (word marking a clause as subordinate to another clause), advcl (adverbial clause modifier) or acl:relcl (relative clause) drop in frequency, further contributing to the observed aDL reduction.

### 5.3.2 Comparison 1900 vs. 1950

Comparing 1900 and 1950 (Figure 8b), the most notable difference to 8a is the increase of compounds, followed by a rise in amod, both indicating leftward expansion of the NP. The strongest decrease in frequency can be observed for determiners (det), followed by case markers (case) and nominal modifiers (nmod), typically indicating an NP postmodifying element (prepositional phrases). The frequency development points to a shift of NP modification towards the left side of the nominal head, confirming our hypothesis that leftward expansion is associated with the aDL increase observed in the 20<sup>th</sup> c. We observe an increase in aDL in most of the deprels categorized as belonging to *nominals* according to the UD-framework<sup>3</sup>. More specifi-

cally, the deprels belong to the group of nominal premodifiers. The aDL increase of these deprels can be explained by the increase in compounds since their presence extends the distance between premodifiers (e.g., determiners, numeric modifiers, adjectives) and the nominal head (as in e.g. *the supersonic convective Mach number shear layer*).

aDL decreases for a few clausal deprels, such as relative clauses (acl:relcl) and deprels indicating coordination. Interestingly, conjuncts (conj) increase in frequency but become shorter on average. In UD, conjuncts can refer to both nominal coordination (*Mary and John*) and clausal coordination (*John slept and Paul left*). The aDL decrease of conjuncts suggests a shift towards nominal coordination and away from clausal coordination, which is in line with the literature (e.g. Halliday, 1988).

### 5.3.3 The influence of compounds on aDL

To conclusively verify the influence of compounds on the overall extension of aDL in the 20<sup>th</sup> c., we recalculate aDL per SL for the last 96 years (1900 - 1996), using the code<sup>4</sup> provided by Dyer (2023), treating compounds as single words vs. compounds treated as multi-word expressions (MWEs, cf. Figure 7).

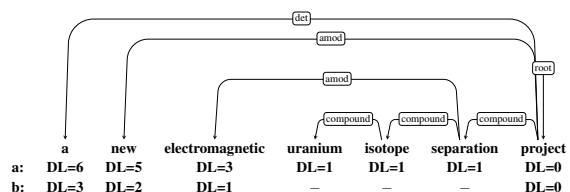


Figure 7: Calculation of DL for compounds as MWEs (a) vs. as single words (b).

<sup>3</sup><https://universaldependencies.org/u/dep/all.html>

<sup>4</sup><https://github.com/andidyer/DependencyLengthSurvey>



Although SL changes when compounds are counted as single words, the visualization of aDL against SL (Figure 8a) shows that, aDL with compounds as single words drops visibly below the curve of compounds treated as MWEs. This shows that the strong presence of compounds in the 20<sup>th</sup> c. leads to a substantial increase in aDL.

Compared to earlier periods, this effect is also visible when comparing aDL calculated with compounds as single words: aDL in the 20<sup>th</sup> c. stays below the aDL in other periods (Figure 8b).

## 6 Discussion

DL is a well-established measure of syntactic complexity that impacts memory-based processing costs. Our results may suggest that an increase in aDL in the 20<sup>th</sup> c. makes scientific English increasingly difficult to process. However, this interpretation should be approached with caution.

First, it is important to recognize that sentence length (SL) is the most influential factor in aDL. The overall decreasing trend in both mean SL and mean aDL (without controlling for their interaction) suggests that sentence complexity has primarily decreased due to shorter sentence structures. Shorter sentences are generally easier to process than longer ones.

Second, averaging DL per sentence conceals the distinct factors contributing to this aggregate measure. In our case, subordination and premodification both increase aDL on average per sentence. This reflects a broader principle: Any construction that creates an uneven distribution of syntactic dependents relative to the syntactic head tends to increase DL and thus aDL. However, from a processing standpoint, it remains unclear which of these syntactic configurations is inherently more difficult to process.

From a qualitative perspective, we argue that subordination typically conveys more explicit syntactic relations between constituents (e.g., *a new project that focuses on the separation of electromagnetic uranium isotopes*), which may make such structures more accessible to readers. In contrast, highly premodified noun phrases (e.g., *a new electromagnetic uranium isotope separation project*) may be more compact but cognitively demanding due to covert relations that need to be inferred relying on expert knowledge. While we can assume that experts and non-experts differ widely in processing these constructions, ultimately, behavioral

research is required to determine the relative processing difficulty of these configurations.

## 7 Conclusion

We presented a corpus spanning 300+ years of scientific English annotated with UD and DL. We discussed Stanza’s accuracy on our historical data, finding increasing accuracy for more recent data, as well as more accurate parses for shorter sentences.

Our analyses show a generally expected decrease in aDL in scientific English until 1900, while scientific texts from the 20<sup>th</sup> c. display a significant increase in aDL. The initial aDL decrease can be attributed to a decline in subordinate clauses and coordination, as shown by decreasing usage of long (mostly clausal) relations such as relative and adverbial clauses (including their core arguments) as well as coordinate structures becoming much less frequent, paired with an even distribution of information left and right of the nominal head. These results are largely in line with prior observations based on different parsers by Juzek et al. (2020) and Krielke (2024).

The significant increase in aDL in the 20<sup>th</sup> c. seems to be attributed to a rising usage of compounds as well as attributive adjectives expanding the entire length of the nominal group. It is precisely the leftward expansion with premodifying elements increasing in frequency and length that drive the overall aDL increase observed, which was confirmed by comparing aDL calculated with compounds as MWEs and as single words revealing a substantial increase for compounds as MWEs even when controlling for SL.

Our results reflect a shift of complexity from clausal into phrasal structures. Being perfectly in line with previous work on the development of scientific English (e.g. Biber and Gray, 2011, 2016), the shift towards phrasal complexity is also associated with optimization through densification and Dependency Length Minimization, while clausal complexity is rather associated with longer DL. However, our study has shown that, in fact, both clausal and phrasal complexity ultimately have similar effects on aDL. The expansion effect of extremely premodified noun phrases due to excessive compound usage on the overall aDL represents a valuable insight into the development of scientific English syntax and its implications for aDL. We have thus shown that syntactic complexity can be triggered by different syntactic renderings, i.e. on

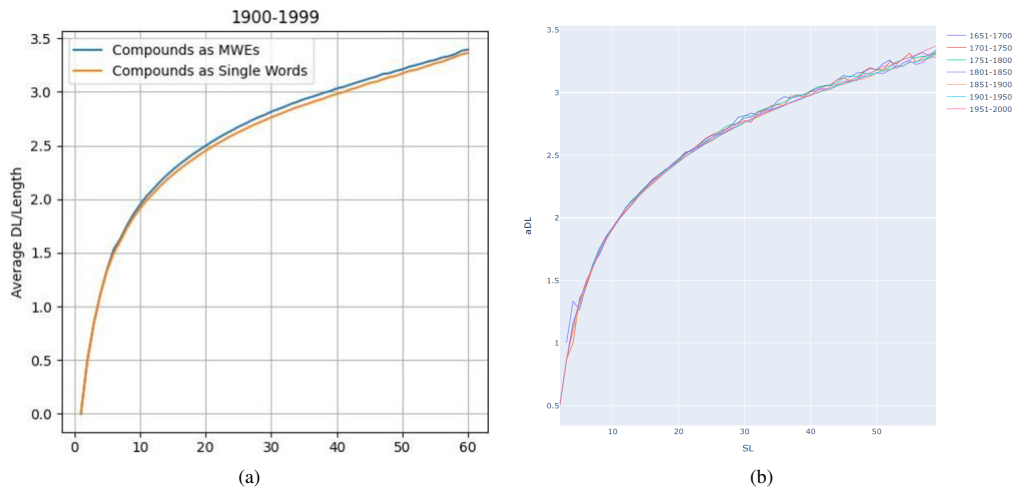


Figure 8: a) aDL per SL in the 20<sup>th</sup> c. with compounds as MWEs vs. single words. b) aDL per 50-year period calculated with compounds as single words (cf. Figure 7).

the phrasal and the causal level Biber and Gray (2016, p.62), and both affect aDL. Hence, phrasal compression through premodification of the nominal head does not necessarily minimize aDL but can rather lead to aDL expansion if used excessively.

As future work, we intend to go beyond DL to analyze trade-offs in syntactic complexity over time (e.g., tree depth, intervener complexity), and to incorporate measures from constituency-parsed corpora (e.g., average branching factor).

## 8 Limitations

This study includes only time as a predictor, despite considerable variation in text type, author, and topic. Future work should control for these factors to assess the robustness of temporal effects. Additionally, the analysis is limited to scientific texts; comparing multiple genres would clarify whether the observed trend is genre-specific or indicative of a broader pattern in English.

## Acknowledgements

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

## References

- Diego Alves, Stefan Fischer, Stefania Degaetano-Ortlieb, and Elke Teich. 2024. [Multi-word expressions in English scientific writing](#). In *Proceedings of the 8th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature (LaTeCH-CLfL 2024)*, pages 67–76.
- David Banks. 2008. *The development of scientific writing: Linguistic features and historical context*. Equinox.
- Alistair Baron and Paul Rayson. 2008. [VARD 2: A tool for dealing with spelling variation in historical corpora](#). In *Proceedings of the Postgraduate Conference in Corpus Linguistics*.
- Douglas Biber and Victoria Clark. 2002. [Historical shifts in modification patterns with complex noun phrase structures](#). *Teresa Fanego, Maria López—Couso and Javier Perez—Guerra (eds.) English Historical Morphology. Selected Papers from*, 11:43–66.
- Douglas Biber and Bethany Gray. 2010. [Challenging stereotypes about academic writing: Complexity, elaboration, explicitness](#). *Journal of English for Academic Purposes*, 9(1):2–20.
- Douglas Biber and Bethany Gray. 2011. [Grammatical change in the noun phrase: the influence of written language use](#). *English Language and Linguistics*, 15(2):223–250.
- Douglas Biber and Bethany Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Studies in English Language. Cambridge University Press.
- Yuri Bizzoni, Stefania Degaetano-Ortlieb, Peter Fankhauser, and Elke Teich. 2020. [Linguistic Variation and Change in 250 years of English Scientific](#)

- Writing: A Data-driven Approach. *Frontiers in Artificial Intelligence, section Language and Computation*.
- Jinho D Choi, Joel Tetreault, and Amanda Stent. 2015. It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 387–396.
- 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, Hannah Kermes, Ashraf Khamis, and Elke Teich. 2016. An information-theoretic approach to modeling diachronic change in scientific English. In Carla Suhr, Terttu Nevalainen, and Irma Taavitsainen, editors, *Selected papers from VARIENG – from data to evidence (d2e)*, Language and computers. Brill.
- Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.
- Andrew Thomas Dyer. 2023. Revisiting dependency length and intervener complexity minimisation on a parallel corpus in 35 languages. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 110–119.
- Stefan Fischer, Katrin Menzel, Jörg Knappen, and Elke Teich. 2020. The Royal Society Corpus 6.0 providing 300+ years of scientific writing for humanistic study. In *Proceedings of the conference on language resources and evaluation (LREC)*.
- 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, Paris, France. Association for Computational Linguistics.
- 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.
- Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.
- Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, pages 95–126.
- Edward Gibson, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, 23(5):389–407.
- Daniel Gildea and David Temperley. 2010. Do Grammars Minimize Dependency Length? *Cognitive Science*, 34(2):286–310.
- Kristina Gulordava, Paola Merlo, and Benoit Crabbé. 2015. Dependency length minimisation effects in short spans: a large-scale analysis of adjective placement in complex noun phrases. 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 477–482. Association for Computational Linguistics.
- Michael A. K. Halliday. 1988. On the language of physical science. In Mohsen Ghadessy, editor, *Registers of written English: Situational factors and linguistic features*, pages 162–177. Pinter.
- Michael A.K. Halliday and James R. Martin. 1993. *Writing science: Literacy and discursive power*. Falmer Press.
- John A. Hawkins. 1994. *A performance theory of order and constituency*. Cambridge University Press.
- John A. Hawkins. 2004. *Efficiency and Complexity in Grammars*. Oxford University Press.
- Marianne Hundt, David Denison, and Gerold Schneider. 2012. Relative complexity in scientific discourse. *English Language and Linguistics*, 16(2):209–240.
- Tom S Juzek, Marie-Pauline Krielke, and Elke Teich. 2020. Exploring diachronic syntactic shifts with dependency length: the case of scientific English. In *Proceedings of the fourth workshop on universal dependencies (UDW 2020)*, pages 109–119.
- Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the association for computational linguistics*, pages 423–430.
- Marie-Pauline Krielke. 2024. Cross-linguistic Dependency Length Minimization in scientific language: Syntactic complexity reduction in English and German in the Late Modern period. *Languages in Contrast*, 24(1):133–163.
- Marie-Pauline Krielke, Luigi Talamo, Mahmoud Fawzi, and Jörg Knappen. 2022. Tracing Syntactic Change in the Scientific Genre: Two Universal Dependency-parsed Diachronic Corpora of Scientific English and German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4808–4816. European Language Resources Association.



- Lei Lei and Ju Wen. 2020. [Is dependency distance experiencing a process of minimization? A diachronic study based on the State of the Union addresses.](#) *Lingua*, 239:102762.
- 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.
- Xueying Liu, Haoran Zhu, and Lei Lei. 2022. [Dependency distance minimization: a diachronic exploration of the effects of sentence length and dependency types.](#) *Humanities and Social Sciences Communications*, 9(1):1–9.
- 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*.
- Jesús Romero-Barranco. 2020. [Linguistic Complexity across Two Early Modern English Scientific Text Types.](#) *Atlantis*, 42(2):50–71.
- Helmut Schmid. 1994. [Probabilistic Part-of-Speech Tagging Using Decision Trees.](#) In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.
- Milan Straka and Jana Straková. 2017. [Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with UDPipe.](#) In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.
- Elke Teich, Peter Fankhauser, Stefania Degaetano-Ortlieb, and Yuri Bizzoni. 2021. [Less is More/More Diverse: On The Communicative Utility of Linguistic Conventionalization.](#) *Frontiers in Communication*, 5:142.
- David Temperley. 2008. [Dependency-length minimization in natural and artificial languages.](#) *Journal of Quantitative Linguistics*, 15(3):256–282.
- David Temperley and Daniel Gildea. 2018. [Minimizing syntactic dependency lengths: Typological/cognitive universal?](#) 4(1):67–80.
- Harry Tily. 2010. [The role of processing complexity in word order variation and change.](#)
- Gui Wang, Hui Wang, Xinyi Sun, Nan Wang, and Li Wang. 2023. [Linguistic complexity in scientific writing: A large-scale diachronic study from 1821 to 1920.](#) *Scientometrics*, 128(1):441–460.
- Yaqin Wang and Haitao Liu. 2017. [The effects of genre on dependency distance and dependency direction.](#) *Language Sciences*, 59:135–147.
- Himanshu Yadav, Ashwini Vaidya, and Samar Husain. 2017. [Understanding constraints on non-projectivity using novel measures.](#) In *Proceedings of the fourth international conference on dependency linguistics (depling 2017)*, pages 276–286.
- George Kingsley Zipf. 1949. *Human behavior and the principle of least effort: An introduction to human ecology.* Addison-Wesley Press.

# Modeling Syntactic Dependencies in Southern Dutch Dialects

Loic De Langhe, Jasper Degraeuwe, Melissa Farasyn, Véronique Hoste

Ghent University, Belgium

Groot-Brittanniëlaan 45, 9000 Ghent, Belgium

firstname.lastname@ugent.be

## Abstract

Dependency parsing of non-normative language varieties remains a challenge for modern NLP. While contemporary parsers excel at standardized languages, dialectal variation – for instance in function words, conjunctives, and verb clustering – introduces syntactic ambiguity that disrupts traditional parsing approaches. In this paper, we conduct a quantitative evaluation of syntactic dependencies in Southern Dutch dialects, leveraging a standardized dialect corpus to isolate syntactic effects from lexical variation. Using a neural biaffine dependency parser with various mono- and multilingual transformer-based encoders, we benchmark parsing performance on standard Dutch, dialectal data, and mixed training sets. Our results demonstrate that incorporating dialect-specific data significantly enhances parsing accuracy, yet certain syntactic structures remain difficult to resolve, even with dedicated adaptation. These findings highlight the need for more nuanced parsing strategies and improved syntactic modeling for non-normative language varieties.

## 1 Introduction

Modern language models demonstrate impressive mastery of both semantics and syntax across most well-studied languages. This success is largely due to abundant training data or effective transfer learning methods (Weiss et al., 2016), which ensure strong alignment for mid-to-high-resource languages. However, processing non-normative variants such as dialects and code-switched speech remains a significant challenge (Jørgensen et al., 2015). In this paper, we present a comprehensive evaluation and analysis of syntactic dependencies in various Southern Dutch language varieties based on a sizable human-annotated corpus of transcribed and to a certain extent standardized dialect speech (Breitbarth et al., 2020; Ghyselen et al., 2020; Breitbarth et al., 2024).

The Southern Dutch dialect (SDD) group encompasses dialects spoken in (i) Dutch-speaking Belgium, (ii) the three southern provinces of the Netherlands (Limburg, Noord-Brabant and Zeeland) and (iii) the Flemish-speaking dialect region in France (Farasyn et al., 2022). A geographical situation of the area, including the main dialect variants, can be found in Figure 1. SDDs can be grouped into four major varieties. Flemish includes West Flemish, East Flemish, Zeeland Flemish (spoken in Zeeland Flanders, i.e. south of the Westerschelde in the province of Zeeland, excluding the Land van Hulst), and the nearly extinct French Flemish (spoken in northern France). Zeelandic is spoken in the other areas of Zeeland. Brabantic covers North Brabant, Antwerp, and Flemish Brabant, while Limburgish is spoken in Belgian and Dutch Limburg.

In this study, the term dialect refers specifically to historically established regional varieties, distinct from *tussentaal* (‘interlanguage’). *Tussentaal* is a linguistic phenomenon where regional speech varieties gradually converge toward the standard language, leading to dialect erosion. Unlike traditional dialects, which have well-defined grammatical, phonological, and lexical features, *tussentaal* blends regional and standard elements (De Caluwe, 2009). While dialect knowledge in the Low Countries has steadily declined since the 1980s, *tussentaal* has not. Recent research shows that *tussentaal* itself displays both regional and social variation, functioning as a stratified cluster of varieties rather than a single intermediate form. At the same time, certain features show signs of supraregional stabilization, particularly in informal spoken registers, with both diversification and convergence depending on social context and speaker group (De Caluwe et al., 2013; Ghyselen, 2015).

Despite the decline in dialect use over the past decades, the linguistic diversity within the original dialect varieties remains a rich and interesting area



Figure 1: Geographical distribution of spoken dialects in the region, highlighting linguistic boundaries and areas of dialect convergence (Farasyn et al., 2022)

of study, both from a historical and computational perspective. The main variants exhibit both shared and individual syntactic particularities, which are not found in the standard Dutch language (Barbiers et al., 2005).

This paper is structured as follows: we first examine key syntactic features of SDDs in Section 2. Next, we benchmark a neural dependency parser with various mono- and multilingual transformer encoders (Section 3). We evaluate models trained on standard Dutch, dialect data, and both combined, focusing on four dialects: West Flemish, East Flemish, Brabantian, and Zeeland Flemish. Our results show that incorporating dialect data significantly improves parser performance, though some syntactic patterns remain challenging, and certain dialects pose difficulties even with dedicated training data (Section 4).

## 2 Related Work

### 2.1 Dependency Parsing

Dependency parsing has always been a popular research topic in the Dutch language domain. Early formal approaches struggled with cross-serial dependencies, illustrating the limitations of context-free grammars (Bresnan et al., 1987). The development of the Alpino Dependency Treebank (Van der Beek et al., 2002; Van Noord, 2006) provided a crucial resource that played a key role in the advancement of rule-based and statistical parsers, driving further progress in the field.

The rise of Universal Dependencies (UD) (Nivre et al., 2016) standardized Dutch dependency annotation, fostering cross-linguistic research and improving multilingual parsing. Dutch UD treebanks,

such as LassySmall UD and Alpino UD (Bouma and van Noord, 2017), have supported data-driven models, including transition-based (Nivre et al., 2006) and graph-based (McDonald et al., 2006) approaches. More recently, biaffine dependency parsing (Dozat and Manning, 2016) combined with transformer encoders such as BERTje (De Vries et al., 2019) and RobBERT (Delobelle et al., 2020) have achieved state-of-the-art results for Dutch. Additionally, more advanced techniques such as self-distillation (de Kok and Pütz, 2020) have also been proposed as methods for further enhancing modern-day neural parsers.

Dependency parsing of non-normative language, such as dialects and code-switching, is challenging due to linguistic variability and scarce annotated data (Jørgensen et al., 2015). While some methods use domain adaptation and transformer models to improve robustness (Jørgensen et al., 2016; Nguyen et al., 2020), processing dialect remains a largely open problem. For Dutch specifically, UD-based dialect treebanks (Braggaar and van der Goot, 2021) and transfer learning have enhanced parsing for northern regional varieties and informal registers (Braggaar and Van Der Goot, 2021). While some progress has been made with these language variants, the limited availability of data remains a significant obstacle. In our own work on SDDs, we aim to tackle this challenge by incorporating the substantial (and partly standardized) GCND corpus (Breitbarth et al., 2024), which helps mitigate two major concerns that commonly affect studies on non-normative language: spelling variation and limited data availability. The corpus includes two transcription layers, both of which use Dutch-based

orthographic conventions to reduce phonological variation. The first layer remains dialectal in morphology, syntax, and vocabulary, while the second ‘dutchified’ layer adds light morphological and lexical normalization to facilitate readability and automatic processing (Ghyselen et al., 2020). All parsing experiments in this paper were conducted on the second transcription layer.

## 2.2 Syntactic Variations in Southern Dutch Dialects

SDDs exhibit considerable syntactic variation (Barbiers et al., 2005), posing challenges for traditional parsers. These dialects differ from standard Dutch in areas such as word order, the usage and placement of function words, pronominal paradigms, negation, complementizer systems and regionally specific conjunctives, all of which result in widely varying parse trees. Additionally, clitic doubling and verb cluster variation further challenge parsing, as standard models struggle to map these structures onto expected patterns. In the paragraphs below, we address six linguistic phenomena that are likely to significantly hinder the performance of a parser trained on standard language.

**Subject doubling (or tripling)** is a phenomenon where the subject of a sentence occurs multiple times, typically involving a combination of a pronoun and a full noun phrase or two pronouns (De Vogelaer, Gunther, 2006). This construction is particularly common in the SDDs, including West Flemish, East Flemish, Brabantic, and Limburgish varieties (Van Craenenbroeck and Van Koppen, 2002). Subject doubling often occurs as part of topic marking or emphasis, distinguishing these dialects from standard Dutch, where such constructions are ungrammatical.

### 1. **Ze** werkt **zij** in Brussel

*EN: She works (she) in Brussels*

Here, the first subject (*Ze*) serves as a topic, while the second subject (*zij*) functions as a resumptive pronoun. In some dialects, particularly in West Flemish and Brabantian, subject doubling can extend even further to subject tripling, where a noun phrase is followed by two pronouns (De Vogelaer and Devos, 2008).

**Negation doubling and tripling** is another syntactic feature observed in various southern dialect varieties (Haegeman and Zanuttini, 1996). This

phenomenon involves the repetition of negation markers within a single sentence, often used for emphasis or to express stronger negation.

### 2. Ik heb dat **nooit niet** gedaan.

*EN: I never (not) did that*

Here, the negation particle *niet* (not) is doubled with the negative indefinit *nooit* (never), which intensifies the negation beyond what is found in standard Dutch. In some dialects, negation can be tripled, further emphasizing its intensity. The phenomenon of negation stacking in dialects has been thoroughly discussed in the literature (Paardekooper, 2015; De Schutter, 2015) and contrasts with the grammatical use where two negative elements typically cancel each other out, creating a positive meaning.

**En negation** In many Southern Dutch dialects, negation can involve a preverbal particle *en*, in addition to a negation particle like *niet* or a negative indefinite like *nooit*. This results in negation doubling or tripling, depending on the combination. For instance, a sentence such as:

### 3. Ik **en** zie niets.

*EN: I (and) don't see anything*

The *en* particle is a remnant of the historical sentential negator, now reinterpreted as a discourse-related element (Breitbarth and Haegeman, 2014, 2015). Its surface form coincides with the standard coordinating conjunction *en* purely by chance since it has a distinct syntactic origin and function. The construction remains productive in many Southern Dutch dialects, particularly in Flemish varieties (Neuckermans, 2008; Barbiers et al., 2007). It poses a challenge for automatic parsing due to its surface ambiguity and non-standard word order.

**The expletive *dat*** (that) can serve as an expletive in spoken dialect Dutch, usually following interrogative pronouns, relative pronouns or subordinating conjunctions, resulting in so-called complementizer doubling, as it overlaps in form with the standard complementizer (Bacskai-Atkari, 2020; Barbiers, 2009). While this complementizer does not change sentence meaning, it does change the word order compared to standard Dutch.

### 5. Ik weet niet waar **dat** hij is.

*EN: I don't know where (that) he is*



This linguistic feature occurs frequently across all Flemish regions, with the notable exception of Southeast Limburg (Barbiers et al., 2007; Taelde-man, 2008).

**Distinct comparative conjunctions** occur widely in the SDDs (Rooy, 1965). Instead of the standard Dutch *dan* (‘than’), many varieties use *als* (‘as’) or *of* (‘or’) in comparative constructions, as in:

4. Hij is groter **als** jou.

*EN: He is bigger than you*

This variation is well attested across dialect areas (see also (Postma, 2006)). The forms are fully grammatical in their dialects, but are likely to confuse parsers trained on standard Dutch.

**Deviating clause introductions** are a final phenomenon commonly observed in dialects and informal speech. In standard Dutch, non-finite clauses with a to-infinitive are introduced either by the complementizer *om* or by a null element. *Om* is mandatory in conditional clauses but optional when the clause functions as a true subject, direct object, or postmodifier of a noun (Vandeweghe, 1971). In non-standard Flemish registers, *voor* and *van* are often used as an alternative to introduce non-finite clauses with a to-infinitive (Barbiers et al., 2005).

6. Ze deed dat **voor** beter te horen.

*EN: She did that to (for) hear better*

### 3 Experiments

#### 3.1 Data

##### 3.1.1 Standard Dutch

The standard Dutch portion of our data consists of two benchmark corpora, annotated for syntactic dependencies. First, the Lassy-Small Corpus (Van Noord et al., 2013) contains a total of 1 million words sourced from the larger D-COI (50 million words) corpus (Oostdijk, 2006). The second part of our standard dataset consists of the Alpino treebank, a collection of over 150,000 words of newspaper data (Van der Beek et al., 2002). Unlike the Lassy Corpus, the Alpino treebank was synthetically created through the use of the eponymous Alpino parser, a HPSG-based linguistic analysis tool for parsing Dutch text (Van Noord, 2006).

For both corpora we use the official UD versions (Bouma and van Noord, 2017) which are made available through the Universal Dependencies project (Nivre et al., 2016).

##### 3.1.2 Southern (Dialect) Dutch

Our dialect data is a subset of the larger Corpus of Spoken Dutch Dialects (GCND) (Breitbarth et al., 2020), which is part of the Voices of the Past project (Hellebaut et al., 2021). This data originates from a collection of audio-recorded interviews with native speakers, conducted over a 13-year period (1963–1976) (Hellebaut et al., 2021). In total, the project amassed over 700 hours of spoken dialect material from more than 500 distinct locations across Flanders and the Netherlands. In recent years, conservation efforts have ensured that the majority of this audio material has been transcribed and normalized for spelling (Ghyselen et al., 2020). As part of this project, a portion of the data has been annotated with POS tags and syntactic dependencies following the Alpino guidelines (Breitbarth et al., 2020). For the annotation of syntactic dependencies this process follows a two-step approach: first, transcriptions are processed using the Alpino syntactic parser, and then they undergo manual correction by human annotators (Farasyn et al., 2022). At this stage, the syntactic annotations adhere to the original Alpino annotation guidelines. To create a usable corpus in UD format, we follow the same process <sup>1</sup> as outlined in previous work (Bouma and van Noord, 2017).

We have access to a total of 26,146 sentences from four dialect (sub)groups: Zeeland Flemish (4,141 sentences), Brabantian (4,496 sentences), East Flemish (8,674 sentences), and West Flemish (8,687 sentences). In these sentences, the language has been standardized to minimize lexical interference, allowing us to focus on syntactic patterns specific to these dialects, such as the ones discussed earlier in Section 2.2. Since dialectal varieties do not exhibit identical syntactic patterns, it is useful to first provide a high-level overview of how these characteristics vary across the different dialects. Figure 2 presents the distribution with which each characteristic appears in each of the four dialects.

#### 3.2 Methodology

We train several versions of Diaparser (Attardi et al., 2021), a transformer-based extension on the standard deep biaffine Dependency Parser by Dozat and Manning (2016). In this framework, the encoder, typically a BERT-based transformer model, generates contextualized token embeddings that are fed into a biaffine network. This network then pre-

<sup>1</sup><https://github.com/rug-compling/alud>

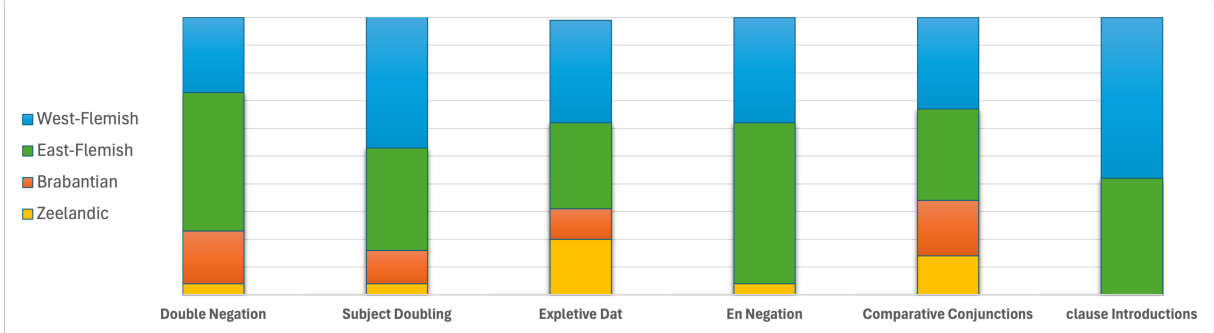


Figure 2: Syntactic pattern distribution across the four dialects.

Training Dataset	Encoder	UCM	LCM	UAS	LAS
Standard	<i>BERTje</i>	62.26	47.42	82.01	75.46
	<i>RobBERT-2023</i>	61.99	54.16	82.45	75.30
	<i>ModernBERT</i>	52.70	37.63	75.32	66.93
	<i>mBERT</i>	57.02	40.88	78.99	71.33
Dialect	<i>BERTje</i>	70.98	58.28	88.12	83.11
	<i>RobBERT-2023</i>	71.66	57.78	88.14	83.11
	<i>ModernBERT</i>	66.73	53.38	85.69	80.03
	<i>mBERT</i>	70.75	57.02	87.22	82.22
Merged	<i>BERTje</i>	74.07	62.91	89.07	85.17
	<i>RobBERT-2023</i>	73.77	60.61	89.26	84.86
	<i>ModernBERT</i>	71.59	57.48	87.84	82.84
	<i>mBERT</i>	72.89	59.66	88.46	83.85

Table 1: Main experimental results. Standard training data includes the Alpino and Lassy corpora; Dialect training data refers to the GCND corpus; the Merged dataset combines all three.

dicts the syntactic tree by simultaneously modeling both head and label prediction tasks. We select a variety of underlying encoder models to be tested on our data. These include monolingual models such the standard Dutch BERTje (De Vries et al., 2019) and the RoBERTa-based RobBERT-2023, a more recent Dutch language model (Delobelle et al., 2020). Additionally, we evaluate a benchmark multilingual encoder (mBERT) (Devlin et al., 2019) and the recently developed modernBERT (Warner et al., 2024).

For evaluation, we use a set of parsing metrics to assess model performance: Unlabeled Complete Match (UCM), Labeled Complete Match (LCM), Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS). UCM and LCM are strict metrics that evaluate whether the entire predicted tree matches the gold-standard tree, with LCM requiring both the structure and labels to match exactly (Lücking et al., 2024). UAS measures the

percentage of tokens with the correct head in the dependency tree, while LAS also considers the correct syntactic labels (Nivre et al., 2004).

### 3.3 Results

Table 1 presents a comprehensive overview of our main experimental results. The inclusion of dialect data leads to considerable performance gains. The best-performing model, BERTje, achieves an average improvement of 8.33% on the dialect dataset across all evaluated metrics, compared to a parser trained solely on standard (normative) Dutch. Furthermore, the best performance (mean improvement of 11.01%) is achieved when combining the both standard datasets (Alpino and Lassy) with dialect sentences, creating a well-balanced corpus that enhances the model’s ability to process non-normative language fluently.

As expected, Dutch encoders outperform their multilingual counterparts, likely due to their spe-

cialized pretraining on Dutch linguistic structures. A key factor contributing to their strong performance is the absence of significant spelling variation, as the dialect datasets were normalized prior to training. While some dialect-specific vocabulary remains, tokenization issues appear to be minimal, especially compared to studies on less standardized dialects, where inconsistent orthography often hinders model performance (Jørgensen et al., 2015).

## 4 Discussion and Analysis

### 4.1 Performance and Syntactic Variation

While many Natural Language Processing studies on dialectal and non-normative data focus on lexical and spelling variation, our primary interest lies in how syntactic variation affects model performance. The SDDs discussed in this paper exhibit several distinct syntactic patterns that are different compared to standard Dutch, posing unique challenges for parsing models. For this analysis, we base ourselves on the most commonly observed characteristics of the SDD group, as described in Section 2.2.

To evaluate whether the trained models can handle these syntactic variations, we begin by filtering sentences that exhibit these patterns from the entire test set ( $n = 2,616$ ). This allows us to create smaller partitions focused on specific syntactic variations for targeted evaluation. We examine sentences that feature the following patterns: subject doubling ( $n = 162$ ), negation doubling ( $n = 43$ ), *en* negation ( $n = 24$ ), deviating comparative conjunctions ( $n = 30$ ), expletive *dat* ( $n = 35$ ) and deviating clause introductions ( $n = 14$ ).

For clarity and comprehensiveness, this discussion is limited to the overall best-performing models from Section 3. Specifically, we focus on the BERTje-based model across all three dataset partitions and evaluate its performance on sentences exhibiting specific syntactic patterns. Table 2 presents the results of model performance and its improvement across different training setups.

Overall, we observe a consistent and significant improvement across all categories when comparing the standard Dutch parser to the dialect-tuned models. As in the main experiments, a combination of standard and dialect Dutch leads to the highest overall performance. The most notable gains occur in areas affecting negation – specifically, negation doubling and *en* negation. This suggests that while these syntactic irregularities pose challenges for

normative language models, it is possible to align them more effectively to handle such data with minimal additional training resources.

An additional challenge for standard models arises in processing alternative comparative conjunctions. As discussed in Section 2.2, certain dialect varieties use markers such as *of* and *als* – typically reserved for disjunctions and conditions respectively – as comparative elements. Unsurprisingly, models trained on normative Dutch struggle with parsing and interpreting sentence structure when these markers are present. However, as with the cases described above, incorporating even a relatively small number of such instances significantly improves the model’s ability to handle this syntactic variation more effectively.

### 4.2 Geographical Variation

The final part of our discussion consists of a formal analysis of the systems per larger dialect group. We split the test set into four smaller partitions. Each partition consists entirely out of sentences from one of the four major dialect varieties in the corpus: West Flemish, East Flemish, Zeeland Flemish and Brabantic. Note that border cases are resolved according to the current provincial/national borders. Note that dividing by provincial borders is not optimal, as many nuanced border cases exist where dialects blend across regions. However, for this preliminary analysis, such a division provides a practical and sufficiently clear framework.

Following the approach outlined in the previous section, we evaluate the best-performing models from each training set (Standard, Dialect and Merged) on the newly created dialect-specific test partitions. This analysis allows us to determine whether certain parsers are better equipped to handle specific dialects or if performance varies across different dialect groups. By comparing results across these partitions, we gain insight into the models’ ability to generalize across dialectal variation. Table 3 presents the performance of each model on the various dialect groups.

As can be seen from the table, overall performance per dialect is consistent with the earlier obtained scores in Section 3 for the dataset as a whole. Once again, models trained on a combination of standard Dutch (Lassy + Alpino) and dialect (GCND) data perform best, which is consistent for each of the four evaluated dialect groups. It should be noted, though, that East Flemish performs markedly worse than the other language va-

Pattern	Dataset	UCM	LCM	UAS	LAS
Subject Doubling	Standard	52.47	20.99	81.91	72.49
	Dialect	60.49	40.12	88.36	82.77
	Merged	67.90	46.30	89.85	84.83
Negation Doubling	Standard	34.88	23.26	74.44	65.31
	Dialect	55.81	41.86	85.99	80.93
	Merged	58.14	46.51	86.95	82.73
En Negation	Standard	21.74	8.70	67.89	53.66
	Dialect	52.17	39.13	84.82	78.60
	Merged	65.22	47.83	90.31	85.27
Expletive Dat	Standard	31.43	24.81	80.38	71.63
	Dialect	28.57	25.71	85.55	82.05
	Merged	42.86	34.29	88.01	84.39
Comparative Conjunctions	Standard	31.03	17.24	69.90	62.46
	Dialect	34.48	24.14	79.37	72.70
	Merged	44.83	37.39	83.18	79.51
Clausal Introduction	Standard	42.86	14.29	83.44	69.94
	Dialect	50.00	21.43	86.36	76.70
	Merged	57.14	28.57	87.71	80.45

Table 2: Performance of the BERTje-based model on syntactic pattern-specific sentences across dataset partitions

rieties in all three training setups. We hypothesize two possible explanations for this phenomenon.

First, the training data for the underlying BERTje encoder (De Vries et al., 2019) consists entirely of Dutch as spoken in the Netherlands (i.e., non-Flemish), which may make it more challenging for the model to align with southern Flemish dialects, leading to lower scores. This hypothesis initially seems plausible, given the high performance on Zeeuws (a dialect native to the Netherlands) and Brabants (generally considered the Flemish dialect most similar to Netherlandic Dutch). However, the performance of West Flemish does not fully support this explanation. Further experimentation with the robbert-2023 model, which was pre-trained on both (dialect) Flemish and standard Dutch, does not yield significant improvements on the East Flemish dataset. The performance gap between East Flemish and the other dialects therefore remains consistent, regardless of the encoder used.

A more plausible explanation lies in the specific syntactic patterns characteristic of each dialect group. Analyzing the distribution of these syntactic features, we find that in the East Flemish test set, 14% of the sentences contain one or more of the patterns discussed in Section 2.2. This is the highest proportion among all dialect varieties (West Flemish: 13%, Zeeland Flemish: 4%, Brabant: 8%) and may account for the model’s decreased

performance. A closer examination reveals that East Flemish has the highest proportional occurrence of double negation markers – almost double that of West Flemish – as well as a high frequency of *en* negation. These are precisely the syntactic patterns on which the standard models struggled (see Section 4). Although, as shown in Table 3, performance improves somewhat with fine-tuning on dialect data, the average scores for these patterns remain among the lowest in the experiments.

## 5 Ablation Studies

While the four main dialect groups in this paper have many lexical and syntactic commonalities, they are still ways in which they are notably distinct. Our approach of training on a combined dataset that includes all dialects may unintentionally obscure important dialect-specific characteristics. A potential problem herein is that the model is encouraged to generalize across shared syntactic patterns rather than capturing fine(r)-grained variations. Therefore, an additional set of experiments of training and evaluating performance on individual dialects is crucial to understanding how well the model can preserve these finer linguistic distinctions.

Concretely, we divide our original training, development, and test sets into four smaller subsets, each corresponding to one of the four main dialect groups: West Flemish, East Flemish, Zeeland Flem-



Training Dataset	Dialect group	UCM	LCM	UAS	LAS
Standard	<i>Brabantic</i>	62.22	49.10	81.45	75.71
	<i>Zeeland Flemish</i>	65.14	50.00	84.45	77.26
	<i>West Flemish</i>	64.42	49.04	84.64	78.48
	<i>East Flemish</i>	58.90	43.78	78.65	71.64
Dialect	<i>Brabantic</i>	73.76	62.44	89.62	85.34
	<i>Zeeland Flemish</i>	70.41	56.88	89.00	82.93
	<i>West Flemish</i>	71.39	58.41	89.14	84.67
	<i>East Flemish</i>	69.54	53.86	86.09	80.64
Merged	<i>Brabantic</i>	77.15	68.10	90.50	87.24
	<i>Zeeland Flemish</i>	75.00	62.84	89.58	84.72
	<i>West Flemish</i>	75.00	64.54	90.79	87.20
	<i>East Flemish</i>	71.22	58.90	86.60	82.52

Table 3: Performance on the partitioned test set for each of the dialect varieties using various training datasets

ish, and Brabantic. Table 4 provides an overview of the dataset sizes after partitioning, detailing the distribution across the new training, development, and test sets for each group.

	# Train	# Dev	# Test
<i>West Flemish</i>	7,034	820	833
<i>East Flemish</i>	6,895	886	893
<i>Zeeland Flemish</i>	3,287	418	436
<i>Brabantic</i>	3,574	480	422

Table 4: Number of sentences present in the training, development and test sets for each of the dialects

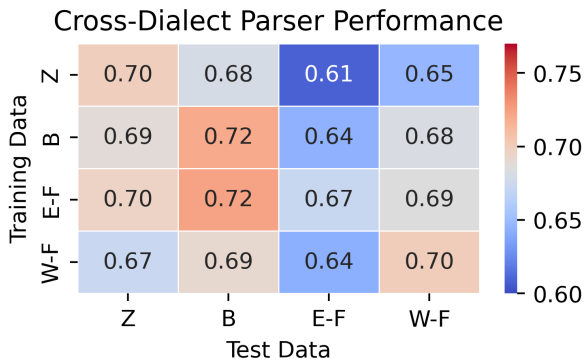


Figure 3: Cross-Dialect results for West Flemish (W-F), East Flemish (E-F), Brabantic (B) and Zeeland Flemish (Z)

Next, we train four dialect-specific parsers, each using the corresponding partitioned dataset. For each dialect, we fine-tune a parser with the optimal encoder identified in Table 1 (BERTje) and evaluate its performance in two ways: (1) on its respective dialect-specific test set and (2) on the other individual test sets to assess generalization

across dialects. In order to provide an interpretable overview, Figure 3 displays a comparative heatmap indicating cross-dialect performance. The scores presented in the figure are the average of the four metrics that were used earlier. A full overview, including all metrics can be found in the Appendix.

From the figure, we infer that irrespective of the amount of available training data, models trained on a particular dialect are consistently the best performing model on that specific evaluation data. We also find that performance for the East Flemish dialect is notably worse, which is consistent with our findings from Section 4.2.

Interesting however, is the fact that the East Flemish model itself generally performs on-par or slightly below the top models for the other dialects. We hypothesize two possible explanations for this. First, the amount of available training data is highest for East Flemish as a whole. The good performance of the model might therefore be explained simply by the fact the model had access to more training data, resulting in better alignment. We also note here that the West Flemish model, which contains a comparable amount of training data, tends to be the second best overall model, supporting this hypothesis. Another plausible explanation is that, due to the sheer number and remarkable diversity of divergent syntactic patterns present within the East Flemish dataset (as illustrated in Figure 2), the model may have developed a broader proficiency in handling non-normative linguistic structures in general. Rather than merely adapting to individual irregularities, the model could be refining its capacity to process and generate language that deviates from standard norms, thereby demonstrating an

overall advantage when working with non-standard linguistic forms. This suggests that exposure to a wide array of syntactic variations enhances the model’s flexibility in parsing, interpreting, and predicting structures that fall outside the boundaries of conventional or prescriptive grammar.

## 6 Conclusion

This paper has explored the difficulties and potential of dependency parsing for SDDs, a linguistically diverse yet underrepresented area. By using a lexically standardized corpus, we focused on syntactic variation across four key dialect groups: West Flemish, East Flemish, Brabantic, and Zeeland Flemish. Our results show that models trained on standard Dutch perform poorly on dialectal input, especially when dealing with ambiguous function words and region-specific structures. Adding dialect data to training significantly boosts performance, though some constructions remain difficult to parse accurately. Our findings suggest that more dialect-sensitive approaches in syntactic modeling could improve parsing accuracy and support the development of more adaptable NLP tools.

## References

- Giuseppe Attardi, Daniele Sartiano, and Maria Simi. 2021. Biaffine dependency and semantic graph parsing for enhanced universal dependencies. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 184–188.
- Julia Bacskai-Atkari. 2020. German v2 and doubly filled comp in west germanic. *The Journal of Comparative Germanic Linguistics*, 23(2):125–160.
- Sjef Barbiers. 2009. Locus and limits of syntactic microvariation. *Lingua*, 119(11):1607–1623.
- Sjef Barbiers, Leonie Cornips, and Jan Pieter Kunst. 2007. The syntactic atlas of the dutch dialects (sand): a corpus of elicited speech and text as an online dynamic atlas. *Creating and Digitizing Language Corpora: Volume 1: Synchronic Databases*, pages 54–90.
- Sjef Barbiers, Johan Van der Auwera, Hans Bennis, Eefje Boef, Gunther De Vogelaer, and Margreet Van der Ham. 2005. *Syntactic atlas of the Dutch dialects*, volume 2. Amsterdam University Press.
- Gosse Bouma and Gerardus van Noord. 2017. Increasing return on annotation investment: The automatic construction of a universal dependency treebank for dutch. In *Proceedings of the nodalida 2017 workshop on universal dependencies (udw 2017)*, pages 19–26.
- Anouck Braggaar and Rob Van Der Goot. 2021. Challenges in annotating and parsing spoken, code-switched, frisian-dutch data. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 50–58.
- Anouck Braggaar and Rob van der Goot. 2021. Creating a universal dependencies treebank of spoken frisian-dutch code-switched data. *arXiv preprint arXiv:2102.11152*.
- Anne Breitbarth, Melissa Farasyn, Anne-Sophie Ghyselen, Lien Hellebaut, Frederic Lamsens, Katrien Depuydt, Jesse de Does, Jan Niestadt, and Koen Mertens. 2024. Gesproken corpus van de zuidelijk-nederlandse dialecten. 1st release october 2024.
- Anne Breitbarth, Melissa Farasyn, Anne-Sophie Ghyselen, and Jacques Van Keymeulen. 2020. Het gesproken corpus van de zuidelijk-nederlandse dialecten. *Handelingen van de Koninklijke Zuid-Nederlandse maatschappij voor taal-en letterkunde en geschiedenis*, 72.
- Anne Breitbarth and Liliane Haegeman. 2014. The distribution and interpretation of preverbal *en* in flemish. In Theresa Biberauer and George Walkden, editors, *Syntax Over Time: Lexical, Morphological, and Information-Structural Interactions*, pages 35–56. Oxford University Press, Oxford.
- Anne Breitbarth and Liliane Haegeman. 2015. ‘en’ en is niet wat we dachten: A flemish discourse particle. In *Proceedings of Moscow Syntax and Semantics (MOSS) 2*, volume 78, pages 41–55, Cambridge, MA. MIT Working Papers in Linguistics.
- Joan Bresnan, Ronald M Kaplan, Stanley Peters, and Annie Zaenen. 1987. Cross-serial dependencies in dutch. In *The formal complexity of natural language*, pages 286–319. Springer.
- Johan De Caluwe. 2009. Tussentaal wordt omgangstaal in vlaanderen. *Nederlandse taalkunde*, 14(1):8–25.
- Johan De Caluwe, Steven Delarue, Anne-Sophie Ghyselen, and Chloé Lybaert, editors. 2013. *Tussentaal: Over de talige ruimte tussen dialect en standaardtaal in Vlaanderen*. Studia Germanica Gandensia. Academia Press, Gent. Spieghel Historiae, themanummer.
- Daniël de Kok and Tobias Pütz. 2020. Self-distillation for german and dutch dependency parsing. *Computational Linguistics in the Netherlands Journal*, 10:91–107.
- Georges De Schutter. 2015. Meervoudige negatie en paardekooper z’n begrip" stapeling". wat heeft de rnd te bieden? *Verslagen & Mededelingen van de Koninklijke Academie voor Nederlandse Taal en Letteren*, 125(3).

- Gunther De Vogelaer and Magda Devos. 2008. On geographical adequacy, or: How many types of subject doubling in dutch. *Microvariation in syntactic doubling*, 36:251–276.
- De Vogelaer, Gunther. 2006. *Subjectmarkering in de Nederlandse en Friese dialecten*. Ph.D. thesis, Ghent University.
- Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.
- Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: a dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*.
- 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.
- Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.
- Melissa Farasyn, Anne-Sophie Ghyselen, Jacques Van Keymeulen, and Anne Breitbarth. 2022. Challenges in tagging and parsing spoken dialects of dutch. *Journal of Historical Syntax*, 6(4-11):1–36.
- Anne-Sophie Ghyselen. 2015. ‘stabilisering’ van tussentaal? het taalrepertorium in de westhoek als casus. *Taal & Tongval*, 67(1):43–95.
- Anne-Sophie Ghyselen, Jacques Van Keymeulen, Melissa Farasyn, Lien Hellebaut, and Anne Breitbarth. 2020. Het transcriptieprotocol van het gesproken corpus van de nederlandse dialecten (gcnd). *BULLETIN DE LA COMMISSION ROYALE DE TOPONYMIE & DIALECTOLOGIE (PRINTED)= HANDELINGEN VAN DE KONINKLIJKE COMMISSIE VOOR TOPONYMIE & DIALECTOLOGIE*, 92:83–115.
- Liliane Haegeman and Raffaella Zanuttini. 1996. Negative concord in west flemish. *Parameters and functional heads. Essays in comparative syntax*, (3):117–197.
- Lien Hellebaut, Anne-Sophie Ghyselen, Melissa Farasyn, Anne Breitbarth, Veronique De Tier, and Jacques Van Keymeulen. 2021. Stemmen uit het verleden: een schat aan informatie voor heemkundigen en andere erfgoedactoren. *HISTORIES MAGAZINE*, (06/07/2021).
- Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the workshop on noisy user-generated text*, pages 9–18.
- Anna Jørgensen, Dirk Hovy, Anders Søgaard, et al. 2016. Learning a pos tagger for aave-like language. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Proceedings of the conference*. Association for Computational Linguistics.
- Andy Lücking, Giuseppe Abrami, Leon Hammerla, Marc Rahn, Daniel Baumartz, Steffen Eger, and Alexander Mehler. 2024. Dependencies over times and tools (dott). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4641–4653.
- Ryan McDonald, Kevin Lerman, and Fernando Pereira. 2006. Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 216–220.
- Neuckermans. 2008. *Negatie in de Vlaamse dialecten volgens de gegevens van de Syntactische Atlas van de Nederlandse Dialecten (SAND)*. Ph.D. thesis, Ghent University.
- Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.
- Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1659–1666.
- Joakim Nivre, Johan Hall, and Jens Nilsson. 2004. Memory-based dependency parsing. In *Proceedings of the eighth conference on computational natural language learning (CoNLL-2004) at HLT-NAACL 2004*, pages 49–56.
- Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Malt-parser: A data-driven parser-generator for dependency parsing. In *LREC*, volume 6, pages 2216–2219.
- Nelleke HJ Oostdijk. 2006. A reference corpus of written dutch. *Corpus design (D-coi 06-01). Persistent identifier: urn: nbn: nl: ui*, pages 22–2066.
- Petrus Cornelis (Piet) Paardekooper. 2015. Meer-voudige uitdrukking (stapel) van negatie, vooral in het nederlands. *Verslagen & Mededelingen van de Koninklijke Academie voor Nederlandse Taal en Letteren*, 125(1-2).
- Gertjan Postma. 2006. Van’ groter dan’ naar’ groter als’ structurele oorzaken voor het verval van het comparatieve voegwoord’dan’. *Nederlandse Taalkunde*, 11(1):2–22.

- J de Rooy. 1965. *Als-of-dat: een semantisch-onomasiologische studie over enkele subordinerende conjuncties in het ABN, de Nederlandse dialecten en het Fries, vergelijkend-synchronisch beschouwd*. Ph.D. thesis, Assen: Van Gorcum.
- Johan Taeldeman. 2008. Zich stabiliserende grammaticale kenmerken in vlaamse tussentaal. *Taal & Tongval*, 60(2).
- Jeroen Van Craenenbroeck and Marjo Van Koppen. 2002. Subject doubling in dutch dialects. In *Proceedings of Console IX*, pages 54–67. Citeseer.
- Leonor Van der Beek, Gosse Bouma, Rob Malouf, and Gertjan Van Noord. 2002. The alpino dependency treebank. In *Computational linguistics in the Netherlands 2001*, pages 8–22. Brill.
- Gertjan Van Noord. 2006. At last parsing is now operational. In *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles. Conférences invitées*, pages 20–42.
- Gertjan Van Noord, Gosse Bouma, Frank Van Eynde, Daniel De Kok, Jelmer Van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. Large scale syntactic annotation of written dutch: Lassy. *Essential speech and language technology for Dutch: results by the STEVIN programme*, pages 147–164.
- Willy Vandeweghe. 1971. Om en rond de (om) tekonstruktie. *Studia Germanica Gandensia*, 13:37–41.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.
- Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data*, 3:1–40.

## A Appendix

<b>Train Dialect</b>	<b>Test Dialect</b>	<b>UCM</b>	<b>LCM</b>	<b>UAS</b>	<b>LAS</b>
Zeeuws	Zeeuws	66.06	49.77	85.54	78.39
	Brabants	62.90	50.68	82.81	76.26
	Oost-vl	56.10	38.63	78.35	70.24
	West-vl	58.23	41.71	83.18	75.44
Brabants	Zeeuws	66.06	49.31	84.18	75.90
	brabants	68.33	53.39	86.10	80.15
	Oost-vl	61.03	42.55	80.79	73.25
	West-vl	62.74	49.04	84.07	77.19
Oost-vl	Zeeuws	66.51	49.31	85.46	77.26
	Brabants	67.87	55.20	86.25	80.26
	Oost-vl	63.72	46.96	82.72	75.80
	West-vl	62.14	48.08	85.22	78.56
West-vl	Zeeuws	64.22	44.72	84.45	75.98
	Brabants	64.93	50.90	83.75	77.18
	Oost-vl	59.57	52.78	81.03	73.75
	West-vl	64.64	49.16	86.42	80.04

Table 5: Full results of ablation experiments.

# Assessing the Agreement Competence of Large Language Models

**Alba Táboas García**  
NLP Group  
Pompeu Fabra University  
alba.taboas@upf.edu

**Leo Wanner**  
Barcelona Supercomputing Center &  
Catalan Institute for Research  
and Advanced Studies (ICREA)  
leo.wanner@bsc.es

## Abstract

While the competence of LLMs to cope with agreement constraints has been widely tested for English, only a very limited number of works deals with morphologically rich(er) languages. In this work, we experiment with 25 mono- and multilingual LLMs, applying them to a collection of more than 5,000 test examples that cover the main agreement phenomena in three Romance languages (Italian, Portuguese, and Spanish) and one Slavic Language (Russian). We identify which of the agreement phenomena are most difficult for which models and challenge some common assumptions of what makes a good model. The test suites into which the test examples are organized are openly available and can be easily adapted to other agreement phenomena and other languages for further research.

## 1 Introduction

Agreement is one of the linguistic phenomena usually invoked to illustrate dependency in language (Mel'čuk, 2009). It reflects the fact that, within a sentence, the wordform  $w_2$  (the *target*) co-varies with a wordform  $w_1$  (the *controller*) with respect to selected morpho-syntactic features.<sup>1</sup> In English, agreement is most obvious in the number covariance in subject–verb constructions. Therefore, it is not surprising that compliance with subject–verb agreement restrictions has become a key diagnosis for the syntactic competence of neural language models. It allows researchers to evaluate whether a model truly captures hierarchical structure rather than merely learning surface-level patterns (Linzen et al., 2016; Goldberg, 2019; Nastase et al., 2024). However, in morphologically rich(er) languages, agreement is a considerably more prominent phenomenon than in English. *Canonical agreement*

<sup>1</sup>Mel'čuk (1993) makes a distinction between the phenomena of *agreement*, *government*, and *congruence*. In our study, we refrain from such a detailed differentiation.

*features* include not only number, but also person, gender, and grammatical case; certain numerals as well as possessive and qualitative adjectives can also act as controllers; and among targets, in addition to verbs and adjectives as in English, we also find pronouns, numerals, adverbs, adpositions, nouns, etc. (Corbett, 2006). The goal of our work is to assess the competence of state-of-the-art neural models in handling a broader range of agreement features than encountered in English. To this end, we selected three Romance languages: Italian, Portuguese and Spanish, and Russian as a representative of Slavic languages. For both the Romance languages triple and for Russian, we define ten different agreement tests. These tests are applied to a number of monolingual and multilingual models. Our experiments show that despite a good overall performance, language models struggle with complex agreement constructions, with monolingual models outperforming multilingual ones. But model size and training data size, which are typically correlated with model performance, do not play a significant role in this case.

## 2 Related Work

Linzen et al. (2016) were among the first to propose using a model's assignment of higher probability to targets with the correct number grammeme (as opposed to the incorrect one) in subject–verb agreement in English as a benchmark for evaluating its syntactic competence. This test methodology, known as *targeted syntactic evaluation*, has since then been extended to test other syntactic structures in English, such as, e.g., reflexive anaphora agreement and licensing of negative polarity items (Marvin and Linzen, 2018), filler-gap constructions and island constraints (Wilcox et al., 2018, 2019), garden path effects (Futrell et al., 2018, 2019), or all of the above (Hu et al., 2020).

Early work on LSTMs includes studies on num-



ber and case agreement in Basque (Ravfogel et al., 2018) and long-distance number agreement in Italian, English, Hebrew, Russian, covering both subject–verb agreement and noun–adjective agreement, where applicable (Gulordava et al., 2018). With transformer models, this line of research expanded to German (Zaczynska et al., 2020) as well as French, Hebrew and Russian (Mueller et al., 2020). Spanish has been the focus of broader agreement testing also beyond subject–verb (Pérez-Mayos et al., 2021), while targeted evaluations have also been developed for Galician and Portuguese (Garcia and Crespo-Otero, 2022; de Dios-Flores and Garcia, 2022). More recently, Basque auxiliary verb agreement with its complements and noun-class agreement in Swahili has been studied as well (Kryvosheieva and Levy, 2025). Overall, the studies have shown that the architecture of the tested model plays a role in its performance on agreement tasks. While LSTMs capture syntactic structure under certain conditions, their performance degrades in more complex configurations. Transformer-based models, on the other hand, exhibit a more robust agreement performance, especially in English. However, they are still sensitive to constructions involving agreement attractors or long-distance dependencies and perform worse on languages with a richer morphology, such, e.g., Basque, Hebrew, Russian, or Swahili.

Our work differs from previous works in two key ways. First, our test suites were manually curated by linguists to ensure that all examples are well-formed and semantically plausible, contrasting with other approaches that often rely on synthetically generated stimuli. Second, we provide a comparison across a wide range of monolingual and multilingual language models, which allows us to assess what elements have an impact on their agreement performance.

### 3 Agreement Test Suites

Following Hu et al. (2020); Pérez-Mayos et al. (2021) and others, we group the different tests into *test suites*. Each test suite focuses on a specific agreement rule and contains several *items*. Each item consists of a sentence sample adhering to the given rule and one or more samples that systematically vary from the first sample in the way they violate this rule. All test suites are based on the premise that a model should yield higher *sur-*

*prisal*<sup>2</sup> values for a target whose features fail to match those of its controller than one with correctly matching features.

To assess model performance under more realistic conditions, some test suites include an adversarial sample featuring grammatical constructions that increase the linear distance between the target and the controller. These constructions also incorporate what is commonly referred to in the literature as an *agreement attractor*, i.e., an element that shares its part of speech with the controller but that differs from it in the values for some or all of the agreement features involved in the relation. We paid special attention to select, when possible, agreement attractors that remain semantically plausible in relation to the target. The following examples from one of our Spanish test suites illustrate a regular test sentence and its adversarial counterpart (the controller appears in bold, the correct target in blue, the incorrect target in red, and the attractor is underlined):

- (1) Las **voluntarias** cayeron  
the **volunteer.F.PL** fell  
enfermas/\*enfermos  
ill.F.PL/\*ill.M.PL  
‘The volunteers fell ill.’
- (2) Las **voluntarias** [que ayudaron a los  
the **volunteer.F.PL** who helped to the  
refugiados] cayeron enfermas/\*enfermos  
refugee.M.PL fell ill.F.PL/\*ill.M.PL  
‘The volunteers who helped the refugees  
fell ill.’

In (2), the construction increasing linear distance between the controller and target is the relative clause in brackets, where the agreement attractor is underlined. Note that the attractor semantically fits both the main verb and the target, and that its features match those of the incorrect target, making the test more difficult for the models.

Regarding the **building process**, every example in each test suite was hand-crafted by a linguist fluent in the specific language<sup>3</sup>. Starting from a grammatical sentence, ungrammatical variants were created by altering morphological features involved in

<sup>2</sup>Following the terminology in Information Theory, we use the term *surprisal* to denote the negative log probability of a token (Samson, 1953), and in line with its use in psycholinguistics (Hale, 2001; Hu et al., 2020). Note, however, that to better capture contrasts between matching and mismatching controller wordforms, we rely on a different scoring metric; cf., Section 4.2.

<sup>3</sup>The time required to create comparable test suites varies with the designer’s linguistic expertise and creativity, but our examples can serve as a helpful starting point for future work.

agreement. To reduce frequency effects and bias, tests balanced all relevant feature values. Additionally, we deliberately included both stereotypical and non-stereotypical gender roles (e.g., female lawyers, male nurses) to ensure lexical diversity and further mitigate potential gender bias.

Below, we introduce the tested agreement phenomena, along with a list of all test suites created for them; for a more detailed description and additional examples, see Appendix B.

### 3.1 Italian, Portuguese, and Spanish

In Romance languages, controllers in agreement relations are nominal in nature, they are either nouns or pronouns; targets can be any words that can be inflected, such as finite verbs, participles, adjectives, and determiners; the features involved are gender (masculine and feminine), number (singular and plural) and person (first, second and third). As for the domain, we consider agreement within the noun phrase and agreement within the clause.

We created the test suites for Spanish, Italian and Portuguese based on these four variables, aiming to cover a representative range of their main agreement relations. Within the NP, we test nominal agreement (gender and number) between nouns and articles or possessives (both determiners) and between nouns and adjectives. Within the clause, we test nominal agreement between subject nouns and adjectives or predicate participles, as well as verbal agreement (person and number) between subject nouns or pronouns and the verb. In total, we experiment with ten different test suites for each of the three languages, some of which have already been introduced by Pérez-Mayos et al. (2021) for Spanish. Since these languages share many agreement-related properties, we present the test suites jointly, although separate instances of the test suites are used for each of the three languages, unless stated otherwise. Table 1 lists the different test suites and provides examples.

### 3.2 Russian

Similarly to Romance languages, in Slavic languages, agreement typically occurs between nominal controllers (nouns and pronouns) and targets that can be inflected (determiners, adjectives, participles and finite verbs), within a noun phrase and within a clause. Agreement features include number (singular and plural), gender (feminine, masculine, and neuter) and person (first, second, and third). The first difference to Romance languages

is, however, that Russian has a more pronounced case system. Noun phrases have six different case markings: nominative, accusative, genitive, dative, prepositional, and instrumental. Although it can be argued that case is the morphological manifestation of government and not an agreement feature (Corbett, 2006), we have included it in our test suites for two reasons. First, in nominal agreement, controller and target can take different forms with the same values of their agreement features, depending on their case, so it is impossible to avoid the matter completely. Second, if we assume the grammar to be dependency-based, the noun in a NP may take a specific case due to it being governed by the verb, but other elements in the NP (for instance, determiners or adjectives) take the same case because they are targets in their agreement relation with the noun. Moreover, unlike in Romance languages, the behaviour of Russian verbs regarding agreement depends on tense. In Russian, the verb БЫТЬ ('[to] be') is omitted in the present tense, leaving only the participle to carry past tense marking. As a result, Russian verbs show person and number agreement in present tense, but number and gender agreement in past tense.

Following the same approach as for the Romance languages, we designed for Russian ten different test suites, which aim to cover a range of fundamental agreement phenomena by manipulating the four key variables involved: controller, target, features, and domain. Within the NP, we test nominal agreement (gender and number) between nouns and articles or possessives (both determiners), and between nouns and adjectives. The test suites are grouped by syntactic structure, with all their test items sharing the same structure. Therefore, case, number and gender cannot be grouped together in one test for each controller–target combination. This means that for each combination, we can create six different, but partially repetitive, test suites. To avoid repetitions, we reduce the number of tests and explore noun–adjective agreement in nominative, accusative, and dative case, and noun–determiner agreement in genitive, prepositional, and instrumental case. Within the clause, we test nominal agreement between subject nouns and adjectives or participles in the predicate. Furthermore, we test verbal agreement (person and number) between subject nouns or pronouns and the verb. See Table 2 for the tests and examples.



Table 1: Test suites for Romance languages: agreement phenomena

Test suite	Languages	#Items**	Grammatical example***	Translation
Article–Noun	es, it, pt	32 × 4	(es) <u>El</u> .M.SG <u>gato</u> .(M).SG	‘The cat’
Possessive–Noun	it, pt	32 × 4	(it) <u>Il</u> .M.SG <u>mio</u> .M.SG <u>lavoro</u> .(M).SG	‘My job’
Adjective–Noun	es, it, pt	24 × 4	(es) <u>La tienda vende</u> <u>discos</u> .(M).PL <u>usados</u> .M.PL	‘The store sells second-hand vinyls’
Predicative Attribute*	es, it, pt	32 × 4	(pt) <u>O apartamento</u> .(M).SG <u>está</u> <u>vazio</u> .M.SG	‘The apartment is empty’
Predicative Complement*	es, it, pt	32 × 4	(it) <u>Le attrici</u> .(F).PL <u>ridevano</u> <u>spensierate</u> .F.PL	‘The actresses laughed nonchalantly’
			(es) <u>El conserje dejó</u> <u>la puerta</u> .(F).SG <u>abierta</u> .F.SG	‘The janitor left the door open’
Unaccusative Participle*	it	24 × 2	(it) <u>Il bambino</u> .M.SG <u>è andato</u> .M.SG <u>a scuola</u>	‘The boy went to school’
Passive Participle*	es, pt	24 × 2	(pt) <u>Os livros</u> .(M).PL <u>tem sido</u> <u>publicados</u> .M.PL	‘The books have been published’
Subject–Verb Basic	es, it, pt	24 × 4	(it) <u>Noi</u> .1.PL <u>cuciniamo</u> .1.PL	‘We cook’
Subject–Verb with Subject Relative Clause	es, it, pt	22 × 4	(pt) <u>O encanador</u> .SG [ <u>que ajudou os pedreiros</u> ] <u>trabalha</u> .3.SG <u>de sábado</u>	‘The plumber who helped the bricklayers works on Saturdays’
Subject–Verb with Object Relative Clause	es, it, pt	22 × 4	(es) <u>Los albañiles</u> .PL [ <u>la los que ayudó el fontanero</u> ] <u>trabajan</u> .3.PL <u>los sábados</u>	‘The bricklayers who the plumber helped work on Saturdays’

\* These test suites have an adversarial version (with approximately the same number of items, but only 2 examples per item).

\*\* Number of items × number of examples per item.

\*\*\* Target and controller are underlined, with their agreement features in small capital letters. Features in brackets are inherent to the word.

Table 2: Russian test suites: agreement phenomena

Test suite	#Items**	Grammatical example***	Lit. translation
Determiner–Noun Genitive	21 × 6	Машина <u>твоего</u> .GEN.M.SG <u>отца</u> .GEN.(M).SG	‘Car your father’s’
Determiner–Noun Instrumental	21 × 6	Я обедал со <u>своей</u> .INS.F.SG <u>сестрой</u> .INS.(F).SG	‘I had-lunch with my sister’
Determiner–Noun Prepositional	21 × 6	На <u>том</u> .PREP.M.SG <u>столе</u> .PREP.(M).SG	‘On that table’
Adjective–Noun Nominative	21 × 6	<u>красивая</u> .NOM.F.SG <u>женщина</u> .NOM.(F).SG <u>спит</u>	‘Beautiful woman sleeping’
Adjective–Noun Accusative	21 × 6	Президент примет <u>серьезное</u> .ACC.N.SG <u>решение</u> .ACC.(N).SG	‘President will-make serious decision’
Adjective–Noun Dative	21 × 6	Работодатель позвонит <u>лучшему</u> .DAT.M.SG <u>кандидату</u> .DAT.(M).SG	‘Employer will-call best candidate’
Predicative Attribute*	28 × 3	<u>Квартира</u> .(F).SG <u>кажется</u> <u>пустой</u> .F.SG	‘Apartment seems empty’
Predicative Complement*	31 × 3	<u>Учительница</u> .(F).SG <u>ушла</u> <u>сердитая</u> .F.SG	‘Teacher left angry’
Subject–Verb Present/Future	24 × 4	<u>Я</u> .1.SG <u>читаю</u> .1.SG <u>книгу</u>	‘I am-reading book’
Subject–Verb Past*	27 × 3	<u>Вода</u> .(F).SG <u>повредила</u> .F.SG <u>посевы</u>	‘Water damaged crops’

\* These test suites have an adversarial version (with approximately the same number of items, but only 2 examples per item).

\*\* Number of items × number of examples per item.

\*\*\* Target and controller are underlined with their agreement features in small capital letters. Features in brackets are inherent to the word.

## 4 Experimental Setup

We evaluated 25 different monolingual and multilingual models (cf. Table 6 in Appendix A), using the metric presented in Section 4.2 below and the *minicons* library (Misra, 2022), which allows for easy surprisal and probability computations. For bidirectional models, we applied the modified scoring technique proposed by Kauf and Ivanova (2023), which masks all tokens to the right of the target word (within the same word) to prevent over-estimation in multi-token words. For causal models with tokenizers that mark the beginning of words (and not the continuation), we applied the correction suggested by Pimentel and Meister (2024).

### 4.1 The Models

For monolingual models, we tested for **Spanish** BETO (Canete et al., 2020), the base version of RoBERTa from the MarIA family of models (Gutiérrez-Fandiño et al., 2022), the open-source GPT2-Spanish model from DeepESP<sup>4</sup>, and two lightweight models, alBETO and DistilBETO, introduced by Cañete et al. (2022). For **Italian**, we included an open-source Italian BERT model from the Bavarian State Library<sup>5</sup>, and its distilled version BERTino<sup>6</sup> trained by indigo.ai. We also tested

<sup>4</sup><https://huggingface.co/DeepESP/gpt2-spanish>

<sup>5</sup><https://huggingface.co/dbmdz/bert-base-italian-xxl-cased>

<sup>6</sup><https://huggingface.co/indigo-ai/BERTino>

an Italian RoBERTa model from the Osiria project<sup>7</sup>, UmBERTo, another RoBERTa-based model<sup>8</sup>, and the GPT-based model GePpeTto (Mattei et al., 2020). For **Portuguese**, we considered BERTimbau (Souza et al., 2020) and its distilled version<sup>9</sup>, as well as Tucano-160m, a LLaMA-based model (Corrêa et al., 2024), and GPorTuguese-2, a GPT-based model<sup>10</sup>. For **Russian**, we evaluated RuBERT (Kuratov and Arkipov, 2019), which is the large version of RuRoBERTa, the small version of RuGPT3 (based on GPT2) from the family of models pre-trained by Zmitrovich et al. (2024), and DistilBERTru, a model derived from mBERT via language reduction (Abdaoui et al., 2020).

For **multilingual models**, we tested the base versions of mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and the smallest version of XGLM (Lin et al., 2022), representing architectures based on BERT, RoBERTa and GPT3, respectively. Additionally, we included three larger and more recent decoder-only Transformer models: LLaMA-3.2-1B<sup>11</sup>, Bloom-7B (Workshop, 2023) and Salamandra-7B (Gonzalez-Agirre et al., 2025).

Information about the size of the selected models can be found in Table 3. Further technical details about the models are provided in Appendix A.

## 4.2 Evaluation Metric

Traditional targeted syntactic evaluations (see the references in Section 2) assess a model’s success in binary terms, i.e., whether it assigns a higher probability (or lower *surprisal*) to the correct word or sentence than to the incorrect one, without considering the magnitude of the difference.<sup>12</sup> This means that a model assigning nearly identical probabilities to both versions, with a slight preference for the correct one, would receive the same score as another model that strongly favors the correct choice. To address this limitation, we use a metric that accounts for the magnitude of the difference between the probabilities assigned by the model to each of the versions:

<sup>7</sup><https://huggingface.co/osiria/roberta-base-italian>

<sup>8</sup><https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1>

<sup>9</sup><https://huggingface.co/adalbertojunior/distilbert-portuguese-cased>

<sup>10</sup><https://huggingface.co/pierreguillou/gpt2-small-portuguese>

<sup>11</sup><https://huggingface.co/meta-llama/Llama-3.2-1B>

<sup>12</sup>Some works on targeted syntactic evaluation score single words while others score the full sentence.

Model	Number Params	Data Size	Model	Number Params	Data Size
<b>Spanish</b>			<b>Italian</b>		
BETO	110M	16GB	ItalianBERT	110M	81GB
RoBERTa-BNE	125M	570GB	UmBERTo	110M	70GB
GPT2-Spanish	125M	11.5GB	ItalianRoBERTa <sup>1</sup>	125M	30GB+
DistilBETO	67M	16GB	GePpeTto	117M	13GB
alBETO	12M	16GB	BERTino	68M	12GB
<b>Portuguese</b>			<b>Russian</b>		
BERTimbau	109	17.5GB	RuBERT <sup>2</sup>	178M	150GB+
Tucano-160M	160M	589GB	RuRoBERTa	355M	250GB
GPorTuguese-2 <sup>1</sup>	124M	1GB+	RuGPT3	125M	450GB
DistilBERTimbau <sup>3</sup>	66M	??	DistilBERTru <sup>4</sup>	55M	–
<b>Multilingual models</b>					
DistilBERT	134M	~75GB	LLaMA3.2	1.23B	~40TB
mBERT	178M	~75GB	Bloom	7.07B	1.5TB
XLM-R	270M	2.5TB	Salamandra	7.77B	~20TB
XGLM	564M	9TB			

<sup>1</sup> ItalianRoBERTa and GPorTuguese-2 were not trained from scratch.

<sup>2</sup> RuBERT was adapted from mBERT by training a new tokenizer and replacing the embedding layer.

<sup>3</sup> No information was found about DistilBERTimbau’s training data.

<sup>4</sup> DistilBERTru was created from mBERT via language reduction, there was no ulterior training.

Table 3: Number of parameters and dataset size of selected models.

$$Score(item) = \frac{1}{n} \sum_{x_i \in I} \frac{p(x_t|c)}{p(x_t|c) + p(x_i|c)}$$

where  $p$  represents the model’s probability distribution,  $x_t$  is the target<sup>13</sup> word with matching morphological features,  $x_i$  is an incorrect word with all or some mismatching features,  $I$  is a set of  $n$  possible incorrect words for this item<sup>14</sup>, and  $c$  represents the context (left context for causal models, and both left and right context for bidirectional ones).

This metric provides an estimation of the model’s probability of choosing the correct word over an incorrect one, and then averages this probability across a set of incorrect alternatives. A value over 0.5 means the model assigned (on average) higher probability to the correct word than the incorrect ones, the higher the value, the bigger the difference.

## 5 Results and Discussion

Table 4 presents the average evaluation scores achieved by the individual models on our agreement test suites. Overall, it can be stated that the models have a reasonable agreement competence, although some significant differences can be ob-

<sup>13</sup>The term *target* is used here to refer to the expected or correct word, not to be confused with the grammatical concept of *target* in an agreement relation, as introduced in Section 1

<sup>14</sup>The number of possible incorrect words is one less than the number of examples per item, which is provided in Tables 1 and 2

Model	Agreement Score	Model	Agreement Score
<b>Spanish</b>		<b>Italian</b>	
BETO	0.9127	ItalianBERT	0.9009
RoBERTa-BNE	0.9167	UmBERTo	0.7581
GPT2-Spanish	<b>0.9223</b>	ItalianRoBERTa	0.8354
DistilBETO	0.7703	GePpeTto	0.8818
alBETO	0.7930	BERTino	<b>0.9112</b>
<b>Portuguese</b>		<b>Russian</b>	
BERTimbau	<b>0.9451</b>	RuBERT	0.8941
Tucano-160M	0.8967	RuRoBERTa	0.9078
GPorTuguese-2	0.8117	RuGPT3	<b>0.9159</b>
DistilBERTimbau	0.5126	DistilBERTru	0.7220
<b>Multilingual models</b>			
DistilMBERT	0.7257	LLaMA3.2-1B	0.8568
mBERT	0.8036	Bloom-7B	0.8585
XLM-R	0.8402	Salamandra-7B	<b>0.9331</b>
XGLM	0.8664		

Table 4: Models’ average agreement score.

served. As expected, monolingual models generally outperform multilingual ones. However, the multilingual Salamandra achieves an average score across all four languages that is comparable to the score of the best monolingual models.

In what follows, we take a closer look at the performance of monolingual and multilingual models, emphasizing the most significant findings.

### 5.1 Monolingual Models

The **Spanish** models BETO, RoBERTa-BNE, and GPT2-Spanish achieve very similar scores, with GPT2-Spanish performing the best, reaching a 0.92 score. The two lightweight models perform worse, but still achieve reasonable scores over 0.77. For **Italian**, the distilled model BERTino outperforms all the rest at 0.91. GePpeTto and ItalianBERT perform similarly, with scores above 0.88. Italian-RoBERTa and UmBERTo score slightly lower, at 0.84 and 0.76, respectively. **Portuguese** models achieve the highest overall scores. BERTimbau is the strongest performer, reaching a 0.95 score, followed by Tucano-160m (0.90) and GPorTuguese-2 (0.81). The latter is particularly noteworthy, as it was fine-tuned from English GPT2-small using only 1GB of Portuguese data. The distilled version of BERTimbau performs significantly worse, with a score of 0.51. Finally, among **Russian** models, RuGPT3 leads with 0.92, closely followed by RuRoBERTa (0.91) and RuBERT (0.89). Once again, the lightweight model scores slightly lower at 0.72.

The performance of the models allows for some interesting conclusions that challenge common as-

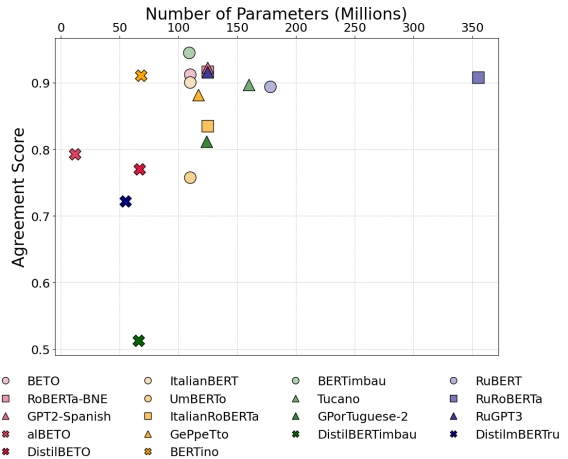


Figure 1: Average agreement score vs. model size (for monolingual models).

sumptions concerning model size, the size and quality of the training data, and model architecture. Thus, we find that model size (in terms of parameter count) has a weaker effect on agreement performance than one might expect (see Figure 1). For example, in Italian, BERTino, despite being half the size of GePpeTto and ItalianBERT, achieves a higher score. A similar pattern emerges in Spanish, where the lightweight alBETO model – despite being only 10% the size of GPT2-Spanish – still achieves 86% of its performance. In Russian, RuGPT3 outperforms RuRoBERTa despite having just over a third of its parameters; similarly, in Portuguese, BERTimbau surpasses Tucano-160m while using a third fewer parameters. **Training data size** also shows a limited impact on performance. In Spanish, GPT2-Spanish, trained on only 2% the amount of data used for RoBERTa-BNE, still outperforms it. Likewise, BERTimbau, trained on just 3% of the data used for Tucano-160m, achieves a considerably higher score in Portuguese. Italian follows the same pattern: GePpeTto, despite being trained on just 16% of the data used for ItalianBERT, performs at a comparable level. While it is not feasible to systematically assess within the scope of this study the linguistic **quality of the training data** for all models, we hypothesize that exposure to well-written, grammatically correct texts (e.g., Wikipedia, books, and news articles) likely contributes to better performance on linguistically demanding tasks such as ours. Testing this hypothesis directly would require a dedicated analysis beyond the current work.

**Model architecture** does not reveal a clear per-

formance trend either. Both encoder–decoder architectures like BERT and RoBERTa and decoder-only architectures like the GPT variants can achieve high agreement scores. In Spanish, BERT, RoBERTa-BNE, and GPT2-Spanish perform similarly, as do ItalianBERT, BERTino, and GePpeTto in Italian, and RuBERT, RuRoBERTa, and RuGPT3 in Russian. However, in Portuguese, the highest performance is achieved by BERTimbau, a BERT-based model. Neither the **tokenizer strategy** nor the **vocabulary size** appears to have a strong impact on agreement scores: the best models for Spanish and Russian use a BPE tokenizer with a 50K vocabulary, while those for Italian and Portuguese perform best with WordPiece and a 30K vocabulary. The top-performing multilingual model, meanwhile, uses SentencePiece.

## 5.2 Multilingual Models

Table 5 presents additional information on the performance of the multilingual models, including average scores for each language and the proportion of training data allocated to each language in each of the models. Overall, multilingual models perform well, but somewhat weaker than monolingual models, with the exception of Salamandra and Bloom for Spanish. Salamandra leads the rankings at an average score of 0.93, followed by XGLM at 0.87, and both Bloom and LLaMA-3.2 at 0.86. Looking at the results by language, Salamandra achieves the highest scores for Italian (0.92), Portuguese (0.94), and Russian (0.92), and Bloom stands out as the best-performing multilingual model for Spanish (0.96). Interestingly, Salamandra outperforms the best monolingual models for Spanish, Italian and Russian while remaining highly competitive for Portuguese. Similarly, Bloom surpasses the top Spanish monolingual model and nearly matches the best Portuguese one. Large multilingual models clearly benefit from **transfer learning** across languages, as evidenced by Bloom’s results. Despite not being trained on any Italian or Russian data, it still performs reasonably well on these languages, most likely due to typological proximity.

Unlike for monolingual models, for multilingual models, **model size** appears to have a stronger impact on agreement competence, with larger models generally achieving better results, although it should be noted that we are now comparing much bigger sizes (see Figure 2). However, there are exceptions: XGLM, despite being half the size

Model	Our Score			
	Spanish	Italian	Portuguese	Russian
DistilBERT	~4.7% 0.691	~4.7% 0.741	~2.3% 0.744	~4.7% 0.727
mBERT	~.7% 0.795	~4.7% 0.803	~2.3% 0.806	~4.7% 0.811
XLM-R	2.1% 0.834	1.2% 0.832	2.0% 0.839	11.1% 0.856
XGLM	4.3% 0.893	2.0% 0.809	1.8% 0.878	12.0% 0.885
LLaMA	?? 0.894	?? 0.827	?? 0.844	?? 0.862
Bloom	11.1% <b>0.960</b>	0% 0.751	5.2% 0.938	0% 0.785
Salamandra	16.1% 0.953	2.1% <b>0.919</b>	2.2% <b>0.944</b>	5.6% <b>0.917</b>

Table 5: Multilingual models’ score by language

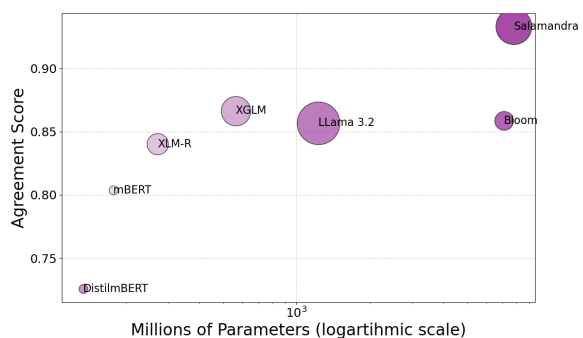


Figure 2: Average agreement score vs. model size (for multilingual models). Bubble size reflects training data size.

of LLaMA-3.2, outperforms it; and XLM-R with a fourth of LLaMA-3.2’s size comes quite close. Something similar happens with **training data size**: models trained on more data tend to perform better. Yet again, there is an exception: although LLaMA-3.2 was trained on twice the amount of data as Salamandra (the second-largest model in terms of training data), it is still outperformed by it.

As far as **model architecture** is concerned, no definitive conclusions can be drawn since all larger models in the study are decoder-only Transformers, whereas the smaller ones follow an encoder–decoder architecture.

## 5.3 Results by Test Suite

Figure 3 presents the agreement scores across all individual test suites for all models and languages. As anticipated from the averages, the overall performance remains reasonably high. However, a few noteworthy observations stand out.

Although the **Article—Noun** test suite is relatively simple and should thus not be particularly



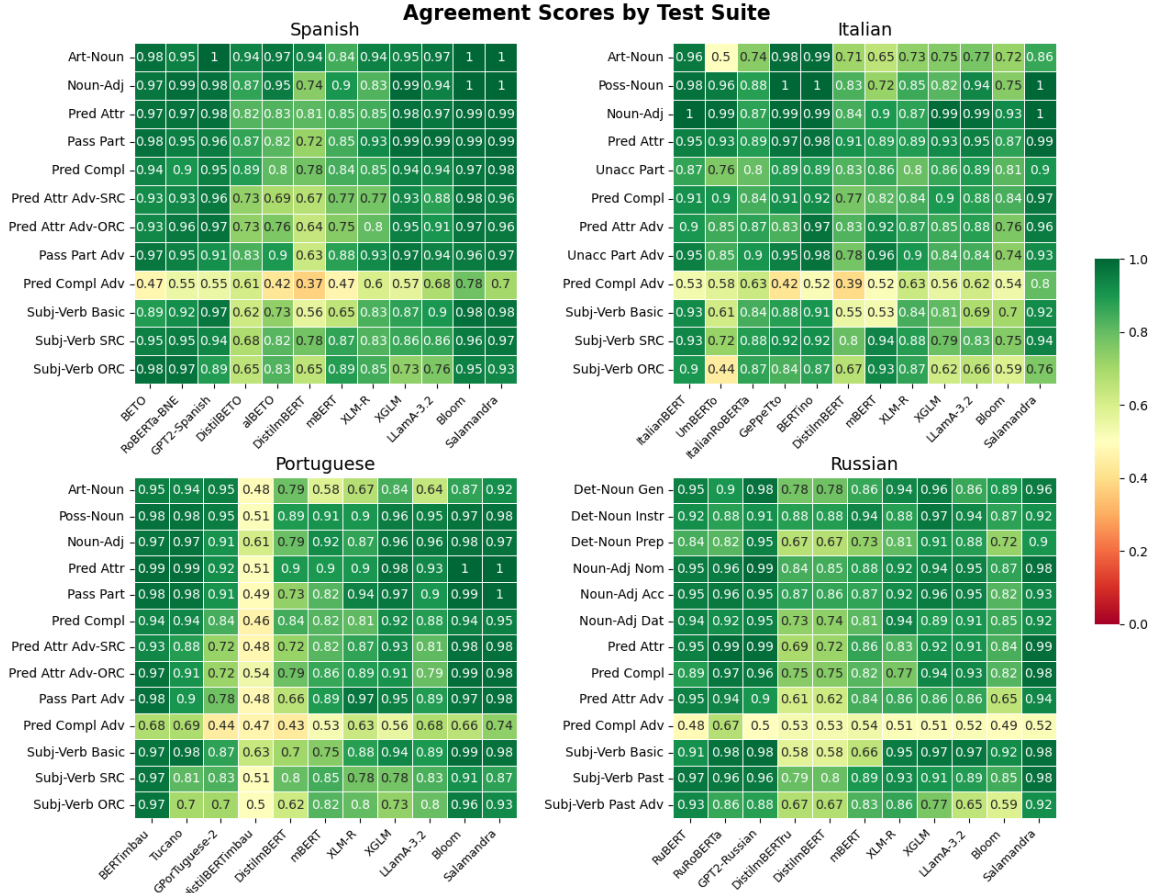


Figure 3: Agreement scores by test suite for all models and languages.

demanding, multilingual models seem to struggle with it for Italian and Portuguese. This may be due to the brevity of articles in these languages, often made up of just one or two letters (e.g., *o* and *a* as singular definite articles in Portuguese for masculine and feminine nouns). The models might misinterpret them as the start of a word (perhaps in another language), leading to errors. Similarly, while the **Subject–Verb Basic** test suite is also expected to be straightforward, it proved problematic for some models. Notably, mBERT exhibited a consistent difficulty across all four languages.

Surprisingly, most **adversarial versions** of the test suites were not radically more demanding than their standard counterparts. The only exception was the **Predicative Complement** test suite, which was already somewhat more complex than the other nominal agreement tests within the clause. In its adversarial form, it posed a substantial challenge, even for the strongest monolingual and multilingual models.

Both Subject–Verb with Relative Clause test suites should be considered inherently adversar-

ial, despite lacking a standard equivalent. Adapted from English (Marvin and Linzen, 2018), they focus exclusively on number agreement with third-person subjects. These constructions proved difficult for many models, particularly multilingual ones, though even some monolingual models struggled, as can be observed in the results for Portuguese.

Looking at the lighter columns in Figure 3, which generally correspond to distilled models, we observe that **distillation** tends to negatively impact models’ agreement performance, with the Italian model BERTino being a notable exception. In particular, the Spanish lightweight model alBETO, despite having only a fifth of DistilBETO’s parameters, performs more robustly.

Although one might expect models to perform worse in **Russian** due to its more complex case system, the opposite could also be argued, as case markings provide additional grammatical cues. Moreover, relative clauses, used in adversarial test suites to increase the linear distance between the controller and the target, are typically enclosed by

commas in Russian, offering further structural hints that may increase the performance of the model.

Finally, **Salamandra** once again demonstrates exceptional robustness, facing real difficulty only in the Predicative Complement adversarial test suite for all languages, especially Russian, where it scored 0.52. It also encountered mild challenges in the Subject–Verb with Object Relative Clause for Italian: the only other instance where its score fell below 0.8.

#### 5.4 Adversarial Test Suites

As our results show, the adversarial test suites were generally more difficult than their standard counterparts, though not as consistently as expected; some models even performed better on the adversarial versions.

Adversarial suites were carefully constructed to include semantically plausible agreement attractors, with incorrect targets deliberately chosen to agree with them rather than with the correct controller. Despite this, several models reliably assigned higher probabilities to the grammatically correct target, prompting for further investigation.

In the context of adversarial suites, we also developed alternative versions of the Subject–Verb with Relative Clauses test suites in Spanish. In the **subject** relative clause variant, we removed the preposition preceding the attractor<sup>15</sup>, hypothesizing that its absence (potentially serving as a syntactic cue) would increase the difficulty. In contrast, for the **object** relative clause version, we repositioned the subject of the relative clause before the subordinate verb<sup>16</sup>, expecting this to simplify the task by distancing the attractor from the main verb. Interestingly, the results contradicted our expectations: most models performed better on the revised subject relative clause suite and worse on the modified object relative clause suite. The improvement in the former may be due to the attractor’s reduced semantic plausibility (being inanimate), while the decline in the latter might stem from the unusual surface structure, specifically, the sequence of two adjacent verbs confusing the model’s internal rep-

<sup>15</sup>For example, *La bióloga que colabora con los veterinarios tiene mucha experiencia* (‘The biologist(f) who collaborates with the veterinarians has a lot of experience’) became *La bióloga que comprueba los equipos tiene mucha experiencia* (‘The biologist who tests the equipment has a lot of experience’).

<sup>16</sup>We replaced *Los albañiles a los que ayudó el fontanero* with *Los albañiles a los que el fontanero ayudó* (‘The bricklayers who the plumber helped’).

resentation of syntactic relations.

The key takeaway from these findings is that language models do not rely heavily on surface adjacency between controller and target. Instead, they seem to leverage other cues, including semantic compatibility. At the same time, the models’ tendency to fail in the presence of non-canonical verb sequences suggests that their grasp of syntactic structure remains limited. Rather than encoding abstract syntactic relations robustly, they may depend more on frequent patterns and shallow heuristics.

## 6 Conclusions

Our experiments with 25 different Large Language Models, applied to a number of different test suites for Italian, Portuguese and Spanish on the one side and Russian on the other side, have shown that, in general, the models show a reasonable competence in agreement across languages, although monolingual models tend to perform better than multilingual ones. No significant difference has been noted between the Romance languages and Russian. However, all models still struggle to a certain extent with more complex syntactic constructions with attractors, for instance, in relative clauses. Interestingly, and contrary to common assumptions, model size, architecture, and training data volume had only a limited impact on agreement performance.

The test suites can be freely accessed and downloaded for research purposes<sup>17</sup>. Our approach to the creation of test suites can be extended to other languages, provided it remains focused on agreement phenomena. Creating test suites for a new language requires identifying potential *targets* and *controllers*, determining the relevant agreement *features*, and defining the *domain* where agreement occurs within the language’s grammatical structure. The initial goal should be to cover core agreement phenomena by combining these four factors, which can then be expanded by incorporating more challenging or marginal cases, as well as semantic agreement and agreement resolution. This is precisely the direction we aim to take in the next phase of our work, broadening the range of languages covered and increasing the level of difficulty in our tests.

<sup>17</sup><https://huggingface.co/datasets/albalbalba/SyntacticAgreement>

## Limitations

Despite a considerably broader coverage of agreement phenomena and different state-of-the-art neural language models than in previous work, our study has some obvious limitations. First, our study covers only four languages, three of which are closely related Romance languages. A broader typological coverage, particularly including languages from non-Indo-European families, languages with a rich inflectional morphology, or those with unique agreement systems, would provide a more comprehensive picture. Unfortunately, the development of high-quality test suites requires both linguistic expertise in the target languages and access to language models trained on them, both of which are often lacking for under-resourced or less commonly studied languages.

Second, the test suites used in this study primarily target core and relatively canonical agreement phenomena. While this allows for consistent and controlled evaluation, it may also underestimate the challenges that arise in more marginal or exceptional agreement cases—e.g., agreement across clause boundaries, or cases involving semantic factors. Future work should aim to incorporate such phenomena, both to test deeper syntactic and semantic understanding and to better represent the complexity found in natural language.

Third, while our test suites were carefully constructed by expert linguists to ensure that grammatical and ungrammatical examples are well-formed and contrastive, we did not complement them with human acceptability judgments. Given the controlled design and linguistic motivation behind each example, we expect them to be generally reliable. However, incorporating human judgments in future work could provide additional insights into whether model predictions align with speaker intuitions, and help clarify to what extent observed model behavior reflects genuine linguistic competence.

Fourth, we observed some unexpected patterns, most notably, the strong performance of the lightweight Italian model BERTino and the generally lower scores of Italian models overall. We consider it unlikely that these differences arise from language-specific properties, given the close typological similarity among the Romance languages examined. Instead, the differences are more plausibly due to the variation in model pretraining or training data, which are not well documented for

the Italian models. This highlights the need for more transparency in model development and warrants further investigation beyond the scope of the present study.

Finally, the range of language models we evaluated was constrained by the available computational resources. Many of the most recent state-of-the-art models are prohibitively large for independent researchers to use, even in a zero-shot evaluation setting. This limits the ability to fully explore the impact of scale and architecture. As model sizes continue to increase, the need for more equitable access to these technologies will become increasingly pressing, not just for training, but also for systematic evaluation.

## References

- Amine Abdaoui, Camille Pradel, and Grégoire Sigel. 2020. [Load what you need: Smaller versions of multilingual BERT](#). In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 119–123, Online. Association for Computational Linguistics.
- José Canete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. *PML4DC at ICLR*, 2020.
- José Cañete, Sebastian Donoso, Felipe Bravo-Marquez, Andrés Carvallo, and Vladimir Araujo. 2022. [ALBETO and DistilBETO: Lightweight Spanish language models](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4291–4298, Marseille, France. European Language Resources Association.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Greville G. Corbett. 2006. *Agreement*. Cambridge University Press, Cambridge, UK.
- Nicholas Kluge Corrêa, Aniket Sen, Sophia Falk, and Shiza Fatimah. 2024. [Tucano: Advancing neural text generation for portuguese](#). *Preprint*, arXiv:2411.07854.
- Iria de Dios-Flores and Marcos Garcia. 2022. [A computational psycholinguistic evaluation of the syntactic abilities of galician bert models at the interface of dependency resolution and training time](#). *Procesamiento del lenguaje natural*, 69:15–26.



- 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.
- Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. [RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency](#). *arXiv preprint arXiv:1809.01329*.
- Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. [Neural language models as psycholinguistic subjects: Representations of syntactic state](#). 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 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.
- Marcos Garcia and Alfredo Crespo-Otero. 2022. A targeted assessment of the syntactic abilities of transformer models for galician-portuguese. In *Computational Processing of the Portuguese Language*, pages 46–56, Cham. Springer International Publishing.
- Yoav Goldberg. 2019. Assessing BERT’s syntactic abilities. *arXiv preprint arXiv:1901.05287*.
- Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lancunza, Jorge Palomar, Júlia Falcão, Lucía Tormo, Luis Vasquez-Reina, Montserrat Marimon, Oriol Pareras, Valle Ruiz-Fernández, and Marta Villegas. 2025. [Salamandra technical report](#). *Preprint*, arXiv:2502.08489.
- Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. [Colorless green recurrent networks dream hierarchically](#). 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 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.
- Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodríguez Penagos, Aitor Gonzalez-Agirre, and Marta Villegas Montserrat. 2022. [Maria: Spanish language models](#). *Procesamiento del Lenguaje Natural*, 68:39–60.
- John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*.
- Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. [A systematic assessment of syntactic generalization in neural language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.
- Carina Kauf and Anna Ivanova. 2023. [A better way to do masked language model scoring](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada. Association for Computational Linguistics.
- Daria Kryvosheieva and Roger Levy. 2025. [Controlled evaluation of syntactic knowledge in multilingual language models](#). In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 402–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yuri Kuratov and Mikhail Arhipov. 2019. [Adaptation of deep bidirectional multilingual transformers for russian language](#). *Preprint*, arXiv:1905.07213.
- Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Nanman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O’Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. [Few-shot learning with multilingual generative language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. [Assessing the ability of LSTMs to learn syntax-sensitive dependencies](#). *Transactions of the Association for Computational Linguistics*, 4:521–535.
- Rebecca Marvin and Tal Linzen. 2018. [Targeted syntactic evaluation of language models](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.
- Lorenzo De Mattei, Michele Cafagna, Felice Dell’Orletta, Malvina Nissim, and Marco Guerini. 2020. [Geppetto carves italian into a language model](#). In *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769. CEUR-WS.org. Italian Conference on Computational Linguistics 2020, CLiC-it 2020 ; Conference date: 01-03-2021 Through 03-03-2021.



- Igor Mel'čuk. 1993. Agreement, governance and congruence. *Linguisticae investigationes*, 17(2):307–373.
- Igor Mel'čuk. 2009. Dependency in natural language. In *Dependency in Linguistic Description*, pages 1–110. John Benjamins Publishing Company.
- Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.
- Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. [Cross-linguistic syntactic evaluation of word prediction models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.
- Vivi Nastase, Giuseppe Samo, Chunyang Jiang, and Paola Merlo. 2024. [Exploring syntactic information in sentence embeddings through multilingual subject-verb agreement](#). In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 631–643, Pisa, Italy. CEUR Workshop Proceedings.
- Laura Pérez-Mayos, Alba Táboas García, Simon Mille, and Leo Wanner. 2021. [Assessing the syntactic capabilities of transformer-based multilingual language models](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3799–3812, Online. Association for Computational Linguistics.
- Tiago Pimentel and Clara Meister. 2024. [How to compute the probability of a word](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375, Miami, Florida, USA. Association for Computational Linguistics.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. 2018. [Can LSTM learn to capture agreement? the case of Basque](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107, Brussels, Belgium. Association for Computational Linguistics.
- Edward W. Samson. 1953. Fundamental natural concepts of information theory. *ETC: A Review of General Semantics*, 10(4):283–297.
- Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. [Bertimbau: Pretrained bert models for brazilian portuguese](#). In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, page 403–417, Berlin, Heidelberg. Springer-Verlag.
- Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. [What do RNN language models learn about filler–gap dependencies?](#) In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.
- Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. 2019. [Structural supervision improves learning of non-local grammatical dependencies](#). 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 3302–3312, Minneapolis, Minnesota. Association for Computational Linguistics.
- BigScience Workshop. 2023. [Bloom: A 176b-parameter open-access multilingual language model](#). *Preprint*, arXiv:2211.05100.
- Karolina Zaczynska, Nils Feldhus, Robert Schwarzenberg, Aleksandra Gabryszak, and Sebastian Möller. 2020. [Evaluating german transformer language models with syntactic agreement tests](#). *CoRR*, abs/2007.03765.
- Dmitry Zmitrovich, Aleksandr Abramov, Andrey Kalmykov, Vitaly Kadulin, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergei S. Markov, Vladislav Mikhailov, and Alena Fenogenova. 2024. [A family of pretrained transformer language models for Russian](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 507–524, Torino, Italia. ELRA and ICCL.

## A Models' Size and Training Information

Table 6 provides detailed information about the language models selected for evaluation: model size, architecture, training dataset(s) and its size, tokenizer strategy and vocabulary size, as well as the average score obtained on our test suites.

## B Description of test suites

This appendix is a description of all test suites created to assess the ability of LLMs to identify sentence samples that violate the rules of agreement in three Romance languages (Italian, Portuguese and Spanish) and one Slavic language (Russian). Each test suite contains sentence samples of a specific agreement rule and systematic variations of these samples violating that given rule. All of them are based on the premise that an element in an agreement relation whose features match those of the other element in the relation should yield higher

Table 6: Model information and average agreement score.

Model	Lgg	Architecture	Number Params	Training Data	Dataset Size	Tokenizer	Vocabulary Size	Our Score <sup>1</sup>
BETO	es	BERT	110M	Wikipedia, OPUS	16GB	WordPiece	31K	0.9127
RoBERTa-BNE	es	RoBERTa	125M	BNE BookCrawl	570GB	BPE	50K	0.9167
GPT2-Spanish	es	GPT2	125M	Wikipedia, Books	11.5GB	BPE	50K	<b>0.9223</b>
DistilBETO	es	DistilBERT	67M	Wikipedia, OPUS	16GB	SentencePiece	31K	0.7703
alBETO	es	alBERT	12M	Wikipedia, OPUS	16GB	SentencePiece	31K	0.7930
ItalianBERT	it	BERT	110M	Wikipedia OPUS, OSCAR	81GB	WordPiece	31K	0.9009
UmBERTo	it	RoBERTa	110M	OSCAR	70GB	SentencePiece	32K	0.7581
ItalianRoBERTa <sup>2</sup>	it	RoBERTa	125M	CommonCrawl	30GB+	BPE	50K	0.8354
GePpeTto	it	GPT2	117M	Wikipedia, ItWaC	13GB	BPE	30K	0.8818
BERTino	it	DistilBERT	68M	Paisa, ItWaC	12 GB	WordPiece	31K	<b>0.9112</b>
BERTimbau	pt	BERT	109M	brWaC	17.5GB	WordPiece	30K	<b>0.9451</b>
Tucano-160M	pt	LLaMA	160M	GigaVerbo	589GB	BPE	32K	0.8967
GPorTuguese-2 <sup>3</sup>	pt	GPT2	124M	Wikipedia	1GB+	BPE	50K	0.8117
DistilBERTimbau	pt	DistilBERT	66M	??	??	WordPiece	30K	0.5126
RuBERT <sup>4</sup>	ru	BERT	178M	Wikipedia, News	150GB+	BPE	120K	0.8941
RuRoBERTa	ru	RoBERTa	355M	Wikipedia News(Corus), Books	250GB	BPE	50K	0.9078
RuGPT3	ru	GPT2	125M	Wikipedia(ru+en) News, Books, C4	450GB	BPE	50K	<b>0.9159</b>
DistilmBERTru <sup>5</sup>	ru	BERT	55M	-	-	WordPiece	14K	0.7220
DistilmBERT	multi	DistilBERT	134M	Wikipedia	~50-100GB <sup>6</sup>	WordPiece	120K	0.7257
mBERT	multi	BERT	178M	Wikipedia	~50-100GB <sup>6</sup>	WordPiece	120K	0.8036
XLM-R	multi	RoBERTa	270M	CommonCrawl	2.5TB	Unigram	150K	0.8402
XGLM	multi	GPT3	564M	CommonCrawl Books, Wikipedia	9TB	Unigram	250K	0.8664
LLaMA-3.2	multi	LLaMA	1.23B	??	~40TB (9T tokens) <sup>7</sup>	Unigram	128K	0.8568
Bloom	multi	Bloom	7.07B	ROOTS Corpus	1.5TB	BPE	250K	0.8585
Salamandra	multi	Salmandra	7.77B	Colossal OSCAR FineWeb-Edu	~20TB (13T tokens) <sup>7</sup>	SentencePiece	256K	<b>0.9331</b>

<sup>1</sup> The score for each monolingual model is the average calculated over all the test suites for each specific language. The score for multilingual models is the average score for the four languages evaluated. The best results for each language appears in bold.

<sup>2</sup> ItalianRoBERTa was not trained from scratch, but initialized with XLM-R weights.

<sup>3</sup> GPorTuguese-2 was not trained from scratch, but fine-tuned from the English GPT-2-small (Radford et al., 2019).

<sup>4</sup> RuBERT was adapted from mBERT by training a new tokenizer and replacing the embedding layer.

<sup>5</sup> DistilmBERTru is not really a distilled model, rather, it has been obtained by language reduction from mBERT as proposed by Abdaoui et al. (2020); there was no ulterior training performed.

<sup>6</sup> Dataset size was estimated considering the size of Wikipedia dumps at the time the model was released.

<sup>7</sup> Information about the dataset size was in number of tokens; size in TB was estimated by assuming an average token size of 5 bytes.

“?”: No information was found about training data and/or its size.

probability values than one with mismatching features. In total, the test suites contain over 5,000 test samples.

In all test suites, the element that is systematically modified to violate the rule is the *target* of the agreement relation. The probabilities used to calculate the metric described in Section 4.2 are the ones assigned by the model to the correct and incorrect versions of this element. However, to be able to apply all our test suites to both bidirectional and causal LLMs, there are a couple of exceptions in which the element that varies is the target, but the probability is measured for the *controller*. This happens with the determiner–noun agreement relation, in which the target is to the left of the controller. Note that although the controller is the same, the probabilities are not, because the target in its left context has been modified.

### B.1 Italian, Portuguese, and Spanish

Since Italian, Portuguese and Spanish share many agreement-related properties, we present most of the test suites for them jointly, with an example in one of the three languages. Note, however, that in these cases, each test also has a version in all three of them, unless explicitly stated otherwise.

In total, we experiment with ten different test suites for these languages. We explore nominal agreement–gender and number–within the noun phrase (three suites) and within the clause (four suites), and verbal agreement–person and number–within the clause (three suites).

#### • Article–Noun Agreement

To test article–noun agreement, the four possible forms of the definite article (‘sg.’ vs. ‘pl.’ × ‘fem.’ vs. ‘masc.’) are paired with different nouns to capture all four forms. Cf. the following example in Spanish, with the masculine noun *gato* ‘cat’:

- (3) El/\*La/\*Los/\*Las gato  
the.M.SG/\*F.SG/\*M.PL/\*F.PL cat  
‘The cat’

#### • Possessive–Noun Agreement

Unlike in Spanish, in Italian and Portuguese, the possessive determiner has the four possible forms ‘sg.’ vs. ‘pl.’ × ‘fem.’ vs. ‘masc.’, as the definite article. This test pairs them with nouns that reflect these forms; cf. an example in Italian:

- (4) Il mio lavoro  
the.M.SG my.M.SG job

- (5) \*La mia / \*[I miei] /  
\*[the.F.SG my.F.SG] / \*[the.M.PL my.M.PL] /  
\*Le mie lavoro  
\*[the.F.PL my.F.PL] job  
‘My job’

• **Adjective–Noun Agreement** This test pairs a noun with the ‘sg.’ vs. ‘pl.’ × ‘fem.’ vs. ‘masc.’ forms of an adjective that modifies it. To avoid providing extra information to the model, the test uses constructions without a determiner; cf., an example in Spanish:

- (6) La tienda vende discos  
the store sells discs  
usados/\*usado/\*usadas/\*usada  
used.M.PL/M.SG/F.PL/F.SG  
‘The store sells second-hand discs’

#### • Predicative Attribute Agreement

In Romance languages, the predicative attribute in copulative constructions must agree with the grammatical subject in gender and number. Consider an example in Portuguese:

- (7) O apartamento está  
the.M.SG apartment is  
vazio/\*vazios/\*vazia/\*vazias  
empty.M.SG/\*M.PL/\*F.SG/\*F.PL  
‘The apartment is empty’

For Spanish and Portuguese, this suite has two adversarial versions one with object and one with subject relative clauses. For Italian, there is only one adversarial version in which the intervening material can be a relative clause or a prepositional phrase. Since the role of the intervening material is to increase the linear distance between the co-varying words and to include an agreement attractor, we decided to condense the two versions into one, as long as it maintains these two important factors. Here is an example in Italian:

- (8) L’ appartamento che guarda  
the.M.SG apartment that look.3.SG  
verso la spiaggia è  
towards the.F.SG beach is  
vuoto/\*vuoti/\*vuota/\*vuote  
empty.M.SG/\*M.PL/\*F.SG/\*F.PL  
‘The apartment facing the beach is empty’

#### • Predicative Complement Agreement

In Italian, Portuguese, and Spanish, an attribute that functions as a predicative complement to the grammatical subject or the object must agree with it in gender and number; cf. an example in Spanish for a complement to the object:

- (9) El tenista dejó la raqueta  
 the tennis-player left the.F.SG racket  
 destrizada/\*destrizadas/\*destrizado/\*destrizados.  
 destroyed.F.SG/\*F.PL/\*M.SG/\*M.PL  
 ‘The tennis player left his racket destroyed.’

This suite has an adversarial version with intervening material in the form of a relative clause or a prepositional phrase; cf. this example for a complement to the subject:

- (10) Las voluntarias que ayudaron a  
 the.F.PL volunteer.F.PL who helped to  
 los refugiados cayeron  
 the.M.PL refugee.M.PL fell  
 enfermas/\*enferma/\*enfermos/\*enfermo.  
 ill.F.PL/\*F.SG/\*M.PL/\*M.SG  
 ‘The volunteers who helped the refugees fell ill.’

#### • Participle Agreement

In contrast to Portuguese and Spanish, in Italian, unaccusative verbs in past tense are conjugated with the auxiliary verb *essere* (‘[to] be’) and their past participle form. The past participle must agree in gender and number with the grammatical subject; cf.:

- (11) Il bambino è andato/\*andata a scuola.  
 the child is gone.M.SG/\*F.SG to school  
 ‘The child has gone to school.’

This suite has an adversarial version as well, with relative clauses or prepositional phrases as intervening material.

- (12) Il bambino che ha litigato con sua  
 the child who has fought with his.F.SG  
 sorella è andato/\*andata a scuola.  
 sister is gone.M.SG/\*F.SG to school  
 ‘The child who had a fight with his sister has gone to school.’

#### • Passive Participle Agreement

In passive constructions of Portuguese and Spanish, the past participle must agree in gender and number with the grammatical subject; cf. a Portuguese example:

- (13) Os encontros serão  
 the.M.PL matches will.be  
 transmitidos/\*transmitidas ao vivo  
 broadcast.M.SG/\*F.SG to.the live  
 ‘The matches will be broadcast live.’

This suite also has an adversarial version for both languages.

- (14) Os encontros de qualificação para  
 the.M.PL matches of qualification for  
 as semifinais serão  
 the.F.SG semifinal will.be  
 transmitidos/\*transmitidas ao vivo  
 broadcasted.M.SG/\*F.SG to.the live  
 ‘The qualifying matches for the semifinals will be broadcast live.’

• **Basic Subject–Verb Agreement**

In Italian, Portuguese and Spanish, the finite verb tense and mood forms must agree in person and number with the grammatical subject, as in the Italian example below:

- (15) Tu cucini  
you.2SG cook.2SG
- (16) \* Tu cucinate/cucino/cucinano  
you.2SG cook.2PL/1SG/3PL  
‘You cook’

• **Subject–Verb Agreement with Subject Relative Clause**

This test suite, which has been adapted from the English test introduced by [Marvin and Linzen \(2018\)](#), focuses on number agreement. The subject relative clause includes an *agreement attractor* differing in number with the subject. The model is expected to assign higher probability to the verb agreeing with the subject (instead of the attractor), in both singular and in plural; cf. an example from Portuguese:

- (17) O encanador que ajudou os  
the.SG plumber that helped.3SG thePL  
pedreiros trabalha/\*trabalham de sábado.  
bricklayers work.3SG/3PL of saturday.  
‘The plumber who helped the bricklayers works/\*work on Saturdays.’
- (18) Os encanadores que ajudaram o  
the.PL plumbers that helped.3SG thePL  
pedreiro \*trabalha/trabalham de sábado.  
bricklayer work.3PL/3SG of saturday.  
‘The plumbers who helped the bricklayer \*works/work on Saturdays.’

• **Subject–Verb Agreement with Object Relative Clause**

As the previous test suite, this test is also on number agreement, only that it contains an object instead of a subject relative clause. Furthermore, in view of the stricter subject–verb order in Brazilian Portuguese, we introduce two versions of this test, one for Brazilian Portuguese and one for Italian and Spanish. The one for Portuguese follows the same pattern as the English version:

- (19) Os pedreiros que o encanador  
the.PL bricklayers that theSG plumber  
ajudou \*trabalha/trabalham de sábado.  
helped.3SG work.3SG/3PL of saturday.  
‘The bricklayers who the plumber helped \*works/work on Saturdays.’

- (20) O pedreiro que os encanadores  
the.SG bricklayer that the.PL plumbers  
ajudaram trabalha/\*trabalham de sábado.  
helped.3PL work.3SG/3PL of saturday.  
‘The bricklayer who the plumbers helped works/\*work on Saturdays.’

In the one for Italian and Spanish, the agreement attractor within the relative clause is adjacent to the critical region where the main verb is located. In this case, the agreement attractor is the subject of the relative clause, and thanks to these languages’ flexibility, it can appear post-posed to the subordinate verb and hence adjacent to the main one, as shown in the Spanish example below.

- (21) Los albañiles a los que ayudó  
the.PL bricklayers to the.PL that helped.3SG  
el fontanero \*trabaja/trabajan los  
theSG plumber work.3SG/3PL the  
sábados.  
saturdays.  
‘The bricklayers who the plumber helped \*works/work on Saturdays.’
- (22) El albañil al que ayudaron  
the.SG bricklayer to-the.SG that helped  
los fontaneros trabaja/\*trabajan los  
the.PL plumbers work.3SG/3PL the  
sábados.  
saturdays.  
‘The bricklayer who the plumbers helped works/\*work on Saturdays.’

**B.2 Russian**

For Russian, we experiment with ten different test suites. We explore nominal agreement–gender and number–within the noun phrase (six suites) and within the clause (two suites), and verbal agreement–person and number–within the clause (two suites).

• **Determiner–Noun Agreement in Genitive**

In this test, a noun in a genitive construction is paired with a possessive or demonstrative determiner that modifies it:

- (23) Машина твоего отца  
car your.GEN.M.SG father.GEN  
‘Your father’s car’
- (24) \* Машина  
car  
твоих/твоей/твой  
your.GEN.PL/GEN.F.SG/NOM.M.SG  
отца  
father.GEN

- (25) \* Машина твоя/твоими  
car your.GEN.PL/NOM.F.SG/INS.PL  
отца  
father.GEN

• **Determiner-Noun Agreement in Instrumental**

Analogous to the previous test suite. Here, a noun preceded by a specific preposition or verb that requires it to appear in instrumental case is paired with a determiner (a possessive or a demonstrative) that modifies it.

- (26) Я обедал со своей  
I had.lunch with my.INS.F.SG  
сестрой  
sister.INS.(F).SG  
'I had lunch with my sister.'
- (27) \* Я обедал со  
I had.lunch with  
своими/своим/своя  
my.INS.PL/INS.M.PL/NOM.F.SG  
сестрой  
sister.INS.(F).SG
- (28) \* Я обедал со своём/своих  
I had.lunch with my.PREP.M.SG/ACC.PL  
сестрой  
sister.INS.(F).SG

• **Determiner-Noun Agreement in Prepositional**

The same as before, but with a noun preceded by a preposition that requires prepositional case and a determiner (a possessive or a demonstrative) that modifies it.

- (29) На том столе  
on that.PREP.M.SG table.PREP.(M).SG  
'On that table'
- (30) \* На тех/той/тому  
on that.PREP.PL/PREP.F.SG/DAT.M.SG  
столе  
table.PREP.(M).SG
- (31) \* На то/те  
on that.ACC.N.SG/NOM.PL  
столе  
table.PREP.(M).SG

• **Adjective-Noun Agreement in Nominative**

This test suite pairs a noun in the subject position (hence in nominative case) with an adjective that modifies it:

- (32) красивая женщина спит.  
beautiful.NOM.F.SG woman.NOM sleeps.  
'A beautiful woman is sleeping.'

- (33) \* красивые/красивый/красивую  
beautiful.NOM.PL/NOM.M.SG/ACC.F.SG/  
женщина спит.  
woman sleeps.

- (34) \* красивого/красивым женщина  
beautiful.GEN.M.SG/DAT.PL woman  
спит.  
sleeps.

• **Adjective-Noun Agreement in Accusative**

This suite is analogous to the previous one, but with the noun in the object position (hence in accusative case):

- (35) Медсестра держала маленькое  
nurse held small.ACC.N.SG  
чадо.  
child.ACC.(N).  
'The nurse was holding a small child.'
- (36) \* Медсестра держала  
nurse held  
маленьких/маленькую/маленькому  
small.ACC.PL/ACC.F.SG/DAT.N.SG  
чадо.  
child.ACC.(N).
- (37) \* Медсестра держала  
nurse held  
маленькой/маленькими чадо.  
small.GEN.F.SG/INS.PL child.ACC.

• **Adjective-Noun Agreement in Dative**

This test suite is analogous to the previous two, but here the noun occupies a position (e.g., indirect object) in the sentence that requires dative case.

- (38) Старик радуется солнечному  
old-man enjoys sunny.DAT.N.SG  
утру.  
morning.DAT.(N)  
'The old man enjoys the sunny morning.'
- (39) \* Старик радуется  
old-man enjoys  
солнечным/солнечной/солнечном  
sunny.DAT.PL/DAT.F.SG/PREP.N.SG  
утру.  
morning.DAT.(N)
- (40) \* Старик радуется  
old-man enjoys  
солнечная/солнечными  
sunny.NOM.F.SG/INS.PL  
утру.  
morning.DAT.(N)

• **Predicative Attribute Agreement**

This test suite is similar to the corresponding test suite for Spanish, Italian and Portuguese. A noun is paired with an adjective through a copulative

construction. The main difference comes from the fact that in Russian the gender feature is neutralized in plural. This means that to be able to capture mismatches in gender, only singular subjects are to be used:

- (41) Квартира           кажется  
apartment(F).SG seems  
старой/\*старыми/\*старым.  
empty.F.SG/\*PL/\*M.SG  
'The apartment seems empty.'

Note that gender and number cannot disagree at once (as it happened with Spanish, Italian and Portuguese), since gender is not apparent in plural.

This suite has an adversarial version, with a relative clause (sometimes a reduced one) serving as modifier for the grammatical subject. The modifier includes an agreement attractor differing in gender or number with the subject:

- (42) Квартира,           которая была  
apartment.(F).SG which   was  
обставлена моим братом,           кажется  
furnished my brother.(M).SG seems  
пустой/\*пустым/\*пустыми.  
empty.F.SG/\*M.SG/\*PL  
'The apartment that my brother furnished seems empty.'

#### • Predicative Complement Agreement

This test suite is also similar to the corresponding test suite for Italian, Portuguese, and Spanish. The subject is paired with an adjective functioning as a predicative complement. Again, the main difference is that to be able to capture mismatches in gender, only subjects in singular are used.

- (43) Ребенок приехал  
kid           arrived  
счастливый/\*счастливая/\*счастливые.  
happy.M.SG/\*F.SG/\*PL  
'The kid arrived happy.'

As the previous one, this suite also has an adversarial version.

- (44) Ребенок, которого похвалила  
kid           who           was.praised  
воспитательница, приехал  
teacher                   arrived  
счастливый/\*счастливая/\*счастливые.  
happy.M.SG/\*F.SG/\*PL  
'The kid who was praised by the teacher arrived happy.'

#### • Basic Subject–Verb Agreement in Present/Future Tense

Finite verbs in present/future tense and indicative mood have six inflected forms according to person and number features. The verb's features must agree with the subject's:

- (45) Я           читаю книгу.  
I.1SG read.1SG book  
'I am reading a book.'
- (46) \* Я           читаем/читаешь/читают  
I.1SG read.1PL/2SG/3PL  
книгу.  
book

#### • Subject–Verb Agreement in Past Tense

In contrast, finite verbs in past tense and indicative mood have four inflected forms according to gender and number features (person is not involved). This applies to any person, but personal pronouns without context do not provide gender information, so the test only includes subjects in the third person singular (recall that gender feature is not apparent in plural).

- (47) Учитель           прочитал поэму в классе.  
teacher.(M).SG read.M.SG poem   in class  
'The teacher read a poem in class.'
- (48) \* Учитель           прочитала/прочитали  
teacher.(M).SG read.F.SG/PL  
поэму в классе.  
poem   in class

There is also an adversarial version of this test suite, as shown below:

- (49) Учитель,           которого ненавидели  
teacher.(M).SG who           was.hated  
девочки, прочитал поэму в классе.  
girl.PL   read.M.SG poem   in class  
'The teacher who the girls hated read a poem in class.'
- (50) \* Учитель,           которого ненавидели  
teacher.(M).SG who           was.hated  
девочки, прочитала/прочитали  
girl.(PL   read.F.SG/PL  
поэму в классе.  
book   in class



# Introducing KIParla Forest: seeds for a UD annotation of interactional syntax

**Ludovica Pannitto**  
University of Bologna  
ludovica.pannitto@unibo.it

**Eleonora Zucchini**  
University of Bologna  
eleonora.zucchini2@unibo.it

**Silvia Ballarè**  
University of Bologna  
silvia.ballare@unibo.it

**Cristina Bosco**  
University of Torino  
cristina.bosco@unito.it

**Caterina Mauri**  
University of Bologna  
caterina.mauri@unibo.it

**Manuela Sanguinetti**  
University of Cagliari  
manuela.sanguinetti@unica.it

## Abstract

The present project endeavors to enrich the linguistic resources available for Italian by introducing *KIParla Forest*, a treebank for the KIParla corpus - an existing and well-known resource for spoken Italian. This article contextualizes the project, describes the treebank creation process and design choices, and highlights future plans for next improvements.

## 1 Introduction

Today, the Universal Dependencies (henceforth, UD, de Marneffe et al. 2021) body of resources<sup>1</sup> counts 296 treebanks for 168 languages. While many different genres are represented among the corpora, ranging from news to fiction to legal texts, spoken language is surely underrepresented. This aspect strikes as counterintuitive if one thinks that language resources should mirror language use; however, it is actually in line with a tendency in the Natural Language Processing (NLP) community to rely on what is, or was, easily accessible for processing rather than truly representative. Spoken language, in fact, poses unique challenges when it comes to its representation for processing, some of which derive from the long-standing but unstated assumption that NLP is primarily Written Language Processing (Linell, 2019; Chrupała, 2023). As a result, while there is a shared consensus on the primacy of spoken over written language, only approximately 20 out of the 168 UD languages have a dedicated spoken treebank (Dobrovolic, 2022), and Italian is not among those. A greater availability of spoken treebanks would open the path to large-scale studies on phenomena typical of interactional data, such as conversational patterns, discourse markers, and syntactic variation, which are hard to scale above the lexical level with available resources. The NLP community has only

recently begun to focus on spoken languages, taking into account not only institutional languages but also dialects and endangered languages (Bird and Yibarbuk, 2024). The great diversity of these languages and their wide distribution make starting to study them particularly urgent. From the NLP perspective, accuracy rates of currently available pipelines drastically drop when running on spoken language varieties, and no spoken resource is currently available to train accurate annotation pipelines tailored to speech data (see, among others, Liu and Prud’hommeaux 2023).

We therefore introduce *KIParla Forest*, the first Universal Dependency treebank of Spoken Italian, derived from the KIParla corpus project (Mauri et al., 2019a; Ballarè et al., 2020). In this paper, we examine the motivations and major design choices taken in the first phases of the creation of the resource, focusing in particular on the pipeline from segmentation to syntactic annotation. *KIParla Forest* is planned for release in UD 2.17 in November 2025. Because of their complexity and the need for linguistic glosses, most examples are reported in Appendix A.

## 2 Universal Dependencies for Spoken Language

Increased attention to the syntactic annotation of spoken varieties within the Universal Dependencies framework is attested by the fact that the number of treebanks including or completely dedicated to spoken language is on the rise. UDv2.0 already included UD\_Slovenian-SST (Dobrovolic and Nivre, 2016), a treebank composed entirely of spoken data, and some spoken data in mixed-genre treebanks. Despite the fact that UDv2.16 sees now 48 treebanks counting both spoken-specific and mixed-genre treebanks that contain spoken data, a full set of guidelines dedicated to spoken-specific phenomena is yet to be released. Currently, a dedi-

<sup>1</sup>[www.https://universaldependencies.org/](https://universaldependencies.org/)

cated taskforce within the UniDive COST Action<sup>2</sup> is dedicated to analyzing and harmonizing current practices for morphosyntactic annotation of speech-specific phenomena. Currently, in fact, treebank curators took different directions in the creation of their resources, which could impact on derived measures or performance on downstream tasks (see Table 1 for an overview). Most spoken treebanks include information about alignment and metadata about speakers and language variety. As far as capitalization and punctuation are concerned, some take a written-derived approach, normalizing the transcription with added capitalization and written-like punctuations, while others (for instance, UD\_Beja-Autogramm Kahane et al. 2022) employ it to represent prosodic traits. Fillers and filled pauses are reported in most treebanks, mostly with conventionalized transcriptions (e.g., *eah* in French, *e* in Norwegian or *ähm* in Turkish-German), either marked as  $X$  or INTJ (we choose the latter) and generally labeled as discourse or discourse:fillers, attaching to the root of the sentence. Discourse markers are generally marked according to their syntactic category (they could be verbal, adverbial, interjections, etc). They are generally labeled as discourse, while Naija NSC (Caron et al., 2019), Slovenian SST (Dobrovoljc and Nivre, 2016) and Turkish-German SAGT (Çetinoğlu and Çöltekin, 2019) use `parataxis:discourse` for distinguishing clausal markers.

	Beja	Cantonese	Chinese	Chukchi	ParisStories	Rhapsodie	Frisian-Dutch	Komi-Zyrian	Naija	Norwegian	Slovenian	Turkish-German	KIParlaForest
Sound file ID	yes	no	no	yes	yes	no	no	no	yes	no	no	no	no*
Text-sound alignment	yes	no	no	yes	no	no	no	no	yes	no	no	no	yes*
Speaker ID	no	no	no	no	yes	yes	yes	no	yes	yes	no	no	yes
Language variety	no	no	no	no	no	no	yes	yes	no	yes	no	yes	yes
Standard orthography	no	no	yes	yes	yes	yes	yes	no	no	yes	yes	yes	yes
Capitalization	no	no	yes	no	no	no	no	yes	no	no	no	yes	yes*
Pronunciation	yes	no	no	yes	no	no	no	no	no	no	yes	no	no
Speaker overlap	no	no	no	no	no	yes	no	no	no	no	yes	no	yes
Final punctuation	yes	yes	yes	yes	yes	yes	no	yes	yes	yes	no	yes	no
Other punctuation	yes	yes	yes	no	yes	yes	no	yes	yes	yes	yes	yes	no
Incomplete words	no	no	no	yes	yes	yes	no	no	yes	yes	yes	yes	yes
Fillers	no	no	no	no	yes	yes	yes	no	yes	yes	yes	yes	yes
Silent pauses	yes	no	no	no	no	no	no	no	yes	yes	yes	no	yes
Incidents	no	no	no	no	no	no	no	no	no	no	yes	no	yes

Table 1: The table, adapted from (Dobrovoljc, 2022), compares features to be found in spoken UD treebanks and features that will be available in KIParlaForest (rightmost column). The table is not exhaustive as, since Dobrovoljc’s paper in 2022, new treebanks of spoken data have appeared within the UD family of resources.

### 3 Universal Dependencies for Italian

Italian has a very solid tradition in the UD enterprise, with resources already appearing in the first release dating back to 2015 (Nivre et al., 2015). The first UD-based treebank ever released for Italian is ISDT (Italian Stanford Dependency Treebank), originally developed for the dependency parsing shared task of EVALITA-2014 (Bosco et al., 2014). In the current release, ISDT contains approximately 298K tokens and includes texts pertaining to the legal domain, or harvested from news and Wikipedia. Another Italian treebank, UD-VIT (Alfieri et al., 2016) was obtained by semi-automatically converting the Venice Italian Treebank (Delmonte et al., 2007), which included approximately 60K words of spoken data in its original version. However, to the best of our knowledge, this portion was not ported into the UD resource.

Spoken data, or as we could better define it ‘conceptually-written’ (Koch and Oesterreicher, 2012) or ‘spoken-written’ (Nencioni, 1976) language, is also collected in the ParlaMint corpus (Agnoloni et al., 2022; Alzetta et al., 2024), built from stenographic verbatim records of parliamentary speeches. Whereas, as the authors say, ‘debates of the COVID-19 period are mostly characterised by traits specific to the spontaneous speech’, no detailed description of such features is provided and no measures are described to adapt UD guidelines to such genre. Similarly, a section of the parallel treebank ParTUT (Sanguinetti and Bosco, 2014, 2015) features annotated data from the Europarl corpus (Koehn, 2005), a collection of texts from the proceedings of the European Parliament.

Two resources that are not specific for spoken language but are still relevant for our work are PoSTWITA-UD (Sanguinetti et al., 2018) and TWITTIRÒ-UD (Cignarella et al., 2019), which contain collections of tweets: in these cases, explicit choices were made to extend UD guidelines to non-standard productions, in particular extending the `parataxis` relation to systematically cover a class of juxtaposition phenomena. Many of these guidelines are collected in Sanguinetti et al. (2023), that describes annotation choices for user-generated content. Lastly, among the resources of interest to our domain, is MarkIT (Paccosi et al., 2023), which contains around 800 sentences, extracted from students’ essays, covering seven types of marked constructions, many of which are also typical of spoken data, such as for instance hang-

<sup>2</sup>CA21167, <https://www.cost.eu/actions/CA21167/>

ing topic sentences or sentences with presentative *there*. In this scenario, KIParla Forest would thus represent the first attempt to develop a fully-spoken treebank for Italian. The following section will outline the corpus from which this treebank originates.

## 4 Data

The KIParla corpus<sup>3</sup> (Mauri et al., 2019a; Ballarè et al., 2020) is a resource for the study of spoken Italian and is a product of a collaborative effort between the Universities of Bologna and Turin. It is structured in an incremental and modular fashion that allows the addition of new corpus modules over time. To date, KIParla encompasses a diverse range of Italian spoken varieties and involves participants of various age, genders and backgrounds and with different professional and educational achievements. As a whole, the KIParla counts ca. 228 hours of recordings and approximately 2M transcribed tokens. At the time of writing, the corpus is freely available for consultation through a custom noSketchEngine service<sup>4</sup>, that provides transcriptions, carried out manually following Jefferson guidelines (Jefferson, 2004), aligned with audio files; access to full transcripts is also provided. Preliminary linguistic annotation efforts on the KIParla corpus were initiated during the EVALITA<sup>5</sup> evaluation campaign in 2020. The KIPoS task<sup>6</sup> (Bosco et al., 2020) precisely focused on Part-of-Speech tagging of KIParla data, comprising approximately 200K tokens automatically annotated with UDpipe and partially manually revised. KIParla contains recordings collected in different conversational settings. To create the core of KIParla Forest, a balanced sample of such data was selected to showcase syntactic annotation of conversations presenting different degrees of interactional freedom, and including various number of speakers. Then, the chosen conversations were organised based on interactional levels identified in the KIParla corpus, ranging from *free interaction* (free conversations), to *partially free interaction* (semi-structured interviews), *rigid interaction* (university exams and office hours) and situations with *almost no interaction* (lectures).

When selecting conversations, we made sure we

<sup>3</sup>[www.kiparka.it](http://www.kiparka.it)

<sup>4</sup><https://search.corpuskiparla.it/corpus/crystal/#open>

<sup>5</sup><https://www.evalita.it/>

<sup>6</sup><http://www.di.unito.it/~tutreeb/kipos-evalita2020/index.html>

CODE	TOD1005b6	BOD2018	PBB004	BOA3017
TYPE	lecture	interview	interview	free conversation
INTERACTION LEVEL	almost none	partially free	partially free	free
N. TOKENS	6788	4634	5898	4551
DURATION	00:50:44	00:28:08	00:35:54	00:30:22
PARTICIPANTS	1	2	3	4
KIPoS	yes	yes	no	yes

Table 2: Conversations selected for the first release of KIParla Forest.

included those that had been already manually annotated during the KIPoS task, in order to capitalize on the gold part of speech annotations already in place<sup>7</sup>. The final selection is reported in Table 2. All summed up, the treebank counts 21.871 tokens.

### 4.1 Data preparation

KIParla conversations are manually transcribed through ELAN<sup>8</sup> (Max Planck Institute for Psycholinguistics, The Language Archive, 2024) and stored in .eaf format. The native transcription format includes a Jefferson-inspired set of conventions to represent features of spoken language (intonation, pace, pauses, overlaps, repair...). The first step towards the construction of the treebank consisted, therefore, in fully transforming the current notation into a columnar format, therefore isolating orthographic tokens from prosodic features annotated in Jefferson notation. Since not all Jefferson features will be included in the UD treebank, we made sure that each orthographic token bears a unique token identifier (TID) in order to retrieve, in combination with sent\_id, more specific features and to ensure backward compatibility with the KIParla resource. As a result of our normalization process, each conversation is represented in a conll-like format. The conversation is divided into Transcription Units (TUs), manually identified by transcribers and aligned with audio. TUs are then split into orthographic tokens, each annotated with Jefferson-derived features.

### 4.2 Speech-specific metadata

Most spoken treebanks include speech-specific metadata such as links to audio files, information about the speaker and on language variety. As audio access is restricted to registered users, for privacy reasons, an explicit link to the audio file cannot be provided as of today in KIParla Forest. All audio files and speaker-specific metadata are available upon request, only for research purposes.

<sup>7</sup>see Section 7 regarding the modifications that were implemented.

<sup>8</sup><https://archive.mpi.nl/tla/elan>

Two attributes (`AlignBegin` and `AlignEnd`, expressed in milliseconds), typically attributed to the first and last token of each TU, are provided at token level in the `MISC` field of the CoNLLU files. Differently from other spoken corpora, that provide speaker information at the maximal unit level, in our treebank each token bears a special `SpeakerID` feature that contains the id of the speaker as a value. Each speaker is then described through its metadata (including data such as gender, age, origin, education level, profession) in a separate json file available in the treebank repository. The same applies to conversation-specific metadata (i.e., number of participants, place and date of recording, type of interaction). The resource also contains information about *overlaps*: these represent a particularly challenging feature both to annotate and to parse, as single tokens can participate in more than one overlapping span, and overlaps can happen among two or multiple speakers. We have adopted a special `Overlap` feature in the `MISC` column, attributed to all tokens that participate in overlapping spans. The feature value is composed as a comma-separated list of ranges, where each range has format: `idX-idY@sent_id-n+...+idT-idS@sent_id-m` (see Figure 1). Figure 2 shows how the feature is rendered in the different overlapping scenarios. Figures 5 and 6 in Appendix A demonstrate the annotation.

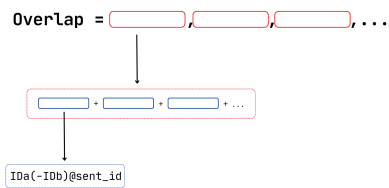


Figure 1: The figure shows the composition of the `Overlap` feature. The feature value is composed as a comma-separated list of pointers (top tier of the scheme), where each range is in turn composed of a + separated list of ranges. Each range is then a reference to a specific token or sequence of tokens, identified by their CoNLLU IDs and sentence identifier.

The next sections describe the design choices we made to transform such preliminary data collection into proper UD-compatible data and to operationalize certain annotation decisions, starting from the basic segmentation steps up to the syntactic level.

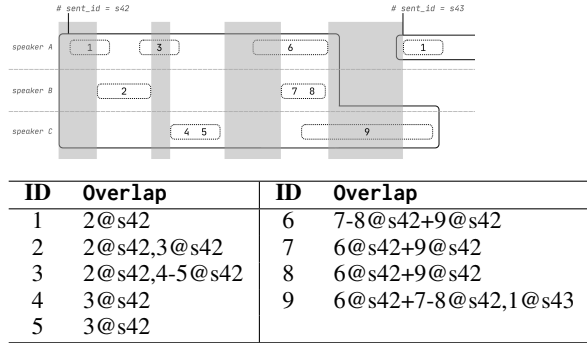


Figure 2: The figure shows some cases of overlaps. The example shows an extract of a conversation where three speakers (A, B and C) are involved. Dotted boxes represent TUs, while numbers represent orthographic tokens as they would be numbered in CoNLLU. TUs are in fact grouped in two maximal units, namely `s42`, composed by 9 tokens, and `s43`, where only token 1 is visible. Token 2 in sentence `s42`, for instance, overlaps partially with token 1 and partially with token 3: its `Overlap` feature would therefore be constituted by two different span references. Token 6, on the other hand, participates to a complex span, where all three speakers overlap. This translates into a + separated sequence of references, accounting for the fact that overlap with tokens 7-8 and with token 9 is simultaneous. Token 9 shows a combination of the two overlapping situations.

## 5 Segmentation into maximal units

Segmentation of spoken data into maximal units has already been tackled by existing spoken language treebanks in UD; however, the documentation mostly lacks information on the formal criteria that were adopted (Dobrovoljc, 2022): a popular choice is that of illocutionary units (IUs, Cresti et al. 1995; Pietrandrea et al. 2014), defined as speech segments that correspond to a single speech act, or linguistic units serving to express a single primary idea. On the other hand, it is well known that, especially in conversational settings, syntactic affordances are exploited across turns by different speakers, to co-construct the structure of discourse (Du Bois, 2014). As a consequence, given that syntactic relations can be observed only within the sentence boundaries, their identification should be carried out in a way that does not obscure relations within a broader linguistic context and allows to natively represent syntactic co-construction among speakers (see Figure 7 in Appendix A). The need for a more careful definition of the maximal unit, i.e., the domain in which syntactic relations hold, is also demonstrated by the introduction of the feature `AttachTo` in treebanks of spoken language (Kahane et al., 2021b),



for cases of co-construction.

When dealing with the transcriptions of the KIParla corpus, it is necessary to consider that TU boundaries cannot be treated as reliable basis for segmentation since they do not represent sentence-like units or utterances in any meaningful way: their boundaries remain highly subjective and there are many examples where core grammatical relations hold between elements of different TUs (see Figure 8 in Appendix A). As a result, we decide for the KIParla Forest to explore a different perspective: we begin with annotating dependencies between words, establishing a unit boundary whenever no dependency link can be found. Such an approach allows boundaries to emerge bottom-up, purely based on the existence of grammatical relations, rather than the reverse; moreover, it enables us to distribute syntax along the speech stream and represent the cooperative - potentially inter-speaker and inter-turn - organization of syntax. While in fact we acknowledge the centrality of IUs in speech, we do not believe they are useful tools to identify syntactic maximal units, as their speaker-bound nature can obscure this cooperative nature of spoken syntax.

This said, for technical reasons we needed a preliminary sentence-like segmentation in order to run an automatic pipeline to pre-annotate parts of speech and also to visualize text in a reasonable way on the annotation tool of our choice (i.e., ArboratorGrew, Guibon et al. 2020). Therefore, we decided to employ `wtpsplit` (Minixhofer et al., 2023), an unsupervised multilingual sentence segmentation system that does not rely on punctuation marks to segment a textual stream. We applied the model to the entire conversation, ordering tokens based on their time of utterance, regardless of speaker. The output of `wtpsplit` constitutes the basis for further steps of analysis and annotation: as it was done in KIPoS, automatic PoS tagging and syntactic parsing was performed through UD-Pipe (Straka, 2018), based on the model trained on PoSTWITA-UD data from release 2.15 (Straka, 2024). The model was chosen in continuity with what was done during KIPoS task, under the assumption that social media data would share some of the features of spoken language, such as for instance the extended use of discourse markers and increased use of parataxis relations. This constitutes the basis on which annotators will then be free to merge or split such automatically-identified units in the syntactic annotation phase, based on

the identification of local grammatical relations. The segmentation phase is thus divided into an automatically-driven one, initial and purely operational, and a manually revised one, which emerges out of the identification and annotation of locally relevant grammatical relations. As a consequence, the identification of maximal units is not anterior to the annotation stage in our approach, but rather emerges from it (see Section 8 for more details).

## 6 Tokenization and lemmatization

Aside of the Jefferson notation, the KIParla project uses standard orthography and spelling, which, in the case of Italian is not particularly problematic. The only difference between UD-like tokenization and the one natively available through the transcription is the case of multiword tokens, which are used in Italian treebanks for article-preposition contractions and cases of clitics attached to verbs. These were split by our pre-annotation pipeline, which introduced multiword tokens. Metadata such as SpeakerID, TID, and Jefferson-derived features remain on the multiword token, while new syntactic tokens receive distinct TIDs for backward compatibility with KIParla (these are created during the parsing step and kept in an intermediate pivot file that allows to cross-reference the corpus and the treebank).

The KIParla resource includes both code-switching and dialectal variation, which is currently identified at TU level by the introduction of a # symbol in the Jefferson transcription, at the beginning of each TU. The information about variation remains therefore available among the metadata of each maximal unit (# contains\_variation). As new KIParla modules will involve L2 speakers and showcase examples of code-switching, we are experimenting the ‘Code-switched analysis’ currently proposed by UD guidelines<sup>9</sup>: we therefore differentiate between cases where the foreign material is borrowed and incorporated in Italian, by fully considering them same as Italian material, and cases where we apply the analysis (either morphological or syntactic) of the target language. Such cases are marked by the feature Lang=CODE in MISC. Ambiguous cases will be annotated as foreign language only when considering them Italian is impossible; unknown cases will be coded with Lang=UNKNOWN<sub>ISO</sub>. Dialects, for the moment,

<sup>9</sup><https://universaldependencies.org/foreign.html#option-1-code-switched-analysis>

have been coded with Lang=NO\_ISO\_CODE for the lack of a dedicated ISO-639 code (see Figure 3). Furthermore, KIParla contains special tokens to represent non-linguistic behaviour and instances of anonymization (home addresses; work places and the like). Non-linguistic behaviour includes short pauses, tags expressing actual non linguistic behaviour (NLB, e.g., ((*laughter*))), annotations expressing modality of utterance (e.g., ((*reading*))), events happening outside of the interaction (OOI, e.g., ((*phone ringing*))) and notes (e.g., ((*recording interrupted*))). These cases are treated differently when imported into the treebank. More specifically, short pauses are transformed into a PauseAfter=Yes feature in MISC. Cases of true non linguistic behaviour are only kept when relevant to the syntactic construction of the units, with their forms and lemmas uppercased and a feature Type=NLB is added to the MISC column (see Figure 9 in Appendix A). Modalities are not included in the treebank as tokens, but a feature Type=reading|singing|... is added in the MISC column on the relevant linguistic tokens (see Figure 10 in Appendix A). OOI events and annotations are kept as metadata at the maximal unit level. Finally, as far as anonymized tokens are concerned, as done in PoSTWITA-UD, instances of anonymized tokens are prefixed by @: examples include cases such as '@nomepaese' (en. '@villagenam'). It is worth mentioning that all personal first names (except for recognizable names e.g., celebrities) are pseudonymised: they are replaced with a different name of approximately the same length; therefore, such instances are considered as normal tokens.

Concerning lemmatization, a few choices need to be discussed. While the original transcription contains no capital letters at all, all proper nouns' lemmas have been capitalized in order to facilitate downstream tasks that might require named entity recognition. Words interrupted during speech (i.e., false starts) are lemmatized as their complete version whenever the context is informative enough, either because there is a repetition surrounding the interrupted word (see Example 1) or because there is compatible syntactic context preceding or following the token, as in Example 2. We did not trust semantic predictability to be informative enough, so we did not lemmatize cases as the one in Example 3. In this case semantics would suggest 'person' (en. 'people') as the interrupted lemma, but we excluded these cases as a clear repetition was missing. A feature Interrupted=Yes is reported

```

961, BOI012 >lo so che bologna< è basket
              3SG.OBJ know.1SG that Bologna is basket
                                                    *
city ma::
city but
*
'I know that Bologna is "basket city" but'

(a)
#pa se vuoi fazzu eu
da(d) if want:2SG do:1SG 1SG
              * *
'dad if you want I can do it'

(b)

```

Figure 3: Both examples show code-switching phenomena, example 3a includes English elements while 3b includes elements from an Italo-Romance dialect. Tokens marked with \* have features Foreign=Yes and Lang=eng in 3a, and feature Lang=NO\_ISO\_CODE in 3b. (from conversations PBB004 and KPS001)

in MISC in all cases.

- (1) vabbè scusa è **sta**~ è **stato**  
well sorry AUX.3SG be~ AUX.3SG been  
più...  
more  
'sorry it's be~ it's been more...' (conversation BOA3017)
- (2) e non è una città **vic**~ che è vicino a  
and NEG is INDEF city cl~ that is close to  
tante possibilità  
many possibilities  
'and it is not a city cl~ that is close to many possibilities' (conversation BOD2018)
- (3) generalmente: [non] conosco person~  
generally NEG know:1SG peop~  
famiglie: che bolognesi che abitano in  
families that from.bologna that live:3PL in  
centro  
centre  
'I generally don't know peop~ families that from Bologna that live in the city centre' (conversation BOD2018)

Acronyms are transcribed through their phonetic realization, at the form level, and they are lemmatized as their dictionary entries ('RSA', en. 'nursing home', lemmatized as such but transcribed as *erreessea*, its phonetic realization). Interjections and ideophones are transcribed but are normalized, at lemma level, to the lexical entry that can be found in Italian dictionaries (for instance, 'okay' is kept as such at the form level but lemmatized as 'ok').



## 7 PoS tagging and morphology

The KIPoS task, the first attempt at PoS annotation on the KIParla corpus, was carried out using a tagset only inspired by UD tagset, that included also PoS labels introduced on purpose. Specifically, NEG was employed for sentence negation, PARA for particles pertaining to paraverbal communication, DIA and LIN as subtypes of any UPOS to mark Italo-Romance dialectal variation and languages other than Italian<sup>10</sup>. Therefore, in our process, after automatic annotation we aligned our data with KIPoS gold datasets, having restored the UPOS labels, in order to retrieve as much gold annotation as possible. This was then the basis for manual correction. Annotation (both for morphosyntax and syntax) was performed collaboratively by the authors through ArboratorGrew.

We operated by the following criteria. Fillers and filled pauses, which include cases such as *beh*, *eh*, *ehm* and *mh*, are marked as INTJ. Interrupted words are tagged either with the PoS of their repair (see Section 6) or with X. We align with French Rhapsodie, ParisStories (Kahane et al., 2021a) and Naija NSC (Caron et al., 2019) marking them with the  $\sim$  symbol at the end of the form in order to avoid any possible overlap with Italian words that contain an hyphen. We adopt a rather conservative approach when assigning PoS labels, sticking to the main category of each word, even though they perform a different function in the syntactic context. An example may be the word *basta* (en. ‘stop’), which is an inflected form of the verb *bastare* (en. ‘to stop’) but also works as discourse marker meaning ‘that’s it’. In line with choices taken by other spoken language treebanks, all discourse markers are marked according to their morphological category (e.g. verbs, adverbs, interjections, etc.). We specifically questioned the annotation of determiners: we restricted the use of the DET label only for articles, demonstratives and quantifiers, while considering any other elements of the noun phrase, both preceding and following the head, as modifiers of the noun (be they *adjectival*, *numeral* or *possessive*). This allows for a consistent annotation of diatopical variation concerning, for instance, the position of elements such as *mio*, *tuo*, *suo...* (en. *my*, *your*, *his/her...*), which may precede or follow the nominal head (e.g. the use of ‘*il mio libro*’ over

<sup>10</sup>Moreover, because of technical issues, the data employed for the KIPoS task was not entirely identical to the current version of the corpus.

‘*il libro mio*’ can depend on a number of factors, which also include simple diatopical variation with no implications on the linguistic relation between ‘*mio*’ - *my* - and ‘*libro*’ - *book*). However, in cases where modifiers such as possessives exclude the presence of a properly defined determiner (e.g., ‘*mia mamma*’, en. *my mom*), these are tagged as DET. We manually revise morphological features while we do not annotate XPOS.

We computed Cohen’s  $\kappa$  to evaluate inter-annotator agreement (Artstein and Poesio, 2008) on UPOS labeling, obtaining almost perfect agreement (above 0.87<sup>11</sup>) in all our scenarios. For the agreement task, an external annotator was asked to annotate UPOS labels on approximately 1500 tokens from each of our four conversations. We provided the annotator files pre-annotated with the PoSTWITA UDPipe model and set up a dedicated project on ArboratorGrew. We also instructed the annotators with the criteria described in this section. As expected, most disagreement is registered between CCONJ and ADV, which are the ones more prone to develop discourse functions.

## 8 Syntax

Dependency trees in the UD formalism are directed acyclic graphs that have tokens as nodes and *grammatical relations* as edges, with no notion of constituency or bracketing allowed. However, not all edges allowed in UD represent syntactic relation in the strict sense (Mel’cuk, 1988; Tesnière, 2015): there exist relations like *flat* or *goeswith* that aim at representing exocentric constructions or at allowing to treat phenomena that are more pertinent to the form in which data presents itself, rather than its actual linguistic structure (de Marneffe and Nivre, 2019). UD is also already equipped with a set of relations that seem to have been introduced with speech in mind: in particular *discourse* that is used for interjections and other discourse particles and elements not clearly linked to the structure of the sentence, *reparandum* for disfluencies overridden in a speech repair and *parataxis*, whose cases of applications are manifold and include discourse relations in linking clauses and tag questions. In designing our treebank, we tried to take advantage of these already defined relations, while questioning the need for more fine-grained analysis of spoken-language specific phenomena. More specif-

<sup>11</sup>More precisely,  $\kappa = 0.87$  for BOD2018,  $\kappa = 0.88$  for BOA3017,  $\kappa = 0.88$  for TOD1005bis,  $\kappa = 0.91$  for PBB004.

73, BOI013 se perdo un autobus poi devo  
 if lose:1SG INDEF bus after must:1SG  
 s~ (a)spectare un'altra ora  
 w~ wait another hour  
 'if I miss a bus then I have to wait for another hour'

74, BOR005 >ah okay **quindi**< la mobilità è molto  
 ah okay **so** DEF mobility is very  
 ridotta  
 reduced  
 'ah okay **so** mobility is very reduced'

Figure 4: A case where ‘quindi’ (e. ‘so’), usually used as a connective, develops a discourse function as the antecedent of the connective is missing. (from conversation PBB004)

ically, we label interjections and filled pauses as *discourse*, attaching them to the closest projective head. The relation *reparandum* is to be used for disfluencies and self-repairs, that concern both individual words or longer chunks. In this case, the false start or interrupted token is linked to its repair. The biggest issues arise when dealing with clause-linking criteria and, therefore, in relation to segmentation. As described in Section 5, we choose not to segment conversations based on a priori criteria, but we rather start from the annotation of local, purely syntactic dependencies and establish a boundary whenever no further dependency can be found. Such an approach to segmentation seems to be more adherent to how speakers construct syntactic relations, that is, incrementally, by progressive, unplanned expansions (see Figure 11 in Appendix A) that exploit syntactic connections to keep discourse tightly interwoven and cohesive (e.g. through relative clauses, conjunctions, lists - see Masini et al. 2018; Mauri et al. 2019b -, etc.). Such an approach, however, comes with consequences, both for segmentation and relation labeling. The first problem regards connectives, because they frequently develop discourse functions (see Example 4, the connective is rendered in bold) and it may be difficult to tell cases in which they create a local syntactic link from cases in which they indicate some general anaphoric discourse relation; however, choosing among the two strongly affects segmentation.

We are still working on a precise set of criteria to deal with such cases. Two parameters that we currently rely on are prosody (i.e., the syntactic annotation task has to be performed while listening to the recording) and the identifiability of a clear head to which the connective should be attached.

If the connective is linked to a larger portion of conversation and/or it is not possible to identify a clear preceding head, then we set a maximal unit boundary before the connective itself. In this latter case, the connective is identified as having a *discourse marker* function. Discourse markers are indeed typical of spoken language and are connected to their head through the *discourse* relation. Interestingly, discourse markers frequently involve more than one word. Following what is done with complex prepositions, we chose to identify well established cases through the *fixed* relation. A further problem in unit boundaries identification occurs when the grammatical relationship is not overtly marked, as in parataxis. In such cases (Figure 11), in particular in the case of ‘eh uno studio sui ‘riti magici’ (en. ‘well, a study on magic rites’), it may be difficult to decide whether we are dealing with listing or implicit reformulation occurring within the same unit, or with independent segments. Prosody here plays a crucial role, because independent segments typically correlate to separate prosodical contours (Mithun, 1988), while there is rich evidence in the literature for clearly identifiable intonational patterns associated to lists (Masini et al., 2025). Moreover, spoken data require to find a specific way to treat feedback phenomena; since the extent of such phenomena varies greatly, we have hypothesized the following solutions. We consider as internal to the maximal units all feedback phenomena that do not interfere with the main syntactic flow (see Figure 12), regardless of the speaker who is uttering them; in such cases, we propose to link the expression providing feedback to the unit head through the *ad hoc* label *discourse:feedback*; no maximal unit boundary is thus identified. In case of feedback phenomena that interrupt the syntactic flow, which may or may not cause replanning, we have to proceed being aware that we are dealing with a continuum: we set a maximal unit boundary if the portion following feedback has no clear syntactic relation to the portion preceding it. Obviously, there may be cases that are more difficult to assign to one of these two types; as with the other pending issues, we are working on testing the validity of our hypothesis on larger sets of data.

## 9 Conclusion and Future work

One might wonder: why taking the trouble to create a treebank for spoken language, if most of the

categories seem to be ill-defined when applied to spoken data? We asked ourselves the same question while working on this project, and we came to the following conclusion. On the one hand, we hope to lay the foundations of a future discussion on the categories themselves, giving a contribution from the perspective of ecological spoken data, i.e. naturally-occurring spoken data, collected in real-life communicative contexts, rather than in artificial or experimental settings. This perspective emphasizes the importance of capturing language as it is actually used by speakers in their everyday social interactions, preserving the features of spontaneity, interaction, variation, and context-dependence that characterize real-world speech. On the other hand, the creation of a spoken treebank is, in our case, also aimed to offer an additional level for accessing and querying the resource: when it comes to spoken data, in fact, interfaces are typically limited to form-based queries, highly restricting the range of possible data explorations. For these reasons, the choices we made about what kind of relations are considered *grammatical relations* are tailored to represent the interactional architecture of ecological spoken data. In our work, our design choices try to follow the competing criteria taken as design choices for the UD formalism (de Marneffe et al., 2021), without favoring one perspective over the others (we comment here on the relevant ones):

- *UD needs to be reasonably satisfactory on linguistic analysis of individual languages:* Italian shows great internal variation, not only wrt. to written vs. oral modalities, but also in terms of regional and register variation. The development of KIParla Forest aims to move a step forward in the process of representing intra-linguistic diversity;
- *UD needs to be good for linguistic typology:* a treebank of spoken language avoiding *a priori* segmentation, based instead on local (and incremental) syntactic relations, allows to represent and extract phenomena of syntactic chains along speech. This allows for better typological comparability with (often purely oral) languages showing, for instance, clause-chaining phenomena (Mansfield and Barth, 2021);
- *UD must be suitable for rapid and consistent annotation:* we kept the modifications to the UD annotation procedure to the minimum, to

favor consistency and rapidity and limit the need to learn new rules;

- *UD must support well downstream tasks:* while the role of resources in the LLM era is a challenging discussion topic, we believe that our choices could be fit to support tasks that require rich semantic representations, based on larger discourse context, as well as open the path to benchmarks dedicated to interactional fluency.

While our morphosyntactic choices have been satisfactorily validated through inter-coder agreement, future work is needed on the validation of syntactic criteria. Moreover, as more and more spoken treebanks are being released, we foresee a broader discussion within the community to agree on common annotation guidelines for spoken-specific phenomena.

## 10 Limitations

The paper describes initial steps towards the release of a new resource. We therefore acknowledge that many of our statements are to be considered preliminary and are likely to be rediscussed and updated as new data are integrated in the resource. Moreover, UD is still lacking stable and shared guidelines on the annotation of spoken data. We will participate in the community debate to develop shared guidelines and update our choices accordingly.

## 11 Ethical considerations

The KIParla corpus is compliant with current European data protection regulations (Data protection - European Commission<sup>12</sup>); all data are recorded with overt microphones and speakers provide a written consent to the collection and usage of the data for research purposes. Before upload, audio tracks and transcriptions are pseudonymized, by removing all sensitive information. Metadata regarding speakers and conversations are stored and shared in an aggregated format that prevents speakers' recognition. The treebank is automatically linked to the original data, and the choices taken ensure the possibility of removing data, should speakers revoke their consent.

<sup>12</sup>[https://commission.europa.eu/law/law-topic/data-protection\\_en?prefLang=it](https://commission.europa.eu/law/law-topic/data-protection_en?prefLang=it)

## Acknowledgments

The research leading to these results has received funding from Project "DiverSIta-Diversity in spoken Italian", prot. P2022RFR8T, CUP J53D23017320001, funded by EU in NextGenerationEU plan through the Italian "Bando Prin 2022 - D.D. 1409 del 14-09-2022". This work also received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology). We want to specifically thank student that helped and will help up annotating the treebank.

## References

- Tommaso Agnoloni, Roberto Bartolini, Francesca Frontini, Simonetta Montemagni, Carlo Marchetti, Valeria Quochi, Manuela Ruisi, and Giulia Venturi. 2022. Making italian parliamentary records machine-actionable: The construction of the parlamint-it corpus. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 117–124.
- Linda Alfieri, Fabio Tamburini, and 1 others. 2016. (almost) automatic conversion of the venice italian treebank into the merged italian dependency treebank format. In *CEUR WORKSHOP PROCEEDINGS*, volume 1749, pages 19–23. AAccademia University Press.
- Chiara Alzetta, Simonetta Montemagni, Marta Sartor, and Giulia Venturi. 2024. *Parlamint-it: an 18-karat UD treebank of Italian parliamentary speeches*. *Language Resources and Evaluation*.
- Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.
- Silvia Ballarè, Caterina Mauri, and 1 others. 2020. La creazione del corpus kifarla: criteri metodologici e prospettive future. *RID, RIVISTA ITALIANA DI DIALETTOLOGIA*, 44:53–69.
- Steven Bird and Dean Yibarbuk. 2024. Centering the speech community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics - Volume 1: Long Papers*, page 826–839. ACL.
- Cristina Bosco, Silvia Ballare, Massimo Cerruti, Eugenio Gorla, and Caterina Mauri. 2020. Kipos@evalita2020: Overview of the task on kifarla part of speech tagging. *EVALITA Evaluation of NLP and Speech Tools for Italian-December 17th, 2020*.
- Cristina Bosco, Felice Dell’Orletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. *The Evalita 2014 Dependency Parsing task*. *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*, pages 1–8. Publisher: Pisa University Press.
- Bernard Caron, Marine Courtin, Kim Gerdes, and Sylvain Kahane. 2019. A surface-syntactic ud treebank for naija. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 13–24. Association for Computational Linguistics.
- Özlem Çetinoğlu and Çağrı Çöltekin. 2019. Challenges of annotating a code-switching treebank. In *Proceedings of the 18th international workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 82–90.
- Grzegorz Chrupała. 2023. *Putting Natural in Natural Language Processing*. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7820–7827, Toronto, Canada. Association for Computational Linguistics.
- Alessandra Teresa Cignarella, Cristina Bosco, and Paolo Rosso. 2019. Presenting twittirò-ud: An italian twitter treebank in universal dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 190–197.
- Emanuela Cresti and 1 others. 1995. Speech act units and informational units. In *E. Fava (haz.), Speech Acts and Linguistic Research, Proceedings of the Workshop, Center for Cognitive Science, State University of New York at Buffalo, Nemo, Padova*, pages 89–107.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. *Universal Dependencies*. *Computational Linguistics*, 47(2):255–308.
- Marie-Catherine de Marneffe and Joakim Nivre. 2019. *Dependency grammar*. *Annual Review of Linguistics*, 5(Volume 5, 2019):197–218.
- Rodolfo Delmonte, Antonella Bristot, and Sara Tonelli. 2007. Vit-venice italian treebank: Syntactic and quantitative features. In *Sixth International Workshop on Treebanks and Linguistic Theories*, volume 1, pages 43–54. Northern European Association for Language Technologies.
- Kaja Dobrovoljc. 2022. Spoken language treebanks in universal dependencies: An overview. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1798–1806.
- Kaja Dobrovoljc and Joakim Nivre. 2016. The universal dependencies treebank of spoken slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1566–1573.
- John W. Du Bois. 2014. *Towards a dialogic syntax*. *Cognitive Linguistics*, 25(3):359–410.



- Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *LREC 2020-12th Language Resources and Evaluation Conference*.
- Gail Jefferson. 2004. [Glossary of transcript symbols with an introduction](#). In Gene H. Lerner, editor, *Pragmatics & Beyond New Series*, volume 125, pages 13–31. John Benjamins Publishing Company, Amsterdam.
- Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021a. Annotation guidelines of ud and sud treebanks for spoken corpora. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages pp–35. Association for Computational Linguistics.
- Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021b. [Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal](#). In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 35–47, Sofia, Bulgaria. Association for Computational Linguistics.
- Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 2022. A morph-based and a word-based treebank for beja. In *TLT 2021-20th International Workshop on Treebanks and Linguistic Theories. 21-25 March 2021, Sofia, Bulgaria*, pages 48–60.
- Peter Koch and Wulf Oesterreicher. 2012. Language of immediacy-language of distance: Orality and literacy from the perspective of language theory and linguistic history.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.
- Per Linell. 2019. [The Written Language Bias \(WLB\) in linguistics 40 years after](#). *Language Sciences*, 76. Publisher: Elsevier Ltd.
- Zoey Liu and Emily Prud’hommeaux. 2023. [Data-driven parsing evaluation for child-parent interactions](#). *Transactions of the Association for Computational Linguistics*, 11:1734–1753.
- John Mansfield and Danielle Barth. 2021. [Clause chaining and the utterance phrase: Syntax–prosody mapping in matukar panau](#). *Open Linguistics*, 7(1):423–447.
- Francesca Masini, Claudia Roberta Combei, and Roberta Cicchirillo. 2025. [The prosody of list constructions](#). In Kiki Nikiforidou and Mirjam Fried, editors, *Multimodal Communication from a Construction Grammar Perspective*, Constructional Approaches to Language, pages 116–151. John Benjamins Publishing Company.
- Francesca Masini, Caterina Mauri, Paola Pietrandrea, and 1 others. 2018. List constructions: Towards a unified account. *Italian Journal of Linguistics*, 30(1):49–94.
- Caterina Mauri, Silvia Ballarè, Eugenio Gorla, Massimo Cerruti, and Francesco Suriano. 2019a. Kiparla corpus: A new resource for spoken Italian. In *Proceedings of the 6th Italian Conference on Computational Linguistics CLiC-it*.
- Caterina Mauri, Eugenio Gorla, and Iliaria Fiorentini. 2019b. Non-exhaustive lists in spoken language: A construction grammatical perspective. *Constructions and frames*, 11(2):290–316.
- Max Planck Institute for Psycholinguistics, The Language Archive. 2024. ELAN (version 6.9) [computer software]. <https://archive.mpi.nl/tla/elan>. Retrieved from <https://archive.mpi.nl/tla/elan>.
- Igor Aleksandrović Mel’čuk. 1988. *Dependency Syntax: Theory and Practice*. SUNY Press.
- Benjamin Minixhofer, Jonas Pfeiffer, and Ivan Vulić. 2023. [Where’s the point? self-supervised multilingual punctuation-agnostic sentence segmentation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7215–7235, Toronto, Canada. Association for Computational Linguistics.
- Marianne Mithun. 1988. The grammaticization of coordination. In John Haiman and Sandra Thompson, editors, *Clause combining in grammar and discourse*, pages 331–60. Benjamins, Amsterdam; Philadelphia.
- Giovanni Nencioni. 1976. Parlato-parlato, parlato-scritto, parlato-recitato.
- Joakim Nivre, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, Yoav Goldberg, Jan Hajič, Jenna Kanerva, Veronika Laippala, Alessandro Lenci, Teresa Lynn, Christopher Manning, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Slav Petrov, and 7 others. 2015. [Universal dependencies 1.0](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Teresa Paccosi, Alessio Palmero Aprosio, and Sara Tonelli. 2023. [Adding a Novel Italian Treebank of Marked Constructions to Universal Dependencies](#). *IJCoL. Italian Journal of Computational Linguistics*, 9(1). Number: 1 Publisher: Accademia University Press.
- Paola Pietrandrea, Sylvain Kahane, Anne Lacheret-Dujour, and Frédéric Sabio. 2014. The notion of sentence and other discourse units in corpus annotation. *Spoken corpora and linguistic studies*, pages 331–364.

- Manuela Sanguinetti and Cristina Bosco. 2014. Converting the parallel treebank partut in universal stanford dependencies. 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*, pages 316–321. Pisa University Press.
- Manuela Sanguinetti and Cristina Bosco. 2015. Partut: The turin university parallel treebank. *Harmonization and development of resources and tools for italian natural language processing within the parli project*, pages 51–69.
- Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamel Seddah, and Amir Zeldes. 2023. [Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations](#). *Language Resources and Evaluation*, 57(2):493–544.
- Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018. Postwita-ud: an italian twitter treebank in universal dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- 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. 2024. [Universal dependencies 2.15 models for UDPipe 2 \(2024-11-21\)](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Lucien Tesnière. 2015. *Elements of Structural Syntax*. John Benjamins Publishing Company, Amsterdam.



## A Examples

BO147 (285)		id:93 [eh]
BO145 (320)		id:92 [mh]
BO139 (417)	id:91 volete stare in bisca fino alle quattro [del ma]tt[ino stano]tte?	

91, BO139 volete stare in bisca fino alle quattro [del ma]tt[ino stano]tte?  
 want:2PL stay in casino until to.DEF four of.DEF morning tonight  
 ‘do you want to stay up until four in the morning tonight?’

92, BO145 [mh]  
 mh  
 ‘mh’

93, BO147 [eh]  
 eh  
 ‘eh’

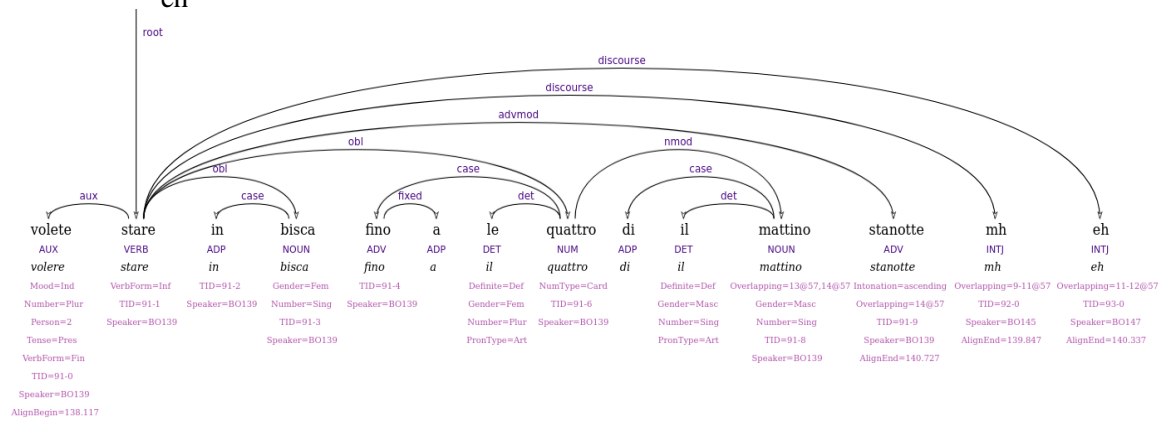


Figure 5: A case where a token (i.e., ‘mattino’, en. ‘morning’) is participating in two distinct overlapping spans. Its Overlapping feature is in fact composed by two distinct references, separated by a comma. (from conversation BOA3017)

BO147 (285)	id:129 [>in<fatti te po]tresti fare il sotto~ quello che fa i sotto[titoli]	
BO145 (326)		id:130 [sottotitolato]re
BO139 (437)	[sta]	
BO146 (533)		id:131 [pure te]

129, BO147 [>in<fatti te po]tresti fare il sotto~ quello che fa i sotto[titoli]  
indeed you.SG could:2SG do DEF sub~ the.one that does DEF subtitles  
'indeed you could be a subtit~ the person that makes subtitles'

130, BO145 [sottotitolato]re  
subtitler  
'subtitler'

131, BO146 [pure te]  
also you.SG  
'you too'

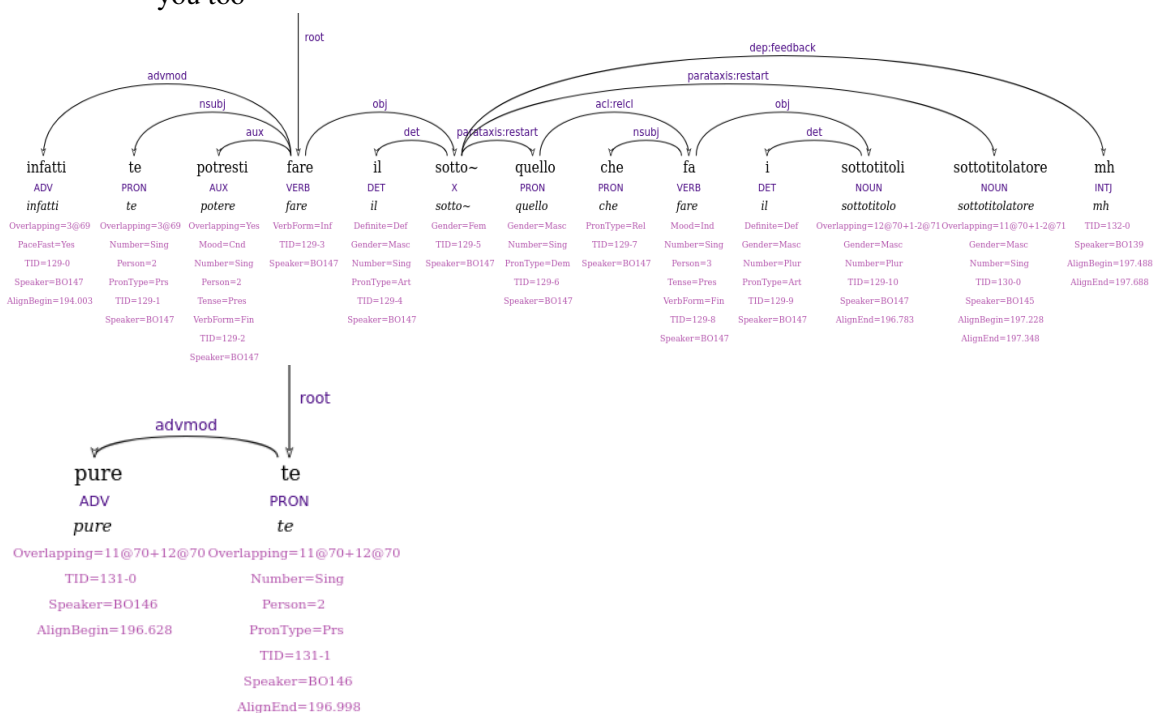


Figure 6: A case of overlap among multiple (i.e., more than two) speakers. The token 'sottotitoli' (en. 'subtitles'), with TID=129-10, overlaps with token 'sottotitolatore' (en. 'subtitler', TID=130-0 in the first sentence) and 'pure te' (en. 'you too', in the second one). This is expressed by means of the + symbol in the Overlap feature. (from conversation BOA3017)

38, BO147 (xx ma io da pallotti ci piglio le paste >cioè:< ci prend[o: =mh le  
 UNK but I by Pallotti LOC take:PRS.1SG DEF pastries that.is LOC take:PRS.1SG mh DEF  
 brio~])  
 crois(sants)  
 ‘but from Pallotti I buy pastries I mean I buy croissants’

39, BO146 [le lasagne]  
 DEF lasagne  
 ‘lasagna’

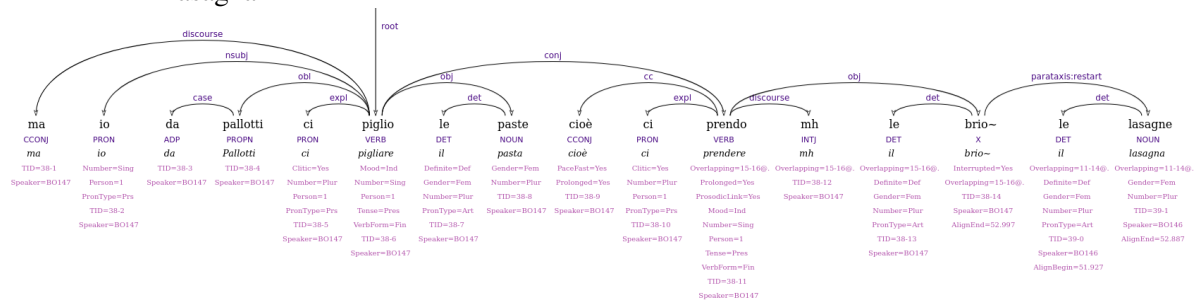


Figure 7: A case of co-construction, where the second speaker provides material (i.e., ‘le lasagne’) that is syntactically dependent on the verb uttered by the first speaker (i.e., ‘ci prendo’). (from conversation BOA3017)

4, BO140 allora la mia casa:: è:: una: villa::  
 so DEF my house is INDEF villa  
 ‘so my house is a villa’

5, BO140 mh: in mezzo alla natura,  
 mh in middle of.DEF nature  
 ‘mh, immersed in nature’

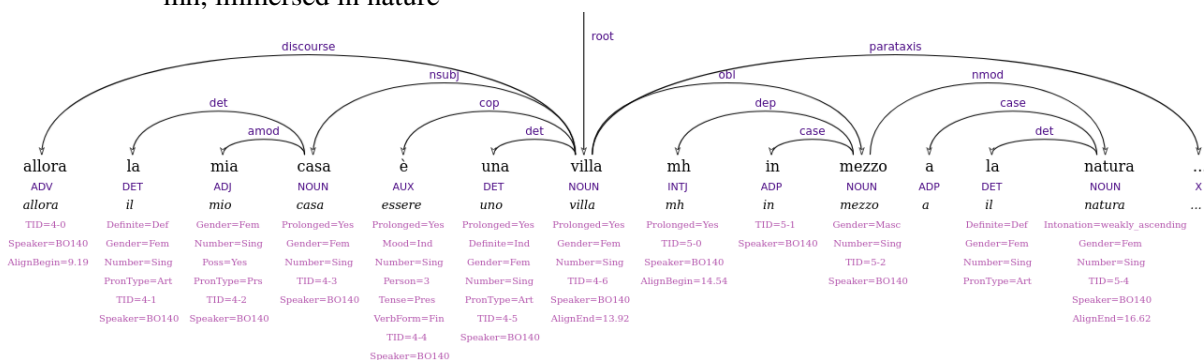


Figure 8: A case where, during transcription, a TU boundary was introduced breaking an intra-phrase relation: the nominal modifier ‘in mezzo alla natura’ would be separated from its head ‘villa’ if segmentation was performed based on TU boundaries. For space reasons, only the first part of the full unit is shown. (from conversation BOD2018)

- 342, BO139 ho scoperto chi è perché era  
 have.PRS.1SG found.out who is why was  
 ‘I found out who that was because they was’
- 343, BO139 un un  
 INDEF INDEF  
 ‘a a’
- 344, BO147 {tossisce}  
 coughs\*  
 ‘{coughs}’
- 345, BO139 x un’ in[tervista in cui x=ava]  
 UNK INDEF interview in which  
 ‘in an interview where’
- 346, BO145 [(ti) strozzi]  
 you.OBJ choke:2SG  
 ‘you’re choking’
- 347, BO139 salute  
 bless.you  
 ‘bless you’
- 348, BO147 {tossisce}  
 coughs\*  
 ‘{coughs}’
- 349, BO139 era tipo (un) ri[cevimento]  
 it.was like INDEF meeting  
 ‘it was like a meeting’
- 350, BO147 [trascrivi] {ride} {tossisce}  
 transcribe laughs<sup>-</sup> coughs\*  
 ‘transcribe {laughter} {coughs}’
- 351, BO139 {ride}  
 laughs<sup>-</sup>  
 ‘{laughter}’
- 352, BO145 cough cough  
 cough cough  
 ‘cough cough’
- 353, BO147 cou[gh cough] {ride}  
 cough cough laughs<sup>-</sup>  
 ‘cough cough {laughter}’
- 354, BO139 [tossisce] {ride}  
 coughs laughs<sup>-</sup>  
 ‘coughs’
- 355, BO139 chiusa parentesi  
 closed parenthesis  
 ‘parenthesis closed’

Figure 9: Curly brackets mark non-linguistic behavior in Jefferson notation. In the glosses, elements marked with \* indicate NLBs that are annotated with feature Type=NLB in the treebank, while elements marked by – have not been ported as tokens to the treebank. (from conversation BOA3017)

435, BO147 che cosa vuoi da me? {cantando} {ride}  
 what thing want.2SG from me singing laughs  
 ‘what do you want from me’

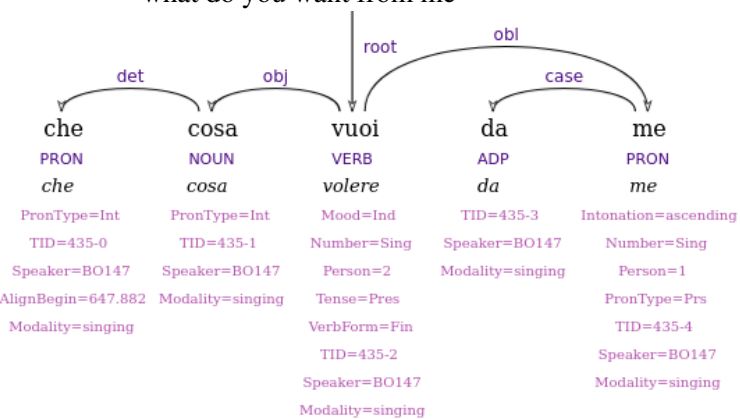


Figure 10: A case where a token originally present in the resource (i.e., {cantando}, en. ‘singing’, with TID=435-5) has been transformed into the `Modality=singing` feature on the tokens that were uttered while singing. Consequently, the token is not included in the treebank as a syntactic token. The example also shows a case where an NLB token (i.e., {ride}, en. ‘laughs’) is removed as not syntactically relevant. (from conversation BOA3017)

- 925, BO139 la prima c'e~ {ride}  
 DEF first there.wa~ laughs  
 'the first one there wa~ {laughter}'
- 926, BO139 una ragazza che: raccontava di ex coinquiline cristiane che tipo  
 INDEF girl that talked of ex roommates christian who like  
 'a girl that was talking about ex christian roommates that like'
- 927, BO139 le avevano rubato un pr~ un =eh non mh un dildo che lei aveva nel =eh  
 to.her had:3PL stolen INDEF pr~ INDEF eh not mh INDEF dildo that 3SG.F had in.DEF eh  
 {ride}  
 laughs  
 'they stole her a co~ mh a dildo that she had in {laughter}'
- 928, BO147 {ride}  
 laughs  
 '{laughter}'
- 929, BO139 nel comodino  
 in.INDEF bedside.table  
 'in her bedside table'
- 930, BO139 perché col sospetto che lei lei stava studiando delle cose di magia nel  
 because with.DEF suspect that 3SG.F 3SG.F was studying INDED.PL things of magic in.DEF  
 sen[so >ciòè< {P} di antropologia]  
 sense I.mean PAUSE of anthropology  
 'because suspecting she was studying something about magic, I mean, anthropology'
- 931, BO145 [a:h me l' hai raccontato]  
 ah to.me it have.2SG told  
 'oh you told me'
- 932, BO139 eh  
 eh  
 'eh'
- 933, BO139 uno studio sui riti magici nella sicilia tipo dell' ottoce~ metti una roba  
 INDEF study on.tDEF rituals magic in.DEF sicily like of.DEF eight.hundr~ take:2SG a thing  
 del genere  
 of.DEF genre  
 'a study on magic rites in the Sicily of ninet~ century, for example, something like that'

Figure 11: The example shows how syntax develops incrementally and not necessarily planned in advance by the speaker. The syntactic cohesion of the discourse portion is also marked by the final 'una roba del genere' (en. 'something like that') that can be identified as the closing element of a listing. Figure 13a shows the parsed tree. (from conversation BOA3017)



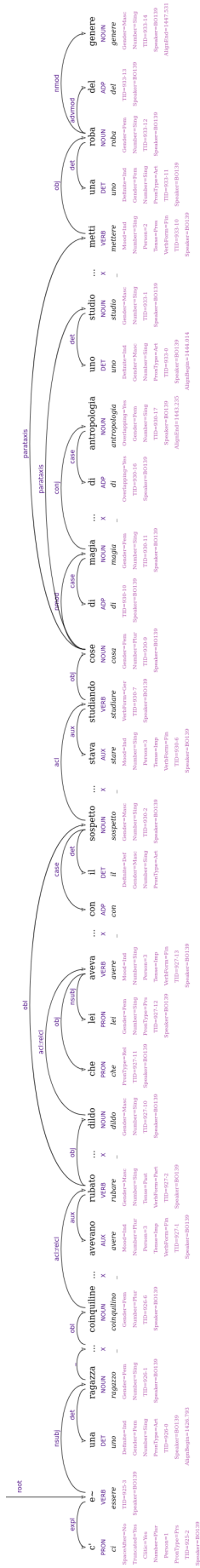
BO140 <small>[262]</small>	id:136 dai zambo:ni:, [ca]stiglio:ne, san vitale, insomma cioè quello è il centro: piazza maggiore così.
BO118 <small>[190]</small>	id:137 [si]

136, BO140 dai zambo:ni:, [ca]stiglio:ne, san vitale, insomma cioè quello è il centro: piazza  
 come.on Zamboni Castiglione san Vitale to.sum.up I.mean that is the center square  
 maggiore così.  
 major so

‘come on, Zamboni, Castiglione, San Vitale I mean. That is DEF city centre, Piazza Maggiore,  
 like that’

137, BO118 sì  
 yes  
 ‘yes’

Figure 12: A case where feedback from speaker BO118 (‘sì’, en. ‘yes’) does not interrupt the syntactic construction of speaker BO140. Figure 13b shows the parsed tree. (from conversation BOD2018)



(a) Parsing tree for Example in Figure 11.



(b) Parsing tree for Example in Figure 12.

Figure 13

# Head-Initial and Head-Final Coordinate Structures in Two Annotation Schemes of Dependency Grammar

Timothy John Osborne  
Zhejiang University  
tjo3ya@yahoo.com

Chenchen Song  
Zhejiang University  
cjs021@zju.edu.cn

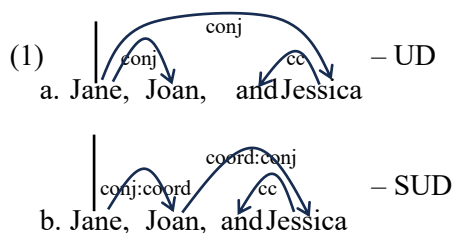
## Abstract

The *Universal Dependencies* (UD) and *Surface-Syntactic Universal Dependencies* (SUD) annotation schemes view coordinate structures as head-initial. This contribution argues that a more flexible approach to coordinate structures is linguistically motivated, one that sees coordinate structures as head-initial in greater head-initial structures and as head-final in greater head-final structures. Support for this flexible approach comes from two areas: dependency distance and a nearness effect. In addition, two arguments that have been produced supporting the strictly head-initial approach are examined and refuted.

## 1 Introduction

The *Universal Dependencies* (UD: de Marneffe et al. 2014; Nivre et al. 2019) and *Surface-Syntactic Universal Dependencies* (SUD: Gerdes et al., 2018, 2019; Kahane et al. 2021) annotation schemes agree to an extent in their annotation choices concerning coordination.<sup>1</sup> They both view coordinate structures as head-initial, the initial (i.e. leftmost) conjunct being head over the following conjunct(s). In doing so, they are following other DGs (e.g. Engel, 1982; Mel'čuk, 1988; Groß, 1999; Eroms, 2000). The two annotation schemes also disagree in an important way, however, concerning the hierarchical status of non-initial conjuncts. The next two trees serve to illustrate major points of agreement and disagreement (conj =

conjunction, coord = coordinate, cc = coordinate conjunction):



UD annotation assumes a *bouquet* structure, whereby the non-initial conjuncts are equi-level dependents of the initial conjunct. SUD annotation, in contrast, chooses a more right-branching structure such that each successive conjunct is an immediate dependent of the immediately preceding conjunct. The two schemes agree insofar as coordinate structures are head-initial, the initial conjunct being head over the following conjuncts. They disagree, however, in all cases where the coordinate structure at hand contains three or more conjuncts, UD choosing the flatter bouquet structure, and SUD the more layered one.

The intent of this contribution is to critique the strictly head-initial approach to coordinate structures that both annotation schemes espouse. In doing so, the message delivered is similar to the messages of other recent accounts of the dependency analyses of coordinate structures (Kanayama et al. 2018; Przepiórkowski and Wóznik 2023; Przepiórkowski et al. 2024a, Przepiórkowski et al. 2024b; Stempniak 2024). More specifically, the account here pursues the approach to coordination

<sup>1</sup> The claims about the UD and SUD accounts of coordinate structures are based mainly on the guidance provided in the UD website (<https://universaldependencies.org/u/overview/complex-syntax.html>) and the SUD website ([https://surfacesyntacticud.github.io/guidelines/u/oral\\_language/conj\\_coord/](https://surfacesyntacticud.github.io/guidelines/u/oral_language/conj_coord/)). Note that the UD claims in the area seem contradictory. The UD website states that

“coordinate structures are in principle symmetrical, but the first clause is by convention treated as the parent (or “technical head”) of all subsequent coordinated clauses via the conj relation.”

This statement is then followed by examples in which the standard dependency arcs are provided, with the initial conjunct shown as dominating the non-initial conjuncts.

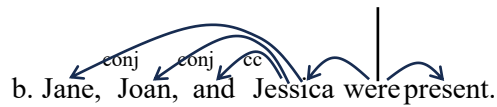
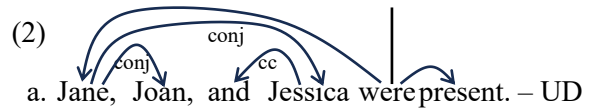
developed in Osborne and Groß (2017); this approach sees coordinate structures as head-initial or head-final within one and the same language depending on the greater structure in which the given coordinate structure appears. Two types of support are presented in favor of this flexible approach, the one being based on dependency distance and the other on a nearness effect having to do with mismatches in form. The account also examines and refutes two arguments supporting the strictly head-initial approach to coordinate structures.

There are two important points about the UD and SUD annotation choices concerning coordination and the proposal here that must be stated and acknowledged before proceeding. The first has to do with the tendency among the authors of UD and SUD to emphasize that their schemes are *not* intended to be theoretically stringent, linguistically unimpeachable analyses of sentence structures. They emphasize that the necessity to create easily implementable annotation guidelines has forced difficult decisions in the interest of practicality. Given this concession, it can be emphasized here from the outset that the proposal presented and defended below is easily implementable, for there is nothing complex or difficult about it, nothing that would prevent it from being adopted as a simple improvement to existing annotation guidelines.

The second point concerns the challenge posed by various phenomena of coordination, e.g. *gapping*, *right node raising* (RNR), *non-constituent conjuncts*, etc. The discussion of coordination presented here does not attempt to present coherent accounts of these phenomena, since doing so would require much more space than is available. The discussion concentrates instead on the core issue, which is the head-initial vs. head-final accounts of coordinate structures.

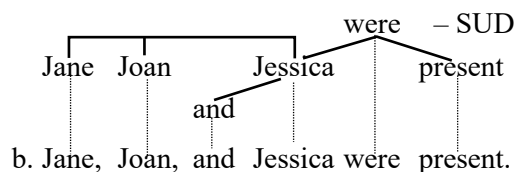
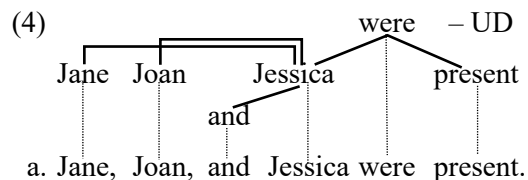
## 2 The proposal and the convention

The core proposal presented and defended here is now illustrated with the sentence *Jane, Joan, and Jessica were present*. The current UD and SUD annotation choices for this sentence are given next as the a-trees, and the alternative trees of the current proposal are given as the b-trees.



In all those cases where the coordinate structure precedes its head (*were* here), the coordinate structure is in fact head-final instead of head initial. In all those cases where the coordinate structure follows its head, the coordinate structure continues to be head-initial as shown in examples (1a-b).

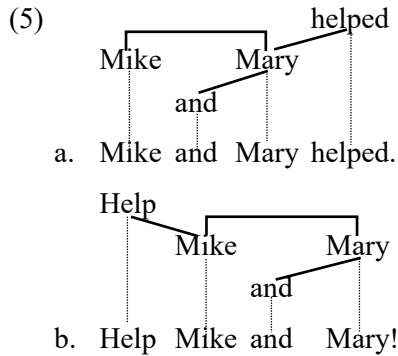
A special graphic convention for rendering coordinate structures shall henceforth be employed here: completely horizontal edges. These edges require little space and capture well the nature of the conjuncts of coordinate structures. The dependency structures are henceforth rendered as follows:



The benefit of this tree convention is that the conjuncts now clearly appear on the same level of the structure. The syntactic functions on the dependency edges (e.g. CONJ or COORD) are no longer necessary to help indicate the presence of coordination. Observe as well that the edges in (4a-b) are all still directed due to the fact that the tree is rooted. From a graph-theoretic standpoint, *Jessica* dominates *Joan* and *Jane* (and *and*) in (4a) because *Jessica* is linearly closer to the root *were* than *Joan* and *Jane* (and *and*). Similarly, *Jessica* dominates *Joan* in (4b) because *Joan* is linearly closer to the root *were* than *Joan*, and *Joan* dominates *Jane* because *Joan* is linearly closer to the

root *were* than *Jane*. Trees (4a) and (4b) are therefore isomorphic to trees (2b) and (3b), respectively.

Given this new convention, the proposal here is that coordinate structures that precede a shared head are head-final and coordinate structures that follow a shared head are head-initial. The next trees illustrate the proposal with respect to the coordinate structure *Mike and Mary*.



There are two benefits to these annotation choices. Both are due to the fact that the coordinate structure is now linked into the greater sentence at the closest point. This reduces dependency distances and accommodates the aforementioned nearness effect.

Note next that the proposal here is, as stated above, in line with the recent accounts of coordination that present corpus-based reasoning against inflexible “asymmetric” approaches to coordinate structures, the current basic UD and SUD annotation schemes being such inflexible approaches (i.e. Kanayama et al. 2018; Przepiórkowski and Wóznia 2023; Przepiórkowski et al. 2024a, Przepiórkowski et al. 2024b; Stempniak 2024). The proposal here is that coordinate structures are in

fact asymmetric, but asymmetric in the special way suggested by Przepiórkowski and Wóznia (2023: 15501): coordinate structures can be head-initial or head-final within one and the same language depending on the position of the shared governor with respect to the conjuncts of the coordinate structure.

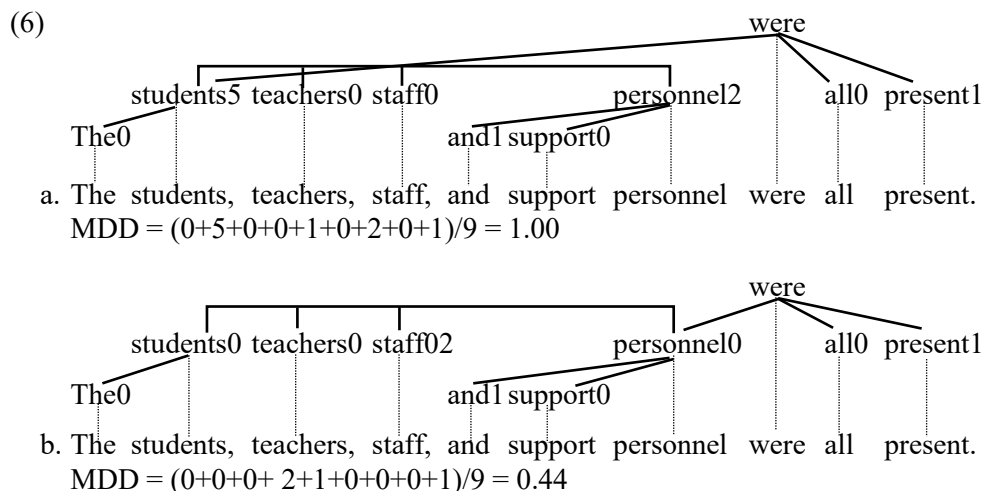
Finally, just the trees analogous to SUD annotation are produced henceforth in order to save space. All conclusions reached apply equally to UD and SUD annotations, though.

### 3 Two arguments

The following two sections present the two arguments just mentioned in favor of the flexible approach to coordination.

#### 3.1 Dependency distance

The first source of motivation for the proposal comes from dependency distance (cf. Hudson, 1995; Temperley, 2007; Liu, 2008; Liu, et al. 2017; Wang and Liu, 2017). Attaching the shared head of a coordinate structure to the closest conjunct can significantly reduce dependency distances. This is particularly true of head-final structures where the shared head follows the coordinate structure, as is frequently the case in head-final languages (cf. Kanayama et al. 2018: 81; Stempniak 2024). This occurs in English, for instance, with coordinated subject phrases. The *mean dependency distance* (MDD) of the next sentence is significantly reduced when the shared head reaches just to the closest conjunct of the coordinate structure:

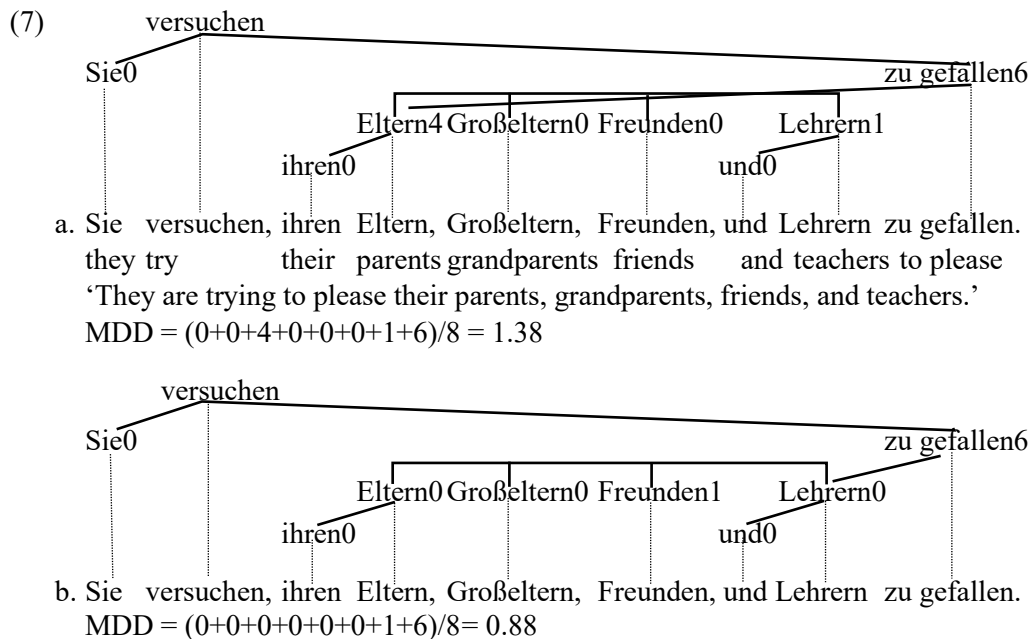


The number immediately after each node gives the dependency distance measured in terms of intervening words from that word to its head.

The dependency in (6a) reaching from *students* to *were* is long, pushing the mean score for the entire sentence up to 1.00, a score high enough

that one might expect the sentence to be difficult to process. In contrast, linking *personnel* directly to *were* cuts the MDD score in more than half; this much lower score matches the ease with which the sentence is processed.

A second example can further illustrate the benefit of attaching a shared head to the closest conjunct of the coordinate structure. The next sentence from German contains a post-dependent *zu*-infinitive phrase that has a four-conjunct coordinate structure preceding the *zu*-infinitive:



The MDD score of 1.38 is quite high, so high that one might expect processing difficulty with such a sentence. In contrast, linking the shared head *zu gefallen* ‘to please’ to the closest conjunct significantly reduces the MDD score, down to 0.88. The lower score is more consistent with the relative ease with which the sentence is processed.

The lower MDD values of the current proposal align well with the widespread and well-documented tendency for shorter conjuncts to precede longer conjuncts in English, English in general having more head-initial than head-final structures. Crucially in this area, none of the four established annotation schemes cited in [Przepiórkowski and Wóznia \(2023\)](#) – not the “symmetric” nor the “asymmetric” ones – can achieve the overall mean dependency distance scores that are as low as those of the current asymmetric proposal, because none of those annotation schemes is flexible in the manner of the current proposal, allowing for both head-initial and head-final coordinate structures within one and the same language.

Given that coordinate structures occur frequently in most languages, the current proposal can have a significant impact on overall dependency distance values. This is particularly

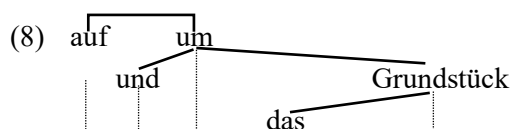
true of those languages that have many head-final structures. Consider in this regard that approximately 45% of the world’s languages are deemed SOV, which means they are more head-final than head-initial.

### 3.2 Nearness effect

The second source of support for the current proposal comes from mismatching forms that occur with coordinate structures. Material that appears outside of a given coordinate structure is shared by the conjuncts of the coordinate structure. This shared material tends to be congruent in form with the closest conjunct. The concord can be much weaker or non-existent with the conjunct(s) that are further removed. The mismatches that occur in this area involve number, gender, case, and definiteness, as well as subcategorization requirements more broadly.

The first example is from German and involves case. It is taken from Müller (1990: 253):

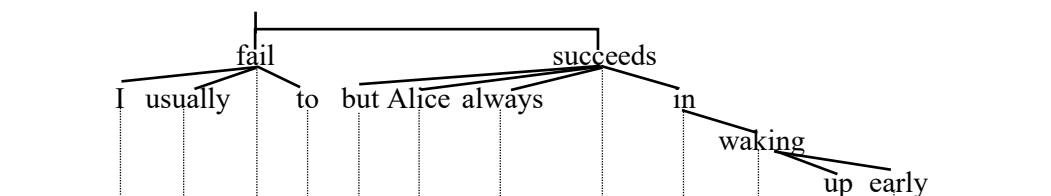


- (8) 
- a. auf und um das Grundstück  
on and around the.ACC property
- b. \*auf und um dem Grundstück  
on and around the.DAT property

The preposition *auf* ‘on’ requires its complement to appear in the dative case, whereas the preposition *um* ‘around’ demands a complement in accusative case. Thus, the fact that *das Grundstück* ‘the property’ is accusative marked sees case concord occurring with the closest preposition only. The ungrammaticality of (8b) demonstrates that case concord cannot occur

with the preposition that is further removed. In other words, there is a clear nearness effect concerning case concord.

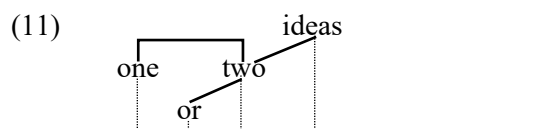
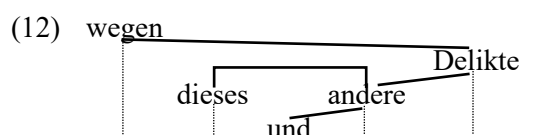
The next two example pairs are instances of so-called *right node raising* (RNR) – examples (9a-b) are taken from Belk et al. (2023: 690). The numbers to the right document informant responses concerning the grammatical acceptability of the sentences. All four sentences were tested in the crowdsourcing service Amazon Mechanical Turk. The number before the slash is the number of informants that judged the sentence, and the number after the slash is the average score the sentence received on a four-point scale from 1 (perfectly grammatical) to 4 (quite ungrammatical):<sup>3</sup>

- (9) 
- a. ?I usually fail to, but Alice always succeeds in, waking up early. 18/2.22
- b. \*I usually fail to, but Alice always succeeds in, wake up early. 18/3.38
- (10) a. ??Henry is going to, and Alice will soon be, working with every student. 18/2.44
- b. \*Henry is going to, and Alice will soon be, work with every student. 18/3.56

The particle *to* in (9) subcategorizes for a bare infinitive (*wake*), whereas the preposition *in* subcategorizes for a nominal form (the gerund *waking*). Similarly, the particle *to* in (10) subcategorizes for a bare infinitive (*work*) whereas the auxiliary *be* subcategorizes for a present participle (*working*). The scores for the a-sentences reveal that they are marginally possible, whereas the scores for the b-sentences demonstrate that they are clearly ungrammatical. There is hence again a nearness effect concerning the preferred form of the shared material.

The examples (8-10) involve a coordinate structure that precedes a shared dependent. In this regard, the nearness effect that they illustrate is not inconsistent with existing UD and SUD annotation choices, since the coordinate structures assumed are head-final in all cases under consideration. The next examples, in contrast, involve a coordinate structure that precedes a shared head. It is precisely in this area that the current proposal differs from UD and SUD annotation choices. The first examples are

from English and German; they involve number concord. The German example is from Müller (1990: 248):

- (11) 
- a. one or two ideas
- b. \*one or two idea
- (12) 
- a. wegen dieses und andere Delikte  
due.to this.SG and other.PL crimes.PL
- b. \*wegen dieses und andere Delikt  
due.to this.SG and other.PL crime.SG

<sup>3</sup> Note that the coordinate structure in (9) is deemed to be head-initial. This is the default assumption when the roots of the conjuncts are the roots of the tree.

A further similar example from German is taken from [Lobin \(1993: 226\)](#):

- (13) die vielen dicken und der  
 the.PL many fat and the.SG  
 eine dünne Mann/\*Männer  
 one thin man men

Note the order of the singular and plural conjuncts in this case, the plural conjunct preceding the singular one. Note as well that the dependency tree is not provided; it is not because of the difficulties analyzing such examples, the three words in each conjunct being equi-level dependents of the noun. The nearness effect is quite pronounced in all of these cases; the shared head agrees in number with the closest conjunct only.

The next eight examples sentences are from English and they again involve number concord. They were tested in Mechanical Turk using the same procedure as for examples (9-10) above. The head-final coordinate structure is presented first, followed by the corresponding head-initial one:

- (14)
- 
- a. ?Lee or Lee's kids were present. 18/1.72  
 b. \*Lee or Lee's kids was present. 18/3.50  
 c. ?Lee's kids or Lee were present. 18/1.94  
 d. ??Lee's kids or Lee was present. 18/2.83

- (15) Were
- 
- a. Were Lee or Lee's kids present? 18/1.39  
 b. ??Was Lee or Lee's kids present? 18/2.50  
 c. Were Lee's kids or Lee present? 18/1.38  
 d. \*Was Lee's kids or Lee present? 18/3.61

On the whole, there is strong preference for plural agreement when at least one of the conjuncts is plural. Nearness is, though, also quite apparently a factor influencing the clarity of the judgment. The sentence at hand is most felicitous when the verb is plural and the closest conjunct is plural, and it is most ungrammatical when the

<sup>4</sup> The dependency trees are not provided in these cases due to the lack of space on the page. They would not fit.

verb is singular and the closest conjunct is plural.

The next examples are from Hungarian; they are taken from [Kiss \(2012: 1052\)](#) and have been adapted slightly for ease of presentation. They involve verb-object agreement in terms of definiteness/indefiniteness:<sup>4</sup>

- (16) a. Melyik professzort és hány diákot  
 Which professor and how.many students  
 ültessünk / \*ültessük  
 make.sit-INDEF.1PL / \*make.sit-DEF.1PL  
 le egymással szemben?  
 down each-other opposite  
 'Which prof and how many students shall we make sit down opposite each other?'
- b. Hány diákot és melyik professzort  
 how.many students and which professor  
 \*ültessünk / ültessük  
 \*make.sit-INDEF.1PL / make.sit-DEF.1PL  
 le egymással szemben?  
 down each-other opposite  
 'How many students and which prof shall we make sit down opposite each other?'

The object conjunct *hány diákot* 'how many students' is indefinite, whereas *melyik professzort* 'which professor' is definite. The verb agrees with the closest conjunct each time and cannot agree with the conjunct further removed.

The final example considered in this section is from Japanese and is taken from [Vermeulen \(2006: 417\)](#). The particle *-to* 'and' in Japanese cliticizes to a preceding noun. When it does so, it appears between conjoined nouns:<sup>5</sup>

- (17)
- 
- John-ga Mary-to Bill-o mita.  
 John-NOM Mary-and Bill-ACC saw  
 'John saw Mary and Bill.'

The aspect of this example that illustrates the nearness effect is the presence of the accusative case marker *-o* on the final conjunct. This case marker does not appear on the initial conjunct (cf. [Kanayama et al. 2018: 77](#)). Thus there is an asymmetry among the conjuncts and the current account accommodates this asymmetry insofar as the case-marked noun can receive case

<sup>5</sup> Vermeulen's account is in terms of phrase structures. The dependency analysis is my addition.

directly from the governing verb. On the UD and SUD accounts, in contrast, the initial noun would have to serve as an intermediary, passing the case marker down through the hierarchy to the following noun. Note further that coordinate structures in Japanese must be head-final as shown here because (almost) all dependencies are head-final in Japanese to begin with. That Japanese is a strictly head-final language is a widely acknowledged fact.

## 4 Two counterarguments

The next two sections consider and refute two counterarguments against the proposal here.

### 4.1 Irreversible conjuncts

Mel'čuk (1988: 26-28) argues that coordinate structures must be head-initial because there is often a logical relationship between the conjuncts such that their order is fixed; they are not reversible. Mel'čuk illustrates the point with the next pairs:

- (18) a. He stood up and gave me the letter.  
 b. He gave me the letter and stood up.
- (19) a. not only a good worker but also a nice man  
 b. not only a nice man but also a good worker

These a- and b-examples do not mean quite the same thing, of course. There is a chronological relationship between the conjuncts of (18a-b) such that reversing their order changes the meaning (cf. Kanayama et al. 2018: 78). Similarly, reversing the order of the conjuncts across (19a-b) shifts the pragmatic focus. Mel'čuk therefore concludes that the dependency hierarchy must be sensitive to meaning in this area; the hierarchy is fixed, with the initial conjunct being head over the non-initial conjuncts.

While Mel'čuk's point is of course correct regarding the meaning difference that often result from reversing the order of the conjuncts, this fact should not be construed as an indication about the dependency relationship between the conjuncts. Mel'čuk's reasoning is influenced by his multi-stratal approach to syntax. All aspects of meaning are to be captured in the two *Meaning to Text* (MTT) levels of syntax (deep and surface), both of which lack linear order. Thus his system cannot appeal to linear

order alone to capture the differences in meaning associated with conjunct order; these differences must be located in the hierarchy of structure instead.

The approach espoused here is monostratal in syntax; there is just one level of syntax, surface syntax, where linear order is indeed present. Given just surface syntax, syntactic units are organized in both dimensions simultaneously, in the hierarchical dimension and in the linear dimension. When syntactic unit U1 precedes a syntactic unit U2, U1 is more prominent than U2 in the linear dimension. This prominence is sufficient to capture the logical and pragmatic relationships between the conjuncts that Mel'čuk observes. There is hence no reason to also put these relationships into the hierarchical dimension. Observe further that the manner in which the conjuncts are organized in the linear dimension helps account for the fact that the coordinator usually introduces just the final conjunct of a coordinate structure.

### 4.2 Omission of final conjunct

UD, SUD, and the current DG agree concerning the status of coordinators such as *and*, *or*, *but*, etc. in English and related languages: these coordinators belong to the following conjunct rather than to the preceding one. This point is illustrated in many of the examples above insofar as the coordinator each time is shown as a dependent of the following conjunct root rather than of the preceding one. The conjuncts of coordinate structures such as *onions and rice* can therefore be understood in the following manner: [*onions*] [*and rice*]. This understanding of the conjuncts suggests that omission can be a diagnostic for discerning the hierarchical relationship between the conjuncts. Gerdes and Kahane (2015) construe the fact that the final conjunct can, but the initial cannot, be omitted at times without affecting grammaticality as an argument supporting the head-initial approach to coordinate structures.

The next examples illustrate the extent to which the one or the other conjunct of a coordinate structure can be omitted:

- (20) A: What did you eat?  
 B: a. I ate onions and rice.  
 b. \*I ate onions and.  
 c. \*I ate and rice.  
 d. I ate onions  
 e. I ate rice.

The fact that one can omit *and rice* yielding the answer *I ate onions* in (20d) but one cannot omit *onions* yielding the answer *\*I ate and rice* in (20c) suggests that of the two conjuncts [*onions*] and [*and rice*], the latter is dependent on the former. In other words, the coordinate structure is head-initial. Note that the grammaticality of the answer *I ate rice* in (20e) should not be construed as suggesting that [*onions and*] is a conjunct because we already know that *and* forms a conjunct with *rice* rather than with *onions*, as established in the previous paragraph.

There are a number of problems with this argument in favor of head-initial coordinate structures. One difficulty is the general fact that omission as employed in (20) is an imperfect test for identifying sentence structure because contrary to expectation, certain heads are known to be omissible, e.g. the subordinator *that* in the sentence *Sam said (that) he would do it* and the preposition *on* in the sentence *Sam departs (on) Tuesday*. A second problem is that with certain coordinate structures, neither conjunct can be omitted, e.g. *Jack and Jane were present* vs. *\*Jack were present* and *\*And Jane were present*. A third problem is related to the second. Neither conjunct can be omitted when correlative coordinators are involved, as illustrated with the next examples from French:

- (21) a. Nous n'avons vu ni Jean ni Marie.  
'We saw neither Jean nor Marie.'
- b. \*Nous n'avons vu ni Jean.  
'We saw neither Jean.'
- c. \*Nous n'avons vu ni Marie.  
'We saw neither Marie.'

A fourth problem is that in languages in which the coordinator is a clitic that attaches to the initial conjunct, the non-initial conjunct clearly cannot be omitted, as with the Japanese sentence from above, example (17), reproduced here as (22):

- (22) a. John-ga Mary-to Bill-o mita.  
John-NOM Mary-and Bill-ACC saw  
'John saw Mary and Bill'.
- b. John-ga Bill-o mita.
- c. \*John-ga Mary-to mita.

If the behavior of the coordinator in head-final languages such as Japanese were a clue about the hierarchical status of the conjuncts with respect to each other, then one has to assume that coordinate structures in head-final languages

are in fact all head-final. The UD and SUD annotation schemes do not do this (but cf. Kahane et al. 2021).

Taken together, the four arguments just enumerated seriously weaken the strength of omission as an argument in favor of the stance that all coordinate structures are head-initial in English and related languages. A more plausible reason why the sentence *\*I ate and rice* is ungrammatical is that the appearance of a coordinator is only possible if a coordinate structure is present, hence for the coordinator *and* to appear, at least two conjuncts must be discernible. Apparent exceptions to this requirement, e.g. *And I ate onions*, are not really exceptions because in such cases, the conjuncts are complete sentences in discourse. Or in certain cases, the element at hand (e.g. *and*, *or*, *but*) is actually best construed as an adverb or subordinator rather than as a coordinator.

## 5 Conclusion

This contribution has drawn attention to an aspect of two prominent annotation schemes in the area of coordination. It has argued that a flexible account of coordination is preferable to the currently prevailing rigid approach. Instead of viewing all coordinate structures as head-initial, a linguistically more plausible approach allows flexibility of structure. A coordinate structure that appears in a greater head-initial structure is itself head-initial, and a coordinate structure that appears in a greater head-final structure is itself head-final. This flexible approach is motivated in two areas, with respect to dependency distance and the nearness effect. Two counterarguments suggesting that all coordinate structures are head-initial were discussed and revealed to be faulty.

## 6 References

- Zoë Belk, Ad Neeleman, and Joy Philip. 2023. [What divides, and what unites, right-node raising](#). *Linguistic Inquiry*, 54.4:685-728.
- Ulrich Engel. 1982. *Syntax der deutschen Gegenwartssprache*. 2<sup>nd</sup> edition. Berlin: Erich Schmidt Verlag.
- Hans-Werner Eroms. 2000. *Syntax der deutschen Sprache*. Berlin: Walter de Gruyter.

- Kim Gerdes and Sylvain Kahane. 2015. [Non-constituent coordination and other coordinative constructions as dependency graphs](#). In *Proceedings of the Third International Conference on Dependency Linguistics*, (Depling 2015), pages 101–110. Uppsala, Sweden, August 24–26 2015.
- Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. [SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD](#). In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74 Brussels, Belgium, November 1, 2018.
- Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2019. [\\_Improving Surface-syntactic Universal Dependencies \(SUD\): surface-syntactic relations and deep syntactic features](#). In TLT 2019.
- Thomas Groß. 1999. *Theoretical Foundations of Dependency Syntax*. Munich: Iudicium.
- Richard Hudson. 1995. Measuring syntactic difficulty. Ms., University College London, London.
- Sylvain Kahane, Bernard Caron, Emmett Strickland, Kim Gerdes. 2021. [Annotation guidelines of UD and SUD treebanks for spoken corpora: a proposal](#). In TLT 2021.
- Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 2021. [A morph-based and a word-based treebank for Beja](#). In *Proceedings of the Third International Conference on Dependency Linguistics*. Sofia, Bulgaria.
- Hiroshi Kanayama, Na-Rae Han, Masayuki Asahara, Jena D Hwang, Yusuke Miyao, Jinho D Choi, and Yuji Matsumoto. 2018. [Coordinate structures in Universal Dependencies for head-final languages](#). In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 75–84, Brussels, Belgium, November 1, 2018.
- Katalin Kiss. 2012. [Patterns of agreement with coordinate noun phrases in Hungarian](#). *Natural Language and Linguistic Theory*, 30:1027–1060.
- 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.
- Henning Lobin. 1993. *Koordinationsyntax als prozedurales Phänomen*. *Studien zur deutsche Grammatik 46*. Tübingen: Gunter Narr Verlag.
- Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. [Universal Stanford Dependencies: A cross-linguistic typology](#). In *Proceedings of LREC*.
- Nicolas Mazziotta. 2011. [Coordination of verbal dependents in Old French: coordination as a specified juxtaposition or apposition](#). In *Proceedings of the First International Conference on Dependency Linguistics*. Pages 28–37, Pompeu Fabra University, Barcelona.
- Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. New York: State University of New York Press.
- Wolfgang Müller 1990. Die real existierenden grammatischen Ellipsis und die Norm: eine Bestandesaufnahme. *Sprachwissenschaft*, 15:241–266.
- Joakim Nivre and al. 2019. [Universal dependencies 2.4](#). LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Timothy Osborne and Thomas Groß. 2017. [Left node blocking](#). *Journal of Linguistics* 53, 641–688.
- Adam Przepiórkowski and Michał Wozniak. 2023. [Conjunct lengths in English, dependency length minimization, and dependency structure of coordination](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15494–15512, July 9–14, 2023.
- Adam Przepiórkowski, Magdalena Borysiak, and Adam Głowacki. 2024a. [An argument for symmetric coordination from dependency length minimization: A replication study](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1021–1033, 20–25 May, 2024.



- Adam Przepiórkowski, Magdalena Borysiak, Adam Okrasinski, Bartosz Pobożniak, Wojciech Stempniak, Kamil Tomaszek, and Adam Głowacki. 2024b. [Symmetric dependency structure of coordination: Crosslinguistic arguments from dependency length minimization](#). In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*, pages 11–22, December 5-6, 2024, Hamburg, Germany.
- Wojciech Stempniak. 2024. [Dependency structure of coordination in head-final languages: a dependency-length-minimization-based study](#). *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*, pages 65–75, Hamburg, Germany. Association for Computational Linguistics.
- David Temperley. 2007. [Minimization of dependency length in written English](#). *Cognition*, 105.2:300-333.
- Reiko Vermeulen. 2006. [Case and coordination in Japanese](#). In *Proceedings of the 25th West Coast Conference on Formal Linguistics*, pages 417-425. Somerville, MA: Cascadilla Proceedings Project.
- Yaqi Wang and Haitao Liu. 2017. [The effects of genre on dependency distance and dependency direction](#). *Language Sciences*, 59:135–147.



# Genre Variation in Dependency Types: A Two-Level Genre Analysis Using the Czech National Corpus

**Xinying Chen**

University of Ostrava  
Ostrava, Czech Republic  
cici13306@gmail.com

**Miroslav Kubát**

University of Ostrava  
Ostrava, Czech Republic  
miroslav.kubat@gmail.com

## Abstract

This paper examines how dependency type distributions vary across genres in the Czech National Corpus (SYN2020). Using a two-level genre classification, broad categories and fine-grained subgenres, we identify genre-sensitive syntactic patterns through relative frequency analysis. The results show that some dependency types (e.g. Atr 'attribute') vary consistently across genres, while others (e.g. ExD 'part of discourse ellipsis') show sensitivity only at the subgenre level. Our dependency-based approach extends common multidimensional analyses based on lexical-grammatical co-occurrences, directly capturing syntactic evidence and improving interpretability. Our findings also highlight the importance of fine-grained genre distinctions in revealing syntactic variation.

## 1 Introduction

Syntactic structure plays a central role in how information is organized and interpreted across different communicative contexts. One of the most important contextual factors that influence language use is the genre or style of the text (Biber and Conrad, 2009). While it is well recognized that genres impose different communicative goals and stylistic conventions, most of the existing work in stylometry studies focuses primarily on lexical features, such as word frequencies, stylistic markers, or vocabulary richness (Stamatatos, 2009; Kestemont, 2014), leaving syntactic variation relatively underexplored. However syntax, particularly as represented through dependency relations, provides valuable insights into how information is structured and presented differently across genres (Nivre, 2005; Roland et al., 2007; Webber, 2009).

This study contributes to the relatively unexplored area by investigating how the usage of dependency types varies across genres in the Czech National Corpus (SYN2020) (Křen et al., 2020;

Jelínek et al., 2021; Křivan and Šindlerová, 2022). Specifically, we take advantage of the corpus's hierarchical genre structure. It categorizes texts into three broader groups: fiction, non-fiction, newspapers and magazines. At the same time, it also provides more fine-grained subcategories, such as novels, short stories, scientific literature, professional literature, newspapers, leisure magazines, etc. This hierarchical genre organization provides a unique opportunity to explore genre sensitivity in syntactic structures both across and within broader genre categories.

Our primary objective is to identify which dependency types remain stable across genres and which ones display genre-sensitive variation. We examine the relative frequencies of all dependency types in both broader and fine-grained genre categories, using descriptive statistics such as maximum-minimum differences and standard deviation to quantify variability. This two-level analysis allows us to highlight dependency types that are structurally central and consistent, as well as those that are more genre-dependent. By comparing genre sensitivity at two levels of granularity, we provide a more nuanced understanding of how syntactic preferences are shaped by genre.

While previous influential genre analyses, notably Biber and Conrad (2009), utilized multidimensional analysis primarily focusing on English and lexical-grammatical feature co-occurrences, our study explicitly employed dependency tag, providing direct, transparent syntactic evidence across hierarchical genre distinctions. This approach enables deeper cross-linguistic and syntactic insights that complement and extend their foundational work.

Therefore, our results contribute to the growing body of research that integrates syntactic analysis into genre studies (Biber and Conrad, 2009; Oostdijk, 1998; Kubát et al., 2021, 2024; Chen and Kubát, 2024) and highlight the importance of us-

ing hierarchical genre structures in corpus-based syntactic research.

## 2 Data and Methodology

This study is based on data from the Czech National Corpus, specifically the SYN2020 subcorpus (Křen et al., 2020; Jelínek et al., 2021; Křivan and Šindlerová, 2022), a representative collection of contemporary written Czech containing approximately 100 million words mainly from 2015 to 2019. SYN2020 is annotated with morphosyntactic and syntactic information, including dependency relations. The syntactic annotation follows the principles outlined in the Prague Dependency Treebank framework (Hajič et al., 2020) and is performed automatically using a neural network-based parser from the NeuroNLP toolkit (Ma et al., 2018). The parser was trained on data from the analytical layer of the Prague Dependency Treebank and the syntactically annotated FicTree fiction corpus (Jelínek, 2017). The automatic syntactic annotation achieves a labeled attachment score (LAS) of 88.73% and an unlabeled attachment score (UAS) of 92.39% on test data<sup>1</sup>. While not manually verified, SYN2020 represents a significant improvement over previous corpus versions and provides consistent annotation across the entire 100-million-word corpus, ensuring reliable frequency comparisons across genres.

The corpus is organized into three primary groups: fiction(FIC), non-fiction(NFC), newspapers and magazines (NMG). Each of these groups contains multiple subcategories, including novels (NOV), short stories (COL), poetry (VER), drama or screenplays (SCR), scientific literature (SCI), professional (PRO) and popular writing (POP), memoirs and autobiographies (MEM), administrative documents (ADM), newspapers (NEW), and leisure magazines (LEI). Table 1 presents the structure of the corpus, and additional details are available in SYN2020 website.

For our analysis, we extracted frequency data for all dependency types occurring in the corpus. We first gathered absolute counts of each dependency type in all broad genre categories and subcategories. These counts were converted into relative frequencies within each genre, enabling fair comparisons across genres of differing sizes. This normalization was particularly important for the subgenre-level

<sup>1</sup>For detailed documentation of the automatic annotation pipeline, including morphological tagging and syntactic parsing procedures, see [https://wiki.korpus.cz/doku.php/cnk:syn2020:automaticka\\_anotace](https://wiki.korpus.cz/doku.php/cnk:syn2020:automaticka_anotace).

analysis, where category sizes varied considerably, see Table 1.

To assess genre sensitivity, we employed two descriptive statistical measures:

- **Maximum-Minimum Difference (Max-Min):** This value captures the absolute difference between the highest and lowest relative frequency of a dependency type across genres. A high Max-Min value suggests that a dependency type is used very differently depending on the genre.
- **Standard Deviation (SD):** This measure reflects the overall spread of the relative frequency of each dependency type across the fine-grained subgenres. Unlike Max-Min, SD accounts for all intermediate values and provides a more balanced view of variability. Due to the limited number of broader genre categories (only three), SD was not computed for the broader genre level.

We calculated these measures separately for the broader genre categories and the more fine-grained subgenres. To address possible internal heterogeneity within broad genre categories, we also examined variation across subgenres using standard deviation for each dependency type. This allowed us to quantify the extent to which syntactic usage diverges within genre groups, especially in categories like non-fiction where textual functions vary widely. It also enables us to compare how certain dependency types behave in both coarse and more fine-grained genre classifications, providing a systematic view of syntactic variability shaped by genre.

## 3 Analysis of Broader Genre Categories

In the first stage of analysis, we examined the relative frequency distributions of the dependency types in the three broader genre categories: FIC, NFC, and NMG. The metric used to assess variation was the Max-Min in relative frequency across the three genres. Dependency types with higher Max-Min values were interpreted as more genre-sensitive, while those with smaller values were considered more stable across genres. Table 2 presents the observed dependency variation (Max-Min > 0.5) across 3 broad genres, where Atr represents attributive relations (nominal modifiers), Obj marks direct objects, AuxK indicates sentence-ending punctuation, and ExD captures elliptical

Text-Type/Genre Group	Text-Type/Genre	Number of Word Tokens	Number of Sentences
FIC: fiction	NOV: novels	26,059,743	2,360,348
	COL: short stories	5,350,850	442,758
	VER: poetry	1,002,449	178,844
	SCR: drama, screenplays	1,003,033	156,750
NFC: non-fiction	SCI: scientific literature	9,284,751	459,801
	PRO: professional literature	7,013,611	405,546
	POP: popular literature	13,431,550	801,639
	MEM: memoirs and autobiographies	4,030,874	270,245
	ADM: administrative	348,920	24,795
NMG: newspapers and magazines	NEW: newspapers	20,393,309	1,402,548
	LEI: leisure magazines	13,601,340	1,049,767

Table 1: Text-type structure of SYN2020.

constructions. For definitions of the dependency type abbreviations, please refer to the Table 3.

Dependency	FIC	NFC	NMG	Max-Min
Atr	12.55	24.58	22.74	12.04
Obj	9.77	6.74	7.83	3.02
AuxK	7.72	4.70	5.84	3.02
Adv	12.72	9.91	11.04	2.81
AuxT	2.78	1.41	0.16	2.62
Pred_Co	5.10	2.90	3.26	2.20
Atr_Co	0.92	3.00	2.02	2.08
AuxP	7.83	9.36	9.86	2.02
ExD	2.72	1.41	1.55	1.31
Pred	4.35	3.05	3.96	1.30
AuxG	3.43	3.08	2.29	1.14
AuxC	2.47	1.58	1.57	0.90
AuxX	6.02	5.41	5.21	0.82
Sb	6.04	5.62	6.43	0.80
AuxV	1.72	1.16	1.17	0.55

Table 2: Dependency variation across 3 broad genres.

The most genre-sensitive dependency type was Atr (attribute), where non-fiction and news texts show markedly higher frequencies compared to fiction (Max-Min = 12.04). This pattern reflects the tendency of non-fiction and journalistic writing to make extensive use of noun modifiers in order to express specific, technical, or formal information. This result aligns with previous findings in English that associate high syntactic density and nominal modification with informational density in informational genres (Biber, 1988; Biber and Conrad, 2009). However, by explicitly analyzing syntactic dependencies rather than lexical co-occurrence patterns, our analysis provides direct syntactic evidence and a more precise characterization of these genre distinctions.

Dependency	Definition
Atr	Attribute (adjective)
Obj	Object
AuxK	Sentence-ending punctuation
Adv	Adverbial (adverbial determination)
AuxT	Reflexive particle 'se' in inherently reflexive verbs
Pred	Predicate
AuxP	Preposition
ExD	Part of discourse ellipsis
AuxG	Other graphic symbols that do not end a sentence
AuxC	Subordinating conjunction
AuxX	Comma
Sb	Subject
AuxV	Auxiliary verb být (to be)
Coord	Coordination node
AuxZ	Emphatic word
AuxY	Adverbs and particles that cannot be classified elsewhere
Pnom	Nominal part of a verbonominal predicate
Apos	Apposition (main node)

Table 3: Definitions of dependency type abbreviations in Czech syntactic analysis. Dependency ending `_Co` is for tokens that are coordinated and ending `_pa` is for part of parentheses. For example, coordinated attributes are assigned the function `Atr_Co`. For more information, please check the introduction page of the [Czech National Corpus](#) and [Prague Dependency Treebank Annotation Manual](#).

Obj (object) and AuxK (sentence-ending punctuation) also demonstrated notable genre sensitivity. Although their Max-Min values were smaller than Atr, this does not imply insignificant variation. Instead, it suggests subtler stylistic variation across genres. For instance, the relatively higher frequency of Obj in fictional texts (Max-Min = 3.02) reflects their narrative-driven syntax, emphasizing events and actions. Regarding AuxK (Max-Min = 3.02), its frequency directly corresponds to the number of sentences in the corpus. Narrative texts (FIC) typically contain shorter sentences and frequent dialogues, resulting in a higher number of sentences and thus increasing the relative fre-

quency of sentence-ending punctuation. In contrast, informational texts (NFC and NMG) often feature longer sentences designed to convey complex ideas, leading to fewer sentences overall and consequently reducing the frequency of AuxK. Therefore, the observed genre variation in AuxK primarily represents differences in sentence segmentation, sentence count, and syntactic complexity across genre categories.

The majority of dependency types such as Sb (subject), AuxV (Auxiliary verb *být* 'to be'), and many other dependency types ( $\text{Max-Min} \leq 0.5$ ) demonstrated relatively stable distributions across all three genre categories. These dependency types exhibit consistent usage patterns across genres, suggesting they fulfill fundamental syntactic roles in Czech that are relatively unaffected by stylistic variation.

The analysis of the broader genre categories reveals both structural constants and genre-sensitive syntactic choices. Types like Atr, Obj, and AuxK demonstrate meaningful variation across genres and can thus serve as indicators of broader stylistic tendencies in written Czech.

#### 4 Analysis of Fine-Grained Genre Subcategories

While the analysis of broader genre categories revealed general patterns of syntactic variation, it is at the subgenre level that genre sensitivity becomes more evident and interesting. Our subgenre analysis reveals dramatically increased variation, with ExD frequencies ranging from 0.20 in leisure magazines to 14.03 in poetry, a 70-fold difference invisible at the broader genre level.

To account for variation in the size of these subcategories, we used normalized relative frequencies and applied both Max-Min and sample standard deviation (SD) as metrics of variability. Whereas Max-Min highlighted extreme contrasts in usage across subgenres, SD allowed us to capture more distributed forms of variation. The results of analysis ( $\text{Max-Min} > 0.5$ ) are presented in Table 4.

Several dependency types emerged as highly sensitive at this level of analysis. Atr once again topped the list. The consistently high frequency of Atr in ADM and SCI genres highlights their requirement for syntactic density and precision. Administrative texts, characterized by formality and specificity, rely heavily on noun modifiers to express precise legal or bureaucratic concepts. Similarly,

scientific literature employs dense noun phrases extensively to contain technical details and methodological precision clearly and concisely, aligning with the informational focus of these genres. This reinforces the role of Atr as a marker of dense, information-heavy discourse (Biber and Conrad, 2009).

In contrast, ExD (part of discourse ellipsis) showed marked sensitivity at the subgenre level, contrasting its relative stability at the broader genre level. Its distribution was notably uneven, peaking in literary texts, especially poetry and drama. ExD refers specifically to elements omitted from sentences because they can be inferred from the context. This high variation likely reflects genre-specific stylistic conventions related to brevity, informality, and implied meaning. For example, poetry frequently utilizes elliptical constructions to create ambiguity, enhance rhythmic conciseness, or engage readers in interpreting implicit meanings. Similarly, dramatic texts commonly feature discourse ellipsis to simulate natural speech patterns, spontaneous dialogues, or emotional intensity by omitting linguistic elements clearly understood from the conversational context. The relatively low frequency of ExD in more formal or informational genres, such as scientific literature and administrative texts, aligns with their explicitness and precision, which discourage reliance on contextual inference. This pattern aligns with previous findings highlighting genre-specific syntactic phenomena distinguishing narrative and expressive texts from expository and formal writing (Biber and Conrad, 2009).

Other types such as Pred\_Co (coordinated predicates) and AuxP (preposition) also ranked highly in variability. Pred\_Co exhibited notable differences across genres. This variation reflects stylistic preferences for predicate coordination, which are more frequent in conversational or literary subgenres probably due to their use of compound predicates that facilitate narrative flow or rhythmic expression. Conversely, scientific and administrative texts tend toward simpler predicate structures to enhance precision and clarity, thus explaining their lower frequencies of Pred\_Co. AuxP variation highlights genre differences in prepositional phrase complexity and density. Technical genres like scientific literature and administrative documents often exhibit a higher frequency of AuxP due to their reliance on prepositional phrases to precisely convey complex information, whereas literary genres typically use



Dependency	NOV	COL	VER	SCR	SCI	PRO	POP	MEM	ADM	NEW	LEI	Max-Min	SD
Atr	12.73	18.21	2.83	9.44	28.66	30.07	23.94	18.78	34.02	23.84	22.01	31.19	9.32
ExD	2.57	2.98	14.03	6.51	1.76	1.36	1.43	1.15	3.37	1.28	0.20	13.83	3.92
Obj	10.31	1.21	13.60	9.61	6.19	6.30	7.40	0.98	5.85	7.59	8.56	12.62	3.69
Adv	13.26	1.71	2.06	9.93	9.49	9.42	10.70	13.90	7.14	10.68	12.11	12.19	4.04
AuxK	7.93	9.55	1.31	11.84	4.42	4.92	4.88	6.63	4.68	5.58	6.52	10.52	2.82
AuxP	8.07	10.74	1.70	5.70	10.19	10.64	9.22	10.01	12.09	10.16	9.82	10.39	2.91
Sb	6.22	0.81	10.44	6.41	5.79	5.95	5.95	6.28	4.78	6.76	6.18	9.63	2.22
Pred	4.46	5.10	12.23	5.63	2.88	3.55	3.10	3.84	3.49	4.01	4.07	9.34	2.61
AuxX	6.33	8.33	0.67	4.16	5.91	4.89	5.69	6.87	0.45	5.04	5.70	7.88	2.41
Coord	4.37	6.01	8.33	4.44	4.73	0.45	4.35	4.77	4.72	3.86	4.90	7.88	1.84
Pred_Co	5.27	6.96	7.82	4.95	0.25	0.24	3.23	4.80	0.18	2.79	4.17	7.64	2.65
ExD_Co	1.22	1.87	5.87	3.80	1.68	1.05	1.13	1.12	2.09	0.99	1.72	4.88	1.50
AuxG	3.69	4.05	0.27	3.63	4.43	2.94	2.80	2.35	0.07	2.24	0.25	4.36	1.58
Atr_Co	0.09	1.55	0.15	0.47	0.42	3.72	2.62	1.74	4.24	2.02	2.11	4.15	1.41
AuxC	2.61	3.14	3.63	1.98	0.14	1.18	0.18	2.55	0.77	1.47	0.18	3.48	1.25
AuxT	2.94	3.50	3.82	2.20	1.31	1.08	1.54	2.46	0.64	1.48	1.84	3.19	1.02
AuxV	1.80	2.23	0.22	0.18	0.10	1.08	1.04	2.70	1.09	1.16	1.24	2.60	0.83
Obj_Co	0.11	1.59	1.26	0.68	1.47	1.50	1.39	0.14	1.87	1.32	1.61	1.76	0.60
AuxZ	0.13	1.84	0.27	1.04	1.65	1.72	1.71	1.66	0.86	1.77	0.19	1.71	0.70
AuxY	0.78	1.13	1.23	0.77	0.94	0.72	0.87	0.91	0.04	0.70	0.83	1.19	0.30
Adv_Co	0.69	1.11	0.10	0.38	1.08	0.94	0.94	0.94	1.27	0.72	0.96	1.17	0.34
Coord_Co	0.51	0.83	0.96	0.60	0.05	0.36	0.44	0.55	0.40	0.34	0.50	0.91	0.24
Phom	1.30	1.72	2.18	1.72	1.63	1.70	1.56	1.57	1.34	1.35	1.61	0.88	0.24
Sb_Co	0.42	0.65	0.65	0.33	1.13	1.02	0.84	0.70	0.92	0.80	0.85	0.79	0.24
ExD_Pa	0.31	0.38	0.61	0.93	0.45	0.22	0.20	0.17	0.35	0.15	0.19	0.78	0.23
Apos	0.18	0.29	0.39	0.47	0.07	0.55	0.45	0.35	0.63	0.36	0.04	0.59	0.19

Table 4: Dependency variation across fine-grained genres.

fewer dense prepositional constructions.

At the same time, several core syntactic functions, such as Sb (subject), Coord (coordination), and AuxX (auxiliary in coordinated constructions), continued to show low variability across subgenres. This consistency supports the notion that certain syntactic dependencies remain relatively unaffected by genre, functioning as part of the grammatical infrastructure of Czech syntax.

Our fine-grained analysis further extends beyond the level of detail achievable through lexical-grammatical multidimensional analyses, as employed by [Biber and Conrad \(2009\)](#). The dependency tag explicitly reveals subtle yet important stylistic differences and subgenre-specific syntactic variations, such as the distinctive high frequency ExD in literary subgenres. It provides syntactic insights that indirect co-occurrence analyses may not capture.

Together, these findings reveal that while some dependency types maintain stability across genre levels, others become more genre-sensitive when fine-grained distinctions are considered. The results underscore the importance of using hierarchical genre structures in syntactic analysis to avoid

averaging out meaningful stylistic variation.

These findings also highlight internal genre heterogeneity. For instance, the non-fiction category includes both scientific texts and memoirs, which show sharply different syntactic profiles, Atr ranges from 18.78% in memoirs to over 34% in administrative texts. Similarly, genres like leisure magazines and newspapers differ in ExD usage despite both falling under NMG. These internal divergences show that traditional genre groupings may mask important syntactic variation, which supports the need for finer-grained or data-driven genre modeling.

## 5 Comparison and Interpretation

Having examined dependency type distributions across both broader genre categories and fine-grained subcategories, we now compare the findings to better understand how syntactic variation is shaped by different levels of genre granularity. This comparison offers insights into the types of dependencies that are consistently genre-sensitive, those that are genre-neutral, and those whose variability becomes more apparent at the subgenre level.

To facilitate this comparison, we ranked all dependency types according to their Max-Min values

in both levels of analysis.<sup>2</sup> Our results reveal significant granularity effects: Atr maintains rank 1 at both levels but with obvious value changes (Max-Min: 12.04 broad, 31.19 fine), while ExD shifts from rank 9 (Max-Min: 1.31) to rank 2 (Max-Min: 13.83).

To visualize these dynamics, we created a scatterplot comparing coarse-grained and fine-grained sensitivity ranks (Figure 1). Points along or close to the diagonal indicate consistent genre sensitivity across both classification levels. Points deviating more from the diagonal represent dependency types whose genre sensitivity becomes more apparent at finer granularity. For instance, ExD, positioned more far from the diagonal than Atr, demonstrates that its sensitivity to genre is not as evident at the broader level but becomes more pronounced within subgenre distinctions, highlighting the importance of analyzing detailed subgenre classifications to uncover syntactic patterns. This visual comparison demonstrates how coarse categorization can sometimes obscure important syntactic variability.

Figure 1 shows that while many dependency types exhibit stable genre sensitivity regardless of granularity, a smaller subset displays considerable divergence between the two levels. These divergences are particularly important, as they highlight constructions that are sensitive to more subtle communicative or stylistic demands found only in specific subgenres. Importantly, the most genre-sensitive types (i.e., those with the lowest ranks) cluster in the lower-left corner of the plot. Identifying these genre-sensitive dependencies such as Atr, ExD, and Obj has practical implications for computational linguistics applications, particularly automated genre classification. Dependency types sensitive to genre differences can improve the accuracy of classifiers by incorporating genre-specific syntactic features, thus enhancing linguistic modeling in computational frameworks.

In a second visualization, we plotted Max-Min values against SD for each dependency type in the fine-grained analysis, see Figure 2. The overall distribution in Figure 2 reveals a positive relationship between Max-Min and SD: dependency types that show stronger sensitivity to genre distinctions also tend to fluctuate more across subgenres. This confirms that genre-sensitive types are not only skewed toward specific contexts but also exhibit greater instability, further reinforcing their role as

stylistically responsive constructions.

Using the mean values of Max-Min and SD as thresholds, we divided the space into four interpretive zones:

- Low Range / Low Spread (bottom-left): 113
- High Range / High Spread (top-right): 19
- High Range / Low Spread (bottom-right): 4
- Low Range / High Spread (top-left): 0

The vast majority of dependency types fall into the low range / low spread quadrant, indicating that they are largely genre-neutral and stable across subgenres. In contrast, a small but crucial set of types cluster in the high range / high spread zone. These include types such as Atr, ExD, and Sb, which exhibit both strong genre sensitivity and high variability. This identifies them as key indicators of subgenre-specific syntactic preferences.

Interestingly, the absence of types in the low range / high spread quadrant suggests that high variability almost never occurs without accompanying genre sensitivity. That is, wide fluctuations in usage typically correspond to meaningful genre-driven effects, rather than random variation.

This analysis further supports the importance of examining genre at multiple levels of resolution. While many dependency relations remain stable regardless of context, a focused view on fine-grained distinctions reveals important dimensions of syntactic variability that would otherwise remain hidden.

This comparative approach demonstrates explicitly how dependency-based syntactic analysis provides methodological depth and granularity beyond previous multidimensional analyses (Biber and Conrad, 2009). By directly mapping syntactic patterns onto genre distinctions at both coarse and fine-grained levels, our method explicitly identifies syntactic features sensitive to subtle genre differences, thus notably enriching the theoretical and methodological scope of genre analysis.

## 6 Discussion

The results of this study underscore the importance of incorporating genre structure into syntactic analysis. The comparison between broader and fine-grained genres revealed both the stability and variability of dependency types. While majority of syntactic relations showed consistent distributions regardless of genre, others such as Atr, ExD, and

<sup>2</sup>Tied values were given the same lowest possible rank.



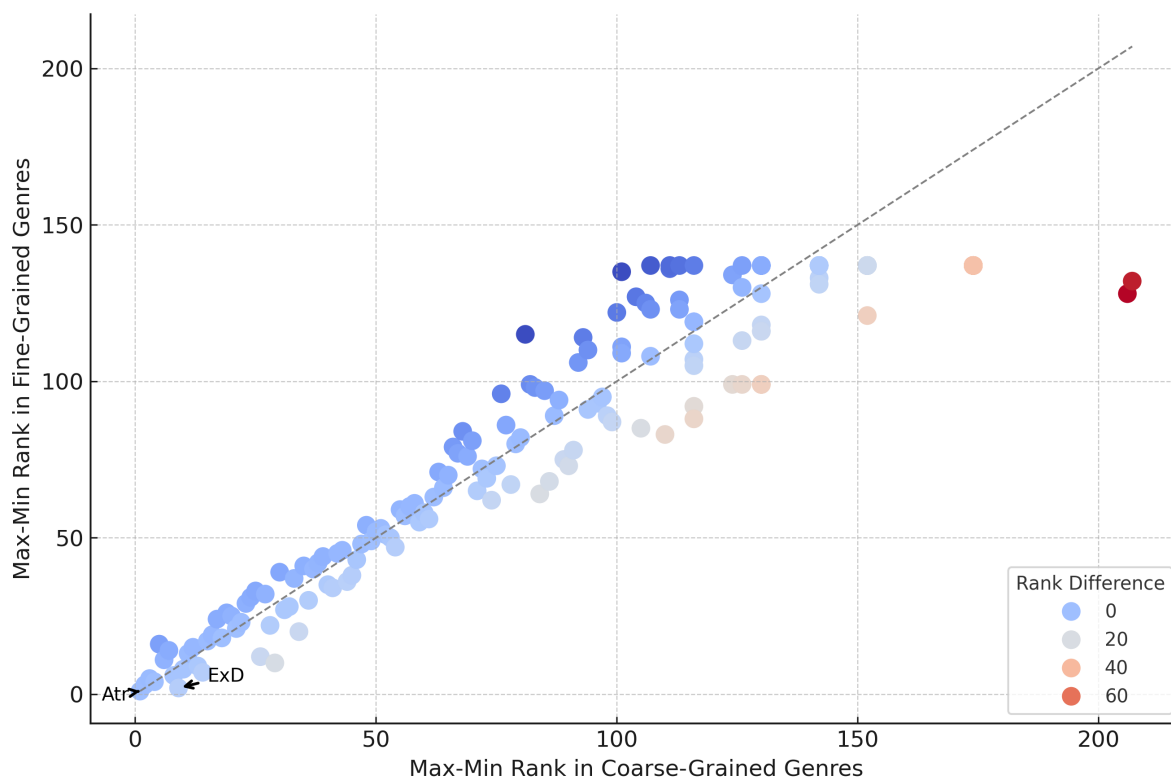


Figure 1: Comparison of dependency sensitivity ranks between coarse- and fine-grained genres.

Obj were markedly sensitive to the communicative and stylistic demands of specific genres.

Importantly, the internal heterogeneity observed within broad genre groups is not a limitation of genre-based analysis but an opportunity for refinement. Our two-level analysis illustrates that genre categories are often composed of syntactically distinct subtypes. Rather than assuming genre homogeneity, our approach enables empirical evaluation of genre cohesion and reveals when fine-grained distinctions are warranted. This supports a more dynamic, corpus-driven model of genre.

Crucially, these findings question the adequacy of relying exclusively on broad genre categories for syntactic analyses, as such coarse classifications may level up subtle yet important stylistic features. Our results align with [Biber and Conrad \(2009\)](#), emphasizing that communicative, stylistic, and functional differences in language frequently manifest at fine-grained levels of genre variation. For instance, dependency types like ExD, clearly more sensitive at the subgenre level, illustrate precisely the kind of stylistic phenomenon that broader classifications may obscure. Thus, incorporating multiple granularity levels is essential not only theoretically but also practically for linguistic analyses

that aim for accuracy and depth.

The use of both Max-Min and SD measures further allowed us to differentiate between types that exhibit obvious shifts and those not. This dual perspective provided a richer view of how syntactic preferences are shaped across the genre spectrum. Moreover, the visual analyses confirmed that variability is not uniformly distributed; some types are tightly linked to fundamental syntactic roles of language, while others are more responsive to genre-specific stylistic conventions. These results offer practical implications for areas such as genre-aware syntactic parsing, authorship attribution, and language modeling, where understanding genre-specific syntactic tendencies can improve performance and interpretability.

## 7 Conclusion

This study has presented a two-level genre analysis of dependency type distributions in the Czech National Corpus. By examining both broader genre categories and fine-grained subgenres, we identified which dependency relations are structurally stable and which are sensitive to genre distinctions. The analysis demonstrated that certain types, such as Atr, show strong and consistent variation across

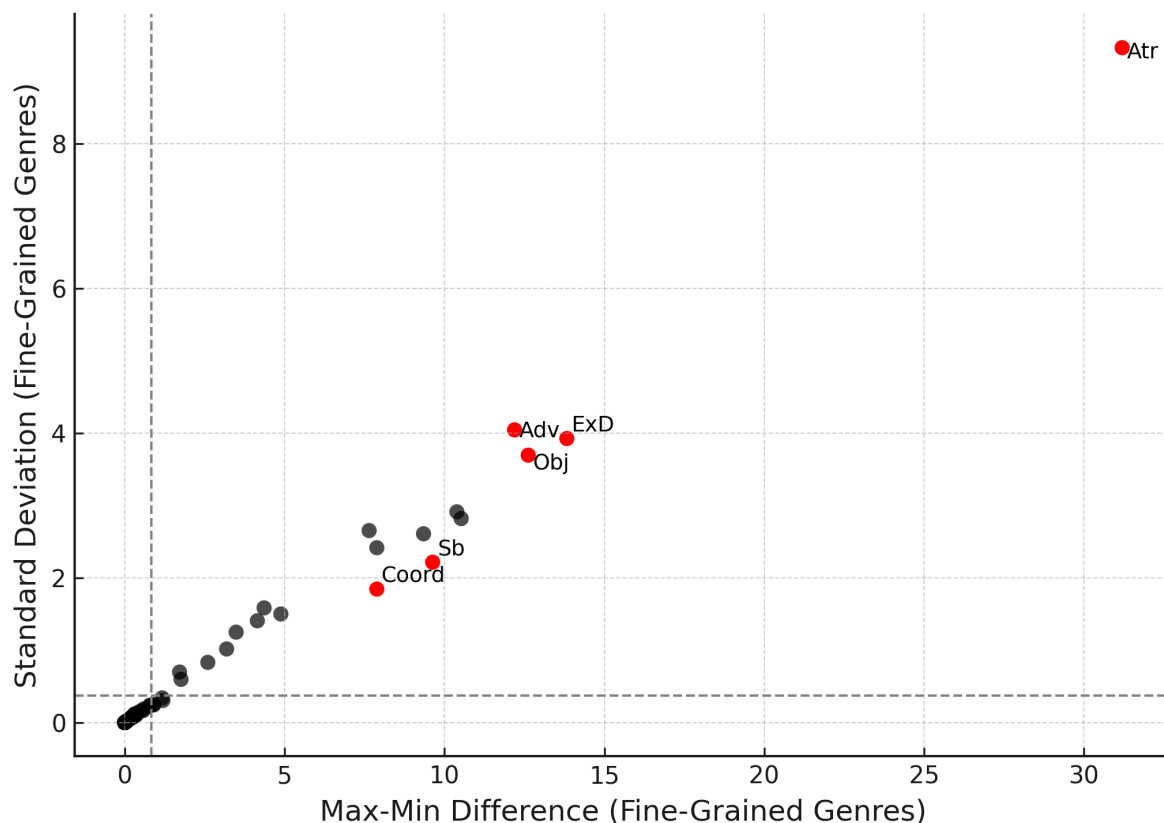


Figure 2: Scatter plot of dependency types based on Max-Min and SD across fine-grained genres. Red points are selected samples for illustration. Dashed lines indicate the mean values for each axis, dividing the plot into four interpretive regions.

both levels, serving as indicators of informational density in formal genres. Others, like ExD, revealed their stylistic specificity more at the sub-genre level, highlighting the expressive features of narrative and literary texts.

Overall, the findings confirm that genre plays a critical role in shaping syntactic preferences and that this role can only be fully appreciated by analyzing data at multiple levels of granularity. By explicitly examining dependency relations across multiple genre levels, our study substantially complements foundational multidimensional analyses (Biber and Conrad, 2009), offering direct syntactic evidence and enhanced theoretical insights that deepen our understanding of genre-specific linguistic variability.

Future research could beneficially extend these analyses cross-linguistically or explore computational approaches to utilize genre-sensitive syntactic patterns in natural language processing applications. Investigating the consistency of these findings across different languages and genre classification frameworks would further clarify the rela-

tionship between syntactic variation and communicative context.

### Limitations

Despite the insights provided, several limitations should be acknowledged. First, this study exclusively relies on the SYN2020 from the Czech National Corpus, which covers texts mainly from 2015–2019. Consequently, the findings may not generalize to other time periods or linguistic contexts, as language use can evolve considerably even within relatively short spans.

Second, while SYN2020 offers a robust genre classification scheme, the hierarchical categorization used in this analysis may still obscure more complex stylistic variations. Certain genres could contain internal heterogeneity, and additional sub-classifications might yield further insights.

Last but not the least, dependency frequency measures alone do not capture the complexity of syntactic variation fully. Including additional linguistic features such as dependency distance might enhance the understanding of genre variations.

## Acknowledgments

This work was supported by the Czech Science Foundation (GAČR), project No. 22-20632S.

## References

- Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press, Cambridge, UK.
- Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge University Press, Cambridge, UK.
- 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.
- Jan Hajič, Eduard Bejček, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. [Prague dependency treebank - consolidated 1.0](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5208–5218, Marseille, France. European Language Resources Association.
- Tomáš Jelínek. 2017. [Fictree: A manually annotated treebank of czech fiction](#). In *ITAT 2017 Proceedings*, pages 181–185.
- Tomáš Jelínek, Jan Křivan, Vladimír Petkevič, Hana Skoumalová, and Jana Šindlerová. 2021. [SYN2020: A new corpus of Czech with an innovated annotation](#). In *Text, Speech, and Dialogue. TSD 2021*, volume 12848 of *Lecture Notes in Computer Science*, pages 48–59, Cham. Springer.
- Mike Kestemont. 2014. Function words in authorship attribution. from black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66.
- Michal Křen, Václav Cvrček, Jan Henyš, Milena Hnátková, Tomáš Jelínek, Jan Koček, Dominika Kovářiková, Jan Křivan, Jiří Milička, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Jana Šindlerová, and Michal Škrabal. 2020. [SYN2020: reprezentativní korpus psané češtiny](#). Ústav Českého národního korpusu FF UK, Praha.
- Jan Křivan and Jana Šindlerová. 2022. Změny v morfologické anotaci korpusů řady SYN: nové možnosti zkoumání české gramatiky a lexikonu. *Slovo a slovesnost*, 83(2):122–145.
- Miroslav Kubát, Ján Mačutek, Radek Čech, and Michaela Nogolová. 2024. Automatic genre classification of czech texts based on syntactic functions. In G. Giordano and M. Misuraca, editors, *New Frontiers in Textual Data Analysis*, pages 163–172. Springer.
- Miroslav Kubát, Radek Čech, and Xinying Chen. 2021. [Attributivity and subjectivity in contemporary written czech](#). In *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pages 58–64. Association for Computational Linguistics.
- Xuezhe Ma, Zecong Hu, Jingzhou Liu, Nanyun Peng, Graham Neubig, and Eduard Hovy. 2018. [Stack-pointer networks for dependency parsing](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, page 1403–1414, Melbourne, Australia. Association for Computational Linguistics.
- Joakim Nivre. 2005. Dependency grammar and dependency parsing. Technical Report MSI report 05133, Växjö University, School of Mathematics and Systems Engineering.
- Nelleke Oostdijk. 1998. [A corpus-based model of syntactic variation with applications for english](#). *Literary and Linguistic Computing*, 13(3):147–153.
- Douglas Roland, Frederic Dick, and Jeffrey L. Elman. 2007. [Frequency of basic english grammatical structures: A corpus analysis](#). *Journal of Memory and Language*, 57(3):348–379.
- Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.
- Bonnie Webber. 2009. [Genre distinctions for discourse in the Penn TreeBank](#). In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 674–682, Suntec, Singapore. Association for Computational Linguistics.

# A morpheme-based treebank for Gbaya, an Ubanguian language of Central Africa

**Paulette Roulon-Doko**  
CNRS, LLACAN, France  
paulette.roulon@cnrs.fr

**Sylvain Kahane**  
Paris Nanterre University,  
CNRS, MoDyCo, France  
Institut Universitaire de France  
sylvain@kahane.fr

**Bruno Guillaume**  
Universite de Lorraine,  
CNRS, Inria, LORIA, France  
Bruno.Guillaume@inria.fr

## Abstract

In this paper, we present the first treebank for Gbaya, a language from the under-resourced Niger-Congo family. The language has a rich system of tonal morphemes and virtually no affixes. The dependency analysis is based on a morpheme-based tokenisation and the treebank is also distributed in a word-based Universal Dependencies version. Several constructions are discussed in the paper: genitive construction, clause coordination, sentence particles, adverbial and relative clauses, serial verb constructions, reported speech, topicalization, and focalization.

## 1 Introduction

This paper presents the first treebank for a Gbaya language. We have decided to follow the tokenization proposed by Roulon-Doko (1995). In this previous work, the analysis considers many tones as autonomous units that combine with root lexical tonal pattern for grammatical reasons; there are called grammatical tones (Hyman, 2016; Rolle, 2018). These tones are placed after the determined term. In the case of verbs without lexical tonal pattern, the tone is attached to the verb base and is therefore always indicated beforehand. In order to encode this tokenization level, we use the mSUD framework (Guillaume et al., 2024). mSUD was designed for morpheme-based level annotation; it is a variant of the Surface Syntactic Universal Dependencies (SUD) (Gerdes et al., 2018a) annotation schema.<sup>1</sup> In this paper, we will use the term “morpheme” also for units that represent tones. We will consider three kind of tokens: roots, clitics, and inflectional morphemes (which are mainly tonal tokens).

<sup>1</sup>The mSUD corpus is available at [https://github.com/surfacesyntacticud/mSUD\\_Northwest\\_Gbaya-Autogramm](https://github.com/surfacesyntacticud/mSUD_Northwest_Gbaya-Autogramm).

The corpus is also available<sup>2</sup> in the Universal Dependencies (UD) framework (de Marneffe et al., 2021). To meet the UD tokenisation requirements, we provide automatic conversion to the UD format, in which tones are not expressed as separate tokens.

The corpus is made up of three tales. The tales in Gbaya are a repertoire without specialists. The language used to tell tales is the language of everyday life, with no stylistic form of its own. The storyteller, whether male or female, young or old, takes the floor spontaneously during a storytelling session. These three tales were recorded in 1970 in the village of Ndongué (Central African Republic) during traditional storytelling evenings by Paulette Roulon. Tale T16-C6, which tells the story of the woman fishing at the dam whose baby was swept away by the water after the dam burst, opened the session. It was told by Anna Zàngé, a woman in her forties. This was followed by tale T9-C7, which tells the story of the brother and sister who went on a hunting camp, which was then told by Yvonne Yàì-s̀̀ a young girl of around 17. Tale T24-C59, which tells the story of Wanto and the little cob, was told at another session by Hélène Dũ̀̀, a young woman in her thirties. The Table 1 gives an overview of the length of each tale in the treebank **mSUD\_Northwest\_Gbaya-Autogramm**.

Northwest Gbaya (ISO: GYA, WALS: gbk, Glottocode: nort2775) is part of the main linguistic group Gbaya-Mandja-Ngbaka, an Ubanguian language family (a branch of the Niger-Congo phylum, Adamawa Ubangi). It is spoken in the northwest of the Central African Republic (CAR) and in the central-eastern part of Cameroon. It is subdivided in six dialects: four in CAR [fòddòè, fòkpan, fòjìnà, fùgùì], usually named Gbaya-kara, and two in Cameroon [fòyà, yàáyùwèè]. In

<sup>2</sup>[https://github.com/UniversalDependencies/UD\\_Northwest\\_Gbaya-Autogramm](https://github.com/UniversalDependencies/UD_Northwest_Gbaya-Autogramm)

	audio length	# of sentences	# tokens (morph-based)	# tokens (word-based)
T16-C6	321s	143	1,363	847
T9-C7	178s	79	820	503
T24-C59	359s	181	1,621	1,067
<b>TOTAL</b>	<b>858s</b>	<b>403</b>	<b>3,804</b>	<b>2,417</b>

Table 1: Sizes of the 3 samples of the corpus **mSUD\_Northwest\_Gbaya-Autogramm** (version 2.15).

1996, there were 265,000 speakers: 200,000 in the western part of CAR and 65,000 in the central-eastern part of Cameroon. This paper deals with the Gbáyá bòddè, a Northwest Gbaya dialect spoken in Central African Republic. The annotation is based on a dictionary from Roulon-Doko (2008) and a complete grammar, based on a 4h50 oral corpus of spontaneous speaking, collected in the field between 1970 and 2013, processed with Toolbox and Elan, glossed and translated into French, by Roulon-Doko.

The consonant system has three glottalized consonants (b, d, ʔ), four labio-velars (kp, gb, ngb, ɟm), oral consonants (both voiceless and voiced) and a complete range of nasals and semi-nasals. There are seven oral and five nasal vowels. Gbaya is a tonal language with two levels and four tones (H, L, LH, HL). All vowels carry a tone. Gbaya syllabic structure includes open and closed syllables, but no initial vowel.

Gbaya is an isolating language with very little morphology and no agreement at all. Gbaya relies minimally on derivation but makes strong use of compounding, marked in writing by a hyphen between components (e.g. *gègè-fið* ceremony sp.), which is also used for adjectives-adverbs with a reduplicated structure (e.g. *bàdàm-bàdàm* irregularly arranged).

The lexemes are thus simple, compound, or structurally reduplicated. They are distributed across 18 categories (Roulon-Doko, 2008). Gbaya has verbs, non-verbal predicates, nouns and four subcategories of adjectives. The main lexical categories, VERB (10%), NOUN (50%), and ADJ (32.6%), have unrelated lexical stocks. For nouns, composition is very important (47.6%) and derivation very little used (3.7%).

Word order is very strict. Gbaya is an SVO language, that also makes use of non-verbal predication. The subject is compulsory, except for some construction that will be described below.

In Section 2, we explain how the tone system

is encoded as “morphemes” in the mSUD framework. The Section 3 describes the principles used for defining the maximal units. Several interesting constructions of Gbaya are described in Section 4, together with the proposed analysis in the treebank. The last Section 5 is dedicated to the conversion to the UD treebank.

## 2 The morpheme-based annotation

Verb inflection in Gbaya is marked by tone alternations, possibly accompanied by affixes. Verbal lexemes have no lexical tonal pattern, the tonal pattern they carry systematically comes from the TAM marker and is identical for all. The TAM verbal system is organized in three moods (Realis, Virtual and Command) and two aspects (Perfective and Imperfective). Tense is not marked on the verb in Gbaya (Roulon, 1975; Roulon-Doko, 1994). Verbs always express a process where the obligatory subject is either external to the process (transitive construction = transitive voice) or included in the process (intransitive construction = middle voice).

### 2.1 Tonal tokens

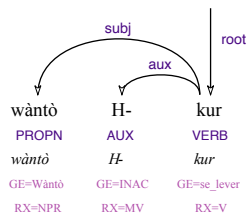
Tonal tokens, or morphotonemes (Meeussen, 1967), are inflectional morphemes that impose tonal patterns or trigger tonal alternations. Three cases of tonal tokens are discussed.

**TAM markers.** The 19 TAM markers are all tonal tokens. They include a verbal pattern that we place in front of the verb root (R) for the sake of regularity, even if they contain suffixal elements. For TAM markers, we follow the UD annotation scheme: they receive **upos=AUX** and depend on the root by a **aux** relation: see sentences (1) and (2).<sup>3</sup>

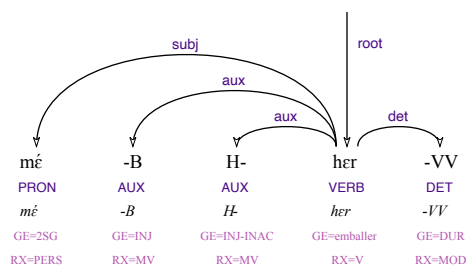
<sup>3</sup>In fact, TAM marker are distributional heads, because they control the distribution of the verb forms, which is different between finite and non-finite forms. In consequence, auxiliaries are generally treated as heads in SUD. Due to the particular status of Gbaya’s TAM markers, which are only inflectional morphemes, we have treated them as dependents, following UD.



- (1) *Wàntò kúr*  
 H-kur  
 NPR IPFV-get\_up  
 ‘Wanto gets up.’



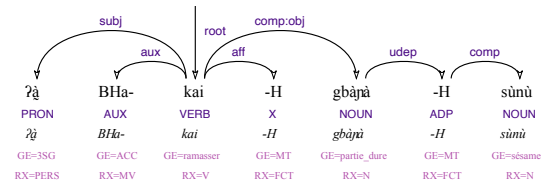
- (2) *mè hér*  
 mé-B H-her-VV  
 2SG-INJ INJ\_IPFV-wrap-DUR  
 ‘Wrap it.’



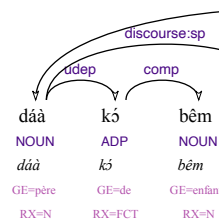
In (1), we show the realis imperfective (IPFV), which is realized by a high tone H-. In (2), we show the injunctive-imperfective (INJ\_IPFV), which is also realized by a high tone H- on the verb, but also imposes a low tone -B on its subject, which must obligatorily be a personal pronoun. Such a construction forces us to make choices when converting the treebank to the word-based UD annotation scheme. The fusion of the personal pronoun and the low tone could be analyzed an alternative form of the pronominal subject (our choice) or with an auxiliary bearing a pronominal index. If the second case is common in many languages, it would have been uncommon in Gbaya, which has no other cases of pronominal indices and where clauses always have an overt subject.

**The floating tone -H.** There are regular tonal modifications linked to the presence of a floating high tone -H, which occurs between two words. It modifies the tonal pattern of the first word depending on the first tone of the second word according to fixed rules, except in the case of a high pattern. This floating high tone can mark grammatical distinctions (functioning as a supra-segmental marker, TM in the glosses). It occurs in two con-

- (3) *ʔá kàýá gbàjá sùnù*  
 BHa-kai-H gbàjá-H sùnù  
 3SG PFV-gather-MT hard\_part-MT sesame  
 ‘He collects sesame seeds.’



- (4) *dáà kó bém*  
 father of child  
 ‘The childs father.’



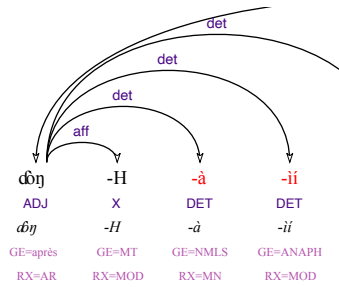
texts. Either (i) it is structurally attached to certain words (preposition, perfective verbal marker, etc.), involving no choice on the part of the speaker, similar to class markers, for instance (as the first -H in (3)); or (ii) it functions as a connective linking the two nouns of a genitive phrase (second -H in (3)) contrasting with the genitival SN with segmental connective *kó* (4).

In the first case, it is analyzed as an affixe of first word, with **upos=X** and linked to the root by a relation **aff** (it represents 75% of the occurrences in the treebank). In the second case, it is analyzed with **upos=ADP**, like the real ADP *kó*. In SUD, ADPs are analyzed as the head of the adpositional phrase and the noun depends on the ADP by a relation **comp**.

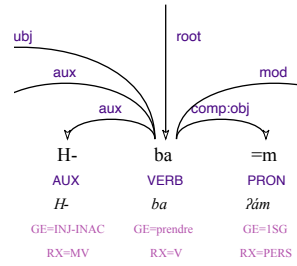
**Multicategory markers.** Multicategory markers are suffixes or morphotonemes that can combine with lexemes of different parts of speech. We treat all of them as suffixes and give them **upos=DET**. They are the anaphoric *-ií*, the locative *-è*, the insistent *-V* polar tone, the durative *-VV*, the suffix *-à* (definite, nominalizer). They are not exclusive from each other. See (5) and (6).



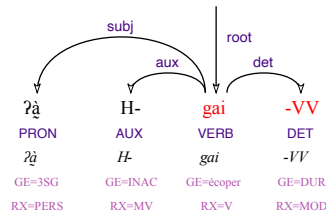
- (5) *dòpáìí*  
 dòp-H-á-ìí  
 behind-MT-NMLS-ANAPH  
 ‘After that’



- (9) *bám*  
 H-ba-ám  
 INJ\_IPFV-catch-1SG  
 ‘Catch me’



- (6) *ʔà gáííí*  
 ʔà H-gai-VV  
 3SG IPFV-bail-DUR  
 ‘She bails (the water) for a long time’



## 2.2 Clitics

Personal pronouns have a free form when they are used as a subject and placed before the verb (cf. *ʔám* in (7)). When they are placed after any term, they cliticized on it and lose the initial consonant when it is the glottal /ʔ/ (cf. =ám in (8) and (9))

All personal pronouns are treated in this way, even those that begin with a different consonant and have the same form in all positions (2SG *mé* vs =*mé*, 3PL *wà* vs =*wà*)

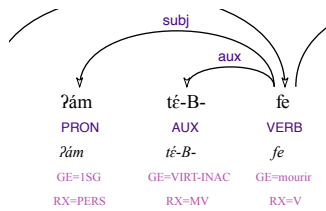
## 3 Sentence segmentation

Every verb has an overt subject, except in two constructions, relative clauses and coordination of clauses: two successive clauses that share the same subject can be coordinated without repeating the subject before the verb in the second clause. See for example sentence (10).

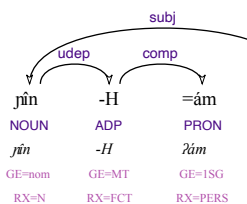
This gives us a very simple criteria to segment our spoken corpus into sentences: a sentence contains one and only one non-subordinated verb with a subject. A coordinated clause with the main clause will be attached the same sentence if and only if the verb has no subject. This follows the prescriptions of previous spoken corpora (Kahane et al., 2021).

Additionally, there’s a case of close coordination without an explicit marker: when two clauses share the same subject, the subject is omitted in the second clause. This subject omission, which goes against the norm of repeating the subject, is annotated with *conj:coord*, linking the second verb (without subject) to the first.

- (7) *ʔám té-fè*  
 ʔám té-B-fe  
 1SG VIRT-IPFV-die  
 ‘I’m going to die.’

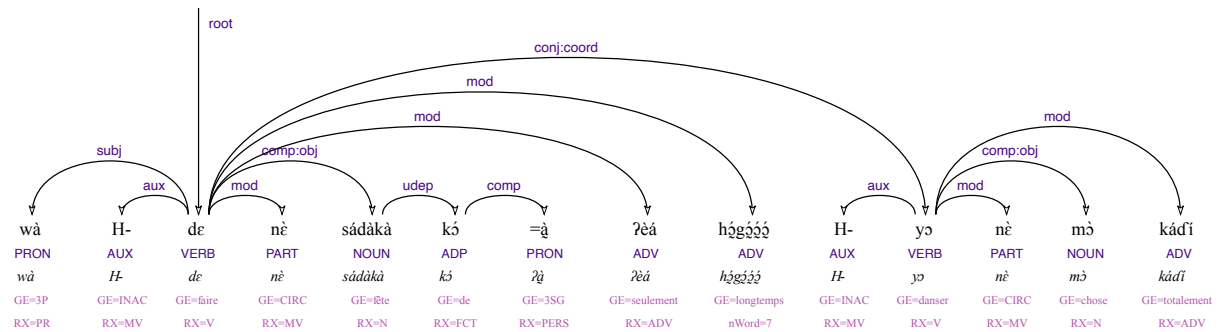


- (8) *ʔínám*  
 ʔínm-H-ám  
 name-MT-1SG  
 ‘My name’

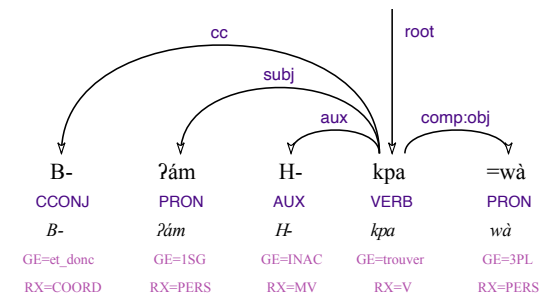


(10) wà dé nè sádàkà kɔ́à ʔéá hɔ́gɔ́ɔ́ yó nè mɔ́  
 3PL IPFV-do CIRC celebration of-3SG only for\_a\_long\_time IPFV-dance CIRC thing  
 kádí  
 wholly

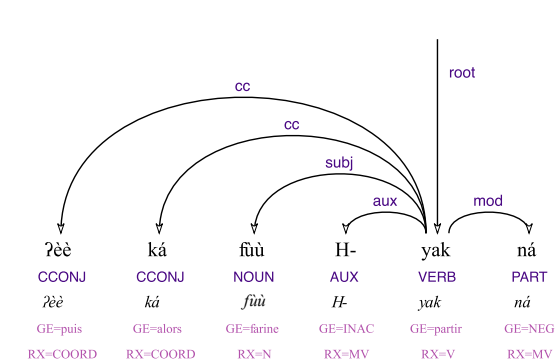
‘They just celebrate her for a long time and they dance until they’re exhausted.’



(11) ʔám kpáwà  
 ʔB-àm  
 and\_so-1SG IPFV-find-3PL  
 ‘And so I met them’



(12) ʔèè ká fùù yák ná  
 coord coord  
 then and\_then flour IPFV-leave NEG  
 ‘Then the flour didnt leave again’



## 4 Some syntactic constructions

We give an overview of several syntactic constructions, some of them unusual in the current UD collection, such as the reported speech construction.

### 4.1 Sentence coordinating conjunction

Clauses are usually coordinated, expressing the speaker’s enunciative choice. There are eight coordinating conjunctions (plus four variants), such as ʔá-nè ‘and now’, ká ‘and then’ or B- ‘and so’ a low tone on the subject personal pronoun (11). The coordinating conjunction ʔèè ‘then’ can combine with all the others coordinating conjunctions (12).

### 4.2 Sentence particles

In addition to the coordinating conjunctions, Gbaya has a system of nine enunciative particles, which are always placed at the end of a clause and serve an enunciative function by specifying the speakers point of view. We categorize them

as **PART** and introduce the relation **discourse:sp** (**sp** for sentence particle), already used for Chinese treebanks (Leung et al., 2016).

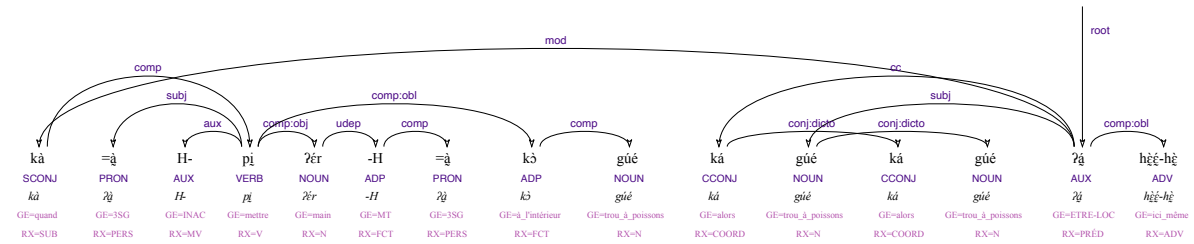
The sentence particles must be distinguished from the discourse markers, which are analyzed as interjections (**INTJ**) and receive the function **discourse** (without an extension).

### 4.3 Adverbial clauses

In SUD, subordinating conjunctions are analyzed as the head of the clause, with the verb as a **comp** (13).

In Gbaya, however, many subordinate clause relationships are not marked by a subordinating element; rather, it is the verb form itself that marks the subordination. These are the TAM-bound forms. We provide the example of the hypothetical, which renders the clause dependent on a main

- (13) *kàà pí ?éráà kò gué ká gué ká gué ?á hèè-hè*  
 when-3SG IPFV-put hand-MT-3SG in fish\_hole and\_then fish\_hole and\_then fish\_hole BE-LOC just\_here  
 ‘When she puts her hand in the fish hole, it’s a fish hole all right.’



clause that must follow (see example (14) where the subordinative clause is in blue).

#### 4.4 Relative clauses

There is no relative pronoun in Gbaya, but only one pure relativizer *nè*, which is analyzed as a subordinating conjunction. In 11 of the 18 examples of our treebank, the subject is extracted and the clause has a gap in the subject position (see (15) where the relative clause is in blue and the domain name in orange).

Object relative clauses have also a gap in the object position, but in locative relative clause there is an adverb in the extracted position. Gbaya has also an original construction with a nominalized verb as the antecedent, which must be repeated in the relative clause (16).

#### 4.5 Serial verb construction

Serial verb constructions consists of the expansion of a verb in predicative position by another verb in the infinitive (perfective, imperfective, or virtual), within the limit of a sequence of three verbs V1.TAM V2.INF [V3.INF]. In the treebank, V1 functions as the root, V2 is linked to V1 by a **compound:svc** relation (17).

#### 4.6 The reported speech

Reported speech constitutes an original construction in Gbaya, combining two clauses and two enunciations. These two clauses are interdependent neither subordinated, as in subordinative clauses, nor sequential, as in coordinated clauses. The first clause, referred to as the “quoting speech”, introduces the speaker; the second, the “quoted speech”, provides the content of the discourse. Without going into the specifics of this construction, it is important to note that the quoted speech alone is sufficient to establish reported discourse. The quoting speech is rarely verbal and

is often limited to the speakers name (eventually followed by a particle, the most common being *ndé*). The quoted speech may begin with a mention of the addressee, which can be followed by the particle *nà*. Note that *ndé* have the primary role of being a sentence particle expressing interrogation, and *nà* have the primary role of being a verb modifier expressing negation. We analyze both the speaker and the addressee with a new relation which we call **discourse:participant**. See (18) and (19).

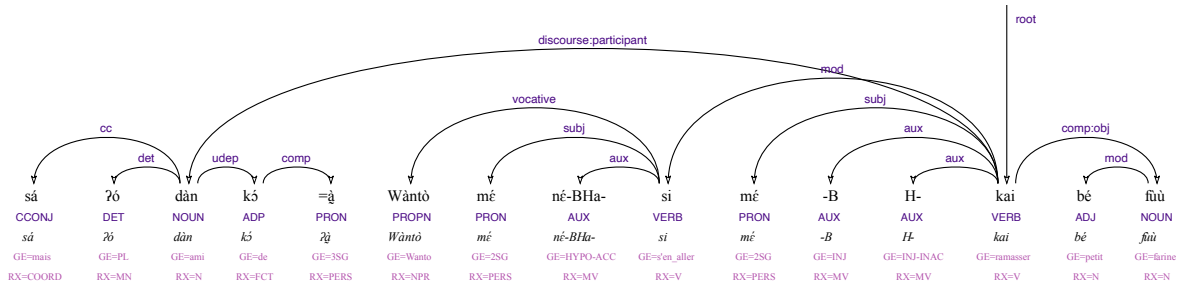
#### 4.7 Topicalization

In Gbaya, topicalization brings a noun or noun phrase (the topic) to the front of the sentence, followed without pause by the comment. When the subject occupies this initial position, specific markers (a resumptive pronoun, a topicalization particle, or both) are required to indicate its topical status. Moving a direct object or oblique complement to the front of the sentence is enough to topicalize it, sometimes reinforced by a topicalization marker. If the object is animate, a resumptive pronoun follows the verb; if it is inanimate, the position remains empty. Another construction topicalizes the verb by placing a verbal noun first, followed immediately by the same verb.

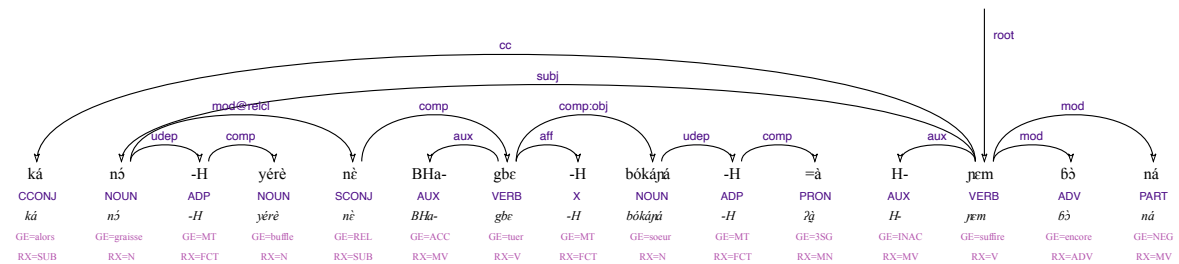
#### 4.8 Focalization

In Gbaya, noun focalization is marked by the identifiers *né* or *mè-né / mò-né*. The sentence begins with this identifier, followed by a clause in which the focused element functions as subject, object, or circumstantial complement. Such a construction exhibits a cleft extraction that splits the verbal predication into two parts. Only *né* combines with the potential verb form to express negation, unlike *mè-né*, which occurs only in affirmative clauses. Verbal focus uses *né*, functioning like an applicative suffix placed after the conjugated verb, target-

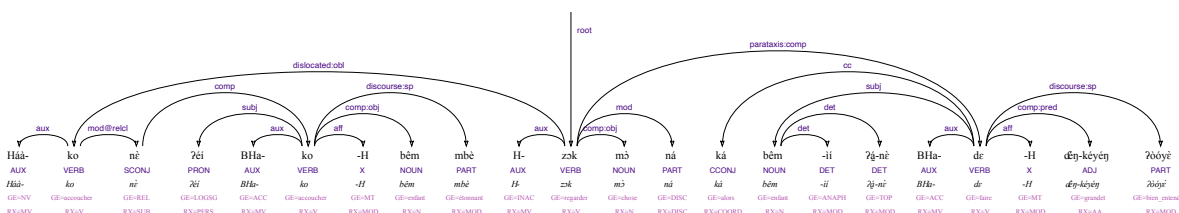
(14) *Wántò mé né-sià mè kái bé fùù*  
 NPR 2SG HYPO-PFV-return 2SG-INJ INJ\_IPFV-collect little flour  
 ‘Wanto if you return, take some flour.’



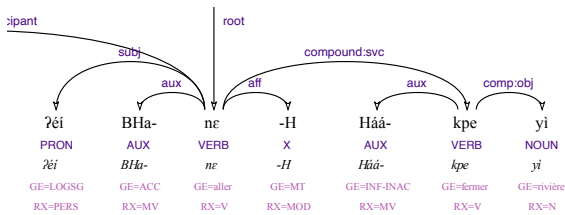
(15) *ká nó yéré nè gbèè bókápáà jém bò ná*  
 and\_then fat buffalo REL PFV-kill-MT sister-MT-3SG IPFV-suit yet NEG  
 ‘The way buffalo fat dulled his sisters senses is just unbelievable.’



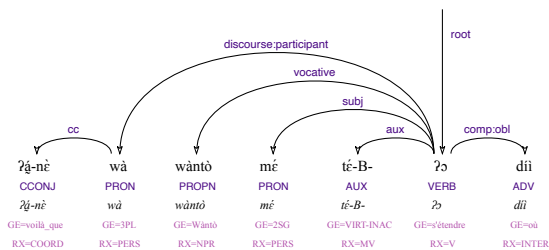
(16) *kóáà nè ǵéí kòò bém mbè zók mò ná ká*  
 NV-give\_birth REL LOGSG PFV-give\_birth-MT child amazing IPFV-see thing DISC and\_then  
*bémí ǵá-nè dèè déj-kéyè ǵóyè*  
 child-ANAPH TOP PFV-dmt tall\_(6\_8\_years\_old) of\_course  
 ‘Having surprisingly given birth to a child, she realizes that the child is of course grown up [6-8 years].’



(17) *ʔéí nèè kpéé yi*  
 V1 V2  
 LOGSG PFV-go-MT INF\_IPFV-close river  
 ‘I’ve gone to block the river (scoop fishing).’



(18) *ʔá-nè wà wántò mé té-ʔò*  
 and\_now they Wanto you lying\_down  
*dû*  
 where  
 ‘and now they (say): Wanto, where do you want to sleep?’



ing the immediately following element: the object (with transitive verbs), a circumstantial complement (with intransitives), or, if no complement follows, a verbal noun derived from the same verb.

## 5 UD version of the treebank

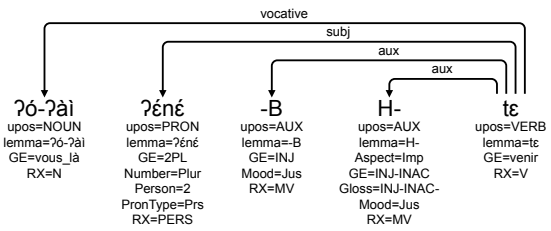
Like for the mSUD annotation, the UD version of the Gbaya treebank is produced in two steps. First, the word tokenisation is modified to match the syntactic word level expected by UD. Then the regular conversion for SUD to UD (Gerdes et al., 2018b) is used to produce the UD version.

### 5.1 Word-level tokenisation

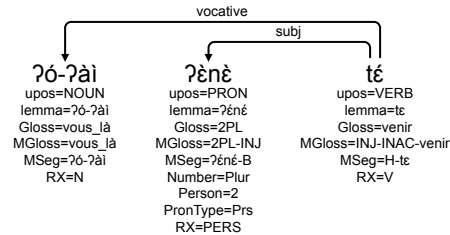
Word-level tokenisation is obtained from the original annotation by merging inflectional morphemes with the tokens to which they are attached.

There are few cases where the inflectional morpheme is not directly syntactically related to its neighbour token. In the example below, the suffix -B is not syntactically linked to the previous token *ʔéné*. We have built a dedicated heuristic to compute the necessary syntactic structure after the fusion: in the example, we want to keep the **subj**

relation in the final structure.



After affixes merging, the example above is converted into:



Note that in the word-based version, the morpheme-based analysis is indicated in the **MSeg** and **MGloss** features.

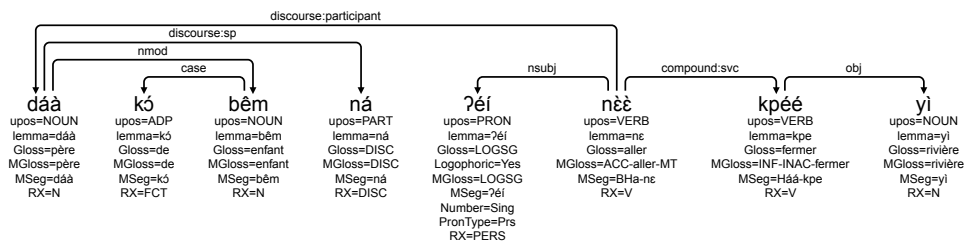
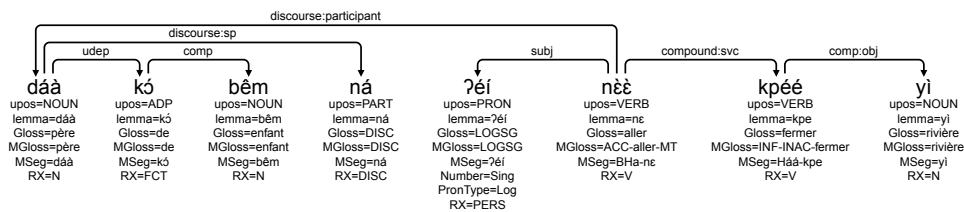
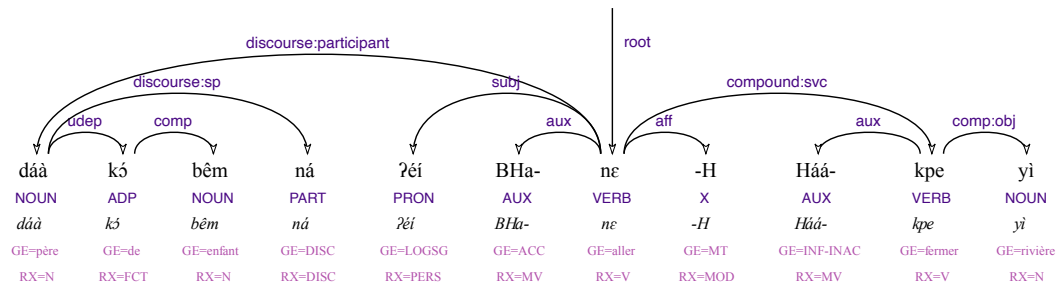
## 5.2 Conversion to UD

The annotation obtained in the previous step is not a regular SUD format. Nevertheless, the universal conversion from SUD to UD can be applied because the Gbaya word-level format is between SUD and UD: adpositions are heads, as in SUD, and auxiliaries are dependent, as in UD. Thus, only a subset of the conversion rules is activated during the conversion process. With sentence (19) the 3 annotation formats are given: first the annotated format at morpheme level, then the SUD annotation at word level and then UD annotation (also at word level).

## 6 Conclusion

The Gbaya treebank is the first treebank of a new genus within the Niger-Congo family, and only the Fourteenth treebank (along with Tswana, Yoruba and Wolof) of the greater Niger-Congo family, which comprises over 1,500 languages, including the 400 Bantu languages. Gbaya differs from Bantu languages in the absence of nominal classes and agreement rules. It has a highly developed tonal inflection system, but virtually no affixes. We have chosen to develop a morpheme-based treebank in order to highlight the elegance of this language’s grammatical system and to have a resource that combines the interlinear glosses previously developed by Paulette Roulon-Doko with

(19) *dáà ké bêm ná ?éí nèè kpéé yì*  
 father of child DISC LOGSG PFV-go-MT INF\_PFV-close river  
 ‘[She says] To the child’s father I go to block the river.’



syntactic analysis. The treebank is also distributed in a word-based UD version, where morphological information remains accessible in the **MSeg** and **MGloss** features.

## Acknowledgements

We would like to thank the three reviewers for their constructive feedback on the paper. This work has been supported by the French National Research Project Autogramm (ANR-21-CE38-0017).

## References

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

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018a. [SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD](#). In *Universal Dependencies Workshop 2018*, Brussels, Belgium.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018b. [SUD or surface-syntactic Universal Dependencies: An annotation scheme near-](#)

[isomorphic to UD](#). In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium. Association for Computational Linguistics.

Bruno Guillaume, Kim Gerdes, Kirian Guiller, Sylvain Kahane, and Yixuan Li. 2024. [Joint annotation of morphology and syntax in dependency treebanks](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9568–9577, Torino, Italia. ELRA and ICCL.

Larry M. Hyman. 2016. Lexical vs. grammatical tone: Sorting out the differences. In *Proceedings of Tonal Aspects of Language (TAL) 2016*, pages 5–11, Buffalo, New York. ISCA Archive.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021. [Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal](#). In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 35–47, Sofia, Bulgaria. Association for Computational Linguistics.

Herman Leung, Rafaël Poiret, Tak-sum Wong, Xinying Chen, Kim Gerdes, and John Lee. 2016. [Developing Universal Dependencies for Mandarin Chinese](#). In *Proceedings of the 12th Workshop on Asian*



- Language Resources (ALR12)*, pages 20–29, Osaka, Japan. The COLING 2016 Organizing Committee.
- A. E. Meeussen. 1967. Bantu grammatical reconstructions. *Africana Linguistica*, 3(1):79–121.
- Nicholas R. Rolle. 2018. [Grammatical tone: Typology and theory](#). *Berkeley Papers in Formal Linguistics*, 1(1).
- Paulette Roulon. 1975. *Le verbe gbaya, étude syntaxique et sémantique (R.C.A.)*. Number 51-52 in Bibliothèque de la SELAF. Paris. 2 cartes.
- Paulette Roulon-Doko. 1994. L'expression de la qualification (l'exemple du gbaya 'bodoe de centrafricain). In T. Geider and R. Kastenholz, editors, *Sprachen und Sprachzeugnisse in Afrika*, pages 345–356. Rüdiger Köppe Verlag, Köln.
- Paulette Roulon-Doko. 1995. Le système verbal gbaya. In R. Boyd, editor, *Le système verbal dans les langues oubanguiennes*, number 07 in LINCOS Studies in African Linguistics, pages 25–80. LINCOS, München.
- Paulette Roulon-Doko. 1998. La prédication non processive en gbaya 'bodoe. In P. Roulon-Doko, editor, *Les manières d'être et les mots pour le dire dans les langues d'Afrique Centrale*, pages 111–134. Lincos Europa, München-Newcastle.
- Paulette Roulon-Doko. 2008. *Dictionnaire Gbaya-Français (République Centrafricaine), suivi d'un dictionnaire des noms propres et d'un index français-gbaya*. Coll. Dictionnaires et langues. Karthala, Paris.

# Dative alternations in less-researched syntactic patterns of standard Croatian

Matea Birtić, Siniša Runjaić, Robert Sviben<sup>1</sup>

<sup>1</sup>Institute for the Croatian Language  
{mbirtic, srunjaic, rsviben}@ihjj.hr

## Abstract

Dative alternation in double object constructions is a frequently researched syntactic phenomenon, having been investigated across world languages. Consequently, even relatively smaller and under-resourced languages like Croatian have seen influential studies on the topic. Recent syntactic and semantic analyses of verbs in standard Croatian have identified less-explored instances of dative alternation. This contribution aims to describe the alternation between dative case and prepositional phrase for the non-agentive and intransitive uses of the verb *služiti* ('to serve'), as well as the dative alternation for the agentive and transitive uses of the verb *izbjeći* ('to avoid').

## 1 Introduction

The phenomenon of dative alternation (dative shift) closely related to *double object construction* (DOC) has a prominent role in linguistic theory, particularly within generative grammar (Chomsky 1955/1975, Larson 1988, Pesetsky 1995), though it is also explored in other linguistic frameworks (Goldberg 1995). Since Chomsky (1955/1975) and earlier it has been noted that there exists a class of English verbs that show up in two different syntactic patterns and pose a problem for syntactic theory on multiple levels (case-marking of two 'bare' objects, two syntactic patterns for the same meaning). Verbs like *give*, for instance, may appear in the canonical ditransitive pattern (*John gave a letter to Mary* > Agent[NP] V Theme[NP] Recipient[PP]) or in the DOC pattern, where two prepositionless objects are required (*John gave Mary a letter* > Agent[NP] V

Recipient[NP] Theme[NP]). The phenomenon has been extensively studied in English and has given rise to the stipulation of layered structure for the Verb Phrase (vP-shell, Larson 1988). Terminological preferences vary – some authors prefer dative shift (Larson 1988), while others opt for dative alternation (Levin and Rappaport Hovav, 2005). A similar construction has been noted in Croatian, though scholars advise against applying the term *double object construction* in this context. Instead, they refer to it as dative alternation (Zovko Dinković 2007). As a case-marking language, Croatian expresses grammatical relations and semantic roles through morphological case, and two objects of a single verb are usually marked differently. In the prototypical situation which describes a transfer of a Theme from an Agent to a Recipient, and is expressed through “give”-verbs, an Agent is realized as nominative NP, Theme as accusative NP and Recipient as dative NP. Interestingly, a small group of Croatian verbs (approximately eight) also allow an alternative syntactic pattern in which the Recipient is marked with the accusative case (as a direct object), while the Theme appears in the instrumental case. This construction resembles English dative alternation in that the Recipient can appear in two different morphological forms. In Croatian Recipient alternates between indirect and direct object as in 1a-1b.<sup>1</sup>

1a)			
Lena	je poslužila	gostima	čaj.
Lena	serve <sub>3SG.PAST</sub>	guests <sub>DAT.PL</sub>	tea <sub>ACC.SG</sub>
'Lena served tea to the guests.'			
1b)			
Lena	je poslužila	goste	čajem
Lena	serve <sub>3SG.PAST</sub>	guests <sub>ACC.PL</sub>	tea <sub>INST.SG</sub>
'Lena served guests the tea.'			

<sup>1</sup> The same is sometimes claimed for English too (Larson 1998).

Zovko Dinković (2007) points out that the dative alternation in Croatian serves to increase the affectedness of the Recipient which is achieved by shifting it to the direct object position i.e. by using the accusative case. Belaj and Tanacković Faletar (2017) further argue that this syntactic shift serves to topicalize and focalize the Recipient, and also suggest that the instrumental-marked NP can carry two semantic roles: Theme and Instrument. In the pattern with the dative case in its prototypical semantic role of Recipient, the construction is unmarked. In contrast, the alternation where the direct object in the accusative case takes a semantic role of a Recipient, accompanied by the instrumental-marked Theme, is syntactically and stylistically marked. Various theoretical accounts have been proposed to explain these alternations. Some researchers argue that the constructions differ semantically and thus correspond to two distinct verb entries with different subcategorization frames (Oehrle 1976, according to van Gelderen 2013), while others claim that one construction is derived from the other (Larson 1988, Baker 1997), assuming a single lexical entry with approximately the same semantic roles.

## 2 Analysis of the non-agentive intransitive use of the verb *služiti* ('to serve')

Introductory analysis of the ditransitive dative alternation of the verb *služiti* ('to serve') in Croatian, where a direct object (Theme) in the accusative case and an indirect object (Recipient) in the dative case (pattern Agent[NP\_Nom] V Theme[NP\_Acc] Recipient[NP\_Dat]) can alternate with relatively similar constructions featuring a direct object (Recipient) in the accusative case and an indirect object (Theme) in the instrumental case (pattern Agent[NP\_Nom] V Recipient [NP\_Acc] Theme[NP\_Inst]), pertains to the agentive use of this verb in its primary meaning. This meaning can be broadly defined as 'to offer something to guests or clients' and could semantically be classified within the "give" verbs (according to Levin 1993).

However, prior analyses have focused exclusively on the agentive ditransitive construction within the semantic group of "give"-verbs. The verb *služiti* ('to serve') is polysemous, and its meaning as a "give"-verb – along with this agentive ditransitive usage – is not its only one. A relatively common non-agentive use exists, meaning 'to be

serviceable or suitable for something', as in the following grammatical sentence 2a):

2a)  
Sport služi za jačanje  
svijesti.  
Sport serve<sub>3SG</sub> for strengthening<sub>ACC.SG</sub>  
awareness<sub>GEN.SG</sub>

'Sport serves to strengthen awareness.'

On the level of syntactic realization, we can formally represent this realized pattern as Theme[NP\_Nom] V Purpose[PP\_za+Acc], seemingly without an expressed semantic role of Recipient. This prepositional phrase (*za* 'for' + *jačanje*<sub>ACC.SG</sub> 'strengthening') is interchangeable with a noun phrase in the dative case (*jačanju*<sub>DAT.SG</sub>), while the sentence remains grammatical and retains the same meaning, as in

2b):

2b)  
Sport služi jačanju  
svijesti.  
Sport serve<sub>3SG</sub> strengthening<sub>DAT.SG</sub>  
awareness<sub>GEN.SG</sub>

'Sport serves to strengthen awareness.'

The semantic role of Purpose, which is prototypically realized in (2a) by the prepositional phrase *za* 'for' + accusative NP, is realized in (2b) with the dative case (Theme[NP\_Nom] V Purpose[NP\_Dat]). The dative case is more commonly associated with the semantic roles of the Recipient, but in this instance, the appearance of the dative NP does not alter the semantic role of Purpose.

We hypothesize that the reason for this lies in the fact that, in such sentences, the Recipient of the action of serving a purpose for someone is not overtly expressed in the sentence. Semantically, at a logical argument-structure level of the event, an action in which no one benefits from the notion of "serving" is impossible. Thus, it can be presumed that, in its full form (Theme[NP\_Nom] V Recipient[NP\_Dat] Purpose[PP\_za+Acc]), the sentence is as in (2c):

2c)  
Sport služi državi za  
jačanje svijesti.  
Sport serve<sub>3SG</sub> state<sub>DAT.SG</sub> for  
strengthening<sub>ACC.SG</sub> awareness<sub>GEN.SG</sub>

'Sport serves the state to strengthen awareness.'

The alternation between (2a) and (2b) could therefore be explained as an instance of grammatical metonymy (Ruiz de Mendoza Ibáñez

& Pérez Hernández, 2001: 334). Specifically, since an essential participant in the action (in this case, the Recipient in the dative case) is unexpressed as an argument, the next argument in the structure (though it fulfils a completely different semantic role) appears in unexpected grammatical form due to recategorization, replacing the prototypical prepositional phrase. Thus, this alternation between the dative case and the prepositional phrase in the meaning of the imperfective verb *služiti* (and its perfective counterpart *poslužiti*) is both possible and grammatical.

### 3 Dative alternations in specific group of motion verbs

In Section 2, it was shown how the dative case alternates with a prepositional phrase, covering the semantic role of Purpose in the intransitive syntactic pattern when an important member (Recipient) of the argument structure in non-agentive usage was not expressed. In this chapter, we will analyze in more detail the uses of the verb *izbjeci* ('avoid, escape', perfective), a polysemous verb that is relatively frequent in the Croatian language. At the surface level, we can say that this is a motion verb, but upon deeper classification, the first two distinct meanings identified in the analysis of examples (3a-3b) and (4a-4b) fall under "avoid"-verbs (according to Levin 1993). In the first distinct meaning, which we can define as 'to do everything to prevent anything unpleasant or unwanted from happening', a possible alternation between accusative and dative case in the same structural position was observed without a change in the semantic role. Thus, in example (3a):

3a)  
Taj krvolok        izbjegao je        progon.  
That bloodsucker escape<sub>3SG.PAST</sub>    persecution<sub>ACC.SG</sub>  
'That bloodsucker escaped persecution.'

The prototypical pattern is transitive and has the Theme marked with the accusative case (Agent[Nom] V Theme[Acc]), but the Theme can also appear in the dative case (Agent[Nom] V

Theme[Dat]), which is not prototypical role for dative case.

3b)  
Taj krvolok        izbjegao je        progonu.  
That bloodsucker escape<sub>3SG.PAST</sub>    persecution<sub>DAT.SG</sub>  
'That bloodsucker escaped persecution.'

The same valency patterns were observed in the analysis of examples for the second meaning, defined 'to move suddenly from current location' ("avoid"), as in (4a) and (4b):

4a)  
On        će izbjeci        kamen.  
He        avoid<sub>3SG.FUT</sub>        stone<sub>ACC.SG</sub>  
'He will avoid the stone.'

4b)  
On        će izbjeci        kamenu.  
He        avoid<sub>3SG.FUT</sub>        stone<sub>DAT.SG</sub>  
'He will avoid the stone.'

We emphasize that the analysis of this dative alternation for single-object "avoid"-verbs cannot be entirely identical to the analysis of dative shift in double-object constructions for "give"-verbs. Although the prototypical uses in both meanings involve the Theme in the accusative case, we believe that alternation with the dative is possible due to a combination of two reasons:

1. The argument with a dative case in south Slavic languages is usually used with certain types of motion verbs<sup>2</sup>, making it somewhat ingrained in speakers' usage memory.
2. The alternation is enabled by the process of conceptualizing the object in the sentence: while the prototypical accusative usage focuses on the event or entity being avoided as a whole, the dative shifts the focus more toward the direction or goal of avoidance.

The third distinct meaning, 'to go in the opposite direction from where a certain danger is coming' ("flee, fly, escape"), is slightly different from the first two because the English translation equivalents would not apply to the previous two meanings, and we would not primarily classify that meaning semantically as an "avoid"-verb, but rather as an "escape"-verb (according to the VerbNet (Class Hierarchy), Kipper Schuler and Palmer 2005). Regardless of this semantic shift,

<sup>2</sup> A detailed study of dative arguments with verbs of motion in South Slavic languages can be found in Palić 2010, pp. 239-267.

the alternation between the accusative and dative in examples (5a) and (5b) could be analyzed in the same way as the alternations in examples (3a-3b) and (4a-4b).

5a)

Stanovnici sela	izbjegli su	rat.
Villagers	escape <sub>3PL.PAST</sub>	war <sub>ACC.SG</sub>

‘Villagers escaped the war.’

5b)

Stanovnici sela	izbjegli su	ratu.
Villagers	escape <sub>3PL.PAST</sub>	war <sub>DAT.SG</sub>

‘Villagers avoided the war.’

However, due to the semantic difference from the first two meanings, another syntactic alternation appears, as in example (5c), where a prepositional phrase is used (something not expected for the first two described meanings).<sup>3</sup>

5c)

Stanovnici sela	izbjegli su	od	rata.
Villagers	escape <sub>3PL.PAST</sub>	from	war <sub>GEN.SG</sub>

‘Villagers fled from the war.’

In this case, we would not draw a parallel between the dative alternation in (5b) and the prepositional phrase in (5c) by explaining it through the process of grammatical metonymy, as in examples (2a-2b) for the verb *služiti* ('serve'), but rather through semantic proximity to prepositional semantics because *od* ('from') in (5c) introduces a spatial or abstract separation, which can overlap with the directional or goal-oriented meaning of the dative case when used with motion verbs in general. The small semantic difference can be related to perspective: the dative NP may focus more on the direction, while the PP focuses more on the source or the process of separation.

#### 4 Conclusion and outlook

This paper presents initial analyses of less frequent and rarely described dative alternations in the Croatian language. While the phenomenon of dative alternation with ditransitive verbs semantically belonging to the “give”-verb group has been precisely described (Zovko 2007, Belaj and Tanacković Faletar 2017), this contribution focuses on other patterns and examples. In the first part of the analysis, we described the possible

dative alternation with a prepositional phrase in the non-agentive use of the verb *služiti* ('serve'), even though it cannot be considered part of the “give”-verb group. In the second part, we examined different levels of dative alternations across all three distinct meanings of the frequent motion verb *izbjeći* ('avoid').

By comparing the conclusions from both analyses, we observed that:

1. The cause of the syntactic alternation between dative and accusative in the examples with the (mono)transitive verb *izbjeći* ('avoid') significantly differs from the cause of this phenomenon in the agentive use of the ditransitive verb *služiti* ('serve'). However, a key similarity is that, in both cases, the alternation arises due to focalization and the topicalization of objects in the sentence.
2. The phenomenon of grammatical metonymy emerged as a logical cause for the syntactic alternation between the dative and prepositional phrase ('*za*' + Acc) in the non-agentive use of *služiti* ('serve'). This process differs significantly in speakers' cognition from the cause of the alternation between the dative and prepositional phrase ('*od*' + Gen) in the examples for the third meaning of *izbjeći* ('avoid'). Here, the dative case and prepositional phrase alternate due to the semantic similarity between the source preposition *od* ('from') and the prefix '*iz-*' in the prefixed perfective verb *izbjeći* (both the preposition and the prefix carry the same meaning 'from').

These are fundamental and preliminary analyses obtained using traditional methods on small corpus samples. It would be preferred to develop an automated method for extracting this type of argument structure alternation and similar patterns from corpora for the purposes of further, diverse linguistic analyses. However, existing syntactically and, especially, semantically annotated treebanks for Croatian are limited in size and genre-specific, which poses significant challenges for conducting large-scale quantitative studies. Moreover, current

<sup>3</sup> A search of the hrWaC corpus (Ljubešić and Klubička 2014) confirmed that this alternation also applies to less frequent synonyms of that particular sense of *izbjeći*, such as *odbjeći*, *umaći/umaknuti*, or *izmaći/izmaknuti*. These are relatively low-frequency synonyms, all also perfective in

aspect with the core semantics of avoidance, but there is no evidence of the described alternations in the aspectual counterpart *izbjegavati* (imperfective). Thus, it can be assumed that in future research, verbal aspect will continue to play an important role.

frameworks, such as Universal Dependencies, continue to face difficulties in distinguishing between adverbial arguments and adjuncts, as well as in providing sufficient semantic role annotation. For analyses of this kind, it is probably necessary to introduce additional layers of grammatical representations – for instance, Enhanced Dependencies (Schuster and Manning 2016) – into larger treebanks. A possible direction for future work is therefore to explore ways in which the less-researched syntactic patterns of argument structure in Croatian could be systematically identified and described – for example, by applying the method of neutralizing argument alternations (Candito et al. 2017) to Croatian treebanks.

### Acknowledgments

This work has been supported by the Croatian Science Foundation under the projects *Semantic-Syntactic Classification of Croatian Verbs* (SEMTACTIC) (HRZZ-IP-2022-10-8074) and *Croatian Prepositions in Use – Semantic and Syntactic Analysis* (HRPA) (HRZZ-IP-2022-10-6867).

### References

- Mark C. Baker. 1997. Thematic roles and syntactic structures. In *Elements of grammar: Handbook in generative syntax*. Kluwer Academic Publishers, pages 73-137.
- Branimir Belaj and Goran Tanacković Faletar. 2017. *Kognitivna gramatika hrvatskoga jezika: sintaksa jednostavne rečenice*. Disput, Zagreb.
- Marie Candito, Bruno Guillaume, Guy Perrier and Djamel Seddah. 2017. Enhanced UD Dependencies with Neutralized Diathesis Alternation. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 42-53, Pisa, Italy. Linköping University Electronic Press.
- Noam Chomsky. 1975. (1955.) *The logical structure for linguistic theory*. Springer Publishing, New York, NY.
- Elly van Gelderen. 2013. *Clause structure*. Cambridge University Press, Cambridge, UK.
- Adele E. Goldberg. 1995. *Constructions: a construction grammar approach to argument structure*. The University of Chicago Press, Chicago, IL.
- Karin Kipper Schuler and Martha S. Palmer. 2005. *Verbnet: a broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.
- Richard K. Larson. 1988. On the double object construction. *Linguistic Inquiry*, 19(3): 335-391. <https://www.jstor.org/stable/25164901>.
- Beth Levin. 1993. *English verb classes and alternations*. The University of Chicago Press, Chicago, IL.
- Beth Levin and Malka Rappaport Hovav. 2005. *Argument realization*. Cambridge University Press, Cambridge, UK.
- Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr} WaC – Web Corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9<sup>th</sup> Web as Corpus Workshop (WaC-9)*. Association for Computational Linguistics, pages 29-35.
- Francisco J. Ruiz de Mendoza Ibáñez and Lorena P. Hernández. 2001. Metonymy and the grammar: motivation, constraints and interaction. *Language & Communication*, 21(4): 321-357. <https://www.sciencedirect.com/science/article/pii/S0271530901000088>.
- Richard T. Oehrle. 1976. *The grammatical status of the English dative alternation*. Doctoral dissertation, MIT, Cambridge, MA.
- Ismail Palić. 2010. *Dativ u bosanskom jeziku*. Bookline, Sarajevo.
- David Pesetsky. 1995. *Zero syntax*. MIT Press, Cambridge, MA.
- Sebastian Schuster and Christopher D. Manning. 2016. Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2371-2378, Portorož, Slovenia. European Language Resources Association (ELRA).
- Irena Zovko Dinković. 2007. Dative alternation in Croatian. *Suvremena lingvistika*, 33/63: 65-83.



# Distance and Projectivity as Predictors of Sentence Acceptability in Free Word Order Languages

**Kirill Chuprinko**

Center for Cognitive Science  
of Language

University of Nova Gorica  
Vipavska 13, 5000 Slovenia

kirill.chuprinko@ung.si

**Artem Novozhilov**

Center for Cognitive Science  
of Language

University of Nova Gorica  
Vipavska 13, 5000 Slovenia

artem.novozhilov@ung.si

**Arthur Stepanov**

Center for Cognitive Science  
of Language

University of Nova Gorica  
Vipavska 13, 5000 Slovenia

arthur.stepanov@ung.si

## Abstract

This study investigates how two core metrics rooted in Dependency Grammar, Mean Dependency Distance (MDD) and projectivity, predict sentence acceptability in Russian and Serbo-Croatian. Using exhaustive word order permutations in controlled five-word sentences, we model how these metrics relate to acceptability judgments in two psycholinguistic experiments. While MDD has been widely studied as a processing constraint, projectivity violations have received less attention in experiments, and particularly in acceptability modeling. We demonstrate that both metrics have a significant independent impact on judgments, with projectivity playing a surprisingly strong role. In addition, Serbo-Croatian's rigid clitic placement provides a natural test case for disentangling grammatical from processing constraints. Our findings offer a computationally precise, dependency-based model of acceptability that advances cognitively grounded language modeling for free word order languages.

## 1 Introduction

Sentence acceptability reflects how natural or well-formed a sentence appears to native speakers, bridging linguistic competence and real-world performance (Chomsky, 1965). While judgments of acceptability are shaped by multiple factors such as semantic plausibility, discourse coherence, and real (for listening mode) or potential prosody, they are fundamentally influenced by two key forces: i) grammaticality (conformity to internalized rules of grammar) and ii) processing load (constraints arising during sentence comprehension and production). Cognitively informed models of sentence acceptability must capture both of these dimensions. Yet much of the current research, particularly in the evaluation of neural language models, tends to conflate them or focus on surface-level performance metrics (Warstadt et al., 2019; Zhang et al., 2024). This limits our understanding of the underlying mechanisms that drive acceptability.

One well-established processing constraint is *dependency distance*, rooted in the framework of Dependency Grammar (Mel'čuk, 2009). Prior work based on Universal Dependencies (Futrell et al., 2015; Choi, 2007; Ros et al., 2015) demonstrated a general tendency for speakers for shorter syntactic dependencies across languages. This is formalized in the "Minimize Dependency Distance" principle (MDDP) and operationalized in Mean Dependency Distance (MDD) as a metric. MDD quantifies how far apart syntactically related words appear in linear order. Increased distance is thought to increase processing cost and, by extension, reduce acceptability. This naturally aligns with memory-based theories of sentence processing (Gibson, 1998, 2000).

However, several questions remain open. First, how does MDD interact with grammatical constraints, especially in languages with relatively free word order? Second, is MDD the only processing-related factor influencing acceptability, or are other surface structural properties, such as projectivity violations, where dependency arcs cross (Testelefs, 2001; Gildea and Temperley, 2010; Liu et al., 2017; Yadav et al., 2020, 2022)? While MDD has been well integrated into psycholinguistic models of sentence processing, projectivity remains largely underexplored in experimental linguistics, especially in the context of acceptability judgments.

This study addresses these gaps by systematically modeling sentence acceptability across all possible word order permutations in five-word sentences in Russian and Serbo-Croatian. The choice of these two languages is not accidental. Both languages allow documented high word order flexibility, with some theoretical studies claiming that any permutation of words in a clause is acceptable (Kallestinova, 2007; Stjepanović, 1999). At the same time, only Serbo-Croatian poses a hard grammatical constraint on word order: clitics must be in the second position. In both languages, permuting words in a sentence leads to variation in syn-

tactic dependency lengths which can be captured by MDD but also be gauged by other word-order related metrics. By systematically varying word order while controlling for lexical and structural factors, the role of different dependency metrics can be isolated. Furthermore, a comparison of sentence acceptability profiles in these two otherwise close languages allows us to disentangle the grammatical factor from processing effects.

We evaluate several dependency-based metrics including MDD, projectivity violations and a number of other structural and processing complexity measures, on their ability to predict sentence acceptability. Our results show that both MDD and projectivity violations significantly contribute to acceptability ratings, with projectivity playing a stronger role than previously assumed. These findings offer a computationally precise, cognitively grounded model of acceptability that applies core principles of Dependency Grammar to a psycholinguistic context, while deepening our understanding of word order preferences in free word order languages.

## 2 Minimize Dependency Distance Principle and Its Measures

The MDDP is rooted in cognitive constraints associated with working memory and retrieval processes during sentence comprehension and production. The key idea is that syntactically related words should be placed closer together to facilitate efficient processing and reduce cognitive load.

The first intuitions suggesting the existence of a cognitive mechanism responsible for favoring shorter dependencies in sentence structure can be traced back to the early 20th century. In his descriptive study, Behaghel (1909) noted that in German ditransitive constructions, longer noun phrases typically follow shorter ones. This observation can retrospectively be attributed to the MDDP (Staub et al., 2006).

The second major step toward the formulation of the principle came from psycholinguistic research in the 1970s and 1980s, when psychologists (Perfetti and Lesgold, 1977; Daneman and Carpenter, 1980) showed that comprehension difficulty is affected by the amount of information that needs to be actively maintained and retrieved, depending on working memory capacity. These findings provided the empirical foundation upon which the principle was explicitly formulated in the 1990s in a range of

grammatical and cognitive approaches (Hawkins, 1994; Hudson, 1995; Gibson, 1998).

Since its formal articulation in the 1990s, the principle has received substantial empirical and theoretical support from studies on sentence processing, working memory, and syntactic dependency structures (Gibson, 2000; Ferrer-i Cancho, 2004; Choi, 2007; Liu, 2008; Futrell et al., 2015). One of the most influential formulations is Gibson’s Dependency Locality Theory (DLT) (Gibson, 2000), which posits that processing difficulty increases with the linear distance between syntactically dependent elements, as longer dependencies demand more memory resources to maintain. This perspective is further elaborated by retrieval-based approaches to parsing, which argue that sentence processing involves cue-based retrieval from memory. According to Lewis and Vasishth (2005), greater syntactic distance increases the likelihood of interference and retrieval failure, thereby raising processing cost. Minimizing dependency length, therefore, enhances the accessibility of syntactically related words, reduces interference, and facilitates more efficient parsing (Grodner and Gibson, 2005; Lewis and Vasishth, 2005).

While the conceptual foundation of the MDDP is widely accepted, its mathematical operationalization varies across studies. Researchers propose different ways to quantify dependency length. For example, Gibson (2000), working in the context of phrase structure grammar, computes the incremental integration cost of each dependency as the number of intervening discourse-referent words and adds this to the concurrent storage cost (i.e. the number of yet-to-be-resolved dependencies being held in memory). Within the dependency grammar tradition, it is more common to use the mean dependency distance as proposed by Liu (2008), which averages the absolute linear distances between heads and dependents. A related, though less widespread, approach comes from Ferrer-i Cancho (2004), who suggests using the mean Euclidean distance, defined as the average of  $\sqrt{(P(\text{head})_i - P(\text{dep})_i)^2}$  – a method more common in computational linguistics (Futrell et al., 2020). Although Ferrer-i-Cancho’s approach involves squaring and taking the square root of the difference, in one-dimensional space (i.e., linear word order), it produces the same results as Liu’s simpler absolute-value method.

For this reason, in this work we adopt a uniform and computationally straightforward definition of

MDD (see above) along the lines of Liu (2008), calculated as follows:

$$\text{MDD} = \frac{1}{n-1} \sum_{i=1}^n |P(\text{head})_i - P(\text{dep})_i| \quad (1)$$

where  $n$  is the number of words in the sentence and  $P$  denotes the position of a word in the linear sequence.

### 3 Russian and Serbo-Croatian Word Order

Word order in Slavic languages is often described as relatively free, with variations influenced by both grammatical constraints and processing considerations. The canonical (i.e., most frequent or “default”) word order in both Serbo-Croatian and Russian is Subject–Verb–Object (SVO) (Urošević et al., 1986; Bailyn, 1995). Adjectives typically precede nouns in both languages. Additionally, while Russian employs almost no tense auxiliaries except for the Future Imperfective form of ‘to be,’ Serbo-Croatian extensively uses the auxiliary ‘to be’ to form both past and future tenses. In both languages, auxiliaries follow the subject and precede the verb. A key difference, however, is that in Serbo-Croatian, tense auxiliaries are clitics obeying Wackernagel’s Law, meaning they must always appear in the second position in the sentence (Bošković, 2001). However, the canonical sentence structure in both languages is the same: Subj Aux Verb [Adj Obj] (Bailyn, 1995; Bošković, 2001; Bošković, 2005).

How does the MDD fit into this picture? Suppose we define syntactic dependencies as follows: Verb–Auxiliary, Verb–Subject, Verb–Object, Object–Adjective (de Marneffe et al., 2014), or alternatively, Auxiliary–Verb, Auxiliary–Subject, Verb–Object, Object–Adjective (Groß and Osborne, 2015). From the perspective of processing cost (see above), a reasonable hypothesis is that, for a given sentence, the greater the MDD, the lower its acceptability rating. However, this reasoning does not take into account independent grammatical constraints on word order. Due to the strict second-position requirement for clitics, any deviation from Wackernagel’s Law in Serbo-Croatian (but not in Russian) would cause a downgrade in acceptability, regardless of the MDD. Because the categorical second-position rule and the gradient, dependency-based memory pressures captured by MDD pull in different directions in Serbo-Croatian,

but not in Russian, the two languages jointly provide an ideal testbed for disentangling how rigid grammatical constraints and processing costs on syntactic dependencies shape word-order acceptability.

## 4 Other Word-Order Related Metrics

As pointed out above, the MDD is not the only metric that can account for patterns in acceptability judgments in the so-called “free word order” languages. The literature proposes several additional ways to quantify how much a sentence deviates from its canonical order, yet none has been systematically evaluated in an experimental setting. In this work we explore five additional metrics: *Number of Displaced Words*, *Total Path of Displaced Words*, *Number of Projectivity Violations*, *Word Order Penalty Score* and *Special Status Residuals and Processing Penalties*. These metrics were developed and/or adapted for the present study drawing on insights from a wide range of psycholinguistic and syntactic literature. Each of these metrics may explain high or low acceptability through different mechanisms and principles embedded in syntactic parsing. A key criterion in their selection was ensuring that none of the metrics exhibited a correlation higher than 0.5 with any of the others, thereby maintaining their independence in terms of explanatory power.

### 4.1 Number of Displaced Words

This metric is our operational adaptation of the optimality-theoretic model proposed by Kallestinova (2007). While Kallestinova did not formulate an explicit metric, her analysis draws on the Linearity-IO constraint (McCarthy and Prince, 1995), which penalizes deviations from canonical word order. Within this framework, each displacement from the base order is interpreted as a violation that increases processing cost and reduces acceptability.

To translate this into a measurable form, we defined the metric as the number of displaced words relative to the canonical [Subj Aux V Adj Obj] order. For instance, [Subj Aux V Adj Obj] incurs 0 displacements, while [Subj Aux Obj Adj V] incurs 2, as both the object and the verb are misaligned with their canonical positions. The metric does not consider the direction or length of the displacement – only the number of misordered elements.

## 4.2 Total Path of Displaced Words

This metric extends Kallestinova’s approach by focusing exclusively on the linear distance of displacement. The rationale for this metric lies in the idea that longer displacements require greater cognitive effort for processing and greater distortion from canonicity thereby reducing sentence acceptability.

The metric evaluates the cumulative linear distance of displacements relative to the canonical [Subj Aux V Adj Obj]. For each displaced element, the distance is calculated as the absolute difference between its ordinal position in the canonical order and its actual position in the sentence. For instance, given the canonical order Subject (1), Aux (2), Verb (3), Adjective (4), Object (5), a sentence like "Subject Aux Verb Object Adjective" would involve the following calculation:  $|4 - 5| + |5 - 4| = 2$ . Thus, the higher the total score, the lower the acceptability rating.

## 4.3 Number of Projectivity Violations

This metric is rooted in dependency grammar and functional approaches to language (Liu et al., 2017). Dependency grammar operates under four fundamental rules as outlined by Robinson (1970):

1. One and only one element is independent.
2. All other elements depend directly on some element.
3. No element depends directly on more than one other.
4. If A depends directly on B and some element C intervenes between them in the linear order of the string, then C must depend directly on A, B, or another intervening element.

The fourth rule defines projectivity, stipulating that dependency arcs must not cross each other or the root node.

Here we followed the Dependency grammar principles applying two separate approaches, as in MDD calculations: one assuming the verb as the root node, based on de Marneffe et al. (2014), and the other assuming tense auxiliary as the root node, as in Groß and Osborne (2015). In our dependency calculation process, each crossing of a dependency arc is counted as one violation (Type I / strong violation) (Lu et al., 2016; Testelets, 2001). Similarly, any crossing between an arc and the root node is counted as one violation (Type II / weak violation). The final metric is the sum of both violation types.

To illustrate, in Figure 1, an auxiliary-root struc-

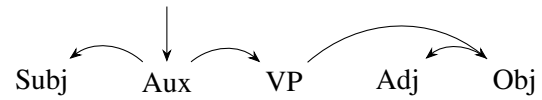


Figure 1: An example of a sentence structure with no projectivity violations.

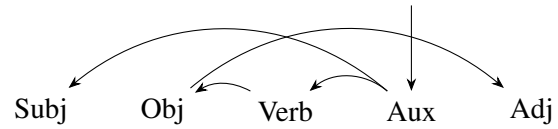


Figure 2: An example of a sentence structure with strong and weak projectivity violations.

ture adheres to projectivity constraints, resulting in zero violations. In Figure 2, the arcs crossing introduces one strong violation, while the crossing between the arc and the root node (Aux) adds one weak violation, totaling 2 violations. Again, a higher number of violations corresponds to a lower acceptability rating.

## 4.4 Word Order Penalty Score

Building on the work of (Urošević et al., 1986), the Word Order Penalty Score metric quantifies the cognitive cost of deviations from the canonical SVO structure in Slavic languages. Unlike metrics based on projectivity violations, this metric focuses on the linear disruption of canonically adjacent syntactic pairs, regardless of whether their dependency arcs cross others or the root. Experimental findings show that noncanonical word orders (e.g., VSO, OSV) are generally processed more slowly than canonical ones. This suggests that unlicensed deviations from SVO incur additional processing costs. Urošević and colleagues (1986) attribute this to the rarity of verb- or object-initial sequences in Serbian. Similarly, a corpus-based study by Slioussar and Makarchuk (2022) shows that Russian exhibits a comparable word order frequency distribution, allowing conclusions from Serbo-Croatian to extend to Russian.

The Word Order Penalty Score metric formalizes these observations by assigning penalties to deviations from canonical SVO structure. Sentences retain 0 points if they follow SVO, while noncanonical orders are penalized: object-before-subject (+1), verb-before-subject (+1), object-before-verb (+1). Additional penalties apply for disrupting (an occurrence of an intervening element inside the pair) syntactic units such as noun-modifier pairs or auxiliary-verb combinations (+1 each), as these



interruptions increase cognitive cost and memory load by breaking expected grouping patterns (Sek-  
erina, 1999). Thus, the Word Order Penalty Score  
assigns cumulative penalties for each structural and  
group-level deviation:

- Object preceding subject: +1 point
- Verb preceding subject: +1 point
- Object preceding verb: +1 point
- Adjective + Object disruption: +1 point
- Auxiliary + Verb disruption: +1 point

#### 4.5 Special Status Residuals and Processing Penalties

This metric integrates elements of formal syntax and psycholinguistics. One frequently observed factor influencing the acceptability of a given word order is its ontological status within the grammar of a particular language, that is, whether it is recognized as a distinct grammatical phenomenon rather than a surface variant. Formal grammatical approaches and empirical studies of word order variability (e.g. Bader and Meng, 1999; Miyamoto and Takahashi, 2002; Sekerina, 1999; Hyönä and Hujanen, 1997) have been especially effective in identifying such configurations. Building on this work, we focus on four phenomena relevant for Slavic languages: (1) Left-Branch Extraction (LBE) in wide sense (extracting Adjective to any leftward position), (2) Noun Scrambling (Object Displacement), (3) violations of canonical clitic or auxiliary placement, captured as NotClitic2 for Serbo-Croatian and NotBudet2 for Russian, and (4) verb topicalization, reflected in the NotSV metric, which detects disruptions of subject-verb adjacency.

A key observation here is that the special discourse status of these constructions may persist even outside of context, allowing them to be perceived as marked permutations and to receive higher acceptability ratings in comparison with their unmarked permuted counterparts. However, when multiple such operations co-occur, processing difficulty increases substantially and tends to override any residual interpretive coherence, resulting in strong acceptability penalties (Novozhilov et al., 2025). This makes it possible to use this metric as a proxy for interactions between processing cost and discourse licensing, capturing how far constructions deviate from both canonical structure and context-sensitive grammatical norms.

The metric assigns penalties to quantify the cumulative processing costs associated with these deviations. Penalties are applied as follows:

- NotBudet2: If present, +1 point; if absent, 0 points.
- Noun Scrambling: If present, +1 point; if absent, 0 points.
- NotSV: If present, +1 point; if absent, 0 points.
- LBE: If present, +1 point; if absent, 0 points.

## 5 Experiment 1 (Russian)

### 5.1 Participants

79 adult native Russian speakers (mean age = 30.1) took part in Experiment 1. Participation was voluntary and uncompensated.<sup>1</sup>

### 5.2 Materials and Procedure

Sentences were constructed by producing all 120 word order permutations of a five-word kernel structure in canonical order: Subj Aux V Adj Obj. The analytical future tense auxiliary *budet* 'be.3SG' / *budut* 'be.3PL' was used in all sentences. The subject was a noun in the nominative case, and the adjective unambiguously modified the direct object. Subject and object features were systematically varied in animacy, gender, and number. For instance, if the subject was animate, feminine, and singular, the object was inanimate, masculine, and plural, and vice versa. 120 lexical content variants (lexicalizations) were created.

Lexicalizations were distributed across six experimental lists, each containing 20 word orders.<sup>2</sup> Each order was represented by six different lexicalizations, totaling 120 sentences per list. The distribution was randomized.

Participants accessed the experiment via the PCIbex platform (Zehr and Schwarz, 2018). Before starting, they completed a brief demographic questionnaire (age, gender, education level, native language) and signed consent forms. Participants rated the sentences for acceptability on a 5-point

<sup>1</sup>The Russian and Serbo-Croatian experiments were approved by the Ethics Committee of the University of Nova Gorica, protocol no. 4/2024-6.

<sup>2</sup>In our design, participants were presented with one sentence at a time; all preceding and subsequent sentences served as fillers relative to that target. The randomization procedure ensured that different lexicalizations of the same word order never appeared consecutively. Lexical variation was used throughout to maximize filler-like effects. We excluded from analysis all trials in which participants rated the canonical (SVO) sentence below 4, which served as a baseline check for syntactic norm adherence and task engagement. We opted for this design because using actual fillers in the context of studying all possible word order permutations of a five-word long kernel sentence would result in an enormous logistical problem or an unfeasibly large number of trials.

Likert scale in a speeded-acceptability task, with a 7-second limit per sentence.<sup>3</sup> After every 20 sentences, participants could take a short break. Sentences were presented one by one on the computer screen and participants typed numerical acceptability responses. Participation was restricted to PCs and laptops.

## 6 Experiment 2 (Serbo-Croatian)

### 6.1 Participants

118 adult self-reported native speakers of Serbo-Croatian participated in the study (mean age=33). Participation was voluntary and uncompensated.

### 6.2 Materials and Procedure

In Experiment 2 the structure of kernel sentences was the same as in Experiment 1, but only 3 different sentence lexicalizations were used resulting in 3 experimental lists containing the set of all 120 permutations of a single lexicalization each. In all sentences, the grammatical subjects were animate, while the direct object could be animate or inanimate. Gender and number features of the subjects and objects varied across stimuli. All sentences were in past tense and had the clitic *je* (be.3SG) or *su* (be.3PL) as an auxiliary verb. There was no time limit for answers and all stimuli sentences were presented at the same time to the participants. 63 participants from our pool evaluated the first experimental list, 40 participants evaluated the second experimental list, and 15 participants evaluated the third one. Other aspects of the experiment design were identical to Experiment 1.

## 7 Modeling Results: Processing and Grammar Interaction

First, we examined whether any demographic or lexical factors significantly predicted acceptability ratings. A cumulative link mixed model implemented via the ordinal package (Christensen, 2023) in R (R Core Team, 2021) was used for this purpose. The results indicated that only subject animacy was a significant predictor. Consequently, it was added as a covariate to all subsequent models including the null models, that served as baseline for comparison of models. Subject Animacy was included only

<sup>3</sup>We opted for the speeded acceptability task in order to (a) avoid potential satiation effects and (b) explore its advantage as better reflecting initial parsing difficulty (Sprouse, 2008; Weskott and Fanselow, 2011)

in Russian related models, since in Serbo-Croatian stimuli all subjects were animate

Model fit was evaluated using likelihood-ratio tests against the null model, with p-values adjusted via the Holm–Bonferroni correction. Table 1 presents the results of model comparisons for Russian, ranked by AIC (Akaike, 1970) and pseudo- $R^2$  (McFadden, 1974). The models with the respective tested metrics were numerically coded in the table as follows:

1. Projectivity violations (verb as root),
2. Projectivity violations (auxiliary as root),
3. MDD (auxiliary as root),
4. Word Order Penalty Score,
5. MDD (verb as root),
6. Number of Displaced Words,
7. Total Path Length,
8. Residuals and Processing Penalties.

As shown in Table 1, the best-fitting predictors are projectivity-based metrics (with either verb or auxiliary as root) and MDD in both variants. Although several other models also differ significantly from the null, AIC differences of over 200 strongly favor the top-ranked models, providing clear evidence of superior model fit.

For Serbo-Croatian, we first examined the effect of Wackernagel Law violations (NotClitic2) by comparing the subset of sentences containing such violations to the rest of the dataset. A cumulative link mixed model was fitted using the ordinal package in R (Christensen, 2023) with sentence rating as the dependent variable, NotClitic2 as a fixed effect, and random intercepts for lexicalization and participant (subject\_id). Sentences with clitic-placement violations were rated significantly lower ( $p < 0.001$ ), with a mean rating of 1.8 compared to 3.44 for the rest of the dataset. All these findings were further confirmed through cross-validation, following (Barth and Kapatsinski, 2018).

The results for Serbo-Croatian highlight several key points. First, the goodness-of-fit patterns across metrics partially align with those observed for Russian. However, the metric based on projectivity violations with the verb as root clearly outperforms all others, with an AIC advantage of over 250 points compared to the next best metric, MDD (verb as root). In contrast, projectivity (auxiliary as root) and MDD (auxiliary as root) fail to reach significance thresholds. This may be attributed to the fact that, after subsetting, we only have sentences with auxiliaries remaining in a fixed position, which results in lower overall word order



Model	AIC	ps.- $R^2$	p-value	p-bonf
Null	24952.8	NA	NA	NA
1	24420.5	0.02	<0.001	<0.001
2	24536.6	0.01	<0.001	<0.001
3	24538.5	0.01	<0.001	<0.001
4	24739.4	0.009	<0.001	<0.001
5	24739.8	0.009	<0.001	<0.001
6	24934.9	<0.001	<0.01	<0.01
7	24945.9	<0.001	<0.05	0.023
8	24953.6	<0.001	0.3	1.00

Table 1: Comparison of metrics’ fit for Russian data

Model	AIC	ps.- $R^2$	p-value	p-bonf
Null	7478.68	NA	NA	NA
1	7220.93	0.04	<0.001	<0.001
2	7479.43	<0.001	0.26	1.00
3	7480.15	<0.001	0.47	1.00
4	7480.47	<0.001	0.64	1.00
5	7401.95	0.01	<0.001	<0.001
6	7440.82	0.005	<0.01	<0.01
7	7404.36	0.01	<0.001	<0.001
8	7451.93	0.004	<0.01	<0.01

Table 2: Comparison of metrics’ fit for Serbo-Croatian data, clitic-second condition.

variability than in Russian. In contrast, main verbs in Serbo-Croatian exhibit greater positional flexibility, making the verb-based projectivity metric more sensitive and explanatory in this context.

These findings emphasise that language-specific grammatical constraints must be modelled explicitly in “free” word-order systems. When discourse cues are deactivated (as in out-of-context judgements), ease of processing alone cannot compensate for categorical violations. Omitting grammar–processing interactions can therefore blur metric performance (see Table "AIC\_and\_R2\_SC\_data" in our [OSF](#) repository). We argue, therefore, that predictive models should include such constraints as independent factors. A promising next step for future work is a weighted framework that gives grammar maximal weight under decontextualised conditions, for example, assigning grammatical constraints 100% of the weight when no contextual licensing is available, then letting processing metrics assume a larger role when discourse support is present.

## 8 A Closer Look at the Projectivity

While the difference in projectivity-based metrics for Serbo-Croatian has a straightforward explanation, the analogous divergence in Russian presents a conundrum. A projectivity violation is defined based on the number of crossing dependencies of various types. But why should two implementations of the same general principle, with either the verb or the auxiliary as root, differ so markedly in model performance?

We propose that the discrepancy arises not solely from the quantity of crossings, but from the type of dependency relation whose arc performs the crossing, a subtle but crucial factor<sup>4</sup>. Consider the dependency structure when the Verb is taken as root: Verb  $\rightarrow$  Object, Verb  $\rightarrow$  Subject, Verb  $\rightarrow$  Auxiliary, Object  $\rightarrow$  Adjective. Since arcs emanating from the same node cannot intersect, only one arc in this configuration is capable of producing projectivity violations of both types – Object  $\rightarrow$  Adjective.

Now consider the structure when the Auxiliary is taken as root: Auxiliary  $\rightarrow$  Subject, Auxiliary  $\rightarrow$  Verb, Verb  $\rightarrow$  Object, Object  $\rightarrow$  Adjective. In this configuration, two dependencies: Verb  $\rightarrow$  Object and Object  $\rightarrow$  Adjective – are structurally eligible to cross other arcs. Thus, the auxiliary-root model includes a distinct type of projectivity violation not present in the verb-root model. It is plausible that these dependency types differ in processing cost, leading to divergent model performance despite comparable violation counts.

To evaluate this hypothesis, we conducted three additional statistical analyses for Russian. First, we removed all sentences in which projectivity violations involved only VO-type dependencies (Verb  $\rightarrow$  Object), and compared Aux-root and V-root models once again. The resulting AICs equaled 19035,4 for Aux-as-root, and 19050,5 for V-as-root. We assessed the robustness of this difference using a non-parametric bootstrap. Given that the two models under comparison are non-nested and AIC values can be sensitive to data perturbation, we applied a bootstrapping procedure to evaluate whether the observed  $\Delta$ AIC is stable across resampled datasets. The original dataset consisted of 9123 observations. For each of 1000 bootstrap iterations, we resampled the full dataset with replacement. The mean difference in AICs was 14.2

<sup>4</sup>We are grateful to an anonymous reviewer for raising this possibility.

in favor of the Aux model. However, the 95% confidence interval [-29.43, 45.59] included 0 and  $p = 0.712$ , indicating that the observed difference is not statistically robust. We therefore conclude that the apparent advantage of the Aux model is not reliably supported across resampled datasets.

In addition, these results indicate that a VO-dependency is easier to process than an Object-Adjective(AO) dependency. To further test this hypothesis and also test the significance of violation type (weak vs strong) we ran a cumulative link mixed-effects model on acceptability values testing an interaction term  $dependency\_type * violation\_type$ , where dependency type ranges over VO, AO, or both and violation type values were weak, strong and weak+strong (since a violation can be weak and strong at the same time). Random effects included participant and lexicalization. This was followed by pairwise comparisons with Tukey correction, whose results are reported in Table 3.<sup>5</sup>

Comparison	$\beta$	SE	z	p
VO.s-VO.w	-0.085	0.12	-0.68	0.999
VO.s-VO.ws	-0.38	0.14	-2.8	0.113
VO.s-AO.s	1.24	0.1	11.7	<b>&lt;.0001</b>
VO.s-AO.w	0.61	0.15	3.9	<b>0.0025</b>
VO.s-AO.ws	1.02	0.12	8.5	<b>&lt;.0001</b>
VO.w-VO.ws	-0.3	0.14	-2.1	0.458
VO.w-AO.s	1.32	0.11	11.7	<b>&lt;.0001</b>
VO.w-AO.w	0.69	0.16	4.4	<b>0.0004</b>
VO.w-AO.ws	1.11	0.12	9.4	<b>&lt;.0001</b>
VO.ws-AO.s	1.62	0.13	12.3	<b>&lt;.0001</b>
VO.ws-AO.w	0.99	0.17	5.9	<b>&lt;.0001</b>
VO.ws-AO.ws	1.41	0.14	9.9	<b>&lt;.0001</b>
AO.s-AO.w	-0.63	0.14	-4.5	<b>0.0003</b>
AO.s-AO.ws	-0.21	0.1	-2.1	0.49
AO.w-AO.ws	0.42	0.14	2.9	0.09

Table 3: Pairwise comparisons of dependency and violation types, ( $p < 0.05$ ) are in bold in the text.

Table 3 shows that the factor  $dependency\_type$  plays a substantial role in shaping acceptability judgments. Specifically, violations caused by VO arcs differ significantly from their AO counterparts: even weak violations, where the AO dependency crosses the root, are rated significantly lower than any VO violation. In addition, the type of viola-

<sup>5</sup>We excluded comparisons involving the "both" condition from the presentation for reasons of brevity. The full model output is available on the [OSF](#)

tion interacts with the dependency type. In the VO condition, the distinction between weak, strong, and combined violations does not yield significant differences in ratings. In contrast, within AO structures, weak violations are rated significantly higher than strong and marginally than combined violations, although the difference between AO\_strong and AO\_weak+strong is not statistically significant. Moreover, the effect of weak violation seems to be blurred in mixed types.

Importantly, however, the question of how dependency type and violation type work together to predict acceptability is still far from resolved. Among the 120 permuted sentences, some involved mixed violations – for instance, [Obj Subj V Aux Adj] includes both an AO-arc (a weak violation) and a VO-arc (a strong violation). Investigating how such combinations of violations affect acceptability represents a promising direction for future research, as the effects appear to be cumulative and potentially multidirectional.

Finally, to further disentangle the influence of violation number from dependency type, we conducted an ordinal regression on a subset of sentences containing only strong violations involving the AO dependency. The model used the number of strong violations as the only predictor (with the same random effects structure). The results showed that sentences with a single strong violation were rated significantly higher than those with two violations ( $p = 0.0007$ ), strengthening the conjecture that the number of violations independently contributes to acceptability judgments.

## 9 The Interaction of MDD and Projectivity

The final question is whether MDD and projectivity violation metrics capture overlapping or complementary aspects of word order processing—i.e., whether combining them improves predictive accuracy across languages compared to using either metric alone. To address this, we recalculated AIC and pseudo- $R^2$  values using projectivity violations as the baseline, given that earlier tests indicated that MDD alone performed worse. For consistency, and because the bootstrap analysis showed no significant difference between Verb-as-root and Aux-as-root calculations, we adopted the Verb-as-root configuration for both languages. The resulting model comparisons are shown in Tables 4 and 5 for Russian and Serbo-Croatian, respectively.

Model Name	AIC	Pr(>Chisq)
Proj. (V-Root)	24420.54	NA
Proj. + MDD	24330.29	<0.001

Table 4: Comparison of Mixed Models: MDD vs. Projectivity + MDD (Verb as Root), Russian

Model Name	AIC	Pr(>Chisq)
Proj. (V-Root)	7220.93	NA
Proj. + MDD	7192.32	<0.001

Table 5: Comparison of Mixed Models: MDD vs. Projectivity + MDD (Verb Root), Serbo-Croatian

These findings indicate that MDD and projectivity violation metrics are not redundant, but rather complementary. Their combined use enhances the model’s ability to account for the distribution of acceptability ratings in free word order languages. Importantly, the joint model demonstrates greater descriptive and predictive power.

To further illustrate this, consider two example sentence types: *Adj Aux Subj Obj V* and *V Aux Adj Obj Subj*. Both have MDD = 2.25, yet differ in projectivity violations (2 vs. 0) and acceptability ratings (2.33 vs. 2.97). The ordinal regression showed significant differences in their ratings,  $p < 0.05$ . This suggests that the two metrics capture distinct cognitive pressures: MDD reflects general processing economy, while projectivity violations relate to structural predictability and locality. Together, they offer a more comprehensive account of how sentence structure is evaluated during comprehension.

## 10 General Discussion

This study yields several key findings. First, the best-performing model was the one that combined Mean Dependency Distance (MDD) and projectivity violation metrics, outperforming models that used either metric alone. This finding broadens the scope for future inquiry across different theoretical frameworks, including those within the MDD/MDDP tradition (Boston et al., 2011; Ferrer-i Cancho, 2004; Gibson, 1998; Futrell et al., 2015; Gibson, 2000) as well as those focusing on projectivity and its violations (Ferrer-i Cancho, 2017; Yadav et al., 2020, 2022).

Second, our findings suggest that MDD and projectivity encode distinct cognitive mechanisms involved in syntactic parsing. MDD indexes the

memory load of cue-based retrieval: longer linear distances force the parser to keep more items active, increasing interference and cost. Projectivity violations, by contrast, disrupt a stack-based incremental parser, penalizing non-projective structures where immediate attachment is not possible (Frazier and Fodor, 1978; Frazier, 1979). In transition-based parsing, projective dependency trees are precisely those that can be derived with a single push-down stack, whereas non-projective trees need extra SWAP operations or an auxiliary stack (Nivre, 2003, 2009). Our final regression model further supports this interpretation. Further research is needed to compare these mechanisms directly, particularly in cases where the two principles overlap or diverge in their predictions.

Third, our data indicate that the type of violation matters, but only for certain dependency types. The contrast we observed between AO and VO dependencies may point to effects of structural embeddedness that may increase the cognitive cost of disrupting the canonical configuration. In both dependency and phrase-structure based grammars, the AO pair is structurally more embedded than the VO relation. Again, targeted studies are required to test this hypothesis explicitly.

Finally, our modeling results underscore the complex interaction of multiple factors contributing to acceptability. Grammatical constraints appear to carry the greatest weight, followed by projectivity violations. Within the latter, it would be valuable to further explore how dependency type, violation type, and number of violations contribute independently and interactively. The third layer is MDD, which acts as a general processing principle affecting sentence structure evaluation. Understanding how these layers interact and compete within the acceptability space is a promising direction for future formal and experimental modeling.

## Limitations

One key limitation of this study is that it relies on acceptability judgments, which, while informative, provide only an indirect measure of real-time processing. Also, although two processing principles were modeled (MDD and projectivity violations), other cognitive and discourse-level factors, such as information structure, prosody, or thematic prominence, were not explicitly controlled or integrated into the models, leaving open questions about their interaction with structural constraints.

## Acknowledgments and Data Availability

We thank three anonymous reviewers of this paper for their useful feedback. This research has received funding from the Slovenian Research and Innovation Agency (ARIS) under project no. J6-4615. All materials and experimental results from this study are available in our OSF repository at [https://osf.io/7p29c/?view\\_only=ce1cc0a390a24865b76558f9974dceef](https://osf.io/7p29c/?view_only=ce1cc0a390a24865b76558f9974dceef).

## References

- Hirotsugu Akaike. 1970. **Statistical Predictor Identification**. *Annals of the Institute of Statistical Mathematics*, 22:203–217.
- Markus Bader and Michael Meng. 1999. **Subject-object Ambiguities in German Embedded Clauses: An Across-the-Board Comparison**. *Journal of Psycholinguistic Research*, 28(2):121–143.
- John F. Bailyn. 1995. *A configurational approach to Russian “free” word order*. Ph.D. thesis, Cornell University, Ithaca.
- Danielle Barth and Vsevolod Kapatsinski. 2018. **Evaluating Logistic Mixed-Effects Models of Corpus-Linguistic Data in Light of Lexical Diffusion**. In Dirk Speelman, Kris Heylen, and Dirk Geeraerts, editors, *Mixed-Effects Regression Models in Linguistics. Quantitative Methods in the Humanities and Social Sciences*, 1 edition, volume 1, pages 99–116. Springer International Publishing, Cham, Switzerland.
- Otto Behaghel. 1909. **Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern**. *Indogermanische Forschungen*, 25(1909):110–142.
- Marisa F. Boston, John T. Hale, Shravan Vasishth, and Reinhold Kliegl. 2011. **Parallel processing and sentence comprehension difficulty**. *Language and Cognitive Processes*, 26(3):301–349.
- Željko Bošković. 2005. **Left branch extraction, structure of np, and scrambling**. In Joachim Sabel and Mamoru Saito, editors, *The free word order phenomenon: Its syntactic sources and diversity*, pages 13–73. Mouton de Gruyter, Berlin. Accessed 13 June 2025.
- Željko Bošković. 2001. *On the Nature of the Syntax-Phonology Interface: Cliticization and Related Phenomena*, volume 60 of *North Holland Linguistic Series: Linguistic Variations*. Brill, Leiden. Accessed 13 June 2025.
- Hyo-Woon Choi. 2007. **Length and order: A corpus study of Korean dative-accusative construction**. *Discourse and Cognition*, 14(3):207–227. Accessed 13 June 2025.
- Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. The MIT Press.
- Rune H. B. Christensen. 2023. *ordinal-Regression Models for Ordinal Data*. R package version 2023.12-4.1.
- Meredith Daneman and Patricia A. Carpenter. 1980. **Individual Differences in Working Memory and Reading**. *Journal of Verbal Learning & Verbal Behavior*, 19(4):450–466.
- Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katrin Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. **Universal Stanford Dependencies: A Cross-Linguistic Typology**. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. Accessed 13 June 2025.
- Ramon Ferrer-i Cancho. 2004. **Euclidean distance between syntactically linked words**. *Phys. Rev. E*, 70(5):056135.
- Ramon Ferrer-i Cancho. 2017. **The Placement of the Head that Maximizes Predictability. An Information Theoretic Approach**. *Preprint*, arXiv:1705.09932. ArXiv preprint, version 3.
- Lyn Frazier. 1979. *On Comprehending Sentences: Syntactic Parsing Strategies*. Ph.D. thesis, University of Connecticut.
- Lyn Frazier and Janet Dean Fodor. 1978. **The Sausage Machine: A New Two-Stage Parsing Model**. *Cognition*, 6(4):291–325.
- Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. **Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing**. *Cognitive Science*, 44(3):e12814.
- 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 of the United States of America*, 112(33):10336–10341.
- Edward Gibson. 1998. **Linguistic Complexity: Locality of Syntactic Dependencies**. *Cognition*, 68(1):1–76.
- Edward Gibson. 2000. **The Dependency Locality Theory: A Distance-Based Theory of Linguistic Complexity**. In Alec Marantz, Yasushi Miyashita, and Wayne O’Neil, editors, *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 94–126. The MIT Press.
- Daniel Gildea and David Temperley. 2010. **Do grammars minimize dependency length?** *Cognitive Science*, 34(2):286–310.
- Daniel Grodner and Edward Gibson. 2005. **Consequences of the serial nature of linguistic input for sentential complexity**. *Cognitive Science*, 29(2):261–290.



- Thomas Groß and Timothy Osborne. 2015. [The Dependency Status of Function Words: Auxiliaries](#). In *International Conference on Dependency Linguistics*.
- John A. Hawkins. 1994. *A Performance Theory of Order and Constituency*, volume 73 of *Cambridge Studies in Linguistics*. Cambridge University Press, Cambridge.
- Richard Hudson. 1995. Measuring syntactic difficulty. Unpublished paper. Available at <http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf>. Accessed 15 June 2025.
- Jukka Hyönä and Heli Hujanen. 1997. [Effects of Case Marking and Word Order on Sentence Parsing in Finnish: An Eye Fixation Analysis](#). *The Quarterly Journal of Experimental Psychology Section A*, 50(4):841–858.
- Elena D. Kallestinova. 2007. *Aspects of Word Order in Russian*. Ph.D. thesis, University of Iowa, Iowa City. Accessed 8 December 2024.
- Richard L. Lewis and Shravan Vasishth. 2005. [An activation-based model of sentence processing as skilled memory retrieval](#). *Cognitive Science*, 29(3):375–419.
- 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.
- Qun Lu, Chunshan Xu, and Haitao Liu. 2016. [Can chunking reduce syntactic complexity of natural languages?](#) *Complexity*, 21:33–41.
- John J. McCarthy and Alan Prince. 1995. [Faithfulness and reduplicative identity](#). In *University of Massachusetts Occasional Papers in Linguistics 18: Papers in Optimality Theory*. University of Massachusetts.
- Daniel McFadden. 1974. [Conditional Logit Analysis of Qualitative Choice Behavior](#). In Paul Zarembka, editor, *Economic Theory and Mathematical Economics*, pages 105–142. Academic Press, New York. Accessed 13 June 2025.
- Igor’ A. Mel’čuk. 2009. [Dependency in natural language](#). In Alain Polguère and Igor’ A. Mel’čuk, editors, *Dependency in Linguistic Description*, pages 1–110. John Benjamins Publishing.
- Edson T. Miyamoto and Shoichi Takahashi. 2002. The Processing of Wh-Phrases and Interrogative Complementizers in Japanese. In *Japanese/Korean Linguistics*, volume 10, pages 62–75.
- Joakim Nivre. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of IWPT-2003*, pages 149–160, Nancy, France.
- Joakim Nivre. 2009. Non-projective dependency parsing in expected linear time. In *Proceedings of ACL-IJCNLP 2009*, pages 351–359, Singapore.
- Artem Novozhilov, Kirill Chuprisko, and Arthur Stepanov. 2025. Dense sentence sets induce an anchor-and-baseline strategy in likert scale acceptability judgments. In *Proceedings of the 47th Annual Meeting of the Cognitive Science Society*. To appear.
- Charles A. Perfetti and Alan M. Lesgold. 1977. Discourse Comprehension and Sources of Individual Differences. In Marcel A. Just and Patricia Daneman, editors, *Discourse Comprehension and Sources of Individual Differences*, pages 141–183. Pittsburgh University.
- R Core Team. 2021. [R: A Language and Environment for Statistical Computing](#).
- Jane J. Robinson. 1970. [Dependency Structures and Transformational Rules](#). *Language*, 46(2):259–285.
- Irene Ros, Marta Santesteban, Kazuko Fukumura, and Itziar Laka. 2015. [Aiming at shorter dependencies: The role of agreement morphology](#). *Language, Cognition and Neuroscience*, 30(9):1156–1174.
- Irina A. Sekerina. 1999. [The Scrambling Complexity Hypothesis and Processing of Split Scrambling Constructions in Russian](#). *Journal of Slavic Linguistics*, 7(2):265–304.
- Natalia Slioussar and Ilya Makarchuk. 2022. [SOV in Russian: A Corpus Study](#). *Journal of Slavic Linguistics*, 30(3):1–14.
- Jon Sprouse. 2008. [The effect of task demands on acceptability judgments](#). *Journal of Linguistics*, 44(2):387–408.
- Adrian Staub, Charles Clifton, and Lyn Frazier. 2006. [Heavy NP shift is the parser’s last resort: Evidence from eye movements](#). *Journal of Memory and Language*, 54(3):389–406.
- Sandra Stjepanović. 1999. *What do second-position cliticization, scrambling and multiple wh-fronting have in common?* PhD dissertation, University of Connecticut, Storrs.
- Jakov G. Testeleš. 2001. *Introduction to General Syntax [Vvedenie v obshchii sintaksis]*. Russian State University for the Humanities.
- Zoran Urošević, Claudia Carello, Milan D. Savić, Georgije Lukatela, and Michael T. Turvey. 1986. [Some word-order effects in Serbo-Croat](#). *Language and Speech*, 29(2):177–195.
- Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. [Neural Network Acceptability Judgments](#). *Transactions of the Association for Computational Linguistics*, 7:625–641.

- Thomas Weskott and Gisbert Fanselow. 2011. [On the informativity of different measures of linguistic acceptability](#). *Language*, 87(2):249–273.
- Himanshu Yadav, Samar Husain, and Richard Futrell. 2022. [Assessing Corpus Evidence for Formal and Psycholinguistic Constraints on Nonprojectivity](#). *Computational Linguistics*, 48(2):375–401.
- Himanshu Yadav, Ashwini Vaidya, Vishakha Shukla, and Samar Husain. 2020. [Word order typology interacts with linguistic complexity: A cross-linguistic corpus study](#). *Cognitive Science*, 44(4):e12822.
- Jeremy Zehr and Florian Schwarz. 2018. [PennController for Internet-Based Experiments \(IBEX\)](#).
- Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2024. [MELA: Multilingual Evaluation of Linguistic Acceptability](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2658–2674, Bangkok, Thailand. Association for Computational Linguistics.



# UD Annotation of Experience Clauses in Tigrinya

**Nazareth Amlesom Kifle**  
Østfold University College  
nazareth.a.kifle@hiof.no

**Michael Gasser**  
Indiana University  
gasser@iu.edu

## Abstract

We are developing a treebank for Tigrinya within the Universal Dependency (UD) framework. UD proposes a set of universal grammatical relations such as *nsubj*, *obj* and *iobj* to capture dependency relations between words in any language. However, for some classes of verbs it is not a straightforward matter to know what grammatical relations the verbs are categorized for. In this paper we discuss the decisions we have had to make for the annotation of arguments of experience verbs in the Semitic language Tigrinya, which exhibit a number of unusual morphosyntactic properties. We describe a classification of experience verb roots in the language, based on the various ways in which the core experiencer and stimulus arguments are realized syntactically and morphologically and on which valence-changing operations the roots permit. We supplement our analysis with data from a morphologically analyzed Tigrinya corpus.

## 1 Introduction

We are developing a morphologically rich Universal Dependency (de Marneffe et al., 2021) treebank for the Semitic language Tigrinya. In addition to the extensions required to accommodate dependencies within as well as between words, we face several annotation challenges because of the mismatch between morphology and syntax and the unusual behavior of some verbs.

In this paper we focus on the category of experience verbs. Such verbs possess arguments that undergo some sort of mental, emotional or sensory experience and which exhibit variation in their morphosyntactic encoding in a wide variety of languages (Belletti and Rizzi, 1988; Næss, 2007; Psetsky, 2000; Croft, 1993, 577-580). Experiencer predicates are typically categorized for an experiencer, the argument that experiences the mental state, and a stimulus, the argument that instigates

the experience. Some predicates express the experiencer as a subject, while others express it as an object. This syntactic variation can be illustrated by the English predicates *fear* and *like*, on the one hand, and *frighten* and *please*, the other hand, where the experiencer corresponds to the subject 'I' in 'I fear snakes.', but with the object 'me' in 'Snakes frighten me'.

In this study we aim to outline a classification of experience verb roots on the basis of the syntax and morphology of their base forms as well as their passive and causative forms, where applicable. To our knowledge, there are no studies dedicated to Tigrinya experiencer verbs, only one that briefly describes the constructions (Kifle, 2011, 128-133). There is some work on experience verbs in the closely related language, Amharic (Amberber, 2005; Workneh, 2019) and, where relevant, we look at how Tigrinya clearly differs from Amharic.

In our study we rely not only on the native-speaker intuitions of one of us, but also on a morphologically analyzed corpus, which reveals statistical tendencies for particular roots and subcategories within the categories we propose.

This paper is divided into eight sections. Following this introduction, in section 2, we give a brief description of the morphologically enriched treebank we are developing. Section 3 presents a brief introduction to the morphosyntactic properties of Tigrinya, indicating in general how we annotate syntactic and morphological dependencies in our treebank. In Section 4, experiencer clauses are briefly described. Section 5 presents our morphologically analyzed corpus and the corpus data. Section 6 covers the method we used to categorize Tigrinya experience roots. In Section 7, we present the four categories of Tigrinya experience roots. Finally, in Section 8, we summarize our conclusions and outline future work to cover other possible arguments of experience verbs in the language and to automatically classify experience roots.

## 2 A morphologically enriched treebank

In the Tigrinya treebank we are creating, we segment morphologically complex words, treating all inflectional morphemes as tokens with their own parts-of-speech, lemmas, features, and dependencies. We do not separate derivational morphemes. We also maintain the distinction between subword tokens and morphologically complex words, making use of the CoNNL-U extension for handling multi-token expressions (<https://universaldependencies.org/format.html#words-tokens-and-empty-nodes>) for this purpose, as is done in the existing Amharic (<https://universaldependencies.org/am/index.html>) and Yupik (<https://universaldependencies.org/ess/index.html>) UD treebanks. Other UD treebanks that treat inflectional morphemes as tokens with their own relations to stems but do not maintain a separate word level that groups subword units together are the Beja (<https://universaldependencies.org/bej/index.html>) and Japanese (<https://universaldependencies.org/ja/index.html>) treebanks.

One of our goals in making relations explicit at both the morphological (within-word) and syntactic (between-word) levels is to explore and elucidate the complex ways in which participants are encoded both within a verb and as explicit nominals. The mapping between categories of pronominal affixes on verbs and case marking on nominals is not one-to-one in the language (Kifle, 2011, 66ff.). For example, the object pronominal suffix that typically marks definite accusative objects can also code applicative objects of intransitive verbs that are understood as affected participants which negatively experience the action of the verb. Moreover, pronominal suffixes serve as embedded pronouns instead of merely being agreement makers. In addition, the object case marker codes both accusative, dative and applicative objects. A further reason for segmenting verbs, nouns, and adjectives stems from our interest in using the treebank to train linguistically enriched language models and machine translation systems. The treebank will provide linguistically motivated subword units as an alternative to the segments generated by statistical methods such as byte-pair encoding that are the norm for such models (Gezmu, 2023).

## 3 Tigrinya morphosyntax

Tigrinya belongs to the family of Semitic languages spoken in Ethiopia and Eritrea. Like the other languages in this family, it is written in the Ge'ez abugida writing system. In our treebanks, we make use of Ge'ez orthography only, including for the morphological segmentation of words, but for the purposes of this paper, we add phonetic transcriptions and indicate the segmentation of words only when this is necessary to make a point.

The Ethiopian-Eritrean Semitic languages share many of the properties of other Semitic languages (e.g., template-based morphology, obligatory subject agreement, object agreement) as well as a number of properties of their own (e.g., verb final clause structure) (Feleke, 2021; Demeke, 2003; Hetzron, 1972). In this section, we describe morphological and syntactic properties of verbs and nominals that are relevant for the annotation of experience verbs and their arguments.

### 3.1 Verbs and valence-changing derivation

Tigrinya verbs consist of a stem and affixes coding subject and object agreement. Subordinate verbs take additional prefixes representing conjunctions and, for relative verbs, optional adpositions representing the case of the modified nominal.

- (1) ሰለዝረአዩቶ  
*silā-zī-rəʔay-ət-to*  
since-that-see.PFV-SB3SF-OB1,3SM  
‘since she saw him’

As in other Semitic languages, verb stems are in turn derived from a root consisting of a series of consonants and a template consisting of a pattern of vowels inserted between the consonants and sometimes the gemination of one of the consonants. The language distinguishes four basic tense-aspect-mood categories, differing in their templates and their subject agreement affixes.

In addition to its base (simplex) form, each root can also appear in one or more derived forms, traditionally called *ʔaʕimad* (ሳዕጣድ), corresponding to the *binyanim* of Hebrew and the *ʾawzaan* of Arabic verbs. Each *ʔaʕimad* has separate templates for each of the language’s four tense-aspect-mood categories. As is usual for Arabic, we will refer to the different *ʔaʕimad* possibilities as “forms.” For a given root, there may be as many as eight forms, in addition to the base form. In this paper we consider

only three of these: the BASE, the PASSIVE and the CAUSATIVE. Note that the specific interpretation of what we are calling PASSIVE and CAUSATIVE varies with the root. For example, some roots have no BASE form, and it is the PASSIVE or CAUSATIVE form that functions as the base form for these roots (Kifle, 2011, 61). We will refer to verb roots and stems using the 3rd person singular masculine perfective, as is conventional for Semitic languages.

### 3.2 Subject and object agreement

As in other Afro-Asiatic languages, verbs in Tigrinya are obligatorily inflected for subject person-number-gender agreement. In our morphologically enriched treebanks, we segment off subject agreement affixes and annotate the dependency joining the verb stem to them with the relation *nsubj*, adding the sub-relation *:aff* to distinguish them from the syntactic relations with the same label, as is done by Kahane et al. (2021, 51) for their morpheme-based treebank for Beja.

Tigrinya does not have a neuter gender, and 3rd person singular masculine (3SM) agreement is used to refer both to inanimate nouns that are lexically masculine and to unspecified dummy entities. As in modern Hebrew (Halevy, 2023, 10-12), it does not also have a locative or a demonstrative expletive, such as *there* in ‘There is water in the glass.’, or a dummy subject pronoun, such as *it* in ‘It is hot’, as in example (2).<sup>1</sup>

- (2) ሞዖቁ  
*moyq-u*  
 be.hot.PFV-SB3SM  
 ‘It got/is hot.’

Verbs in Tigrinya may also take object agreement suffixes, also called “object suffix pronouns.” These appear in two types, which we refer to as “object1” (OBJ1) and “object2” (OBJ2), following Kifle (2011, 104). The suffixes may refer to both objects (direct and indirect) and to applicative arguments, for example, *-to* (O1,3SM) in ርእየቶ *riʔyat-to*, ‘she saw him’; *-llu* (O2,3SM) in ርእየቶ *riʔiya-llu*, ‘she saw for/on him’. While OBJ1 most often represents the direct or indirect object of a transitive

<sup>1</sup>We use the following abbreviations in interlinear glossing. 1: 1st person, 2: 2nd person, 3: 3rd person, AUX: Auxiliary, CAUS: Causative, DEF: Definite, F: Feminine, IPFV: Imperfective, M: Masculine, O1: Object1, O2: Object2, OBJ: Objective, PASS: Passive, PST: Past, PFV: Perfective, P: Plural, POSS: Possessive, PRS: Present, REL: Relative, S: Singular, SB: Subject.

verb, it may also represent a malefactive argument of an intransitive verb, for example, *-to* (O1,3SM) in ሞይታቶ *moytat-to*, ‘she died on him/to his detriment.’ (Kifle, 2007, 2011, p.119).

Each verb may take at most one object suffix; thus, speakers must choose between the objective and applicative suffixes when both are applicable to the arguments of a verb. In our treebank, we segment off the object suffixes. Since neither category of suffix corresponds directly to the UD *obj* relation, we annotate dependencies from the stem to the two types of suffixes with the special morphological relations *obj1:aff* and *obj2:aff*.

### 3.3 Nominals and case

Subjects in Tigrinya are not marked for case. Direct and indirect objects may take the objective prefix ጎ- *n-*.<sup>2</sup> Definite objects are obligatorily marked for case. The objective case marker also functions as the dative case marker (Kifle, 2007; Kievit and Kievit, 2009), marking arguments we annotate as *iobj*.

- (3) a. ኣሰቴር ጎዮሴፍ ርእየቶ [Accusative]  
*ʔaster ni-yosef riʔiy-at-to*  
 Aster OBJ-Yosef see.PFV-SB3SF-O1,3SM  
 ‘Aster saw Yosef.’
- b. ኣሰቴር ጎዮሴፍ ህዖብ [Dative]  
*ʔaster ni-yosef hiyab*  
 Aster OBJ-Yosef gift  
*ሃባቶ*  
*hib-at-to*  
 gave.PFV-SB3SF-O1,3SM  
 ‘Aster gave Yosef a gift.’

Figure 1 shows the syntactic and morphological dependencies within sentence (3a). Co-referential nouns and verb affixes are indicated with the same color. Note that the co-reference relations are not explicit in the dependency graph.

The language has a set of adpositions that mark different semantic roles such as instrumental (*-n bi*), locative (*ኣብ ʔab*), associative (*ምስ mis*) and elative (*ካብ kab*). As we shall see in some of the examples below, there is no simple one-to-one or one-to-many mapping between the categories of object affixes on verbs and the case markers and adpositions.

<sup>2</sup>This prefix is normally referred to as “accusative,” but we prefer “objective” because it can mark indirect as well as direct objects.

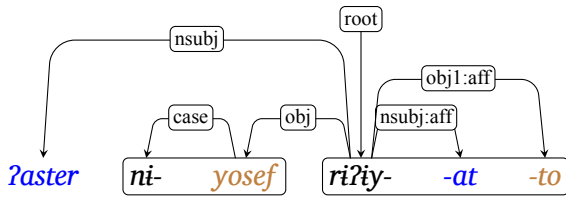


Figure 1: Dependencies in (3a). Blue tokens represent the subject, brown tokens the direct object.

## 4 Experience clauses

Experience clauses contain experiencer predicates, also known as “psychological predicates” (Postal, 1971, chapter 6) and “mental verbs (Croft, 1993, 55), that denote events that affect the consciousness of the experiencer such as its emotional or mental state or bodily sensation (Verhoeven, 2014, 130). Experience clauses are characterized by the presence of at least one of the two core semantic roles, the animate participant undergoing the experience, the EXP(ERIENCER), and the event or entity causing the experience, the STIM(ULUS) (Dowty, 1991; Croft, 1993; Klein and Kutscher, 2015). Languages have different means for leaving either the EXP or the STIM unspecified, for example, in the English sentences *this film is depressing*, which foregrounds the STIM, and *I’m depressed*, which foregrounds the EXP. As we will also see, experience predicates expressed by adjectives, such as *be quiet*, in English are normally expressed by verbs in Tigrinya.

With particular experience predicates, additional semantic roles are possible. Sometimes an external CAUSER needs to be distinguished from the STIM argument, for example, *news* in the sentence *the news made her dislike her teacher*.

Because experiencers can be perceived with different degrees of control over the experienced states and events, it is common in the world’s languages for experience clauses to deviate from prototypical transitivity (Næss, 2007, 196). Experiencer nominals commonly appear as both subjects and objects, and when they are objects, they may be characterized by unusual case marking patterns.

## 5 Corpus Data

As we have access to a morphological analyzer for Tigrinya (<https://github.com/hltdi/HornMorpho>), we are able to assess how much information a morphologically analyzed corpus of sentences can provide about the statistical

tendencies characterizing particular roots in the different categories of experience verbs we will be proposing.

First, we consider what morphological agreement features we expect for the EXP and STIM arguments. Experiencers are normally people, so all three persons, including in particular 1st and 2nd, should be possible features of the affixes agreeing with the EXP argument. Thus the absence of 1st and 2nd person agreement features for a particular affix can indicate that it does not refer to an experiencer. Stimulus features, on the other hand, are relatively unconstrained: experiences can be caused by people as well as inanimate objects and events. Impersonal verbs with “dummy” subjects are a special case; they always take 3SM subject agreement.

We can estimate a root’s transitivity by looking at the proportion of instances that have an OBJ1 suffix, but this is only an indication of transitivity because (1) the suffix is only obligatory for definite objects and (2) though this is by far the most common use of the suffix, it can also function as a malefactive applicative agreement marker on intransitive verbs. Another measure of transitivity is the occurrence and frequency of the PASSIVE form of the root.

We ran a dedicated morphological analyzer on 1,000,000 Tigrinya sentences from the TLMD corpus (<https://zenodo.org/records/5139094>). For each verb root occurring in at least 10 unambiguous words, for each of the three forms under consideration, BASE, PASSIVE, and CAUSATIVE, we counted the occurrences of different subject and object agreement features.

For comparison we ran a morphological analyzer for the related Amharic language on 100,000 sentences from the CACO corpus (<https://github.com/andmek/CACO>).

## 6 Method

We start with the basic distinction between verbs taking EXP subjects and those taking EXP objects (Fleischhauer, 2016, 263-285). Because we are concerned with the annotation of the arguments of experience verbs, it is experience clauses, rather than simply experience verbs, that we will be discussing.

For each experience verb root that we consider, we will examine each of the three main forms that occur for that root: BASE, PASSIVE, and CAUSATIVE. For each of these forms, we will look



at how EXP and STIM are coded both morphologically and syntactically, and we will classify the roots on the basis of these properties. The result will be up to three morphosyntactic schemas for each root. For each schema we will be concerned with how the canonical roles are realized syntactically and morphologically and which UD relations we use for annotating each argument, both morphological agreement affixes and explicit nominal arguments of the verb.

## 7 Tigrinya Experience Verbs

The analysis of Tigrinya experiencer verbs reveals four categories which are outlined below.

### 7.1 Subject-experiencer verbs

Subject-experiencer (SE) verbs fall into two categories, intransitive verbs, which leave the STIM unexpressed, and transitive verbs, which code the STIM as direct object.

#### 7.1.1 Intransitive SE verbs

Experiencer verbs such as ሰንባደ *sənbədə* ‘be shocked’, ሓዘኻ *ħazəna* ‘be sad,’ and ኅገበ ናገጭ *ħagəbe* ‘be satisfied’ are typical examples of the intransitive subject experiencer (ISE) class. The BASE form of this class is illustrated in (4).<sup>3</sup>

- (4) ኣሰፍር ሰንባደ [Tir]  
*ʔaster sənbid-a*  
 Aster be.shocked.PFV-SB3SF  
 ‘Aster is shocked.’

ISE roots such as ሰንገደ *sənbədə* typically have a CAUSATIVE form in addition to the intransitive BASE SE form. The CAUSATIVE form takes the STIM as subject and the EXP as direct object. This is illustrated in (5).

- (5) ኣፍ ወረ ንኣሰፍር  
*ʔiti wərə ni-ʔaster*  
 the news OBJ-Aster  
 ኣሰንባደ-ዓ  
*ʔasənbid-u-wa*  
 be.shocked.CAUS-SB3SM-O1,3SF  
 ‘The news shocked Aster.’

The fully segmented dependency tree for the sentence is shown in Figure 2.

<sup>3</sup>In all of our examples with an explicit EXP, this will be the feminine participant Aster.

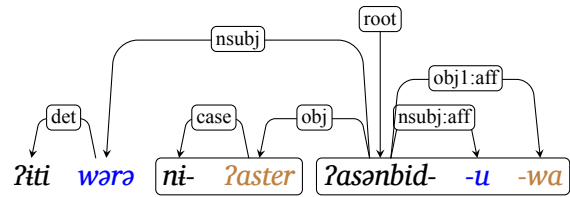


Figure 2: Dependencies in (5). Blue tokens represent the subject, brown tokens the direct object. Segmented words are surrounded by rectangles with rounded corners.

These CAUSATIVE forms can also appear without an explicit object EXP, where the focus is on the experience, independent of any particular EXP. We will refer to such clauses as ‘stimulus only’ clauses. (6) is an example with the CAUSATIVE of the root *sənbədə* ‘be shocked’.

- (6) ኣፍ ሰንባደ ኣዘዩ  
*ʔiti məbrəx' ʔazyu*  
 the lightning very  
 ኣሰንባደ ንጻፋ  
*yəsənbid nəyru*  
 be.shocked.CAUS.IPFV.SB3SM AUX.PST  
 ‘The lightning was very shocking.’

The CAUSATIVE forms of these roots with relative subordinating morphology correspond to causative experiential adjectives in languages such as English: ሰንገደ *zəsənbid* be.shocked.REL.CAUS.IPFV.SB3SM ‘shocking’ (lit., ‘that which causes shock’).

ISE verb roots normally have no PASSIVE form. The exceptions are roots that lack a BASE form. For these verbs, the PASSIVE form behaves like the BASE form of a verb like ሰንገደ *sənbədə*, as in (4). Examples are the intransitive roots ተሓገሰ *təħagʷəsə* ‘be happy’, ተጣዕሰ *tətʷaʃlə* ‘regret’, and ተጼጥዐ *təxʷətʷə* ‘be angry.’

Corpus data for four ISE verbs confirm what we expected: that there is a frequent CAUSATIVE but no PASSIVE form, that OBJ1 suffixes are rare with the BASE forms but common with the CAUSATIVE forms, and that 1st and 2nd person subjects are frequent with the BASE form.

#### 7.1.2 Transitive SE verbs

Tigrinya also has a set of transitive SE verbs (TSE) taking the STIM as direct object in the BASE form. Examples are ፈርኻ *fərħə* ‘fear’, ናፈቅ *nafəxʷə* ‘miss’, ሓፈረ *ħafərə* ‘be embarrassed (over),’ and

ጸልሐ ስ'ላሊ?ə 'hate.' As expected for transitive verbs, these roots usually have PASSIVE as well as BASE forms. We annotate the EXP subjects of these PASSIVE verbs as *nsubj:pass*.

These roots also have derived CAUSATIVE forms related to the BASE forms in the manner of pairs like English *fear* and *frighten*. The CAUSATIVE of ፈርሐ *fərhə* 'fear' is illustrated in (7).

- (7) የ'ሴፍ            ጎሳሰቴር  
*yosef*            *ni-ጎaster*  
 Yosef            OBJ-Aster  
 ኣፍሪሐ-ዋ  
*ጎafriḥ-u-wa*  
 fear.CAUS.PFV-SB3SM-O1,3SF  
 'Yosef frightened Aster.'

The CAUSATIVE forms appear frequently in the stimulus only pattern, like the CAUSATIVE of intransitive SE verbs, as illustrated in (8).

- (8) ዝብላ            የፍርሐ  
*zibiጎi*            *yəfirriḥ*  
 hyena            fear.CAUS.IPFV.SB3SM  
 ኣዩ  
*ጎጃጎጎ*  
 AUX.PRS.SB3SM  
 'A hyena is scary.'

The corpus data reveal that the roots in this class differ significantly with respect to transitivity, with ሳፈቕ *nafəx'ə* 'miss' taking an OBJ1 suffix in 57% of the sentences in the BASE form, while this is true for only 21% of the sentences with the BASE form of ፈርሐ *fərhə* 'fear.' On the other hand, ፈርሐ *fərhə* 'fear' has a common PASSIVE form, whereas there are no instances of the PASSIVE form of ሳፈቕ *nafəx'ə* 'miss' in the data.<sup>4</sup>

## 7.2 Object-experiencer verbs

Because EXPs are not prototypical agents and may be construed with varying degrees of control, they often appear as objects of different sorts and in many languages, for example, Icelandic (Barðdal,

<sup>4</sup>In fact the passive form of ሳፈቕ *nafəx'ə* is possible in the language, for example, ኣቲ ዝሓለፈ ጊዜ ተሳፊቕ ኣሎ። *ጎiti zihələfə gize tənafix'u ጎallo* 'The past time has been missed.' This shows that we need to be cautious about concluding that a form is not possible simply because it fails to occur in the data.

1999), Faeroese (Barnes, 1986), and Greek (Landau, 2009), have quirky properties not characteristic of canonical transitive sentences.<sup>5</sup>

### 7.2.1 OE verbs with ambient stimuli

For one set of object-experiencer (OE) roots in the BASE form, Tigrinya shows a mismatch in case and pronominal marking: the EXP is treated morphologically as the object of the verb but syntactically it shows a split transitivity combining subject and object properties which according to Malchukov (2005) arises from a functional tension to foreground the most prominent argument, i.e. the experiencer.

In (9) the EXP, Aster is optionally marked with the objective case. When it appears as a bare noun, superficially like a canonical subject because it lacks the objective prefix that is normally obligatory for a definite direct object, but agrees with the 3rd person singular feminine OBJ1 suffix on the verb.

- (9) (?) ኣሰቴር ጸጫኡ-ዋ  
*ni-ጎaster s'əmi-u-wa*  
 Aster            be.thirsty.PFV-SB3SM-O1,3SF  
 'Aster is thirsty.'

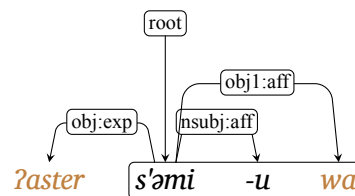


Figure 3: Dependencies in (9). Brown tokens represent the EXP.

Because the EXP nominals in such sentences are optionally marked with the objective affix and agree with the obligatory morphological object, we annotate them with the *obj* relation but add the sub-relation *:exp* to distinguish them from

<sup>5</sup>Landau (2009) identifies three types of languages based on a quiriness scale. The first group comprises languages that have various options to code the EXP as dative, accusative and genitive quirky subjects, with Icelandic, Faeroese and Greek as typical examples. The second group allows only dative EXPs as subjects, with languages such as Italian, Spanish and Dutch showing this pattern. The third group does not allow quirky EXPs; that is, only nominative subjects can be used for EXP. Such languages include English, French and Hebrew. As we will see, Tigrinya is closest to the second group.



canonical direct objects, which require the objective prefix when definite. Roots of this type include *ጸሞኦ ስ'ጠጠጠ* ‘be thirsty’, *ጠሞየ ጥ'ጠጠጠ* ‘be hungry’, *ደኸሞ ልጸጸጠ* ‘be tired’, *ጸሞሠ ስ'ጠጠጠጠ* ‘feel lonely’, *ሰልጥሠ ስጠጠጠ* ‘be bored’, and *ኣሞሞ ከጠጠጠ* ‘be sick, hurt’.

Morphologically, the subjects of these OE verbs are 3SM, similar to what Pesetsky (1995, 111) refers to as the unspecified stimuli behind “emotional weather,” and, on the surface at least, identical to what Amberber (2005, 295), describing Amharic, calls “ambient causers”. We will refer to clauses of this type as “ambient stimulus” (AS) clauses.

The picture is complicated by the fact that many of these roots also belong to the ISE category; for this reason Kifle (2011) treats them as applicative alternations. For example, the English gloss in (9) has another possible translation in Tigrinya, illustrated in (10).

- (10) ኣሰቲር ጸሞኦ  
*?aster s'ጠጠጠ-a*  
 Aster be.thirsty:PFV-SB3SF  
 ‘Aster is thirsty.’

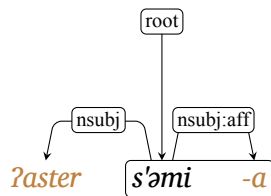


Figure 4: Dependencies in (10). Brown tokens represent the EXP.

In (10) the root *ጸሞኦ ስ'ጠጠጠ* behaves like an SE root; the verb’s subject agrees with the EXP, Aster. In (9), on the other hand, the same root behaves like an ASOE root; the verb’s object suffix agrees with the EXP and the EXP nominal has no case marker.

Examining the corpus data, we discover that the roots in this category differ strikingly with respect to their frequency of occurrence in the ISE and ASOE patterns. While the BASE forms *ጸሞሠ ስ'ጠጠጠጠ* ‘feel lonely’ and *ሰልጥሠ ስጠጠጠ* ‘be bored’ have 3SM subjects with OBJ1 suffixes on the verb (indicating the ASOE pattern) in 40% and 82% of the instances, respectively, these proportions drop to 2.4% for *ደኸሞ ልጸጸጠ* ‘be tired’,

2.2% for *ጠሞየ ጥ'ጠጠጠ* ‘be hungry’, and 0.9% for *ኣሞሞ ከጠጠጠ* ‘be sick’.

Interestingly, two of the clearly related roots in Amharic exhibit quite different patterns: the proportion of 3SM subjects with object suffixes in the BASE form is 23% for *ደኸሞ ልጸጸጠ* ‘be tired’ and 90% for *ኣሞሞ ስጠጠጠ* ‘be sick.’ Another notable difference is that the Tigrinya roots in this category have no PASSIVE forms, while the PASSIVE forms for Amharic roots like *ኣሞሞ ስጠጠጠ* ‘be sick’ not only exist but are quite common.

For at least some of the roots that belong to both the ASOE and ISE categories, a further argument representing a generic stimulus is possible. In (11), *ሞይ ጠጠ* ‘water’ adds no information at all about the nature of the stimulus behind the state. Though not related etymologically to the verb, such an argument is analogous to “cognate objects” in other languages, for example *death in he died a peaceful death* (Austin, 1982; Jones, 1988; Pesetsky, 1995; Börjars and Vincent, 2008). We will refer to it as an “internal object.”

- (11) ኣሰቲር ሞይ ጸሞኦ  
*?aster ጠጠ ስ'ጠጠጠ-ሀ-ሠ*  
 Aster water be.thirsty:PFV-SB3SM-O1,3SF  
 ‘Aster is thirsty (for water).’

At least for this root, the internal object is also possible when the root appears in the ISE pattern.

- (12) ኣሰቲር ሞይ ጸሞኦ  
*?aster ጠጠ ስ'ጠጠጠ-a*  
 Aster water be.thirsty:PFV-SB3SF  
 ‘Aster is thirsty (for water).’

We annotate the internal object as *obl:internal* in both (11) and (12).

Some of the roots in this category permit an explicit STIM argument that takes the form of the subject, so these then resemble the roots described in the next section. There is apparently a limited set of possible STIM subject arguments for these roots. With the BASE form of the root *ደኸሞ ልጸጸጠ* ‘tire, be tired,’ the noun *ኣዞል ከጎላ* ‘strength’ with a possessive suffix is a common subject, as illustrated in (13). As in the AS pattern, we annotate the EXP in such sentences as *obj:exp* because it does not require the objective prefix when definite.

- (13) ኣሰቲር ሓይላ ደኺሙዋ  
*?aster hayl-a daxim-u-wa*  
 Aster strength-her tire.PFV-SB3SM-O1,3SF  
 ‘Aster is tired.’

For the root ሓመመ *haməmə* ‘sicken, be sick, hurt’, the body part where the experience is centered may appear as the subject, as illustrated in (14). Again the EXP takes the form of an obj : exp, without the normal obligatory case marking.

- (14) ኣሰቲር ርእሳ ሓመዋ  
*?aster riʔis-a him-u-wa*  
 Aster head-her hurt.PFV-SB3SM-O1,3SF  
 ‘Aster’s head hurts.’

For other roots in this category, if the speaker wants to refer to an explicit STIM, the CAUSATIVE form must be used, with the EXP in the form of a canonical object, that is, with an OBJ1 agreement suffix on the verb and the obligatory objective prefix on the nominal if definite. We annotate the EXP argument as obj. Sentences in this pattern may also include the internal object, which we annotate as iobj : internal in this case. See (15), in which ጨው *č’əw* ‘salt’ is the nsubj, ኣሰቲር *?aster* is the obj, and ማይ *may* ‘water’ is the iobj : internal.

- (15) እቲ ጨው ንኣሰቲር ማይ  
*?iti č’əw ni-?aster may*  
 the salt OBJ-Aster water  
 ኣጽግኡዋ  
*?as’mi-u-wa*  
 thirsty.CAUS.PFV-SB3SM-O1,3SF

‘The salt made Aster thirsty (for water).’

### 7.2.2 OE verbs with explicit stimuli

For other OE roots, such as ገረመዋ *gəramə* ‘surprise, be surprising,’ an explicit STIM subject is possible with the root’s BASE form. This is illustrated in (16).

- (16) ሰርሑ ንኣሰቲር  
*sirhu ni-?aster*  
 action.his OBJ-Aster  
 ገረመዋ  
*gərim-u-wa*  
 surprise.PFV-SB3SM-O1,3SF  
 ‘His action surprised Aster.’

These roots may also appear in the ambient stimulus pattern, in which case the EXP, if definite, no longer requires the objective prefix.

- (17) ኣሰቲር ገረመዋ  
*?aster gərim-u-wa*  
 Aster surprise.PFV-SB3SM-O1,3SF  
 ‘Aster is surprised.’

With such roots, the BASE form may also appear in the stimulus only pattern, as illustrated in (18).

- (18) ሰርሑ ይገርም እዩ  
*sirhu yigərrim ?əyyu*  
 action.his surprise.IPFV-SB3SM AUX-SB3SM  
 ‘His action is surprising.’

Other roots in this category include ጨነቕ *č’anəx’ə* ‘worry,’ ኣገመ *s’əggəmə* ‘trouble’, and ሃወኸ *hawwəxə* ‘disturb.’

Not surprisingly, the roots in this category have a PASSIVE form, in which the EXP is the subject and the STIM is an ob1, as in (19).

- (19) ኣሰቲር ብትዕግስቲ  
*?aster bi-tiʔgistu*  
 Aster by-patience.his  
 ተገረማ  
*təgərrim-a*  
 surprise.PFV.PASS-SB3SM-O1,3SF  
 ‘Aster is surprised by/with his patience.’

But many of the roots also have a CAUSATIVE form, in which the EXP and STIM are realized as with the BASE form, as object and subject respectively. See 20, in which the subject, ትዕግስቲ *tiʔgistu*, functions as STIM.

- (20) ትዕግስቲ ንኣሰቲር  
*tiʔgistu ni-aster*  
 his.patience OBJ-Aster  
 ኣገረመዋ  
*?agərrim-u-wa*  
 surprise.PFV.CAUS-SB3SM-O1,3SF  
 ‘His patience surprised Aster.’

There are cases where the BASE and CAUSATIVE forms of such roots are interchangeable but others in which the experience CAUSER and STIM are separated, as in (21). In these cases the CAUSATIVE

form of the root is required, the CAUSER is the nsubj, and the STIM is realized as an obl argument.

- (21) የሴፍ ንእስቴር ብትዕግስቱ  
*yosef ni-aster bi-tiṣgistu*  
 Yosef OBJ-Aster by-his.patience  
 ኣገረሙዋ  
*?agərrim-u-wa*  
 surprise.PFV.CAUS-SB3SM-O1,3SF

‘Yosef surprised Aster with his patience.’

### 7.3 Summary of categories

Here we summarize the morphosyntactic schemas we have described for experience clauses in Tigrinya. Syntactic and morphological relations are separated by a slash when they are different. To simplify, the morphological subrelation :aff is not included.

- Subject Experiencer: Intransitive  
 Example: ሰንበደ *sənbədə* ‘be alarmed’
  - Base: ሰንበደ *sənbədə*
    - \* EXP: nsubj
  - Causative: ኣሰንበደ *?asənbədə*
    - \* EXP: obj/obj1
    - \* STIM: nsubj
- Subject Experiencer: Transitive  
 Example: ጸልኦ *s’əl?ə* ‘hate’
  - Base: ጸልኦ *s’əl?ə*
    - \* EXP: nsubj
    - \* STIM: obj/obj1
  - Passive: ተጸልኦ *təs’əl?ə*
    - \* STIM: nsubj
  - Causative: ኣጸልኦ *?as’li?ə*
    - \* EXP: obj/obj1
    - \* STIM: nsubj
- Object Experiencer: Ambient Stimulus  
 Example: ጸመወ *s’əmməwə* ‘feel lonely’
  - Base: ጸመወ *s’əmməwə*
    - \* EXP: obj : exp/obj1
  - Causative: ኣጸመወ *?as’əmməwə*
    - \* EXP: obj/obj1
    - \* STIM: nsubj
- Object Experiencer: Explicit Stimulus  
 Example: ገረሙ *gəramə* ‘surprise’

- Base: ገረሙ *gəramə*
  - \* EXP: obj/obj1
  - \* STIM: nsubj
- Passive: ተገረሙ *təgərrəmə*
  - \* EXP: nsubj
- Causative: ኣገረሙ *agərrəmə*
  - \* EXP: obj/obj1
  - \* STIM: nsubj

## 8 Conclusions and Future Work

Our investigation has uncovered two categories within each of the Subject Experiencer and Object Experiencer classes of experience verbs in Tigrinya, each defined by a schema for each of the two or three forms that occur for the roots in the category. We have also seen that many roots belong to more than one category. In particular, roots such as ደኸሙ *dəxəmə* ‘tire, be tired’ occur in both the Ambient Stimulus Object Experiencer and Intransitive Subject Experiencer categories. We might guess that the use of these roots in the Subject Experiencer pattern implies a more active role for the EXP, something we plan to explore in future work. We have also learned that specific roots may permit arguments that are not possible with others in the same category, for example, body part subjects with ሓመመ *haməmə* ‘hurt’ and internal objects with ጸምኦ *s’əm?ə* ‘be thirsty’.

We have not exhausted all of the possibilities for the arguments of experience predicates in the language. For example, the verb ሓመመ *haməmə* ‘be sick, hurt’ can take a malefactive argument representing a participant who is harmed by the experiencer’s pain or ailment. In future work we plan to investigate the morphosyntax associated with these arguments and propose UD relations for them.

Finally, the existence of morphological data on thousands of Tigrinya roots opens up the possibility of classifying experience roots on the basis of their similarity to the categories we have outlined in the paper.

## References

- Mengistu Amberber. 2005. Differential subject marking in Amharic. In Mengistu Amberber and Helen de Hoop, editors, *Competition and Variation in Natural Languages: The Case for Case*. Elsevier.
- Peter Austin. 1982. Transitivity and cognate objects in Australian languages. In *Studies in transitivity*, pages 37–47. Brill.

- Jóhanna Barðdal. 1999. The dual nature of Icelandic psych-verbs. *Working papers in Scandinavian syntax*, 64:79–101.
- Michael Barnes. 1986. Subject, nominative and oblique case in Faroese. *Scripta Islandica*, 37:13–46.
- Adriana Belletti and Luigi Rizzi. 1988. Psych-verbs and  $\theta$ -theory. *Natural Language & Linguistic Theory*, pages 291–352.
- Kersti Börjars and Nigel Vincent. 2008. Objects and obj. In *Proceedings of the LFG08 Conference*, pages 150–168.
- William Croft. 1993. Case marking and the semantics of mental verbs. In *Semantics and the Lexicon*, pages 55–72. Springer.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*.
- Girma Awgichew Demeke. 2003. *The clausal syntax of Ethio-Semitic*. Ph.D. thesis, University of Tromsø.
- David Dowty. 1991. Thematic proto-roles and argument selection. *language*, 67(3):547–619.
- Tekabe Legesse Feleke. 2021. [Ethiosemitic languages: Classifications and classification determinants](#). *Ampersand*, 8:100074.
- Jens Fleischhauer. 2016. *Degree Gradation of Verbs*. Düsseldorf University Press.
- Andargachew Mekonnen Gezmu. 2023. *Subword-based neural machine translation for low-resource fusion languages*. Ph.D. thesis, Otto-von-Guericke-Universität Magdeburg.
- Rivka Halevy. 2023. Non-subject oriented existential, possessive and dative-experiencer constructions in modern hebrew—a cross-linguistic typological approach. *STUF-Language Typology and Universals*, 76(4):545–585.
- Robert Hetzron. 1972. *Ethiopian Semitic: studies in classification*. 1. Manchester University Press.
- Michael Allan Jones. 1988. Cognate objects and the case-filter. *Journal of Linguistics*, 24(1):89–110.
- Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 2021. A morph-based and word-based treebank for Beja. In *20th International Workshop on Treebanks and Linguistic Theories*.
- Dirk Kievit and Saliem Kievit. 2009. [Differential object marking in tigrinya](#). *Journal of African Languages and Linguistics*, 30(1):45–71.
- Nazareth Amlesom Kifle. 2007. Differential object marking and topicality in Tigrinya. In *Proceedings of the LFG*, volume 7, pages 5–25.
- Nazareth Amleson Kifle. 2011. *Tigrinya Applicatives in Lexical Functional Grammar*. Ph.D. thesis, University of Bergen.
- Katarina Klein and Silvia Kutscher. 2015. *Lexical economy and case selection of psych-verbs in German*. Institut für Deutsche Sprache, Bibliothek.
- Idan Landau. 2009. *The locative syntax of experiencers*. MIT press.
- Andrej Malchukov. 2005. Split intransitives, experiencer objects and ‘transimpersonal’ constructions:(re-) establishing the connection.
- Åshild Næss. 2007. *Prototypical Transitivity*. John Benjamins.
- David Michael Pesetsky. 1995. *Zero syntax: Experiencers and cascades*. 27. MIT press.
- David Michael Pesetsky. 2000. *Phrasal movement and its kin*. MIT press.
- Paul Martin Postal. 1971. *Cross-over phenomena*. Holt, Rinehart and Winston.
- Elisabeth Verhoeven. 2014. Thematic prominence and animacy asymmetries. Evidence from a cross-linguistic production study. *Lingua*, 143:129–161.
- Desalegn Workneh. 2019. *The Voice System of Amharic*. Ph.D. thesis, Arctic University of Norway.

# A Corpus-driven Description of OV Order in Archaic Chinese

Qishen Wu<sup>1</sup> Santiago Herrera<sup>1</sup> Pierre Magistry<sup>2</sup> Sylvain Kahane<sup>1</sup>

<sup>1</sup>MoDyCo, Université Paris Nanterre

<sup>2</sup>ERTIM, INALCO, Paris

{qishen.wu, sherrera, skahane}@parisnanterre.fr, pierre.magistry@inalco.fr,

## Abstract

This paper presents a quantitative study of Object-Verb (OV) order in Archaic Chinese based on a Universal Dependencies (UD) treebanks. Treating word order as a binary choice (OV vs VO), we train a sparse logistic-regression classifier that selects the most salient syntactic features needed for an accurate prediction to investigate the specific syntactic contexts allowing OV word order and to identify to what extent do these factors favor this order. The ranked features are understood as interpretable rules, and their coverage and precision as quantitative properties of each rule. The approach confirms earlier qualitative findings (e.g. pronoun object fronting and negation favor OV) and uncovers new contrasts in word order between different reflexive pronouns. It also identifies annotation errors that we corrected in the final analysis, illustrating how the quantitative models, combined with fine-grained corpus analysis, can improve treebank quality. Our study demonstrates that lightweight machine-learning techniques applied to an existing syntactic resource can reveal fine-grained patterns in historical word order and this can be reapplied to other languages.

## 1 Introduction

In this paper, we investigate Object-Verb (OV) word order in Archaic Chinese, following the historical linguistic periods of Chinese language proposed by Wang (1980). Wang defines Archaic Chinese as the historical stage extending from the Shang Dynasty (circa 1600 - 1046 BCE) and ending in the early Han Dynasty (206 BCE - 25 CE). This historical stage covers oracle bone inscriptions, bronze inscriptions, and classical texts from the pre-Qin period up to early Han period. Our study primarily focuses on representative texts from Eastern Zhou Dynasty (circa 771 - 256 BCE) to the early Han Dynasty. Our main goal is to

systematically explore syntactic structures, particularly the conditions favoring or promoting OV order.

Most scholars currently agree that a stable SVO word order was already established by the time of Archaic Chinese. For instance Wang (1980) and Feng (2013) argued that during the Pre-Archaic period, the basic word order of Pre-Archaic Chinese was likely SOV, but as the language evolved, an SVO word order was established by the Archaic Chinese period. Ma (1898) highlight that object is placed after verbs in Archaic Chinese. More recently, Peyraube (1997) argue that Archaic Chinese was always a SVO language and OV word order is used only in specific syntactic contexts that are strictly constrained. Djamouri (2014) also states that the SVO word order had already become the dominant order in the Shang oracle bone inscriptions, based on a detailed study of 5,500 complete sentences from the Shang Dynasty (ca. 1600–1046 BCE), among which 94% followed the SVO order, while only 6% exhibited an SOV order.

Therefore, based on the research of these scholars, we can assume that Archaic Chinese is a stable SVO language. However, when exploring Archaic Chinese corpora, we find that although the OV word order is not dominant, it appears to be far from uncommon. In many syntactic contexts, OV word order is allowed and sometimes even preferred, which means that the VO order may remain productive in Archaic Chinese. In this regard, Yu (1981) highlight certain features of SOV language in Pre-Archaic Chinese, such as modifiers occurring before heads, which partially explain the possible origins of SOV word order in Archaic Chinese. Peyraube (1997) identifies four cases of OV word order in Archaic Chinese. Finally, Pan and Jiao (2023) conclude that OV order in Archaic Chinese can be roughly classified into unmarked object-fronting and marked object-fronting, after consulting previous studies on object-fronting in Archaic

Chinese by several different linguists. They also provide more detailed syntactic context in which OV word order can be used.

However, existing research remains predominantly qualitative in nature, mainly focusing on identifying potential syntactic environments and describing by hand the conditions that allow OV word order in Archaic Chinese. Besides, previous studies have suggested that there is no absolute grammatical rule enforcing OV order; instead, OV structures represent a possible syntactic choice within certain contexts. Therefore, given the observed variability in scientific statements about OV order, a quantitative exploration of syntactic tendencies becomes essential. Our aim is to undertake a more rigorous exploration, grounded in corpus data, to ascertain the specific manifestations of OV word order within Archaic Chinese corpora. In this paper, building upon previous theoretical research on OV word order, we aim to clarify the following research question: What specific syntactic contexts systematically permit or prefer OV order in Archaic Chinese and to what extent do these factors favor the OV order? To address this question, we adopt quantitative methods from computational linguistics, analyzing a syntactically annotated treebank to precisely identify and quantify the syntactic conditions that lead to object fronting.

## 2 Related Work

**OV order in Archaic Chinese** Former researches has showed that the use of OV word order is typically governed by strict syntactic contexts or specific semantic purposes. Most of these studies have adopted qualitative approaches to explore the occurrences and cases of OV word order in Archaic Chinese, summarizing and analyzing specific cases based on corpora and textual examples. As early as the end of the 19th century, [Ma \(1898\)](#) has noted that in Archaic Chinese, when a verb is preceded by a negation element or when the direct object is an interrogative pronoun, the corresponding direct object must be fronted. Since the 1980s, numerous scholars have analyzed the occurrences of the demonstrative pronoun *shì* 是(this) and the third-person pronoun *zhī* 之(3-person) in OV order when they appear between the direct object and the verb. For instance, [Wang \(1989\)](#) and [Xiang \(2010\)](#) argue that in OV word order, they function as resumptive pronouns referring to the object NP. In contrast, [Han \(1996\)](#) interprets *shì* 是(this) and

*zhī* 之(3-person) as grammatical markers in OV constructions. Towards the end of the 20th century, [Peyraube \(1997\)](#) summarized the instances of OV word order in Archaic Chinese, identifying four main types and the following four examples are taken from [Peyraube \(1997\)](#):

1. The object is an interrogative pronoun;

(1) 子 何 言  
zi hé yán  
you what say  
What do you say?

2. The object is the demonstrative pronoun *shì* 是(this);

(2) 子 孫 是 保  
zi sūn shì bǎo  
son grandson this preserve  
The future generations (will) preserve this.

3. The object is a pronoun in negative sentences;

(3) 不 吾 知 也  
bù wú zhī yě  
negation IPRON understand final-part  
(You) don't know me.

4. The object is a noun phrase (NP) followed by a preverbal object marker *shì* 是(this) or *zhī* 之(this).

(4) 四 方 是 維  
sì fāng shì wéi  
four region objet-marker unite  
(You should) unite the four region.

Very recently, [Pan and Jiao \(2023\)](#) categorize OV constructions into marked and unmarked types. Unmarked OV constructions include the fronting of WH-pronominal objects; in declarative sentences, the fronting of the demonstrative *shì* 是(this); in negative sentences, the fronting of demonstrative pronouns *zhī* 之(this), *shì* 是(this) and *cǐ* 此(this), personal pronouns, and noun phrases (NPs). In addition, marked OV constructions are further subdivided based on the type of marker into *wéi* 唯(only)-type, *shì* 是(this)/*zhī* 之(this)-type, and



wéi 唯(only)... shì 是(this)/ zhī 之(this)-type constructions. For each type, the authors provide specific examples from the corpus and detailed analyses.

### Grammar Study and Computational Linguistics

In recent decades, the intersection of grammar study and computational linguistics has gained increasing attention, leveraging computational tools to deepen the understanding of linguistic structures and phenomena. This interdisciplinary approach has significantly transformed linguistic research. Vlachos and Craven (2010) parse biomedical text to extract features based on syntactic dependency relations, then they feed a sparse Bayesian logistic-regression model with extracted features to classify speculative language. This method improves the model’s ability to recognise phrases that express uncertainty. The development of annotated corpora, such as dependency treebanks of Universal Dependencies (UD) frameworks (de Marneffe et al., 2014; Nivre et al., 2020; de Marneffe et al., 2021) which includes treebanks for more than 160 languages, have provided robust datasets for quantitative and comparative analyses. By analyzing data from the Universal Dependencies project, Gerdes et al. (2021) introduces a quantitative approach to linguistic typology, which move beyond traditional implicational universals to identify quantitative patterns in word order typology. This method allows for a more nuanced understanding of syntactic structures across languages, highlighting statistical tendencies rather than absolute rules. Levshina (2019), promoting a token-based typology, measures word-order variability with Shannon entropy calculated from Universal Dependencies data for about 60 languages. They find that languages fall into three clusters: high-entropy morphologically rich VO/flexible orders, mid-entropy analytic VO languages, and low-entropy OV languages. Chaudhary et al. (2022) developed a framework to assist linguists in the extraction of comprehensible syntactic rules, specifically focusing on morphological agreement, case marking, and word order. This system was validated across multiple languages, demonstrating its capability to generalize and apply linguistic rule extraction effectively in diverse language contexts. More recently, Herrera et al. (2024) introduces a novel method for inferring and mining, in a more exploratory design, detailed syntactic rules from treebanks: by employing sparse logistic regression enhanced with a richer feature

search space, they effectively identify significant grammatical patterns, particularly for agreement and word order in Spanish, French, and Wolof, successfully uncovering both well-known and underexplored syntactic tendencies and rules.

### 3 Corpus and Method

In our study, we use Kyoto University’s UD syntactic treebank of Classical Chinese in its version 2.15 (Yasuoka, 2019; Yasuoka et al., 2022). This corpus comprises an extensive dataset of 86,239 sentences. Our study focuses on the Archaic Chinese subset of the corpus, which includes three Confucian texts "Lún yǔ" 論語, "Lǐ jì" 禮記, "Mèng zǐ" 孟子, one classical poetic text "Chǔ cí" 楚辭 and one historical text "Zhàn guó cè" 戰國策. This sub-corpus consist of a total of 55,632 sentences, providing a rich resource for in-depth linguistic analysis.

Our aim, as stated above, is to identify quantitative and gradient rules or tendencies that favor OV word order in Archaic Chinese. For that, it is essential first to define what constitutes for us a quantitative grammar rule. Inspired by correspondence rules of the Meaning-Text Theory (Melcuk, 1988) and by Chaudhary et al.’s (2022) work, Herrera et al. (2024) formalise a grammatical rule with three elements or patterns: **the scope**  $S$ , which is the domain within which the specific grammatical phenomena under investigation may occur. In our case this consists of occurrences of verb with an object; **the target linguistic phenomenon**  $Q$  that has to be predicted, which is for us the object preceding the verb; and **the predictor pattern**  $P$ , in our study, the syntactic context that allows object fronting. Consequently, our aim is to investigate what are the syntactic contexts  $P$  that allow object fronting in Archaic Chinese, and to what extent ( $\alpha\%$ ) of object fronting is likely to occur under each of these conditions.

$$S \implies (P \xrightarrow{\alpha\%} Q)$$

This formalization captures both the probabilistic nature of grammar and the overlapping relationships between grammar factors, making it highly adaptable to diverse linguistic frameworks and phenomena. By associating features with specific syntactic contexts, this approach offers a quantitative yet interpretable method for modelling grammar.

We then adopted the method described in Herrera et al. (2024) which use a linear classifier to extract the most salient features that predict the

linguistic phenomena. This method has been tested by using syntactic treebank corpora of English, French, Spanish and Wolof, demonstrating its applicability across different languages. This particular method was selected due to its inherent tendency to favor an exploratory approach.

More specifically, to identify the syntactic contexts that most strongly predict OV order, the authors use sparse logistic regression classifier to distinguish between OV and VO constructions based on the syntactic features extracted from the treebank. The feature space employed by the classifier is, as a matter of implementation, determined by manual specification. In our case, it consists for each node defined in the scope, i.e. the verb and the object, the following UD features: part-of-speech tags, dependency relations, morphological features (such as pronoun type), and clause-level modifiers (like presence of negation or sentence particles)<sup>1</sup>. We did not include lexical forms in the initial model to generalize across surface variation, though this remains a direction for future work.<sup>2</sup>

The classifier is tasked with estimating the probability of object fronting or not given a set of features. Once the model is trained, we examine the features that most heavily influence its decision-making process, specially the features corresponding to the syntactic conditions under which OV order is most likely to occur.

The authors use L1-regularisation to rank the most informative features for predicting syntactic phenomena. Specifically, they train the model for  $k + 1$  regularization strengths  $\alpha_i$ ,  $0 \leq i \leq k$ , which controls sparsity through the regularization strength parameter  $\alpha$ . When  $\alpha$  is large, only the most relevant features are retained; as  $\alpha$  decreases, additional features gradually enter the model as their associated weights become non-zero. The  $k$  is set to 100 by default, with  $\alpha_0$  set to 0.01 and  $\alpha_k$  set to 0.001. This built-in feature selection keeps only the most informative syntactic factors and suppresses noisy features. Because the surviving features and their weights are directly inspectable, the model is far more transparent than neural models and therefore well suited to recognising and interpreting the grammatical patterns that govern

<sup>1</sup>features appearing fewer than 5 times in the corpus were excluded to reduce noise

<sup>2</sup>To extract more general features, we initially did not include orthographic forms in our analysis. However, based on the preliminary results, we may consider incorporating orthographic forms as an influencing feature in future research.

object fronting in Archaic Chinese. To test the statistical significance of each grammatical rule, the authors applied a G-test comparing the observed and expected distributions different features. The G statistic approximates a  $\chi^2$  distribution, and p-values were computed accordingly. Features with  $p < 0.01$  were considered statistically significant, indicating that the observed association between the grammatical feature and the target linguistic phenomenon is unlikely to occur by chance.

The authors also compute some statistical measures for each extracted pattern to understand its behaviour within the corpus. These measures are coverage and precision and are calculated as follows:

$$\text{Coverage} = \frac{\#(S \wedge P \wedge Q)}{\#(S \wedge Q)},$$

$$\text{Precision} = \frac{\#(S \wedge P \wedge Q)}{\#(S \wedge P)}$$

The **coverage** indicates among all OV occurrences, how likely the specific grammatical phenomenon occurs, whereas the **precision** measures among all occurrences exhibiting this selected feature, how likely they follow the OV order. For example, consider the grammatical feature "the object is an interrogative pronoun". If its coverage is 37%, this means that interrogative pronouns account for 37% of all OV occurrences of our corpus. If its precision is 73%, it indicates that among all instances where an interrogative pronoun serves as an object, 73% of these are fronted. High coverage suggests that the feature is common among OV constructions, whereas high precision implies that the feature strongly predicts OV word order.

## 4 Results

In this section, we analyze the syntactic factors of OV word order selected by the linear model. Among 55,632 sentences in the corpus, a total of 783 instances of OV word order were identified. The selected syntactic features significantly influencing OV word order are shown in the Table 1.<sup>3</sup>

The linear classifier identified nine grammatical factors that influence OV structures to varying

<sup>3</sup>We implemented our approach using the code provided by Herrera et al. (2024) in their paper. All results and data are included as supplementary material in paper submission portal.

	pattern $P$	occurrences of $P$	occurrences of $Q$	decision	coverage	precision	$\alpha$
1	Object is a pronoun	5509	601	OV	76.8	10.9	0.012
2	Object is an interrogative pronoun	395	292	OV	37.3	73.9	0.007
3	Object is Third-person pronoun	3961	39	VO	10.7	99.0	0.005
4	Verb has an adverbial clause modifier	2183	257	OV	32.8	11.8	0.004
5	Verb has an expletive	210	129	OV	16.5	61.4	0.003
6	Object is a reflexive pronoun	209	89	OV	11.4	42.6	0.001
7	Verb has Degree "Equ"	801	91	OV	11.6	11.4	0.001
8	Verb has a sentence particle	3783	155	OV	19.8	4.1	0.001
9	Verb has a negative modifier	2707	83	OV	10.6	3.1	0.001

Table 1: Top features selected by the classifier favoring OV order.

degrees<sup>4</sup>. Based on the  $\alpha$  values of each grammatical feature, we can observe that pronominal objects, as the most prominent feature, are identified first by the model. The next most significant feature is interrogative pronominal objects. The third significant grammatical feature—third-person pronominal objects—is somewhat special, as it is identified by the model as indicative of VO order<sup>5</sup>. The fourth feature selected is verbs with an adverbial clause modifier, while the fifth is verbs with an expletive. The final four grammatical features are distinguished at an  $\alpha$  value of 0.001, with no clear difference in their level of relevance. Based on this ordering, we can already see that pronouns and pronoun-related features play a significant role in object fronting. We then describe the top nine factors in detail.

Firstly, the data indicates that the first and the most salient factor involves objects that are pronouns. The coverage for fronted pronouns is as high as 76.8%, highlighting a strong focus on pronouns in OV word order in Archaic Chinese. However, the precision for fronted pronouns is only 10.9%, indicating that only 10.9% of clauses with pronouns as objects exhibit fronting. This suggests that while pronoun fronting is a prominent feature of OV word order, OV structures themselves are not dominant in Archaic Chinese, as most verbs with pronoun objects do not exhibit this pattern. The second factor reveals that interrogative pronouns as fronted objects are particularly significant in the corpus. The coverage shows that interrogative pronouns account for 37.3% of all fronted objects, indicating their prominence. Since interrogative pronouns are a subset of pronouns, this finding can be seen as a refinement of the first rule,

<sup>4</sup>All nine factors are statistically significant, with p-values below 0.01 (detailed results are provided in the supplementary material).

<sup>5</sup>In this case, coverage and precision are calculated for  $\neg Q$ : Coverage =  $\frac{\#(S \wedge P \wedge \neg Q)}{\#(S \wedge \neg Q)}$ , Precision =  $\frac{\#(S \wedge P \wedge \neg Q)}{\#(S \wedge P)}$

which shows that approximately half of the fronted pronouns are interrogative pronouns. Furthermore, the precision for this grammatical phenomenon is 73.9%, meaning that in cases where interrogative pronouns serve as objects, they are fronted in the vast majority of instances. This factor has been selected in the second place also highlighting a strong syntactic tendency for the fronting of interrogative pronouns in Archaic Chinese. These two factors are also consistent with Peyraube’s (1997) and Pan and Jiao’s (2023) research on object fronting in Archaic Chinese.

And the third rule reveals a grammatical situation where VO word order is clear favored, i.e. when the object is a third-person pronoun<sup>6</sup>. Coverage indicates that among all post-verbal objects, third-person pronouns account for 10.7%. However, this is an extremely precise rule: 99.0% of third-person pronouns used as objects are in fact postposed. At the same time, the extremely high precision of this rule and the fact that third-person pronouns used as fronted objects appear only 39 times in the corpus push us to consider whether these cases are very special or possibly due to annotation oversights or errors. Examining the occurrences, among the 3,961 occurrences of third-person pronouns used as objects, only two were *qi* 其(3-person), while the rest were *zhi* 之(3-person). Furthermore, all 39 fronted instances are *zhi* 之(3-person), each occurring as a fronted object in negative clauses, as shown in the example 5. This indicates that for the position of third-person pronoun objects, the decisive factor is actually another rule, namely, the influence of negation on object fronting. Moreover, in our corpus (which to some extent reflects features of Classical Chinese), the grammatical feature “third-person pronoun as object” can effectively serve as a criterion for VO word order.

<sup>6</sup>coverage and precision are calculated for  $\neg Q$

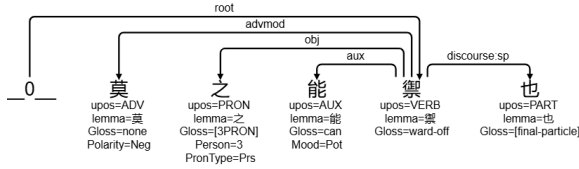


Figure 1: Fronting of *zhī* 之(3-person) in negative sentence.

- (5) 莫 之 能 禦 也  
 mò zhī néng yù yě  
 none 3PRON can ward-off final-particle  
 None can resist it.

The next rule indicates that in adverbial clauses, the object is fronted in 11.8% of cases. Previous qualitative studies on Archaic Chinese have not specifically analyzed object fronting in adverbial clauses. Therefore, we conducted more detailed analyses of this grammatical context in our corpus. After ruling out several clear annotation mistakes, we found that there are 234 instances where the object is a pronoun, in which interrogative pronouns *hé* 何(what) accounts for 47.5% (122/257) and the demonstrative pronoun *shì* 是(this) accounts for 37.4% (96/257). Additionally, the demonstrative pronoun *shì* 是(this) in this context always appears in the fixed structure *shì yǐ* 是以(because of). Therefore, we can conclude that this rule may not seem to hold genuine grammatical significance, and it is other correlated grammatical features that actually constrain OV word order.

We then individually examined the remained 23 instances where the fronted object was a noun. We then discovered two instances of incorrect part-of-speech tagging for the interrogative pronoun *hé* 何(what) and one instances that the noun object is modified by the interrogative word *hé* 何(what) to form a wh-phrases *hé gù* 何故(why), as illustrated in the example 6. The remaining cases involve noun-object fronting in non-negative clauses and non wh-phrases, which drew our attention. We found that in the remaining 20 example clauses, *yǐ* 以(use) appeared as a verb 17 times, while *yún* 云(say) appeared three times. Among them, *yún* 云(say) occurred in a well-structured clause (in example 7), and perhaps the structure and rhythm of the clause prompt the fronting of the object. The case for *yǐ* 以(use) is different. First, *yǐ* 以(use) appeared multiple times as a verb in the examples covered by this rule, forming fixed expressions such as *hé yǐ* 何以(how) and *shì yǐ* 是以(because

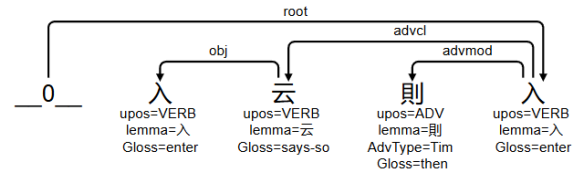


Figure 2: Fronting of object of *yún* 云(say).

of). Additionally, in these instances, there was also the phenomenon of a noun object being fronted. This might suggest that when *yǐ* 以(use) is used as a verb in instrumental constructions with nouns, it may, to some extent, also encourage the fronting of its object. However, these instances are too rare in our corpus to draw a definitive conclusion. We thus offer only a tentative hypothesis here, and a more detailed discussion will require further targeted research in the future with more specific data.

- (6) 隆 何 故 以 東 南 傾  
 dì hé gù yǐ dōng nán qīng  
 earth what reason use east south overturn  
 Why does the earth tilt to the southeast.
- (7) 入 云 則 入 ， 坐 云 則 坐 ，  
 rù yún zé rù , zuò yún zé zuò ,  
 enter say then enter , sit say then sit ,  
 食 云 則 食  
 shí yún zé shí  
 food say then feed  
 when told him to come into his house, he came; when told him to be seated, he sat; when told him to eat, he ate.

The fifth rule indicates that when a verb is modified by an expletive, 42.6% of its object is fronted, which means the 42.6% object is fronted with a marker. By examining coverage, we can see that fronted objects carrying such markers are not very common among all instances of object fronting. However, the precision shows that when a verb does have an expletive modifier, the object is more likely to be fronted. In light of this grammatical situation, we also examined examples from our corpus. After ruling out some clear annotation errors, we found that the primary word functioning as an expletive to modify the verb is *zhī* 之(3-person). There are also a few instances of *shì* 是(this) (12 occurrences, 4 of which are OV) and *sī* 斯(this) (2 occurrences, both VO). However, when we analyze the instances of *zhī* 之(3-person) in the corpus, we discover some subtle annotation issues. Because *zhī* 之(3-person) is a very commonly used syntactic



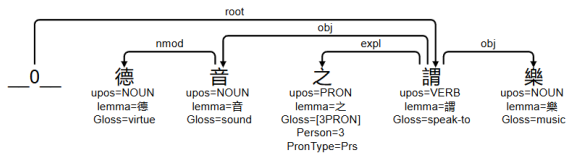


Figure 3: NP before *zhī* 之(3-person) considered as object of verb.

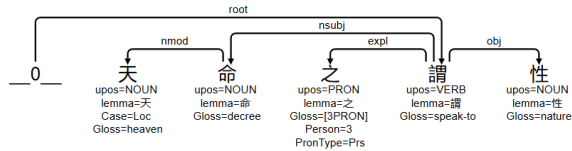


Figure 4: NP before *zhī* 之(3-person) considered as subject of verb.

word in Archaic Chinese with a wide range of grammatical functions, two occurrences containing *zhī* 之(3-person) that appear structurally similar can nevertheless be annotated quite differently in the corpus, as shown by the example 8 and 9. This illustrates that when performing quantitative analyses on an already annotated corpus, the choices made in annotation directly affect the analysis results. Therefore, although our quantitative methods are relatively swift and can straightforwardly provide coverage and precision for a concise set of grammatical phenomena within the corpus, we should still analyze specific corpus instances, as we have done above.

(8) 德 音 之 謂 樂  
 dé yīn zhī wèi yuè  
 virtue sound 3PRON use speak-to  
 The sound of benevolence is what makes true music.

(9) 天 命 之 謂 性  
 tiān mìng zhī wèi xìng  
 heaven decree 3PRON speak-to nature  
 A person's natural endowment is called "nature".

The sixth rule indicates that when the object is a reflexive pronoun, the object is fronted in 42.6% of cases. This rule does not have high coverage, but its precision is relatively high. Regarding the fronting of reflexive pronouns as objects, earlier qualitative research did not provide much discussion on this topic. Therefore, we conducted further research on reflexive pronouns in our corpus. In the corpus, there are two reflexive pronouns, *zì* 自(self) and *jǐ*

*己*(self), with the object *zì* 自(self) appearing 100 times and the object *jǐ* 己(self) appearing 109 times. However, these two reflexive pronouns differ significantly in their positioning when used as objects. After verifying the corpus and its annotations, we have found that all instances in which *zì* 自(self) was annotated as a post-verbal object turned out to be misannotations, all sharing the same structure: *shǐ/yǐ* 使/以(to make/ to use) + *zì* 自(self) + V. In this structure *zì* 自(self) was considered as an object of *shǐ/yǐ* 使/以(to make/ to use), but in fact *zì* 自(self) is the fronted object of the verb following it, and there are 15 such misannotations. Once corrected, *zì* 自(self) consistently appears as a fronted object in the corpus, regardless of whether the context is negative. By contrast, *jǐ* 己(self) as an object tends to be post-verbal, with only four instances of fronting, all of which occur in negative contexts. In comparison, when used as an object, *zì* 自(self) seems more like a reflexive pronoun specifically dedicated to fronting, whereas *jǐ* 己(self) does not exhibit this tendency.

The seventh rule is a more finely specified grammatical feature indicating the degree expressed by the verb, with “Equ” denoting “equal.” In the corpus, the only verbs annotated in this manner are *rú* 如(be-like) and *ruò* 若(be-like). Moreover, upon excluding potential annotation errors and clause structures requiring further analysis, we find that the fronted objects under this rule are all interrogative pronouns. The eighth rule likewise shows a similar pattern. It states that when the verb is modified by a discourse element, specifically a sentence particle, 4.1% of object is fronted. However, both the coverage and precision of this rule are quite low. Upon reviewing the corpus data, we found that in those instances identified by the model, the genuinely decisive grammatical features for OV word order are actually interrogative-pronoun objects, demonstrative-pronoun objects, and object fronting in negative clauses. The classifier identifies this rule primarily because sentence particles occur very frequently in the corpus (nearly one-fifth of OV occurrences contain sentence particles) so the model treats this grammatical phenomenon as a distinct rule. Therefore, in connection with rules 4, 7, and 8, we have identified an aspect of the linear model that requires further refinement. Some grammatical factors identified by the model as influencing OV word order may actually be coincidental by products of large-scale data, and in reality, other grammatical factors are what truly

determine OV word order. We thus need to examine actual corpus instances to validate the model's results.

Finally, let us turn to the ninth rule, which indicates that when a verb is modified by a negative element, 3.1% of the object is fronted. This rule aligns with previous qualitative research, but since its coverage and precision are both low, we also need to analyze the corpus data. In our corpus, the negative elements that can modify a verb include *bù* 不(not), *mò* 莫(none), *wèi* 未(not-yet), *fú* 弗(not), *wú* 無(not-have), *fēi* 非(not), *wù* 勿(don't), and *wú* 毋(don't). The fronted objects under this rule include *zhī* 之(3-person), which is only fronted in negative clauses as mentioned in the third rule, the reflexive pronouns *zì* 自(self) and *jǐ* 己(self) (mentioned in the sixth rule), the first-person pronouns *wǒ* 我, *wú* 吾, *yǔ* 予, *yú* 余, and certain nouns or noun phrases some ending with *zhě* 者(the person/thing that ...). We can see that although this rule has low coverage and precision, it differs from the scenarios in rules 4, 7, and 8, as the instances in question are not entirely determined by other grammatical factors, and the negative element exerts a considerable influence on fronting the 3-person pronoun object *zhī* 之(3-person). The low coverage and precision of this rule may be due to the rarity of such object-fronting cases or because the dataset we used is not sufficiently comprehensive.

## 5 Conclusion

After carefully analyzing the model-generated results and validating them against the corpus, we conducted a detailed exploration of our research questions and obtained satisfactory outcomes: we have provided a corpus-driven quantitative analysis of OV word order, addressing the previously unexplored quantitative dimension of object fronting in Archaic Chinese.

First, we can observe that, the syntactic features automatically selected by the linear model align with the grammatical rules summarized by traditional linguistic studies: we see the importance of pronouns in object fronting in Archaic Chinese, especially interrogative pronouns, the tendency for object fronting in negative clauses, and the different roles of *zhī* 之(3-person) as various constituents in object-fronting constructions. Secondly, the quantitative generalizations from the regression model also helped us identify a new syntactic feature with significant influence on object fronting: the strong

tendency for fronting when “*zì*” 自(self) when used as an object. Besides rather than merely identifying OV occurrences in Archaic Chinese, our results quantitatively demonstrate how common different object-fronting phenomena are, their productivity, and the strength of their tendencies.

Overall, this study demonstrates the significant potential of integrating quantitative computational methods into historical linguistic research, opening new avenues for systematic exploration of syntactic variation and grammatical structures in ancient languages.

## 6 Limitations and Future Research

However, our current approach also has certain limitations.

Firstly, corpus-based research is constrained by the limitations of the corpus itself. When we look closely at the specific content of the corpus, we still find issues arising from annotations: we discovered inconsistent annotations of identical sentence structures involving *zhī* 之(3-person), as well as part-of-speech tagging errors for *hé* 何(what). Moreover, due to the nature of the corpus texts, our current conclusions may be limited by the types of texts included in the corpus.

Secondly, recognizing that the phenomenon of object fronting may be influenced by multiple syntactic factors, such as the stronger tendency for pronoun objects to be fronted in negative clauses, we also attempted to select combinations of different grammatical features. Some meaningful syntactic feature combinations has also been distinguished, such as "object is a personal pronoun; verb has a negation modifier." Although the coverage of this combination was only 7.3%, its precision reached 41.3%, indicating that personal pronouns have a strong tendency to be fronted in negative clauses. However, in this case, we encountered many redundant grammatical combinations. For example, the model selected combinations such as "object is a pronoun; object is an interrogative pronoun" however, this rule essentially indicates that the object is an interrogative pronoun. Yet the model treats "object is an interrogative pronoun" and "object is a pronoun; object is an interrogative pronoun" as separate rules for comparison, which to some extent reduces the weight assigned to other potentially meaningful factors. This may be due to redundancy in the corpus annotation, or it may result from a lack of constraints when selecting features space.



Nevertheless, by combining the quantitative results of the linear classifier with detailed corpus analysis, we are also working toward improving the quality of corpus annotation. Through this integration, we have identified inconsistencies in syntactic annotation and found that attempts to extract combined grammatical rules reveal redundancy in the annotation information. Therefore, in future work, we will first use the current findings to optimize corpus annotation—correcting errors and inconsistencies—to improve the reliability of single-feature factor extraction. At the same time, we will attempt to reduce redundant information in order to support more accurate quantitative analysis of composite grammatical factors. We also note that OV constructions in Archaic Chinese are strongly influenced by lexical factors. Therefore, in future research, we plan to add orthographic features into our analysis.

## Acknowledgments

This work was supported by the China Scholarship Council (CSC). This work is also supported by the Université Paris Nanterre. We gratefully acknowledge the contribution of Ms. Yuxin Zhang during the finalization of the camera-ready version. We also thank beloved cat *Xiaoguai*, and guinea pigs *Coco* and *Chocolat*, whose quiet companionship provided much comfort throughout the writing process. Finally, we sincerely appreciate the anonymous reviewers for their valuable comments and suggestions, which helped improve the quality of this work.

## References

- Aditi Chaudhary, Zaid Sheikh, David R Mortensen, Antonios Anastasopoulos, and Graham Neubig. 2022. [AUTOLEX: An Automatic Framework for Linguistic Exploration](#). *arXiv preprint*.
- Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. [Universal Stanford dependencies: A cross-linguistic typology](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. [Universal Dependencies](#). *Computational Linguistics*, 47(2):255–308.
- Redouane Djamouri. 2014. [Dui shànggǔ hànǔ fǒudìng jù lǐ dàicí bīnyǔ wèizhì de jìnyībù tāolùn](#) (對上古漢語否定句裏代詞賓語位置的進一步討論) [further discussion on positions of the object pronoun in negative sentences in archaic Chinese]. *Lishǐ Yǔyánxué Yánjiū* [Research on Historical Linguistics], (02):47–57. [in Chinese].
- Shengli Feng. 2013. [hàn yǔ yùn lǜ jù fǎ xué](#) (漢語韻律句法學) [Prosodic syntax in Chinese (revised edition)]. Shangwu Yinshuguan, Bei jing. [in Chinese].
- Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2021. [Typometrics: From Implicational to Quantitative Universals in Word Order Typology](#). *Glossa: a journal of general linguistics*, 6(1).
- Xuezhong Han. 1996. [Xiānqín fǒudìng jù zhōng “fǒu+dàibīn+dòng” jiégòu de yǔfǎ tèdiǎn](#) (先秦否定句中“否+代賓+動”結構的語法特點) [grammatical properties of “negator + pronominal object + verb” construction in negative sentences in pre Qin period]. *Běijīngdàxué xuébào(zhèxué shèhuì kēxué bǎn)* [Journal of Peking University (Philosophy and Social Sciences)], (06):103–106. [in Chinese].
- Santiago Herrera, Caio Corro, and Sylvain Kahane. 2024. [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.
- Natalia Levshina. 2019. [Token-based typology and word order entropy: A study based on universal dependencies](#). *Linguistic Typology*, 23(3):533–572.
- Jianzhong Ma. 1898. [ma shi wen tong](#) (馬氏文通) [Ma’s Grammar: The First Systematic Grammar of Chinese]. Shangwu Yinshuguan, Bei jing. [in Chinese].
- Igor A. Melcuk. 1988. *Dependency syntax: theory and practice*. SUNY series in linguistics. State University Press of New York, Albany.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. [Universal Dependencies v2: An evergrowing multilingual treebank collection](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.
- Victor Junnan Pan and Yihe Jiao. 2023. [Object-Fronting in Archaic Chinese](#). In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.
- Alain Peyraube. 1997. [On word order in Archaic Chinese](#). *Cahiers de linguistique - Asie orientale*, 26(1):3–20.
- Andreas Vlachos and Mark Craven. 2010. [Detecting speculative language using syntactic dependencies](#)

- and logistic regression. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, pages 18–25, Uppsala, Sweden. Association for Computational Linguistics.
- Li Wang. 1980. *hàn yǔ shǐ gǎo* (漢語史稿) [*Lectures on the history of the Chinese language*], 3 edition. Zhonghua Shuju, Bei jing. [in Chinese].
- Li Wang. 1989. *Hànyǔ yǔfǎ shǐ* (漢語語法史) [*The history of Chinese grammar*]. Di 3 juan. Shangwu Yinshuguan, Bei jing. [in Chinese].
- Xi Xiang. 2010. *Jiǎnmíng hànyǔ shǐ* (簡明漢語史) [*A Concise History of Chinese (Revised Edition)*]. Shangwu Yinshuguan, Bei jing. [in Chinese].
- Koichi Yasuoka. 2019. **Universal Dependencies Treebank of the Four Books in Classical Chinese**. *DADH2019: 10th International Conference of Digital Archives and Digital Humanities*, pages 20–28.
- Koichi Yasuoka, Christian Wittern, Tomohiko Morioka, Takumi Ikeda, Naoki Yamazaki, Yoshihiro Nikaido, Shingo Suzuki, Shigeki Moro, and Kazunori Fujita. 2022. Designing universal dependencies for classical chinese and its application. *Journal of Information Processing*, 63(2):355–363.
- Min Yu. 1981. *Dàojuàn tàn yuán* (倒句探源) [tracing the object-predicate construction]. *Yǔyán yánjiū*, (00):78–82. [in Chinese].

# Periphrastic Verb Forms in Universal Dependencies

Lenka Krippnerová and Daniel Zeman

Charles University, Faculty of Mathematics and Physics

Institute of Formal and Applied Linguistics (ÚFAL)

Prague, Czechia

lenka.krippnerova@volny.cz, zeman@ufal.mff.cuni.cz

## Abstract

We propose a generalization of the morphological annotation in Universal Dependencies (UD) to phrases spanning multiple words, possibly discontinuous. Our focus area is that of periphrastic tenses, voices and other forms, typically consisting of a non-finite content verb combined with one or more auxiliaries; however, the same approach can be applied to other morphosyntactic constructions. We present a software tool that can detect periphrastic verb forms, extract the relevant morphological features from member words and combine them into new, phrase-level annotation. The tool currently detects periphrastic verb forms in 15 Slavic languages that are represented in UD and it is easily adaptable to other constructions and languages. Both the tool and the processed Slavic data are freely available.

## 1 Introduction

Since the basic annotation units in Universal Dependencies (de Marneffe et al., 2021) are morphosyntactic words, UPOS tags and morphological features always relate to a single word. This can be viewed as a limitation. Many languages have multiword expressions<sup>1</sup> that could be described by feature-value pairs from an inventory similar to morphological features, but the features would apply to the whole expression, and not to any of its member words alone (Zeman, 2023). Periphrastic verb forms are a prime example of this. For example, the English present perfect *I have left* can be described as Mood=Ind, Tense=Pres, Aspect=Perf, Voice=Act, Number=Sing, Person=1; however, in UD annotation some of these features are scattered on individual words and others, such as the aspect, are not annotated at all because they do not characterize any of the words in isolation.

<sup>1</sup>By multiword expression we now mean just an expression of multiple words; no idiosyncrasy is required.

Another issue is that words are not always easy to delimit (Evang and Zeman, 2024). For example, the Japanese writing system does not insert spaces between words, and several approaches have been proposed to break the text up to word-like units (Murawaki, 2019). Consider (1) below:

- (1) 行ってきました  
*ittekimashita*  
'went'

This could be treated as one verb in the polite form of the past tense. It contains two lexical roots, so it could be also considered a compound predicate *itte kimashita*, consisting of a converb of *iku* 'go' and a polite past form of *kuru* 'come'. But in fact, Japanese UD<sup>2</sup> decomposes it into two verbs, two auxiliaries, and one subordinator: *it te ki mashi ta*. Naturally, the selected segmentation directly affects which features can be annotated and where to find them.

In some cases, a word participating in a periphrastic form may even bear a feature that conflicts with the feature of the whole expression. For example, in Czech (2), *řekl jsem* 'I told' is a periphrastic past tense composed of past participle and present auxiliary; in *by přišel* 'he would come', the "past" participle is used in a present conditional construction.

- (2) Řekl jsem mu, a= =by přišel  
told I.have him that would he.come  
V.F.=Part M.=Ind M.=Cnd V.F.=Part  
T.=Past T.=Pres T.=Past  
'I told him to come.'

We present a rule-based software tool that takes the existing UD annotation as input and enriches it with features for periphrastic verb forms. The tool currently covers all 15 Slavic languages in

<sup>2</sup>Japanese GSD in UD 2.15.

ID	FORM	UPOS	MISC
1	Řekl	VERB	Phrase=[1, 2] PhraseAspect=Perf PhraseForm=Fin PhraseGender=Masc  PhraseMood=Ind PhraseNumber=Sing PhrasePerson=1 PhraseTense=Past PhraseVoice=Act
2	jsem	AUX	–
3	mu	PRON	SpaceAfter=No
4	,	PUNCT	–
5-6	aby	–	–
5	aby	SCONJ	–
6	by	AUX	–
7	přišel	VERB	Phrase=[6, 7] PhraseAspect=Perf PhraseForm=Fin PhraseGender=Masc  PhraseMood=Cnd PhraseNumber=Sing PhrasePerson=3 PhraseVoice=Act SpaceAfter=No
8	.	PUNCT	–

Table 1: Sample output; for glosses and translation, see (2) in the text. The new annotations are placed in the MISC column at the head node of the verb form. The Phrase attribute identifies the nodes that belong to the periphrastic form, the other Phrase\* attributes correspond to morphological features as defined for the FEATS column.

	UPOS	VerbForm	Mood	Aspect	Tense	Voice	Number	Person	Gender	Animacy	Clitic	Variant	Phrase
Několik	DET												
jsem	AUX	Fin	Ind	Imp	Pres	Act	Sing	1					*
jich	PRON												
našel	VERB	Part		Perf	Past	Act	Sing		Masc				*
jsem našel	VERB	Fin	Ind	Perf	Past	Act	Sing	1	Masc				[2, 4]
Znalazl-	VERB	Fin	Ind	Perf	Past	Act	Sing		Masc	Hum			*
-em	AUX			Imp			Sing	1			Yes	Long	*
ich	PRON												
kilka	DET												
Znalazłem	VERB	Fin	Ind	Perf	Past	Act	Sing	1	Masc	Hum			[1, 2]

Table 2: Propagation of word features to phrase features shown on a Czech and a Polish sentence with the same meaning: [cs] *Několik jsem jich našel.* / [p1] *Znalazłem ich kilka.* ‘I found several of them.’ Blue color indicates the periphrastic form (phrase) and its features. Orange are the contributing features of the member words. Word features shown in black are not copied to the phrase annotation. The two languages differ in word order. In Czech, the periphrastic form is discontinuous, while in Polish the auxiliary is a clitic on the main verb. The feature profiles are very similar except that the Polish participle expresses Animacy and the Polish auxiliary lacks the Tense=Pres annotation.

UD<sup>3</sup> and it is easily extensible to other languages and other grammatical constructions. The tool has been used to prepare Czech data for the shared task on “Morpho-Syntactic Parsing” that is being organized<sup>4</sup> as part of SyntaxFest 2025.

## 2 The Tool

The tool relies on the Udapi<sup>5</sup> Python framework (Popel et al., 2017). Udapi works as a processing pipeline that reads data in the CoNLL-U format, applies selected processing *blocks* to the data and saves the modified data in CoNLL-U again. We created a number of blocks that take care of verb forms found in Slavic languages. When a periphrastic form is found, the features that describe it are encoded as MISC attributes of the word that heads the periphrastic expression in the

<sup>3</sup>Belarusian, Bulgarian, Croatian, Czech, Macedonian, Old Church Slavonic, Old East Slavic, Polish, Pomak, Russian, Serbian, Slovak, Slovenian, Ukrainian, and Upper Sorbian.

<sup>4</sup><https://unidive.lisn.upsaclay.fr/doku.php?id=other-events:mnp>

<sup>5</sup><https://udapi.github.io/>

future indicative		<i>budu dělat</i>	<i>budu se dělat</i>	<i>budu dělán</i>
present indicative		<i>dělám</i>	<i>dělám se</i>	<i>jsem dělán</i>
past indicative		<i>dělal jsem</i>	<i>dělal jsem se</i>	<i>byl jsem dělán</i>
conditional		<i>dělal bych</i>	<i>dělal bych se</i>	<i>byl bych dělán</i>
imperative		<i>dělej</i>	<i>dělej se</i>	<i>buď dělán</i>
	active			
	middle /			
	× reflexive passive			
participle	passive	<i>dělající</i>	<i>dělající se</i>	<i>dělaný</i>
converb		<i>dělaje</i>	<i>dělaje se</i>	<i>jsa dělán</i>
infinitive		<i>dělat</i>	<i>dělat se</i>	<i>být dělán</i>
		‘to do’	‘to be done’	‘to be done’

Table 3: Overview of verb forms with examples from Czech. A few other forms that are not attested in Modern Czech exist in other Slavic languages: supine (Old Church Slavonic, Slovenian); aorist and imperfect past (Old Church Slavonic, Bulgarian, Macedonian, Upper Sorbian); impersonal non-passive form (Polish, Ukrainian).

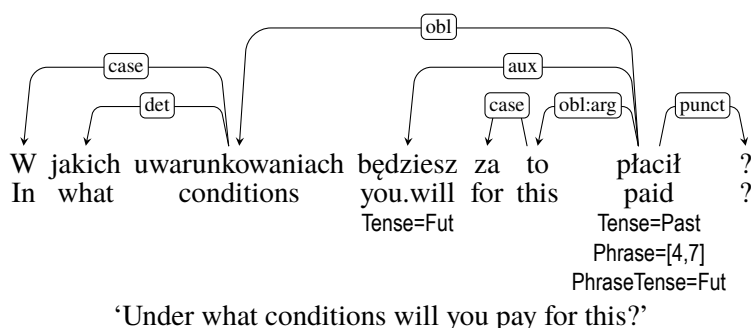


Figure 1: Example of future tense from the Polish PDB treebank. The value of the PhraseTense is copied from the auxiliary.

dependency tree (Tables 1 and 2).<sup>6</sup>

Each of the blocks specializes in finding one of the verb forms expressed in Slavic languages. There are blocks for present tense, future tense, past tense, conditional, imperative, transgressive,<sup>7</sup> and infinitive. The blocks must also handle negation and passivization, as each of these forms can be passivized using either a reflexive marker or a passive participle with an auxiliary. Some verb forms are not periphrastic (they are expressed by one word) but for completeness we capture them, too. This is the case with present tense, transgressive, imperative, and infinitive in the active voice. In some Slavic languages, the past tense and some forms of the future tense are also simple (Zeman, 2016). Periphrastic verb forms include the future and past tenses, the conditional, and all passive forms. Besides standard verbal predicates, the

forms are also marked for non-verbal predicates with copula. Table 3 gives an overview of the forms with examples.

The detection of a periphrastic form normally involves identification of the head word (verb, other word with the VerbForm feature, or a non-verbal predicate) and collection of its children with relations aux, cop, or expl. Each block has its own set of conditions over these nodes to verify whether they represent an instance of the construction the block focuses on. When the conditions are met, it is necessary to determine which features should be propagated to the entire periphrastic phrase. For example, in the periphrastic future tense, the value of the attribute PhraseTense is copied from the Tense feature of the auxiliary (Figure 1). In contrast, in the periphrastic past tense, the value of PhraseTense is copied from the Tense feature of the content verb (Figure 2).

In some cases, a value of a Phrase\* attribute must be added even though it is not present on the head word or any of its dependents. This occurs,

<sup>6</sup>Our blocks are now available directly in the Udapi GitHub (<https://github.com/udapi/udapi-python>). They have their own namespace `msf.slavic`, e.g., `msf.slavic.Future` is the block for future tense.

<sup>7</sup>Also known as gerund or converb.

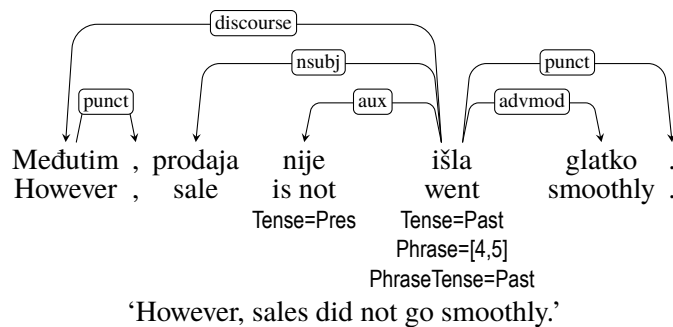


Figure 2: Example of past tense from the Serbian SET treebank. The value of the PhraseTense is copied from the content verb.

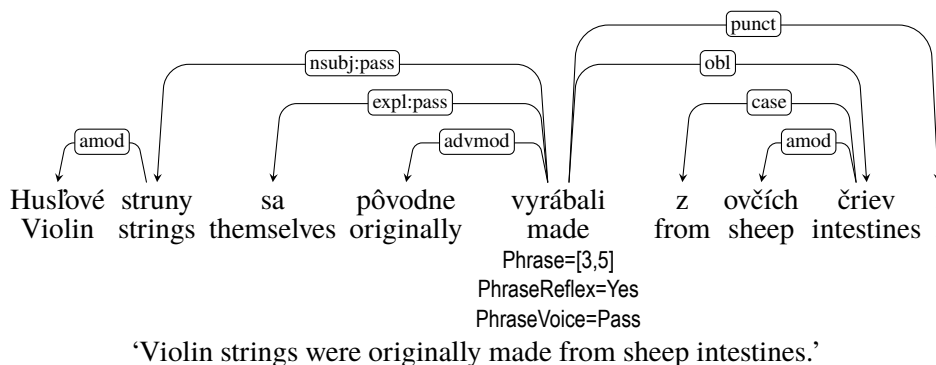


Figure 3: Example of reflexive passive from the Slovak SNK treebank (abridged).

for example, in the case of the reflexive passive (Figure 3). The need to assign `PhraseVoice=Pass` is inferred solely from the presence of a reflexive marker in an `expl:pass` relation. `Voice=Pass` is annotated neither on the content verb nor the reflexive marker; in fact, the content verb is marked with `Voice=Act`.

The blocks also handle negation. In Slavic languages, negation can be expressed in two ways: either with a negative prefix or with a negative particle. In addition to searching for `aux`, `cop`, and `expl` relations among the descendants of the head word, we also look for the presence of a negative particle to determine whether the attribute `PhrasePolarity=Neg` should be generated. If no negative particle is found among the descendants, we then check whether the negation is expressed via a prefix. This can be challenging, as different verb forms may realize the negative prefix in different parts of the verb phrase. For example, in the Czech active past tense, the negative prefix appears on the content verb, whereas in the passive, it can be expressed on the auxiliary, on the content verb, or both.

### 3 Harmonization of Annotations

Even though the annotations in Universal Dependencies are supposed to be consistent, there are still cases across different languages where the annotations are not unified sufficiently. Whenever such discrepancies directly affect the retrieval of periphrastic verb forms, we harmonize them, meaning that even word-level features in our output may differ from the input data. The benefit is twofold: Besides making the identification of verb forms easier, the resulting data is also more suitable for cross-linguistic studies, very much in the UD spirit.

The conditional mood may serve as an example. In Polish, the conditional auxiliary is not tagged with `Mood=Cnd`, but its incoming relation is subtyped as `aux:cnd`. However, in other Slavic languages, the conditional auxiliary is marked with `Mood=Cnd`, therefore we assign this feature to the corresponding auxiliaries in Polish as well (Figure 4).

### 4 Participles

We decided to harmonize the UPOS annotation of participles. Since participles express both verbal



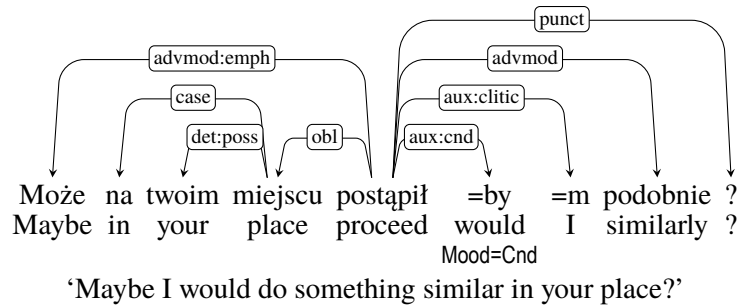


Figure 4: Example of conditional from the Polish PDB treebank. The Mood=Cnd feature was added by the preprocessing block.

past active (l-participle)	VERB	<i>читал</i>	<i>čital</i>	used mainly with auxiliaries
present active	ADJ	<i>читающий</i>	<i>čitajuščij</i>	used mainly as attribute
past active (adjectival)	ADJ	<i>читавший</i>	<i>čitavšij</i>	used mainly as attribute
present passive <sup>8</sup>	ADJ	<i>читаемый</i>	<i>čitaemyj</i>	used mainly as attribute
past passive (long variant)	ADJ	<i>прочитанный</i>	<i>pročitannyj</i>	used as attribute or predicate
past passive (short variant)	ADJ	<i>прочитан</i>	<i>pročitan</i>	used as attribute or predicate

Table 4: Overview of participles with examples from Russian *читать* (*čitat*) ‘to read’. The l-participle is used predicatively, with or without auxiliaries, to form the past tense (or resultative / perfect in old languages), conditional and future tense (in Polish and Slovenian; other languages use the infinitive instead). The past passive participle with an auxiliary forms the passive voice. In addition, Polish and Ukrainian have a special impersonal verb form, which is not considered participle, but it bears some similarities to passive participles.

features (such as aspect) and adjectival features (such as case), they occupy an intermediate position between verbs and adjectives. This leads to inconsistencies in the annotations. In some treebanks, these forms are tagged with UPOS ADJ, while in others they receive UPOS VERB. To resolve this, we apply a simple rule of thumb: participle types that can express case (i.e., all types except so called l-participles<sup>9</sup>) are now annotated as adjectives; it is still easy to recognize them thanks to the VerbForm=Part feature. Despite the UPOS tag, we continue to treat participles as potential members of periphrastic verb forms. The fact that some participles are used attributively rather than predicatively will be visible in syntactic annotation (which we carry over unmodified to the output); in such cases, our Phrase\* features will only reflect the features of the participle itself.

Table 4 exemplifies the various participle types that can be found in Slavic languages (Sussex and Cubberley, 2006).

<sup>8</sup>The present passive participle is found only in Russian, Old Church Slavonic, and Old East Slavic.

<sup>9</sup>Also excluded are converbs, which developed from participles but their forms are frozen w.r.t. Case.

## 5 Reflexive / Middle Voice

One of the Phrase\* attributes placed in the MISC column at the head node of a verb phrase is PhraseReflex. This is a Boolean feature that appears only with the value Yes; when absent, it is interpreted as No. We mark as reflexive only those verb phrases that contain the reflexive marker in an expletive relation. Reflexive pronouns that function as objects or obliques are not considered part of the verb phrase and therefore do not justify reflexive marking.

The expl relation of reflexive pronouns can include subtypes such as expl:pv (pronominal verb), expl:pass (reflexive passive), and expl:impers (impersonal construction). Among these, expl:pass is essential for identifying reflexive passives (Figure 3). However, because this subtype is not distinguished in many treebanks, we are often unable to recognize reflexive passives and must instead annotate such verb phrases with PhraseVoice=Act.

In East Slavic languages (Belarusian, Russian, and Ukrainian, partly also in Old East Slavic), reflexive markers are suffixed on the verb. In such cases, neither the Reflex=Yes feature nor the expl relation is present in the data.<sup>10</sup> Instead, the feature

<sup>10</sup>Old East Slavic contains both suffixed and separate re-

Feature	Precision	Recall	$F_1$ -score
Phrase	1	0.99	0.99
PhraseAspect	1	0.63	0.78
PhraseForm	1	0.99	0.99
PhraseMood	1	0.99	0.99
PhraseNumber	0.94	0.92	0.93
PhrasePerson	1	0.99	0.99
PhraseTense	0.96	0.95	0.95
PhraseVoice	1	0.99	0.99

Table 5: Evaluation of Czech.

Feature	Precision	Recall	$F_1$ -score
Phrase	0.99	0.99	0.99
PhraseAspect	1	1	1
PhraseForm	1	1	1
PhraseMood	0.98	0.96	0.97
PhraseNumber	1	1	1
PhrasePerson	1	1	1
PhraseTense	0.98	1	0.99
PhraseVoice	1	1	1

Table 6: Evaluation of Ukrainian.

Voice=Mid (middle voice) indicates that the verb is reflexive (Figure 5).<sup>11</sup>

When a reflexive verb phrase is identified, we assign `PhraseReflex=Yes`. Additionally, if the head verb form contains the feature `Voice=Mid`, we also assign `PhraseVoice=Mid`.

## 6 Data Release

While we believe that installation and usage of Udapi with our blocks is easy, we are simplifying it even more by releasing the processed Slavic treebanks from UD 2.16 at <http://hdl.handle.net/11234/1-5936>. The blocks are still useful when one wants to process other versions of UD, or even one’s own data processed by an automatic parser.

## 7 Evaluation

We have manually verified the output and calculated the precision, recall, and  $F_1$ -score on 100 Czech and 100 Ukrainian sentences.

### 7.1 Evaluation of Czech

For the evaluation, we used the first 100 sentences of the Czech PUD treebank v2.15 with 275 periphrastic verb forms (Table 5). The recall of the feature `PhraseAspect` is low due to the fact that this feature is often missing in the input data for individual verbs. 90 verb tokens out of 286 lack the `Aspect` feature.<sup>12</sup> Because of this, we have decided to simplify the detection of the future tense. In Czech, perfective verbs have a simple future tense,

flexive markers. When they are separate words, they have `Reflex=Yes` and `expl (:pv)`.

<sup>11</sup>The `Voice=Mid` feature is currently not used in Ukrainian treebanks. To maintain consistency, we add it in our harmonization step (Section 3), based on verb suffixes.

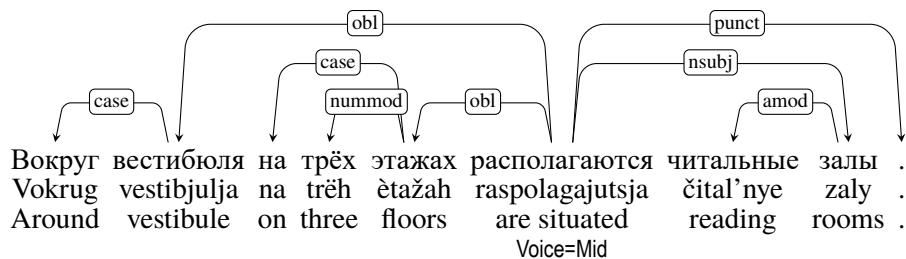
<sup>12</sup>In the rare case of biaspectual verbs, omitting the `Aspect` feature would be legitimate. The verbs in our test sample are not biaspectual.

which looks morphologically like the present and is labeled `Tense=Pres` in the input features; due to the absence of the `Aspect` feature, it is not possible to reliably discriminate present from future in these cases. Therefore, we mark all simple present-like forms as `PhraseTense=Pres`, which decreases precision of `PhraseTense` in Table 5. There are five perfective verbs in the test data that we marked as `PhraseTense=Pres` even though they express the future and the aspect is specified. For 22 verb tokens in the present-like form, the aspect is not specified and we marked all of them as `PhraseTense=Pres`, while three of them should actually be `PhraseTense=Fut`. The precision of the `PhraseNumber` feature is lower because some verb forms have `Number=Plur,Sing` and it is not always easy to decide which number to choose.

### 7.2 Evaluation of Ukrainian

Ukrainian was evaluated on the first 100 sentences of the IU test treebank v2.15 with 250 periphrastic verb forms (Table 6). Unlike the Czech test data, there are no issues with missing `Aspect`. The present-like form of perfective verbs is already tagged `Tense=Fut` in the input data; copying it to `PhraseTense` is all we need to do.

Although there are no errors in future tense, the precision of `Phrase`, `PhraseMood` and `PhraseTense` is less than 1. This is because the conditional mood is not detected correctly. Ukrainian conditional is formed using the past participle of the content verb and a special form of the auxiliary verb  $\delta$  (*b*). However, this auxiliary can be encliticized to a subordinator, forming *уоѠ* (*ščob*) ‘so that’, but these cases are not labeled conditional in the treebank. As a result, the periphrastic verb form is not fully detected and only the content verb in the past tense is recognized. Consequently, the feature `PhraseTense=Past` is generated, but the conditional mood does not express tense, so the feature `PhraseTense` should not be



‘There are reading rooms on three floors around the vestibule.’

Figure 5: Example of middle voice from the Russian GSD treebank.

generated at all. Figure 6 gives an example of such conditional clause with the correct annotation that our tool failed to deliver.

### 7.3 Old Church Slavonic

We do not have the same kind of manual evaluation for Old Church Slavonic as we do for Czech and Ukrainian. Nevertheless, this language’s data is an outlier in many respects and we believe that some observations are worth sharing here. Some of them are related to OCS being different than the other languages; quite a few, however, reflect divergent approaches to annotation of phenomena that are not so different in nature.

Infinitives do not express tense, nevertheless, in Old Church Slavonic they are annotated with `Tense=Pres`. We remove this feature in our pre-harmonization step.

The future tense seems to be the youngest grammaticalized tense (Vepřek, 2015) and in Old Church Slavonic it is often expressed using several pseudo-auxiliary verbs that may still keep a shade of their original lexical meaning. The UD treebank does not distinguish the original present tense meaning of the auxiliaries from periphrastic future. We cannot reliably make this distinction on the fly, so we annotate all such forms as `PhraseTense=Pres`, although it is probably not always correct.

The `Aspect` in modern Slavic languages is lexical: If an imperfective verb has a perfective counterpart, they will have different lemmas and will be considered different lexemes. This is how the `Aspect` feature is handled in languages where its annotation is present.<sup>13</sup> However, in OCS the lexical aspect is not annotated and the feature is used to distinguish the two simple past tenses: imperfect (`Aspect=Imp`) and aorist (`Aspect=Perf`). This generates inconsistency because in the other languages

<sup>13</sup>Aspect annotation is not present in Upper Sorbian, Croatian and Serbian.

where these tenses have been preserved (most notably Bulgarian), the tenses are distinguished by the `Tense` feature (`Tense=Imp` for imperfect, `Tense=Past` for aorist).

Moving from tense to mood, we observe a terminological mismatch: Some authors (Huntley, 2002, p. 156) use the term ‘subjunctive’ for the form that is usually called conditional (`Mood=Cnd`) in Slavic languages including Old Church Slavonic (Vepřek, 2015, 5.17.1). Unfortunately, the authors of the OCS treebank preferred the former term and used `Mood=Sub` instead of `Mood=Cnd`. We eliminate this inconsistency in the preprocessing step.

Passive participles have present and past forms, unlike all the other Slavic languages except Russian. The periphrastic passive (the auxiliary *byti* ‘be’ + passive participle) is difficult to distinguish from a similar deverbative adjective used as a non-verbal predicate with a copula; in the data, most such cases are annotated as `cop` rather than `aux:pass`. There was also the reflexive (medio)passive but again it is not recognizable in the data. The reflexive clitic is always attached as `expl:pv`, although some occurrences should probably receive `expl:pass`.

The periphrastic passive, combined with conditional, is illustrated in Figure 7.

## 8 Extensibility to Other Languages

While the present version readily handles Slavic verb forms, the same approach can be used in other languages and for other phrase-level features. To facilitate such extensions, we have designed a generic Udapi block that reads a configuration file in YAML format. The YAML file defines rules for periphrastic forms in a particular language: how to identify nodes that belong to the form, and how to derive phrasal features from the features of the nodes. The rules have been designed to be simple enough that even a user without programming ex-

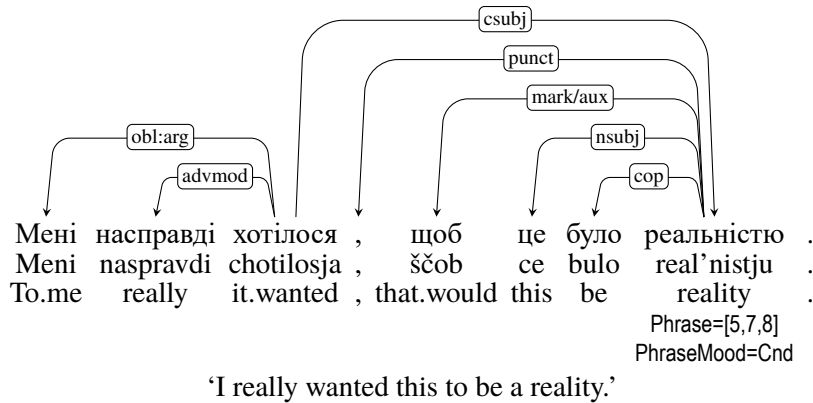


Figure 6: Example of unrecognized conditional from the Ukrainian IU treebank.

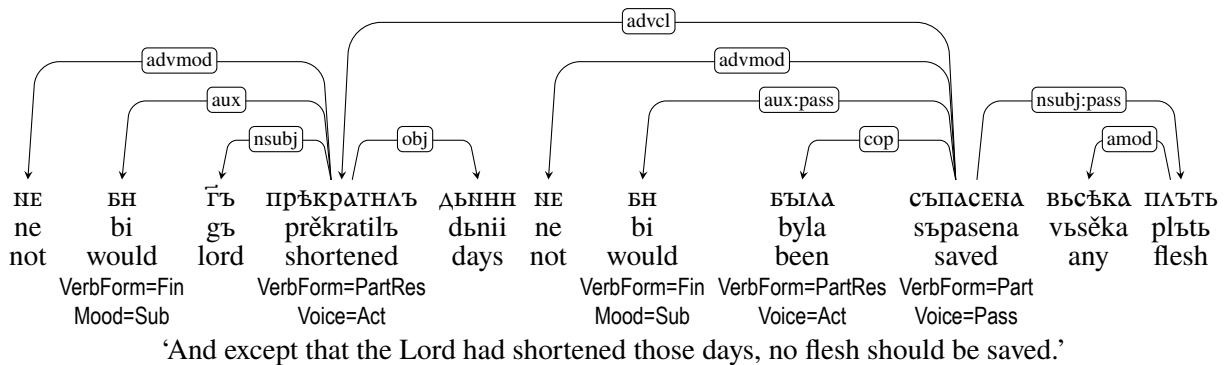


Figure 7: Example of the source annotation in Old Church Slavonic PROIEL (sentence shortened). The l-participles have a language-specific feature `VerbForm=PartRes`. Conditional auxiliaries are tagged `Mood=Sub` instead of `Mood=Cnd`. The second l-participle is incorrectly attached to the passive participle as copula, while it should be the passive auxiliary. The conditional *bi* should be attached as auxiliary but not as passive auxiliary.

perience can create them.

We are currently working on a similar pipeline for Portuguese, Spanish, and Italian. We have created new rules for identifying periphrastic verb forms, based on the grammatical structures of these languages. A different approach to aspect is required: unlike Slavic languages, these languages do not express aspect lexically. Consequently, we introduced new values for the `PhraseAspect` attribute – `ImpProg` and `PerfProg` – to annotate completed and ongoing progressive actions, respectively. The example is illustrated in Figures 8 and 9.

Extending the approach to other languages is easy from the implementation perspective, as the logic of the existing Udapi blocks can be reused. However, when adapting the tool to a new language, it is necessary to develop specific rules for phrase identification. This process is relatively straightforward when rules already exist for a closely related language (for instance, once we developed rules for Portuguese, adapting them for

Spanish was not difficult). In cases where no such rules are available, a careful analysis of the target language’s grammar is necessary to formulate appropriate rules.

## 9 Conclusion

We have presented a software tool that reads UD treebanks and adds phrase-level features for periphrastic grammatical forms. The tool is freely available within the Udapi framework at <https://github.com/udapi/udapi-python>, and its output on UD v2.16 is available at <http://hdl.handle.net/11234/1-5936>.

The tool is ready to analyze verb forms in Slavic languages but it is easily extensible, both to other languages and to constructions other than verb forms. For example, it could be used to unify morphological and periphrastic comparatives (cf. English *smarter* vs. *more intelligent*). The tool can be used for cross-linguistic studies (e.g. the full verbal paradigms in two languages) but also in NLP appli-

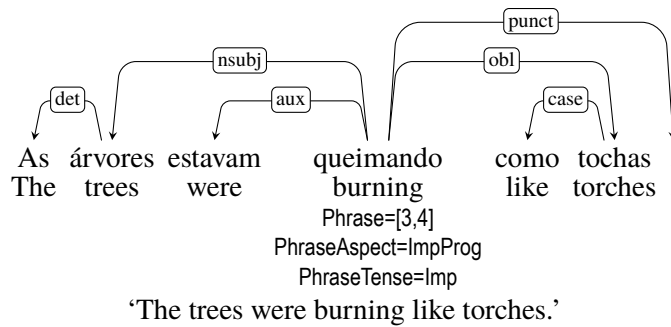


Figure 8: Example of a Portuguese verb phrase with `PhraseAspect=ImpProg` from the Porttinari treebank.

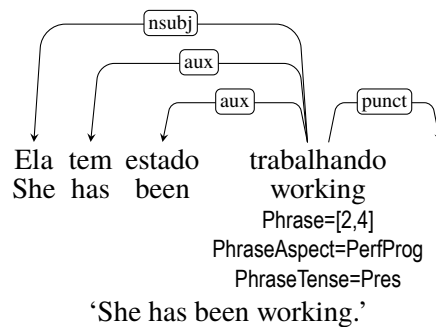


Figure 9: Example of a Portuguese verb phrase with `PhraseAspect=PerfProg`.

cations to overcome the difficulties of defining the word. The tool has been used to prepare Czech data for the UniDive Shared Task on Morphosyntactic Parsing, collocated with SyntaxFest 2025.

## 10 Limitations

For the most part, our tool just takes information from the input data and presents it in a restructured way. Whatever interpretation the tool adds is based on the knowledge of the grammatical rules of the given language as a whole, not on detailed understanding of individual words. Therefore, if some piece of the input annotation is missing or incorrect, it cannot be added or corrected in the output.

## Acknowledgements

The authors are grateful to Adriana Pagano for useful insights when extending the work to Romance languages.

The work described herein has been supported by the grants *Language Understanding: from Syntax to Discourse* of the Czech Science Foundation (Project No. 20-16819X) and *LINDAT/CLARIAH-CZ* (Project No. LM2023062) of the Ministry of Education, Youth, and Sports of the Czech Republic. It has also received support from the CA21167

COST action UniDive, funded by COST (European Cooperation in Science and Technology).

## References

- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. [Universal Dependencies](#). *Computational Linguistics*, 47(2):255–308.
- Kilian Evang and Daniel Zeman. 2024. [Word segmentation in universal dependencies](#). In *UniDive General Meeting in Naples posters*, Napoli, Italy.
- David Huntley. 2002. Old Church Slavonic. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, pages 125–187. Routledge, Oxon, UK.
- Yugo Murawaki. 2019. [On the definition of Japanese word](#). *Preprint*, arXiv:1906.09719.
- Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. [Udapi: Universal API for Universal Dependencies](#). In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden. Association for Computational Linguistics.
- Roland Sussex and Paul Cubberley. 2006. *The Slavic Languages*. Cambridge Language Surveys. Cambridge University Press.

Miroslav Vepřek. 2015. *Komparativní tvarosloví staroslověňštiny a staré češtiny (Comparative Morphology of Old Church Slavonic and Old Czech)*. Univerzita Palackého v Olomouci, Olomouc, Czechia.

Daniel Zeman. 2016. Universal annotation of Slavic verb forms. *The Prague Bulletin of Mathematical Linguistics*, (105):143–193.

Daniel Zeman. 2023. [Subword relations, superword features](#). In *UniDive General Meeting at Paris-Saclay posters*, Orsay, France.



# Word Order Variation in Spoken and Written Corpora: A Cross-Linguistic Study of SVO and Alternative Orders

**Nives Hüll**  
University of Ljubljana  
nives.hull@gmail.com

**Kaja Dobrovoljc**  
University of Ljubljana  
Jozef Stefan Institute, Ljubljana, Slovenia  
kaja.dobrovoljc@ff.uni-lj.si

## Abstract

This study investigates word order variation in spoken and written corpora across five Indo-European languages: English, French, Norwegian (Nynorsk), Slovenian, and Spanish. Using Universal Dependencies treebanks, we analyze the distribution of six canonical word orders (SVO, SOV, VSO, VOS, OSV, OVS). Our results reveal that spoken language consistently exhibits greater word order flexibility than written language. This increased flexibility manifests as a decrease in the dominant SVO pattern and a rise in alternative orders, though the extent of this variation differs across languages. Morphologically rich languages such as Slovenian and Spanish show the most pronounced shifts, while English remains syntactically rigid across modalities. These findings support the claim that modality significantly affects syntactic realizations and highlight the need for typological studies to account for spoken data.

## 1 Introduction

Word order is a fundamental parameter in linguistic typology and syntactic theory. It plays a central role in language classification and shapes our understanding of cross-linguistic variation and syntactic universals. Typological databases such as WALS (Dryer and Haspelmath, 2013) and Grambank (Skirgård et al., 2023) document dominant word order patterns like subject–verb–object (SVO), but these generalizations are typically based on written, formal sources and often fail to capture variation across genres or modalities.

Recent corpus-based work (e.g., Naranjo and Becker, 2018; Levshina, 2019; Baylor et al., 2024) has challenged this categorical view. These studies advocate for a gradient, usage-based approach to word order typology, emphasizing observed token frequencies in syntactically annotated corpora. This shift has enabled a more nuanced classification of languages and has revealed that word order

is not solely a matter of structural constraints, but also reflects contextual factors such as genre, domain, and modality (Levshina et al., 2023; Baylor et al., 2023).

Despite this growing awareness of contextual variation, there remains a lack of systematic, cross-linguistic studies that focus specifically on modality, understood here as the distinction between spoken and written language. While modality is often acknowledged, most existing work incorporates it only indirectly or treats it as a secondary factor within broader genre-based analyses. As a result, cross-linguistic studies that systematically examine the influence of modality on constituent order remain scarce. The extent to which spoken and written language diverge in word order—especially across typologically diverse languages—has yet to be addressed in a comparative framework.

This study offers an exploratory contribution to this gap by examining cross-modal variation in constituent order. We analyze a sample of five Indo-European languages—English, French, Norwegian (Nynorsk), Slovenian, and Spanish—using both spoken and written corpora within the Universal Dependencies framework. Focusing on clause-level syntax, we investigate the distribution of six canonical word order permutations—SVO, SOV, VSO, VOS, OSV, and OVS—as a typological baseline for comparing modalities. While the tendency for spoken language to show more variation is often assumed, we argue that systematically capturing how this plays out across languages adds empirical weight to such claims and exposes patterns missed in writing-based investigations.

Our goal is to assess whether and how word order differs between speech and writing, and whether these differences follow consistent cross-linguistic patterns. The analysis is guided by two central research questions: **(1)** Does word order differ between spoken and written language? and **(2)** If so, how does it differ?

The remainder of the paper outlines our data and methods (Section 2), presents the results (Section 3), and discusses cross-linguistic trends (Section 4), followed by concluding remarks (Section 5).

## 2 Data and Methods

### 2.1 Corpus Selection and Preparation

The analysis includes five Indo-European languages for which both spoken and written data are available in Universal Dependencies v2.15 (Zeman et al., 2024). We focused on languages where spoken and written data are clearly separated, either across different treebanks or within a single treebank. To keep the work manageable and grounded in languages we are familiar with, we limited the study to five treebank pairs listed below.

The *Rhapsodie* (spoken) and *GSD* (written) corpora were used for French (Gerdes et al., 2012; Guillaume et al., 2019); *NynorskLIA* (spoken) and *Nynorsk* (written) for Norwegian (Nynorsk) (Øvrelied et al., 2018; Solberg et al., 2014); *SST* (spoken) and *SSJ* (written) for Slovenian (Dobrovolic and Nivre, 2016; Dobrovolic et al., 2017); *COSER* (spoken) and *GSD* (written) for Spanish (Fernández-Ordóñez, 2005–present; Ballesteros et al., 2024); and the *GUM* treebank for English (Zeldes, 2017), which was manually divided into spoken and written subsets based on genre metadata.<sup>1</sup>

### 2.2 Data Extraction

We conducted a quantitative analysis using the STARK tool, designed for querying syntactic patterns in UD-formatted dependency trees (Krsnik and Dobrovolic, 2025).<sup>2</sup> For each language and modality pair, we extracted all instances in which a finite verb governs both a nominal subject (nsubj) and a direct object (obj), regardless of clause type—this includes main and subordinate, declarative and interrogative clauses alike. Each sentence was then classified based on the linear order of the subject, verb, and object into one of six canonical word orders: SVO, SOV, VSO, VOS, OSV, or OVS. This procedure yielded a dataset containing word order distributions for each corpus, which we compared across modalities. To illustrate the

<sup>1</sup>The spoken subset includes interviews, conversations, podcasts, vlogs, courtroom transcripts, and speeches; the written subset includes news articles, academic texts, fiction, how-to guides, biographies, essays, letters, textbooks, and travel guides.

<sup>2</sup><https://github.com/clarinsi/STARK>

six possible orders, Table 1 provides examples in Slovenian—a language that permits all six permutations—along with their English translations.

With this approach, our analysis aligns with inclusive, usage-based studies (e.g., Gerdes et al., 2019; Östling, 2015; Naranjo and Becker, 2018; Baylor et al., 2024), which aim to capture naturally occurring syntactic variation across clause types. Rather than limiting the analysis to main declarative transitive clauses only (e.g., Levshina, 2019; Dryer, 2013), we include all instances of subject–verb–object structures, regardless of clause type. This enables us to more fully capture modality-sensitive variation, while keeping the analysis straightforward and the results easily interpretable.

Order	Slovenian	Gloss
SVO	<i>Mama kupuje jabolka</i>	mother buys apples
SOV	<i>Mama jabolka kupuje</i>	mother apples buys
VSO	<i>Kupuje mama jabolka</i>	buys mother apples
VOS	<i>Kupuje jabolka mama</i>	buys apples mother
OSV	<i>Jabolka mama kupuje</i>	apples mother buys
OVS	<i>Jabolka kupuje mama</i>	apples buys mother

Table 1: Canonical word order examples in Slovenian. All sentences translate as ‘Mother buys apples’.

## 3 Results

### 3.1 General Observations

Figure 1 summarizes the distribution of six canonical word orders across written and spoken corpora for each language. It shows the relative frequency of SVO, SOV, VSO, VOS, OSV, and OVS, allowing for a direct comparison between modalities.

The results confirm that word order in spoken language differs from written language across all examined languages. In every case, speech exhibits greater variation than writing, with the dominant SVO pattern decreasing in spoken data. Additionally, the degree of flexibility in word order varies across languages, with some showing more pronounced shifts than others. These findings are consistent across the sample.

### 3.2 Language-Specific Findings

**English** shows the least variation. SVO remains dominant, dropping only slightly from 97.4% in written to 93.0% in spoken data. OSV rises modestly from 2.6% to 6.9%, while other patterns remain marginal. This limited shift may suggest that English maintains a relatively high degree of syntactic rigidity even in spontaneous speech. The

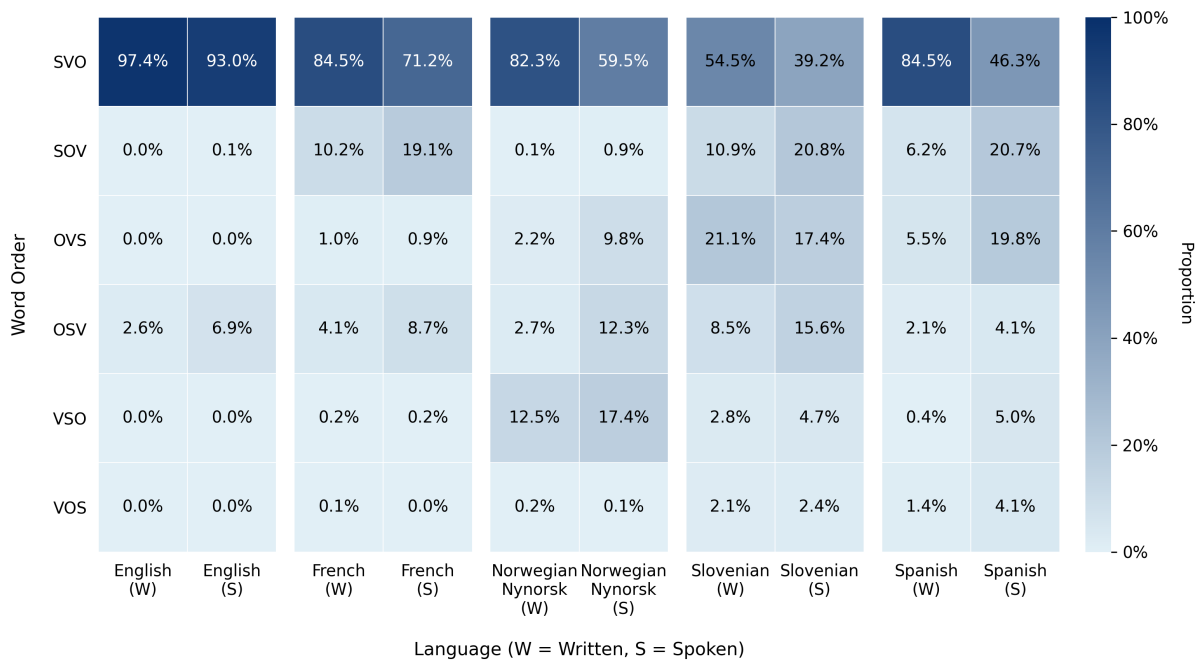


Figure 1: Word order frequencies in written and spoken language across five Indo-European languages.

slight rise in OSV might be due to marked topicalizations or interrogative constructions. For example, *That we did this summer* and *That I give you* illustrate how speakers may foreground the object for emphasis, while questions such as *What did he say?* could also contribute to this pattern.

**French** exhibits a more noticeable change: SVO usage drops from 84.5% to 71.2%, accompanied by increases in SOV (10.2% → 19.1%) and OSV (4.1% → 8.7%). These surface patterns may be influenced by the frequent use of object clitic pronouns and dislocation structures in spontaneous speech. For example, *on l'a mise à l'épreuve* (*we/one her put to the test*) and *je le mets également à l'intérieur* (*I it put also inside*) contain preverbal object clitics, which may lead to an apparent rise in non-SVO orders without indicating a change in underlying syntax.

**Norwegian Nynorsk** displays moderate flexibility. SVO decreases from 82.3% to 59.5%, while OSV and OVS rise significantly (from 2.7% to 12.3% and from 2.2% to 9.8%, respectively). The increase in OSV patterns may partly result from interrogative structures where question elements and pronouns are fronted, as in *Kva den kallast den fóra?* (*What that is called, that feed?*). OVS constructions, such as *Det veit eg ikkje* (*That I don't know*) are common in spoken discourse and may reflect object-fronting for emphasis or information structure.

**Slovenian** is the most flexible language in the sample. SVO accounts for only 39.2% of spoken clauses, with SOV (20.8%), OSV (15.6%), and OVS (17.4%) forming near-equal shares. This distribution may reflect the language's high degree of pragmatic word order variation. SOV patterns, such as *To mi je šlo zelo na živce* (*That really annoyed me*) and *Jaz ti zdaj pomagam* (*I now help you*) may result from object fronting, emphasis, or prosodic rhythm in spontaneous speech. OSV examples like *To jaz nisem* (*This I am not*) are frequently used for contrastive focus, especially in expressive or corrective contexts.

**Spanish** undergoes the strongest shift from a canonical pattern. SVO falls from 84.5% to 46.3%, while SOV (20.7%), OVS (19.8%), and VSO (5.0%) become more frequent. The increased presence of OV patterns may be partly attributed to clitic constructions and discourse-driven reordering. For instance, *las yuntas lo trillaban* (*the oxen it threshed*) shows preverbal clitic placement that results in an apparent SOV order, while *cómo lo hacía su padre* (*how it did his father*) illustrates an OVS structure that may arise in embedded or emphatic contexts.

## 4 Discussion

Our findings confirm that word order varies significantly between spoken and written modalities across the examined languages. Although all are

classified as SVO-dominant in WALS, spoken data consistently exhibit greater flexibility, with higher frequencies of SOV, OSV, and OVS orders. We observe a cross-linguistic rise in postverbal subjects and object-initial configurations—patterns that are rarely captured in typological descriptions based on written sources. The extent of this variation differs by language: it is most pronounced in morphologically rich systems such as Slovenian and Spanish, and more limited in structurally rigid languages like English.

Several factors may account for the greater flexibility observed in speech. As expected, morphological richness plays a central role: languages with robust case marking, such as Slovenian and Spanish, can overtly signal grammatical roles, reducing the need for fixed word order and allowing more pragmatic or prosodically driven constituent placement.

Second, prosodic structure in speech—intonation, rhythm, and stress—can help disambiguate syntactic relations and guide listener interpretation, even in non-canonical orders (Levshina et al., 2023). In morphologically rich languages, prosody may interact with syntax and discourse to license constituent placement (Gerken, 1996).

Third, discourse-related considerations shape spoken word order. The distinction between given and new information often drives constructions like Left-Dislocation, which promote new or contrastive elements to the left periphery of the sentence (Prince, 1981, 1997; Gregory and Michaelis, 2001). This reflects how spoken syntax is sensitive to real-time communicative needs rather than fixed structural defaults.

Finally, cognitive and psycholinguistic constraints influence linearization. Speakers often place accessible or low-load elements earlier in the sentence to ease comprehension and gain time to plan semantically complex constituents (Schouwstra et al., 2022; Levshina et al., 2023). Features typical of spontaneous speech—such as repairs, hesitations, questions, and topic shifts—also encourage deviations from canonical order. These effects are particularly evident in Slovene dialectal discourse (Kumar, 2019), and more generally in languages where grammatical structure permits flexible sequencing (Levshina, 2019).

Taken together, our findings support previous calls for more gradual, context-aware investigations of constituent order that move beyond dominant

patterns and account for variation across modalities (e.g., Baylor et al., 2024; Levshina et al., 2023). In particular, they highlight speech as a crucial communicative context—shaped by a complex interplay of morphosyntactic, prosodic, cognitive, and discourse factors.

Future work should extend this approach to additional languages, including those outside the Indo-European family. With the growing availability of spoken UD treebanks (Dobrovoljc, 2022; Kahane et al., 2021), there is now concrete potential to uncover cross-linguistic patterns that have long remained underdocumented—not only in typological accounts, but in linguistic research more broadly.

## 5 Conclusion

This study highlights the significant impact of modality on SVO word order variation across five Indo-European languages. Spoken language consistently shows greater syntactic flexibility, especially in morphologically rich systems like Slovenian and Spanish. These findings challenge typological generalizations based primarily on written data and underscore the need for future studies to incorporate spoken corpora for a more accurate picture of constituent order variation.

## Limitations

This study focuses on five Indo-European languages, limiting typological diversity. Only the Nynorsk variety of Norwegian was included to ensure consistent comparison between spoken and written data.

The corpora vary in size, balance, and genre coverage, particularly between spoken and written modalities, which may influence pattern distribution. Only clauses with overt nominal subjects and objects were included, following WALS criteria, excluding constructions common in morphologically rich languages where arguments are omitted.

We also restricted the analysis to verbal predicates, excluding nominal and adjectival constructions, and did not distinguish between declarative and interrogative clauses. Our findings are based on quantitative distributions, with no in-depth qualitative analysis.

Lastly, while Universal Dependencies aims for consistency, differences in annotation guidelines or treebank practices may affect comparability.



## Acknowledgment

This work was financially supported by the Slovenian Research and Innovation Agency through the research project Treebank-Driven Approach to the Study of Spoken Slovenian (Z6-4617), and the research program Language Resources and Technologies for Slovene (P6-0411). The authors also made selective use of generative AI tools during the writing phase, but all proposals were independently reviewed and finalized by the authors.

## References

- Miguel Ballesteros, Héctor Martínez Alonso, Ryan McDonald, Elena Pascual, Natalia Silveira, Daniel Zeman, and Joakim Nivre. 2024. Universal dependencies 2.15: Spanish gsd. [https://github.com/UniversalDependencies/UD\\_Spanish-GSD](https://github.com/UniversalDependencies/UD_Spanish-GSD). Accessed: 2025-04-23.
- Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2023. [The past, present, and future of typological databases in NLP](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1163–1169, Singapore. Association for Computational Linguistics.
- 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, St. Julian’s, Malta. Association for Computational Linguistics.
- Kaja Dobrovoljc. 2022. [Spoken language treebanks in universal dependencies: An overview](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1798–1806, Marseille, France. European Language Resources Association.
- Kaja Dobrovoljc, Tomaž Erjavec, and Simon Krek. 2017. The universal dependencies treebank for slovenian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*, pages 33–38, Valencia.
- Kaja Dobrovoljc and Joakim Nivre. 2016. [The Universal Dependencies treebank of spoken Slovenian](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1566–1573. European Language Resources Association (ELRA).
- Matthew S. Dryer. 2013. [Order of subject, object and verb \(v2020.4\)](#). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.
- Matthew S. Dryer and Martin Haspelmath, editors. 2013. [WALS Online \(v2020.4\)](#). Zenodo.
- I. Fernández-Ordóñez. 2005–present. Corpus oral y sonoro del español rural. [urlhttp://www.corpusrural.es/](http://www.corpusrural.es/). Retrieved April 15, 2022.
- Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2019. [Rediscovering greenberg’s word order universals in UD](#). In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 124–131, Paris, France. Association for Computational Linguistics.
- Kim Gerdes, Sylvain Kahane, Anne Lacheret, Arthur Truong, and Paola Pietrandrea. 2012. [Intonosyntactic data structures: The Rhapsodie treebank of spoken french](#). In *Proceedings of the Sixth Linguistic Annotation Workshop (LAW VI)*, Jeju, South Korea. Held in conjunction with ACL-2012.
- LouAnn Gerken. 1996. [Prosody’s role in language acquisition and adult parsing](#). *Journal of Psycholinguistic Research*, 25(3):345–356.
- Michelle L. Gregory and Laura A. Michaelis. 2001. Topicalization and left-dislocation: A functional opposition revisited. *Journal of Pragmatics*, 33(11):1665–1706.
- Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. 2019. [Conversion et améliorations de corpus du français annotés en universal dependencies](#). *Traitement Automatique des Langues*, 60(2):71–95.
- Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021. [Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal](#). In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 35–47, Sofia, Bulgaria. Association for Computational Linguistics.
- Luka Krsnik and Kaja Dobrovoljc. 2025. STARK: A Toolkit for Dependency (Sub)Tree Extraction and Analysis. In *Proceedings of the SyntaxFest 2025*. To appear.
- Danila Zuljan Kumar. 2019. [Word order in slovene dialectal discourse](#). *Slovenski jezik*, 12:53–74. URN:NBN:SI:DOC-VC42NQ7.
- Natalia Levshina. 2019. [Token-based typology and word order entropy: A study based on universal dependencies](#). *Linguistic Typology*, 23(3):533–572.
- Natalia Levshina, Savithry Namboodiripad, Marc Allasonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoyanova. 2023. [Why we need a gradient approach to word order](#). *Linguistics*, 61(4):825–883.

- Matías Guzmán Naranjo and Laura Becker. 2018. [Quantitative word order typology with UD](#). In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, pages 91–104. Linköping University Electronic Press.
- 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.
- Lilja Øvrelid, Andre Kåsen, Kristin Hagen, Anders Nøklestad, Per Erik Solberg, and Janne Bondi Johannessen. 2018. [The LIA treebank of spoken Norwegian dialects](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Ellen F. Prince. 1981. Toward a taxonomy of given-new information. *Radical Pragmatics*, pages 223–256.
- Ellen F. Prince. 1997. On the functions of left-dislocation in english discourse. In Akio Kamio, editor, *Discourse and Functional Linguistics*, pages 117–144. John Benjamins.
- Marieke Schouwstra, Danielle Naegeli, and Simon Kirby. 2022. [Investigating word order emergence: Constraints from cognition and communication](#). *Frontiers in Psychology*, 13.
- Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bower, Patience Epps, Jane Hill, and 85 others. 2023. [Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss](#). *Science Advances*, 9(16).
- Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen, and Janne Bondi Johannessen. 2014. [The norwegian dependency treebank](#). In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. European Language Resources Association (ELRA).
- Amir Zeldes. 2017. [The GUM corpus: Creating multilayer resources in the classroom](#). *Language Resources and Evaluation*, 51(3):581–612.
- Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noémi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, and ... 2024. [Universal dependencies 2.15](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.



# Author Index

- Alves, Diego, 13  
Amlesom Kifle, Nazareth, 120
- Ballarè, Silvia, 54  
Birtić, Matea Andrea, 103  
Bosco, Cristina, 54
- Chen, Xinying, 84  
Chuprinko, Kirill, 108
- De Langhe, Loic, 24  
Degraeuwe, Jasper, 24  
Diaz Hernandez, Roberto A., 1  
Dobrovoljc, Kaja, 150
- Farasyn, Melissa, 24
- Gasser, Michael, 120  
Guillaume, Bruno, 93
- Herrera, Santiago, 130  
Hoste, Veronique, 24  
Hüll, Nives, 150
- Kahane, Sylvain, 93, 130  
Krielke, Marie-Pauline, 13  
Krippnerová, Lenka, 140  
Kubát, Miroslav, 84
- Magistry, Pierre, 130  
Mauri, Caterina, 54
- Novozhilov, Artem, 108
- Osborne, Timothy John, 74
- Pannitto, Ludovica, 54
- Roulon-Doko, Paulette, 93  
Runjaić, Siniša, 103
- Sanguinetti, Manuela, 54  
Song, Chenchen, 74  
Stepanov, Arthur, 108  
Sviben, Robert, 103
- Talamo, Luigi, 13  
Táboas García, Alba, 36
- Wanner, Leo, 36  
Wu, Qishen, 130
- Zeman, Daniel, 140  
Zucchini, Eleonora, 54