UDW 2025

Eighth Workshop on Universal Dependencies
(UDW, SyntaxFest 2025)

Proceedings

August 27, 2025

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

VITASIS

USTANOVA PATRA STANISLAVA ŠKRABCA

UniDive

Ljubljana
Ljubljana Tourism

Mestna občina Ljubljana

Flanders
State of the Art

CLARIN.SI

cjvt
Centre for Language Resources and Technologies

AMEBIS

alpineon

AI4DH
CENTRE OF EXCELLENCE IN AI FOR DIGITAL HUMANITIES

Organized by

UNIVERSITY OF LJUBLJANA
FRI
Faculty of Computer and Information Science

UNIVERSITY OF LJUBLJANA
FF
Faculty of Arts

SDJT
Slovensko društvo za jezikovne tehnologije

As part of SyntaxFest 2025

SYNTAXFEST

©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`

ISBN 979-8-89176-292-3

Introduction

Universal Dependencies (UD) is a framework for cross-linguistically consistent treebank annotation that has so far been applied to over 150 languages. With the goal of enabling both comparative linguistic studies and multilingual natural language processing, the framework aims to capture similarities as well as idiosyncrasies among different languages. Since the first UD workshop in 2017 in Gothenburg, the UD workshops brings together researchers working on UD to reflect on the theory and practice of UD, its use in research and development, and its future goals and challenges. This year's workshop, 8th Workshop on Universal Dependencies (UDW 2025) is held as a part of 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 Cataluna
Marcos Garcia, Universidade de Santiago de Compostela
Kim Gerdes, Université Paris-Saclay
Loïc Grobol, Université Paris Nanterre
Bruno Guillaume, INRIA
Carlos Gómez-Rodríguez, Universidade da Coruña
Eva Hajicova, Charles University
Dag Trygve Truslew Haug, University of Oslo
Santiago Herrera, University of Paris Nanterre
Richard Hudson, University College London
Maarten Janssen, Charles University Prague
Jingyang Jiang, Zhejiang University
Mayank Jobanputra, Universität des Saarlandes
Sylvain Kahane, Université Paris Nanterre
Václava Kettnerová, Charles University Prague
Sandra Kübler, Indiana University
Guy Lapalme, University of Montreal
François Lareau, Université de Montréal
Miryam de Lhoneux, KU Leuven
Zoey Liu, University of Florida

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

Heike Zinsmeister, Universität Hamburg

Keynote
Typologically informed NLP evaluation

Miryam de Lhoneux
KU Leuven, Belgium

Abstract: NLP has a long history of focusing mainly on English. While increasing efforts are being made towards making language technology more multilingual, English remains the language on which NLP technology is developed first, and applied to other languages next, which inevitably leads to degraded performance compared to English. This talk is about reversing this trend and putting multilinguality at the core of NLP, rather than at the periphery. I describe how typology can inform NLP evaluation, using our recently proposed language sampling framework. A strong limitation of the approach is the state of multilingual datasets, which tend to lack coverage, be machine-translated or have questionable quality. UD is an exception, and I emphasize the role it can play in establishing best practices in multilingual NLP evaluation.

Bio: Miryam de Lhoneux is an assistant professor in the department of Computer Science at KU Leuven in Belgium, researching and teaching Natural Language Processing. She heads the LAGoM NLP lab where the focus is on multilingual and interpretable models. Previously, she was a postdoc at the University of Copenhagen in Denmark. She has a PhD from Uppsala University, Sweden, an MSc in Artificial Intelligence from the University of Edinburgh, UK, and a BA and MA in languages and literatures from UCLouvain, Belgium.

Non-Archival Abstract

MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs

Jaap Jumelet[1], Leonie Weissweiler[2] and Arianna Bisazza[1]
[1]University of Groningen
[2]University of Texas at Austin

We introduce MultiBLiMP 1.0, a massively multilingual benchmark of linguistic minimal pairs, covering 101 languages, 6 linguistic phenomena and containing more than 125,000 minimal pairs. Our minimal pairs are created using a fully automated pipeline, leveraging the large-scale linguistic resources of Universal Dependencies and UniMorph. MultiBLiMP evaluates abilities of LLMs at an unprecedented multilingual scale, and highlights the shortcomings of the current state-of-the-art in modelling low-resource languages.

Table of Contents

Reference and Modification in Universal Dependencies
Joakim Nivre and William Croft 1

Annotation of Relative Forms in the Egyptian-UJaen Treebank
Roberto A. Diaz Hernandez and Daniel Zeman 11

UD Treebanks for Esperanto as a natural language
Masanori Oya 22

Universal Dependencies for Suansu
Jessica K. Ivani and Kira Tulchynska 30

Crossing Dialectal Boundaries: Building a Treebank for the Dialect of Lesbos through Knowledge Transfer from Standard Modern Greek
Stavros Bompolas, Stella Markantonatou, Angela Ralli and Antonios Anastasopoulos 39

UD-English-CHILDES: A Collected Resource of Gold and Silver Universal Dependencies Trees for Child Language Interactions
Xiulin Yang, Zhuoxuan Ju, Lanni Bu, Zoey Liu and Nathan Schneider 52

A UD Treebank for Bohairic Coptic
Amir Zeldes, Nina Speransky, Nicholas E. Wagner and Caroline T. Schroeder 59

Negation in Universal Dependencies
Jamie Yates Findlay and Dag Trygve Truslew Haug 70

TreEn: A Multilingual Treebank Project on Environmental Discourse
Adriana Silvina Pagano, Patricia Chiril, Elisa Chierchiello and Cristina Bosco 80

Building UD Cairo for Old English in the Classroom
Lauren Levine, Junghyun Min and Amir Zeldes 97

Universal Dependencies for Sindhi
John Bauer, Sakiina Shah, Muhammad Shaheer, Mir Afza Ahmed Talpur, Zubair Sanjrani, Sarwat Qureshi, Shafi Pirzada, Christopher D. Manning and Mutee U Rahman 105

Universal Dependencies Treebank for Khoekhoe (KDT)
Tulchynska Kira, Job Sylvanus and Witzlack-Makarevich Alena 119

Parallel Universal Dependencies Treebanks for Turkic Languages
Arofat Akhundjanova, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami and Cagri Coltekin 129

Towards better annotation practices for symmetrical voice in Universal Dependencies
Andrew Thomas Dyer and Colleen Alena O'Brien 137

Extending the Enhanced Universal Dependencies – addressing subjects in pro-drop languages
Magali Sanches Duran, Elvis A. de Souza, Maria das Graças Volpe Nunes, Adriana Silvina Pagano and Thiago A. S. Pardo 143

Annotating Second Language in Universal Dependencies: a Review of Current Practices and Directions for Harmonized Guidelines
Arianna Masciolini, Aleksandrs Berdicevskis, Maria Irena Szawerna and Elena Volodina 153

Developing a Universal Dependencies Treebank for Alaskan Gwich'in
Matthew Kirk Andrews and Cagri Coltekin 164

Quid verbumst? Applying a definition of word to Latin in Universal Dependencies
Flavio Massimiliano Cecchini 174

ShUD: the First Shanghainese Universal Dependency Treebank
Qizhen Yang 186

Reference and Modification in Universal Dependencies

Joakim Nivre
Uppsala University
Department of Linguistics and Philology
`joakim.nivre@lingfil.uu.se`

William Croft
University of New Mexico
Department of Linguistics
`wacroft@icloud.com`

Abstract

Is the framework of Universal Dependencies (UD) compatible with findings from linguistic typology? To address this question, we need to systematically review how UD represents linguistic constructions in the world's languages, and how it handles the range of morphosyntactic variation attested in linguistic typology. In this paper, we start this review by discussing reference and modification constructions. The review shows that, although UD can represent all major constructions in this area, there are a number of cases where UD categories do not align systematically with a typological classification of constructions, and where constructional similarity is therefore not transparent across languages. We also identify limitations in the representation of certain morphosyntactic strategies, notably indexation and linkers. To overcome these limitations, we propose a number of revisions that may be considered for future versions of UD.

1 Introduction

Universal Dependencies (UD) is a framework for morphosyntactic annotation, which is designed to be applicable to all human languages in a way that enables meaningful cross-linguistic comparisons (Nivre et al., 2016, 2020; de Marneffe et al., 2021). To find out whether UD meets these requirements, Nivre (2025) proposes to build a constructicon for UD based on the survey of universal constructions and morphosyntactic realization strategies in Croft (2022) and the MoCCA database of comparative concepts derived from it (Lorenzi et al., 2024). In this framework, *constructions* are form-function pairings defined solely in terms of their function (hence universal), while *strategies* are defined by the pairing of a function with some cross-linguistically identifiable morphosyntactic form.

If we can provide a UD analysis for every combination of a construction and a strategy, then we can assess to what extent UD systematically captures cross-linguistic similarities and differences. In the same process, we can also gather evidence of gaps or inconsistencies in the current UD guidelines, and propose improvements for future versions.[1] In this paper, we present a first contribution to this project by discussing one of the most central constructions in the world's languages, that of nominal phrases, or referring expressions, including modification constructions that provide additional information about referents via their properties, their quantities, or their relations to other objects. Reference and modification constructions are discussed in Chapters 3–5 of Croft (2022).

2 Reference Constructions

Referring phrases are used to pick out and identify a referent, and they can be classified semantically into three broad categories, illustrated with a Russian example in (1) (Croft, 2022, p. 66):

(1)	ja	dal	knig-u	Ver-e
	I	gave	book-ACC	Vera-DAT

‘I gave Vera the book’

While *ja* identifies its referent *contextually* as the individual fulfilling the speaker role, *knigu* identifies the referent as belonging to a certain *type* (the book type), and *Vere* refers to a unique *individual*. Given these three semantic categories, we can define three basic reference constructions: *pronouns*, with contextual reference; *nouns*, with type reference; and *proper nouns*, with individual reference.

The basic reference constructions can be further subdivided both with respect to semantic content and information packaging. On the semantic side, nouns are often subdivided according to (degrees of) animacy, and pronoun systems often reflect ontological categories like person, thing, place, time

---

[1] An early review of UD from a typological perspective can be found in Croft et al. (2017).

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 1–10
August 27, 2025 ©2025 Association for Computational Linguistics

Construction	**Strategy**	**UD Annotation**
Pronoun	Zero	–
	Indexation	Predicate[Features]
	Word	PRON/DET
Noun	Word	NOUN
Noun + Determiner	Affix	NOUN[Features]
	Word	NOUN $\xrightarrow{\text{det}}$ DET
Proper Noun	Word	PROPN

Table 1: UD annotation of reference constructions and strategies. Features = indexation features.

and manner (Haspelmath, 1997). On the information packaging side, the most important notion is information status, which concerns the identifiability and accessibility of a referent. Information status is really a continuum, ranging from referents already mentioned in the discourse to purely hypothetical ones, but a broad distinction can be made between definite and indefinite referring expressions, where the defining criterion of the former is that the referent is already known to both speaker and hearer (Croft, 2022, p. 72–99).

The basic constructions are commonly realized on their own, as in (1), but especially (common) nouns are often combined with other expressions to form complex referring phrases. This often involves modification constructions, which will be discussed in Section 3, but it is also common to combine basic reference constructions. For example, contextual reference and type reference are often combined, as in the phrase *this book*, where the demonstrative *this* is used to constrain the type reference of the noun *book*. This combined use of the demonstrative – as opposed to its use as a pronoun on its own – is a subtype of the *determiner* construction, which also includes articles. Determiners are generally used to indicate the information status of the referent. However, while articles only express information status, demonstratives in addition encodes location with respect to the speaker and hearer (Croft, 2022, p. 73–74). Besides determiner constructions, it is common to combine several proper nouns, as in *Susan Smith*, and to combine common and proper nouns, as in *Aunt Susan*. More complex nominal phrases can be formed via modification, as discussed below in Section 3.

When it comes to morphosyntactic strategies, pronouns, nouns and proper nouns are most commonly realized as single words, but pronouns can also appear in reduced forms in many languages. They may be realized as clitics, that is, as morphemes that have the syntactic characteristics of a word but depend phonologically on another word or phrase; they may appear as affixes on the predicate, a strategy known as *indexation*; or they may not be phonetically realized at all, a *zero* strategy. Determiners can be realized as affixes on the noun or as separate words (sometimes with indexation of noun features such as number and gender). In general, strategies for referring expressions can be ranked on an *accessibility scale* (Givón, 1983; Ariel, 1988, 1990), where shorter expressions are preferred for higher accessibility referents.

How are the basic reference constructions represented in UD? Table 1 gives an overview of the constructions and strategies described in this section and their annotation in UD. In the following subsections, we discuss each case in detail and also make some observations about issues with the current guidelines and annotation practice in UD.

2.1 Pronouns in UD

The first thing to note here is that zero pronouns are not represented at all in UD. This is a consequence of the data-driven approach of UD (and most corpus annotation efforts), where the goal is to assign an interpretation to overtly observable forms, rather than to account for the realization of a certain content. This principle, which has been summarized in the slogan "Don't annotate things that are not there!" (Nivre, 2015), does have the drawback that core arguments of a predicate are sometimes not represented at all, but changing it would be a major reorientation of the UD approach to morphosyntactic annotation.

Moving on to pronouns that are realized by indexation on a predicate, also known as agreement, this is captured in the UD annotation through morphological features on the predicate. However,

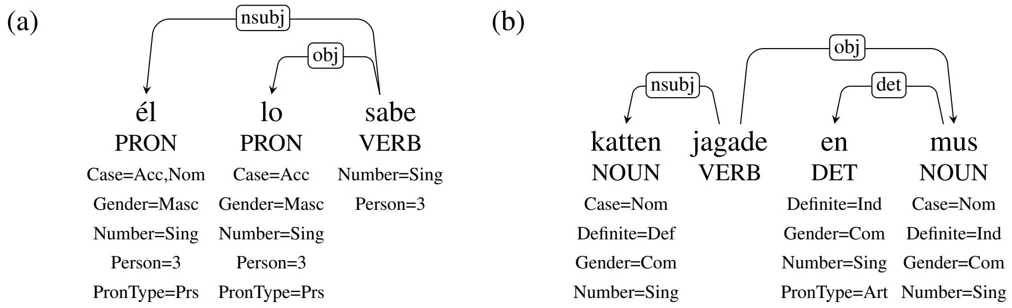


Figure 1: Simplified UD annotation of reference constructions: (a) pronouns, (b) nouns (with determiners).

when the indexation occurs together with an overt pronoun, there is nothing in the representation that links one to the other; and when there is no overt pronoun, there is no explicit information about which argument is being indexed. Thus, in the Spanish example in Figure 1(a),[2] there is nothing to indicate that the feature bundle [Number=Sing, Person=3] on the verb *sabe* (knows) corresponds to the specification of the subject pronoun *él* (he) rather than the object pronoun *lo* (it) (which happens to have the same values for these features). So if the subject pronoun is omitted, which is normal in Spanish if the subject is accessible in the context and not emphasized, it may accidentally look like Spanish has object-verb agreement. A possible improvement in future versions of UD could be to represent not only the features but also the target of the indexation, which would make the annotation more informative regardless of whether there is an overt realization of the target or not.

When a pronoun is realized as a word, finally, it is assigned the part-of-speech tag PRON and further subclassified using features, as illustrated in the Spanish example in Figure 1(a).[3] The feature PronType is used to distinguish major types of pronouns (and determiners) such as demonstrative, personal, interrogative, indefinite, and so on, while features such as Case, Gender, Number, and Person capture contrasting features within a given pronoun paradigm.

A special case of pronouns realized as words are clitics, which have the syntactic characteristics of a word but depend phonologically on another word or phrase, and which may therefore not be clearly recognizable as separate words in the standard orthography of a language. A typical case is Spanish *véase*, which consists of the imperative verb form *vea* (see) and the reflexive pronoun *se*. Cases like these are accommodated in UD through the mechanism of *multiword tokens*, which allows a single orthographic token (*véase*) to be treated as two separate syntactic words in the morphosyntactic annotation. However, multiword tokens are always optional, which means that there may be languages in UD that do not recognize all clitic pronouns as independent words. In such cases, the clitic will essentially be treated as an inflectional morpheme and analyzed by means of features (as the case of indexation discussed earlier).

2.2 Nouns and Determiners in UD

Nouns are almost always realized as independent words. In UD they are tagged with the universal part-of-speech tag NOUN and subclassified using features, as illustrated in the Swedish example in Figure 1(b). Common features for nouns include intrinsic features such as Gender, NounClass and Animacy, as well as inflectional features like Case, Number and Definite.[4]

The Swedish example also illustrates two different strategies for realizing determiners. In the subject noun phrase *katt-en* (cat-DEF), the definite article is realized as an inflectional suffix, captured in the annotation by the feature Definite=Def on the head noun. In the object noun phrase, the indefinite article is realized as an independent word *en*, which is linked to the head noun with the syntactic relation *det* (for determiner). The standalone article is tagged DET and has its own set of fea-

---

[2] In this and all following examples, we simplify the UD representations by omitting (a) lemmas and (b) morphological features that are not relevant for discussion (such as Tense and Mood features on the verb in this example).

[3] Some treebanks keep the tag DET also in pronominal uses for demonstratives regularly used as determiners, such as *all* and *some* in English, which is currently a source of cross-linguistic inconsistency in UD.

[4] The occurrence of nominative case features (Case=Nom) on both the subject and object is due to the Swedish case system having been reduced to just two cases: nominative and genitive.

tures, including PronType=Art and Definite=Ind, which together distinguish indefinite articles from other types of determiners.

The tag DET and the relation *det* are used not only for articles but also for other determiners, including demonstratives (*this book*), quantifiers (*all books*), and sometimes possessives (*my book*), which can be problematic from a typological perspective. We will return to this issue in Section 3.

2.3 Proper Nouns in UD

UD has a special part-of-speech tag PROPN for words that are primarily used as proper nouns, such as *Mary*, *Smith*, *London*, and *Sweden*. When several proper nouns are combined into a referring phrase, such as *Mary Smith*, they are tagged PROPN and combined with the syntactic relation *flat* (unless one of them is clearly distinguishable as the syntactic head).

It should be noted that not all phrases that are used with individual reference are analyzed in this way in UD. Phrases like *the North Sea* and *Gone with the Wind*, which are compositional phrases where the head word may not even be nominal, are annotated according to their internal syntactic structure even when they are used as phrases with individual reference. Hence, the proper noun construction, defined as the coupling of any linguistic form with the function of individual reference, is only partially captured in UD by the part-of-speech tag PROPN. One way of improving the correspondence would be to add the feature ExtPos=PROPN to names with internal syntactic structure.[5]

3 Modification Constructions

The information packaging function of modifiers, or attributive phrases, is to add information to help identify the referent of a referring phrase. Croft (2022) distinguishes six basic modification constructions, exemplified in (2):

(2) a. the *black* dog
b. *five* books
c. the *third* day
d. *a pound of* sugar
e. *Peter's* mother
f. the man *who got away*

In (2a), *black* is an *adjectival modifier*, or simply *adjective* in Croft's terminology, which helps identify the referent of the head noun *dog* by adding information about a *property*, in this case its color. Besides color, adjectival modifiers commonly denote properties such as shape (*round*), age (*old*), value (*good*), and dimension (*big*) (Dixon, 1977). From an information packaging point of view, adjectival modifiers are said to be *subcategorizing*. Adjectival modifiers can be combined with *admodifiers*, like *very* in *a very big house*, which describe semantic operations on the scale denoted by the modifier. Besides intensifiers like *very*, admodifiers can be downtoners (*a rather big house*), comparatives (*a bigger house*), or superlatives (*the biggest house*).

The following three examples (2b–2d) instead belong to the class of *selecting* modifiers, whose information packaging function is to select an instance or set of instances using information about quantity or set membership. This is a diverse class, where three main constructions are distinguished: *numeral quantifiers*, *set-member terms*, and *mensural terms* (Croft, 2022, p. 109–111). Numeral quantifiers include cardinal numerals like *five* in (2b) together with a wide range of different quantifiers like *several* (a vague numeral), *most* (a proportional quantifier), and *each* (a distributive quantifier). Set-member terms include ordinal numerals like *third* in (2c) as well as non-numerical terms like *next*, *last*, and *other*. Numeral quantifiers and set-member terms both presuppose that the head noun denotes an individuated entity. This is not the case for mensural terms, exemplified by the measure term *a pound of* in (2d) and including a diverse set of constructions referring to containers (*cup*), groups (*flock*) or pieces (*slice*), among others (Koptjevskaja-Tamm, 2001).

The next modification construction, exemplified in (2e), is *nominal modification*, where the referent of the head noun is identified by its relation to another referent, such as the kinship relation in *Peter's mother*. The number of relations that can be invoked in this way is in principle unlimited, but other common relations are ownership (*Peter's car*), body-part relations (*Peter's arm*), and figure-ground relations (*the book on the table*) (Kay and Zimmer, 1990). From an information packaging point of view, nominal modifiers have a *situating* function, which makes them semantically similar to the pronoun and determiner constructions discussed in Section 2.

---

[5] The ExtPos feature can be used in UD to specify the part-of-speech category that a multiword expression would get if it were analyzed as a single word.

Construction	**Strategy**	**UD Annotation**
Adjectival modifier	Simple/Indexation	RE $\xrightarrow{\text{amod}}$ ADJ[Features]
Admodifier	Word	ADJ $\xrightarrow{\text{advmod}}$ ADV
	Affix	ADJ[Degree=X]
Numeral quantifier	Simple/Indexation	RE $\xrightarrow{\text{nummod}}$ NUM[Features]
		RE $\xrightarrow{\text{nummod}}$ NUM $\xrightarrow{\text{clf}}$ NOUN
		RE $\xrightarrow{\text{det}}$ DET[Features]
		RE $\xrightarrow{\text{amod}}$ ADJ[Features]
Set-member	Simple/Indexation	RE $\xrightarrow{\text{amod}}$ ADJ[Features]
Mensural quantifier	Simple	RE $\xrightarrow{\text{nmod}}$ NOM
	Flag: Adposition	NOM $\xrightarrow{\text{nmod}}$ RE $\xrightarrow{\text{case}}$ ADP
Nominal modifier	Simple/Indexation	RE $\xrightarrow{\text{nmod}}$ NOM[Features]
		RE $\xrightarrow{\text{det}}$ NOM[Features]
	Flag: Affix	RE $\xrightarrow{\text{nmod}}$ NOM[Case=X]
	Flag: Adposition	RE $\xrightarrow{\text{nmod}}$ NOM $\xrightarrow{\text{case}}$ ADP
	Compounding	RE $\xrightarrow{\text{compound}}$ NOM

Table 2: UD annotation of modification constructions and strategies. RE = referring expression; NOM = nominal (noun, proper noun or pronoun); Features = indexation features (possibly empty).

The final example in (2f) shows *action modification*, where the referent is being identified in relation to an event, and its prototypical realization in the form of a relative clause. Since subordinate clauses will be the topic of a later paper, we will not discuss action modification further in this paper, except to note that the relative clause construction can sometimes be recruited for property modification as an alternative to adjectival modification (Croft, 2022, pp. 113–114).[6]

The morphosyntactic strategies of modification constructions can be divided into four main types Croft (2022, pp. 114–138):

- **Simple:** The modifier is combined with the referring expression without any additional element; the modifier may be realized as an independent word (*juxtaposition*), through *compounding* or *affixation*. Juxtaposition is exemplified in the English examples (2a–c), where the adjective *black*, the cardinal numeral *five*, and the ordinal numeral *third* are all placed next to the head noun without any additional element.

- **Relational:** The combination involves a third element – a *flag* – that encodes the semantic relation between referent and modifier (Malchukov et al., 2010); the flag may be realized as an independent word (*adposition*) or as an affix (*case marker*). The former is exemplified by the preposition *de* in French *la mère de Pierre* (Pierre's mother), the latter by the genitive case inflection in Latin *mater Petri* (Petrus's mother); cf. Figure 4(a–b).

- **Indexical:** The combination involves a third element – an *index* – encoding features of the referent (such as person, number or gender/class) (Croft, 2003). Indices are most commonly realized as affixes, as in French *chien noir* (black dog), where the inflectional form of the adjective *noir* indicates the gender (masculine) and number (singular) of the noun; cf. Figure 2(a). A special case of the indexical strategy is the *classifier* strategy, commonly found with numerals, as in Chrau *du tong aq* (one crossbow), where the classifier *tong* (for long objects) is required with the numeral *du* (one) and the noun *aq* (crossbow) (Thomas, 1971).

- **Linker:** The combination involves a third element – a *linker* – invariant with respect to the features characteristic of flags and indices

---

[6] Recruitment occurs when one construction borrows the morphosyntactic form of another (usually more prototypical) construction (Croft, 2022, p. 53–58).

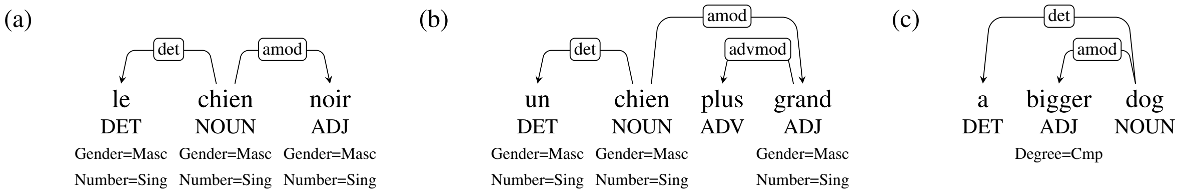


Figure 2: UD annotation of adjectival modifiers: (a–b) with indexation, (b–c) with admodifier.

(that is, it neither indicates a specific semantic relation nor encodes referent features). A typical example is Persian *âb-e garm* (hot water), where the linker *e* appears with the nominal *âb* (water) and the adjectival modifier *garm* (hot) but does not contrast with any other morpheme used to relate these construction elements to each other.

As discussed above, modifiers can furthermore be accompanied by admodifiers, which can be realized as independent words or affixes.

How does UD represent the different modification constructions and strategies? In Table 2, we give a schematic overview of the most important cases to be discussed in following subsections.

3.1 Adjectives and Admodifiers in UD

Adjectival modifiers are normally realized as independent words, which are analyzed in UD by attaching them to their nominal head with the *amod* relation and assigning them the part-of-speech tag ADJ. In an English example like *the black dog*, this is a simple juxtaposition strategy, as there is no additional word or morpheme mediating the relation. However, adjectives are also commonly found with the indexation strategy, where the adjective inflects to agree with the nominal head with respect to features such as gender and number, as illustrated for the corresponding French example *le chien noir* (the black dog) in Figure 2(a). In the latter case, the UD annotation combines the *amod* relation and the ADJ tag with morphological features on the adjective. As noted in Section 2.1, nothing in the annotation indicates that this is indexation, as opposed to the adjective and noun accidentally having identical features values for gender and number, but this is less problematic here because there is only one candidate controller.

Admodifiers realized as independent words are linked to their adjectival modifier heads with the *advmod* relation and are normally assigned the part-of-speech tag ADV, as shown for the French example *un chien plus grand* (a bigger dog) in Figure 2(b). Admodifiers realized as affixes are instead captured by morphological features on the adjective, as seen in the English example *a bigger dog* in Figure 2(c), where the adjective *bigger* is assigned the feature Degree=Cmp. It is worth noting that, while the features used for inflectional admodifiers are quite specific, the use of the *advmod* relation lumps the independent word admodifiers together with a very large and diverse group of expressions, including manner adverbials as well as temporal and locative expressions, among other things.

3.2 Selecting Modifiers in UD

While the annotation of adjectival modifiers in UD appears straightforward, the situation is more complex for the group of selecting modifier constructions. Starting with numeral quantifiers, UD clearly distinguishes *cardinal numerals*, which are annotated with the *nummod* relation and the part-of-speech tag NUM, categories that are not used for any other construction. This is illustrated for the Swedish example *fem böcker* (five books) in Figure 3(a). Other numeral quantifiers, like *many* and *all*, should according to the guidelines be annotated as determiners (with the *det* relation and the tag DET), like the Swedish example *alla böcker* (all books) in Figure 3(b). In practice, however, they are sometimes annotated as adjectival modifiers (with the *amod* relation and the tag ADJ), like the Swedish example *många böcker* (many books) in Figure 3(c).[7] Numeral quantifiers can also be realized using the *classifier* strategy, in which the numeral (or determiner) is combined with a special classifying element (often historically a noun and tagged as such in UD), as shown schematically in Table 2.

Set-member terms include ordinal numerals, which in UD are not analyzed using the *num-*

---

[7] A similar variation between DET and ADJ is found also in other languages, such as English (*all* = DET, *many* = ADJ) and French (*plusieurs* = DET, *tous* = ADJ).

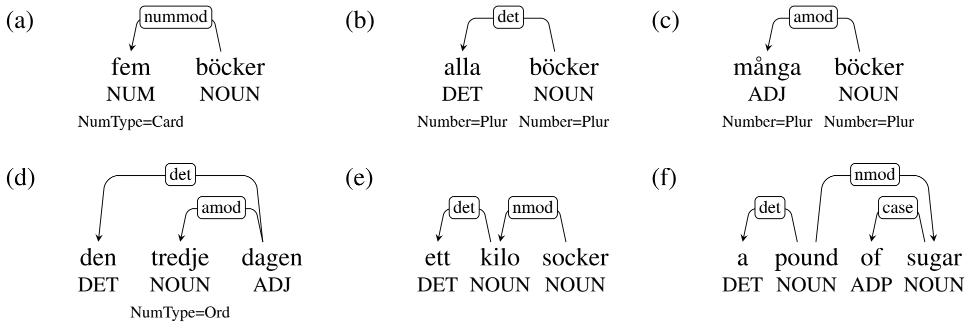


Figure 3: UD annotation of selecting modifiers: (a–c) numeral quantifier, (d) set member, (e–f) mensural quantifier.

mod relation and NUM tag reserved for cardinal numerals. Instead, they are analyzed as adjectival modifiers, as shown for the Swedish example *den tredje boken* (the third book) in Figure 3(d). The same analysis is used for non-numerical terms like *next*, *last*, and *other*, although we suspect that some of the more pronoun-like cases may be analyzed as determiners in some treebanks.

Mensural quantifiers are the construction type that causes the most problems for UD, because some of its strategies give rise to a mismatch between the syntactic and semantic head. The mensural part of this construction is typically realized as a nominal phrase, such as *a pound* or *a bottle*. When this is combined with the referring nominal using simple juxtaposition, as in the Swedish example *ett kilo socker* (a kilo of sugar) in Figure 3(e), the mensural part (*ett kilo*) can be analyzed as a modifier, linked to its head (*socker*) with the *nmod* relation, in which case the syntactic and semantic heads coincide. However, when the combination of the two parts involves an adposition, as in the English example *a pound of sugar* in Figure 3(f), the UD guidelines give priority to the grammatical form, which indicates that the mensural part is the syntactic head. As a result, we fail to capture the common construction in the two languages and end up with *nmod* relations going in opposite directions.

To arrive at a more transparent annotation of constructions and strategies, future versions of UD should consider some revisions of the annotation guidelines. Since the current distinction between *det*, *amod*, and *nummod* does not align well with comparative concepts from linguistic typology, these relations can be replaced by a general *mod* relation.[8] More specific constructions can be distinguished using part-of-speech tags, where we suggest restricting DET to demonstratives and articles, keeping ADJ for property words, and replacing NUM by a broader QNT category that includes quantifiers and set-member terms. If part-of-speech tags are not specific enough, syntactic subtypes may be used as well. In addition, we propose a new relation *admod* for the admodifier construction, which is distinct from other constructions currently covered by the *advmod* relation.

For mensural quantifiers, we advocate an analysis that consistently treats the measured noun as the head, as in the Swedish example in Figure 4(e). This can be seen as a *construction-oriented* analysis, where syntactic relations are aligned to capture constructional similarity whenever possible, as opposed to a *strategy-oriented* analysis, where priority is given to similarities in strategies. To resolve the apparent conflict between form and function in cases like the English example in Figure 4(f), we propose to analyze this as a linker strategy rather than a case marking strategy.

3.3 Nominal Modifiers in UD

Nominal modifiers are typically realized as nominal phrases, headed by a pronoun, noun or proper noun. With a few exceptions to be discussed shortly, they are analyzed in UD by attaching them to their nominal head with the *nmod* relation. Nominal modifiers are often realized using a relational strategy, with a flag that can take the form of an adposition, as in the French example *la mère de Pierre* (Pierre's mother) in Figure 4(a), or an affix, as in the Latin example *mater Petri* (Petrus's mother) in Figure 4(b). In the former case, the flag is attached to the nominal modifier with the *case* relation; in the latter, it is represented by a Case feature on the modifier.

There are two exceptions to the rule that nom-

---

[8] This change has been made in Surface Syntactic Universal Dependencies (SUD) (Gerdes et al., 2018, 2019, 2021).

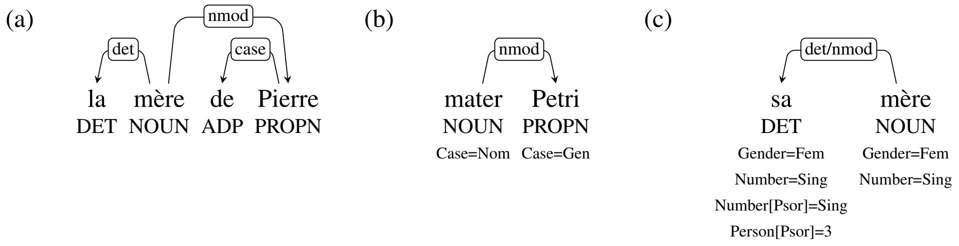


Figure 4: UD annotation of nominal modifiers.

inal modifiers are analyzed as an instance of the *nmod* relation. The first concerns possessive pronouns, as in the French example *sa mère* (his/her mother) in Figure 4(c). Possessive pronouns alternate with full nominal phrases to express nominal modification, so from a constructional point of view it makes sense to subsume them under the *nmod* relations. In many languages, however, they resemble determiners with respect to word order and/or indexation. The UD guidelines therefore currently allow possessives to be analyzed either as nominal modifiers (with the *nmod* relation) or as determiners (with the *det* relation).[9] We propose to use *nmod* consistently to capture constructional parallelism across languages. More generally, we think strategy-oriented relations should be restricted to strategies realized as independent function words, such as adposition flags, while relations between construction elements should be construction-oriented, as discussed above in connection with mensural classifiers.

The second exception is found in nominal modifiers that are realized with a compounding strategy, such as the English *iron pipe*, where the UD annotation prioritizes the information that the two components form a word, rather than a phrase, and attaches the modifier to the head with the *compound* relation. This highlights a more general issue in UD annotation, where construction-oriented relation types like *nmod* and strategy-oriented relation types like *compound* can come into conflict because the UD representation only allows one type for each syntactic relation. In line with the principle of prioritizing construction-oriented relations between construction elements, we propose to use *nmod* instead of *compound* for this kind of nominal modification. We are aware that this will blur the distinction between compounding and phrasal modification, but we think this can be remedied using features or subtyping.[10]

One final observation concerning nominal modification is that it is unclear how the linker strategy should be annotated in UD. If the linker is realized as an affix, a morphological feature can be added to its head, but if it is realized as an independent word, it seems the best we can do currently is to lump linkers and flags together under the *case* relation.[11] For future versions of UD, we therefore propose a new relation *lnk*, which can be used together with the tag PART to annotate linkers.

4 Discussion

While our review has shown that the UD annotation framework can represent all the major constructions and strategies for reference and modification discussed in Croft (2022), we have also seen that the correspondence between comparative concepts and elements of the UD annotation is quite complex. Of course, UD has developed without an explicit aim to represent constructions, let alone distinguish strategies for those constructions across languages. Nevertheless, our review reveals ways in which UD can capture properties of constructions and strategies, using a combination of syntactic relations and part-of-speech tags.

First, some, perhaps most, of the syntactic relations closely match the major information packaging constructions described in Croft (2022), albeit subdivided. For the constructions reviewed here, the main information packaging constructions are modification and admodification. Our suggestions

---

[9] As it happens, the French UD treebanks prefer the *det* analysis, while the English and Swedish UD treebanks use *nmod* instead, a discrepancy that does not appear to be motivated by the linguistic facts.

[10] It is worth noting in this context that the analysis of compounds in UD is not consistent across languages because of differences in orthography. Thus, the Swedish equivalent of *iron pipe*, *järnrör*, is currently not analyzed as a compound at all, because it is written as a single orthographic token.

[11] In UD v2.15, the Tagalog-Ugnayan and Cebuano-GJA treebanks use the *mark* relation for linkers that occur with nominal and adjectival modifiers, which is surprising given that the linkers do not mark relations between clauses.

for future revisions of the relations, specifically distinguishing *admod* from *advmod* and having a single *mod* relation replacing *det*, *amod* and *nummod*, increases this parallelism. Naturally, there will be some mismatches between construction function and morphosyntactic form, but we assume that these are relatively infrequent compared to the general matching of form and function. Another suggestion of ours that makes UD relations more closely match constructions is for a uniform representation of mensural constructions such that the measured noun is the head, which is facilitated by analyzing the etymological adposition in the pseudo-partitive strategy as a linker, not a flag.

Second, some of the part-of-speech categories are close to semantic categories. For example, NUM describes the semantic category of numerals; in fact, in many languages (most famously Russian), different numerals use different modification constructions. The categories NOUN, PROPN and PRON parallel the distinction between type reference, individual reference, and contextual reference. Modification is characterized by a similar information packaging function carried out by different semantic categories of modifiers. Our narrowing of DET to articles and demonstratives, and introducing the new tag QNT, again increases this parallelism.

Third, UD must represent independent words that form part of certain strategies (also known as “function words”), since they are part of morphosyntax. This is currently done using syntactic relations, part-of-speech tags, or both. For a better treatment of modification, we suggest a new *lnk* relation, distinct from *case*, to handle the grammaticalized function word in the pseudo-partitive strategy for mensural constructions, as well as in certain genitive and other modification constructions. The linker itself is assigned the tag PART, a tag that is used for function words in a variety of strategies.

5 Conclusion

In this paper, we have taken a first step towards a constructicon for UD, in the sense of Nivre (2025), by reviewing the way UD annotates constructions and strategies for reference and modification, following the taxonomy of Croft (2022). The constructicon is shown in Tables 1 and 2, where we outline how these constructions and strategies are currently annotated according to the UD guidelines. (In passing, we have also remarked on a few inconsistencies in the way that these guidelines are applied across languages.) On the positive side, we have found that UD can represent almost all constructions and strategies discussed in the survey. On the negative side, we have found that UD categories do not always align systematically with comparative concepts from typology, that there is sometimes a conflict between annotating elements of constructions and strategies, respectively, and that some strategies are not well captured in the UD framework.

In some cases, we have made concrete proposals for future revisions of UD, revisions that would improve the correspondence between UD categories and comparative concepts. It is worth clarifying that these revisions are incompatible with the current version of the UD guidelines (v2), since they involve changes to the set of syntactic relations and part-of-speech tags, which means that they could only be considered for v3 of the guidelines. Moreover, these proposals need to be evaluated also from other perspectives, since UD is designed as “a very subtle compromise between a number of competing criteria” (de Marneffe et al., 2021, p. 302) and should be suitable for language-specific analysis as well as typological language comparison, and should be accessible to non-experts and suitable for processing by computers. Finally, the discussion needs to be informed by a more comprehensive review of the UD framework, covering all major types of constructions and strategies. It is our goal to continue this review in a series of future publications.

Acknowledgments

We are grateful to members of the UniDive COST Action (CA21167) for useful discussions, and acknowledge the support of the Swedish Research Council (grant no. 2022-02909). We thank André Coneglian, Leonie Weissweiler and three anonymous reviewers for comments and suggestions that helped improve the article.

References

Mira Ariel. 1988. Referring and accessibility. *Journal of Linguistics*, 24:65–87.

Mira Ariel. 1990. *Accessing Noun Phrase Antecedents*. Routledge.

William Croft. 2003. *Typology and Universals. Second Edition*. Cambridge University Press.

William Croft. 2022. *Morphosyntax: Constructions of the World's Languages*. Cambridge University Press.

William Croft, Dawn Nordquist, Katherine Looney, and Michael Regan. 2017. Linguistic typology meets Universal Dependencies. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, pages 63–75.

Roger M. W. Dixon. 1977. Where have all the adjectives gone? *Studies in Language*, 1:19–80.

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.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2019. Improving Surface-Syntactic Universal Dependencies (SUD): Surface-syntactic relations and deep syntactic features. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories*, pages 126–132.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2021. Starting a new treebank? Go SUD! Theoretical and practical benefits of the surface-syntactic distributional approach. In *Proceedings of the 6th International Conference on Dependency Linguistics*, pages 35–46.

Talmy Givón. 1983. Topic continuity in discourse: An introduction. In Talmy Givón, editor, *Topic Continuity in Discourse*, pages 1–41. John Benjamins.

Martin Haspelmath. 1997. *Indefinite Pronouns*. Oxford University Press.

Paul Kay and Karl Zimmer. 1990. On the semantics of compounds and genitives in English. In S. L. Tsohatzidis, editor, *Meanings and Prototypes: Studies in Linguistic Categorization*, pages 239–246. John Benjamins.

Maria Koptjevskaja-Tamm. 2001. 'a slice of the cake' and 'a cup of tea': partitive and pseudo-partitive constructions in the Circum-Baltic languages. In Östen Dahl and Maria Koptjevskaja-Tamm, editors, *Circum-Baltic Languages: Their Typology and Contacts*, pages 523–568. John Benjamins.

Arthur Lorenzi, Peter Ljunglöf, Ben Lyngfelt, Tiago Timponi Torrent, William Croft, Alexander Ziem, Nina Böbel, Linnéa Bäckström, Peter Uhrig, and Ely A. Matos. 2024. MoCCA: A model of comparative concepts for aligning constructicons. In *Proceedings of the 20th Joint ACL – ISO Workshop on Interoperable Semantic Annotation*, pages 93–98.

Andrej Malchukov, Bernard Comrie, and Martin Haspelmath. 2010. Ditransitive constructions: a typological overview. In Andrej Malchukov, Bernard Comrie, and Martin Haspelmath, editors, *Studies in Ditransitive Constructions: A Comparative Handbook*, pages 1–64. Mouton de Gruyter.

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

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 3–16. Springer.

Joakim Nivre. 2025. Constructions and strategies in Universal Dependencies. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 419–423.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Dan Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 1659–1666.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Dan Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, pages 4034–4043.

David D. Thomas. 1971. *Chrau Grammar*. University of Hawaii Press.

Annotation of Relative Forms in the Egyptian-UJaen Treebank

Roberto Antonio Díaz Hernández[1], **Daniel Zeman**[2]
[1] Univeristy of Jaén (radiaz@ujaen.es)
[2] Charles University (zeman@ufal.mff.cuni.cz)

Abstract

Relative forms are adjective finite verb forms that can be used in an attributive or a nominal function. They pose a challenge when annotating them according to the Universal Dependencies approach, for they have morphological features of both verbs and adjectives, yet they can also be used syntactically as nouns. The aim of this paper is to discuss the morphosyntactic methodology applied to their annotation in the Egyptian-UJaen treebank.

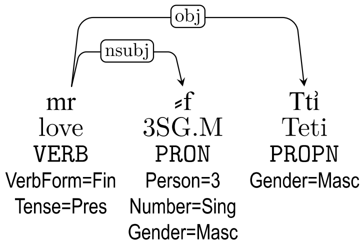
1 Introduction

The Egyptian-UJaen treebank in Universal Dependencies (hereafter UD-EUJA treebank) contains 21,945 words and 2,192 sentences,[1] most of which are taken from the Pyramid Texts written in Old Egyptian (ca. 2700–2000 BC).[2] Since the beginning of this UD treebank in March 2024, the annotation of Egyptian morphosyntactic features according to the Universal Dependencies approach has been a challenge, not only because of the idiosyncrasy of Old Egyptian, but also because of the lack of dependency grammar works by Egyptian philologists.[3]

The present paper is the first of a series of studies on the annotation of Egyptian morphosyntactic features in the UD-EUJA treebank. The aim of these studies is to discuss issues arising from the application of the UD approach (De Marneffe, Manning, Nivre and Zeman, 2021) on Egyptian syntax.

The Old Egyptian relative forms have been chosen as the subject of this paper because of their idiosyncratic nature.[4] They have certainly been described as hybrid forms, half participle, half personal verb forms, which deviate more than any other verb form from the categories we know from the classical languages.[5] A relative form differs from a verb form in two ways. Firstly, the antecedent of a relative form is the dependent of a verb form, for example the antecedent of a relative form can be the object of a verb form. Secondly, the relative form agrees in gender and number with its antecedent, cf.:

1) Example of a verb form followed by its subject and object, PT 535a T (EUJA-1607):



'(...) he loves Teti.'

---

[1] The present paper is based on the latest version of the treebank published in UD release 2.16 (May 2025).

[2] 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.

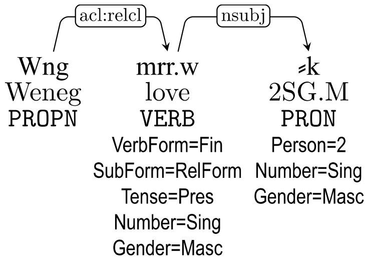
[3] For issues related to the inception and development of the UD-EUJA treebank see Díaz Hernández and Passarotti 2024.

[4] Relative forms disappeared in later stages of Egyptian. In Coptic they were replaced by constructions with the relative pronoun ⲉⲧ-, ⲉⲧⲉ-, ⲛ̄ⲧ-, see Loprieno, 1995, 98.

[5] "... cette forme hybride, moitié participe, moitié mode personnel, et qui s'écarte plus que toute autre des catégories que les langues classiques nous ont rendues familières." Sottas, 1925, 264, quoted by Polotsky, 1976, 2, footnote 2. English translation: "... this hybrid form, half-participle half-personal form, deviates more than any other from the categories that classical languages have made familiar to us".

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 11–21
August 27, 2025 ©2025 Association for Computational Linguistics

2) Example of a masculine relative form, PT 607d T (EUJA-1771):



'(...) Weneg whom you love.'

This paper deals with the following issues:

- General features of the relative forms and their annotation in the UD-EUJA treebank (2).
- Morphological features of the relative forms: gender, number and the *i̓* prefix (also called "prothetic") (3).
- A semantic classification of four relative verb forms—a past/perfective relative form (or past-2 relative form), a present/imperfective relative form, a future/perfect relative form, and Clère's past relative form (or past-1 relative form) (4).
- A syntactic analysis (5) consisting of two parts based on the dual function of the relative forms. These can be attributive relative forms, modifying an antecedent, or nominal relative forms that function syntactically as nouns. Special focus will be laid on the change of the dependency relationship in the nominal relative forms.
- A conclusion (6).

2 General features of the relative forms

A relative form is an adjective *finite* verb form (Polotsky, 1976, 4, 13 and Schenkel, 2012, 279). It is adjective-like for it modifies and agrees with a noun, and it is verb-like for it is finite and has a subject. In Earlier Egyptian, which includes Old Egyptian and Middle Egyptian, relative forms contrast with participles, which are adjective *non-finite* verb forms, i.e. they lack a personal inflection.[6]

The UD-EUJA treebank contains currently 190 relative forms. The following features are used to annotate relative forms in the UD-EUJA treebank:

1) UPOS column: VERB.

2) FEATS column:

- Gender of the relative form (Gender=Masc/Gender=Fem).
- Number of the relative form (Number=Sing/Number=Plur).
- Nominal function (Nominal=Yes) (see 5.2, below).
- A prefix *i̓* (Hebrew *yodh*) occasionally used in relative forms (see Table 3, Appendix). It is annotated with the key "Prefix=Yodh".
- Verb class to which the verb of the relative form belongs.[7] As in other languages such as Arabic, the Earlier Egyptian verb system can be classified into verb classes according to the morphology of the verb. There are verbs with 1 to 6 radicals (Schenkel 2012: 183). If all radicals are strong, the abbreviation *lit* (literal) is used, for example *č̣ṭ* "say" is a *2-lit.* verb. If the last radical is weak, the abbreviation *inf* (*infirma* i.e. weak) is used, for example *mri̯* "love" is a *3ae-inf* verb. Verbs with a reduplicated radical are labelled *red.* (*reduplicata*) while causative verbs bear the abbreviation *caus*.
- Verb tense to which the Egyptian tense corresponds (see features in MISC column, below). While "Tense=Pres" and "Tense=Fut" usually matches in MISC column and FEATS column, "Tense=Past" in FEATS column can be used for "Tense=Past-1" and "Tense=Past-2" in MISC column.
- The key SubForm=RelForm identifies a relative form.

3) MISC column:

- Egyptian tense indicated by four keys (see section 4, below): Tense=Past-1, Tense=Past-2, Tense=Pres, Tense=Fut.

---

[6] A morphosyntactic comparison of relative forms and participles should be taken into account when analysing the annotation of participles in the UD-EUJA treebank.

[7] See the list of Egyptian verb classes in https://universaldependencies.org/egy/feat/VerbClass.html#examples.

3 Morphological features

In its current state the UD-EUJA treebank contains 88 masculine relative forms. 83 of which are singular and 5 are plural. The masculine singular relative forms usually lack an ending. However, a *w* ending referring to a masculine singular antecedent is occasionally added when the tense of the relative form is present (Gardiner, 1957, 297/§ 380; see examples taken from the UD-EUJA treebank in Table 1, Appendix). [8] In contrast, masculine plural relative forms occasionally feature a *w* in past-2 relative forms (see examples taken from the UD-EUJA treebank in Table 2, Appendix)

2) Example of a masculine relative form, PT 607d T (EUJA-1771, see diagram, above):

Wng	*mrr.w*	*⸗k*
Weneg-GN	love:REL.PRES-M.SG	=2SG.M

"(...) Weneg whom you love."

If the antecedent is feminine, the ending of a relative form is *t*. There are 102 feminine singular relative forms in the UD-EUJA treebank. No feminine plural relative form has been annotated so far. The *t* ending is also used in relative forms with a neuter meaning (Gardiner, 1957, 297/§ 380 and Loprieno, 1995, 86; cf. ex. 16, below). The gender of relative forms with neuter meaning is annotated as feminine in the FEATS column, as their morphological gender remains feminine.

3) Example of a feminine relative form, PT 38c W (EUJA-179):

ỉr.t	*Ḥr.w*	*sẖn.t*	*⸗k*
eye-F.SG	Horus-GN	embrace:REL.FUT-F.SG	=2SG.M

"(...) The eye of Horus which you shall embrace (...)"

An *ỉ* prefix, also called "prothetic" (Uljas, 2023 and Stauder, 2014, 256), is occasionally added to the relative forms. The *ỉ* prefix only appears with certainty in past-2 and present relative forms (see Table 3, Appendix).[9] The causative verb with the second reduplicated radical (caus. 2ae-red.) *ś:fkk.tn* and the verbs with the fourth weak radical (4ae-inf.) *nśbi̯* and *ẖẖmi̯* occur with the *ỉ* prefix in the past-2 relative form, while the verb with the third reduplicated radical *śẖꜣꜣ* (*3ae-red*), the verb with two strong radicals (*2-lit*) *č̣t* and the verb with the third weak radical (*3ae-inf*) *šmi̯* are used with the *ỉ* prefix in the present relative form.

4 Semantic function

Since relative forms are verb forms, they are characterised by verb features. However, there is no general agreement on the terminology and semantic function of relative forms. While some authors emphasise their aspect feature using the terms "perfective", "imperfective" and "perfect" (Allen 2010: 354, Borghouts 2010: 277–278), others consider them tenses and classify them as past, present and future relative forms (Ockinga 2005: 67–69 and Schenkel 2012: 280). Despite this disagreement, most grammars concur that there are three relative forms:[10]

a) A past/perfective relative form (or past-2 relative form) characterised by the ending *n* and annotated with the keys "Tense=Past-2" in the MISC column and "SubForm=RelForm" in the FEATS column of the UD-EUJA treebank, e.g.:

4) PT 77a W (EUJA-1552):

ỉr.t	*Ḥr.w*	*ỉṯḥ.tn*	*⸗f*
eye-F.SG	Horus-GN	pull out:REL.PAST-2-F.SG	=3SG.M

"(...) the Eye of Horus which he pulled out."

b) A present/imperfective relative form characterised by the reduplication of the second radical in verbs with a third weak radical (*3ae-inf*), such as *mri̯* "to love". They are annotated with the keys "Tense=Pres" in the MISC column and "SubForm=RelForm" in the FEATS column of the UD-EUJA treebank, e.g.:

5) PT 15 N (EUJA-83):

sꜣ	*⸗k*	*mrr.w*	*⸗k*
son-M.SG	=2SG.M	love:REL.PRES-M.SG	=2SG.M

"(...) Your son whom you love (...)"

---

[8] The PT column provides the reference to the hieroglyphic text in Sethe (1908–1922) and the EUJA column indicates the id sentence in the Egyptian treebank.

[9] The instance of *ỉ:nśb.t* annotated as a past-1 relative form is probably another instance of a past-2 relative form with omission of the ending *n* (cf. *ỉ:nśb.tn* in PT 98c).

[10] The past/perfective and present/imperfective relative forms are known from Erman's and Sethe's works at the end of the 19th century and the beginning of the 20th century, see Erman, 1894, 164–166 (§§ 394–399) and Sethe, 1899, 327 (§ 754 foll.). The future/perfect relative form was discovered later by Gunn (1924, 1–25).

c) A future/perfect relative form characterised by the ending *y* and the absence of a reduplicated second radical in verbs with a third weak radical (*3ae-inf*). They are annotated with the keys "Tense=Fut" in the MISC column and "SubForm=RelForm" in the FEATS column of the UD-EUJA treebank, e.g.:

6) PT 625d T (EUJA-1822):

bw	*nb*	*mr.y*	*⸗k*	*ỉm*
place-M.SG	any (M.SG)	love:REL.FUT-M.SG	=2SG.M	there:ADV

"(...) any place where you will desire to be (lit: you will desire in it)."

In addition, Clère (1949) and Schenkel (2010) argued for the existence of a fourth relative form with a past tense (past-1 relative form). It occasionally features a *w* ending in verbs with the third weak radical (*3ae-inf*), such as *mśỉ* "to bear, to give birth" and in verbs with three strong radicals (*3-lit*), such as *śnḳ* "to suck". Contrary to the present relative form, the second radical of verbs with the third weak radical is not reduplicated in the past-1 relative form, e.g.:

7) PT 623a T (EUJA-1814):[11]

ꜣḫ	*mś.w*	*Nw.t*	*śnḳ.w*	*Nb.t-ḥw.t*
spirit-M.SG	give birth:REL.PAST-1-F.SG	Nut-GN	suckle birth:REL.PAST-1-F.SG	Nephthys-GN

"(...) The spirit whom Nut bore, whom Nephthys suckled (...)"

This relative form is annotated with the keys "Past-1" in the MISC column and "SubForm=RelForm" in the FEATS column of the UD-EUJA treebank. According to Schenkel, the past-1 relative form is an archaic feature that was gradually replaced by the past-2 relative form with an *n* ending (Schenkel, 2010, 66).

5 Syntactic analysis

5.1 Attributive function

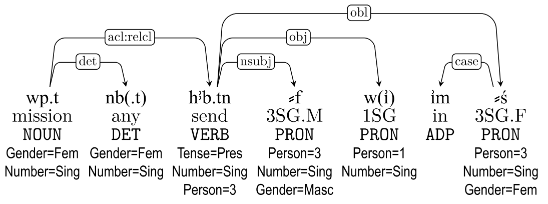
There is a general consensus that adjective relative forms form relative clauses in Earlier Egyptian (Gardiner, 1957, 148/§195, Collier, 1991, 24, Allen, 2010, 353, Schenkel, 2012, 282). As an attributive relative form is used as a relative clause that modifies a nominal antecedent, its dependency relationship is annotated as "acl:relcl" in the UD-EUJA treebank.[12] The UD-EUJA treebank contains 126 relative forms in attributive function, which can be classified into two types according to the syntactic construction:

a) The antecedent corresponds to the object of a relative form, in which no resumptive pronoun is used to refer to the antecedent (Schenkel, 2012, 283, adjective form type II), see examples 2–5 and 6, above.

b) The antecedent corresponds to any other (potential) dependent of the relative form different from its object (Schenkel, 2012, 283, adjective form type III). In this case, the antecedent is referred to by a resumptive pronoun in a non-core (oblique) argument or adjunct governed by the relative form (Polotsky, 1976, 10). The type of argument or adjunct determines the correspondence between antecedent and resumptive pronoun (Polotsky, 1976, 11):

- If the antecedent corresponds to an adverbial adjunct (obl), a resumptive pronoun, attached to a preposition and governed by the relative form, refers to the antecedent:

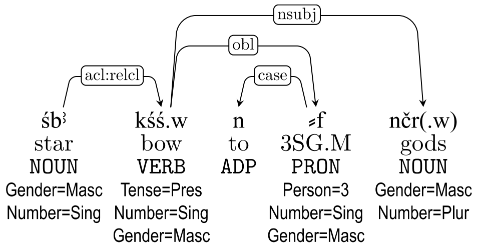
8) Edel 2008 (eds. Seyfried/Vieler): pl. XXXIII–XXXIV, cols. 6–7 (EUJA-33):



'every mission on which he sent me'

- If the antecedent corresponds to an oblique argument with a dative function (obl:arg), a resumptive pronoun, attached to the preposition *n* "to/for" and governed by the relative form, refers to the antecedent:

9) PT 537b T (EUJA-1611):



'a star to whom the gods bow'

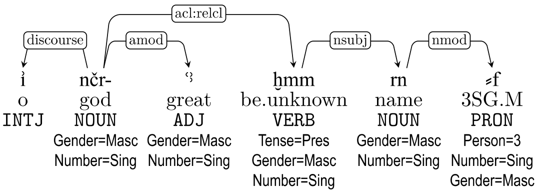
---

[11] See further examples in Schenkel, 2010, 70.

[12] There are other types of Earlier Egyptian relative clauses that are formed without relative forms. However, these are outside the scope of the present paper.

— If the antecedent corresponds to a possessive nominal modifier, a resumptive pronoun referring to the antecedent is attached to a noun governed by the relative form:

10) PT 27a W (EUJA-893):

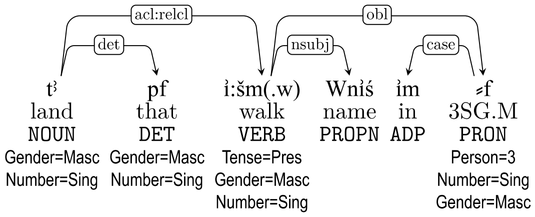

acl:relcl
discourse
amod
nsubj
nmod
ỉ o INTJ
nčr- god NOUN Gender=Masc Number=Sing
ꜥꜣ great ADJ Gender=Masc Number=Sing
ḫmm be.unknown VERB Tense=Pres Gender=Masc Number=Sing
rn name NOUN Gender=Masc Number=Sing
⸗f 3SG.M PRON Person=3 Number=Sing Gender=Masc

‘o great god whose name is unknown’

Lit.: “O great god that his name is unknown (...)”

Intransitive verbs are used with resumptive pronouns when the antecedent corresponds to any other dependent of the relative form different from its object (Gardiner, 1957, 299/§ 384, Polotsky, 1976, 8, type IIBb):

11) PT 382a W (EUJA-1193):


acl:relcl
obl
det
nsubj
case
tꜣ land NOUN Gender=Masc Number=Sing
pf that DET Gender=Masc Number=Sing
ỉ:šm(.w) walk VERB Tense=Pres Gender=Masc Number=Sing
Wnỉś name PROPN
ỉm in ADP
⸗f 3SG.M PRON Person=3 Number=Sing Gender=Masc

‘that land in which Unas walks’

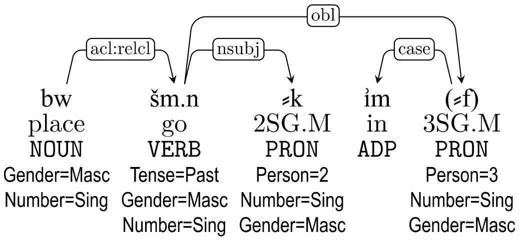
It should be noted that the resumptive pronoun is often omitted after *ỉm* (Gardiner, 1957, 300/§ 385), especially when it is a third person masculine or feminine suffix pronoun, for example:

12) PT 615a–b T (EUJA-1794):

bw	*šm.n*	*⸗k*	*ỉm*
place-M.SG	go:REL.PAST-1-M.SG	=2SG.M	?

“(...) the place where you have gone.”

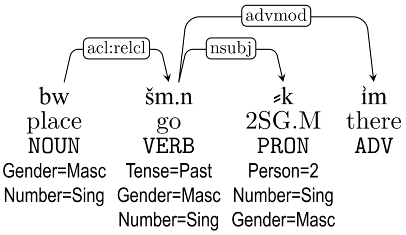
In this case the question arises: Is *ỉm* the preposition *m* “in” in the pronominal state and its suffix pronoun is omitted in the hieroglyphic text, even though it was spoken? If the answer is yes, the suffix pronoun should be added in brackets and linked to the relative form in the tree diagram:


obl
acl:relcl
nsubj
case
bw place NOUN Gender=Masc Number=Sing
šm.n go VERB Tense=Past Gender=Masc Number=Sing
⸗k 2SG.M PRON Person=2 Number=Sing Gender=Masc
ỉm in ADP
(⸗f) 3SG.M PRON Person=3 Number=Sing Gender=Masc

‘the place where you have gone’

Lit.: “(...) the place which you have gone in (it).”

However, if *ỉm* is the adverb *ỉm* “there, thereof” (Allen 2010: 136), then no suffix pronoun should be added and *ỉm* is linked to the relative form:


advmod
acl:relcl
nsubj
bw place NOUN Gender=Masc Number=Sing
šm.n go VERB Tense=Past Gender=Masc Number=Sing
⸗k 2SG.M PRON Person=2 Number=Sing Gender=Masc
ỉm there ADV

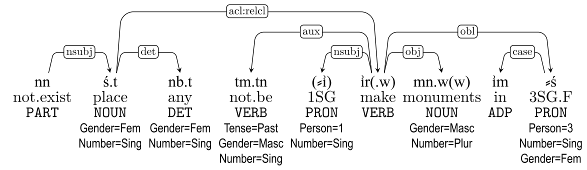
‘the place where you have gone’

Lit.: “(...) the place which you have gone there.”

The second option has been preferred for the annotation of relative forms followed by *ỉm* in the UD-EUJA treebank.

Relative forms are negated in Middle Egyptian by means of the verb *tm* “not be” used as an auxiliary. The following example taken from a Middle Egyptian text has been exceptionally annotated in the UD-EUJA treebank [13] as no instance of *tm* used to negate a relative form is found in Old Egyptian texts (Edel, 1955/64, 340):

13) Gardiner, 1957, 315/§397, (EUJA-2036):


acl:relcl
aux
obl
nsubj
det
nsubj
obj
case
nn not.exist PART
ś.t place NOUN Gender=Fem Number=Sing
nb.t any DET Gender=Fem Number=Sing
tm.tn not.be VERB Tense=Past Gender=Masc Number=Sing
(⸗ỉ) 1SG PRON Person=1 Number=Sing
ỉr(.w) make VERB
mn.w(w) monuments NOUN Gender=Masc Number=Plur
ỉm in ADP
⸗ś 3SG.F PRON Person=3 Number=Sing Gender=Fem

‘There was not any place in which (I) did not make monuments.’

“There was not any place in which (I) did not make monuments.”

As this example shows, *ś.t nb.t* “any place” is the antecedent of the past-2 relative form *tm.tn (⸗ỉ)* “(I) did not (lit.: was not).[14] Since this negative relative form is used as an auxiliary and, following the UD approach, the lexical verb *ỉr(.w)* “make” used as a

---

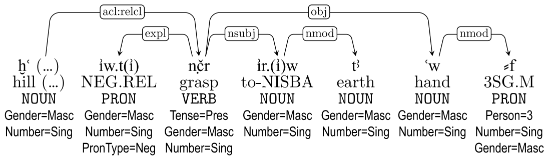
[13] The systematic annotation of Middle Egyptian texts in the UD-EUJA treebank will begin after having finished annotating the Old Egyptian text corpus.

[14] The “negatival complement” is a non-finite adverbial verb form used after negative verbs, see Schenkel, 2012, 272–273.

negatival complement[15] is the head of the relative clause, *ỉr(.w)* is linked to the antecedent in the tree diagram.

In Old Egyptian the negative relative pronoun *ỉw.tỉ*[16] occurs instead of the negative auxiliary *tm* used as a relative form to express a negative relative clause. Although *ỉw.tỉ* agrees in gender and number with its antecedent, its relationship to the verb form of the relative clause is unclear, as its syntactic function differs from that of the core arguments (subject, object and indirect object) and an oblique argument. For this reason, *ỉw.tỉ* is provisionally annotated as an expletive or pleonastic nominal (expl) in the UD-EUJA treebank, e.g.:

14) PT 1022a P (EUJA-2037):


acl:relcl
expl
obj
nsubj
nmod
nmod
ḫꜥ (...) hill (...) NOUN Gender=Masc Number=Sing
ỉw.t(ỉ) NEG.REL PRON Gender=Masc Number=Sing PronType=Neg
nč̣r grasp VERB Tense=Pres Gender=Masc Number=Sing
ỉr.(ỉ)w to-NISBA NOUN Gender=Masc Number=Sing
tꜣ earth NOUN Gender=Masc Number=Sing
ꜥw hand NOUN Gender=Masc Number=Sing
⸗f 3SG.M PRON Person=3 Number=Sing Gender=Masc

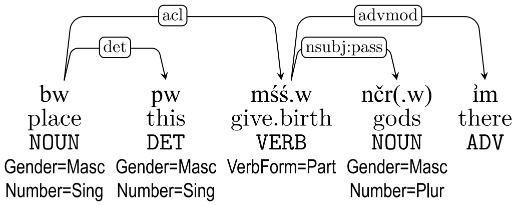
‘(...) hill (...) whose hand no earthlings grasp.’

“(...) The hill (...) whose hand no earthlings grasp (lit.: the hill (...) which no earthlings grasp his hand).”

Here the verb form *nč̣r* “grasp” has its own subject (*ỉr.(ỉ)w tꜣ* “earthlings (lit.: those concerning the earth”)[17] and object (*ꜥw ⸗f* “his hand”), while the negative relative pronoun *ỉw.tỉ* acts as a negative connector between its antecedent (*ḫꜥ* “hill”) and the verb form *nč̣r*.

As for the passive, there are no passive relative forms (Polotsky, 1976, 8) [18] because passive participles are used instead, e.g.:

15) PT 353b T (EUJA-1121):


acl
det
advmod
nsubj:pass
bw place NOUN Gender=Masc Number=Sing
pw this DET Gender=Masc Number=Sing
mśś.w give.birth VERB VerbForm=Part
nčr(.w) gods NOUN Gender=Masc Number=Plur
ỉm there ADV

‘this place where the gods are born’

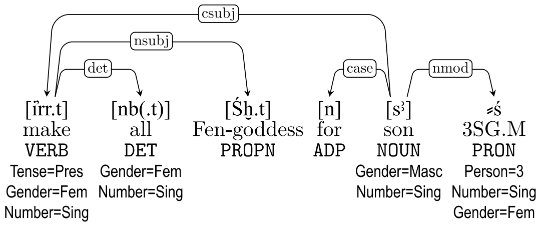
In this sentence *mśś.w* “born” is a present passive participle, *nčr(.w)* “gods” its subject (nsubj:pass) and *ỉm* “there” its oblique argument (obl). The sentence literally means “(...) This place, the gods are born there (...)”.

5.2 Nominal function

As with adjectives and participles, relative forms can syntactically have a nominal function, when they occur in a position where a noun is expected (Sethe, 1899, 328). The hieroglyphic spelling of a relative form features no morphological difference in attributive and nominal functions. There are 61 nominal relative forms in the UD-EUJA treebank.

A nominal relative form can be followed by an adjective, such as *nb* “every” (Sethe, 1899, 328):[19]

16) PT 555d T (EUJA-1652):


csubj
nsubj
det
case
nmod
[ỉrr.t] make VERB Tense=Pres Gender=Fem Number=Sing
[nb(.t)] all DET Gender=Fem Number=Sing
[Śḫ.t] Fen-goddess PROPN
[n] for ADP
[sꜣ] son NOUN Gender=Masc Number=Sing
⸗ś 3SG.M PRON Person=3 Number=Sing Gender=Fem

‘[Whatever the Fen-goddess makes is for] her [son] (...)’

Lit.: “[All what the Fen-goddess makes is for] her [son] (...)”

Although Sethe (1899, 327–329) identified some functions of nominal relative forms, such as subject, predicate (i.e. root), object and nomen

---

[15] The “negatival complement” is a non-finite adverbial verb form used after negative verbs, see Schenkel, 2012, 272–273.

[16] Although *ỉw.tỉ* is originally a nisba adjective (Gardiner, 1957, 152/§ 202), it functions as a relative pronoun (Schenkel, 2012, 342). For the term “nisba” see footnote 17.

[17] *ỉr.(ỉ)w* “those concerning” is a nisba derived from the preposition *r* “concerning”. 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, for example the nisba adjective *n.ỉ* “belonging to” derives from the preposition *n* “for”.

[18] Relative forms with *tw* (Old Egyptian *ti)* are used to express a “man-impersonal” active corresponding to “one” in English, see Borghouts, 2010, 283 (79d).

[19] Square brackets ([]) indicate that the text within is restored according to a preserved witness, i.e. a textual variant.

rectus in a “direct genitive”,[20] there are no further detailed studies of the syntactic use and dependency relation of nominal relative forms in Egyptological literature. This section partly fills such a gap by extending the functions of relative forms and offering a first morphosyntactic analysis thereof, which will hopefully lay the foundation for future works.

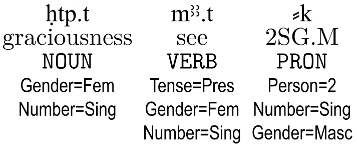
It should be noted that nominal relative forms are considered to be clauses rather than nouns because they consist of a verb that governs a subject, forming a clause. From a theoretical perspective, the annotation of a relative form as a noun is problematic, since it would perform the function of another noun in a sentence, for example considering *mꜣꜣ.t ⸗k* to be the subject (nsubj) of *ḥtp.t* in example 17 (see below) implies that the sentence has two subjects because *⸗k* is the subject (nsubj) of *mꜣꜣ.t*, which would break the syntactic rule that state that a child of an nsubj relation should not be the parent of another nsubj relation.

Following nominal functions of relative forms are annotated in the current state of the UD-EUJA treebank: clausal subject (csubj), clausal complement (ccomp), adverbial clause (advcl), adnominal clause (acl), dislocated element, vocative and root.

5.2.1 Clausal subject

The root of a nominal sentence can govern a nominal relative form used as a topic and clausal subject. The relative form retains its own subject and it can be placed after or before the root. There are 10 relative forms used as clausal subjects in the UD-EUJA treebank, e.g.:

17) PT 34c W (EUJA-162):



‘What you see is graciousness.’

18) PT 192b W (EUJA-612):

wnm.tn	*⸗k*	*ỉr.t*
eat :REL.PAST-2-F.SG	=2SG.M	eye-F.SG

“What you ate is the eye.”

Likewise, a nominal relative form appears as a clausal subject in an adverbial sentence[21] where the root is a noun after a preposition:

19) PT 417a W (EUJA-1300):

ỉr.(y)t	*⸗k*	*ỉr*	*⸗k*
do:REL.FUT-F.SG	=2SG.M	against:PREP	=2SG.M

“What you will do is against you.”

A nominal relative form can also be used as the clausal subject of a verb form, for example:

20) PT 275a W (EUJA-883):

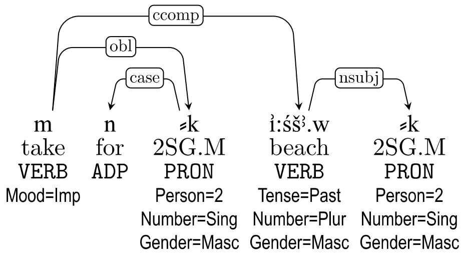
wꜥb.n	*wꜥb.n*	*(⸗ỉ)*
be purified: NMLZ-PAST	purify:REL.PAST-2-M.SG	=1SG

“‘The one (I) purified is purified (...)’”

5.3 Clausal complement

A nominal relative form can adopt the direct object function like a noun. In this case, it is used as a clausal complement of a verb. There are 10 relative forms used as clausal complements in the UD-EUJA treebank, e.g.:

21) PT 82a W (EUJA-299):



‘(...) take (lit.: for yourself) those you beached (...)’

22) PT 624b T (EUJA-1817):

gm	*⸗k*	*ḫśf*	*⸗k*	*ỉm*
find:SBJV	=2SG.M	meet:REL.FUT-M.SG	=2SG.M	there:ADV

“(...) You may find the one you will meet there.”

23) PT 625a T (EUJA-1819):

ỉr	*⸗k*	*ỉrr.t*	*Wśr(.w)*
do:SBJV	=2SG.M	do:REL.PRES-F.SG	Osiris-GN

“You should do what Osiris does.”

5.4 Adverbial clause

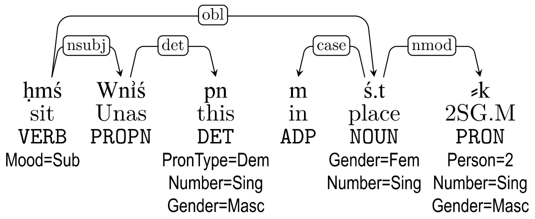
When a nominal relative form takes the place of an oblique nominal consisting of a noun after a preposition, the dependency relationship on the head changes from an oblique nominal to an adverbial clause, cf.:

---

[20] See a similar list in Ockinga, 2005, 69.

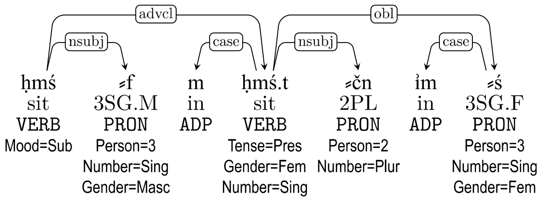
[21] An adverbial sentence is a nonverbal sentence with an adverbial predicate.

24) PT 367b W (EUJA-1155), an oblique nominal (preposition *m* + a noun):


obl
nsubj
det
case
nmod
ḥmś sit VERB Mood=Sub
Wnỉś Unas PROPN
pn this DET PronType=Dem Number=Sing Gender=Masc
m in ADP
ś.t place NOUN Gender=Fem Number=Sing
⸗k 2SG.M PRON Person=2 Number=Sing Gender=Masc

‘(...) this Unas may sit in your place.’

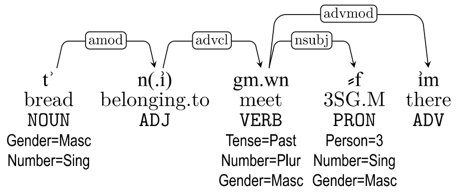
25) PT 129b W (EUJA-422), adverbial clause with a nominal relative form:


advcl
nsubj
case
obl
nsubj
case
ḥmś sit VERB Mood=Sub
⸗f 3SG.M PRON Person=3 Number=Sing Gender=Masc
m in ADP
ḥmś.t sit VERB Tense=Pres Gender=Fem Number=Sing
⸗čn 2PL PRON Person=2 Number=Plur
ỉm in ADP
⸗ś 3SG.F PRON Person=3 Number=Sing Gender=Fem

‘(...) he may sit on what you sit on (...)’

Likewise, the dependency relationship of a relative form is that of an adverbial clause when it is used as a noun modifying the nisba adjective *n.ỉ* in an “indirect genitive construction”:

26) PT 132d W (EUJA-451):


advmod
amod
advcl
nsubj
t’ bread NOUN Gender=Masc Number=Sing
n(.ỉ) belonging.to ADJ
gm.wn meet VERB Tense=Past Number=Plur Gender=Masc
⸗f 3SG.M PRON Person=3 Number=Sing Gender=Masc
ỉm there ADV

‘(...) the bread of (lit.: belonging to) those he found there (...)’

“(...) The bread of (lit.: belonging to) those he found there (...)”

In this structure the relative form replaces a noun used as an oblique nominal of an adjective, cf.:

27) PT 242a-b W (EUJA-785):

t’	*n(.ỉ)*	*ỉt(ỉ)*	*⸗k*
bread-M.SG	belonging to-NISBA	father-M.SG	=2SG.M

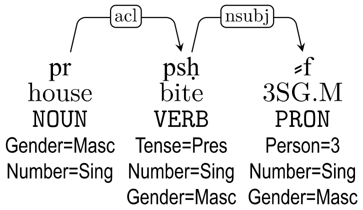
“(...) The bread of (lit.: belonging to) those he found there (...)”

There are 17 relative forms used as adverbial clauses in the UD-EUJA treebank.

5.5 Adnominal clause

The current state of the UD-EUJA treebank contains an instance of a nominal relative form used as an adnominal clause in a “direct genitive”:[22]

28) PT 247b W (EUJA-798):


acl
nsubj
pr house NOUN Gender=Masc Number=Sing
psḥ bite VERB Tense=Pres Number=Sing Gender=Masc
⸗f 3SG.M PRON Person=3 Number=Sing Gender=Masc

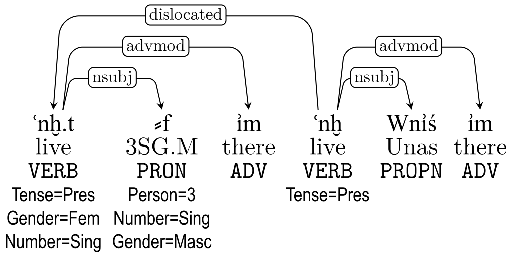
‘(...) the house of the one it should bite (...)’

The nominal relative form *psḥ ⸗f* is used as a clausal modifier (acl) of the noun *pr* “house” in a “direct genitive”.[23]

5.6 Dislocated element

A relative form can be used as a dislocated noun placed in the periphery of the root. There are 10 relative forms used as dislocated elements in the UD-EUJA treebank, e.g.:

29) PT 133d W (EUJA-455):


dislocated
advmod
nsubj
advmod
nsubj
ꜥnḫ.t live VERB Tense=Pres Gender=Fem Number=Sing
⸗f 3SG.M PRON Person=3 Number=Sing Gender=Masc
ỉm there ADV
ꜥnḫ live VERB Tense=Pres
Wnỉś Unas PROPN
ỉm there ADV

‘What he lives on, Unas lives on.’

A resumptive pronoun used as a subject or an object may refer to the dislocated nominal relative form, cf.:

30) PT 682d T (EUJA-1976), a resumptive pronoun (*⸗f)* used as a subject refers to a dislocated nominal relative form (*mꜣꜣ*):

mꜣꜣ	*Ttỉ*	*nỉ*	*ꜥnḫ*	*⸗f*
see:REL.PRES-M.SG	Teti-KN	NEG	live:FUT	=3SG.M

“He whom Teti sees, he will not live.”

---

[22] A “direct genitive” is a possessive construction consisting of two nouns—the first is the *nomen regens* and the second is the *nomen rectum*.

[23] Specifically, a relative form is used as a possessive adnominal clause when modifying a noun in a direct genitive. This syntactic dependency should be annotated as acl:poss.

31) PT 407c T (EUJA-1271), a resumptive pronoun used as an object (*św)* refers to a dislocated nominal relative form:

gm.y	*⸗f*	*m*	*wꜣ.t*	*⸗f*
find:REL.FUT-M.SG	=3SG.M	in:PREP	way-F.SG	=3SG.M

wnm	*⸗f*	*n*	*⸗f*	*św*
devour:FUT	=3SG.M	for:PREP	=3SG.M	3SG.M

"Whoever he will find in his way, he will devour him (lit. for himself)."

5.7 Vocative

When a dislocated relative form is used as a noun addressing a dialogue participant, it acts as a vocative. There is an instance of a nominal relative form used as a vocative in the UD-EUJA treebank:

32) PT 428a–b W (EUJA-1321):[24]

nnỉ	*mw.t*	*⸗f (...)*
turn away:REL.PRES-M.SG	mother-F.SG	=3SG.M

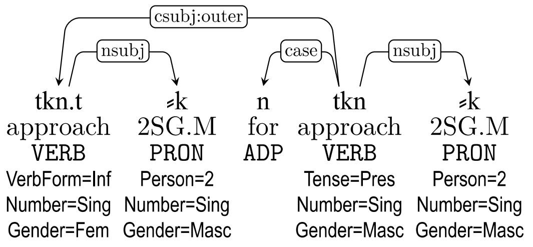
ỉ(w)	*⸗k*	*rr*	*m*	*nn*
PART	=2SG.M	PART	like:PREP	DEM:M.SG

"(O) the one whose mother turned (him) away (...), you are really like this."

5.8 Root

A relative form can be used instead of a noun as a root in an adverbial sentence, e.g.:

33) PT 433b W (EUJA-1334):



'Your approach is for the one you approach.'

The adverbial predicate of this sentence consists of the preposition *n* "for" followed by the nominal relative form *tkn ⸗k* "the one you approach".

6 Conclusion

The annotation of Old Egyptian relative forms in the UD-EUJA treebank was challenging due to their morphological particularities and dual syntactic function. There are four types of relative forms according to their tense—past-1, past-2, present and future relative forms. The dual syntactic function of relative forms has also been taken into account when annotating them in the UD-EUJA treebank. On the one hand, attributive relative forms are defined as "acl:relcl" because they form relative clauses that modify an antecedent. On the other hand, nominal relative forms are annotated as nominal clauses with different syntactic functions.

The morphosyntactic analysis of the Old Egyptian relative forms has demonstrated the effectiveness and accuracy of the UD-EUJA treebank in finding all instances of a given grammatical feature or structure. It contributes to the analysis of morphological particularities, such as the occasional masculine ending *w* in the present relative form (see Table 1, Appendix) or the use of the *ỉ* prefix in relative forms (see Table 3, Appendix). When the annotation of the Pyramid Texts and the historical-biographical texts of the Old Kingdom is completed, it will be possible to carry out a thorough analysis of each Old Egyptian morphosyntactic feature, including the relative forms.

The development of the UD-EUJA treebank will enrich general linguistics, as Egyptian is characterised by idiosyncratic features. A good example of this is the use of relative forms, which are rarely found in other languages, notably Indo-European languages.

Future studies using the UD-EUJA treebank will focus on the dual function of adjective features in Old Egyptian, such as the participles that allow attributive and nominal functions as do relative forms.

Acknowledgments

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

---

[24] For this example, see Allen, 2017, 188 and 345.

References

James P. Allen. 2010. *Middle Egyptian. An Introduction to the Language and Culture of Hieroglyphs*. Cambridge.

James P. Allen. 2017. *A Grammar of the Ancient Egyptian Pyramid Texts. Volume I: Unis*. Eisenbrauns, Winona Lake, Indiana.

James P. Allen. 2013. *A New Concordance of the Pyramid Texts*. 6 volumes. Brown University.

Joris F. Borghouts. 2010. *Egyptian. An Introduction to the Writing and Language of the Middle Kingdom*. Peeters, Leuven.

Jacques J. Clère. 1949. Une nouvelle forme verbale relative en égyptien. In *Actes du XXIe Congrès international des Orientalistes*, Paris: 64–66.

Mark Collier. 1991. *The Relative Clause and the Verb in Middle Egyptian*. The Journal of Egyptian Archaeology 77:23–42.

Roberto Antonio 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/.

Elmar Edel. 1955/1964. *Altägyptische Grammatik*. Roma.

Elmar Edel (K. Seyfried/G. Vieler, eds.) 2008. *Die Felsgräbernekropole der Qubbet el-Hawa bei Assuan*, Paderborn, Munich, Vienna, Zürich.

Adolf Erman. 1894. *Ägyptische Grammatik*. Verlag von Reuther & Reichard. Berlin.

Alan H. Gardiner. 1957^3. *Egyptian Grammar. Being an Introduction to the Study of Hieroglyphs*. Griffith Institute, Oxford.

Battiscombe Gunn. 1924. *Studies in Egyptian Syntax*. Paris.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre and Daniel Zeman. 2021. Universal Dependencies. Association for Computational Linguistics 47(2):255–308. https://doi.org/10.1162/COLI_a_00402.

Antonio Loprieno. 1995 *Ancient Egyptian. A Linguistic Introduction*. Cambridge University Press.

Boyo G. Ockinga. 2005. *Mittelägyptische Grundgrammatik*. Verlag Philipp von Zabern. Mainz am Rhein.

Hans J. Polotsky. 1976. *Les Transpositions du Verbe en Égyptien Classique*. Israel Oriental Studies 6:1–50.

Wolfgang Schenkel. 2010. *Die Clèresche Relativform*. Zeitschrift für ägyptische Sprache und Altertumskunde 137: 66–90.

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. 1899. *Das Aegyptische Verbum im Altaegyptischen, Neuaegyptischen und Koptischen. Zweiter Band. Formenlehre und Syntax der Verbalformen*. Hinrich'sche Buchhandlung, Leipzig.

Kurt Sethe. 1908–1922. *Die altägyptischen Pyramidentexte nach den Papierabdrücken und Photographien des Berliner Museums*. 4 volumes. Hinrich'sche Buchhandlung, Leipzig.

Henri Sottas. 1925. *B. Gunn.—Studies in egyptian syntax. (Review)*. Bulletin de la Société de linguistique de Paris 6:263–265.

Andréas Stauder. 2014. *Interpreting Written Morphology: the sḏm.n⸗f in the Pyramid Texts*. Journal of Near Eastern Studies 73(2):253–271.

Sami Uljas. 2023. The So-Called Old Egyptian Prothetic ỉ—Part III: The Participles. In M. Victoria Almansa-Villatoro (*et al.*). In the House of Heqanakht. Brill, Leiden/Boston:387–399.

A Appendix: Tables

Lemma	RelForm	EUJA	PT
Tense=Present			
wꜣḥ "lay down"	*wꜣḥ.w* + SF	1994	685d T
wꜥb "be pure"	*ś:wꜥb.{i̯}w* + SF	1406	457b W
pri̯ "go out"	*prr.w* + SF	1506	496b W
mri̯ "love"	*mrr.w* + SF	83 1771	15 N 607d T
nwr "shake"	*nwr.w*	495	143a W
nwḥ "rope"	*nwḥ.w*	2077	719d T
nč̣ "protect"	*nč̣.w*	700	215c W
hꜣb "send"	*hꜣb.w* + SF	1256	402c W
ḥmśi̯ "sit"	*ḥmś.w*	2077	719d T
śẖꜣꜣ "row"	*ỉ:śšꜣ.w* + SF	299	82a W
śṭꜣ "tremble"	*śṭꜣ.w*	495 1611	143a W 537b T
šmi̯ "go"	*ỉ:šm.w*	1194	382a P
šsp "accept"	*šsp.w* + SF	1166	372e W
kśi̯ "bend down"	*kśś.w*	1611	537b T
wṭi̯ "put"	*ṭṭ.w* + SF	1993	685c T
č̣ṭ "say"	*č̣ṭ.w*	824	254c W
č̣ṭ "say"	*ỉ:č̣ṭ.w*	1932	665a T

Table 1. PT = Pyramid Texts; SF = Suffix pronoun; W = Unas; T = Teti; N = Pepi II.

Lemma	RelForm	EUJA	PT
Tense=Past-2			
wṭi̯ “put”	*wṭ.wn* + SF	1428	464b W
gmi̯ “find”	*gm.wn* + SF	451	132d W
čꜣs “tie”	*č(ꜣ)s.wn*	768	234c W

Table 2. EUJA = sentence id; PT = Pyramid Texts; SF = Suffix pronoun; W = Unas.

Lemma	RelForm	EUJA	PT
Tense=Past-1 (?)			
nśbi̯ “lick”	*i̓:nśb.t*[25] + SF	347	104b N
Tense=Past-2			
fkk “devastate”	*i̓:ś:fkk.tn* + SF	1551	51a W
nśbi̯ “lick”	*i̓:nśb.tn* + SF	328	98c W
nśbi̯ “lick”	*i̓:nśb.n* + SF	745	228b W
ḫḫmi̯ (unclear meaning)	*i̓:ḫḫm.tn* + SF	260 1559 2165	61b W 89c W 89c T
Tense=Pres			
śẖꜣꜣ “row”	*i̓:śšꜣ.w + SF*	299	82a W
šmi̯ “go”	*i̓:šm + SF*	1932	665c T
šmi̯ “go”	*i̓:šm.w*	1194	382a P
č̣ṭ “say”	*i̓:č̣ṭ* + SF	627	195b W
č̣ṭ “say”	*i̓:č̣ṭ.t* + SF	1491	491d W
č̣ṭ “say”	*i̓:č̣ṭ.w*	1932	665a T

Table 3. EUJA = sentence id; PT = Pyramid Texts; SF = Suffix pronoun; W = Unas; T = Teti; P = Pepi I; N = Pepi I

---

[25] *ʾI:nśb.t<n>* (?).

UD Treebanks for Esperanto as a Natural Language

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

Abstract

This paper describes the details of UD-based morphological and syntactic annotations on Esperanto texts to construct its small-scale UD treebank. Though it was created as an international auxiliary language, Esperanto has increasingly been studied as a natural language both in linguistics and in NLP. This paper introduces the detail of the manual annotation of UD morphological and relational tags and describes how the frequencies of these tags differ across the treebanks and discusses the possibility of future research of Esperanto as a natural language.

1 Introduction

Esperanto is a constructed language created by L.L. Zamenhof in 1887 for the purpose of international communication. Since its creation, it has been spreading across the globe, attracting many language enthusiasts and creating a variety of communities using it as a means of communication. Though it has been criticized for several reasons, it has now developed into one of the natural languages which is worthy of linguistic research and natural language processing (NLP). In this context surrounding Esperanto, this paper introduces small-scale UD treebank for Esperanto. This is the first contribution of Esperanto to UD community (if not the first attempt to construct an Esperanto treebank (Bick 2018)), and it is expected to facilitate not only NLP research of Esperanto but linguistic research in general.

This paper is organized as follows: Section 2 is a summary of previous studies which treat Esperanto not as an artificial language but as one of the natural languages, and those which approach it from the viewpoint of NLP. Section 3 summarizes the process of annotating Esperanto texts with UD POS tags and UD relation tags, along with a brief description of its morphological and syntactic features. Section 4 is a discussion on the issue of how UD POS tags and UD relation tags are used in different genres of Esperanto texts, for the further development of future research on morphological and syntactic properties of Esperanto as a natural language.

2 Previous studies of Esperanto

2.1 Esperanto in linguistics

Goodall (2023) summarizes the history of the relationship between Esperanto and linguistics. He points out decades of hostilities between them due to the overemphasis on the role of native speakers' intuition on naturalness of language. He claims that Esperanto today can be one of the topics of ordinary linguistic research. Moreover, even though Esperanto was constructed as an international auxiliary language, there are researchers who consider it as one of the natural languages and investigate its linguistic characteristics which are similar with those of other natural languages. In Gledhill (2000), which is a comprehensive grammar of Esperanto, he pointed out that it has developed as a natural language, going beyond Zamenhof's original design. Manaris, Pellicoro, Pothering and Hodges (2006) applied Zipf-based metrics (word distribution, word-distance distribution, word-bigram distribution, etc.) to six languages including Esperanto, and found similar statistical proportions between Esperanto and the other five languages. Parkvall (2010)'s typological study based on the World Atlas of Language Structures (Haspelmath 2005, henceforth WALS) revealed that Esperanto has certain features characteristic to European languages, yet it is not a typical one. Stria (2015) assessed Esperanto as a natural language in terms of a variety of criteria. Koutny (2015) conducted a

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 22–29
August 27, 2025 ©2025 Association for Computational Linguistics

typological description of Esperanto on WALS database and concluded that Esperanto has characteristics which make it accessible for people with different linguistic backgrounds.

As these studies indicate, Esperanto in the 21st century can be considered as a natural language which deserves linguistic investigation just like all the other natural languages, even though its status as an international auxiliary language may not be as realistic as it was intended to be when it was created in the 19th century.

2.2 Esperanto in NLP research

Along with the trend where Esperanto has been one of the topics of linguistic research, several researchers have focused on Esperanto from the point of view of natural language processing (NLP). Minnaja and Paccagnella (2000) proposed a PoS tagger for Esperanto. Bick (2007, 2009) introduced *EspGram*, a Constraint-Grammar-based parser for Esperanto. Bick (2016) dealt with the internal structure of complex Esperanto words (CWs) and reported the construction of a CW dictionary. Bick (2019) presented a method for the automatic generation and semantic evaluation of Esperanto sentences for pedagogical purposes. Poncelas, Buts, Hadley and Way (2020) focused on Esperanto for improvements of the performance of a machine translation system for low-resource languages. Bick (2018, 2020) introduced *Arbobanko*, the first treebank for Esperanto. Arbobanko is a dependency treebank with 52,000 tokens. The texts were taken from *Monato,* an Esperanto news magazine, randomly selected from the year 2000-2010. Words in Arbobanko are annotated with tags for several features, such as lemma, part-of-speech, inflection, syntactic function, dependency links, and verb frames categories. Arbobanko also provides CoNLL-format version, yet it is not annotated with UD Pos tags and UD relation tags.

Since the POS and relation tags in Arbobanko are created only for Esperanto, it cannot be employed for multi-lingual comparison and contrast straightforwardly. We need multi-lingual treebank for that purpose, in which different languages are annotated with the same POS and relation tags, in order to develop Esperanto NLP and linguistic investigations.

3 This work: UD Treebank for Esperanto

3.1 Data

In the context mentioned in the previous section, this study attempts to construct a UD treebank for Esperanto. The first text chosen for this study is *Manifesto de Prago* (Prague Manifesto), drafted at the World Esperanto Congress in Prague in 1996. This manifesto promotes more democratic communication, language rights not only for Esperanto but for all languages on this planet, importance of language diversity and its preservation, and language education facilitated by Esperanto. The original text is available online (https://uea.org/teko/praga_manifesto/pm_esperanto). For this study, the text was annotated according to the UD Guidelines manually by the author of this study; automatic parsing was not conducted because the parsers for Esperanto available at present do not yield parse output in the format of CoNLL-U, which means that we need to convert the output into CoNLL-U, manually or automatically, and this is not the main topic of this study (yet this will be a topic of future study, naturally). The annotated text, *eo_prago-ud-text.conllu* (henceforth *Prago*), contains 1023 lines.

The second text chosen for this study is the Cairo CICLing sentences (https://github.com/UniversalDependencies/cairo) translated from the original English into Esperanto, following the UD guideline for a new treebank (https://universaldependencies.org/release_checklist.html). The annotated text, *eo_cairo-ud-text.conllu* (henceforth *Cairo_eo*), contains 257 lines. Table 1 summarizes the number of tokens, types, and type-token ratios of these texts. Though Prago is more than four times larger than Cairo in terms of the number of tokens, their type-token ratios do not differ substantially.

	Tokens	Types	T/T Ratio
Prago	839	321	.383
Cairo_eo	256	103	.402

Table 1: The numbers of types, tokens, and type-token ratio of *eo_prago_ud-test.conllu* (indicated as Prago) and *eo_cairo_ud-text.conllu* (indicated as Cairo_eo).

	Prago		*Cairo_eo*		*Cairo_en*	
	freq.	%	freq.	%	freq.	%
ADJ	108	12.87	4	2.26	5	2.67
ADP	124	14.78	5	2.82	6	3.21
ADV	35	4.17	13	7.34	7	3.74
AUX	18	2.15	6	3.39	12	6.42
CCONJ	32	3.81	8	4.52	10	5.35
DET	75	8.94	10	5.65	20	10.70
NOUN	196	23.36	23	12.99	24	12.83
NUM	10	1.19	1	0.56	0	0.00
PART	4	0.48	3	1.69	6	3.21
PRON	44	5.24	24	13.56	22	11.76
PROPN	8	0.95	14	7.91	15	8.02
PUNCT	104	12.40	28	15.82	25	13.37
SCONJ	12	1.43	5	2.82	3	1.60
VERB	69	8.22	33	18.64	32	17.11
	839		177		187	

Table 2: The frequencies and the percentages of the UPOS tags in *eo_prago_ud-test.conllu* (indicated as *Prago*), *eo_cairo_ud-text.conllu* (indicated as *Cairo_eo*), and *cairo/en.conllu* (indicated as *Cairo_en*)

3.2 UPOS and morphological annotation of Esperanto

This section summarizes the part-of-speech and morphological features of Esperanto along with brief explanations on how they are annotated in the Esperanto text.

According to Zamenhof (1887), an international auxiliary language should be easy to learn for anybody, and therefore the grammar of Esperanto is characterized by a high degree of regularity. As a result of this, it contains only a few exceptions (of course, the fact that there are only a small number of exceptions in a language does not necessarily mean that the language is simple and easy to learn).

Zamenhof's idea of regularity (if not simplicity) is reflected in the fact that the word-ending derivational morpheme of Esperanto indicates the part of speech of the majority of Esperanto words (Zamenhof 1887). Nouns end with *-o* (when nominative singular), infinitive verbs with *-i*, adjectives with *-a*, and adverbs with *-e* (Zamenhof 1887). For example, with the root *parol-*, *parolo* means “speaking”, *paroli* means “to speak”, *parola* means “oral” or “spoken”, and *parole* means “verbally.”

Esperanto uses all the UPOS tags (universal part-of-speech tags), and no XPOS tags (optional language-specific part-of-speech/morphological tags). For the details of UPOS and XPOS tags, refer to the Web page of CoNNL-U format (https://universaldependencies.org/format.html). The table below summarizes the frequencies and the percentages of the UPOS tags in *Prago* and *Cairo_eo*, and those in Cairo/en.conllu, for comparison:

The most obvious differences between Prago and Cairo are the percentage of NOUNs, PRONs, and VERBs. The percentage of ADV in Cairo is almost twice as high as that in Cairo_en, and that of DET in Cairo_en is almost twice as high as that in Cairo.

Table 3 summarizes the morphological features used in Esperanto, categorized according to UPOS tags:

UPOS	Features
ADJ	Case=Acc, Nom; Degree=Pos; Number=Plur, Sing

AUX	Mood=Imp, Ind, Sub; Tense=Fut, Past, Pres;
	VerbForm=Fin, Inf; Voice=Act, Pass

DET	Definite=Def; PronType=Art, Dem, Tot
NOUN	Case=Acc, Nom; Number=Plur, Sing
NUM	NumForm=Word; NumType=Card

PRON	Case=Acc, Nom; Number=Plur, Sing;
	Number[psor]=Plur, Sing;
	Person=1, 2, 3; Poss=Yes;
	PronType=Dem, Int, Prs, Rel, Tot; Reflex=Yes

PROPN	Case=Acc, Nom; Number=Plur, Sing

VERB	Mood=Imp, Ind, Sub; Tense=Fut, Past, Pres;
	VerbForm=Fin, Inf; Voice=Act, Pass

Table 3: The morphological features of Esperanto

Suffixes on nouns and adjectives indicate their case and number, which are summarized in Table 4 and 5:

	Singular	Plural
Nominative	*-o*	*-oj*
Accusative	*-on*	*-ojn*

Table 4: Esperanto nominal suffixes (based on Zamenhof 1887)

	Singular	Plural
Nominative	*-a*	*-aj*
Accusative	*-an*	*-ajn*

Table 5: Esperanto adjectival suffixes (based on Zamenhof 1887)

For example, in an Esperanto phrase *interesaj artikoloj* “interesting articles”, the noun *artikoloj* is annotated with “Case=Nom|Number=Plur,” and the adjective *interesaj* is annotated with “Case=Nom|Degree=Pos|Number=Plur.”

DET is annotated on the definite article *la* (Esperanto does not have indefinite articles), and some correlatives when preceding a noun, such as *tiu artikolo* “that book”, *tiuj artikoloj* “those books”, or *kies artikolo* “whose book”. The definite article *la* is annotated with “Definite=Def|PronType=Art.”

Esperanto pronouns are annotated with POS tag PRON, and with morphology tags, according to their number, person, gender (for 3rd person), and case. For example, the morphology tag on the 3rd-person masculine pronoun *li* is annotated with “Case=Nom|Gender=Masc|Number=Sing|Person=3rd|PronType=Prs.”

Some correlatives used not as determiners but as demonstrative pronouns are annotated with POS tag PRON. For example, the correlative *tiuj* in *Tiuj estas interesaj* “Those are interesting” functions as a pronoun, and it is annotated with “Case=Nom|Number=Plur|PronType=Dem.”

Esperanto verbs do not agree with their subject, and they inflect according to their mood and tense, summarized in the table below:

	suffixes
Infinitive	*-i*
Present	*-as*
Past	*-is*
Future	*-os*
Subjunctive	*-us*
Imperative	*-u*

Table 6: Inflectional morphemes of Esperanto verbs (based on Zamenhof 1887)

For example, the verb *verkis* in a sentence *Mi verkis tiujn artikolojn* “I wrote those articles” is annotated with “Mood=Ind|Tense=Pres|VerbForm=Fin.”

Derivational morphemes for participles and gerunds are summarized in the table below:

	Past	Present	Future
Active	*-int-*	*-ant-*	*-ont-*
Passive	*-it-*	*-at-*	*-ot-*

Table 7: Derivational morphemes of Esperanto verbs (based on Zamenhof 1887)

These morphemes are followed by another morpheme for part of speech (nominal *-o*, adjectival *-a*, and adverbial *-e*), and for plural *-j*.

Gerunds and participles are annotated with not only those tags related to verbs but also those related either to nouns or to adjectives. First, gerunds are accompanied by nominal morphemes indicating their case and number, so they are annotated with those tags related to nouns and verbs. For example, the gerund *lernantoj* “learners” is derived from a verb root *lern-*, with the active present *-ant-*, the nominal morpheme *-o*, and the plural morpheme *-j*. Therefore, it is annotated with “Case=Nom|Number=Plur|Tense=Pres|VerbForm =Part|Voice=Act”.

Adjectival participles agree with the noun they modify when used as modifiers or with the subject noun when used as a predicate with the copula *esti*, so they are annotated with those tags related to adjectives and verbs. For example, in *La lernantoj estas legantaj la librojn* “The learners are reading the books”, the present active plural adjectival participle *legantaj* is annotated with “Case=Nom|Number=Plur|Tense=Pres|Verbform=Part|Voice=Act”.

Adverbial participles do not agree with case and number, so they are annotated with those tags related to verbs. For example, in *Legante la libron, la lernantob manĝas* “Reading the book, the learner is eating”, the adverbial participle *Legante* is annotated with “Tense=Pres|VerbForm=Part|Voice=Act”.

The fact that gerunds in Esperanto have often been lexicalized (e.g. *Esperanto* is originally a present active gerund derived from the verb root *esper-* “to hope” plus present active *-ant-* and nominal *-o*, meaning “one who hopes”) while participles are not always lexicalized leads us to categorize gerunds as nouns while participles as verbs, and POS annotation follows this principle.

3.3 Syntactic annotation

Esperanto allows free word-order (e.g., a corpus investigation by Gledhill (2000)). The case morphemes on nouns and adjectives indicate the grammatical relationship between verbs and their direct objects, so changing the order of the verb and its subject and object in a sentence does not change their grammatical relationships. For example, consider an example sentence *Mi verkis tiujn artikolojn* “I wrote those articles”. The object can precede the subject and the verb, such that *Tiujn artikolojn mi verkis*, without changing the grammatical relationship between the verb and its object noun phrase.

Kråkmo (2022) conducted a survey which showed that Esperanto prefers the SVO word order, though this fact does not mean that other word orders such as SOV are ungrammatical in Esperanto.

The gloss and the UD syntactic annotation on an example sentence *Mi verkis tiujn artikolojn, kiuj ne altiris la atenton de aliaj esploristoj.* "I wrote those articles which did not attract other researchers' attention." are shown below:

(1) *Mi verk-is tiu-j-n artikolo-j-n, kiu-j ne altir-is la atento-n de alia-j esploristo-j.*
I write-*pst* those-*pl-acc* article-*pl-acc,* which-*pl* not attract-*pst* the attention-*acc* of other-*pl* researcher-*pl*

```
# sent_id = 1
# text = Mi verkis tiujn artikolojn, kiuj altiris la atenton de aliaj esploristoj.
# text_en = I wrote those articles which did not attract other researchers' attention.
1    Mi       mi       PRON   _    Case=Nom|Number=Sing|Person=1|PronType=Prs 2 nsubj  _   _
2    verkis   verki    VERB   _    Mood=Ind|Tense=Past|VerbForm=Fin   0       root     _   _
3    tiujn    tiu      DET    _    Case=Acc|Number=Plur|PronType=Dem          4       det  _  _
4    artikolojn artikolo NOUN _    Case=Acc|Number=Plur     2        obj      _   _
5    ,        ,        PUNCT  _    _        8        punct    _   _
6    kiuj     kiu      PRON   _    Case=Nom|Number=Plur|PronType=Rel 8        nsubj    _   _
7    ne       ne       ADV    _    _        8        advmod   _   _
8    altiris  altiri   VERB   _    Mood=Ind|Tense=Past|VerbForm=Fin   4       acl:relcl
9    la       la       DET    _    _        10       det      _   _
10   atenton  atento   NOUN   _    Case=Acc|Number=Sing     8        obj      _   _
11   de       de       ADP    _    _        13       case     _   _
12   aliaj    alia     ADJ    _    Case=Nom|Number=Plur|Degree=Pos   13      amod     _   _
13   esploristoj       esploristo NOUN  _     Case=Nom|Number=Plur    10      nmod     _   _
14   .        .        PUNCT  _    _        2        punct    _   _
```

Figure 1: The UD annotation on an example sentence *Mi verkis tiujn artikolojn, kiuj altiris la atenton de aliaj esploristoj.* "I wrote those articles which did not attract other researchers' attention."

In the example above, the pronoun *Mi* depends on the verb *verkis* as its subject. The determiner *tiujn* agrees with the noun *artikolojn* in terms of case (accusative) and number (plural). The relative pronoun *kiuj* refers to *artikolojn*, and it depends on the verb *altiris* as its subject. The negative adverb *ne* depends on the verb *altiris,* which depends on the noun *artikolojn* as a relative clause. This dependency is typed as *acl:relcl*. The noun *atenton* depends on the verb *altiris* as its object. The preposition *de* depends on the noun *esploristoj* with the dependency type *case*, and the noun *esploristoj* depends on the noun *atenton* with the dependency type *nmod*.

The table below summarizes the frequencies and the percentages of the UPOS tags in Prago and Cairo:

	Prago		*Cairo_eo*		*Cairo_en*	
	Freq.	%	Freq.	%	Freq.	%
acl	11	1.31	0	0	1	0.53
acl:relcl	8	0.95	1	0.56	0	0
advcl	4	0.48	2	1.13	1	0.53
advmod	35	4.17	13	7.34	5	2.67
amod	96	11.44	1	0.56	1	0.53
appos	3	0.36	3	1.69	1	0.53
aux	1	0.12	2	1.13	9	4.81
aux:pass	1	0.12	2	1.13	2	1.07
case	121	14.42	6	3.39	8	4.28
cc	29	3.46	8	4.52	10	5.35
cc:preconj	3	0.36	0	0	0	0
ccomp	8	0.95	2	1.13	3	1.60
compound	0	0	1	0.56	2	1.07
conj	34	4.05	8	4.52	7	3.74
cop	16	1.91	2	1.13	4	2.14
det	78	9.30	10	5.65	20	10.7
mark	16	1.91	5	2.82	5	2.67
mwe	0	0	0	0	1	0.53
name	0	0	0	0	2	1.07
neg	3	0.36	2	1.13	2	1.07
nmod	94	11.2	5	2.82	9	4.81
nmod:poss	4	0.48	5	2.82	0	0
nsubj	55	6.56	22	12.43	23	12.30
nsubj:pass	1	0.12	2	1.13	2	1.07
nummod	1	0.12	0	0	0	0
obj	31	3.69	14	7.91	12	6.42

obl	29	3.46	0	0	0	0
orphan	0	0	3	1.69	0	0
parataxis	0	0	1	0.56	0	0
punct	104	12.4	28	15.82	25	13.37
remnant	0	0	0	0	6	13.37
root	46	5.48	20	11.3	20	10.7
vocative	0	0	1	0.56	1	0.53
xcomp	7	0.83	8	4.52	5	2.67
	839		177		187	

Table 8: The frequencies and the percentages of the Universal Dependency Relation tags in *eo_prago_ud-test.conllu* (indicated as *Prago*), *eo_cairo_ud-text.conllu* (indicated as *Cairo_eo*), and *cairo/en.conllu* (indicated as *Cairo_en*)

The noun phrase in the nominative case dependent on a verb functions as nominal subject and its dependency is annotated with *nsubj*. An infinitive verb can function as the subject of another verb, and its dependency is also annotated with *nsubj*. A finite clause introduced by *ke* can function as the subject of another verb (e.g., *Ke la studento legos ĉi tiun libron, ne surprizas min* "That the student will read this book does not surprise me"), and its dependency will be annotated with *csubj*, yet it is missing both in Prago and Cairo. The passive voice of Esperanto is expressed by a participle accompanied by the auxiliary *esti* in its finite form, such as *Tiu ĉi libro estas legata de studento* "This book is read by a student" and the subject of a participle in a passive sentence will be labeled with *nsubj:pass* (or *csubj:pass* if the subject is a finite clause), yet it is missing both in Prago and Cairo.

A noun with the accusative case morpheme -*on* in singular or -*ojn* in plural dependent on a verb is its object, hence the dependency between them is annotated with *obj*. The dependency between a verb and a prepositional phrase is annotated with either *obl* or *nmod*. Since the distinction between them remains uncertain in Esperanto and at present based on the annotator's intuition, this issue must be addressed in future study. According to the guideline of Universal Dependency Relations, the verb *esti* "be" is either treated as the auxiliary when it depends on a verbal predicate with the dependency type *aux* or *aux:pass*, or as the copula when it depends on a non-verbal predicate with the type *cop*. The verbs *devi* "must" and *povi* "can" are treated not as auxiliaries but as full verbs in Prago and Cairo.

4 Discussion

4.1 Comparison and contrast between Prago and Cairo_eo

Due to the large difference in the token size between Prago and Cairo_eo, we need to be cautious about drawing any conclusion from the comparisons of the frequencies of UD POSs and UD relations. Provided that, based on the findings in Prago and Cairo_eo, we can point out that some POSs are worthy of focus, such as those for content words (ADJ, NOUN, VERB) and function words (PRON). The same logic applies to UD relations such as *amod*, *nmod*, *nsubj*, and *obj*. Further studies using UD-annotated text data will reveal differences in style of Esperanto texts, which will contribute to the understanding of Esperanto as a natural language.

4.2 Comparison and contrast between Cairo_eo and Cairo_en

The same size of tokens (however small it is) and the same sentential meaning between Cairo_eo and Cairo_en will provide us with more reliable insight into the difference between English and Esperanto. The most obvious difference between them in terms of UD POS is ADV; Cairo_eo has twice as large number of ADVs as Cairo_en. Also, UD relations *advcl* and *advmod* are more frequent in Cairo_eo than in Cairo_en. This may suggest that Esperanto prefers adverbial expressions than English does. In contrast, Cairo_en has twice as large number of AUXs as Cairo_eo. This must be the result of the annotation policy of Cairo_eo where the verb *devi* "must" and *povi* "can" are annotated as VERB. This policy might be revised in future annotation processes.

4.3 Research in the future

The insights in the preceding subsections lead us to realize the necessity of much larger-scale Esperanto corpus with UD annotation which is also in parallel with other languages. Currently, manual annotation of *Deklaratio pri Homaraniso* "Declaration of Homaranism" (Zamenhof 1913) with UD POS tags and UD relation tags is under way; Comparing and contrasting this new UD-annotated text written by Zamenhof in the early 20th century with existing UD treebank based on *Manifesto de Prago* written in the end of 20th century are expected to give us discovery in diachronic change of Esperanto: Through more than 100 years of its history, Esperanto must have gone through historical changes since its creation, just like other natural languages, and articulating how it has actually changed in terms of usage of morphemes and dependency relations is of linguistic interest. In order to extend this line of research, large-scale UD-annotated Esperanto treebank will function as the ample data for such studies. In addition to this, adding Esperanto version into Parallel Universal Dependencies (PUD) will give us opportunities for multi-lingual contrastive research of Esperanto. However, it is time-consuming, and therefor unrealistic, to manually translate the 1,000 English original sentences in PUD into Esperanto by one researcher and to annotate them with UD POS tags and UD relation tags all manually. We need to conduct these tasks on a team basis, developing an Esperanto UD parser and evaluating its performance with a UD-annotated gold-standard Esperanto texts. These issues must be addressed in future studies, which will contribute to the development not only of Esperanto NLP, but also of understanding of Esperanto as a genuine natural language.

5 Conclusion

This paper described the details of UD-based morphological and syntactic annotations on Esperanto texts to construct its small-scale UD treebank. After reviewing the previous research on Esperanto as a natural language, the detail of manual annotation of UD morphological and relational tags is described. Then the frequencies of these tags are shown to differ across different genres of Esperanto texts, and between English sentences and their Esperanto translations. It is suggested that larger-scale Esperanto UD treebank will open the possibility of future research which will contribute to the idea that Esperanto is not an artificial language but a natural language.

References

Eckhard Bick. 2007. Tagging and Parsing an Artificial Language: an annotated web-corpus of Esperanto. *Proceedings of Corpus Linguistics* 2007.

Eckhard Bick. 2009. A Dependency Constraint Grammar for Esperanto. *Constraint Grammar and robust parsing: Proceedings of the NODALIDA 2009 workshop*. 8-12.

Eckhard Bick. 2016. A Morphological Lexicon of Esperanto with Morpheme Frequencies. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 1075-1078.

Eckhard Bick. 2018. Arbobanko-A treebank for Esperanto. *International Conference on Computational Linguistics and Intelligent Text Processing*. 248-260. Cham: Springer Nature Switzerland.

Eckhard Bick. 2019. Automatic Generation and Semantic Grading of Esperanto Sentences in a Teaching Context. *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*. 10-19.

Eckhard Bick. 2020. Syntax and Semantics in a Treebank for Esperanto. *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 5120-5127.

Chris Gledhill. 2000. *The Grammar of Esperanto: A Corpus-based description*. München: Lincom Europa.

Grant Goodall. 2023. Esperanto kaj lingvistiko: cent jaroj da (mal)amikeco. "Esperanto and linguistics: A hundred years of friendship and hostility." *Esperantologio / Esperanto Studies*, 4(12). DOI: 10.59718/ees43692

Martin Haspelmath, Matthew Dryer, David Gil and Bernard Comrie, ed. 2005. *The World Atlas of Language Structures*. Oxford: Oxford U. P.

Ilona Koutny. 2015. A typological description of Esperanto as a natural language. Język. *Komunikacja. Informacja "Language. Communication. Information"*. vol.10, 43-62.

Marte Djupvik Kråkmo. 2022. *English vs. Esperanto: A comparative study of clausal word order in a Minimalist framework*. Master's thesis, Department of Foreign Languages and Translation University of Agder.

Bill Manaris, Luca Pellicoro, George Pothering, and Harland Hodges. 2006. Investigating Esperanto's statistical proportions relative to other languages using neural networks and Zipf's law. *Proceedings of the 24th IASTED International Conference on Artificial Intelligence and Applications*. Innsbruck, Austria: Acta press. 102–108.

Carlo Minnaja and Laura Paccagnella. (2000). A Part-of-Speech Tagger for Esperanto oriented to MT. *Proceedings of the International Conference on Machine Translation and Multilingual Applications in the New Millennium: MT 2000*.

Mikael Parkvall. 2010. How European is Esperanto?: A typological study. *Language Problems and Language Planning*, 34(1). 63-79.

Alberto Poncelas, Jan Buts, James Hadley, and Andy Way. (2020) Using multiple subwords to improve English-Esperanto automated literary translation quality. *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, 108–117.

Ida Stria. 2015. Esperanto as a natural language. *Język. Komunikacja. Informacja "Language. Communication. Information"*. vol.10, 32-42.

Ludoviko Lazaro Zamenhof. 1887. *Dr. Esperanto's International Tongue (La Unua Libro)*.

Universal Dependencies for Suansu

Jessica K. Ivani[1,2,3] **Kira Tulchynska**[4]

[1]University of Tübingen [2]University of California Santa Barbara [3]University of Zurich
[4]Hebrew University of Jerusalem

Correspondence: jessica.ivani@uni-tuebingen.de

Abstract

This contribution presents the Naga-Suansu Universal Dependencies (UD) treebank, the first resource of this kind for Suansu, an endangered and underdocumented Tibeto-Burman language spoken in Northeast India. This treebank follows the UD annotation framework. We describe the corpus composition, data sources, and annotation process, outlining the general structure of the treebank. In addition, we highlight morphosyntactic challenges where Suansu grammar does not fit neatly into the UD annotation schema and propose adaptations to better capture its structural properties. As the first Tibeto-Burman language included in the UD project, the Naga-Suansu treebank serves several purposes: it contributes to the documentation and preservation of endangered languages, enables the understanding of cross-linguistic variation, and supports future research efforts in refining UD annotation practices for South and Southeast Asian languages.

1 Introduction

Universal Dependencies (UD) is a framework aiming to provide a unified grammatical annotation standard for a wide range of typologically diverse languages (McDonald et al., 2013; Nivre et al., 2016; de Marneffe et al., 2021). Thanks to its rich annotation capabilities and unified approach, UD has become an increasingly popular resource for cross-lingual NLP research, particularly for syntactic parsing and multilingual language modeling tasks. In addition, large UD corpora have been used in descriptive linguistics to explore internal syntactic variation (Kiss and Thomas, 2019) as well as in cross-linguistic and typologically oriented studies (Naranjo and Becker, 2018; Levshina, 2019).

Although the number of languages and treebanks has increased in the last decade, several language families and regions, including low-resource and endangered languages, are still underrepresented in UD.

This contribution introduces the Universal Dependencies (UD) treebank for Suansu (treebank name: Naga-Suansu[1]), an endangered and under-described language spoken along the Indo-Myanmar border. To our knowledge, this represents both the first treebank corpus for Suansu and the first UD contribution for the Tibeto-Burman language subfamily. Developed in close collaboration with the Suansu-speaking community, this resource aims to facilitate the inclusion of additional genealogically and geographically related languages into UD, enhance the typological diversity of UD corpora, and promote greater visibility and representation for under-resourced languages in language technology research.

The paper is organized as follows. First, we introduce Suansu language, outline its most prominent linguistic features, and clarify the naming conventions used for the treebank (Section 2). Section 3 describes the corpus and data sources, while in Section 4 we detail the data preprocessing steps, describing the orthographic conventions (4.1) and lemmatization process (4.2). In Section 4.3 we present the part-of-speech tags used in the treebank, and Section 4.4 outlines the features in the treebank, with a more detailed discussion on converbs (4.5) and alignment (4.6). In Section 5 we illustrate the dependency relations, with examples from copula clauses (5.1) and verb compounds (5.2). We conclude (6) with a summary and some directions for future research.

2 Suansu

Suansu is a Tibeto-Burman language spoken by approximately 2,000 people living in a small cluster of villages located in Ukhrul district, Manipur state, India, near the Indo-Myanmar border. It

---

[1] The treebank has been released as UD_Naga-Suansu, an official UD treebank, available at `https://universaldependencies.org/treebanks/nmf_suansu/` under the CC BY-SA 4.0 license.

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 30–38
August 27, 2025 ©2025 Association for Computational Linguistics

is an agglutinative language with a strong preference for suffixation and a strict verb-final word order, traits that are typical of languages in this region (Post and Burling, 2017). Prominent morphological processes include verb compounding and serialization. Suansu morphosyntactic alignment does not fit neatly into the common alignment patterns described in the typological literature, a trait it shares with other regional languages such as Meithei (Chelliah, 1997), Poumai Naga (Veikho, 2021), and Mongsen Ao (Coupe, 2007). Suansu argument marking is primarily influenced by semantic and pragmatic factors rather than syntactic ones. Our annotation schema to account for these distributions, developed per the Universal Dependencies guidelines, is described in Section 4.6.

Suansu is an endangered language, and its documentation is ongoing (Ivani, 2023, 2024; Ivani and Zakharko, 2024): cross-village communication, education, and media predominantly use Tangkhul Naga, the regional *lingua franca*. Thus, Suansu lacks a standardized orthography. This paper offers a first orthography for the language, developed jointly with Suansu community members and detailed in Section 4.1.

The decision to name the UD treebank "Naga Suansu" was motivated primarily by logistical considerations. Suansu currently lacks an ISO code (`Glottocode: suan1234`), a requirement for registering UD treebanks. Consequently, the treebank was assigned the ISO code [`nmf`], originally designated for Tangkhul Naga. To clearly differentiate Suansu from Tangkhul within this shared coding scheme, we chose a term that could unambiguously encompass both languages in the future. After consultations with members of the Suansu-speaking community, the term "Naga" was proposed, as it better reflects shared ethnic affiliation while foregrounding linguistic identity. This choice aligns with established naming conventions among nearby languages (e.g., Mao Naga, Poumai Naga). However, readers should note that "Naga" is an ethnic term without inherent linguistic implications. In this paper, we use Naga-Suansu to indicate the UD treebank, and Suansu when we refer to the language.

3 Data

The Naga-Suansu treebank comprises four distinct data sources, each translated into Suansu by native speakers and subsequently manually glossed and annotated. The corpus contains 584 sentences and 3123 words across the four sources. A summary description of the corpus is available in Table 1.

Text	**Sentences**	**Words**
film_Bridge	332	1990
grammar_BivalTyp	131	536
grammar_Cairo	20	159
grammar_ValPal	101	438

Table 1: Corpus information

The most significant portion of the treebank consists of the first 300 lines of subtitles from the film *Bridge of Spies* (2015), forming part of a broader parallel subtitles corpus project (Ebert et al., 2023).

The second source is the ValPal dataset (Hartmann et al., 2013), comprising 101 sentences designed to investigate valency patterns across languages. The last two sources of treebank include 131 sentences from the BivalTyp questionnaire (Say, 2020) and 20 sentences from the Cairo CICLing Corpus[2]. All materials were translated directly from English into Suansu by native speakers, followed by manual interlinear morphological glossing and annotation.

The choice of these sources is motivated by practical considerations, such as the availability of analyzed and glossed materials. At present, the corpus is skewed toward written data. We plan to include conversational data and additional spoken materials in future releases.

4 Annotation

The data was divided into small batches, each containing fewer than 100 sentences. The first batch was converted from interlinear glosses into a CoNLL-U file with tokenized sentences without annotated features and dependencies. Morphosyntactic annotation for this initial batch was carried out entirely manually.

Subsequent batches were processed using the UDPipe 1 pipeline (Straka and Straková, 2017), trained on the manually annotated data from previous batches. The resulting automatic parses were then reviewed and corrected manually.

Throughout the process, any ambiguous or previously unencountered structures were discussed collaboratively by the authors and native speakers to establish a consensus on the appropriate annotation strategy.

---

[2] `https://github.com/UniversalDependencies/cairo/tree/master`

Phonemes	Orthographic representations
/ɬ/	*lh*
/θ/, /tʰ/	*th*
/x/	*hr*
/χ/	*hw*
/ə/	*ā*
nasalized V	*Vhn*

Table 2: Some examples of orthographic representation of Suansu phonemes. Nasalised vowels are phonemic in Suansu.

4.1 Tokenization and Word Segmentation

In the treebank, whitespace is used to mark word boundaries. Lexical units are identified based on phonological independence, which is determined by the presence of primary stress. The orthography employed in this treebank is a preliminary system developed in collaboration with members of the Suansu-speaking community. Based on the Latin script, it draws partially from the Tangkhul Naga orthographic tradition. This Suansu orthography was designed primarily by and for the community, with several key goals in mind: to serve as an initial step toward a standardized writing system; to be accessible and easily typable, facilitating intuitive use in everyday communication, including messaging and chat platforms; and to accurately represent Suansu-specific linguistic features, such as the lateral fricative /ɬ/, that are not present in Tangkhul.

Additionally, the writing system was designed with future refinements in mind, allowing for straightforward updates as needed. While certain phonemic distinctions in Suansu are currently represented by the same grapheme, for instance, *th* is used for both /θ/ and /tʰ/, this overlap does not impede intelligibility among speakers. Table 2 presents illustrative examples of phonemes, some of which are unique to Suansu in contrast to Tangkhul, along with their corresponding orthographic representations in the treebank. Future orthographic developments will address additional features such as diphthongs and tone marking.

4.2 Lemmatization

Certain morphological processes are not explicitly marked in lemma forms. For instance, lexical compounds are treated as single units in the lemma, rather than being segmented into their individual components. Suansu exhibits several productive derivational processes, with compounding and reduplication being particularly prominent. Examples include *themok.nui* 'milk' (literally *themok* 'cow' + *nui* 'breast') and *sui.sui* 'follow' (cf. *sui* 'back'). Such forms are preserved in their derived state in the lemma, including systematic patterns like full reduplication of verb stems or auxiliaries, which is used to express interrogative mood. Similarly, inherently possessed kinship terms remain unsegmented in the lemma, as in *a.thi* 'mother' (literally 'my mother').

Class	UPOS	Total	%
Open	ADJ	51	1.63%
	ADV	191	6.11%
	INTJ	82	2.62%
	NOUN	669	21.41%
	PROPN	260	8.32%
	VERB	635	20.33%
Closed	ADP	28	0.90%
	AUX	176	5.63%
	CCONJ	23	0.74%
	DET	136	4.35%
	NUM	47	1.50%
	PART	50	1.60%
	PRON	311	9.96%
	SCONJ	72	2.30%
Other	PUNCT	387	12.39%
	X	6	0.19%

Table 3: POS tags and their frequencies in the Naga-Suansu treebank.

4.3 Universal Part-of-Speech tags

The Universal Dependencies (UD) framework defines 17 universal part-of-speech (POS) tags. Of these, all except `SYM` (symbol) are attested in the Naga-Suansu treebank. Table 3 summarizes the POS tags used in the treebank along with their respective frequencies. In the following sections, we provide a brief overview of selected parts of speech as they appear in the treebank, highlighting their key properties.

4.3.1 `NOUN`

Suansu nouns take case (1) and number (2) markers, although the number formative can also be found on verbs, where it functions as a nominalizer (cf. below). In pronominal possession, the pronominal form precedes the possessed noun (3).

(1) *Peter lairak-**di** lua-te*
Peter book-**TOP** take-COMPL
'Peter took a book.'

(2) *Tye baneo-**pha** lapui-di chuhn-e*
DET boy-**PL** road-TOP cross-PST
'The boys crossed the road.'

(3) ***A** miszu garhe*
1**SG** person NEG.COP
'Not my guy.'

4.3.2 VERB

Suansu verbs are subject to several inflectional and derivational processes. There is no person or number agreement. Suansu verbs take TAM markers (see above and 4), and derivational morphology includes adverbials (5) and directionals (6), among others, and are further discussed in Section 5.2. Nominalization is extensive in Suansu and encompasses several markers, including plural forms (7). Serialization is highly productive in the language (8).

(4) *Tye duh makh-**e***
DET old.person cough-**PST**
'The old man coughed.'

(5) *A phethe **rai** ga re rung*
1SG eat **COLL** HORT like.that say
wi-le sir
have-PRS sir
'I have a dinner date, sir.'

(6) *Nue-ganan la-di rung **kai,***
ask-CVB AUX-NMLZ say **UP**
Lieutenant
lieutenant
'Just answer the question, Lieutenant.'

(7) *A-va client ba-nan makhwa-da*
1SG-GEN client 3SG-ERG early-ABL
*rung-ha-**pha**-di the-ma ga reha*
say-PST-**NMLZ**-TOP do-NEG EVD REP
lala?
AUX
'...My client is not honoring the claim?'

(8) *Atra thahn kanahn li **la** **ve**-e*
hot place where also **bring** **go**-PST
'I went somewhere hot.'

4.3.3 ADJ and ADV

In Suansu, adjectives and adverbs do not form distinct classes; rather, they are derived from verbs and retain verbal properties (9). Modifiers with adjectival (10) and adverbial (11) functions are typically formed through nominalization, derivational processes, and case marking. In the treebank, such lexemes are annotated as adjectives or adverbs based on their syntactic behavior. However, not all noun modifiers undergo nominalization (12).

(9) *neo-di **szu-e***
child-TOP **be.good-PST**
'The child was good.'

(10) ***ka-szu-e** neo-di hanahn*
NMLZ-be.good-NMLZ child-TOP here
lai-le
be-PRS
'The good child is here.'

(11) *mazohn szuka **szu-nan** thai-le*
all stuff **be.good-ERG** see-PRS
ba-byahn
3SG-BEN
'Everything goes well for him.'

(12) *Peter mobile phone **katha** ska manung-le*
Peter mobile phone **new** one desire-PRS
'Peter wants a new mobile phone.'

4.3.4 PRON

The set of independent personal pronouns in Suansu, as represented in the Naga-Suansu treebank, is shown in Table 4. In addition to personal pronouns, the treebank includes demonstrative pronouns (e.g., *hadi* 'this', *didi* 'that'), quantifiers such as *mazohn* 'all', interrogative pronouns like *thuza* 'who' and *mwe* 'what', as well as indefinite pronouns (e.g., *chatha* 'others'), which frequently function as determiners.

	Singular	**Plural**
1st person	*a*	*ha*
2nd person	*nahn*	*na*
3rd person	*ba*	*bu*

Table 4: Independent personal pronouns in the Naga-Suansu treebank.

4.3.5 `DET`

Determiners in Suansu precede the noun and include modifiers that express definiteness and deixis, such as *hai* ‘this’ (13) and *tye* ‘that’ (14).

(13)

hai	*lairak-di*	*lua*	*dai*
DET	book-TOP	take	JUS

‘Take this book.’

(14)

tye	*lairak-di*	*lua*	*dai*
DET	book-TOP	take	JUS

‘Take that book.’

4.3.6 `AUX`

In the Naga-Suansu treebank, the most prominent auxiliary is the existential verb *la*, which frequently bears TAM inflection in periphrastic constructions, such as those expressing progressive aspect (cf. Section 4.5). We also tag as `AUX` most mood formatives that occur as independent phonological units, with the exception of the imperative suffix *-a*, which is treated as a verbal suffix. Modal forms are treated similarly: the phonologically independent obligative modal *geraha* is annotated as `AUX` (15), whereas abilitive and permissive modals, which are tightly bound to the verb, are treated as verbal suffixes (15). The copula *e* is also tagged as `AUX` and linked via the `cop` relation (see Section 5.1).

(15)

A-nan	*Lynn-di,*	*hai*	*me*	*re*
1SG-ERG	lynn-TOP	PRO	no	REP

*rung-**gam**-ma*	*rahn*	*re,*	*nahn*	*li*
say-**MOD**-NEG	IRR	like.that	2SG	also

me	*re*	*rung*	***geraha***	*matikza*	*thok*
no	REP	say	**MOD**	very	difficult

rahn
IRR

‘I told Lynn, the firm can’t say no, and you’d have a tough time saying no, too.’

4.3.7 `PART`

Forms that do not meet the criteria for other part-of-speech categories are annotated as particles. In the Naga-Suansu treebank, these include the reportative marker *re* (when it appears after quotes), the copular negator *garhe*, and various discourse particles.

4.3.8 `SCONJ`

We use the tag `SCONJ` for nominalizers, including the clause nominalizer *di*, which also functions as a relativizer. This formal overlap between relativization and nominalization is commonly observed in Tibeto-Burman languages. The tag also applies to forms that introduce subordinate clauses, such as the purposive *gase* (16) and the temporal adverbial *ganan*, which additionally serves as a converb in certain aspectual contexts (see Section 4.5).

(16)

Peter	*miaowi-di*	*chokla*	***gase***	*the-ganan*
Peter	cat-TOP	catch	**PURP**	do-CVB

la-le
be-PRS

‘Peter is trying to catch the cat.’

4.3.9 `INTJ`

Expressions primarily used as exclamations or parts of exclamatory utterances are annotated as `INTJ`. This category also includes fillers, such as the numeral *ska* ‘one’ when used in this function, as well as forms like *min* ‘right’ and *ay* ‘yeah’ in backchanneling contexts.

4.4 Features

The morphological features annotated in the Naga-Suansu treebank are listed in Table 5. Most align with existing Universal Dependencies feature definitions, though a few additional features and values have been introduced. In the following section, we briefly discuss these newly added features, along with key annotation decisions involving converbs and alignment.

4.4.1 `Modal=Abil, Obl, Perm`

The Naga-Suansu treebank introduces the feature `Modal`, which groups together modality-related values as follows: `Abil` for abilitive constructions (e.g., expressing physical or cognitive ability, comparable to English *can* or *be able to*), `Obl` for expressions of necessity and obligation, and `Perm` for constructions encoding permission (e.g., *be allowed to* in English). In Suansu, the verbal suffix *-ngam-* is used to express ability, while *-szu-* encodes permission. Both forms occur immediately after the verb stem, and before any TAM marker. Strong and weak necessity are marked by the phonologically independent form *geraha*, which typically appears in clause-final position. We refer to the discussion of `AUX` in 4.3 for the POS annotation of these forms.

Features	Values	Count	%
Abbr	Yes	2	0.04%
Aspect	Imp, Perf, Prog	473	9.56%
Case	Abl, Ben, Dat, DatErg†, Erg, ErgTop†, Gen, GenAbl†, GenTop†, Loc, LocTop†, Top	504	10.18%
Degree	Cmp, Pos	45	0.91%
Deixis	Prox, Remt	151	3.05%
Evidentiality	Fh, Nfh	9	0.18%
ExtPos	ADV, VERB	12	0.24%
Foreign	Yes	6	0.12%
Modal†	Abil†, Obl†, Perm†	14	0.28%
Mood	Des, Hort†, Imp, Ind, Int, Irr, Jus	451	9.11%
Number	Plur, Sing	1218	24.61%
NumForm	Digit, Word	46	0.93%
NumType	Card	47	0.95%
Person	1, 2, 3	263	5.31%
Polarity	Neg, Pos	131	2.65%
PronType	Dem, Ind, Int, Prs, Tot	467	9.44%
Reflex	Yes	2	0.04%
Tense	Past, Pqp, Pres	365	7.38%
VerbForm	Conv, Fin, Inf, Vnoun	743	15.01%

Table 5: Features in the Naga-Suansu treebank. † marks new features and values introduced in the treebank.

4.4.2 `Mood=Hort`

Suansu exhibits several strategies for expressing speaker attitude toward actions, including a fine-grained encoding of orders, requests, and encouragements. In the Naga-Suansu treebank, we annotate a range of mood categories alongside modal expressions, including imperative, desiderative, and jussive moods. To account for polite forms of encouragement, we introduce the hortative mood (`Mood=Hort`). This is typically expressed by the clause-final form *ga*, which we tag as an auxiliary. An example illustrating the hortative mood is provided below (17), alongside a jussive example (18), the latter used for polite directives, to highlight the functional and contextual differences between the two.

(17) *Mazwehn, Doug, hai-da phethe* ***ga***
alright doug 1PL-ABL eat HORT
'Alright Doug, you can join us for dinner.'

(18) *Kapiu matha* ***dai***
search start **JUS**
'Start searching the place.'

4.4.3 `Case=DatErg, ErgTop, GenAbl, GenTop, LocTop`

The Naga-Suansu treebank includes instances of case stacking, or *Suffixaufnahme* (Plank, 1995), where two case suffixes co-occur on a single noun phrase. The case combinations attested include genitive + ablative (`Case=GenAbl`, 19), dative + ergative (`Case=DatErg`, 20), ergative + topic (subject) (`Case=ErgTop`), and locative + topic (object) (`Case=LocTop`). Dependency relations are annotated based on the syntactic context and the argument structure of the associated verb root.

(19) *Hai letter-di Peter-****va-da***
DET letter-TOP Peter-**GEN-ABL**
'This letter is from Peter.'

(20) *A-****la-nan*** *phabta-ma*
1SG-**DAT-ERG** understand-NEG
'It's not clear to me.'

4.5 Converbs

Suansu employs several converbs to express a range of functions, including adverbial subordination, as widely attested in the typological literature (Haspelmath, 1995). One such form, *ganan*, is used to indicate simultaneous or immediately consecutive actions or states. It follows an uninflected verb, while the final verb in the clause, which denotes the subsequent action or state, carries TAM marking (21). Additionally, *ganan* functions in a periphrastic construction to express progressive aspect, where it attaches to the main verb and is followed by the existential auxiliary *la*, which is inflected for tense (22). These distinct uses are differentiated in our Features annotation. Similar annotation strategies have been adopted in the Tatar (Taguchi et al., 2016) and Uyghur (Eli et al., 2016) Universal Dependencies treebanks, where comparable converb forms serve analogous functions.

(21) *Neo pha wiza-va rung-e*
child PL teacher-GEN say-NMLZ
*chu-**ganan** mazohnna-e*
hear-**CVB** sit.down-PST
'The children obeyed the teacher and sat down.'
`VerbForm=Conv`

(22) *Peter guitar khohn-**ganan** la-le*
Peter guitar play-**CVB** be-PRS
'Peter is playing the guitar.'
`Aspect=Prog|VerbForm=Conv`

4.6 Alignment

Suansu alignment system is complex; it is motivated by semantic rather than morpho-syntactic factors and does not fit neatly into the typological classification proposed in the literature. Semantic Agents are generally marked with the ergative marker *nan*, while Patients and Experiencers are often marked with *di*. However, the presence of these markers is not obligatory and is heavily influenced by pragmatic factors. Syntactic subjects of intransitive sentences are typically unmarked.

In our annotation scheme, we label the syntactic subjects of transitive clauses with the ergative case marker *nan* (23). The marker *di*, on the other hand, is annotated as a topic marker, and its syntactic role is determined contextually (23, 24). Previous work (Ivani, 2023, 2024; Ivani and Zakharko, 2024) has shown that *di* is strongly associated with discourse pragmatics; it serves to highlight or single out a referent, drawing the listener's attention to it within the discourse. This function aligns more closely with topicalization than with core argument marking. By tagging *di* as `Topic`, we reflect its pragmatic role in discourse, while syntactic dependencies are annotated separately based on each construction. This approach allows us to remain consistent with UD guidelines without resorting to misleading labels like 'absolutive' or 'accusative', while also preserving the unique alignment properties of Suansu.

(23) *Hui-nan Peter-**di** maka-j*
dog-ERG Peter-**TOP** bite-PST
'A dog bit Peter.'
`Case=Top|DepRel=obj`

(24) *Peter-**di** Maria thrwa-le*
Peter-**TOP** Maria be.similar-PRS
'Peter resembles Maria.'
`Case=Top|DepRel=nsubj`

5 Dependency relations

The Universal Dependency v2 lists 37 syntactic relations. Of these 37 dependency relation types, thirty are used in the Naga-Suansu treebank[3]. Dependencies and related counts are reported in Table 6. We briefly discuss syntactic relations pertaining to copula clauses (5.1) and verb compounding (5.2).

Dependency	**Count**	**%**
root	584	18.69%
nsubj	427	13.67%
punct	387	12.39%
obj	289	9.25%
advmod	159	5.09%
aux	156	4.99%
obl	153	4.90%
det	137	4.39%
nmod:poss	95	3.04%
mark	84	2.69%
xcomp	52	1.66%
discourse	46	1.47%
advcl	44	1.41%
flat	42	1.34%
conj	41	1.31%
nummod	37	1.18%
nmod	32	1.02%
amod	32	1.02%
vocative	32	1.02%
...		
other	295	9.44%

Table 6: Dependency relations used in the Naga-Suansu treebank.

5.1 Copula clauses

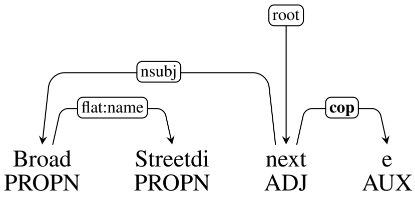
Identity clauses in Suansu are expressed via the copula form *e*, which functions as a linker in these structures and does not carry tense or aspect information. It is tagged in the treebank as `AUX` and annotated via a cop relation, as shown by the example in Figure 1).

5.2 Verb compounding

Verb compounding processes are abundant in Suansu. In the treebank, we distinguish `compound:prt` for compounds involving verbs and

---

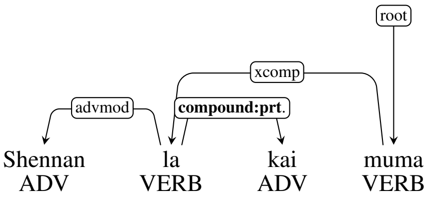
[3] Syntactic dependencies absent in the current Naga-Suansu treebank are `clf`, `dep`, `dislocated`, `expl`, `fixed`, `goeswith` and `list`.



‘Broad Street is next’

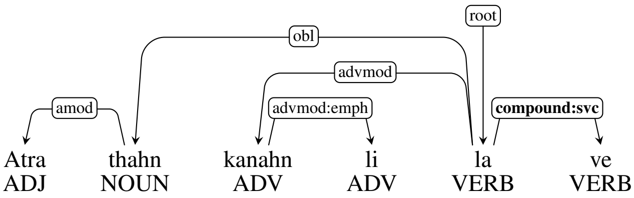
Figure 1: Annotation of copula clauses.

directional particles, and `compound:svc` for serial verb constructions. The syntactic relation `compound:prt` is illustrated in Figure 2a, while the syntactic relation `compound:svc` is exemplified in Figure 2b. These dependency relations reflect both morphological and syntactic integration: directional particles in `compound:prt` are tightly bound to the verb, forming a single semantic unit, whereas in serial verb constructions (`compound:svc`), multiple verbs occur in sequence and jointly express a complex event. Our annotation captures this distinction by analyzing the first verb in the sequence as dependent on the second, more semantically central verb.



‘I’m not sure I want to pick it up’

(a) Annotation of verb compounds with directional particles.



‘I went somewhere hot’

(b) Annotation of serial verb constructions.

6 Conclusion

We have presented the first treebank for Naga-Suansu, the first resource of its kind for Suansu, an endangered language, and the first known contribution to Universal Dependencies from a language within the Tibeto-Burman subfamily. Alongside the treebank, we introduced our approach to orthographic standardisation and detailed the part-of-speech tags, morphological features, and dependency relations used in annotation, most of which align closely with the UD guidelines. Where necessary, we proposed extensions, such as the addition of `Mood=Hort` and modality features, and offered a proposal for representing morphosyntactic alignment specific to Suansu. We also addressed dependency relations for copular clauses and compounding.

Our future goals include expanding the dataset, by completing the UD annotation of the movie *Bridge of Spies*, which currently covers the first 300 sentences. We also plan to include annotated conversational data, further broadening the range of linguistic contexts represented in the treebank. Finally, we plan on incorporating code mixing and multilingual influences, such as borrowings from the regional lingua franca Tangkhul Naga. Through this work, we hope to encourage the development of additional UD treebanks for other Tibeto-Burman languages and to support a growing body of resources for endangered languages within the UD ecosystem.

Acknowledgments

This research was funded by the Global Strategy and Partnerships Funding Scheme of the University of Zurich, project *Event Packaging in Language*, which we gratefully acknowledge here. We extend our gratitude to the Suansu-speaking community for their invaluable contributions, and especially to Jason M. Vashum for his generous assistance with translation and annotation. We also thank the three anonymous reviewers for their constructive comments on the draft of this paper, and Taras Zakharko for insightful comments and formatting support.

References

Shobhana Lakshmi Chelliah. 1997. *A Grammar of Meithei*, volume 17 of *Mouton Grammar Library*. Mouton de Gruyter, Berlin.

Alexander Robertson Coupe. 2007. *A Grammar of Mongsen Ao*, volume 39 of *Mouton Grammar Library*. Mouton de Gruyter, Berlin.

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

Christian Ebert, Natalia Levshina, and Paul Widmer. 2023. Partree - parallel treebanks: A multilingual corpus of movie subtitle.

M. Eli, W. Mushajiang, T. Yibulayin, K. Abiderexiti, and Y. Liu. 2016. Universal dependencies for Uyghur. In *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*, pages 44–50, Osaka, Japan. The COLING 2016 Organizing Committee.

Iren Hartmann, Martin Haspelmath, and Bradley Taylor, editors. 2013. *Valency Patterns Leipzig (ValPaL) Online Database*. Max Planck Institute for Evolutionary Anthropology. Accessed: 2025-04-16.

Martin Haspelmath. 1995. The converb as a cross-linguistically valid category. In Martin Haspelmath and Ekkehard König, editors, *Converbs in Cross-Linguistic Perspective: Structure and Meaning of Adverbial Verb Forms – Adverbial Participles, Gerunds*, pages 1–56. Mouton de Gruyter, Berlin.

Jessica K. Ivani. 2023. Suansu language from northeastern India: A field report. *Linguistics of the Tibeto-Burman Area*, 46(1):138–163.

Jessica K. Ivani. 2024. Caritive expression in Suansu. *Himalayan Linguistics*, 23(3):28–40.

Jessica K. Ivani and Taras Zakharko. 2024. Phasal polarity in Suansu. *Linguistics of the Tibeto-Burman Area*, 47(2):318–342.

Angelika Kiss and Guillaume Thomas. 2019. Word order variation in Mbyá Guaraní. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 121–129, Paris, France. Association for Computational Linguistics.

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

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

M. Guzmán Naranjo and Laura Becker. 2018. Quantitative Word Order Typology with UD. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, pages 91–104, Oslo, Norway. Linköping University Electronic Press.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 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, Portorož, Slovenia. European Language Resources Association (ELRA).

Frans Plank. 1995. *Double Case: Agreement by Suffixaufnahme*. Oxford University Press, New York.

Mark W. Post and Robbins Burling. 2017. The Tibeto-Burman languages of Northeastern India. In Graham Thurgood and Randy J. LaPolla, editors, *The Sino-Tibetan Languages*, pages 213–242. Routledge, London.

Sergey Say. 2020. Bivaltyp: Typological database of bivalent verbs and their encoding frames. Accessed: 2025-04-16.

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, Vancouver, Canada. Association for Computational Linguistics.

Chihiro Taguchi, Sei Iwata, and Taro Watanabe. 2016. Universal dependencies treebank for Tatar: Incorporating intra-word code-switching information. In *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*, pages 44–50, Osaka, Japan. The COLING 2016 Organizing Committee.

Sahiinii Lemaina Veikho. 2021. *Grammar of Poumai Naga (Poula): A Trans-Himalayan Language of North-East India*. Brill, Leiden.

Crossing Dialectal Boundaries: Building a Treebank for the Dialect of Lesbos through Knowledge Transfer from Standard Modern Greek

Stavros Bompolas[1], **Stella Markantonatou**[1,2], **Angela Ralli**[1,3],
Antonios Anastasopoulos[1,4]
[1]Archimedes, Athena Research Center, Greece
[2]ILSP, Athena Research Center, Greece [3]University of Patras, Greece
[4]George Mason University, USA
Correspondence: s.bompolas@athenarc.gr

Abstract

This paper presents the first treebank for the dialect of Lesbos, a low-resource living northern variety of Modern Greek (MG), annotated according to the Universal Dependencies (UD) framework. So far, the only dialectal treebank available for Greek developed with cross-dialectal knowledge transfer is an East Cretan one, which belongs to the same southern branch as Standard Modern Greek (SMG). Our study investigates the effectiveness of cross-dialectal knowledge transfer between dialectologically less similar varieties of the same language by leveraging knowledge from SMG to annotate the northern dialect of Lesbos. We describe the annotation process, present the resulting treebank, inject additional linguistic knowledge to enhance the results, and evaluate the effectiveness of cross-dialectal knowledge transfer for active annotation. Our findings contribute to a better understanding of how dialectal variation within language families affects knowledge transfer in the UD framework, with implications for other low-resource varieties.[1]

1 Introduction

The Universal Dependencies (UD) project (de Marneffe et al., 2021) has established consistent syntactic representations across 168 languages, but non-standard varieties are underrepresented. This gap arises from challenges in text collection, the scarcity of qualified annotators, and the expertise needed to adapt existing guidelines (Blaschke et al., 2024). Documenting less-used dialects not only preserves linguistic diversity but also provides valuable insights for contrastive linguistic analysis while serving as a testbed for computational approaches in data-scarce scenarios.

Two UD treebanks exist for Standard Modern Greek (SMG), and recent efforts aim to develop dialectal ones, including the Lesbian treebank. However, Modern Greek (MG) dialects remain largely unexplored. Moreover, research on cross-dialectal knowledge transfer in NLP for MG is limited. To date, the only effort in this direction has focused on the East Cretan dialect—an intuitively favorable case, given its relative proximity to SMG as a fellow southern MG variety.

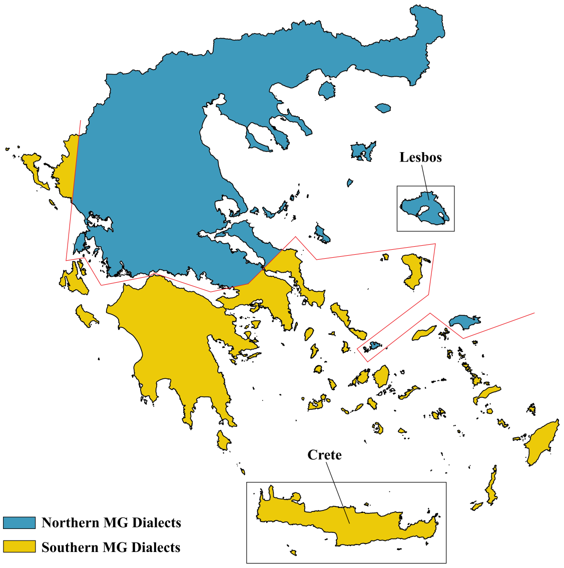


Figure 1: The geographic position of Lesbian (and Cretan) and the isogloss delineating northern and southern MG dialects.

This paper addresses this gap with three key contributions. (a) We present the first UD treebank for a northern MG dialect, focusing on Lesbian. (b) We introduce dialect-specific annotation guidelines for Lesbian and a method for integrating orthographic conventions into the annotation scheme, which is adaptable to other (northern MG) dialects without standardized orthography. (c) We extend previous research on knowledge transfer between SMG and southern MG di-

---

[1] The treebank is available at https://github.com/UniversalDependencies/UD_Greek-Lesbian, released as part of UD v2.16 (May 15, 2025).

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 39–51
August 27, 2025 ©2025 Association for Computational Linguistics

alects (i.e., Cretan) to investigate cross-dialectal transfer between linguistically less similar varieties, specifically from SMG, a southern variety, to a northern one (i.e., Lesbian). In particular, we explore how the injection of linguistic knowledge can improve cross-dialectal transfer.

2 Related Work

Recent work has increasingly focused on developing dialectal treebanks and exploring knowledge transfer for low-resource languages. Treebanks have been created for varieties such as Egyptian Arabic (Maamouri et al., 2014), Norwegian (Øvrelid et al., 2018; Kåsen et al., 2022), Occitan (Miletic et al., 2020), and Bavarian (Blaschke et al., 2024), among others. For surveys of cross-lingual transfer methods in dependency parsing, see Magueresse et al. (2020), Das and Sarkar (2020), Hedderich et al. (2021), and Pakray et al. (2025).

2.1 SMG Treebanks

The UD framework has been applied to SMG through two treebanks. The first, UD_Greek-GDT (Prokopidis and Papageorgiou, 2017), is based on the Greek Dependency Treebank (http://gdt.ilsp.gr). The more recent and comprehensive treebank, UD_Greek-GUD (GUD), follows UD.v2 morphological guidelines, ensuring improved consistency and coverage (Markantonatou et al., 2025).

These treebanks provide a solid foundation for dialect-oriented NLP by allowing cross-dialectal knowledge transfer from SMG to under-resourced dialects. They also serve as important benchmarks, as new resources for MG should align with their guidelines and validation standards for compatibility and interoperability within the UD ecosystem.

2.2 MG Dialectal Treebanks

Three UD treebanks have been developed for MG dialects recently: two for Cappadocian—UD_Cappadocian-AMGiC (Sampanis and Prokopidis, 2021) and UD_Cappadocian-TueCL (Vligouridou et al., 2024) and one for Eastern Cretan, UD_Greek-Cretan (Vakirtzian et al., 2025). Among these, only the work focusing on Cretan has investigated cross-dialectal knowledge transfer from SMG, leveraging the linguistic proximity between these two southern MG varieties.

In contrast, our work addresses cross-dialectal transfer across less similar dialect groups, specifically from SMG, a southern MG variety, to Lesbian, a northern dialect. To our knowledge, this is the first study within the UD framework to examine knowledge transfer between more distantly related MG dialects.

3 The Lesbos Dialect

As shown in Figure 1, the Lesbos dialect belongs to the northern MG dialect group, which is characterized by the so-called "northern vocalism"—specifically, the raising of unstressed mid vowels /e/ and /o/ into [i] and [u], respectively (e.g., *πιδί* [piˈði] instead of SMG *παιδί* [peˈði] 'child', *κάτου* [ˈkatu] instead of SMG *κάτω* [ˈkato] 'down'), and the deletion of unstressed high vowels /i/, /u/ (e.g., *φίδ* [ˈfið] instead of SMG *φίδι* [ˈfiði] 'snake', *βνό* [ˈvno] instead of SMG *βουνό* [vuˈno] 'mountain'). These features distinguish the dialect of Lesbos from southern dialects, including SMG (Chatzidakis, 1905).

The dialect has been shaped by extensive historical contact with Italo-Romance and Turkish (Ralli, 2015, 2019a,c; Alexelli, 2021). During Italo-Romance rule (1355-1462), numerous loanwords and morphological elements of Venetian origin were introduced, such as the diminutive suffix *-ελ(ι)* [-el(i)] (Melissaropoulou and Ralli, 2010). The subsequent Ottoman period (1462-1912) further enriched the dialect with Turkish borrowings. Around the 16th century, speakers from Lesbos settled in the nearby Asia Minor areas of Kydonies and Moschonisia, where the dialects share many similarities with Lesbian. After the Asia Minor Catastrophe (1922) and the Treaty of Lausanne (1923), refugees from these regions permanently resettled in dialectal enclaves in Lesbos, resulting in a complex linguistic system on the island, characterized by intra-dialectal variation and features absent from SMG.

Today, unlike most MG dialects, Lesbian remains vital, serving as the primary means of communication on the entire island.

4 The Treebank

This section outlines the procedure for creating the Lesbian Treebank. As a UD_Greek treebank, it broadly follows the existing annotation guidelines for SMG and adopts the same set of UFeats, UPOS tags, and dependency relations. Accordingly, we focus on dialect-specific deviations and annotation decisions driven by dialectal features not addressed in the SMG guidelines. This overview is not intended to be exhaustive (for a recent overview on the dialectal variation in Lesbos, see Alexelli, 2021 and the linguistic atlas of Lesbos, Ralli, 2019b).

4.1 Source Materials

The corpus draws from six main sources representing different text types and dialectal variants from the island of Lesbos: (a) example sentences extracted from three comprehensive dialectal dictionaries (Papanis and Papanis, 2004; Ralli, 2017; Anagnostopoulou, 2021), and (b) sentences of oral nature taken from three additional texts of contemporary Lesbian literature (Tsokarou-Mitsioni, 1998; Anagnostou, 2014; Tsokarou-Mitsioni, 2019), including humorous tales, plays, narratives, and personal accounts written in the dialect.

These sources capture internal variation within the Lesbian dialect, including sub-dialectal and stylistic differences across narrative and conversational contexts. They span a broad stylistic range—from standardized dictionary entries to informal, orally styled literary texts—exhibiting significant orthographic and grammatical variation.

One of the major challenges in working with these materials is the lack of a standardized orthography for Lesbian, which, like many dialects, is primarily an oral variety (for an overview of the issues that arise when representing spoken varieties in UD treebanks, see Dobrovoljc, 2022). In the aforementioned sources, the general trend in orthographic representation leans toward conformity with SMG, but several issues arise:

1. Despite the adherence of the authors to SMG spelling, the texts contain a considerable number of orthographic errors.
2. Significant inconsistencies also emerge in the representation of the northern vocalism, particularly regarding the raising of unstressed /e/ to [i] and the deletion of unstressed high vowels /u, i/. Within SMG orthography, vowel /i/ corresponds to multiple graphemic representations (<η, ι, υ, ει, οι>), leading authors to reflect vowel raising with notable inconsistency, influenced by both stylistic considerations and etymological factors. Similarly, the deletion of unstressed high vowels is occasionally marked with an apostrophe (’), yet this orthographic strategy, when employed, exhibits irregular application, resulting in orthographic variation for identical lexical items. For example, the Lesbian counter-form for the SMG word *απόμεινε* [aˈpomine] ‘remained’ is attested as *απόμνι*, *απόμνει*, *απόμ’νι*, and *απόμ’νει* [aˈpomni].
3. These orthographic challenges are further complicated by distinctive phonological features of the Lesbian dialect that lack standardized representation in SMG orthography, such as euphonic sound insertion and consonant voicing phenomena (e.g., SMG *κοντά του* [konˈda tu] ‘close to him/it’ > Lesbian *κουντά τ* [kunˈda t] *vs. κουντά ντ* [kunˈda d]), thus introducing additional complexity to orthographic standardization efforts.

4.2 Annotation

The annotation of the Lesbian dialect has primarily followed the UD annotation guidelines established for GUD, complemented by grammatical descriptions (Anagnostou, 1903) and dialect dictionaries (Papanis and Papanis, 2004; Ralli, 2017; Anagnostopoulou, 2021). We use the same set of dependency (sub)relations as defined for SMG. Only deviations, new constructs, and forms have been documented in supplementary guidelines specific to the Lesbian treebank, which are listed as comments on the GUD guidelines. Our main annotations remain compatible with existing (S)MG treebanks in UD, facilitating comparative research. It should be noted, however, that our lemmatization respects Lesbian phonology and morphology rather than conflating with SMG lemmas. As a result, our treebank passes SMG validation rules with minimal exceptions for certain auxiliaries that differ only due to the phonological application of northern vocalism (i.e., the lemmas *έχου* [ˈexu] ‘have’ instead of SMG *έχω* [ˈexo]; *είμι* [ˈimi] ‘be’ instead of SMG *είμαι* [ˈime]).

4.3 Tokenization

Following earlier (S)MG treebanks, we segment adposition-determiner contractions; for example, *στο* [sto] ‘in/to the’ is tokenized as two syntactic words, *σ* [s] ‘in/to’ and *το* [to] ‘the’. Unlike SMG treebanks, we needed to split not only contracted forms but also clitics that are frequently attached to verbs in written dialectal texts (e.g., *τάμπλιξις* [ˈtabliksis] ‘(you) mixed them up’ > *τά* + *μπλιξις*). The same approach applies to possessive pronouns frequently attached to nouns (e.g., *πατέρασιτς* [paˈterasits] ‘her father’ > *πατέρας* + *ιτς*).

While this decision aligns with guidelines from earlier MG treebanks regarding tokenization handling, it differs from GUD's approach of pretokenizing contracted forms. In our treebank, we maintain merged word sequences as written in the original sources and treat such cases as multiword tokens. Additionally, we have respected the original written sources by not merging tokens that were erroneously split, using instead the "goeswith" relation for these instances.[2]

4.4 Lemmatization

For lemmatization, we relied primarily on dialectal dictionaries. As already stated, the words of Lesbian that diverge from their SMG counterparts were assigned a lemma form that bears the dialectal characteristics. Nevertheless, this aspect of annotation required significant effort due to several challenges:

1. As previously mentioned, phonological phenomena are not spelled uniformly across or even within sources due to orthographic inconsistencies. Consequently, the same words are often spelled differently in the legacy texts.
2. Although northern vocalism is a defining feature of northern dialects, it is not applied uniformly even within the same source, creating inconsistencies for the same lexical items (e.g., SMG *φοβάμαι* [foˈvame] 'I am afraid' > Lesbian *φοβάμι* [foˈvami] *vs.* *φουβάμι* [fuˈvami]).
3. Sources frequently contain orthographic errors unrelated to dialectal features.

To address these issues:

For 1.: We eliminated apostrophes from lemmas, since these are not used consistently in the texts. This decision allowed the lemma to serve as a unifying element across all texts, regardless of whether they systematically used apostrophes, didn't use them at all, or used them inconsistently. This standardization approach represents a significant contribution to MG dialectal text processing, as apostrophe usage has been a persistent challenge across Greek dialect documentation efforts. Our systematic treatment of this orthographic feature, combined with the annotation approach developed for this treebank, offers a replicable methodology that can benefit future computational work on MG dialects with similar orthographic variation.

For 2.: We consistently used the dialectologically expected form (with vowel raising and deletion) as the lemma, even when texts did not systematically apply these features; i.e., for both *φοβάμι* and *φουβάμι* we assigned the lemma *φουβάμι*.

For 3.: Regardless of orthographic errors in the source materials, we applied standardized (MG) orthography to lemmas.

4.5 Morphology

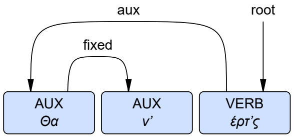
The Lesbian dialect's morphology broadly aligns with SMG, though surface forms differ due to northern vocalism. Notable morphological differences include:

1. Use of the definite article *η* [i] 'the' for masculine nouns in the nominative case of the singular number, phonetically matching the feminine definite article. The SMG definite article /o/ (raised to [u]) occasionally appears in free variation with the latter.
2. 3rd-person plural present and future active verb forms take the inflectional suffix *-in* instead of SMG *-un* (e.g., Lesbian *χάν-ιν* [ˈxan-in] *vs.* SMG *χάν-ουν* [ˈxan-un] 'they lose').
3. Distinctive diminutive suffixes, particularly the highly productive *-ελ(ι)* [-el(i)], attach to bases of all genders, loanwords, and proper names (instead of SMG *-άκι* [-aci]). For these forms, we lemmatize to the base word without the diminutive suffix.

4.6 (Morpho-)Syntax

Based on available written sources and oral material, few syntactic divergences have been identified between the Lesbian dialect and SMG (Ralli, 2019a), documented in our sources and annotated in the treebank:

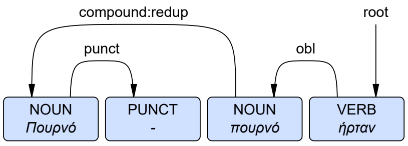
1. Alternation between genitive and accusative case in examples such as *τ*.GEN *έδουσι ένα δικάρ* [t ˈeðusi ˈena ðiˈkar] *vs.* *τουν*.ACC *έδουσι ένα δικάρ* [tun ˈeðusi ˈena ðiˈkar] 'he/she gave him a dim'.
2. In addition to the SMG future particle *θα* [θa] 'will', the form *θα ν(α)* [θa n(a)] 'will' is frequently used in the dialect, which is not attested in SMG. We annotated these structures as follows:
   - *Θα ν' έρτ' ς* [θa n ˈerts] 'you will come'



---

[2] https://universaldependencies.org/u/overview/typos.html#wrongly-split-word

3. Although reduplication is also attested in SMG, it occurs with notably higher frequency in Lesbian, likely due to the dialect's prolonged contact with Turkish (0 occurrences in 1807 sentences of GUD and 12 occurrences among 270 sentences in the Lesbian treebank). These reduplicated forms typically involve two identical lexical elements with occasional morphophonological alternations. Following the UD guidelines,[3] reduplication is annotated as follows in the Lesbian treebank:
    - *Πουρνό-πουρνό ήρταν* [purˈno purˈno ˈirtan] 'They came very early in the morning'

    Additionally, we annotate the head of the compound construction with the UFeat Degree=Aug, as these constructions primarily function to express intensity or augmentation of the core meaning.
4. Given the oral nature of some sources, we frequently employ the parataxis relation, which is much less common in GUD (26 occurrences in GUD and 82 occurrences in the Lesbian treebank). This relation connects pairs of potentially standalone sentences treated as a single sentence. In spoken corpora, this occurs naturally as sentence boundaries often align with utterance turns. When more than two sentences join in this manner, we make all subsequent sentences dependents of the first one, reflecting the structural parallel between parataxis and conjunction relations.

4.7 Voicing and Euphonics

Following Vakirtzian et al. (2025, 780-781), we integrated voicing and euphonic annotations in the MISC (10th) column, explicitly documenting these phenomena to support comparative analysis, e.g., dialectometry.

Euphonics are vowels or consonants that appear within, between, or at the end of words. In Lesbian (and MG more broadly), they create open syllables and avoid hiatuses. We tag these elements as "euphonic" using the MSeg|MGloss format, allowing us to treat them as separate tokens and distinguish them from the rest of the word.

Unlike SMG, where voicing phenomena are not represented orthographically, dialectal texts often explicitly spell voiced consonants that reflect dialect-specific voicing patterns. These dialectal voicing patterns, which may differ from those in SMG, create additional orthographic complexity when represented in writing. To annotate this phenomenon, the corresponding unvoiced form is used as the lemma, and we add the feature-value pair Voicing=Voiced in the MISC column. This approach contributes to annotation consistency and facilitates knowledge transfer from SMG, which typically uses the unvoiced version of these lemmas.

4.8 Standardization and Translation

Sometimes the text underlying a UD treebank does not conform to canonical spelling or other grammatical rules of the language. In most situations, it is desirable to preserve the error because taggers and parsers that learn their models from the data should learn how to deal with noisy input too. On the other hand, it is also desirable to mark such places as errors and to show the correct spelling so that an application can hide bad sentences or present their correct version when necessary.

Working with dialectal text sources presents a significant challenge due to the absence of standardized orthography. The literary and folkloric texts in our corpus use inconsistent spelling conventions to represent dialectal features, complicating computational processing. While we decided against normalizing the original data sources to an artificial Lesbian standard—as no such written or spoken standard exists and would contradict our goal of curating diverse sources—we still needed to address orthographic inconsistency. Therefore, to facilitate language modeling and enhance cross-dialectal knowledge transfer, we implemented a standardization process integrated with the UD annotation scheme, following the guidelines for handling non-standard forms described in the UD documentation.[4] In our approach (Examples 1-2 in Appendix B):

1. Original dialectal forms were preserved in the FORM (2nd) column of the CoNLL-U.
2. Standardized forms (i.e., correct forms closer to SMG orthography or systematized spelling for

---

[3] https://universaldependencies.org/u/dep/all.html#compound-redup-reduplicated-compounds

[4] https://universaldependencies.org/u/overview/typos.html

northern vocalism) were annotated in the MISC (10th) column following the annotation guidelines of UDs.

3. We developed a dedicated script that processes the CoNLL-U files, extracting the standardized forms from the MISC column and inserting them into the FORMs[5].

This standardization was crucial for the subsequent active annotation cycles, as it allowed our models to better leverage lexical and morphological knowledge from SMG while preserving the original characteristics of the resources.

Furthermore, we have incorporated translations of each sentence in SMG, produced by the annotators. These translations maintain maximum fidelity to the original sentences while adhering to SMG conventions, thereby establishing a parallel corpus that may facilitate future comparative research and computational applications.

5 Transfer Experiments

This section presents baseline experiments for evaluating dependency parsing performance on the Lesbian treebank. The experimental setup aligns with recent approaches to NLP in low-resource settings, as surveyed by Hedderich et al. (2021).

5.1 Active Annotation

To annotate the Lesbian treebank, we employed active annotation (Vlachos, 2006). In order to facilitate comparative analysis with Vakirtzian et al.'s (2025) prior research and results on knowledge transfer between SMG and a southern MG dialect (Cretan), and to investigate the extent to which SMG can contribute to modeling a northern MG dialect (Lesbian), we replicated their experimental regime.

Initially, 40 unlabeled Lesbian samples were annotated with a model trained on GUD (that represents SMG). In each subsequent cycle, 40 samples from the model's output were corrected, out of which 30 were allocated to the training set and 10 to the development set. The corrected samples were incorporated into the existing datasets, and the model was retrained on the augmented data. For evaluation, we used a test set of 30 manually annotated samples. All samples were randomly selected to ensure unbiased representation. To enhance cross-dialectal transfer, we utilized standardized forms from the MISC column (Section 4.8).

	1st	2nd	3rd	4th	5th	6th
Sentences						
Train	30	60	90	120	150	180
Dev	10	20	30	40	50	60
Test	30	30	30	30	30	30
Tokens						
Train	392	810	1220	1604	1994	2381
Dev	166	308	449	595	695	826
Test	396	396	396	396	396	396

Table 1: Lesbian sentences and tokens per round of active annotation.

During the development of the UD treebank for the Lesbian dialect, annotation guidelines evolved alongside our research, with revisions consistently applied across all datasets to ensure consistency.

5.2 Reducing Dialectal Distance

While Lesbian's morphosyntax is similar to SMG, phonological variation, especially northern vocalism, complicates cross-dialectal knowledge transfer. To address this, we used simple linguistic rules to reduce the dialectal gap by generating synthetic data (Aufrant et al., 2016).

Specifically, we created a Python script to transform the GUD treebank according to key features of northern MG phonology.[6] This resulted in a "northernized" version of GUD (NGUD) that more closely approximates the phonological profile of Lesbian.

The transformation involved two primary modifications (Examples 3-4 in Appendix B):

1. **Applying northern vocalism rules**, including the raising of unstressed /e/ to [i] and /o/ to [u], and the deletion of unstressed high vowels /i, u/, implemented within the constraints of MG orthography.
2. **Altering definite articles**: masculine nominative singular *ο* was systematically replaced with *η*, reflecting patterns attested in northern MG dialects.

Importantly, all syntactic structures and dependency relations were preserved, ensuring compatibility with UD annotation standards.

This synthetic data augmentation strategy enabled us to test whether the reduction in orthographic distance between SMG and Lesbian Greek due to the phonological distance between these two dialects can improve the effectiveness of cross-dialectal transfer. It also allowed us to isolate the

---

[5] https://github.com/stavros-bompolas/conllu-correct-forms

[6] https://github.com/stavros-bompolas/ngud-transformer

role of phonological divergence as a potential barrier to transfer between dialects belonging to different clusters—southern (SMG) *vs.* northern (Lesbian).

5.3 Models

For the experiments, we used the open-source Stanza package (Qi et al., 2020) with three distinct settings:[7]

1. **Lesbian-only**: A model trained exclusively on the Lesbian data that increased at each round by 40 samples (30 in the training set and 10 in the development set).
2. **GUD+Lesbian**: A model trained on the combination of GUD (1,807 sentences, 25,493 tokens) and Lesbian data, with the Lesbian component increasing exactly as in the Lesbian-only model.
3. **NGUD+Lesbian**: A model trained on the northernized GUD treebank plus the Lesbian data, to test whether reducing dialectal distance through synthetic data enhances knowledge transfer.

In all settings, we fine-tuned the Greek BERT model (Koutsikakis et al., 2020).

5.4 Results

Figure 2 displays the precision metrics for UPOS, Lemmas, UFeats, and LAS in six evaluation rounds. The remaining metrics can be found in Appendix A.[8] In presenting the results, we compare our findings with Vakirtzian et al.'s (2025) work on knowledge transfer between SMG and Cretan (a southern MG dialect) to examine how SMG contributes to modeling Lesbian (a northern MG dialect). We refer to Vakirtzian et al.'s GUD+Cretan model as "GUD+Cretan" and their Cretan-only model as "Cretan-only" throughout our discussion.

Overall, results show that both knowledge transfer approaches significantly outperform the dialect-only model.

UPOS: NGUD/GUD+Lesbian both reach 89.62%, compared to 82.24% for the Lesbian-only model. However, GUD+Cretan achieves 92.90%, suggesting more effective transfer for the southern dialect.

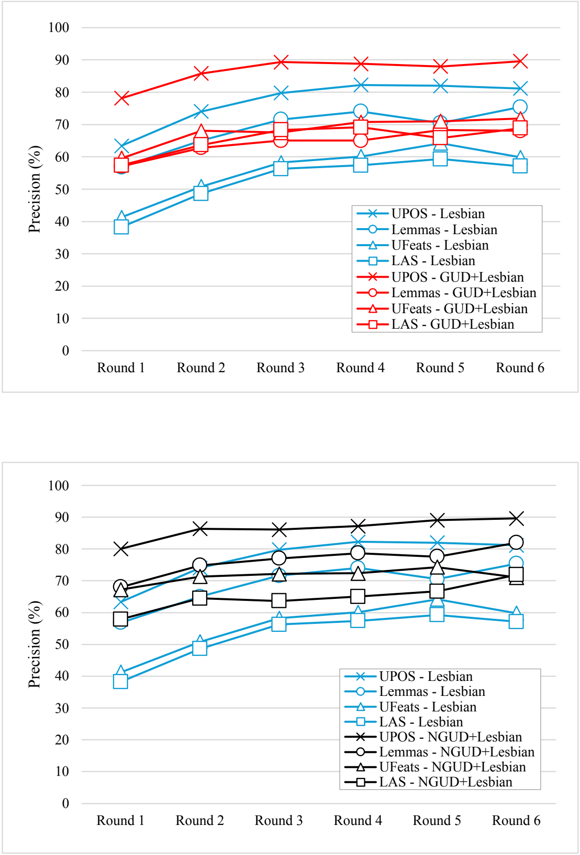


Figure 2: Precision scores between the Lesbian and GUD+Lesbian (top) / NGUD+Lesbian (bottom).

UFeats: NGUD+Lesbian (74.32%) outperforms GUD+Lesbian (71.86%) and Lesbian-only (64.21%), but lags behind both Cretan-only (78.70%) and GUD+Cretan (87.22%), highlighting the greater challenge for northern dialects in knowledge transfer.

Lemmas: NGUD+Lesbian achieves the highest performance (81.97%), dramatically outperforming GUD+Lesbian (68.31%) and matching Cretan-only (81.34%), demonstrating the impact of synthetic data augmentation.

LAS: NGUD+Lesbian (71.86%) outperforms both Lesbian-only (59.29%) and Cretan-only (67.75%), though GUD+Cretan still leads with 78.50%, indicating that dialectal proximity remains a key factor in successful knowledge transfer.

6 Discussion

Comparing knowledge transfer between SMG and a northern (Lesbian) *vs.* a southern (Cretan) dialect reveals several key insights:

---

[7] https://osf.io/yacxu/?view_only=37105ee21ef64ee297109b99dc875c38

[8] For evaluation, we used the standard CoNLL shared task evaluation script, which computes precision, recall, F1-score, and accuracy metrics. We report the precision values from this script's output. Available at: https://github.com/universaldependencies/tools?tab=readme-ov-file#evalpy

The impact of dialectal distance: A southern dialect benefits more from SMG knowledge transfer than a northern one, as evidenced by the superior performance of GUD+Cretan over GUD+Lesbian across all metrics. This stems from Cretan's closer vocalic system to SMG, while Lesbian's northern vocalism introduces greater surface differences at multiple linguistic levels, demonstrating that phonological distance between northern and southern dialects limits cross-dialectal transfer, regardless of orthographic consistency (see Vakirtzian et al. 2024 for similar results in ASR for MG dialects; see also Faisal and Anastasopoulos 2022).

The effectiveness of adaptation through synthetic data: Our synthetic data approach significantly narrows this gap, enabling the northern dialect model to match or outperform standalone southern models across several metrics, particularly lemmatization. Its effectiveness correlates with the degree of transformation: of the 25,493 tokens processed in the GUD treebank, only 39.82% of surface forms and 31.82% of lemmas remained unaffected by the script, with the rest undergoing northern vocalism adaptations. These results underscore the potential for further improvements through expanded rules and additional resources such as dictionaries (Zhao et al., 2009) and parallel corpora (Yarowsky et al., 2001).

The importance of available resources for the standard variety: The effectiveness of (N)GUD highlights the crucial role of robust standard variety resources in cross-dialectal transfer. Even for less similar dialects, such as Lesbian, high-quality SMG resources enhance performance, especially when combined with appropriate adaptation techniques. High-resource standard varieties cover greater linguistic variability, providing valuable baselines for transfer without the challenges of dialectal resource development (Snæbjarnarson et al., 2023).

The role of source characteristics in knowledge transfer: The Lesbian treebank integrates diverse resources, introducing variation that complicates cross-dialectal transfer but enhances representativeness (Dobrovoljc, 2022). In contrast, Cretan's data are more uniform as they come from a single speaker. Additionally, the orthography used to transcribe Cretan oral material is identical to SMG orthography, which likely facilitated knowledge transfer from the SMG treebank.

7 Conclusions

In this paper, we presented the first UD treebank for a northern MG dialect (Lesbian), along with tailored annotation guidelines and cross-dialectal transfer experiments.

Our findings suggest that effective cross-dialectal knowledge transfer depends on several factors: (a) greater dialectal distance reduces transfer effectiveness; (b) the nature and diversity of source materials affect performance; (c) simple rule-based transformations of high-resource varieties can substantially improve performance for distant dialects.

This research extends to other MG dialects, particularly northern varieties that share similar phonological features with Lesbian. Hence, the treebank provides a foundation for developing resources for other northern dialects and contributes significantly to advancing dialectal diversity in Greek NLP.

Our work highlights an inherent paradox in cross-dialectal knowledge transfer: while we aim to leverage pre-existing resources from high-resource varieties, doing so effectively often requires developing additional dialect-specific technologies—the very situation we sought to avoid through transfer learning. Similarly, a tension exists between preserving dialectal characteristics and adapting linguistic representations to enhance cross-dialectal transfer. These contradictions underscore the complex relationship between linguistic authenticity and technological pragmatism in developing NLP resources for dialectal varieties.

8 Limitations

Several limitations should be acknowledged:

First, the current treebank is small (270 sentences), limiting model performance and comprehensive linguistic documentation.

Second, our data relies solely on written sources that may reflect authors' idealized forms rather than authentic dialectal usage. Although we have recently collected oral Lesbian dialectal data for future incorporation, the current resource lacks this direct representation.

Third, our rule-based approach to northern vocalism adaptation applies changes deterministically, whereas actual dialectal usage shows considerable variation in the application of these phonological rules. A probabilistic transformation model might better capture this natural variation.

Fourth, our experimental design, while enabling direct comparison with previous work on Cretan, may not represent the optimal approach for cross-dialectal transfer between distant varieties. Alternative methods such as leveraging larger pre-trained language models specifically fine-tuned for dialectal variation could potentially yield better results.

Finally, we did not explore the extent to which our findings generalize to other dialectal pairs with similar degrees of distance, either within Greek or in other languages with comparable dialectal landscapes. Such comparative analysis would provide a more comprehensive understanding of the relationship between dialectal distance and transfer effectiveness.

Acknowledgments

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. It also received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

References

Vasileia Alexelli. 2021. *Chartografisi tis glossikis poikilias tis Lesvou [Mapping the linguistic variety of Lesbos]*. Ph.D. thesis, University of Patras, School of Humanities and Social Sciences, Department of Philology, Linguistics Section.

Maria Ach. Anagnostopoulou. 2021. *Thematiko Lexiko tis Lesviakis Dialektou [Thematic Dictionary of the Lesbos Dialect]*. Mythos BOOKS, Mytilene.

Spyridon Anagnostou. 1903. *Lesviaka, iti, Sylogi laografikon peri Lesvou pragmateion [Lesbian Studies, or, Collection of Folkloric Treatises on Lesbos]*. From the Press of the "Anestis Konstantinidis" establishments, Athens.

Vasilis Tz. Anagnostou. 2014. *Tsi sta th'ka mas: Komodia sta k'stariot'ka [Tsi sta th'ka mas: Comedy in the K'stariot'ka Dialect]*, first edition. Estia Technon Skoutarou "T'Apono to Scholio".

Lauriane Aufrant, Guillaume Wisniewski, and François Yvon. 2016. Zero-resource Dependency Parsing: Boosting Delexicalized Cross-lingual Transfer with Linguistic Knowledge. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 119–130, Osaka, Japan. The COLING 2016 Organizing Committee.

Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank. 2024. MaiBaam: A Multi-Dialectal Bavarian Universal Dependency Treebank. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10921–10938, Torino, Italia. ELRA and ICCL.

Georgios Chatzidakis. 1905. *Mesaionika kai Nea Ellinika [Medieval and Modern Greek]*, volume 1–2. Sakellarios, Athens.

Ayan Das and Sudeshna Sarkar. 2020. A survey of the model transfer approaches to cross-lingual dependency parsing. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 19(5):67:1–67:60.

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

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.

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-inspired adaptation of multilingual models to new languages. *Preprint*, arXiv:2205.09634.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. *Preprint*, arXiv:2010.12309.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. GREEK-BERT: The Greeks visiting Sesame Street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, pages 110–117, New York, NY, USA. Association for Computing Machinery. Event-place: Athens, Greece.

Andre Kåsen, Kristin Hagen, Anders Nøklestad, Joel Priestly, Per Erik Solberg, and Dag Trygve Truslew Haug. 2022. The Norwegian Dialect Corpus Treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4827–4832, Marseille, France. European Language Resources Association.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. 2014. Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2348–2354, Reykjavik, Iceland. European Language Resources Association (ELRA).

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *CoRR*, abs/2006.07264.

Stella Markantonatou, Vivian Stamou, Stavros Bompolas, Katerina Anastasopoulou, Irianna Linardaki Vasileiadi, Konstantinos Diamantopoulos, Yannis Kazos, and Antonios Anastasopoulos. 2025. VMWE identification with models trained on GUD (a UDv.2 treebank of Standard Modern Greek). In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 14–20, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Dimitra Melissaropoulou and Angela Ralli. 2010. Greek derivational structures: restrictions and constraints. *Morphology*, 20(2):343–357.

Aleksandra Miletic, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade, and Jean Sibille. 2020. A Four-Dialect Treebank for Occitan: Building Process and Parsing Experiments. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 140–149, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Dimitris Papanis and Giannis D. Papanis. 2004. *Lexiko tou Agiasotikou Glosikou Idiomatos (Erminetiko - Etimologiko) [Dictionary of the Agiasos Dialect (Explanatory - Etymological)]*, 3rd improved and expanded edition. Private edition, Mytilene.

Prokopis Prokopidis and Haris Papageorgiou. 2017. Universal Dependencies for Greek. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 102–106, Gothenburg, Sweden. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Angela Ralli. 2015. Strategies and Patterns of Loan Verb Integration in Modern Greek Varieties. In Angela Ralli, editor, *Contact Morphology in Modern Greek Dialects*, pages 73–88. Cambridge Scholars Publishing.

Angela Ralli. 2017. *Lexiko dialektikis poikilias. Kydonion-Moschonision kai Voreioanatolikis Lesvou. Orthografiko-Proforas-Erminetiko-Christiko-Etymologiko-Synonymon [Dictionary of dialectal variety. Kydonies-Moschonisia and Northeastern Lesbos. Orthographic-Pronunciation-Interpretive-Practical-Etymological-Synonyms]*. Hellenic Foundation for Historical Studies, Athens.

Angela Ralli. 2019a. Affixoids and Verb Borrowing in Aivaliot Morphology. In Angela Ralli, editor, *The Morphology of Asia Minor Greek*, pages 221–254. BRILL.

Angela Ralli. 2019b. Glossiki chartografisi: O ilektronikos dialektikos atlas tis Lesvou [Linguistic mapping: The electronic dialect atlas of Lesbos]. In Grammatiki A. Karla, Io Manolessou, and Nikolaos Pantelidis, editors, *Lexeis: Timitikos tomos gia tin Christina Basea-Bezantakou [Words: Festschrift for Christina Basea-Bezantakou]*, pages 435–456. Kardamitsa, Athens.

Angela Ralli. 2019c. Greek in Contact with Romance. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Konstantinos Sampanis and Prokopis Prokopidis. 2021. Asia Minor Greek in Contact (AMGiC): Towards a dialectal Treebank comprising contact-induced grammatical changes. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 86–95, Sofia, Bulgaria. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on faroese. *Preprint*, arXiv:2304.08823.

Eustratia Tsokarou-Mitsioni. 1998. *Palies Istories ap tn Agiasiou [Old Stories from Agiasio]*, 2nd edition. Private Edition, Mytilene.

Eustratia Tsokarou-Mitsioni. 2019. *Prosfygiá [Refugeehood]*, first edition. D. Doukas & Sia O.V.E.E. / Eustratia Tsokarou-Mitsioni.

Socrates Vakirtzian, Vivian Stamou, Yannis Kazos, and Stella Markantonatou. 2025. Dialectal treebanks and their relation with the standard variety: The case of East Cretan and Standard Modern Greek. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 776–784, Tallinn, Estonia. University of Tartu Library.

Socrates Vakirtzian, Chara Tsoukala, Stavros Bompolas, Katerina Mouzou, Vivian Stamou, Georgios Paraskevopoulos, Antonios Dimakis, Stella Markantonatou, Angela Ralli, and Antonios Anastasopoulos. 2024. Speech Recognition for Greek Dialects: A Challenging Benchmark. In *Interspeech 2024*, pages 3974–3978. ISCA.

Andreas Vlachos. 2006. Active Annotation. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)*.

Eleni Vligouridou, Inessa Iliadou, and Çağrı Çöltekin. 2024. A Treebank of Asia Minor Greek. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1715–1721, Torino, Italia. ELRA and ICCL.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

Hai Zhao, Yan Song, Chunyu Kit, and Guodong Zhou. 2009. Cross language dependency parsing using a bilingual lexicon. 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 55–63, Suntec, Singapore. 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).

A Appendix: Precision Metrics

	1st	**2nd**	**3rd**	**4th**	**5th**	**6th**
UPOS	63.39	74.04	79.78	82.24	81.97	81.15
UFeats	41.26	50.82	58.20	60.11	64.21	59.84
AllTags	37.16	46.72	56.01	57.38	61.20	58.74
Lemmas	56.83	65.03	71.58	74.04	70.49	75.41
UAS	64.21	69.40	75.41	76.23	75.14	74.86
LAS	38.25	48.63	56.28	57.38	59.29	57.10
CLAS	24.06	33.18	44.44	42.40	46.58	44.80
MLAS	6.60	14.55	23.11	23.96	29.22	26.24
BLEX	9.91	17.27	26.67	28.11	30.59	31.22

Table 2: Precision metrics across rounds (Trained on Lesbian sentences only).

	1st	**2nd**	**3rd**	**4th**	**5th**	**6th**
UPOS	78.14	85.79	89.34	88.80	87.98	89.62
UFeats	59.56	68.03	67.49	70.77	71.04	71.86
AllTags	56.28	66.12	65.85	68.03	69.13	70.22
Lemmas	57.10	62.84	65.03	65.03	68.31	68.03
UAS	73.22	77.87	82.79	83.06	80.33	82.51
LAS	57.38	63.66	68.31	69.13	65.85	68.85
CLAS	44.86	52.13	57.08	59.91	51.38	56.81
MLAS	24.77	32.70	35.38	37.74	34.40	39.44
BLEX	22.90	33.18	36.79	38.68	36.24	38.50

Table 3: Precision metrics across rounds (Trained on GUD+Lesbian sentences).

	1st	**2nd**	**3rd**	**4th**	**5th**	**6th**
UPOS	80.05	86.34	86.07	87.16	89.07	89.62
UFeats	67.21	71.31	72.13	72.40	74.32	71.04
AllTags	63.11	67.76	68.85	69.40	72.13	68.85
Lemmas	68.03	74.86	77.05	78.69	77.60	81.97
UAS	79.51	80.87	80.60	80.87	82.51	86.34
LAS	57.92	64.48	63.66	65.03	66.67	71.86
CLAS	45.97	52.34	51.18	54.38	54.07	60.37
MLAS	28.91	34.58	32.23	36.87	38.28	39.63
BLEX	30.33	35.98	37.44	41.94	41.63	45.62

Table 4: Precision metrics across rounds (Trained on NGUD+Lesbian sentences).

B Appendix: Examples

```
# text = Πατέρασιτς άμα τν είδει καταφαρμακόστσει .
# text_el = Ο πατέρας της, όταν την είδε, καταφαρμακώθηκε.
# text_en = When her father saw her, he was devastated.
1-2     Πατέρασιτς _ _ _ _ _ _ _ _
1       Πατέρασ Πατέρας NOUN _ Case=Nom|Gender=Masc|Number=Sing 6 nsubj _ _
2       ιτς μ PRON _ Case=Gen|Gender=Fem|Number=Sing|Person=3|Poss=Yes|PronType=Prs 1 nmod _ MGloss=euphonic-her|MSeg=ι-τς
3       άμα άμα SCONJ _ _ 5 mark _ _
4       τν ιγώ PRON _ Case=Acc|Gender=Fem|Number=Sing|Person=3|PronType=Prs 5 obj _ _
5       είδει βλέπου VERB _ Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|Typo=Yes|VerbForm=Fin|Voice=Act 6 advcl _ CorrectForm=είδι
6       καταφαρμακόστσει καταφαρμακώνου VERB _ Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|Typo=Yes|VerbForm=Fin|Voice=Pass 0 root _ CorrectForm=καταφαρμακώστσι
7       . . PUNCT _ _ 6 punct _ _
```

Example 1: Integration of orthographic standardization for the dialect in the 10th (MISC) column.

```
# text = Πατέρασιτς άμα τν είδει καταφαρμακόστσει .
# text_el = Ο πατέρας της, όταν την είδε, καταφαρμακώθηκε.
# text_en = When her father saw her, he was devastated.
1-2     Πατέρασιτς _ _ _ _ _ _ _ _
1       Πατέρασ Πατέρας NOUN _ Case=Nom|Gender=Masc|Number=Sing 6 nsubj _ _
2       ιτς μ PRON _ Case=Gen|Gender=Fem|Number=Sing|Person=3|Poss=Yes|PronType=Prs 1 nmod _ MGloss=euphonic-her|MSeg=ι-τς
3       άμα άμα SCONJ _ _ 5 mark _ _
4       τν ιγώ PRON _ Case=Acc|Gender=Fem|Number=Sing|Person=3|PronType=Prs 5 obj _ _
5       είδι βλέπου VERB _ Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|Typo=Yes|VerbForm=Fin|Voice=Act 6 advcl _ CorrectForm=είδι
6       καταφαρμακώστσι καταφαρμακώνου VERB _ Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|Typo=Yes|VerbForm=Fin|Voice=Pass 0 root _ CorrectForm=καταφαρμακώστσι
7       . . PUNCT _ _ 6 punct _ _
```

Example 2: Orthographic standardization applied automatically from the MISC column via processing script.

```
# text = Ο υπάλληλος σ την είσοδο κουνάει το κεφάλι του , όταν μαθαίνει το σκοπό της επίσκεψής μας.
# text_en = The employee shakes his head at the entrance when he learns the purpose of our visit.
1       Ο ο DET _ Case=Nom|Definite=Def|Gender=Masc|Number=Sing|PronType=Art 2 det _ _
2       υπάλληλος υπάλληλος NOUN _ Case=Nom|Gender=Masc|Number=Sing 6 nsubj _ _
3       σ σε ADP _ _ 5 case _ _
4       την ο DET _ Case=Acc|Definite=Def|Gender=Fem|Number=Sing|PronType=Art 5 det _ _
5       είσοδο είσοδος NOUN _ Case=Acc|Gender=Fem|Number=Sing 6 obl _ _
6       κουνάει κουνώ VERB _ Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act 0 root _ _
7       το ο DET _ Case=Acc|Definite=Def|Gender=Neut|Number=Sing|PronType=Art 8 det _ _
8       κεφάλι κεφάλι NOUN _ Case=Acc|Gender=Neut|Number=Sing 6 obj _ _
9       του μου PRON _ Case=Gen|Gender=Masc|Number=Sing|Person=3|Poss=Yes|PronType=Prs 8 nmod _ _
10      , , PUNCT _ _ 12 punct _ PunctType=Comm
11      όταν όταν SCONJ _ _ 12 mark _ _
12      μαθαίνει μαθαίνω VERB _ Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act 6 advcl _ _
13      το ο DET _ Case=Acc|Definite=Def|Gender=Mas|Number=Sing|PronType=Art 14 det _ _
14      σκοπό σκοπός NOUN _ Case=Acc|Gender=Masc|Number=Sing 12 obj _ _
15      της ο DET _ Case=Gen|Definite=Def|Gender=Fem|Number=Sing|PronType=Art 16 det _ _
16      επίσκεψής επίσκεψη NOUN _ Case=Gen|Gender=Fem|Number=Sing 14 nmod _ _
17      μας μου PRON _ Case=Gen|Number=Plur|Person=1|Poss=Yes|PronType=Prs 16 nmod _ SpaceAfter=No
18      . . PUNCT _ _ 6 punct _ PunctType=Peri
```

Example 3: Example from the GUD treebank prior to applying the transformation script.

```
# text = Η υπάλληλους σ τν είσουδου κνάει του κιφάλ τ , όταν μαθαίν του σκουπό τς ιπίσκιψής μας.
# text_en = The employee shakes his head at the entrance when he learns the purpose of our visit.
1       Η η DET _ Case=Nom|Definite=Def|Gender=Masc|Number=Sing|PronType=Art 2 det _ _
2       υπάλληλους υπάλληλους NOUN _ Case=Nom|Gender=Masc|Number=Sing 6 nsubj _ _
3       σ σι ADP _ _ 5 case _ _
4       τν η DET _ Case=Acc|Definite=Def|Gender=Fem|Number=Sing|PronType=Art 5 det _ _
5       είσουδου είσουδους NOUN _ Case=Acc|Gender=Fem|Number=Sing 6 obl _ _
6       κνάει κνώ VERB _ Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act 0 root _ _
7       του η DET _ Case=Acc|Definite=Def|Gender=Neut|Number=Sing|PronType=Art 8 det _ _
8       κιφάλ κιφάλ NOUN _ Case=Acc|Gender=Neut|Number=Sing 6 obj _ _
9       τ μ PRON _ Case=Gen|Gender=Masc|Number=Sing|Person=3|Poss=Yes|PronType=Prs 8 nmod _ _
10      , , PUNCT _ _ 12 punct _ PunctType=Comm
11      όταν όταν SCONJ _ _ 12 mark _ _
12      μαθαίν μαθαίνου VERB _ Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act 6 advcl _ _
13      του η DET _ Case=Acc|Definite=Def|Gender=Masc|Number=Sing|PronType=Art 14 det _ _
14      σκουπό σκουπός NOUN _ Case=Acc|Gender=Masc|Number=Sing 12 obj _ _
15      τς η DET _ Case=Gen|Definite=Def|Gender=Fem|Number=Sing|PronType=Art 16 det _ _
16      ιπίσκιψής ιπίσκιψ NOUN _ Case=Gen|Gender=Fem|Number=Sing 14 nmod _ _
17      μας μ PRON _ Case=Gen|Number=Plur|Person=1|Poss=Yes|PronType=Prs 16 nmod _ SpaceAfter=No
18      . . PUNCT _ _ 6 punct _ PunctType=Peri
```

Example 4: Transformation of Example 3 after applying the script with rules to convert GUD to northernized GUD (NGUD).

UD-English-CHILDES: A Collected Resource of Gold and Silver Universal Dependencies Trees for Child Language Interactions

Xiulin Yang[⚘] **Zhuoxuan Ju**[⚘] **Lanni Bu**[⚘] **Zoey Liu**[❀] **Nathan Schneider**[⚘]
[⚘]Georgetown University
[❀]University of Florida
{xy236, zj153, lb1437, nathan.schneider}@georgetown.edu
liu.ying@ufl.edu

Abstract

CHILDES is a widely used resource of transcribed child and child-directed speech. This paper introduces UD-English-CHILDES, the first officially released Universal Dependencies (UD) treebank. It is derived from previously dependency-annotated CHILDES data, which we harmonize to follow unified annotation principles. The gold-standard trees encompass utterances sampled from 11 children and their caregivers, totaling over 48K sentences (236K tokens). We validate these gold-standard annotations under the UD v2 framework and provide an additional 1M silver-standard sentences, offering a consistent resource for computational and linguistic research.

1 Introduction

The Child Language Data Exchange System (CHILDES) (MacWhinney, 2000) has long been a key resource for research in language acquisition, computational modeling of child language, and the evaluation of Natural Language Processing (NLP) tools. However, many analyses rely on different grammatical assumptions (e.g., Pearl and Sprouse, 2013; Szubert et al., 2024; Liu and Prud'hommeaux, 2021; Gretz et al., 2015; Sagae et al., 2007), and therefore adopt divergent annotation frameworks or standards. While most existing annotations use syntactic dependencies—in part due to the relative simplicity of annotation and parsing and the growing adoption of the Universal Dependencies (UD) framework (Nivre et al., 2016, 2020)—annotation practices remain inconsistent across datasets. This is largely due to the lack of a unified guideline for annotating children's speech, which presents unique challenges not fully addressed by existing UD documentation.

As UD treebanks have become valuable resources in both NLP (e.g., Jumelet et al., 2025; Opitz et al., 2025) and language acquisition research (e.g., Clark et al., 2023; Hahn et al., 2020), there have been increasing efforts to parse CHILDES data using tools such as `stanza` (Liu and MacWhinney, 2024). However, the resulting annotation quality is often inconsistent and cannot be guaranteed. In this paper, we compile, harmonize, and manually correct major UD-style annotations of CHILDES data into a consistent, unified UD format, resulting in a gold-standard treebank of 48K sentences and 236K tokens (including, e.g., the tree in Figure 1). In addition, we construct a larger silver-standard treebank of 1M sentences and 6M tokens produced by `stanza`[1] and report parser accuracy estimates. We publicly release both datasets.[2]

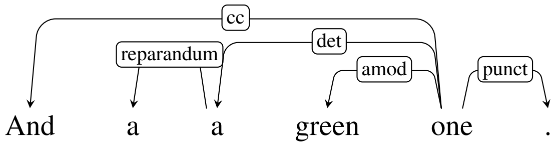


Figure 1: UD tree for a child utterance from Lily (Providence corpus, sentID=16916280)

2 Related Work

2.1 CHILDES Corpora

CHILDES is a collection of child–adult conversations recorded in naturalistic or laboratory settings. It has played a central role in both language acquisition research and the development of NLP tools. In addition to specialized corpora—such as clinical datasets (Gillam and Pearson, 2004), naturalistic family interactions (Gleason, 1980), and controlled laboratory studies (Newman et al.,

---

[1] `stanza 1.9.2 (combined model)`

[2] Official gold UD release: `https://github.com/UniversalDependencies/UD_English-CHILDES` Note: Due to a postprocessing error, the gold UD release from the `main` branch is missing approximately 10K sentences. For complete access to the data, please use the `dev` branch. The `main` branch will be updated in the next official release scheduled for November 2025.
Silver release: `https://github.com/xiulinyang/UD-CHILDES`.

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 52–58
August 27, 2025 ©2025 Association for Computational Linguistics

Corpus	**Children**	**Speakers**	**Trees**	**UPOS**	**Feats**	**Utterances**		**Tokens**	
						Gold	**Silver**	**Gold**	**Silver**
S+24	Adam	Adults	Gold	Gold	Convrtd	17,233	0	91,114	0
LP21	Eve	All	Gold	Silver	Silver	2,207	0	8,497	0
LP23	10 Children	All	Gold	Silver	Silver	34,530	0	168,284	0
UD-English-CHILDES	11 Children	All	Gold/Silver	Gold/Silver	N/A	**48,183**	**1,197,471**	**236,941**	**6,892,314**

Table 1: Overview of CHILDES-based UD treebanks compiled in this paper and our newly-released UD-English-CHILDES treebank. Source corpus labels (S+24, LP21, LP23) are defined in §3. Note that there is overlap in the Adam data: S+24 figures are counts from the original dataset; for our version, these were filtered to avoid duplicates and merged with corresponding LP23 utterances. The heading **Gold** refers to the subset of utterances for which trees and UPOS have been manually corrected according to the UD v2 framework; **Silver** refers to the subset with fully automatic annotations from `stanza`.

Child	**Corpus**	**Child age range**	**Gold sents**	**Gold toks**	**Silver sents**	**Silver toks**
Laura	Braunwald (Braunwald, 1971)	1;3-7;0 (1;3-7;0)	4,622	21,079	41,862	205,427
Adam	Brown (Brown, 1973)	1;6-5;2 (1;6-5;2)	16,736	84,643	93,315	452,348
Eve	Brown	1;6-5;1 (1;6-5;2)	2,207	8,497	108,044	532,319
Abe	Kuczaj (Kuczaj, 1977)	2;4-5;0 (2;4-5;0)	4,167	22,437	38,630	230,489
Sarah	Brown	1;6-5;2 (1;6-5;2)	5,347	23,233	104,926	517,654
Lily	Providence (Demuth et al., 2006)	0;11-4;0 (0;11-4;0)	1,499	6,337	79,573	422,245
Naima	Providence	1;3-3;11 (0;11-4;0)	2,534	14,360	236,350	1,422,543
Violet	Providence	0;11-4;0 (0;11-4;0)	721	1,857	32,801	164,975
Thomas	Thomas (Lieven et al., 2009)	2;0-4;11 (2;0-4;11)	4,240	20,333	313,550	2,039,132
Emma	Weist (Weist and Zevenbergen, 2008)	2;2-4;10 (2;1-5;0)	2,423	13,730	74,825	474,460
Roman	Weist	2;2-4;9 (2;1-5;0)	3,653	20,557	73,595	467,633

Table 2: Detailed statistics for each child, including counts of gold and silver annotations and their corresponding age ranges in months. Ages in the silver corpus are shown in parentheses. For source corpus URLs see Appendix A.

2016)—CHILDES supports a wide range of approaches to developmental linguistics. Many of its corpora inform foundational theories of language acquisition, particularly the poverty of the stimulus hypothesis (Chomsky, 1976). Researchers frequently use child-directed speech from CHILDES to quantify the distribution of linguistic structures that are central to these theories, such as wanna contraction (Getz, 2019), anaphoric one (Foraker et al., 2009; Pearl and Mis, 2011), auxiliary fronting (Perfors et al., 2011), and syntactic islands (Pearl and Mis, 2011). It has also been used in computational models of language acquisition (e.g., Abend et al., 2017).

CHILDES has also emerged as a valuable resource for NLP tool benchmarking and language model pretraining. Following the work of Huang (2016), studies such as Liu and Prud'hommeaux (2023) have highlighted the challenges faced by UD parsers when applied to child-directed speech, showing substantial performance gaps compared to adult data. CHILDES also supports recent research on pretraining dynamics (Feng et al., 2024) and the development of efficient language models, including in initiatives like the BabyLM Challenge (Choshen et al., 2024; Charpentier et al., 2025).

2.2 Spoken Language Treebanks

Overview The development of UD project has fostered the development of spoken language annotations across a wide variety of languages, such as Beja (Kahane et al., 2021) and Japanese (Omura et al., 2023), as documented in Dobrovoljc (2022). For English, the GUM corpus (Zeldes, 2017) incorporates several spoken genres.

CHILDES Dependency Treebanks Early dependency parsing research on English CHILDES data utilized a custom inventory of grammatical relations (GR; Sagae et al., 2004, 2005). These gradually evolved to address CHILDES-specific challenges (Sagae et al., 2007), and were applied to the entire English CHILDES corpus using a supervised parser (Sagae et al., 2010).

More recently, UD-style annotations have been introduced to CHILDES. Liu and MacWhinney (2024) release an automatically parsed version of the English CHILDES corpus, annotated with UD trees using `stanza`. Liu and Prud'hommeaux (2021) used a semi-automatic method to convert previous GR-based annotations into UD trees, focusing on child-produced speech (ages 18–27 months) from the Eve data within the Brown corpus (Brown, 1973). Subsequently, Szubert et al. (2024)

```
# sent_id = 22497 (normalized sentence ID across corpora; used to avoid
                  collisions since some corpora share identical sentence IDs)
# original_sent_id = 946255 (original sentence ID from the corpus, as assigned
                              in childsr)
# childes_toks = who's that (original token string from childsr)
# child_name = Abe
# corpus_name = Kuczaj
# gold_annotation = True
# speaker_age = 43.72369042485472 (child's age in months)
# speaker_gender = male (child's gender)
# speaker_role = Father (speaker role in conversation)
# type = question (sentence type annotation)
# text = Who's that?
1-2   Who's _     _     _     _     _     _     _     _
1     Who   who   PRON  WP    _     0     root  0:root      _
2     's    be    AUX   VBZ   _     1     cop   1:cop _
3     that  that  PRON  DT    _     1     nsubj 1:nsubj     SpaceAfter=No
4     ?     ?     PUNCT ?     _     1     punct 1:punct     _
```

Figure 2: Example of a gold-annotated CoNLL-U sentence from the CHILDES-Providence corpus, with added parenthetical explanations of sentence-level metadata. Enhanced UD (EUD) relations are added deterministically by the script at `https://github.com/amir-zeldes/gum/blob/master/_build/utils/eng_enhance.ini`.

developed gold-standard UD annotations by automatically transforming GR annotations and manually correcting them. Their dataset includes child-directed speech from the Adam data of the Brown corpus and the Hebrew Hagar corpus (Berman, 1990), addressing spoken-language-specific phenomena such as repetitions and non-standard vocabulary, as well as a mapping to semantics.

Building upon these efforts, Liu and Prud'hommeaux (2023) significantly expanded UD annotations to cover utterances from 10 children aged 18–66 months (Adam from the Brown corpus as well as 9 children from other corpora), incorporating both child and caregiver speech. Their work tackles complex spoken-language features, including speech repairs and restarts.

Although Liu and Prud'hommeaux (2021, 2023) provide manually corrected UD trees, their annotations are inconsistent with the UD v2 framework, lack Universal Part-of-Speech (UPOS) tags, and have not been independently verified. Szubert et al. (2024) offer verified data, but they follow the UD v1 annotation guidelines. To date, there is no official UD release for CHILDES speech data.

3 Annotations

3.1 Data Source & Statistics

This work leverages three existing UD treebanks: Szubert et al. (2024) (henceforth **S+24**), Liu and Prud'hommeaux (2021) (**LP21**), and Liu and Prud'hommeaux (2023) (**LP23**), summarized in Table 1. As these treebanks were already annotated, our human annotation efforts focused primarily on correcting errors and harmonizing annotations across corpora. We present post-compilation statistics in Tables 1 and 2. Table 1 summarizes the full corpus and its source contributions, and Table 2 provides per-child statistics.

In the official UD release, we divide the corpus based on the children's names and genders. The training and dev splits (90% and 10%, respectively) are constructed from the data of Adam, Lily, Naima, Sarah, Roman, Laura, and Abe. The corpora of Eve, Violet, Emma, and Thomas are reserved for the test split. Details are reported in Table 3.

3.2 Annotation Pipeline

Following Liu and Prud'hommeaux (2023), we collect CHILDES corpora using the R package `childesr` (Sanchez et al., 2019).[3] Sentence normalization can be found in the paper. As the data from LP21 and LP23 are only parsed but not tagged yet, sentences with existing dependency annotations are identified and automatically tagged with UPOS using `stanza` (Qi et al., 2020), while unannotated sentences are assigned both UPOS and dependency trees. Our current work focuses on correcting previously human-annotated data. To ensure conformity with UD guidelines, we run all processed sentences through the UD validation tool[4] and manually fix those that fail validation. The correction work is performed by three linguistics graduate students trained in UD annotation. In

---

[3] `https://langcog.github.io/childes-db-website/`

[4] `https://github.com/UniversalDependencies/tools/blob/master/validate.py`

Split	Children	Corpus	Gold Sents
Train	Adam, Lily, Naima, Sarah, Roman, Laura, Abe	Brown, Providence, Weist, Kuczaj, Braunwald	34,732
Dev	Adam, Lily, Naima, Sarah, Roman, Laura, Abe	Brown, Providence, Weist, Kuczaj, Braunwald	3,860
Test	Eve, Violet, Emma, Thomas	Brown, Providence, Weist, Thomas	9,591

Table 3: Data splits for the official UD_English-CHILDES with associated children, corpora, and gold-standard sentence counts.

total, we made approximately 8,000 corrections.

Many of the errors stem from mismatches between UPOS tags and dependency labels (as LP21 and LP23 used automatic UPOS tagging). In addition, we address format issues such as multiword tokens, spacing mismatches (e.g., `SpaceAfter`), and deprecated dependency relations not supported by current UD guidelines (e.g., `compound:svc`, `obl:about_like`, `nmod:over_under`). The 5 most common linguistic issues were as follows:

`advmod` tagged as `ADP` This error commonly appears with phrasal verbs such as *get up* and *take over*. The original annotation assigns `advmod` as the dependency relation to phrasal verbs with POS tag `ADP`. We revise these to `compound:prt`, in accordance with the UD treatment of phrasal particles.

Auxiliaries tagged as `VERB` or `PART` Auxiliaries such as *be* and *have* are frequently misclassified as main verbs or particles. In some cases, lemmas are also mislabeled—most notably, the lemma of contracted forms like *'s* is incorrectly assigned as *'s* rather than the appropriate auxiliary *be*. We correct both the POS and lemma annotations in these cases.

Lexical items tagged as `PUNCT` The `stanza` parser often mislabels disfluent word fragments in spontaneous speech as punctuation marks (e.g., *OK/INTJ Adam/PROPN ride/VERB dat/**PUNCT** ./PUNCT*. We reassign these tokens appropriate UPOS labels based on context and speaker intent, often as interjections.

Determiner misrecognition Ambiguous or reduced forms of determiners—such as *de* —are frequently misidentified as proper nouns (`PROPN`). We manually review these cases and reannotate them as `DET` when appropriate.

Function word heads with dependents In previous treebanks, words appearing in functional relations such as `case`, `mark`, and `aux` have been assigned children, which violates UD's constraint that these words should be leaf nodes. We reassign the erroneous dependents to the appropriate content heads, ensuring the structure conforms to UD's projectivity and function word constraints.

3.3 Harmonization

Each treebank follows its own annotation guidelines, which are largely based on UD but not fully compliant. We performed a series of normalization steps to harmonize them into a consistent format. Our unified format is primarily based on LP23, with several adaptations described below.

Metadata In our normalized CoNLL-U files, we include the following metadata fields with an example provided in Figure 2: `sent_id` (normalized sentence IDs); `original_sent_id` (utterance ID retrieved via the `childesr` R package); `childes_toks` (tokenized utterance); `corpus_name` (original corpus name); `gold_annotation` (indicates whether the sentence is manually annotated); `speaker_gender`, `speaker_role`, and `speaker_age` (speaker/child metadata); `text` (the text aligned with the tree), and `type` (sentence type). Table 4 summarizes the distribution of the main sentence types and compares them with those in the UD 2.15 release of GUM (Zeldes, 2017), a multi-genre English corpus. Notably, questions occur in the CHILDES conversations at a much higher rate—they are nearly half (45%) as frequent as declarative utterances, as opposed to 9% in GUM.

Type	CHILDES			GUM
	CDS	CS	Overall	Overall
declarative	16,112	15,884	31,996	7,695 (`decl`)
question	2,882	11,413	14,295	716 (`q`, `wh`)
imperative emphatic	509	288	797	1,326 (`imp`, `intj`)
others	601	494	1095	2,409

Table 4: Sentence type counts in gold CHILDES and GUM corpora. **Question** includes *question*, *self interruption question*, *trail off question*, and *interruption question*. **Others** encompasses less frequent categories: *trail off*, *interruption*, *self interruption*, and *quotation next line*.

Punctuation To bring the transcripts in line with written English conventions, we capitalize the first word of each utterance and infer sentence-final

Metrics	Children's speech	Parents' speech	Overall
LAS	81.2	86.3	84.2
UAS	87.2	91.0	89.5

Table 5: LAS and UAS scores for children's speech, parents' speech, and overall performance.

punctuation at the end of each sentence based on the sentence type provided in the metadata.[5]

Reparandum Each of the three treebanks defines its own subtypes for the `reparandum` and `parataxis` relations. For example, S+24 includes labels such as `parataxis:repeat` not present in the current UD guidelines. Similarly, LP21 and LP23 annotate `reparandum` with subtypes such as `restart` and `repetition` to mark special utterance features of children's speech. To ensure consistency across treebanks, we move all such subrelation information to the `MISC` column.

Others As S+24 was annotated using the UD guidelines version 1.0, we convert the annotation using UD version 2.0 with a script[6] and manual annotation. For example, we shifted the head-dependent direction of `flat` in the annotations.

Since S+24 and LP23 overlap in the Adam corpus, we merged the annotations from these two treebanks. 3375 sentences are repetitive in S+24. We removed these sentences from our corpus.[7]

To ensure a more linguistically plausible analysis, we also diverged from Liu and Prud'hommeaux (2023) in our treatment of interjections. Instead of annotating utterances consisting solely of interjections (e.g., *Ha ha ha ha*) as `conj`, we used the `flat` relation.

3.4 Silver Data Assessment

To create silver-standard annotations, we apply `stanza` to the utterances that were not sampled by the previous treebanks (but were from the same CHILDES datasets, i.e. conversations involving the 11 children in Table 2). To estimate the quality of these silver annotations, we evaluate the parser's performance on the gold-standard data. We report Labeled and Unlabeled Attachment Scores (LAS/UAS) in Table 5. The parser achieves an overall LAS of 83.3. Performance is higher on parents' speech (86.3 LAS) than on children's speech (81.2 LAS), likely due to the greater syntactic regularity and lower frequency of disfluencies in adult utterances. The overall high-quality data can be more easily verified by human annotators than annotated from scratch. It also provides valuable training data for improving parsers on spoken language.

4 Conclusion & Future Work

In this paper, we present the first harmonized UD treebanks for CHILDES, covering 11 corpora and over 48k sentences from both child-directed and child-produced speech. The three datasets we compiled do not preserve conversational structure, and as a result, the finalized gold-standard treebank lacks coherent dialogue sequencing. Preserving such structure would require additional manual annotation to make sure all sentences are gold. However, since our annotations include the `original_sent_id` field, reconstructing the conversation structure is straightforward. Furthermore, morphological features have not been annotated or independently verified. Future work will focus on further corrections to the silver-standard data and the continued expansion of the treebanks. We welcome collaboration on this ongoing effort.

Acknowledgments

We acknowledge Ida Szubert, Omri Abend, Samuel Gibbon, Louis Mahon, Sharon Goldwater, Mark Steedman, and Emily Prud'hommeaux for their contributions to the original UD treebanking efforts. We also thank Brian MacWhinney for helpful discussions and anonymous reviewers for their suggestions.

References

Omri Abend, Tom Kwiatkowski, Nathaniel J. Smith, Sharon Goldwater, and Mark Steedman. 2017. Bootstrapping language acquisition. *Cognition*, 164:116–143.

Ruth A. Berman. 1990. On acquiring an (S)VO language: subjectless sentences in children's Hebrew. *Linguistics*, 28(6):1135–1166.

Susan R Braunwald. 1971. Mother-child communication: the function of maternal-language input. *Word*, 27(1-3):28–50.

Roger Brown. 1973. *A first language: The early stages*. Harvard University Press.

---

[5] The original data transcribes various kinds of prosodic information such as pauses. At present we do not retain this information or attempt to infer corresponding punctuation like commas and parentheses.

[6] `https://github.com/UniversalDependencies/tools/tree/master/v2-conversion`

[7] 883 sentences from S+24 could not be merged because S+24 and LP21 are using different data sources, and were therefore removed from our treebank as well.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Hu, Jaap Jumelet, Tal Linzen, Jing Liu, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Wilcox, and Adina Williams. 2025. BabyLM turns 3: Call for papers for the 2025 BabyLM workshop. *arXiv preprint arXiv:2502.10645*.

Noam Chomsky. 1976. *Reflections on language*. Temple Smith London.

Leshem Choshen, Ryan Cotterell, Michael Y Hu, Tal Linzen, Aaron Mueller, Candace Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. 2024. [Call for papers] the 2nd BabyLM challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2404.06214*.

Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger Levy. 2023. A cross-linguistic pressure for Uniform Information Density in word order. *Transactions of the Association for Computational Linguistics*, 11:1048–1065.

Katherine Demuth, Jennifer Culbertson, and Jennifer Alter. 2006. Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language and speech*, 49(2):137–173.

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.

Steven Y. Feng, Noah Goodman, and Michael Frank. 2024. Is child-directed speech effective training data for language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22055–22071, Miami, Florida, USA. Association for Computational Linguistics.

Stephani Foraker, Terry Regier, Naveen Khetarpal, Amy Perfors, and Joshua Tenenbaum. 2009. Indirect evidence and the poverty of the stimulus: The case of anaphoric *one*. *Cognitive Science*, 33(2):287–300.

Heidi R Getz. 2019. Acquiring *wanna*: Beyond Universal Grammar. *Language Acquisition*, 26(2):119–143.

Ronald Bradley Gillam and Nils A Pearson. 2004. *Test of narrative language*. Pro-ed Austin, TX.

Jean Berko Gleason. 1980. The acquisition of social speech routines and politeness formulas. In *Language*, pages 21–27. Elsevier.

Shai Gretz, Alon Itai, Brian MacWhinney, Bracha Nir, and Shuly Wintner. 2015. Parsing Hebrew CHILDES transcripts. *Language Resources and Evaluation*, 49:107–145.

Michael Hahn, Dan Jurafsky, and Richard Futrell. 2020. Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, 117(5):2347–2353.

Rui Huang. 2016. An evaluation of POS taggers for the CHILDES corpus. *CUNY Academic Works*.

Jaap Jumelet, Leonie Weissweiler, and Arianna Bisazza. 2025. MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2504.02768*.

Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 2021. A morph-based and a word-based treebank for Beja. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 48–60, Sofia, Bulgaria. Association for Computational Linguistics.

Stan Kuczaj. 1977. The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, 16(5):589–600.

Elena Lieven, Dorothé Salomo, and Michael Tomasello. 2009. Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20(3):481–507.

Houjun Liu and Brian MacWhinney. 2024. Morphosyntactic analysis for CHILDES. *Language Development Research*, 4(1).

Zoey Liu and Emily Prud'hommeaux. 2021. Dependency parsing evaluation for low-resource spontaneous speech. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 156–165, Kyiv, Ukraine. Association for Computational Linguistics.

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.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press.

Rochelle S Newman, Meredith L Rowe, and Nan Bernstein Ratner. 2016. Input and uptake at 7 months predicts toddler vocabulary: the role of child-directed speech and infant processing skills in language development. *Journal of Child Language*, 43(5):1158–1173.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 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, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.

Mai Omura, Hiroshi Matsuda, Masayuki Asahara, and Aya Wakasa. 2023. UD_Japanese-CEJC: Dependency relation annotation on corpus of everyday Japanese conversation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 324–335, Prague, Czechia. Association for Computational Linguistics.

Juri Opitz, Shira Wein, and Nathan Schneider. 2025. Natural language processing RELIES on linguistics. *Computational Linguistics*, pages 1–23.

Lisa Pearl and Benjamin Mis. 2011. How far can indirect evidence take us? Anaphoric *one* revisited. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Lisa Pearl and Jon Sprouse. 2013. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, 20(1):23–68.

Amy Perfors, Joshua B. Tenenbaum, and Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. 2007. High-accuracy annotation and parsing of CHILDES transcripts. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 25–32, Prague, Czech Republic. Association for Computational Linguistics.

Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. 2010. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37(3):705–729.

Kenji Sagae, Alon Lavie, and Brian MacWhinney. 2005. Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 197–204, Ann Arbor, Michigan. Association for Computational Linguistics.

Kenji Sagae, Brian MacWhinney, and Alon Lavie. 2004. Adding syntactic annotations to transcripts of parent-child dialogs. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Alessandro Sanchez, Stephan C. Meylan, Mika Braginsky, Kyle E. MacDonald, Daniel Yurovsky, and Michael C. Frank. 2019. childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, 51(4):1928–1941.

Ida Szubert, Omri Abend, Nathan Schneider, Samuel Gibbon, Louis Mahon, Sharon Goldwater, and Mark Steedman. 2024. Cross-linguistically consistent semantic and syntactic annotation of child-directed speech. *Language Resources and Evaluation*.

Richard M Weist and Andrea A Zevenbergen. 2008. Autobiographical memory and past time reference. *Language Learning and Development*, 4(4):291–308.

Amir Zeldes. 2017. The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

A Sources of the Coprora

In this work, we include the sources from the following corpora:

- `https://childes.talkbank.org/access/Eng-NA/Braunwald.html`
- `https://childes.talkbank.org/access/Eng-NA/Brown.html`
- `https://childes.talkbank.org/access/Eng-NA/Kuczaj.html`
- `https://phon.talkbank.org/access/Eng-NA/Providence.html`
- `https://childes.talkbank.org/access/Eng-UK/Thomas.html`
- `https://childes.talkbank.org/access/Eng-NA/Weist.html`

A UD Treebank for Bohairic Coptic

Amir Zeldes
Georgetown University
amir.zeldes@georgetown.edu

Nina Speransky
Hebrew University of Jerusalem
gitchinama@gmail.com

Nicholas Wagner
Duke University
nicholas.wagner@duke.edu

Caroline T. Schroeder
University of Oklahoma
ctschroeder@ou.edu

Abstract

Despite recent advances in digital resources for other Coptic dialects, especially Sahidic, Bohairic Coptic, the main Coptic dialect for pre-Mamluk, late Byzantine Egypt, and the contemporary language of the Coptic Church, remains critically under-resourced. This paper presents and evaluates the first syntactically annotated corpus of Bohairic Coptic, sampling data from a range of works, including Biblical text, saints' lives and Christian ascetic writing. We also explore some of the main differences we observe compared to the existing UD treebank of Sahidic Coptic, the classical dialect of the language, and conduct joint and cross-dialect parsing experiments, revealing the unique nature of Bohairic as a related, but distinct variety from the more often studied Sahidic.

1 Introduction

1.1 Coptic

Coptic was the indigenous spoken and written language of Egypt during the Late Roman, Byzantine, and early Islamic periods. As the final stage of the Ancient Egyptian branch of the Afro-Asiatic language family, Coptic concludes a linguistic tradition with the longest continuous written record in human history, which includes three millennia of Hieroglyphic, Hieratic and Demotic Egyptian writing, as well as over a millennium of writing in Coptic itself, a form of the same language written mainly in Greek letters.

Initially a very low resource language, recent efforts to digitize and annotate data for Coptic have resulted in now substantial resources for Sahidic Coptic, the classical dialect of the language, with corpora (Schroeder and Zeldes, 2020), a machine readable dictionary (Feder et al., 2018) and a UD dependency treebank (Zeldes and Abrams, 2018). At the same time, other forms of Coptic, namely dialects beyond Sahidic, continue to have little or no annotated resources. In this paper, we aim to address this gap by introducing a UD treebank for a second, very significant dialect of the language: Bohairic Coptic. In this section we offer a brief summary of Coptic and its dialects, highlighting especially some of the main differences between the classical Sahidic, and Bohairic Coptic, which we explore in more detail later on using our data.

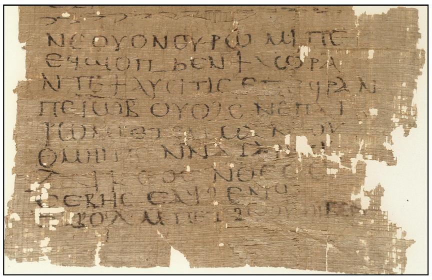


Figure 1: Excerpt from a papyrus containing Bohairic Job 1:1 (P.Mich.inv. 926 recto, TM no. 107875). Image: University of Michigan Library Digital Collections.

Geographic diversity in late ancient Egypt resulted in a range of regional dialects rather than one standardized form of Coptic. Scholars identify six principal dialects, but two of these, Sahidic and Bohairic, were the most influential (Kasser, 1991). Sahidic dominated as the literary language from the third to the ninth century CE before being gradually replaced by Bohairic, a northern variety that continues to serve today as a heritage and liturgical language among Coptic communities in Egypt and the diaspora. Despite the diversity of surviving evidence for Coptic dialects, print and digital resources for the language have remained limited and have focused almost exclusively on Sahidic. The textual evidence for the dialect of Bohairic, though sizeable (and primarily literary), is accessible mostly in facsimiles (see Figure 1 for a papyrus manuscript example) or in older print publications, and many works remain unpublished (Shisha-Halevy, 2007). Using terms from Joshi et al. (2020), Bohairic is at rank 0 of the language technology hierarchy, a 'left-behind' language.

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 59–69
August 27, 2025 ©2025 Association for Computational Linguistics

1.2 Bohairic and Sahidic

Expanding digital corpora to include Bohairic texts presents substantial challenges, as existing tools for even basic preprocessing operations, such as word segmentation and lemmatization, let alone syntactic parsing, require tools trained on dialect-specific data (see Section 5 for some evaluation). Although Sahidic and Bohairic are dialectal manifestations of a single language – sharing a broadly consistent grammatical architecture and much of their lexical inventory, their phonological, orthographic and morphosyntactic systems diverge in ways that make NLP tools trained on Sahidic unsuitable for processing Bohairic texts.

At the most elementary level, the two dialects diverge already in their orthographic systems. Unlike earlier Egyptian – written in hieroglyphic and hieratic scripts over the preceding three millennia – Coptic adopted a modified version of the Greek alphabet, with additional characters of Demotic origin (ultimately derived from hieroglyphs) to represent sounds absent from Greek. Both dialects make use of six such additional letters, including the letter hore (ϩ, Unicode U+03E9 lowercase / U+03E8 uppercase) for the voiceless glottal fricative /h/, but Bohairic distinguishes hore /h/ from khei /x/ (ⳉ , Unicode U+2CC9 lowercase / U+2CC8 uppercase). This distinction is semantically consequential – while the Sahidic word ⲉϩⲣⲁⲓ (ehrai) can confusingly mean both 'up' or 'down', and can only be disambiguated in context, in Bohairic we see distinct forms that had merged in Sahidic: ⲉⳉⲣⲏⲓ (exrēi) 'down' and ⲉϩⲣⲏⲓ (ehrēi) 'up'.

As an agglutinative language, Coptic combines multiple morphosyntactic elements into units known as bound groups. Following modern editorial conventions, these groups are defined by the presence of a single stressed lexical item at their core (Layton, 2011). Material that would normally be tokenized into separate words in annotated corpora often appears conflated into a single space-delimited string in Coptic. Such fusion is common, for example, in noun phrases or prepositional phrases, as in (1), or in auxiliaries and clitics attaching to verbs, as in (2), both examples in the Bohairic dialect.

(1) ϩⲓⲧⲉⲛ-ⲫ-ⲟⲩⲱϣ ⲙ-ⲫ-ⲛⲟⲩϯ
hiten-p^{h}-wōš m-p^{h}-nuti
by-the-will of-the-god
"by the will of God"

(2) ⲁ-ⲥ-ⲑⲣⲉ-ϥ-ⲛⲁϩⲙ-ⲟⲩ
a-s-t^{h}re-f-nahm-u
PAST-3.SG.F-CAUS-3.SG.M-hear-3.PL
"She made him save them"

In this paper we use hyphens to split such space-delimited word forms into tokens, which correspond to units that can receive independent parts of speech, such as nouns and verbs, articles and prepositions, etc. Note that such separators do not exist in source texts and are represented in the treebank directly through the tokenization. For words containing derivational affixes as in (3), and compound nouns or verbs (4), we additionally provide a segmentation into component parts in an additional annotation called MSEG in the conllu format MISC field, following existing practices in UD treebanks.[1]

(3) ⲙⲉⲧⲁⲧϫⲱⲛⲧ
MSeg=ⲙⲉⲧ-ⲁⲧ-ϫⲱⲛⲧ
met-at-čōnt
less-ness-anger
"angerlessness"

(4) ⲉⲣϩⲁⲗ
MSeg=ⲉⲣ-ϩⲁⲗ
er-hal
do-service
"serve"

Such words are easy to recognize since they have distinct, recurring forms (known affixes, special reduced forms of verbs with incorporated objects), lack internal syntactic markers (e.g. incorporated objects like *hal* in (4) appear without articles or other modifiers, the verb 'ire' meaning "do" is reduced to 'er') and are considered single nouns or verbs in terms of parts of speech, as well as single dictionary entries in terms of lemmatization. Due to these multiple levels of complexity, bound groups must first be analyzed and segmented before we can digitize Coptic texts in a way that allows for searchability. Each token can then be lemmatized and tagged to enable structured querying and lexical lookup, or linking to resources such as the Coptic Dictionary Online (CDO, Feder et al. 2018).

2 Previous work

The Bohairic UD treebank joins a growing body of typologically diverse languages analyzed using the UD framework (de Marneffe et al., 2021), in-

---

[1] See `https://universaldependencies.org/misc.html#mseg`

cluding recent treebanks of related Afro-Asiatic languages such as Biblical and Modern Hebrew (Swanson and Tyers, 2022; Zeldes et al., 2022), Arabic (Taji et al., 2017), and very recently, Ancient Egyptian (Díaz Hernández and Carlo Passarotti, 2024), resources we consider in the development of comparable annotation guidelines (and to a lesser extent, treebanks converted to UD, e.g. for Hausa, Caron 2015).

The most important previous resource we model our work on is the existing UD treebank for Sahidic Coptic (Zeldes and Abrams, 2018), which contains around 60K words from a range of works in a number of genres. In particular, the Sahidic treebank contains some Biblical material which is in part also available in Bohairic (see Section 3 below). By selecting the same Biblical books and chapters where possible, we are able to conduct some direct comparisons between the dialects which only target parallel passages (see Section 4). At the same time, for texts in the Sahidic treebank that are unavailable in Bohairic, we select substitutes from similar genres, offering a similar range of language usage.

In terms of annotation scheme, we closely follow and adapt to Bohairic the Coptic Scriptorium guidelines for annotating parts of speech, lemmatization, sentence splitting and UD dependency relations. In a recent paper, Crellin (2025) criticizes the choice of UD as a treebanking framework, among other languages for Coptic, as unmotivated, stating "no overt discussion of the rationale for choosing [...] Universal Dependencies" could be found (Crellin, 2025, 100). We would therefore like to explicitly motivate the choice of UD for the Bohairic Treebank beyond the obvious benefit of comparability with other resources (see Section 4), and outline some of the key decisions in treebanking Coptic.

The most crucial decision we follow is treatment of lexical verbs as heads of their clauses, despite their etymology in many tenses in earlier Egyptian as subordinated, nominalized infinitives. For example, the verb "hear" in (5) is the only possible head for the clause, which otherwise contains only the subject. Meanwhile in (6), the past auxiliary 'a' in 'a-f-sōtem' would have been analyzed as the head in earlier Late Egyptian, as the construction is derived from a periphrasis, Late Egyptian *jr⸗f sḏm*, lit. "did-he hear", or more freely, "he did hearing".

(5) ϥ-ⲥⲱⲧⲉⲙ
f-sōtem
he-hear
"he hears"

(6) ⲁ-ϥ-ⲥⲱⲧⲉⲙ
a-f-sōtem
PST-he-hear
"he heard"

We motivate the choice of the lexical verb as the head by the basic UD principle of lexico-centrism, which also provides a more parallel analysis for subjects in the durative and past tenses shown above. Since Coptic, synchronically an agglutinative language, has a broad range of tenses and constructions, but only a handful of etymological sources for agglutinative morphemes (almost always forms of the Late Egyptian verbs for "do", "give" or "know"), choosing a non-lexico-centric scheme would result in an analysis in which Coptic has only a handful of distinct verbs. The choice of UD is therefore quite conscious, and especially given the benefits of comparability, without obvious superior alternatives.

We also match our native Coptic part of speech tagging scheme and its mapping to Universal POS tags to those used in the Sahidic treebank, as well as using Multiword Tokens (MWTs) to represent bound groups in the conllu format (i.e. examples (5) and (6) would be one MWT each, containing two and three word forms respectively).

3 Data

The textual data in the Bohairic UD corpus consists of selections from works in multiple genres. We include selections from hagiography (saints' lives) of two prominent figures in early Christian Egypt: Shenoute, the leader of a federation of male and female monasteries in the fourth-fifth centuries; and Isaac, Patriarch (or Coptic Pope) of Alexandria from 686 to 689 CE. The Life of Shenoute is a compilation of panegyric works about Shenoute written in Bohairic Coptic over decades after his death and compiled into the genre of a saint's life (see Lubomierski 2008 and Berno 2019c). The Life of Isaac is a saint's life written in Bohairic Coptic possibly in the seventh century after Isaac's death, focusing mostly on his adult life as a monk, priest, and patriarch during the early Islamic period (Berno, 2019b). The Lausiac History is a narrative text originally written in Greek consisting of anecdotes about fourth-century monks and so has hagiographical elements as well as generic elements from travel narratives (Berno, 2019a). Three biblical texts are also translations from Greek: the Gospel of Mark and 1 Corinthians from the New

source	genre	chapters	tokens	sentences
1 Corinthians	Biblical Epistle	1–7	4,789	164
Gospel of Mark	Biblical Narrative	1–9	11,091	373
Book of Habbakuk	Poetic	1–3	1,988	56
Life of Shenoute	Hagiography	1–26	4,970	148
Life of Isaac	Hagiography	1–19	5,433	143
Lausiac History	Hagiography	1–16	4,453	117
Total:			32,724	1,001

Table 1: Data in the Bohahiric Treebank.

Testament, and the Christian Old Testament book of Habakkuk. Bohairic Habakkuk is likely a translation from the Septuagint (the Greek version of the Hebrew Bible). Mark is a gospel or ancient biography, and 1 Corinthians is a letter by Paul the apostle, meaning our Biblical data also spans multiple genres.

The text of each of these works is derived from previous digital or print editions. The *Life of Shenoute* comes from a digital edition created by Hany Takla of the St. Shenouda the Archimandrite Coptic Society, based primarily on the edition published by Leipoldt (1906). Dr. Lydia Bremer-McCollum of the Coptic Scriptorium project[2] extracted digitized text using optical character recognition (OCR) from the public domain edition of the *Life of Isaac* and the public domain edition of the Bohairic *Lausiac History*, both edited by Amélineau (1887, 1890), followed by manual correction of OCR errors. The Gospel of Mark and 1 Corinthians digital texts come from the Marcion Project;[3] both ultimately derive from the print edition of Horner (1905). The text of Habakkuk is from the working digital edition in-progress created by the Göttingen Coptic Old Testament project.[4] All of the digital versions were processed by the Coptic Scriptorium project's natural language processing tools (Zeldes and Schroeder, 2016), which have normalized the text, including normalizing variant spellings and expanding any abbreviations; for this paper, all NLP processing has been manually checked by one or more of the authors before treebanking.

3.1 Inter-annotator agreement

To evaluate the quality of our annotations, we double-annotated 166 sentences (6,207 tokens) from two different texts, the Life of Shenoute and the Lausiac History. Table 2 shows the results for Cohen's κ and mutual F1 score, for both dependency relations (labels) and dependency heads. In order to avoid inflating κ due to a large range of possible labels for heads (i.e. all numbers attested as dependency heads), we represented heads as offsets from the position of the child (i.e. if token 37 has token 35 as its parent, we tally the value as -2). This increases the probability of chance agreement and prevents an inflated metric due to label proliferation. For dependency relations, we use the full label including subtypes (see Appendix A for details).

	Labels		**Heads**	
	Kappa	**F1**	**Kappa**	**F1**
Life of Shenoute	92.91	93.49	93.84	94.77
Lausiac History	95.53	95.78	94.61	95.53
Macro average	94.22	94.64	94.23	95.15
Micro average	94.79	95.12	94.39	95.29

Table 2: Cohen's Kappa (κ) and mutual F1 score in two texts (166 sentences) for dependency relation labels and heads, the latter represented as offsets relative to child tokens.

The results show very high agreement, substantially in excess of initial (pre-adjudication) scores in the 80s for the original version of the Sahidic treebank (Zeldes and Abrams, 2018, 199). This is likely due to the fact that annotators in this case were post-graduate researchers with substantial Coptic annotation experience, as opposed to the novice student scores reported in the Sahidic paper.

The data above constitutes the first openly available morphosyntactically annotated corpus of Bohairic, and allows for a number of quantitative comparisons with the Sahidic, to which we now turn below.

---

[2] `https://copticscriptorium.org/`

[3] `https://marcion.sourceforge.net/`

[4] `https://coptot.manuscriptroom.com/`

4 Comparing UD Bohairic and Sahidic

Like other projects adding UD treebanks in low-resource languages for which closely related languages already have treebanks (Jobanputra et al., 2024), one of our goals is to explore the ways in which Bohairic diverges from other varieties of Coptic – ideally, we would like to have treebanks of all Coptic dialects, but for the present we must focus on the comparison with Sahidic.

Dialects and closely related languages can differ in two different ways: they can have categorically distinct constructions, such as different auxiliaries, distinct argument structures for equivalent verbs etc., or they can use similar constructions but in quantitatively different usage patterns. While categorical differences between Bohairic Coptic and the better studied Sahidic are relatively well understood, quantitative differences are more elusive, but can show up in a corpus analysis.

On the lexical level, we can note that of the approximately 2,800 unique words attested in the Sahidic treebank, and around 2100 unique words in the Bohairic treebank, only about 600 are shared, and even among these, identical forms do not always translate to identical meanings. For example while **ⲉⲧ-** ‘et-’ can be a form of the relative marker in both Sahidic and Bohairic, it can also represent the precursive form meaning ‘after’ in Bohairic. Among the disjoint, dialect-specific vocabulary, many words have corresponding words in the other dialect which differ only due to pronunciation, but some words are totally unique, such as **ⲃⲁⲕⲓ** ‘baki’, which means ‘town’ in Bohairic, but does not exist in Sahidic.

On the syntactic level, we cannot find differences and commonalities as easily, but thanks to the existence of UD trees in both dialects, we can still leverage annotated data in a straightforward way. To find some of the clearest differences that our data reveals, we extract relative proportions of the dependency relations attested in the Bohairic and Sahidic Coptic treebanks, an excerpt of which is presented in Table 3, which is sorted by the ratio of frequencies.

As the table shows, striking differences are present for example in the frequencies of `dislocated` arguments (much more common in Bohairic) and `iobj` (indirect objects, much more common in Sahidic). We also note that some control labels, which we would expect to align across datasets, are quite comparable, such as `cop` for copulas, or `nsubj` for subjects.

	Sahidic		**Bohairic**		
	count	**per 1K**	**count**	**per 1K**	**ratio**
iobj	84	1.471	36	1.100	0.747
. . .					
cop	500	8.757	291	8.892	1.015
nsubj	5,549	97.185	3,275	100.073	1.029
. . .					
dislocated	889	15.569	670	20.473	1.314

Table 3: Frequencies and their ratios for some Bohairic and Sahidic dependency relations.

4.1 Subject dislocation

While all Coptic dialects use a basic SVO word order for tensed clauses with lexical verbs (Loprieno, 2000), both left and right dislocations are well attested, with left-dislocation of any argument (e.g. subject or object) bearing no special marking, as in (7).[5] Here again, the choice of UD (contra Crellin 2025) means we can easily find these, thanks to the UD label `dislocated`.

(7)

ⲡⲓ-ⳉⲁⲓ	**ⲅⲁⲣ**	**ϥ-ⳉⲱⲧⲉⲃ**
pi-sxai	gar	f-xōteb
the-scripture	for	it-kill

“For scripture, it kills” (i.e. scripture kills)

By contrast, right dislocation or extraposition is obligatorily marked for subjects using the particle **ⲛϫⲉ** ‘nče’ in Bohairic, paralleled in Sahidic by the form **ⲛϭⲓ** ‘nk^{j}i’. This particle has been analyzed as a post-verbal nominative case marker by Grossman (2015), who notes that “postverbal subjects are more frequently new referents” in Sahidic, but rarely so in Bohairic. Examples (8)–(9) demonstrate the use of the marker for right dislocation in Sahidic and Bohairic respectively:

(8)

ⲁ-ϥ-ⲃⲱⲕ	**ⲉϩⲟⲩⲛ**	**ⲉ-ⲡⲉϥ-ⲏⲓ**	**ⲛϭⲓ-ⲓⲱϩⲁⲛⲛⲏⲥ**
a-f-bōk	ehoun	e-pef-ēi	nk^{j}i-iōhannēs
PST-he-go	inside	to-his-house	PTC-John

“He went into his house, (that is) John”

(9)

ⲁ-ϥ-ϭⲓⲥⲓ	**ⲛϫⲉ-ⲫ-ⲣⲏ**
a-f-k^{j}isi	nče-ph-rē
PST-it-exalt	PTC-the-sun

“It was exalted, (that is) the sun”

One question we can immediately explore using our data and the existing UD Sahidic treebank is whether the constructions are used comparably often in the two dialects. Since segmentation, tagging

---

[5] In the following, some examples from the treebank are abbreviated for space and clarity.

and parsing guidelines match across the treebanks, we can be confident that all relevant cases can be found using equivalent searches – the results are shown in Table 4.

	matches	**total words**	**freq per 1K**
		All data	
Boh. **ⲛϫⲉ**	258	32,724	7.88
Sah. **ⲛϭⲓ**	206	57,097	3.61
		Parallel only	
Boh. **ⲛϫⲉ**	122	15,880	7.68
Sah. **ⲛϭⲓ**	66	15,619	4.22

Table 4: Bohairic and Sahidic frequencies for right dislocated subjects.

The table shows that right subject dislocation is much more common in Bohairic, with more hits in total in the smaller Bohairic treebank, and more than double the relative frequency. The bottom of the table shows results for the same queries, restricted to only the parallel data available in both corpora – these numbers are slightly less reliable since they are based on less material, but should be expected to be much closer since they derive from the same chapters of the Bible. Here too, the gap remains very substantial, despite the fact that if anything Biblical style should represent a conservative and less free or colloquial style.[6]

Looking at the Sahidic data in more detail, it becomes clear that the construction responsible for the discrepancy is nominal subjects in the canonical position between auxiliaries and verbs, as in (10), which is paralleled in Bohairic by (11) – both examples render 1 Corinthians 2:10.

(10)	**ⲁ-ⲡ-ⲛⲟⲩⲧⲉ**	**ϭⲟⲗⲡ-ⲟⲩ**	**ⲛⲁ-ⲛ**	**ⲉⲃⲟⲗ**
	a-p-noute	k^{j}olp-u	na-n	ebol
	PST-the-god	reveal-3Pl	to-1Pl	out

“God revealed them to us”.

(11)	**ⲁ-ϥ-ϭⲟⲣⲡ-ⲟⲩ**	**ⲛⲁ-ⲛ**	**ⲉⲃⲟⲗ**	**ⲛϫⲉ-ⲫ-ⲛⲟⲩϯ**
	a-f-k^{j}orp-u	na-n	ebol	nče-p^h-nuti
	PST-3Sg-reveal-3Pl	to-1Pl	out	PTC-the-god

“He revealed them to us, that is God”.

This construction is much rarer in Bohairic, and indicates that the Bohairic data represents a further step in the grammaticalization of the pronominal subject + auxiliary group, which forces subjects to be realized before or after the verbal complex. This tendency is well known from other languages in the Afro-Asiatic family, such as Hausa, which has similar subject + auxiliary structures, but can only realize a nominal subject outside of the verbal complex, with a ‘duplicate’ pronoun mirroring the subject – a pronominal TAM marker within the verbal complex is mandatory (see Crysmann 2012, 331, and Hartmann 2006 on fronting in Hausa).

4.2 Indirect objects

Following the Sahidic treebank, we use the `iobj` label primarily to indicate the possessor in the predicative possession constructions with the predicates **ⲟⲩⲟⲛ(ⲧⲉ)** “there exists” and **ⲙⲙⲟⲛ(ⲧⲉ)** “does not exist”, which are used with a subject expressive the possessum. Thus the Sahidic construction in (12) with the corresponding Sahidic form **ⲟⲩⲛⲧⲉ** represents an oblique predication “exists to your father something except sins” (or etymologically more precisely, “to the hands of your father”).

(12)	**ⲟⲩⲛⲧⲉ-ⲡⲉⲕ-ⲉⲓⲱⲧ**	**ⲗⲁⲁⲩ**	**ⲛⲥⲁ-ϩⲉⲛ-ⲛⲟⲃⲉ**
	wnte-pek-iōt	la’u	nsa-hen-nobe
	EXIST-your-father	thing	beyond-some-sin

“Does your father have anything except sins?”

Although Sahidic also prefers pronouns to nouns in the possessor position, this is more extremely the case in Bohairic. What is more, although the same constructions as in Sahidic are possible in Bohairic, the Bohairic data shows a tendency to postpone possessors to a later prepositional phrase, leaving the indirect object slot immediately after the existence predication empty. For example, we can contrast the constructions from Mark 4:9 in Bohairic (13) and Sahidic (14):

(13)	**ⲫⲏ**	**ⲉⲧⲉ-ⲟⲩⲟⲛ**	**ⲟⲩ-ⲙⲁϣϫ**	**ⲙⲙⲟ-ϥ**
	phē	ete-won	u-mašč	mmo-f
	DEM	REL-EXIST	ear	of-him

“he who has ears (to listen, let him listen)”.

(14)	**ⲡ-ⲉⲧⲉ-ⲩⲛⲧ-ϥ**	**ⲙⲁⲁϫⲉ**	**ⲙⲙⲁⲩ**
	p-ete-wnt-f	ma’če	mmau
	the-REL-EXIST-him	ear	there

“he who has ears (to listen, let him listen)”.

The postponed cases of possessors mediated by prepositions thus appear to be the main driver of the lower frequency of `iobj` dependencies in the Bohairic data. As far as we are aware, this finding has not been published on to date.

---

[6] An anonymous reviewer has also asked about comparable breakdowns by genre, but we feel that for some of the texts, comparisons would be hard to establish, and these would also split the already small treebank to the point where counts would be substantially less reliable.

4.3 Focus and preterit marking

Coptic belongs to the group of languages that employ morphological devices known as 'converters' (Layton, 2011, 319-366), which are applied to entire clauses, converting them for example into information structurally marked focalized clauses (focus conversion), signal anterior past tense (the preterit conversion, creating imperfect readings from present clauses or pluperfects from perfect clauses), among others.

For focus marking, two competing strategies are found: Cleft Sentences as in (15) and the morphological focus converter (sometimes referred to as the Second Tense, where the focalized present tense is called the 'Second Present' etc.) as in (16).

(15)

ⲟⲩ	ⲡ-ⲉⲧ-ⲟⲩ-ⲓⲣⲓ	ⲙⲙⲟ-ϥ	ϧⲉⲛ-ⲛⲓ-ⲥⲁⲃⲃⲁⲧⲟⲛ
ou	p-et-ou-iri	mmo-f	xen-ni-sabbaton
what	COP-REL-3PL-do	ACC.it	on-the.PL-sabbath

"What is it that they do on Sabbaths?"

(16)

ⲉⲧ-ⲁ-ⲓ-ϯ-ⲱⲙⲥ	ⲛⲱ-ⲧⲉⲛ	ϧⲉⲛ-ⲟⲩ-ⲙⲱⲟⲩ
et-a-i-ti-ōms	nō-ten	xen-u-mōw
FOC-PST-1SG-give-baptism	to-you	in-a-water

"I christened you with WATER"

The presence of the morpheme glossed as FOC in (16) indicates that a constituent is focalized in the sentence, in this case 'water' ('I christened you with WATER' rather than something else).

It has been observed that these strategies are represented unequally in Sahidic and Bohairic, with the preference for the focalizer in the former dialect, and for the cleft sentence in the latter (Müller, 2021). However, up until now such observations have not been backed up with precise and reproducible quantitative data. The UD treebanks make it possible to find the Sahidic-to-Bohairic ratio of focus markers in various parts of the New Testament texts. Since the use of these markers is heavily context and content dependent, we restrict our search to just the Biblical sources available in both treebanks: For the Gospel of Mark, this ratio roughly equals 2.44, and for 1 Corinthians, it is 3.58, both clearly favoring the prevalence of the focalizer in Sahidic.

A less stark, but more surprising result can be found for the preterit marker in the two dialects: as the only formal category denoting anterior past, we could logically expect its equal representation in identical texts in Sahidic and Bohairic. Yet, the UD treebank statistics show that the preterit marker occurs in Bohairic almost twice as often as in Sahidic (the ratios for Mark and 1 Corinthians are 2.21 and 1.76 favoring Bohairic, respectively). These numbers show that there is a substantial difference in how the tense systems of the two dialects are constructed, though we leave a more detailed study of what stands in place of the preterit in Sahidic to future work.

5 Parsing

5.1 Cross-corpus experiments

To evaluate the possibility of using the treebank to train an effective parser for Bohairic Coptic, we train and test several models in different scenarios:

1. Training on just the Bohairic train-set
2. Training on just the Sahidic train-set
3. Joint training on the Bohairic and Sahidic UD treebank train-sets
4. Balanced training on the smaller Bohairic train-set, and an equal amount of Sahidic data

The balanced setting is meant to evaluate whether joint training is more feasible if we ensure Bohairic examples are not overwhelmed by the larger amount of data available for Sahidic. In this case we take care not to include the same document (e.g. the same Bible chapter) from both dialects, and otherwise randomly select Sahidic documents from the appropriate partition, until the Bohairic data size has been reached.

To run the experiments, we use DiaParser (Attardi et al., 2021), a neural biaffine dependency parser, using the default hyperparameters (see the Appendix for exact values). As input embeddings we use the MicroBERT architecture and Sahidic Coptic embeddings made available by Gessler and Zeldes (2022), and train corresponding MicroBERT embeddings for Bohairic using all available Bohairic Coptic data from Coptic Scriptorium's online repository;[7] the new Bohairic transformer embeddings will be released publicly via Hugging Face.

Table 5 shows the results for labeled and unlabeled attachment scores on the respective test sets in each setting, along with train and test partition sizes in thousands of tokens. The results initially reveal that, unsurprisingly, training and testing across dialects (red numbers) produces very poor results, with LAS and UAS scores around 73 and 62 respectively – the scores are rather comparable in both

---

[7] https://github.com/CopticScriptorium/corpora

	test			
	Bohairic (11K tokens)		Sahidic (10.3K tokens)	
train (tokens)	**LAS**	**UAS**	**LAS**	**UAS**
Bohairic (16.5K)	86.349	89.486	62.205	74.633
Sahidic (35.8K)	62.683	73.178	**89.760**	**92.489**
Joint (52.3K)	**89.929**	**92.677**	88.449	91.602
Balanced (36K)	88.989	91.927	86.858	90.628

Table 5: Labeled (LAS) and Unlabeled (UAS) Attachment Scores in each setting when testing on each dialect. Within and across dialect scores are in green and red respectively. The best settings for each dialect are bolded.

directions, despite the availability of almost double the data when training on Sahidic. This indicates that the parser is unable to generalize when surface forms vary, since as we noted above, even auxiliaries and prepositions look quite different across dialects.

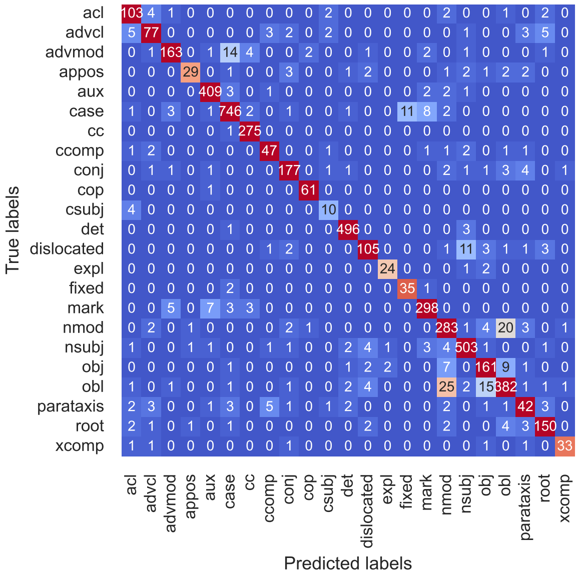
Single dialect models (green numbers) reveal a gap between the smaller Bohairic data and its larger Sahidic counterpart: while Sahidic obtains a LAS of 89.76, Bohairic lags behind with 86.349 (2.5 point difference), with an even starker difference in UAS (92.489 vs. 89.486, about 3 points). Given that the dialects and texts available in them are rather similar, this suggests that more Bohairic data is likely to have an impact in a single-dialect setting.

Moving to the joint models, both the balanced and full-joint scenario improve the score on Bohairic, suggesting that although word forms are different, syntactic structures are similar enough to generalize across the datasets. In fact, the JOINT setting outperforms BALANCED, suggesting that simply having more data is better, as long as there is a core of Bohairic examples to inform the parser about pivotal Bohairic word forms. In absolute terms, the joint Bohairic scores even slightly surpass the Sahidic single-dialect model scores, though these numbers are not strictly comparable, since the test sets contain different documents. We suspect this means the Bohairic test set may be slightly easier overall than the Sahidic one, but we also take it to be an indication that our annotations match the Sahidic guidelines closely.

By contrast to the Bohairic benefit from joint training, both joint scenarios perform worse on Sahidic than the Sahidic-only model. This suggests that given the amount of data available in Sahidic, the infusion of the smaller Bohairic data is not helpful. In the Sahidic case BALANCED is unsurprisingly the worse setting, since there is less total Sahidic data involved.

5.2 Error analysis

To better understand what is challenging about Bohairic parsing, we perform quantitative and qualitative error analysis. Figure 2 shows the confusion matrix for dependency labels in the Bohairic test set for the Bohairic-only model (merging subtypes of the same major relations and omitting labels with fewer than 10 occurrences in the test set).



	acl	advcl	advmod	appos	aux	case	cc	ccomp	conj	cop	csubj	det	dislocated	expl	fixed	mark	nmod	nsubj	obj	obl	parataxis	root	xcomp
acl	103	4	1	0	0	0	0	0	0	0	2	0	0	0	0	0	2	0	0	1	0	2	0
advcl	5	77	0	0	0	0	0	3	2	0	2	0	0	0	0	0	0	1	0	0	3	5	0
advmod	0	1	163	0	1	14	4	0	0	2	0	0	1	0	0	2	0	1	0	0	0	1	0
appos	0	0	0	29	0	1	0	0	3	0	0	1	2	0	0	0	1	2	1	2	2	0	0
aux	0	0	0	0	409	3	0	1	0	0	0	0	0	0	0	2	2	1	0	0	0	0	0
case	1	0	3	0	1	746	2	0	1	0	0	1	0	0	11	8	2	0	0	0	0	0	0
cc	0	0	0	0	0	1	275	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ccomp	1	2	0	0	0	0	0	47	0	0	1	0	0	0	0	1	1	2	0	1	1	0	0
conj	0	1	1	0	1	0	0	0	177	0	1	1	0	0	0	0	2	1	1	3	4	0	1
cop	0	0	0	0	1	0	0	0	0	61	0	0	0	0	0	0	0	0	0	0	0	0	0
csubj	4	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0
det	0	0	0	0	0	1	0	0	0	0	0	496	0	0	0	0	0	3	0	0	0	0	0
dislocated	0	0	0	0	0	0	0	1	2	0	0	0	105	0	0	0	1	11	3	1	1	3	0
expl	0	0	0	0	0	0	0	0	0	0	0	0	0	24	0	0	0	1	2	0	0	0	0
fixed	0	0	0	0	0	2	0	0	0	0	0	0	0	0	35	1	0	0	0	0	0	0	0
mark	0	0	5	0	7	3	3	0	0	0	0	0	0	0	0	298	0	0	0	0	0	0	0
nmod	0	2	0	1	0	0	0	0	2	1	0	0	0	0	0	0	283	1	4	20	3	0	1
nsubj	1	0	0	1	1	0	0	1	1	0	0	2	4	1	0	3	4	503	1	0	0	1	0
obj	0	0	0	0	0	1	0	0	0	0	0	1	2	2	0	0	7	0	161	9	1	0	0
obl	1	0	1	0	0	1	0	0	1	0	0	2	4	0	0	0	25	2	15	382	1	1	1
parataxis	2	3	0	0	1	3	0	5	1	0	1	2	0	0	0	0	2	0	1	1	42	3	0
root	2	1	0	1	0	1	0	0	0	0	0	0	2	0	0	0	2	0	0	4	3	150	0
xcomp	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	33

Figure 2: Confusion matrix for collapsed major dependency relations for the Bohairic-only model on the Bohairic test set (labels with <10 occurrences are omitted).

As with most parsers, the most common confusion is between the `obl` and `nmod` labels, indicating problems with classic PP-attachment ambiguities. Overprediction of low adnominal attachment (`nmod` for true `obl`) is slightly more common than the opposite. Additional relatively common confusions occur for `dislocated` subjects versus regular subjects (`nsubj`), which is not surprising, and for `advmod` and `case` being confused with `case` and `fixed` respectively. The latter two are due to ambiguous phrasal verbs, illustrated in (17)–(18):

(17)	ⲁ-ϥ-ⲱϣ	ⲉⲃⲟⲗ	ϧⲉⲛ-ⲟⲩ-ⲥⲙⲏ	ⲙ-ⲡⲣⲟⲫⲏⲧⲓⲕⲟⲛ
	a-f-ō-š	ebol	xen-u-smē	m-prophētikon
	PST-he-cry out		in-a-voice	of-prophetic
	‘He called out in a prophetic voice’			

(18)	ⲉⲧ-ⲁ-ⲡⲓ-ⲁⲗⲟⲩ	ⲓ	ⲉⲃⲟⲗ	ϧⲉⲛ-ⲡ-ⲏⲓ
	et-a-pi-alu	i	ebol	xen-p-ēi
	PRC-PST-the-boy	came	out	in-the-house
	‘after the boy came out of the house’			

Like English phrasal verbs, some Coptic verbs combine with postposed adverbs to form a unique meaning – for example in (17), cry + out means ‘cry out, call out’ much like in English. In the example, ‘out’ is coincidentally followed by the preposition ‘in’, for the phrase “cry out in a prophetic voice”. However Coptic also has several frequent fixed combination of adverbs with following prepositions, such as **ⲉⲃⲟⲗ ϧⲉⲛ**, lit. ‘out + in’ but actually meaning ‘out of’ (similar to the English fixed expression ‘out of’ sometimes spelled as ‘outta’). In these cases, our guidelines annotate the second token as `fixed`, and the first token takes on the expected `case` label – this ambiguity of adverbs next to prepositions, in which adverbs may belong together with a verb or be part of a multi-word preposition, causes the relatively frequent label confusion in Figure 2. We note that in the interest of comparability with other languages and following UD guidelines (see Ahrenberg 2024 for discussion), the list of fixed expressions of this kind is kept small and is meant to be exhaustive, covering as of writing 25 unique combinations of lemma pairs.

6 Conclusion

In this paper we presented the first morphosyntactically annotated corpus of Bohairic Coptic, containing over 30K word forms from a range of texts. By adopting the same guidelines as the existing UD Sahidic Coptic Treebank, we have been able to perform some first studies of more subtle quantitative differences between the dialects, complementing the better known categorical differences between them. We also ran parsing experiments which indicated that models trained on both dialects jointly were able to boost performance on the lower resource Bohairic dialect, but not on Sahidic.

We are hopeful that this corpus will represent a starting point for further expansion of annotated data for Bohairic Coptic in particular, and Coptic dialects in general. We are confident there is much room for both improving NLP tools for Coptic using such data, and for studying Coptic dialects individually and comparatively.

Limitations

By its nature, this study is based on specific texts which lead to specific results. Although the attempt has been made to select somewhat diverse texts for the corpus, it is always possible that a different selection would have led to different results. In particular, the inclusion of translated texts, such as material from the Bible, is not ideal for some types of research, but as is often the case in resources for historical languages with limited attestation, this is somewhat inevitable. We are hopeful that as new data becomes available, additional studies may revisit some of our findings and either validate or relativize these results.

Acknowledgments

This work was made possible by a grant from the National Endowment for the Humanities Preservation and Access Humanities Collections Reference and Resources Program (PW-290519-23). We thank our postdoctoral fellow, Lydia Bremer-McCollum, for digitizing some of the text used in this paper via OCR, and Wolfgang Jentner and the Data Institute for Societal Challenges at the University of Oklahoma for their server support. We would also like to acknowledge work done by other members of the Coptic Scriptorium project, the Marcion project and the Coptic Old Testament project, as well as by Hany Takla of the St. Shenouda the Archimandrite Coptic Society, which was instrumental in making the data we annotated in this paper available in digital formats. All resources developed in this project will be released under an open license in the hope that we can contribute and match their generosity.

References

Lars Ahrenberg. 2024. Fitting fixed expressions into the UD mould: Swedish as a use case. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 33–42, Torino, Italia. ELRA and ICCL.

Emile Amélineau, editor. 1887. *De Historia Lausiaca: adjecta sunt quaedam hujus Historiae coptica fragmenta inedita*. Ernest Leroux, Paris.

Emile Amélineau, editor. 1890. *Histoire du patriarche copte Isaac: étude critique, texte et traduction*. Ernest Leroux, Paris.

Giuseppe Attardi, Daniele Sartiano, and Maria Simi. 2021. Biaffine dependency and semantic graph parsing for EnhancedUniversal 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, Online. Association for Computational Linguistics.

Francesco Berno. 2019a. Historia lausiaca. In *Archaeological Atlas of Coptic Literature*, volume paths.works.245.

Francesco Berno. 2019b. Vita isaac ep. alexandriae. In *Archaeological Atlas of Coptic Literature*, volume paths.works.225.

Francesco Berno. 2019c. Vita Sinuthii. In *Archaeological Atlas of Coptic Literature*, volume paths.works.461.

Bernard Caron. 2015. Hausa grammatical sketch. In Amina Mettouchi, Martine Vanhove, and Dominique Caube, editors, *Corpus-based Studies of Lesser-described Languages. The CorpAfroAs corpus of spoken Afro-Asiatic languages*. John Benjamins, Amsterdam & Philadelphia.

Robert S. D. Crellin. 2025. Considerations for the design of dependency treebanks for linguistic research in Biblical Hebrew. In Aaron D. Hornkohl, Nadia Vidro, Janet C. E. Watson, Eleanor Coghill, Magdalen M. Connolly, and Benjamin M. Outhwaite, editors, *Interconnected Traditions: Semitic Languages, Literatures, Cultures – A Festschrift for Geoffrey Khan*, pages 99–130. Open Book Publishers.

Berthold Crysmann. 2012. HaG: A computational grammar of Hausa. In *Selected proceedings of the 42nd annual conference on African linguistics (ACAL 42)*, pages 321–337.

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

Roberto Antonio 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)*, pages 1–10, Hamburg,Germany. Association for Computational Linguistics.

Frank Feder, Maxim Kupreyev, Emma Manning, Caroline T. Schroeder, and Amir Zeldes. 2018. A linked Coptic dictionary online. In *Proceedings of LaTeCH 2018 - The 11th SIGHUM Workshop at COLING2018*, pages 12–21, Santa Fe, NM.

Luke Gessler and Amir Zeldes. 2022. MicroBERT: Effective training of low-resource monolingual BERTs through parameter reduction and multitask learning. In *Proceedings of the 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pages 86–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Eitan Grossman. 2015. No case before the verb, obligatory caseafter the verb in Coptic. In Eitan Grossman, Martin Haspelmath, and Tonio Sebastian Richter, editors, *Egyptian-Coptic Linguistics in Typological Perspective*, Empirical Approaches to Language Typology [EALT] 55, pages 203–225. De Gruyter Mouton, Berlin.

Katharina Hartmann. 2006. Focus constructions in Hausa. In Valéria Molnár and Susanne Winkler, editors, *The Architecture of Focus*, pages 579–607. De Gruyter Mouton, Berlin.

George W. Horner. 1905. *The Coptic version of the New Testament in the northern dialect*, volume 3. Clarendon Press.

Mayank Jobanputra, Maitrey Mehta, and Çağrı Çöltekin. 2024. A Universal Dependencies treebank for Gujarati. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 56–62, Torino, Italia. ELRA and ICCL.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Rodolphe Kasser. 1991. Dialects: Grouping and major groups of. In Aziz S. Atiya, editor, *The Coptic Encyclopedia*, volume 8, pages 97–101. Macmillan, New York.

Bentley Layton. 2011. *A Coptic Grammar*, third edition, revised and expanded edition. Porta linguarum orientalium 20. Harrassowitz, Wiesbaden.

Johannes Leipoldt, editor. 1906. *Sinuthii Archimandritae Vita et Opera Omnia. Vol. 1.* Corpus Scriptorum Christianorum Orientalium 41. Imprimerie nationale, Paris.

Antonio Loprieno. 2000. From VSO to SVO? Word order and rear extraposition in Coptic. In Rosanna Sornicola, Erich Poppe, and Ariel Shisha-Halevy, editors, *Stability, Variation and Change of Word-Order Patterns over Time*, Current Issues in Linguistic Theory 213, pages 23–40. John Benjamins, Amsterdam/Philadelphia.

Nina Lubomierski. 2008. The Coptic life of Shenoute. In Gawdat Gabra and Hany N. Takla, editors, *Christianity and Monasticism in Upper Egypt: Volume 1, Akhmim and Sohag*, pages 91–98. American University in Cairo Press, Cairo and New York.

Matthias Müller. 2021. *Grammatik des Bohairischen*. Lingua Aegyptia 24. Widmaier, Hamburg.

Caroline T. Schroeder and Amir Zeldes. 2020. A collaborative ecosystem for digital Coptic studies. *Journal of Data Mining and Digital Humanities. Special Issue on Collecting, Preserving, and Disseminating Endangered Cultural Heritage for New Understandings through Multilingual Approaches*, pages 1–9.

Ariel Shisha-Halevy. 2007. *Topics in Coptic Syntax: Structural Studies in the Bohairic Dialect*. Orientalia Lovaniensia analecta 160. Peeters, Leuven.

Daniel G. Swanson and Francis M. Tyers. 2022. A Universal Dependencies treebank of Ancient Hebrew. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2353–2361, Marseille, France. European Language Resources Association.

Dima Taji, Nizar Habash, and Daniel Zeman. 2017. Universal Dependencies for Arabic. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 166–176, Valencia, Spain. Association for Computational Linguistics.

Amir Zeldes and Mitchell Abrams. 2018. The Coptic Universal Dependency treebank. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 192–201, Brussels, Belgium. Association for Computational Linguistics.

Amir Zeldes, Nick Howell, Noam Ordan, and Yifat Ben Moshe. 2022. A second wave of UD Hebrew treebanking and cross-domain parsing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4331–4344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Amir Zeldes and Caroline T. Schroeder. 2016. An NLP pipeline for Coptic. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 146–155, Berlin, Germany. Association for Computational Linguistics.

A Dependency relations

We use the entire inventory of Universal Dependency relations with the exception of the `clf` relation, since Coptic has no classifiers, and no cases of an underspecified dep relation, for a total of 32 basic relations. In addition, we use the following four subtypes, as used in the Sahidic treebank:

- `acl:relcl` - to distinguish relative clauses from adnominal infinitives and other adnominal clauses
- `nmod:poss` - for adnominal possessive pronouns, including both enclitic pronoun possessors and prenominal possessive pronouns
- `nmod:unmarked` - for adnominal, adverbially used noun phrases, not mediated by a preposition
- `nmod:unmarked` - for adverbially used noun phrases, not mediated by a preposition, when modifying a verbal head

We do not use the subtype `nsubj:pass` since Coptic has no unambiguous actional passive, instead using impersonal third person active syntax (“they built it” = “it was built”). The total distinct labels in the corpus therefore number 36.

B Hyperparameters

The following hyperparameters were used for DiaParser, based on the default parameters combined with the embeddings size of the MicroBERT transformer model:

- BertEmbedding
  - n_layers=4
  - n_out=100
  - max_len=512
- embed_dropout: p=0.33
- LSTM
  - dimensions: 200 x 400 x 3 layers
  - bidirection=True
  - dropout=0.33
- MLP dropouts (arc_d/h, rel_d/h): 0.33
- criterion=CrossEntropyLoss

Negation in Universal Dependencies

Jamie Y. Findlay and **Dag T. T. Haug**
University of Oslo
`j.y.findlay@iln.uio.no, d.t.t.haug@ifikk.uio.no`

Abstract

In this paper we study the representation of negation in UD treebanks. We show that the existing annotations are often inconsistent with the guidelines and that there are ill-motivated differences in annotation of constructions across and even within languages. Moreover, we argue that even if the annotation of the two negation-related features (`Polarity=Neg` and `PronType=Neg`) were consistent, these two features would be inadequate for straightforwardly expressing the semantics of negation because they relate to the word level only and hence to form rather than meaning. We therefore propose to add two features, `Negated=+` and `DoubleNegated=+`, which directly encode when a predicate is *semantically* under negation, and thereby allow a straightforward semantic interpretation of a UD parse in terms of negation.

1 Introduction

Negation is a complex but ubiquitous phenomenon in natural language, involving constraints across the syntax-semantics interface. In this paper, we address the representation of negation in Universal Dependencies (UD: Nivre et al. 2020), point out inconsistencies in the current representation, and make a suggestion for how it can be improved, including an implementation that will create this representation for English UD.

Our ultimate motivation for this work is the use of UD in semantic parsing, i.e. the translation of UD syntactic annotations into semantic representations. In Section 2 we discuss this use case for UD and what it entails – especially for feature representation, which is much less standardised than dependency relations. Section 3 then gives a (simplified) overview of the challenges this task raises at the syntax-semantics interface. Negation is a wide field and we therefore choose to focus on some selected cases all involving sentential negation. Unfortunately, the existing UD protocol and its current implementations in the UD treebanks leave something to be desired when it comes to helping with this challenge. In Section 4 we show that the current annotation of negation is inconsistent across treebanks, and sometimes even within single treebanks. In Section 5 we show that even if the annotation was consistent, the word-level, morphology-dominated features currently used in the UD annotation scheme make it challenging to introduce negation compositionally in the semantics. In Section 6, we propose an addition to the UD annotation scheme whereby features are added to the predicate itself to indicate whether it is negated (or doubly negated). Using the Graph Rewriting formalism (GREW: Guillaume et al. 2012; Bonfante et al. 2018), we provide a ruleset which can add these features automatically to UD annotations for English. However, we suggest that these features should be included in the standard annotation scheme in future, in keeping with recent proposals to add phrase- or clause-level features to UD annotations (Savary et al. 2023; Weissweiler et al. 2024).

2 Motivation

There is a tradition of work using UD representations to create semantic representation (Reddy et al. 2017; Poelman et al. 2022; Findlay et al. 2023). Most of this work focuses on the information available in the tree structure and the labels, which clearly provide important clues to the structure of the semantic representations as well as to the semantic relations between different words. But features also have an important role to play, since they provide *language neutral* representations of certain aspects of lexical information, most often in function words. For example, in the absence of a featural representation of negation, a semantic parser will need language-specific lexical information about what words express negation and how. While high quality semantic parsing will likely re-

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 70–79
August 27, 2025 ©2025 Association for Computational Linguistics

quire fine-grained language-specific information in any case, the process can clearly be simplified to the extent that certain lexically-encoded aspects of meaning can be read off the features in a language neutral way.

This is, however, not possible with the current UD features. Consider tense, for example. Despite the guidelines' claim that `Tense` "is a feature that specifies the time when the action took/takes/will take place",[1] the actual features that we find in UD reflect morphological rather than semantic facts. Part of the reason this is so is that features in UD are necessarily word-bound, and automatic annotation procedures tend not to take the surrounding context into account, so that properties which arise constructionally or periphrastically are not represented. As a concrete example, consider the value of the `Tense` feature on an English verb in the UD treebanks – this depends entirely on the verb's form, and not on its meaning, so *provided* gets `Tense=Past` regardless of whether it is used as a preterite (1), a (present) perfect (2), or a passive of any tense (3):

(1) They also bred small domesticated dogs which [. . .] provided their protein.
`(UD_English-GUM, GUM_textbook_history-37)`

(2) Mary has provided a preface.
`(UD_English-GUM, GUM_letter_arendt-33)`

(3) [T]he tourist card is normally provided on the flight.
`(UD_English-GUM, GUM_voyage_cuba-11)`

It is possible to compute tense values for the verbal complexes in (3), by inspecting other features such as `VerbForm` and `Voice`, but this cannot be done without detailed, language-specific knowledge about the English annotation. This approach is therefore not amenable to large-scale typological investigations, which is one of the oft-cited use cases for UD.

It would be possible to solve this problem "compositionally" by defining universal resolution rules for sets of features coming from both the lexical verb and the auxiliary. A more direct route is to add annotations of *constructions* on top of UD, as argued for by Weissweiler et al. (2024). Here we will take a slightly different route, by allowing head features to represent features of the whole construction. This is largely equivalent to constructional annotations for constructions that have a clear head, which is the case for both verb-auxiliary constructions and verb-negation constructions.[2] Adding features to the head also has a practical advantage by making it possible to directly compare constructional and lexical expressions (although one could of course just admit single word constructions).

Our approach also aligns with the goals of the UniDive shared task on morpho-syntactic parsing (MSP).[3] Following UD, MSP makes a sharp divide between lexical words and function words. The latter are always dependents in UD and can therefore be ignored in the tree structure. Instead they contribute features to their heads. So, in the MSP annotation of *will not go*, only *go* is a node in the tree; *will* and *not* contribute `Fut` and `Neg` to the features of *go*.

Although we do not care about the tree structures and what counts as a node here, we adopt the same principle as MSP that function words can contribute features to their head. This move is in fact what makes it possible for us (and MSP) to take advantage of a much-discussed feature of UD, that function words are always dependents. The main argument for this approach is that it makes annotations more similar across languages that use analytic and synthetic means of expression. As the UD guidelines put it,[4] "Preferring content words as heads maximizes parallelism between languages because content words vary less than function words between languages. In particular, one commonly finds the same grammatical relation being expressed by morphology in some languages or constructions and by function words in other languages or constructions [. . .]". However, the primacy of lexical words only makes the *tree structures* similar, not the *features*, unless we let dependents contribute features to their head.

Consider the example discussed by de Marneffe et al. (2021: 264) shown in Figure 1. The tree topology is identical modulo function words, but the feature representation is not: in Swedish, the subject NP bears both a `Number=Plur` and a

---

[1] See https://universaldependencies.org/u/feat/Tense.html.

[2] Although UD disagrees with most linguistic analyses in taking the auxiliary to be a dependent of the verb, both approaches agree that the structure is headed.

[3] https://www.aclweb.org/portal/content/1st-shared-task-morphosyntactic-parsing

[4] https://universaldependencies.org/u/overview/syntax.html

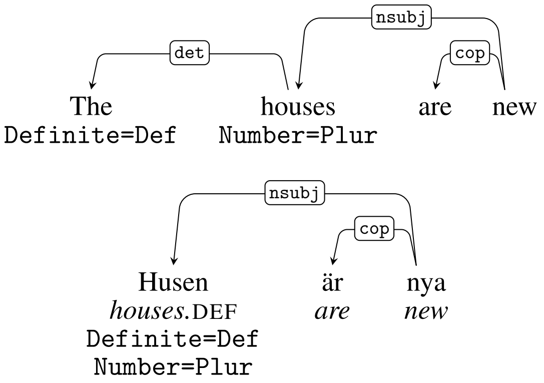


Figure 1: Definiteness in English and Swedish

`Definite=Def` feature, whereas the English subject NP only has the number feature. This means the UD representation cannot be used to study plural, definite noun phrases across languages without some further post-processing. In this case, the rule that is needed to unify the two representations is simple (move the definite feature from the functional dependent to its lexical head), but as we have seen for tense, and as we will see for negation in this paper, this is not always the case. While the percolation of features from functional dependents to lexical heads, as made explicit in MSP, can be seen as a sort of default, the rules of feature construction can be quite involved and often cannot be stated without deep knowledge of the language in question, which is why it should be done by annotators, and not by end users of UD treebanks.

3 The challenges of negation

With all this said, our goal is to explore to what extent the *morphosyntactic* representation of negation in UD treebanks allows for an adequate *semantic* representation of negation in a uniform manner across languages, and to propose amendments to the UD treatment of negation that would facilitate this.

We do not want to commit to a particular meaning representation, so in the following we discuss the inference patterns that negated sentences give rise to, largely following the presentation in Huddleston and Pullum (2002: ch. 9). We do not explicitly discuss the role of tense, but assume, following Partee (1984), that sentences are interpreted relative to some particular time interval that is anaphorically determined in the context, and that this interval is held constant from premises to conclusions in inferencing.

Let us first consider the simple negated transitive clause in (4):

(4) Jason did not eat the porridge.

We take it that this sentence licenses an inference to the disjunction of (5a–c):

(5) a. No eating event took place.
 b. The Agent of the eating event was not Jason.
 c. The Theme of the eating event was not the porridge.

This is a fairly weak semantics for negation, but it is all that can be attributed to the linguistic information alone. Exactly which of the disjuncts turn(s) out to be true will be determined by context and world knowledge.

Notice that the negation scopes over both the verb's event description and the semantic roles associated with the event. We take it that *clausal* negation (as opposed to so-called *constituent* negation) always negates these parts of the sentence: that is what it means for negation to be clausal. However, clausal negation may or may not scope over adjuncts.

Adjunct scope Consider the following extended variant of (4).

(6) Jason didn't eat the porridge because he had already eaten.

World knowledge tells us that in (6), the most plausible reading is that one or more of the disjuncts that we saw in (5) is true, and the *because*-clause gives the reason why this is so. That is, negation does not scope over the adjunct.

But now consider the extension in (7)

(7) Jason didn't eat the porridge because he was hungry.

Here, world knowledge tells us that it is likely none of the disjuncts in (5) is true; instead only the causal relation between John's hunger and his eating is denied. In other words, he ate the porridge, but not because he was hungry.

The linguistic structure itself is insufficient to distinguish between these two cases, and this is once again a situation where context and world knowledge will guide interpretation.

Negative indefinites Another challenging phenomeon is the fact that, in addition to marking negation on the verb, many languages allow negation to be expressed with so-called "negative indefinites" (Haspelmath 1997: ch. 8), which combine sentence negation and an indefinite (existential) quantifier. Hence (8a) can be paraphrased as (8b).

(8) a. Jason ate nothing.
b. It was not the case that Jason ate something.

An adequate representation of negation in these cases must be able to account for this equivalence, i.e. that the negation is clausal in spite of being marked only on the object quantifier. In other words, we want *nothing* to contribute a clausal negation, represented on the clausal head, while at the same time filling the object position. This is a direct challenge to the the strict division adopted by MSP, where function words contribute features to their heads and lexical words are nodes in the tree: words like *nothing* need to be able to do both.

4 Negation in the UD treebanks

In this section we explore the variable representations of negation in the UD treebanks (in the 2.15 release: Zeman et al. 2024). It will be seen that there are unfortunate inconsistencies within and across languages, even taking into account typological variation. Ideally, we would like the UD parse to provide us with sufficient information to produce a schematic meaning representation of the sentence, with at least the right logical structure – even though specific predicate names or thematic relations will ultimately have to be retrieved from a lexicon (see e.g. Reddy et al. 2017 or Findlay et al. 2023 for proposals along these lines). However, inconsistencies and lacunae in the annotations make this more challenging than it needs to be, and negation proves to be a case in point.

UD provides two relevant features for representing negation: `Polarity=Neg` and `PronType=Neg`. According to the guidelines,[5] the former is used for "negating particles and words which inflect for polarity (verbs, adjectives, etc.)", including English *not*, *nor*, and *no* as an interjection, while the latter is used for negative pronouns such as English *nobody* or *nothing*, along with other negative proforms such as *never*, and negative determiners such as *no* (as in *no books*).

When it comes to straightforward clausal negation, the treebanks are largely uniform in correctly employing `Polarity=Neg` alone on the negator, whether that be the predicate itself, inflected with a negative affix, as in e.g. Persian, or a dependent of the negated predicate, as in e.g. Norwegian – see Figure 2

There are exceptions even with this most basic kind of negative construction, however. The `UD_Italian-ISDT` treebank, for example, uses `PronType=Neg` for this purpose instead, contrary to the annotation guidelines, and only uses `Polarity=Neg` for the discourse particle/interjection *no* – see Figure 3. This inconsistency between treebanks is an issue for semantic interpretation, since there is no single feature which is reliably associated with clausal negation.

A different challenge comes from treebanks such as `UD_English-EWT`, which make no use of `Polarity=Neg` whatsoever, such that clausal negation is simply not marked in the syntactic analysis. In fact, only 224 of the 296 treebanks in the 2.15 release (76%) use the `Polarity=Neg` feature, and only 99 (33%) use the `PronType=Neg` feature. Overall, 62 treebanks (21%) use neither feature, meaning over a fifth of the treebanks do not represent negation of any kind, even though negation is a phenomenon present in every natural language (Zeijlstra 2020: 426).[6]

One area of much more cross-treebank consistency is in the treatment of negative conjunctions like *neither* and *nor* or their translation equivalents. Unfortunately, this is because treebanks consistently *fail* to annotate these words with any feature that indicates a negative polarity. This is once again contrary to the annotation guidelines – which explicitly mention English *nor* as meriting the feature `Polarity=Neg`, for example – and is also plainly unsatisfactory from the standpoint of interpretation: (9a) is truth-conditionally equivalent to (9b), and therefore just as clearly negated, but only the latter will contain a `Polarity=Neg` as things stand:

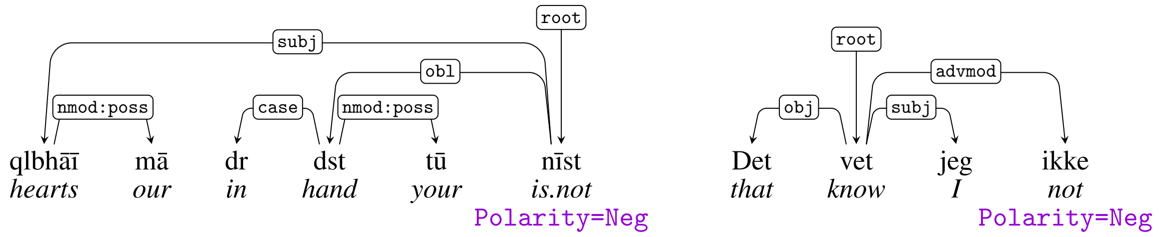
(9) a. Neither did Jonathan.
b. Jonathan didn't either.

In other areas, inconsistencies abound. Neg-

---

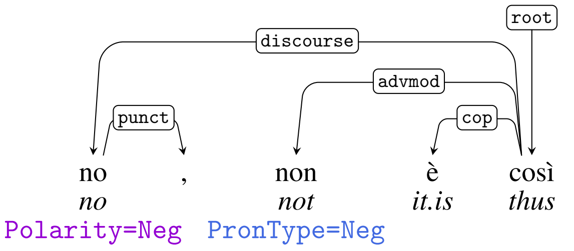
[5] See https://universaldependencies.org/u/feat/PronType.html#Neg and https://universaldependencies.org/u/feat/Polarity.html.

[6] The script used to calculate these figures is available at https://github.com/findlayjy/negation-in-UD/.



(`UD_Persian-Seraji`, `dev-s117`) (`UD_Norwegian-Bokmaal`, `000611`)

Figure 2: UD trees for two negated sentences



(`UD_Italian-ISDT`, `isst_tanl-864`)

Figure 3: The two features in `UD_Italian-ISDT`

ative determiners, for example – the equivalent of English *no* (as in *We have no bananas*) – ought to be annotated with a `PronType=Neg` feature according to the UD guidelines. And some treebanks do just this, such as `UD_Persian-Seraji`. But we also find such words annotated with only a `Polarity=Neg` feature, as in `UD_Norwegian-Bokmaal`, or with neither feature, as in `UD_Turkish-BOUN`. Part of the challenge here comes from the fact that in negative concord languages these determiners can appear under negation, such that they have the (negative polarity item) meaning ‘any’ (as in *We do not have any bananas*). But, as noted above, treebanks generally choose a single annotation for a particular word regardless of context, and this therefore leads to inaccuracies in both directions. For example, Turkish *hiç* can have a directly negative meaning but is consistently annotated as if it cannot, while the Persian cognate *hīč* is explicitly marked as negative but can equally well have the NPI meaning.

Treatment of negative indefinites also varies substantially, even within a single language. We take `UD_English-GUM` as a case study. Table 1 shows the spread of features used across various negative words in the treebank in comparison with what the UD guidelines would lead us to expect. As can be seen, only three of the nine n-words are annotated as the guidelines would suggest. Among the quantifiers/indefinites, only *nobody* is annotated “correctly”; the others all bear an unexpected `Polarity=Neg` feature. But what is also striking is that only two of the four quantifiers bear the expected `PronType=Neg` feature. Perhaps *nobody* and *nothing* are distinguished from *nowhere* and *never* in `UD_English-GUM` because the former are pronouns while the latter are broader proforms (e.g. they can replace PPs), but the guidelines explicitly list all four as examples of words that should carry the `PronType=Neg` feature. However, we can understand the `UD_English-GUM` creators’ decision here, since this use of `PronType=Neg` is somewhat at odds with the higher-level, more general guidelines around the feature `PronType`, which say that, as the name would suggest, it “typically” applies to pronouns and other pronominal items.[7] (Of course, this is presumably one of the “atypical” times, where the more specific guidelines should supersede the broader ones.)

On the other hand, we do not understand the decision to distinguish *nobody* from *nothing* by the former’s lack of a `Polarity=Neg` feature; it is not clear to us what linguistic distinction this could be intended to capture.

In sum, the existing UD treebanks handle negation in an inconsistent manner, both across and within treebanks. Because of this, neither of the two negation-related features can be given a consistent semantic interpretation. This is problematic for our goal of providing a skeletal semantic representation of the logical structure of a sentence, since negation plays an important role in determining this structure.

5 Insufficiency of the annotation scheme

Lack of annotation consistency makes universal interpretation directly from the features impossible. Certainly, clarifying the intended usage of both

---

[7] `https://universaldependencies.org/u/feat/PronType.html`

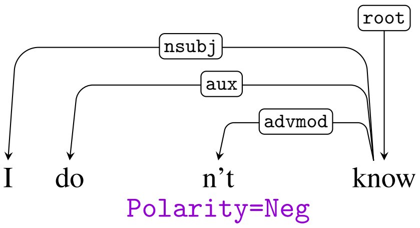
	`Polarity=Neg`		`PronType=Neg`	
	GUM	UD-G	GUM	UD-G
not	✓	✓	–	–
no$_{disc}$	✓	✓	–	–
neither	–	✓	–	–
nor	–	✓	–	–
no$_{det}$	✓	–	✓	✓
nobody	–	–	✓	✓
nothing	✓	–	✓	✓
nowhere	✓	–	–	✓
never	✓	–	–	✓

Table 1: Assignment of negative features to n-words in the English GUM corpus in comparison with the UD guidelines (UD-G)

`Polarity=Neg` and `PronType=Neg` and achieving greater coverage and consistency across the treebanks would help in this regard. Nevertheless, we believe even treebanks perfectly annotated according to the existing guidelines would be inadequate as the basis for semantic interpretation when it comes to negation. Let us see why.

The most elementary clausal negation is associated with the feature `Polarity=Neg`.

(10)

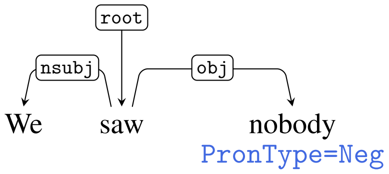


`(English-GUM, GUM_conversation_blacksmithing-26)`

We might therefore suggest that this feature triggers the introduction of negation in the semantics, so that (10) has the interpretation described in Section 3: there is no 'knowing' situation with the speaker as the Experiencer.

What about the other feature, `PronType=Neg`? In languages like English or Norwegian, which lack negative concord, this also sometimes needs to introduce clausal negation into the semantics:

(11)



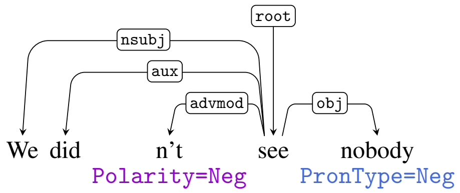
The sentence in (11) is, at least truth-conditionally, equivalent to (12), which has an explicit clausal negation:[8]

(12) We didn't see anybody.

So it seems reasonable to assume that `PronType=Neg` introduces such a meaning in (11) as well.

What is more, when both features are present, we get a double-negation reading (in non-negative concord varieties of English, at least):

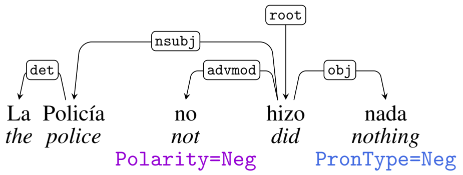
(13)



It therefore seems perfectly well-motivated to have both negation-related features introduce a clausal negation meaning.

Of course, when we come to negative concord (NC) languages, things change. Despite having two negative words, (14) does not express double, but only single negation – 'The police did nothing':

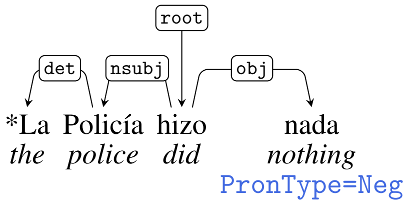
(14)



`(UD_Spanish-AnCora, 3LB-CAST-n1-2-s2)`

What is more, the `PronType=Neg` word cannot express clausal negation alone, at least in this position:

(15)



---

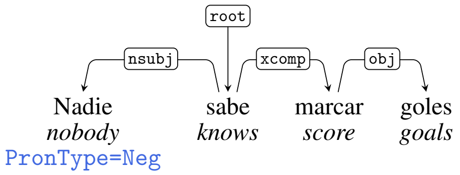
[8] Indeed, in languages like Norwegian, there is a systematic alternation betwen these two types of sentence: the form with an explicit negator is required with compound tenses, while both forms are compatible with simplex tenses:

(i) a. Han har ikke sett noe.
he has not seen something.
'He has not seen anything./He has seen nothing.'
b. Han så { ikke noe | ingenting }.
He saw not something nothing
'He didn't see anything./He saw nothing.'

We could, therefore, parametrise our semantic interpretation rule so that `PronType=Neg` introduces a negative meaning in non-negative concord languages, but does not do so in negative concord languages.

Unfortunately, this is not the whole story. Negative quantifiers in NC languages *do* sometimes introduce clausal negation into the semantics, when they are in subject position; (16) means ‘Nobody knows how to score goals’, i.e. ‘It is not the case that somebody knows how to score goals.’:

(16)



(UD_Spanish-AnCora, 3LB-CAST-d2-3-s1)

So we cannot simply parametrise the `PronType=Neg` feature and ‘switch off’ its semantic contribution in NC languages, since in fact it is sometimes still needed.

6 Proposal

Negation is in a similar bind to tense, as discussed above (in Section 2): the current annotations really relate to form alone, which means that, for example, no accommodation is made for (a) the difference between NC and non-NC languages, and (b) the contrast between different uses of n-words within a single language. Our solution is to introduce a new feature, `Negated=+`, which is annotated directly on a predicate which is *semantically* under negation, regardless of any morphosyntactic concerns. In the next section we explain how this can be interpreted, but for now we can observe that such a move also contributes to an increased universality in representations, since it establishes a clearer parallel between synthetic and analytic expressions of negation: in Figure 4, for example, we see that the English and Czech verbs both bear the `Negated=+` feature, even though the existing `Polarity=Neg` feature appears in different places in the two languages.

In addition to the `Negated=+` feature, we also add a feature `DoubleNegated=+`, for sentences like *We didn’t see nobody* in non-NC varieties, where this truly expresses a double negation. An alternative would be to simply omit the `Negated=+` feature in this case, on the basis that the semantics is equivalent to that of positive *We saw somebody*, but we choose to include more information rather than make the decision that double negations should be interpreted exactly the same way as un-negated sentences – after all, they are clearly pragmatically marked, and we prefer to leave open the space of analytical possibilities. Przepiórkowski and Patejuk (2015: 327–328) suggest that double negation actually consists of one instance of clausal negation and one of constituent negation, and that multiple instances of a single type are not in fact possible. This may be so, but UD does not always enable us to make the distinction between these two types of negation explicit in the parse, so we use the feature `DoubleNegated=+` as a proxy for a more sophisticated analysis.

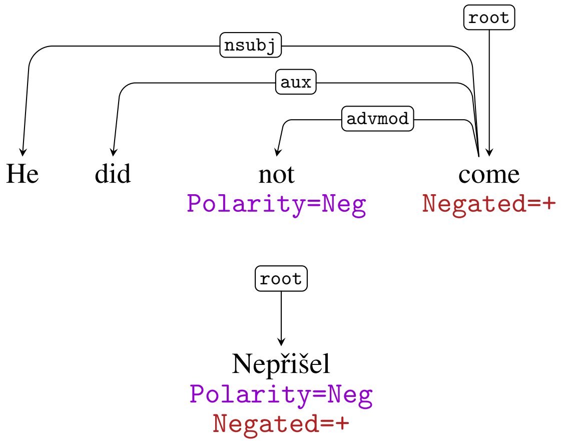


Figure 4: An English sentence and its Czech translation; the verb bears the `Negated=+` feature in both cases, even though `Polarity=Neg` appears in different places in the two languages

For future annotation projects, we would advocate for adding these annotations from the start, and preferably manually, in order to capture complex interactions or edge cases. But given a little language-specific information about different negative lemmata, they can also be added automatically, and we have prepared a script to do just this for the English treebanks, using the Graph Rewriting formalism (GREW: Guillaume et al. 2012; Bonfante et al. 2018).[9]

The `Negated=+` and `DoubleNegated=+` features allow for a straightforward translation from features to semantics: the presence of each of these features adds a new instance of negation to the semantics, one which scopes over the sub-tree rooted by the node bearing the annotation. This is a fairly

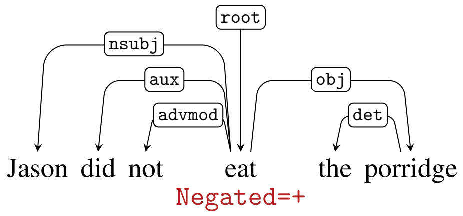
---

[9] This is available at `https://github.com/findlayjy/negation-in-UD`.

simplistic semantics for negation, but has the virtue of being able to be read directly off the UD tree. It will give rise to an interpretation which is consistent with but usually be weaker than the fully contextually-resolved meaning, as discussed in Section 3.

As mentioned above, a clausal negation like (17) entails the disjunction in (18):

(17)



(18) a. No eating event took place.
b. The Agent of the eating event was not Jason.
c. The Theme of the eating event was not the porridge.

We can rephrase this disjunction of negations as a conjunction under a single negation (via De Morgan's law):

(19) It was not the case that
a. an eating event took place, and
b. the Agent of the eating event was Jason, and
c. the Theme of the eating event was the porridge.

It is clear to see how this interpretation follows from the addition of the `Negated=+` feature, since these conjuncts correspond to the semantic contributions of the verb itself, the `nsubj` dependency, and the `obj` dependency, respectively, i.e. the subtree rooted in the *eat* node which hosts the feature.

Of course, context may well provide us with the necessary information to determine a stronger meaning: perhaps there *was* a porridge-eating event, but Jason was not the Agent, so that only the second conjunct is actually false. Nevertheless, the UD parse will never give us sufficient information to determine this, so it is best to provide a weak but at least consistent semantics: even if it is true that a porridge-eating event took place with Isabella as the Agent, this does not diminish the truth of (19), since the *conjunction* of statements remains false.

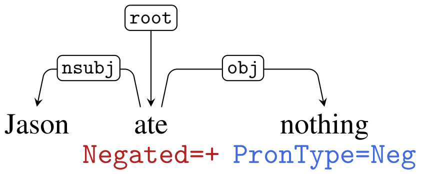
We apply exactly the same logic to more obviously ambiguous sentences like (20):

(20) Jason didn't eat the porridge because he was hungry.

Here the most plausible interpretation, given our world knowledge, is that it is only the Cause relation between Jason's hunger and his eating the porridge which is denied. But this state of affairs is still consistent with the claim that there was no porridge-eating event with Jason as Agent and his hunger as Cause. Since the syntax alone does not tell us which or how many of these conjuncts are true or false, it is reasonable for our syntax-driven interpretation to provide only a weak semantics, allowing the context to narrow it down if possible. It is much easier to see how context might add additional information that narrows the space of possible interpretations, for example, than how it could broaden an already too-narrow interpretation.

An advantage of adding our annotation to the verb itself rather than the explicitly negative element, which may be a dependent, is that often the verb itself should be under the scope of negation, even when the negation is introduced lower down in the tree. Sentence (21) is a case in point:

(21)



The sentence in (21) should most plausibly not mean there was an eating event with no Theme, for example, but rather that there was no eating event (with some Theme) at all. This requires us to introduce the negation meaning at the verb, rather than just at the quantifier *nothing*, which is where the only negation-related feature would be found in the vanilla annotation.

Notice that this implies that a negative quantifier like *nothing* makes the same contribution to the compositional semantics as a positive quantifier like *something*. This means that we can treat NC and non-NC languages the same in this regard; the only difference is a syntactic one, namely when it is that negative quantifiers induce the presence of a `Negated=+` or `DoubleNegated=+` feature on their governor. In a language like English, they always will; in a language like Spanish they will do so only when in subject position.

7 Conclusion

UD's features are restricted to the word level, and this often means that annotators base their decisions on form instead of meaning. This is problematic if one wants to derive a semantic representation from a UD one, since many features which 'ought' to directly express semantic attributes do not. Negation is a case in point, which we have focussed on here. The existing treebanks apply the annotation guidelines incompletely and inconsistently for the two negation-related features (`Polarity=Neg` and `PronType=Neg`), but even if they did not, these two features would be inadequate for straightforwardly expressing the semantics of negation. We propose the addition of two features, `Negated=+` and `DoubleNegated=+`, which directly encode when a predicate is *semantically* under negation. If these are added as part of the annotation process, they will enable a straightforward semantic interpretation of a UD parse in terms of negation, at least for the cases we discussed here, namely basic clausal negation, adjunct scope, and negative quantifiers.

Acknowledgements

We would like to thank Saeedeh Salimifar and Ahmet Yıldırım for useful discussions of Persian and Turkish data.

References

Guillaume Bonfante, Bruno Guillaume, and Guy Perrier. 2018. *Application of Graph Rewriting to Natural Language Processing*. Number 1 in Logic, Linguistics and Computer Science set. ISTE & Wiley, London.

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

Jamie Y. Findlay, Saeedeh Salimifar, Ahmet Yıldırım, and Dag T. T. Haug. 2023. Rule-based semantic interpretation for Universal Dependencies. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 47–57, Washington, D.C. Association for Computational Linguistics.

Bruno Guillaume, Guillame Bonfante, Paul Masson, Mathieu Morey, and Guy Perrier. 2012. Grew : un outil de réécriture de graphes pour le TAL (Grew: a graph rewriting tool for NLP) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 5: Software Demonstrations*, pages 1–2, Grenoble, France. ATALA/AFCP.

Martin Haspelmath. 1997. *Indefinite pronouns*. Oxford Studies in Typology and Linguistic Theory. Clarendon Press, Oxford.

Rodney Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge University Press, Cambridge.

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.

Barbara H. Partee. 1984. Nominal and temporal anaphora. *Linguistics and Philosophy*, 7(3):243–286.

Wessel Poelman, Rik van Noord, and Johan Bos. 2022. Transparent semantic parsing with Universal Dependencies using graph transformations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4186–4192, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Adam Przepiórkowski and Agnieszka Patejuk. 2015. Two representations of negation in LFG: evidence from Polish. In *Proceedings of the LFG15 Conference*, pages 322–336, Stanford, CA. CSLI Publications.

Siva Reddy, Oscar Täckström, Slav Petrov, Mark Steedman, and Mirella Lapata. 2017. Universal semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 89–101, Copenhagen, Denmark. Association for Computational Linguistics.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023. PARSEME meets Universal Dependencies: getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*, 9(1).

Leonie Weissweiler, Nina Böbel, Kirian Guiller, Santiago Herrera, Wesley Samuel Scivetti, Arthur Lorenzi, Nurit Melnik, Archna Bhatia, Hinrich Schütze, Lori Levin, Amir Zeldes, Joakim Nivre, William Croft, and Nathan Schneider. 2024. UCxn: Typologically-informed annotation of constructions atop Universal Dependencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16919–16932, Torino, Italia. ELRA and ICCL.

Hedde Zeijlstra. 2020. Negative quantifiers. In Viviane Déprez and M. Teresa Espinal, editor, *The Oxford handbook of negation*, pages 426–440. Oxford University Press.

Daniel Zeman et al. 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.

TreEn: A Multilingual Treebank Project on Environmental Discourse

Adriana Pagano
Universidade Federal de Minas Gerais
`apagano@ufmg.br`

Patricia Chiril
University of Chicago
`pchiril@uchicago.edu`

Elisa Chierchiello
Università degli Studi di Torino
`elisa.chierchiello@unito.it`

Cristina Bosco
Università degli Studi di Torino
`cristina.bosco@unito.it`

Abstract

The past few decades have seen a simultaneous increase in both the complexity of environmental debates in the media and the language people use to engage with them. While linguistic and communication studies have been pursued on this discourse, the development of computational linguistic tools and resources dedicated to support its analysis and interpretation is still very incipient. For one, no morphological and syntactic resources specific to the environmental domain can be found on major platforms and repositories. This paper introduces TreEn, a multilingual treebank project in progress which compiles texts on environmental discourse produced in different communication contexts, focusing on a set of non-governmental reports included in the first release of TreEn. It reports on the parallel component of the project and discusses issues faced during sentence-level alignment between original and translated texts, annotation of texts following UD guidelines, and labeling entities drawing on an ontology of environmental-related topics. This novel resource is expected to support environmental discourse analysis by providing morphological and syntactic data to enable cross-language and cross-cultural comparisons based on the semantics of the entities annotated in the treebank.

1 Introduction

In 1894, the Swedish Nobel Prize-winning chemist and physicist Svante Arrhenius formulated a climate model that correlated rising atmospheric CO_2 levels with glacier melting, pointing for the first time to human activity as the cause of global warming (Kolbert, 2024). More than a century later, Arrhenius' model remains the subject of heated debate, not only in academia but also in communication and social media, yielding a massive amount of environmental discourse and posing a challenge to analysts and consumers alike.

Environmental discourse is often characterized by its grammatical intricacy (Halliday, 1992), accountable for the emergence of its domain-specialized lexicon. It encompasses different genres and registers, whereby meaning is construed by different stakeholders. Moreover, there is the impact of languages in contact, as translation from major into minor languages is both a frequent means of production and a main source of variation.

As far as the computational analysis of environmental discourse is concerned, resources have only recently begun to be developed, driven by the poor performance of Large Language Models (LLMs) in this domain (Webersinke et al., 2022) and the need for dedicated pre-trained models to efficiently process such texts (Thulke et al., 2024). However, corpora and datasets remain scarce. Stede and Patz (2021) reviewed available datasets on climate-change discourse and highlighted the predominance of news and social media texts in NLP analysis, with approaches primarily relying on corpus linguistics (raw frequency counts, collocations) and out-of-the box tools for topic modelling, sentiment analysis and network analysis. Notably, the authors did not report the existence of any annotated resources, such as treebanks. A more recent

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 80–96
August 27, 2025 ©2025 Association for Computational Linguistics

initiative, however, introduced a multimodal corpus of academic articles and texts from websites (including the International Panel on Climate Change (IPCC), Greenpeace International, and Greenpeace Germany), enriched with metadata and annotations in order to allow for more in-depth discourse analysis (Bartsch et al., 2023).

In line with Bartsch et al. (2023), our work pursues the development of annotated corpora to support the study of environmental discourse. We adopt a multilingual perspective which can inform both theoretical and applied studies in the fields of language comparison and typology, translation studies, discourse analysis, to name but a few. It also provides a valuable resource for computational linguists seeking to fine-tune general-purpose models and create a benchmark for evaluating LLMs.

As part of this wider project, this paper introduces TREEN (Treebanks for Environment), the first multilingual treebank project, comprising treebanks of comparable and parallel texts on environmental discourse. More specifically, it describes the parallel component of TREEN currently including texts extracted from a non-governmental report in four different languages (English, Brazilian Portuguese, Italian and Romanian), annotated for morphology and syntax following the Universal Dependencies (UD) guidelines and enriched with entity annotations drawing on the GEneral Multilingual Environmental Thesaurus (GEMET).[1]

The remainder of the paper is organized as follows. Section 2 provides a brief survey of related work. Section 3 describes the data collection process for the creation of the parallel component of TREEN and outlines the methodological decisions made during text alignment. It also details the annotation guidelines applied to this resource, namely UD and Universal Named Entity Recognition (U-NER). Section 4 presents the statistics of our parallel treebanks and characterizes each of the four languages compiled in TREEN, highlighting key insights from a cross-linguistic comparison. Finally, Section 5 presents concluding remarks and outlines directions for future work.

2 Related Work

Most of the resources hitherto developed for the analysis of environmental discourse consist of raw corpora or datasets labelled for classification tasks such as Sentiment Analysis. Nevertheless, as noted by Ibrohim et al. (2023), Sentiment Analysis resources for environmental texts are currently extremely limited in all languages. While some corpora on environmental topics are available for English, resources in other languages are scarce. Even in dedicated emerging disciplines, as in Ecological Discourse Analysis (EAD), studies and resources are still at an early stage and mostly concentrated in specific countries and languages, as highlighted in a very recent systematic review by Song et al. (2025). Interestingly, some of the countries that are key players in environmental debates, as is the case of Brazil, are not represented in EAD at all.

In computational linguistics, there is growing interest in **environmental data**, as evidenced by the increasing number of events on this topic in recent years, such as the Workshop on *Ecology, Environment and Natural Language Processing*[2] (Basile et al., 2025), *ClimateNLP* (2024 - 2025),[3] and the *Tackling Climate Change with Machine Learning Workshop*,[4] which has been held regularly since 2019 in top-tier conferences. In addition, smaller-scale challenges and data competitions (such as those hosted on platforms like Kaggle[5]) or individual efforts by researchers have made annotated social media corpora related to climate and environmental issues available. These corpora cover a range of tasks related to environmental discourse, with a particular focus on aspects of climate change. Some works focus on verifying whether a given text contains environmental claims (or pertains to climate change) (Varini et al., 2020; Diggelmann et al., 2020; Stammbach et al., 2023), or facilitate expert-informed retrieval from corporate climate disclosures (Schimanski et al., 2024), while others examine the topics discussed within such narratives (Dahal et al., 2019; Duong et al., 2022; Effrosynidis et al., 2022; Vaid et al., 2022). Other studies focus on detecting and analyzing attitudes towards global warming (Luo et al., 2020), as well as identifying neutralization techniques (i.e., rhetorical strategies used in climate change skepticism to justify inaction or promote alternative views) (Bhatia et al.,

---

[1] https://www.eionet.europa.eu/gemet/en/about/

[2] https://econlpws2025.di.unito.it/

[3] https://nlp4climate.github.io/

[4] https://www.climatechange.ai/events#past-events

[5] One such dataset aggregates climate-related tweets, each manually labeled by three annotators as: *news* (factual news about climate change), *pro* (supports the human-caused climate change belief), *neutral* (no stance), or *anti* (rejects the human-caused climate change) belief: https://www.kaggle.com/competitions/climate-change-edsa2020-21/overview.

2021). Another important line of research targets climate change contrarianism, denial, and disinformation (Coan et al., 2021; Piskorski et al., 2022; Bhatia et al., 2020).

As previously mentioned, one of the few **annotated resources** reported in the literature is the InsightsNet Climate Change Corpus (ICCC) introduced by Bartsch et al. (2023). The ICCC comprises academic articles and texts from websites, and is annotated for metadata, morphology, syntax, and Named Entities. The authors used spaCy and StanzaCoreNLP for automatic annotation and do not report any manual check of the output. While the corpus includes texts in English along with their German translations (for most of the texts), no alignment between the originals and translations was carried out.

The availability of parallel texts is relevant for several NLP tasks and linguistic fields. With regard to **parallel treebanks**, there are few available in the UD repository. Among them, the Parallel Universal Dependencies (PUD) treebanks, created for the CoNLL 2017 shared task on *Multilingual Parsing from Raw Text to Universal Dependencies*[6] and including 15 languages (Zeman et al., 2017). For English and Swedish, we can cite LinEs (Ahrenberg, 2015), while Italian, English and French are aligned in ParTUT (Sanguinetti and Bosco, 2014). The fact that these treebanks have been made compliant with the UD guidelines enhances their comparability and makes them a particularly valuable resource for linguists and computer scientists. None of these treebanks, however, is dedicated to environmental discourse.

Several **unannotated parallel corpora** are also available through the OPUS project platform.[7] Among them, some parallel corpora can incidentally include language related to environmental topics, as is the case of EUROPARL,[8] which includes debates on environmental issues held in the European Parliament. However, these corpora are mostly unannotated and no dedicated subcorpus is readily available.

Another important line of work related to multilingual resources for comparability analysis is **entity recognition**, which is currently being promoted by the Universal NER (UNER) project which especially focuses on named entities, usually introduced in text by proper nouns. This community-driven initiative provides gold-standard named entity recognition annotations in a wide variety of typologically and geographically different languages, adding a consistent NER layer to UD corpora (Mayhew et al., 2024). This provides a valuable benchmark for multilingual NER, especially for low-resource languages and for testing inter-annotator agreement and tag distribution across domains.

This general purpose initiative is especially inspiring for our project since a key challenge in **domain-specific NER**—and particularly in **environmental NER**— is the lack of annotated corpora and benchmarks dedicated to the environmental domain. As in most specific domains, entities other than persons, organizations, and locations are relevant in environmental discourse and they are realized by common nouns and noun phrases, some entities being nested inside larger ones, which are not covered by general-purpose NER tools.

To map domain entities, researchers usually turn to **domain ontologies**, which provide a common ground for naming classes and subclasses and the relations holding between them. In the environmental domain, GEMET (General Multilingual Environmental Thesaurus)[9] provides structured multilingual identifiers and relations between environmental concepts (European Environment Agency, 2024). GEMET has been used for tasks such as concept mapping and normalization of environmental terminology; however, its integration into NER pipelines is still experimental. Nevertheless, GEMET is a solid framework for multilingual annotation, which accounts for our decision to adopt it in our treebank annotation, as detailed in the following section.

3 Corpus Compilation and Annotation

To compile our parallel corpus, we downloaded the original version of *WWF's*[10] *2024 Living Planet Report*, available in multiple languages, including English,[11] Brazilian Portuguese,[12] Italian,[13] and Romanian,[14] the languages currently annotated in

---

[6] `http://universaldependencies.org/conll17/`

[7] `https://opus.nlpl.eu/legacy/`

[8] `https://opus.nlpl.eu/legacy/Europarl.php`

[9] `https://www.eionet.europa.eu/gemet/en/about/`

[10] The *World Wildlife Found for Nature* is the global leading organization for the nature conservation with offices in many countries.

[11] `https://livingplanet.panda.org/en-US/`

[12] `https://livingplanet.panda.org/pt-BR/`

[13] `https://www.wwf.it/cosa-facciamo/pubblicazioni/living-planet-report/`

[14] `https://wwf.ro/campanii/raportul-planeta-vie-2024/`

ALIGN	ENGLISH	BRAZILIAN PORTUGUESE
1-2	For example, in Colombia, a 2019 law requires that solar and wind projects transfer a percentage of their sales to communities within the project's area of influence while the Philippines' Renewable Energy Act requires that 80% of project royalties be directed toward subsidizing power costs in affected communities.	Por exemplo, na Colômbia, uma lei de 2019 exige que os projetos de energia solar e eólica transfiram um percentual de suas vendas para as comunidades dentro da "área de influência" do projeto.
		Já a Lei das Energias Renováveis das Filipinas exige que 80% dos royalties dos projetos sejam direcionados para subsidiar os custos de energia nas comunidades afetadas.
2-1	The most recently published plan includes areas of intact ecosystems for human livelihoods and well-being.	O plano recém-publicado inclui áreas de ecossistemas intactos para a subsistência e o bem-estar humanos e teve o cuidado de garantir que a expansão das áreas protegidas contribua para os objetivos de desenvolvimento da África do Sul, fornecendo importantes serviços ecossistêmicos às pessoas.
	Care was taken to ensure the expansion of protected areas would contribute to South Africa's development goals by providing important ecosystem services to people.	

Table 1: Examples of one-to-many and many-to-one sentence alignment.

our parallel treebanks. Other languages for which the WWF report is available are expected to be included in our resource using and the same methodology in a future expansion of the TREEN project.

A comparison of the four versions of *WWF's 2024 Living Planet Report* showed that they all featured the same images and text content. For the purposes of treebank development, images and infographics were removed and plain text files were created. All bibliographic references and footnotes were also removed.

3.1 Alignment

In general, the development of aligned parallel resources is considered a very time-consuming task, mostly due to alignment issues. This poses a challenge when alignment is pursued at sentence or more fine-grained level, as meanings can be realized at different levels in different languages.

In TREEN, alignment was performed at sentence level. Sentences were extracted from each file and pasted in their sequential order onto a spreadsheet, a language pair (original and translated text) per tab. Sentence alignment was manually checked and when no one-to-one alignment was obtained, extra rows were added. This was the case when a single sentence was split into two or more sentences in the translated text or the opposite, i.e. two or more sentences in the original in English were translated as a single one in one of the other languages. Table 1 shows how these two cases were dealt with. The first line (1-2) shows a sentence in English that is split in two sentences in Brazilian Portuguese, while the second line (2-1) shows a couple of English sentences encompassed in a single sentence in Brazilian Portuguese. Often the diversity in segmentation is triggered by punctuation and how this is deployed in the translated text. For example, the two English sentences *"Nature narratives."* and *"Using indicators to understand change over different timescales"*, which are adjacent in the original text of the *WWF's 2024 Report*, are encompassed in the Italian *"Raccontare il declino della natura: utilizzo di indicatori per comprendere il cambiamento su scale temporali diverse."* since the translator interpreted the end of the first sentence as colons.

Sequential id numbers assigned to source sentences in English were used as an alignment reference. This means that in cases of no one-to-one sentence correspondence, the id of the English sentence is assigned to the corresponding sentences in the other languages. Each sequential id number is, moreover, preceded by three further digits. Following the same encoding strategy applied in PUD, we used the digits on the second and third position of each id to encode the original language of the sentence (e.g. *01* for English) and used the first digit to encode the genre of text. In our treebanks, the texts are extracted from a non-governmental report, as they were produced by an activist organization. Hence, the first digit *a* stands for activism. We are also planning a future extension of our resource to other types of texts, namely generated by institutions (governments and policy-makers) and social media users. *i* will be used for governmental institutions and *s* for social media texts.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Transforming	transform	VERB	_	_	0	root	_	_
2	the	the	DET	_	_	4	det	_	_
3	energy	energy	NOUN	_	_	4	compound	_	gemet_id=2712 group=ENERGY
4	system	system	NOUN	_	_	1	obj	_	_

Table 2: Example sentence in CoNLL-U format with ontology-based annotation.

3.2 Morphological and syntactic annotation

As already mentioned, there are only a few resources compiling texts on environmental issues and, fewer still, annotated for morphology and syntax. To provide a comprehensive resource for linguistic analysis and computational uses, we annotated our texts following the UD guidelines (Nivre et al., 2020), which is the *de facto standard* for treebanks. Table 2 shows an example from the `WWF24-Eng` in CoNLL-U format.

To parse our texts following UD guidelines, we used UDPipe parser (Straka, 2018) trained on the best models available for the four languages currently included in our dataset. For English, we used `english-gum-ud-2.15-241121`, for Italian we used `italian-isdt-ud-2.15-241121`, for Romanian we used `romanian-rrt-ud-2.15-241121` and for Brazilian Portuguese, `portuguese-petrogold-ud-2.15-241121`. A sample of 116 sentences of the output generated by UDPipe (corresponding to the preamble to the 2024 WWF Living Planet Report) was uploaded to the Arborator Grew (Guibon et al., 2020) tool and manually checked to have some hints about the correctness of the data and enable preliminary evaluation of their quality. While this validation is suitable for the purpose of the first release of TREEN, it must be extended to a larger set of data in the future development of the treebank.

The annotated sentences will be submitted to be part of UD repository.

3.3 Entity annotation

To annotate environmental content at the level of specific concepts, we designed a pipeline that combines NER with ontology-based matching. Rather than relying on general-purpose resources, our approach is centered entirely on the GEMET ontology, as it provides unique identifiers and thematic groupings for environmental concepts. Environmental discourse involves a wide range of entities that go beyond traditional Named Entities such as *places* or *organizations*. In this domain, the most relevant elements are often abstract or thematic concepts — such as *biodiversity*, *climate change*, or *ecosystem resilience* — which are central to how environmental issues are discussed and framed.

A key insight of our method is the strong connection between syntactic structure and an ontology-based annotation. The UD framework provides detailed syntactic information that allows us to extract meaningful expressions in a principled way. By focusing on noun phrases and their modifiers — as determined by dependency relations like `compound`, `amod`, and `nmod` — we are able to detect linguistically grounded candidate entities.

This syntactic scaffolding is essential: it ensures that entities are not just found, but are consistently defined across different languages and sentence structures. Once identified, these expressions are matched to GEMET concepts through exact or partial alignment, adding an ontology-based layer to the treebank. This enriched structure opens the door to a wide range of multilingual applications, including cross-lingual entity propagation and concept-level analysis (see Figure 1).

Entity Extraction. As a first step, we identified potential entities in each sentence by extracting *noun phrases* (NPs). This was done by combining part-of-speech tagging with syntactic dependency information from the CoNLL-U files. We took into account both expressions composed of a nominal head and some modifiers (not properly multi-word expressions in the formal sense, such as “climate change”) and *single nouns or proper nouns* (like “biodiversity”). To extract meaningful phrases, we used a simple set of heuristics that detect noun heads along with their modifiers (e.g., `compound`, `amod`, `nmod` dependencies), while discarding generic or weak elements such as quanti-

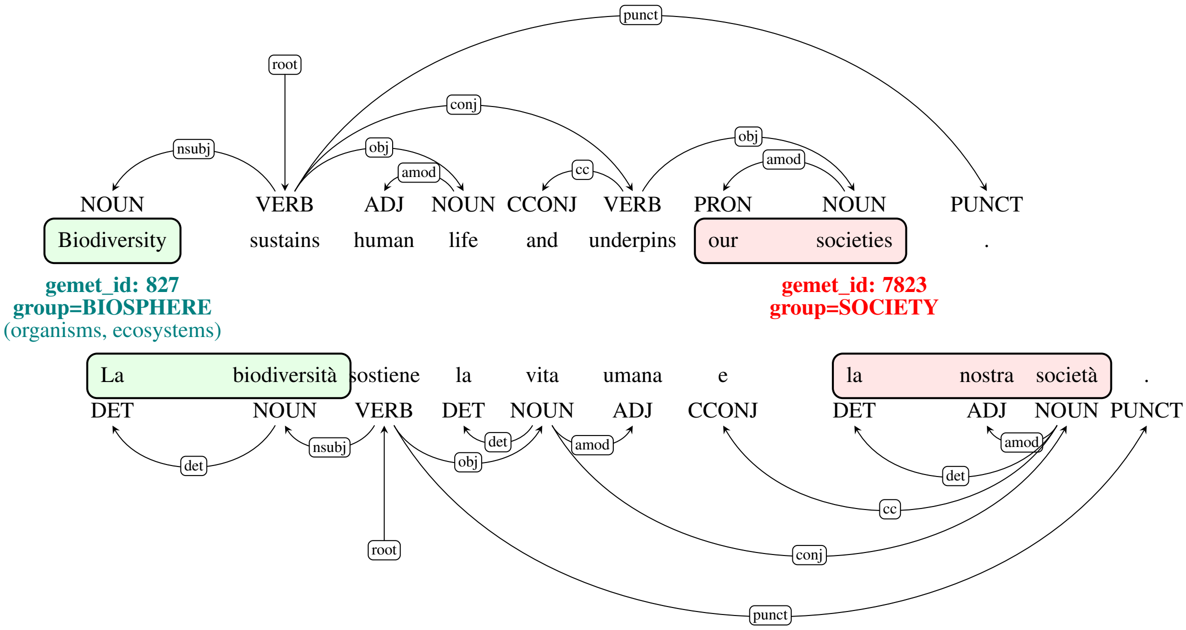


Figure 1: Example of dependency-based annotation of two parallel sentences, in English and Italian, enriched with `GEMET` ontology labels in the CoNLL-U format.

fiers (e.g., many, several) and evaluative adjectives (e.g., important, good).

Ontology Matching in English. Each candidate noun phrase was first normalized and then compared against the set of preferred labels provided by the GEMET ontology, using exact string matching. When a match was found, we associated the corresponding `concept ID` and its `group name` to the phrase. These annotations were added to the `MISC` column in the CoNLL-U file, like in Table 2 (other examples are in Table 4, 5 and 6 in the Appendix). In the case of multiword expressions (e.g., *"climate change"*), the annotation is consistently propagated across all tokens that make up the GEMET entity, ensuring that the concept is clearly represented throughout the span (see e.g. *"greenhouse gas"* in Table 6).

Multilingual Propagation. After annotating the English sentences, we transferred the entity annotations to the aligned sentences in the other languages in our dataset (Italian, Brazilian Portuguese, and Romanian). This projection process relies on sentence-level alignment between English and the target language. For each GEMET concept identified in the English sentence, we verified whether a corresponding noun or noun phrase appeared in the aligned sentence. We relied on two strategies: (i) direct translation of the English label into the target language using the `Google Translate API`[15], and (ii) lexical overlap, where we checked whether the translated term (or its components) matched any of the extracted noun phrases or their tokens in the target sentence. This allowed us to transfer GEMET annotations across languages even when the syntactic realization differs slightly. When a match was found, we assigned the same GEMET concept ID to the relevant tokens in the target language, ensuring consistent cross-lingual annotation.

These experiments were implemented in **Python**,[16] using the `spaCy`[17] and `pandas`[18] libraries for linguistic processing and data manipulation. The GEMET ontology was used as the sole reference for concept matching and a semantics oriented labelling.

An example of the annotation of two corresponding sentences, respectively from WWF-ENG and WWF-ITA, is shown in Figure 1 (for the CoNLL-U of these sentences see Table 4 and 5 in the Appendix). The two sentences were parsed following the UD framework and then annotated for the entities. The upper and lower parts of the figure show

---

[15] `https://cloud.google.com/translate`
[16] `https://www.python.org/`
[17] `https://spacy.io/`
[18] `https://pandas.pydata.org/`

the sentence in English and in Italian respectively, both associated with their parts of speech. The entities extracted according to GEMET can be seen as a semantic bridge between the syntactic structures. In the example, "*biodiversity*", "*life*", and "*societies*" were matched against the GEMET ontology. "*biodiversity*" was linked to concept ID `827` in the `BIOSPHERE` group, while "*society*" was matched to concept ID `7823` in the `SOCIETY` group. On the other hand, not all terms and meaningful expressions extracted as a candidate noun phrase have a direct match in the ontology: for example, "*human life*" did not match any GEMET concept.

Once there is matching, information is added to the `MISC` column of the CoNLL-U file and can be visualized in the syntactic tree. This enriched representation brings together syntactic structure and ontological information, creating a solid foundation for further tasks like recognizing environmental entities or transferring annotations across languages. Table 2 shows how the annotation of an entity is integrated in the CoNLL-U format.

4 Results

We computed the number of sentences, words, and lemmas in the WWF24 parallel corpora using a custom `Python` pipeline built on `pandas` and `spaCy` language-specific models (`{en,pt,it,ro}_core_web_sm`). Table 3 presents the resulting distributions.

As previously mentioned, some of the sentences in the original text in English were split in the translated texts, which accounts for the higher number of sentences in Italian and much higher in Brazilian Portuguese when compared with the source text in English. Regarding number of tokens, the translated texts exhibit a higher number, also accountable for by lexical rendition of some English terms and typological differences between the languages, such as the strong reliance of Romance languages on determiners and prepositions to construe noun phrases. This can be seen, for instance, in terms such as *"greenhouse gas emissions"* (3 tokens) translated into Italian as *"le emissioni di gas serra"* (5 tokens), Brazilian Portuguese as *"as emissões de gases de efeito estufa"* (7 tokens) and Romanian as *"emisiile de gaze cu efect de seră"* (7 tokens). The higher number of unique words and lemmas in the translated texts is also impacted by typological differences as well as by translators' choices in rendering the same word and lemma in English by different synonyms of the equivalent word in their languages. With regard to average sentence length, results point to relatively long sentences, which can be explored as a variable contributing to the overall complexity of this kind of discourse.

4.1 Annotated samples

We analyzed the human-curated annotated sample of 116 sentences from the original text in English and their translation into Romanian, Italian and Brazilian Portuguese (for the full distribution of POS and dependency relations see Table 7 and Table 8 in the Appendix). Both `POS` and `deprel` tags reflect analogous and different annotation guidelines in the four languages. Regarding POS tags, for instance, a source of variation is the annotation of geographical names and institutions as proper nouns (`PROPN`) in some of the languages, the relation *flat* holding between them, and as common nouns (`NOUN`) in others with *nmod* relations holding between them. This result shows different guidelines for English and the Romance languages. With regard to analogous guidelines, differences in the count of `DET`, for example, show typological differences, with English using relatively few determiners when compared to Italian and Brazilian Portuguese, and Romanian expressing determinateness through different morphological markers.

Regarding deprels, most noun phrases in English are annotated with the deprel *compound* while the Romance languages annotate them with the *nmod* tag. Finally, it is worth noting that the high number of *conj* relations in the four languages is an indicator of more complex syntax, also correlated to the average sentence length mentioned above.

4.2 Annotation challenges

Manual check of the parsed output revealed some of the main challenges in automatic annotation. These have to do with nested noun phrases and coordinated phrases and clauses. Our results are in line with those of the studies about the dissemination and communication about environmental matters. The recent literature has indeed problematized the effectiveness of such discourse due to the complexity in the content of this kind of texts (see e.g. (Bosco et al., 2023)). By providing a fine-grained analysis of the language used to communicate environmental issues, computational linguistics can provide crucial information on how individuals, groups of people or entire societies are coping with environmental issues.

Language	sentences	tokens	unique words	lemmas	average sentence length
English	1,042	23,462	3,163	2,346	22,5
Brazilian-Portuguese	1,063	26,471	3,790	2,598	24,9
Italian	1,048	26,976	3,988	2,783	25,7
Romanian	1,042	25,830	4,699	2,968	24,8

Table 3: Basic statistics per language.

As far as more precise syntactic phenomena, nested noun phrases are very common in scientific domains and consist of a noun phrase nested into another, sometimes having a coordinated relation as well. An example retrieved from our corpus is *"a combination of pine bark beetle infestation and more frequent and ferocious forest fires"*. In this example (as shown in Figure 2 in the Appendix), there is coordination between *"infestation"* and *"fires"*, which are both heads of noun phrases. The noun *"infestation"* has three nouns as pre-modifiers, which need to be annotated as a sequence of *compound* relations, where *"infestation"* stands in a *compound* relation with *"beetle"*, which, in turn, is in a *compound* relation with *"bark"*, in turn, in a *compound* relation with *"pine"*. In the case of Portuguese, Italian, and Romanian texts, as seen in Figure 3, 4, and 5, noun modification is established through the dependency relation *nmod*. Interestingly, all four language models used in UDpipe failed to assign the correct *conj* relation between *"infestation"* and *"fires"* and had to be manually edited.

Another challenging case for parsers is discriminating between *obl* and *nmod* relations. For example, in the English sentence *"In the biosphere, the mass die-off of coral reefs would destroy fisheries and storm protection for hundreds of millions of people living on the coasts"*, the UDpipe parser assigned an *obl* relation between *"destroy"* and *"for hundreds of millions of people [...]"*, instead of an *nmod* relation between *"protection"* and *"for hundreds of people"*, which is semantically more plausible in this context.

5 Conclusion and Future Work

This paper presented the first steps in the development of a novel set of aligned treebanks within the Multilingual Treebank Project TREEN, which compiles texts on environmental discourse produced by different stakeholders (institutions, citizens and activists) in different communication contexts. The data included in the treebanks were extracted from the *WWF's 2024 Living Planet Report* in English, Italian, Romanian and Brazilian Portuguese. They were annotated for morphology and syntax in UD format and enriched with the annotation of the entities involved in the discourse according to the GEMET ontology for environmental topics. This methodology is designed to apply to other texts in the ongoing expansion of TREEN also considering e.g. data alignment at fine-grained levels and an ontology matching that takes into account form variation (such as synonyms or paraphrases).

These treebanks and the future development of the project are intended to fill a gap regarding resources dedicated to the environmental discourse in order to offer the community the possibility of conducting a more in-depth analysis of environmental discourse. This is expected to open new perspectives to more reliably study the different narratives present in different languages, mirroring different cultures and perceptions in the discourse about environmental issues. The availability of morphological and syntactic annotation can be very helpful for detecting different narratives, e.g. allowing the identification of different tenses and forms of verbs as indicators of different ways of narrating events (realis vs. irealis). Complementarily, annotating named entities and linking the syntactic structures in the different languages to a common ontological framework such as GEMET helps us in the semantics oriented characterization of the discourse about the environment in different cultures.

Last but not least, we expect to use morphological and syntactic annotation and a deeper analysis of the annotated data according to several different perspectives to explore the level of intricacy and complexity of environmental discourse produced by the different stakeholders, which can yield insightful and actionable information to develop more inclusive forms of environmental discourse.

Limitations

The main aim of this study is to introduce the first release of a novel resource. This release comprises 116 sentences, with human-curated alignment and annotation, the remaining sentences being currently under human validation. This resource is part of a wider project for the study of the discourse about environmental matters. The texts collected in the component of TREEN here reported are drawn from a single source: one non-governmental association (WWF) and in four languages only. We are working on the expansion of the parallel treebanks to include more languages and to incorporate other stakeholders.

While the proposed methodology yields promising results, several aspects of the pipeline can be improved in future work.

In particular, as far as the matching between cited concepts and entities in GEMET is concerned, this is based on exact string matching, that is, the ontology matching relies on exact string comparison, which can miss variations, synonyms or paraphrases.

Also, the coverage of the ontology is partial: although GEMET is a rich resource, it does not cover all relevant environmental expressions, especially newer or more specific domain terms.

The translation strategy applied for the multilingual propagation of entity annotation may also produce inaccurate or overly generic results, affecting the mapping accuracy. The quality of the automatic translation may be improved using tools other than Google Translate.

Additionally, our entity extraction procedure is restricted to noun phrases. As a result, relevant information conveyed through verbs, adjectives, or multi-clause constructions may be overlooked. Incorporating broader syntactic patterns or using semantic role labeling could help address this limitation.

Acknowledgments

Adriana Pagano has a grant from the National Council for Scientific and Technological Development (CNPq 404722/2024-5; 313103/2021-6) and Minas Gerais State Agency for Research and Development (FAPEMIG).
The work of Elisa Chierchiello is funded by the International project *CN-HPC-Spoke1-Future HPC & Big Data, PNRR MUR-M4C2*.

References

Lars Ahrenberg. 2015. Converting an English-Swedish parallel treebank to Universal Dependencies. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 10–19, Uppsala, Sweden. Uppsala University, Uppsala, Sweden.

Sabine Bartsch, Changxu Duan, Sherry Tan, Elena Volkanovska, and Wolfgang Stille. 2023. The insightsnet climate change corpus (iccc). In *BTW 2023*, pages 887–900. Gesellschaft für Informatik e.V., Bonn.

Valerio Basile, Cristina Bosco, Francesca Grasso, Muhammad Okky Ibrohim, Maria Skeppstedt, and Manfred Stede, editors. 2025. *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*. University of Tartu Library, Tallinn, Estonia.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2020. You are right. i am alarmed–but by climate change counter movement. *arXiv preprint arXiv:2004.14907*.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2021. Automatic classification of neutralization techniques in the narrative of climate change scepticism. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2167–2175, Online. Association for Computational Linguistics.

Cristina Bosco, Muhammad Okky Ibrohim, Valerio Basile, and Indra Budi. 2023. How green is sentiment analysis? environmental topics in corpora at the university of turin. In *Proceeding of CLiC-it 2023*, Venice, Italy.

Travis G Coan, Constantine Boussalis, John Cook, and Mirjam O Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific reports*, 11(1):22320.

Biraj Dahal, Sathish AP Kumar, and Zhenlong Li. 2019. Topic modeling and sentiment analysis of global climate change tweets. *Social network analysis and mining*, 9:1–20.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Cuc Duong, Qian Liu, Rui Mao, and Erik Cambria. 2022. Saving earth one tweet at a time through the lens of artificial intelligence. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE.

Dimitrios Effrosynidis, Alexandros I Karasakalidis, Georgios Sylaios, and Avi Arampatzis. 2022. The climate change twitter dataset. *Expert Systems with Applications*, 204:117541.

European Environment Agency. 2024. Gemet - general multilingual environmental thesaurus. `https://www.eionet.europa.eu/gemet/en/about/`. Accessed: 2025-04-11.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5293–5302, Marseille, France. European Language Resources Association.

Michael AK Halliday. 1992. New ways of meaning: The challenge to applied linguistics. *Thirty years of linguistic evolution*, pages 59–95.

Muhammad Okky Ibrohim, Cristina Bosco, and Valerio Basile. 2023. Sentiment analysis for the natural environment: A systematic review. *ACM Computing Surveys*, 56(4).

Elizabeth Kolbert. 2024. *H is for Hope: Climate Change from A to Z*. Oneworld Publications.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. *arXiv preprint arXiv:2010.15149*.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Suppa, Hila Gonen, Joseph Marvin Imperial, Börje Karlsson, Peiqin Lin, Nikola Ljubešić, Lester James Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2024. Universal NER: A gold-standard multilingual named entity recognition benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337, Mexico City, Mexico. Association for Computational Linguistics.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, and other UD contributors. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 4034–4043. European Language Resources Association (ELRA).

Jakub Piskorski, Nikolaos Nikolaidis, Nicolas Stefanovitch, Bonka Kotseva, Irene Vianini, Sopho Kharazi, Jens P Linge, and 1 others. 2022. Exploring data augmentation for classification of climate change denial: Preliminary study.

Manuela Sanguinetti and Cristina Bosco. 2014. PartTUT: The Turin University Parallel Treebank. In Basili, Bosco, Delmonte, Moschitti, and Simi, editors, *Harmonization and development of resources and tools for Italian Natural Language Processing within the PARLI project*. Springer Verlag.

Tobias Schimanski, Jingwei Ni, Roberto Spacey Martín, Nicola Ranger, and Markus Leippold. 2024. ClimRetrieve: A benchmarking dataset for information retrieval from corporate climate disclosures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17509–17524, Miami, Florida, USA. Association for Computational Linguistics.

Xu Song, Kesumawati A Bakar, Azlan Abas, and Wan Fatimah Solihah Wan Abdul Halim. 2025. A systematic literature review on ecological discourse analysis (2014–2023). *Journal of World Languages*, (0).

Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada. Association for Computational Linguistics.

Manfred Stede and Ronny Patz. 2021. The climate change debate and natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 8–18, Online. Association for Computational Linguistics.

Milan Straka. 2018. Udpipe 2.0 prototype at conll 2018 ud shared task. In *Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies*, pages 197–207.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, Taylor Tragemann, Katie Nguyen, Ariana Fowler, Andrew Stanco, Jon Gabriel, Jordan Taylor, Dean Moro, Evgenii Tsymbalov, Juliette de Waal, and 7 others. 2024. ClimateGPT: towards AI synthetizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646*.

Roopal Vaid, Kartikey Pant, and Manish Shrivastava. 2022. Towards fine-grained classification of climate change related social media text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 434–443, Dublin, Ireland. Association for Computational Linguistics.

Francesco S Varini, Jordan Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. 2020. Climatext: A dataset for climate change topic detection. *arXiv preprint arXiv:2012.00483*.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. ClimateBert: A pretrained language model for climate-related text. *ArXiv*, abs/2110.12010.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava

Hlavacova, Václava Kettnerová, Zdenka Uresova, and 43 others. 2017. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

A Appendix

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Biodiversity	biodiversity	NOUN	_	_	2	nsubj	_	gemet_id: 827 group=BIOSPHERE (organisms, ecosystems)
2	sustains	sustain	VERB	_	_	0	root	_	_
3	human	human	ADJ	_	_	4	amod	_	_
4	life	life	NOUN	_	_	2	obj	_	
5	and	and	CCONJ	_	_	6	cc	_	_
6	underpins	underpin	VERB	_	_	2	conj	_	_
7	our	our	ADJ	_	_	8	nmod:poss	_	_
8	societies	society	NOUN	_	_	6	obj	_	gemet_id: 7823 group=SOCIETY
9	.	.	PUNCT	_	_	2	punct	_	_

Table 4: Annotation of GEMET entities in the English sentence in Figure 1.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	La	il	DET	_	_	2	det	_	_
2	biodiversità	biodiversità	NOUN	_	_	3	nsubj	_	gemet_id: 827 group=BIOSPHERE (organisms, ecosystems)
3	sostiene	sostenere	VERB	_	_	0	root	_	_
4	la	il	DET	_	_	5	det	_	_
5	vita	vita	NOUN	_	_	3	obj	_	_
6	umana	umano	ADJ	_	_	5	amod	_	
7	e	e	CCONJ	_	_	10	cc	_	_
8	la	il	DET	_	_	10	det	_	_
9	nostra	nostro	DET	_	_	10	det:poss	_	_
10	società	società	NOUN	_	_	5	conj	_	gemet_id: 7823 group=SOCIETY
11	.	.	PUNCT	_	_	3	punct	_	_

Table 5: Annotation of propagated GEMET entities in the Italian sentence in Figure 1.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Nature	nature	NOUN	_	_	3	compound	_	_
2	-	-	PUNCT	_	_	1	punct	_	_
3	based	base	VERB	_	_	4	amod	_	_
4	solutions	solution	NOUN	_	_	8	nsubj	_	_
5	for	for	ADP	_	_	7	case	_	_
6	climate	climate	NOUN	_	_	7	compound	_	gemet_id=1462 group=ATMOSPHERE (air, climate)
7	mitigation	mitigation	NOUN	_	_	4	nmod	_	_
8	have	have	VERB	_	_	0	root	_	_
9	the	the	DET	_	_	10	det	_	_
10	potential	potential	NOUN	_	_	8	obj	_	_
11	to	to	PART	_	_	12	mark	_	_
12	reduce	reduce	VERB	_	_	10	acl	_	_
13	annual	annual	ADJ	_	_	16	amod	_	_
14	greenhouse	greenhouse	NOUN	_	_	15	compound	_	gemet_id=3763 group=WASTES, POLLUTANTS, POLLUTION
15	gas	gas	NOUN	_	_	16	compound	_	gemet_id=3763 group=WASTES, POLLUTANTS, POLLUTION
16	emissions	emission	NOUN	_	_	12	obj	_	gemet_id=2663 group=WASTES, POLLUTANTS, POLLUTION
17	by	by	ADP	_	_	21	case	_	_
18	10	10	NUM	_	_	21	nummod	_	_
19	-	-	SYM	_	_	20	case	_	_
20	19	19	NUM	_	_	18	nmod	_	_
21	%	%	SYM	_	_	12	obl	_	_
22	,	,	PUNCT	_	_	25	punct	_	_
23	while	while	SCONJ	_	_	25	mark	_	_
24	also	also	ADV	_	_	25	advmod	_	_
25	benefiting	benefit	VERB	_	_	12	advcl	_	_
26	ecosystems	ecosystem	NOUN	_	_	25	obj	_	gemet_id=2519 group=BIOSPHERE (organisms, ecosystems)
27	and	and	CCONJ	_	_	28	cc	_	_
28	improving	improve	VERB	_	_	25	conj	_	_
29	livelihoods	livelihood	NOUN	_	_	28	obj	_	_
30	.	.	PUNCT	_	_	8	punct	_	_

Table 6: Annotation of propagated GEMET entities made up by more than one word.

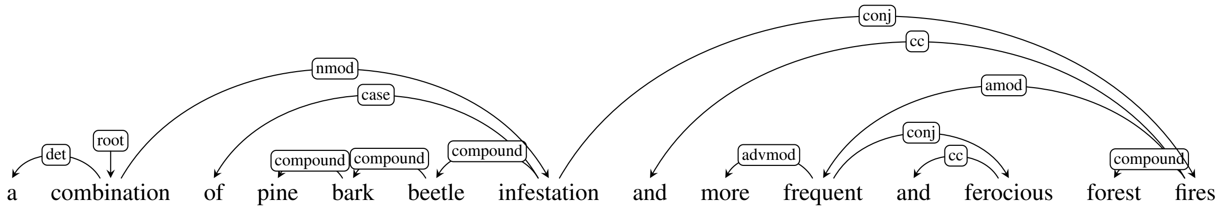

conj
cc
nmod
case
amod
root
det
compound
compound
compound
advmod
conj
cc
compound
a combination of pine bark beetle infestation and more frequent and ferocious forest fires

Figure 2: Dependency relations for the English noun phrase *"a combination of pine bark beetle infestation and more frequent and ferocious forest fires"*.

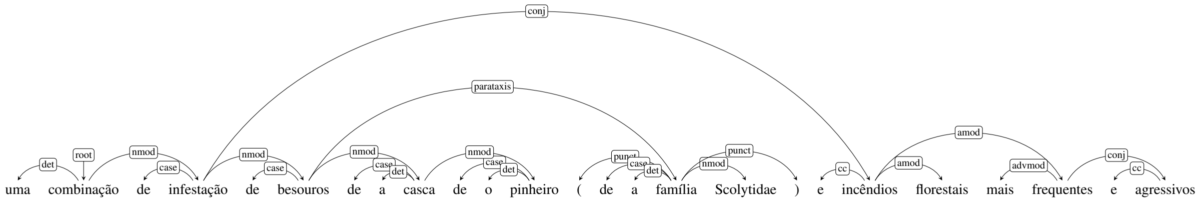

conj
parataxis
amod
det
root
nmod
case
nmod
case
nmod
case
det
nmod
case
det
punct
case
det
punct
nmod
cc
amod
advmod
conj
cc
uma combinação de infestação de besouros de a casca de o pinheiro (de a família Scolytidae) e incêndios florestais mais frequentes e agressivos

Figure 3: Dependency relations for the Brazilian Portuguese noun phrase *"uma combinação de infestação de besouros de a casca de o pinheiro (de a família Scolytidae) e incêndios florestais mais frequentes e agressivos"*.

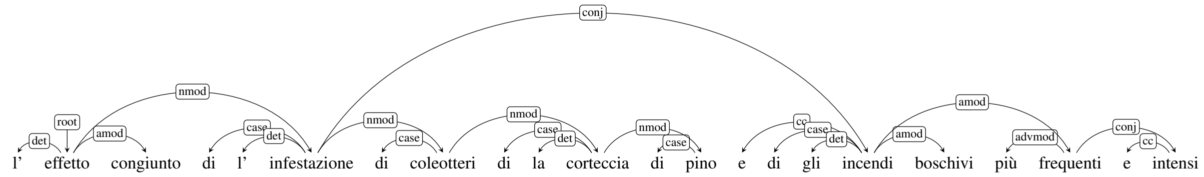

conj
nmod
amod
root
det
amod
case
det
nmod
case
nmod
case
det
nmod
case
cc
case
det
amod
advmod
conj
cc
l' effetto congiunto di l' infestazione di coleotteri di la corteccia di pino e di gli incendi boschivi più frequenti e intensi

Figure 4: Dependency relations for the Italian noun phrase *"l'effetto congiunto di l'infestazione di coleotteri di la corteccia di pino e di gli incendi boschivi più frequenti e intensi"*.

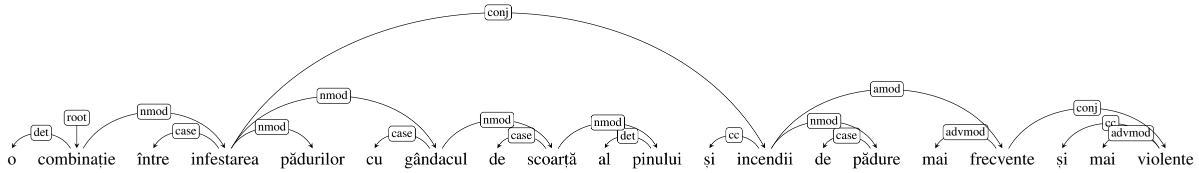

conj
nmod
amod
root
det
nmod
case
nmod
case
nmod
case
nmod
det
cc
nmod
case
advmod
conj
cc
advmod
o combinație între infestarea pădurilor cu gândacul de scoarță al pinului și incendii de pădure mai frecvente și mai violente

Figure 5: Dependency relations for the Romanian noun phrase *"o combinație între infestarea pădurilor cu gândacul de scoarță al pinului și incendii de pădure mai frecvente și mai violente"*.

POS tag	English.conllu	Romanian.conllu	Italian.conllu	Br_Portuguese.conllu
ADJ	270	366	399	325
ADP	280	489	553	564
ADV	109	117	117	134
AUX	127	97	104	80
CCONJ	160	154	152	163
DET	219	175	658	572
NOUN	746	896	795	763
NUM	101	93	83	98
PART	55	78	0	0
PRON	86	94	82	74
PROPN	80	40	62	75
PUNCT	353	323	347	327
SCONJ	43	23	32	30
SYM	48	0	29	45
VERB	349	294	294	317
X	0	6	11	1

Table 7: Distribution of `POS tags` across the four treebanks.

DepRel	English.conllu	Romanian.conllu	Italian.conllu	Br_Portuguese.conllu
acl	27	66	33	42
acl:relcl	23	0	29	27
advcl	57	44	47	59
advcl:relcl	4	0	0	0
advmod	115	140	109	122
amod	271	307	338	262
appos	10	16	14	10
aux	65	43	55	17
aux:pass	22	21	14	20
case	287	421	519	494
cc	159	159	152	168
cc:preconj	1	0	0	0
ccomp	6	30	17	14
ccomp:pmod	0	1	0	0
compound	180	3	21	0
compound:prt	3	0	0	0

Continued on next page

Table 8 – continued from previous page

DepRel	English.conllu	Romanian.conllu	Italian.conllu	Br_Portuguese.conllu
conj	206	217	217	221
cop	38	33	35	42
csubj	7	11	11	9
csubj:pass	0	1	0	0
dep	2	0	0	0
det	213	169	637	558
det:poss	0	0	17	0
det:predet	1	0	4	0
discourse	0	0	0	2
expl	1	4	16	0
expl:impers	0	0	5	0
expl:pass	0	2	1	3
expl:poss	0	1	0	0
expl:pv	0	15	0	7
fixed	2	96	9	45
flat	27	14	0	14
flat:foreign	0	0	5	0
flat:name	0	0	14	43
goeswith	0	0	0	1
iobj	0	15	4	2
mark	86	109	66	69
nmod	149	367	285	278
nmod:poss	22	0	0	0
nsubj	144	134	123	132
nsubj:pass	24	11	12	15
nummod	55	61	60	63
obj	167	123	175	155
obl	122	127	179	155
obl:agent	8	11	10	7
obl:arg	0	0	0	15
obl:pmod	0	12	0	0
obl:unmarked	4	0	0	0
parataxis	14	10	5	13
punct	352	323	347	325
root	116	116	121	117

Continued on next page

Table 8 – continued from previous page

DepRel	English.conllu	Romanian.conllu	Italian.conllu	Br_Portuguese.conllu
xcomp	34	12	12	38

Table 8: Distribution of dependency relations across treebanks.

Building UD Cairo for Old English in the Classroom

Lauren Levine, Junghyun Min, and Amir Zeldes
Georgetown University, Department of Linguistics
`{lel76, jm3743, amir.zeldes}@georgetown.edu`

Abstract

In this paper we present a sample treebank for Old English based on the UD Cairo sentences, collected and annotated as part of a classroom curriculum in Historical Linguistics. To collect the data, a sample of 20 sentences illustrating a range of syntactic constructions in the world's languages, we employ a combination of LLM prompting and searches in authentic Old English data. For annotation we assigned sentences to multiple students with limited prior exposure to UD, whose annotations we compare and adjudicate. Our results suggest that while current LLM outputs in Old English do not reflect authentic syntax, this can be mitigated by post-editing, and that although beginner annotators do not possess enough background to complete the task perfectly, taken together they can produce good results and learn from the experience. We also conduct preliminary parsing experiments using Modern English training data, and find that although performance on Old English is poor, parsing on annotated features (lemma, hyperlemma, gloss) leads to improved performance.

1 Introduction

Treebanking historical languages is a challenging task for multiple reasons – to name just a few: 1. annotators are never native speakers; 2. it is difficult to recruit annotators with sufficient background in the language, in theoretical syntax, and in corpus annotation; 3. guidelines and sample datasets illustrating annotation principles are often tailored to modern languages, with examples that would be implausible in context for ancient texts; 4. there is a scarcity of previous work for annotators to reliably refer to, making examples for contexts that are common in ancient texts difficult to find.

At the same time, tackling ancient language syntax through treebanking can be a rewarding and pedagogically valuable way of confronting students of ancient languages and historical linguistics with authentic syntax in ways that simply looking at an original sentence and its translation cannot (Bamman and Crane, 2007; Mambrini, 2016; Keersmaekers et al., 2019). An additional challenge here is defining a realistic target for treebanking which can be completed as a homework assignment without becoming overwhelming.

In Universal Dependencies (UD, de Marneffe et al. 2021), the most popular treebanking formalism, there is no single template for starting a treebanking project. However, a set of basic sentences is sometimes suggested, known as the Cairo sentences and originally proposed at CICLing 2015 in Cairo.[1] The data has been translated and used as a UD starting point for a number of languages, including Ligurian (Lusito and Maillard, 2021), Luxembourgish (Plum et al., 2024) and Gujarati (Jobanputra et al., 2024). In this paper, we therefore propose and report on a classroom historical treebanking activity creating a UD Cairo-style corpus for Old English (Ald-Ænglisc or OE), including generating the text, treebanking proper, and adding supplementary annotations.

Although some work on treebanking OE and later stages of older English exists – notably The York–Toronto–Helsinki Parsed Corpus of Old English Prose (YCOE; Taylor, 2007), the Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME; Kroch et al., 2004), and their follow-up work (Kroch, 2020; Kulick et al., 2023) – no such data exists in UD. We add to the literature by providing the first OE contribution to UD. We evaluate treebanking from students with and without previous exposure to UD, and run some parsing experiments to reveal how similar or distinct OE syntax is compared to Modern English (ME). We release all data publicly[2] and hope this work in-

---

[1] https://github.com/UniversalDependencies/cairo

[2] https://universaldependencies.org/treebanks/ang_cairo/index.html

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 97–104
August 27, 2025 ©2025 Association for Computational Linguistics

spires others to incorporate more treebanking into classroom activities.

2 LLM-assisted UD Cairo in Ald-Ænglisc

Translating the UD Cairo sentences to Old English is a non-trivial task, especially considering the goal of expressing the basic constructions of the language in the most natural possible way, without introducing translation effects conforming to the Modern English syntax or lexicon. As a first step, we considered simply asking an LLM, in this case GPT4o, to translate the sentences (see Appendix A for the full list of sentences with initial GPT outputs and final versions).

Inspecting GPT's output revealed that while the model was surprisingly capable of translating the Cairo sentences into superficially correct Old English, many problems emerged, especially in morphology and the tendency to mimic the words and word order of the input one for one. For example, consider sentence (1):

(1) a. The girl wrote a letter to her friend
b. Sēo mǣden wrāt ǣnne ærendgecwidd tō hyre frēonde.

The model translated 'the girl' into 'seo mæden' - a correct lexical choice with an incorrect article gender (feminine, when 'mæden' was in fact neuter in OE). The model also tended to insert indefinite articles (here 'ǣnne'), which were not yet grammaticalized or expected in OE, where the Modern English original had them, resulting in superfluous 'ones' (e.g. 'one letter' for 'a letter', where just 'letter' would be correct). We also note that GPT recreates the input word order with the 'to' dative, rather than using the more usual ditransitive construction, ('wrote her friend a letter'), which was used more in OE. Finally, the word for 'letter' is a creative invention, compounded from two attested OE stems, but which is not attested as a word. In the final data, we therefore corrected the gender error, removed the indefinite article, used the ditransitive word order, and used Bosworth Toller's Anglo-Saxon Online Dictionary[3] to select an attested word for 'letter', resulting in (2):

(2) Þæt mæden wrat hyre freonde ærend-writ

At the same time, GPT was adept at proposing creative solutions to vocabulary gaps: for example, Cairo sentence 6 refers to washing a car, to which GPT, unprompted, added: "(Note: "car" is anachronistic, but Old English did have "cār," meaning "chariot" or "cart.")", and selected OE *cræt* 'cart' for its own translation, which we adopted. In some cases, reprompting for additional translation ideas was also an effective strategy.

As a final source to ensure our chosen constructions corresponded to attested Old English syntax, we cross-referenced sentences containing target constructions in a Modern English New Testament translation with the OE Wessex Gospels, the oldest existing manuscript of the Gospels in OE. For example, GPT's initial version of Cairo sentence 17 contains a counterfactual 'she should have been doing', which GPT rendered with *heo scolde þæt ... dyde*, roughly corresponding to a Modern English 'she should that did'. However searching the gospels for 'should have', we quickly realized the most common equivalent in the Wessex Gospels is actually *hit gebyrede þæt heo dyde*, i.e. "it was appropriate that she did...".

In summary, our approach to producing an ancient language version of the Cairo sentences is: 1. ask an LLM to translate the sentence; 2. check the result for morphological errors; 3. ensure chosen words are attested in a dictionary and replace/reprompt for alternatives as needed; 4. check a modern translation for equivalent modern construction instances, then ensure their original counterparts correspond to the selected constructions.

3 Annotation

3.1 UD trees and classroom annotation

Our primary objective was to design a classroom learning activity for an audience that might not be familiar with UD guidelines, rather than to create data for OE parsing. As part of a mid-level, mixed over/undergraduate course on Historical Linguistics, we formulated an assignment in the Old English module of the class in which students were asked to treebank four random OE Cairo sentences. We allowed students to access both UD corpora (using the Grew-match interface, Guillaume 2021) and Modern English parsers – we pointed them to online interfaces for UDPipe (Straka, 2018) and Stanza (Qi et al., 2020), but any parser was allowed.

Since the students mostly had only minimal instruction about the UD guidelines, they were instructed to look in existing UD data for similar constructions to the sentence they were annotat-

---

[3] `https://bosworthtoller.com/`

ing, and to play with the parsers by feeding them what they felt were modern English equivalents (the original Cairo sentences in English were not provided, though students may have found them online too). Use of AI agents was also allowed, though we used the Arborator graphical interface (Guibon et al., 2020), meaning students could not simply let LLMs solve the assignment for them directly.

Figure 1 plots students' accuracy on the assignment in terms of unlabeled and labeled attachment scores (UAS and LAS). We note that some of the students in the class, who had background in Computational Linguistics and previous exposure to UD annotation, unsurprisingly did better on average in terms of accuracy, and therefore we report scores separately for those two annotator subsets.

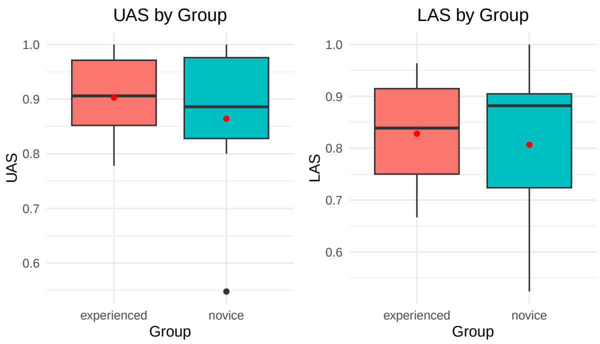


Figure 1: Unlabeled and labeled attachment scores (UAS & LAS) for experienced and novice annotators.

As Figure 1 shows, while the mean scores (indicated with red dots) are somewhat higher for the experienced group, novice students were able to achieve rather comparable scores by using corpus searches and automatic modern English parsers as references, and the median novice LAS score is even higher than the experienced one. At the same time, the range of scores is wider for novices, which we suspect is due to the heterogeneity of their background in general linguistics, with students who had more background in theoretical (but non-UD) syntax achieving better accuracy.

We report inter-annotator agreement (IAA) metrics for our gold reference annotations (accuracy, Cohen's Kappa (Cohen, 1960), and F1-score) between two groups of experienced Modern English UD annotators in Table 1. We also include IAA accuracy on section 23 of the Penn Treebank (Marcus et al., 1993) measured by Berzak et al. (2016) as a point of reference. As our numbers are similar to Penn Treebank ones, we believe experience with ME UD annotation translates non-trivially to performance on OE UD annotation.

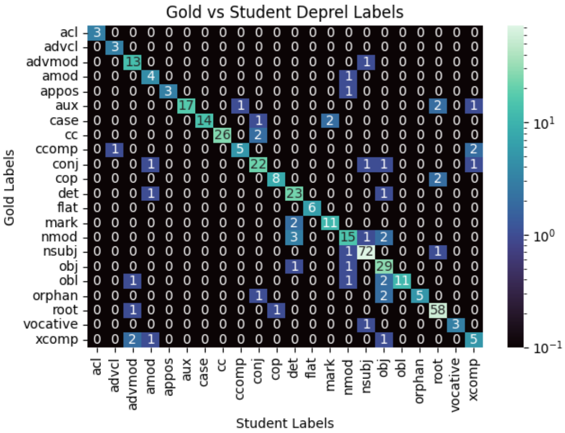


Figure 2: Confusion matrix for gold dependency relation (deprel) labels vs the student annotations. Deprel subtypes are collapsed, only deprel labels that occur > 5 times are included, and `punct` was excluded.

IAA score	**UAS**	**LAS**
UD Cairo OE		
acc.	94.15	88.30
κ	93.38	88.27
F1	96.99	93.79
Reference		
PTB-WSJ (acc.)	93.07	88.32

Table 1: Inter-annotator agreement metrics: accuracy, Cohen's Kappa, and F1-Score. We present an IAA metric on PTB-WSJ by Berzak et al. (2016) for reference.

3.2 Disagreement Analysis

In Figure 2, we show a confusion matrix comparing our final gold labels for the Cairo UD sentences with the student annotations from our classroom activity. In line with our high inter-annotator agreement scores in the previous section, we see that there is strong overlap in the selection of dependency relation labels. However, there is still a smattering of disagreements between the gold labels and the student labels. The dependency relation that was mislabeled by students the most was `nmod`, which was mislabeled 6 times (3 times as `det`, 2 times as `obj`, and 1 time as `nsubj`). These confusions primarily stemmed from students not understanding when to apply the `nmod:poss` subtype when possessive case was involved in the sentence. The label that was most misapplied by students was `obj`, which was misapplied a total of 9 times

(2 times each on `nmod`, `obl`, and `orphan` relations, and 1 time each as on `conj`, `det`, and `xcomp` relations). In Figure 3, we show a dependency analysis for one of the UD Cairo sentences where a student both misapplied the `obj` label rather than utilizing the `orphan` relation, as well as incorrectly using `det` for a possessive pronoun. We see that the student uses a promotion strategy, rather than utilizing `orphan`. Likely the use of more complicated relations, such as `orphan`, could be learned sufficiently with more time devoted to annotator training.

3.3 Additional annotations

In addition to the classroom UD annotations that comprise part of speech tags and dependency labels and their attachments, we provide morphology, hyperlemma, and root annotations (a sample analysis is included in Appendix B).

Morphological features: We follow UD's standard CoNLL-U format and provide annotations on case, number, gender, tense, person, verb-form, and mood. In our corpus, we count four cases (nominative, accusative, genitive, and dative), two numbers (singular and plural), three genders (masculine, neuter, and feminine), two tenses (past and present), three persons (1st, 2nd, and 3rd), three verb-forms (finite, infinitive, and participle), and two moods (indicative and imperative). However, the above list is not exhaustive for Old English morphology. In submitting our corpus to UD, we define Old English's language-specific morphological feature space as outlined in Appendix C.

Gloss: We provide word-level glosses in Modern English, whose uninflected base forms are often equivalent to the hyperlemma described below.

Hyperlemmas: Hyperlemmas, first proposed in Dipper et al. (2004), align word forms and lemmas from different language branches and historical periods that correspond to each other. For example, a hyperlemma *have* would normalize German *haben*, English *have*, Old English *habban* and their inflected forms. We follow several annotated corpora of historic European languages (Meyer, 2011; Kučera and Stluka, 2014; Shokina and Mocken, 2016; Chiarcos et al., 2018) and include hyperlemmas in our annotations.

Roots and language of origin: Roots are another method of normalizing within language family and across historical periods. For example, German *trinken*, English *drink*, and Old English *drincan* can all be associated with their reconstructed Indo-European root *$^*d^h$reńǵ*. While it is more common for Afro-Asiatic corpora to contain root information (e.g. Hajič et al., 2009), we provide root information in addition to hyperlemmas. If the root is of non-Indo-European origin, its language of origin is also specified.

	ELECTRA		**mBERT**	
Input	UAS	LAS	UAS	LAS
GUM 11 dev	95.35	93.17	94.21	91.95
Form	23.98	12.87	41.52	25.73
+Norm	35.09	20.47	41.52	27.49
Lemma	18.13	7.60	38.60	22.81
+Norm	32.16	16.96	39.18	23.39
Hyperlemma	61.99	49.12	70.76	56.73
+Norm	74.27	59.06	83.04	67.84
Gloss	81.87	75.44	85.38	78.36

Table 2: Unlabeled and labeled attachment scores (UAS & LAS) for parsers trained on ME data (GUM V11).

4 Parsing experiments

In order to shed light on how much Old English dependency syntax analysis differs from that of Modern English, we conduct parsing experiments on our OE Cairo sentences using dependency parsers trained on ME. While it may be possible and potentially effective to use dependency parsers trained on other modern and historical Germanic languages, we limit the scope our parsing experiments to English. We train DiaParser (Dozat and Manning, 2017), a neural biaffine dependency parser, using English ELECTRA (Clark et al., 2020) and multilingual mBERT (Devlin et al., 2019) embeddings, in order to compare the impact of monolingual vs multilingual pre-trained inputs. For training data, we use version 11 of the GUM corpus (Zeldes, 2017), which is the largest available UD corpus of Modern English, containing 268k tokens across a variety of genres. We evaluate the ELECTRA and mBERT models on our OE Cairo sentences, using our various token annotations (Form, Lemma, Hyperlemma, and Gloss) as input. The results of these experiments are reported in Table 2, giving the UAS and the LAS for each model. Form refers to the unaltered text of the tokens, and Lemma refers to the uninflected form of the tokens. Hyperlemma and Gloss are as described in Section 3.3. +Norm indicates the addition of orthographic normalization to the annotation of the preceding row – replacing OE letters that no longer exist in

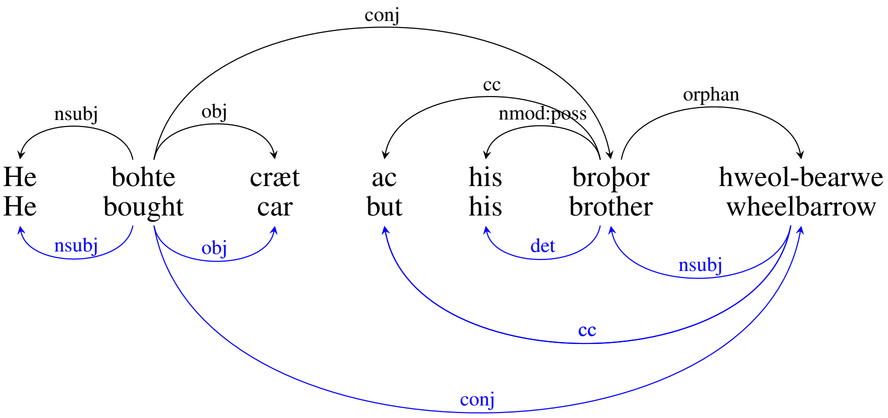


Figure 3: Dependency analysis of a ellipsis construction in OE: *He bought a car but his brother just a wheelbarrow.* ME gloss is included below the OE tokens. The gold analysis is shown on top in black, and a student analysis is shown on the bottom in blue.

ME ("th" for "þ", "e" for "æ", etc.).

Looking at Table 2, we see that although the ELECTRA model has higher performance on the Modern English GUM 11 dev data, mBERT uniformly performs better on the Old English data for all annotation categories. This demonstrates the advantage that multilingual pretrained embeddings provide when a parsing model is used on a different language than the one used during model training. We also note that for the categories of Form and Lemma, orthographic normalization leads to large improvements for the ELECTRA model with monolingual embeddings, but only minor improvement to the mBERT model with multilingual embeddings. However, both models show a large increase in LAS between Hyperlemma and Hyperlemma+Norm, with an increase of 9.94% for the ELECTRA model and an increase of 11.11% for the mBERT model.

For both models, performance increases from Form to Hyperlemma and again to Gloss, for both UAS and LAS, with the mBERT model giving the highest score of UAS 85.38% and LAS 78.36%. Using Hyperlemma and Gloss for the model input allows us to abstract away from the lexical differences between Old English and Modern English. Although parsing the Gloss leads to a large improvement in performance over parsing the Form (LAS Δ 62.65% for the ELECTRA model and LAS Δ 52.63% for the mBERT model), there is still a large drop in performance when compared to the ME dev data (LAS Δ 17.73% for ELECTRA and LAS Δ 13.59% for mBERT). Although some degradation can be expected in cross-corpus parsing (Zeldes and Schneider, 2023), its magnitude indicates that there are notable syntactic differences between OE and ME for dependency parsing, even after controlling for lexical differences. Still, the scores for parsing on glossed tokens of Old English are high enough to suggest that leveraging Modern English models to parse glossed Old English texts could be an effective method to bootstrap the creation of Old English syntactic dependency datasets in the future.

5 Conclusion

In this paper, we presented an Old English version of the UD Cairo sentences, annotated for syntactic dependencies, morphological features, and historically relevant features of hyperlemma and root. In using LLMs for the initial generation of the OE version of the data, we demonstrate how current AI tools can be used in conjunction with quality assurance procedures in the workflow of data creation to produce more reliable results. By including the treebanking task as part of a course in Historical Linguistics, we show that a brief introduction to UD in the classroom, along with providing reference resources, can result in relatively good performance, even for a language with which most of the students have no familiarity. Additionally, through preliminary parsing experiments, we demonstrate that dependency parsers trained on Modern English data may be of use for the creation of new syntactic dependency datasets for Old English. We hope that the sample Old English treebank presented here can serve as a guide for future classroom annotation efforts.

Acknowledgments

We would like to thank the students in the course Corpus Approaches to Historical Linguistics at Georgetown University for participating in the course and the creation of the dataset. The final public repository will include all annotation versions of all students who agreed to release their work and signed an open license agreement. Credit is given by name to all students who wished to be named, as well as anonymously to students who wish to remain anonymous.

References

David Bamman and Gregory Crane. 2007. The Latin dependency treebank in a cultural heritage digital library. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007).*, pages 33–40, Prague, Czech Republic. Association for Computational Linguistics.

Yevgeni Berzak, Yan Huang, Andrei Barbu, Anna Korhonen, and Boris Katz. 2016. Anchoring and agreement in syntactic annotations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2215–2224, Austin, Texas. Association for Computational Linguistics.

Christian Chiarcos, Benjamin Kosmehl, Christian Fäth, and Maria Sukhareva. 2018. Analyzing middle high german syntax with rdf and sparql. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kevin Clark, Thang Luong, Quoc V. Le, and Christopher Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

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

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.

Stefanie Dipper, Lukas Faulstich, Ulf Leser, and Anke Lüdeling. 2004. Challenges in modelling a richly annotated diachronic corpus of german. In *Workshop on XML-based richly annotated corpora*, pages 21–29.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5291–5300, Marseille, France. European Language Resources Association.

Bruno Guillaume. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175, Online. Association for Computational Linguistics.

Jan Hajič, Otakar Smrž, Petr Zemánek, Petr Pajas, Jan Šnaidauf, Emanuel Beška, Jakub Kracmar, and Kamila Hassanová. 2009. Prague arabic dependency treebank 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Mayank Jobanputra, Maitrey Mehta, and Çağrı Çöltekin. 2024. A Universal Dependencies treebank for Gujarati. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 56–62, Torino, Italia. ELRA and ICCL.

Alek Keersmaekers, Wouter Mercelis, Colin Swaelens, and Toon Van Hal. 2019. Creating, enriching and valorizing treebanks of Ancient Greek. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 109–117, Paris, France. Association for Computational Linguistics.

Anthony Kroch. 2020. Penn parsed corpora of historical english.

Anthony Kroch, Beatrice Santorini, and Lauren Delfs. 2004. Penn-helsinki parsed corpus of early modern english.

Seth Kulick, Neville Ryant, and Beatrice Santorini. 2023. Parsing “early English books online” for linguistic search. In *Proceedings of the Society for Computation in Linguistics 2023*, pages 222–242, Amherst, MA. Association for Computational Linguistics.

Karel Kučera and Martin Stluka. 2014. Data processing and lemmatization in digitized 19th-century czech texts. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH ’14, page 193–196, New York, NY, USA. Association for Computing Machinery.

Stefano Lusito and Jean Maillard. 2021. A Universal Dependencies corpus for Ligurian. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 121–128, Sofia, Bulgaria. Association for Computational Linguistics.

Francesco Mambrini. 2016. The Ancient Greek Dependency Treebank: Linguistic annotation in a teaching environment. In M. Romanello and G. Bodard, editors, *Digital Classics Outside the Echo-Chamber*. Ubiquity Press, London.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Roland Meyer. 2011. New wine in old wineskins?—tagging old russian via annotation projection from modern translations. *Russian linguistics*, 35(2):267–281.

Alistair Plum, Caroline Döhmer, Emilia Milano, Anne-Marie Lutgen, and Christoph Purschke. 2024. LuxBank: The first Universal Dependency treebank for Luxembourgish. In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*, pages 30–39, Hamburg,Germany. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Nina Yu Shokina and Susanne Mocken. 2016. A text mining system for creating electronic glossaries with application to research of church slavonic language. *Vychislitel'nyye tekhnologii*, 21(4).

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.

Ann Taylor. 2007. *The York—Toronto—Helsinki parsed corpus of old english prose*, pages 196–227. Palgrave Macmillan UK, London.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes and Nathan Schneider. 2023. Are UD treebanks getting more consistent? a report card for English UD. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 58–64, Washington, D.C. Association for Computational Linguistics.

A GPT initial outputs

GPT's initial renditions of the Old English Cairo sentences are given below, in italics. Note that these include clear errors. For reference, original English sentences from the Cairo dataset in quotes and the final Old English translations for our UD Old English Cairo data are also supplied for easy comparison.

1. *Sēo mǣden wrāt ǣnne ærendgecwidd tō hyre frēonde.*

   "The girl wrote a letter to her friend."

   Þæt mæden wrat hyre freonde ærend-writ.

2. *Ic þence þæt hit rīne.*

   "I think that it is raining."

   Ic þence þæt hit regnaþ.

3. *Hē ġewende tō geswīgenne smōkunge and drincan.*

   "He tried to stop smoking and drinking."

   He gesohte to geswicanne smican and drincan.

4. *Wilt þū gān?*

   "Do you want to go?"

   Wilt þu gan?

5. *Sam, ontyne þæt ēagþȳrel!*

   "Sam, open the window!"

   Sam, ontyne þæt eagþyrel!

6. *Hēo dyde þæt hyre efenġemacca þæt cræt gewēasce*

   "She made her husband wash the car."

   Heo dyde þæt hyre efen-gemacca þæt cræt geweasce.

7. *Pēteres nēahġebūr smearde þone stoccweall rēad.*

   "Peter's neighbor painted the fence red."

   Peteres neah-gebur hæfð þone eodor readne ongemet.

8. *Mīn fæder is cealdra þonne þīn.*

   "My dad is cooler than yours."

   Min fæder is micelre þonne þin.

9. *Maria gewann brons, Pēter seolfor, and Ġeane gold.*

   "Mary won bronze, Peter silver, and Jane gold."

   Maria gewann ær, Petrus seolfor, and Iohanna gold.

10. *Is Iguazu micel ríce oþþe lȳtel?*

    "Is Iguazu a big or a small country?"

    Is Iguazu micel rice oþþe lytel?

11. *Nǣþer Pēter Smīþ nē Maria Brūn mihton bēon gecōrene.*

    "Neither Peter Smith nor Mary Brown could be selected."

    Ne Petrus Smiþ ne Maria Brun mihton beon gecorene.

12. *Hīe nabbaþ nāne ġeþōht hwā hit wrāt.*

    "They have no idea who wrote it."

    Hie nabbaþ nane geþoht hwa hit wrat.

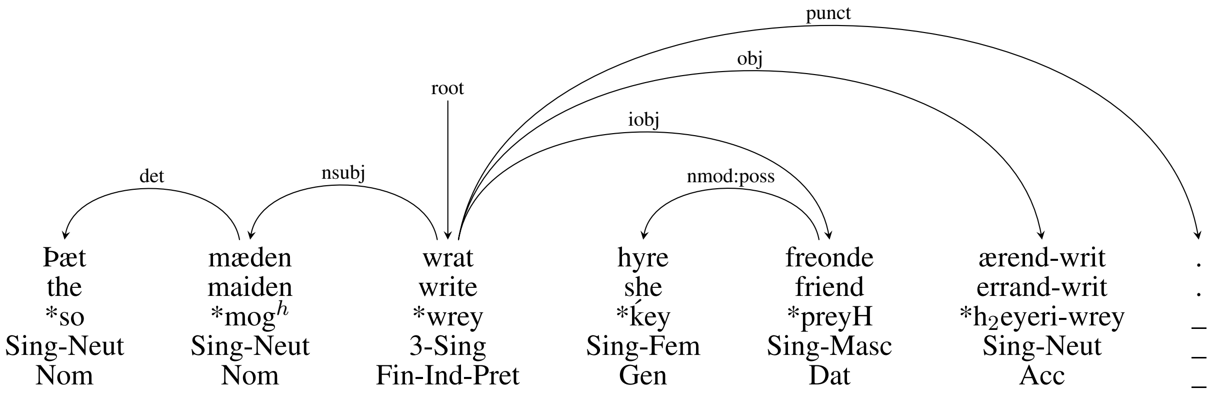

punct
obj
root
iobj
det
nsubj
nmod:poss
Þæt mæden wrat hyre freonde ærend-writ .
the maiden write she friend errand-writ .
*so *mog^h *wrey *ḱey *preyH *h$_2$eyeri-wrey _
Sing-Neut Sing-Neut 3-Sing Sing-Fem Sing-Masc Sing-Neut _
Nom Nom Fin-Ind-Pret Gen Dat Acc _

Figure 4: Dependency analysis of a ditransitive construction in OE: *The girl wrote a letter to her friend.* The glossing contains (from top to bottom) OE form, hyperlemma, root, and morphological feature annotations (person-number-gender and verbForm-mood-tense, as applicable).

13. *Hwæt locast þū on?*

    “What are you looking at?”

    Hwæt locast þu on?

14. *Hwonne þenctst þū þæt þū mōst cuman?*

    “When do you think you can come?”

    Hwonne þenctst þu þæt þu meaht cuman?

15. *Hē bohte cræt ac his brōþor hweól-bearwe.*

    “He bought a car but his brother just a wheelbarrow.”

    He bohte cræt ac his broþor hweol-bearwe.

16. *Petrus and Maria beclypton heom be-tweonen and þā forlētone þā heoþe.*

    “Peter and Mary hugged each other and then left the room.”

    Petrus and Maria beclypton heom be-tweonen and þa forletone þa heoþe.

17. *Hēo scolde þæt hire hēafodġearwung dyde ac for sumre rǣswan hēo nolde on þām dæġe.*

    “She should have been doing her hair but for some reason she wouldn’t that day.”

    Hit gebyrede þt heo dyde hire her, ac for sume wisan nolde seo don swa.

18. *Ic ne mihte fylġan, forþām hē tō hrædlīce arn.*

    “I wasn’t able to keep up, because he ran too fast.”

    Ic ne mihte þurhwunian, forðon he to swiftlice rann.

19. *Þis gewrit is fram Pētere and hit wæs ġebrōht ġierstan dæġe.*

    “This letter is from Peter and it was delivered yesterday.”

    Þis ærend-writ is fram Petere and hit wæs gebroht gierstan-dæg.

20. *Hēo weox on Paris, þǣre heofodburh Franclandes.*

    “She grew up in Paris, the capital of France.”

    Heo weox on Paris, þære heafod-burh Franc-landes.

B Sample OE sentence analysis

We provide a sample Old English sentence analysis, complete with its syntactic features, hyperlemma, Indo-European root, and morphological features in Figure 4.

C Old English specific morphological feature space

Our feature space follows that of Modern English, with the exception of case and gender, with 5 cases (nominative, accusative, genitive, dative, instrumental – the latter not attested in UD Cairo) and 3 genders (masculine, neuter, and feminine). Old English also exhibited dual number in pronouns, though this is not attested in our data. Features in tense, person, verb-form, and mood are equivalent with constraints set in Modern English UD corpora.

1. Case. `Nom, Acc, Gen, Dat, Ins`.
2. Gender. `Masc, Fem, Neut`.
3. Number. `Sing, Dual, Plur`.
4. Person. `1, 2, 3`.
5. VerbForm. `Fin, Inf, Part, Ger`.
6. Mood. `Ind, Imp, Sub`.
7. Tense. `Pres, Pret`
8. Degree. `Pos, Comp, Sup`
9. Poss. `Yes`

Universal Dependencies for Sindhi

John Bauer
HAI
Stanford University
horatio@cs.stanford.edu

Sakiina Shah
Muhammad Shaheer
Mir Afza Ahmed Talpur
Zubair Sanjrani
Isra University
sakiinashah77@gmail.com
shaheer.memon,afzal.talpur,zubair.sanjrani@isra.edu.pk

Sarwat Qureshi
U. Manchester
sarwatqureshi4@gmail.com

Shafi Pirzada
MLtwist
shafi.pirzada@gmail.com

Christopher D. Manning
Linguistics & Computer Science
Stanford University
manning@stanford.edu

Mutee U Rahman
Isra University
muteeurahman@gmail.com

Abstract

Sindhi is an Indo-Aryan language spoken primarily in Pakistan and India by about 40 million people. Despite this extensive use, it is a low-resource language for NLP tasks, with few datasets or pretrained embeddings available. In this work, we explore linguistic challenges for annotating Sindhi in the UD paradigm, such as language-specific analysis of adpositions and verb forms. We use this analysis to present a newly annotated dependency treebank for Universal Dependencies, along with pretrained embeddings and an annotation pipeline specifically for Sindhi annotation.

1 Introduction

Developing a Universal Dependencies (UD) Treebank for Sindhi presents unique challenges due to the language's complicated linguistic features. Despite being spoken by approximately 40 million people in Pakistan and India, Sindhi is a low-resource language compared to other languages of similar population, with only a few tools and datasets available (see section 2).

Sindhi is a split-ergative Indo-Aryan language, written right to left.[1] It is closely related to Punjabi and Saraiki. Both languages have UD datasets as works in progress (Arora, 2022; Alam et al., 2024), but they are not yet publicly released, and many related languages are also low resource, meaning difficult issues in Sindhi annotation have few references from which to work.

Annotating Sindhi for the increasingly widespread Universal Dependencies framework (de Marneffe et al., 2021) is complicated by its complex case system, where case markers themselves undergo further inflection, alongside rich inflectional morphology with nominal and verbal elements. Additional complexities arise from pronominal suffixation with nouns, postpositions, adverbs, and verbs, and a partially free word order that introduces syntactic ambiguity. Furthermore, the language employs intensifiers across various word classes. The multifunctional use of pronouns as determiners also poses challenges for establishing consistent dependency relations. These characteristics require a tailored approach to create an accurate and robust Sindhi UD treebank, with some of the more difficult topics explained in section 4.

To annotate plain text, current annotation pipelines almost universally use neural methods with static or contextual embeddings. As a low-resource language, Sindhi is most commonly represented in such embeddings as part of a multilingual collection, such as the FastText project (Grave et al., 2018), or as part of a collection of related languages (Khanuja et al., 2021). However, embeddings specifically trained for the language in use typically produce better results, a result this work demonstrates applies to Sindhi. In order to train new embeddings, we collect data from a variety of text sources, described in section 5.

We use the annotated dependency treebank and newly trained word embeddings to train an annotation pipeline using the Stanza package (Qi et al., 2020). To facilitate annotation, we build pipelines from partially annotated data and use these labels as silver annotations for the annotators to use as a baseline.

The main contributions of this work are:

---

[1] As Sindhi text is written RtL, annotated trees in this text are written RtL, including the English glosses

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 105–118
August 27, 2025 ©2025 Association for Computational Linguistics

- build a UD dataset for Sindhi, to be available at universaldependencies.org
- use a multilingual transformer to build progressively better crosslingual models, allowing for quicker annotation
- build Sindhi word vectors and transformers, to be available at huggingface.com
- combine these into a Sindhi annotation pipeline, to be integrated into Stanza at stanza.github.io.

2 Related Work

There are existing datasets for building components of Sindhi annotation pipelines, although none of them are complete solutions.

A partially complete analysis of 665 sentences for Sindhi, named Mazhar Dootio (MD) after its primary author, is available at universaldependencies.org (Dootio and Wagan, 2019). This dataset is small and lacks dependency trees. Nevertheless, it provides a useful baseline, especially for features and lemmatization. In this work, we reanalyze these sentences, adding dependencies and updating the tagging and featurization standards.

Ali et al. (2020b) presents an NER dataset for Sindhi, with 26,000 tagged sentences, available on github.[2] We use it here to build an NER model for the annotation pipeline. The same group also produced a POS dataset of 293K words (Ali et al., 2021).

As static and contextual word embeddings are essential for training neural models, there have been multiple works which provide embeddings. For static embeddings, FastText included Sindhi vectors trained on Common Crawl and Wikipedia (Grave et al., 2018). Later, UESTC crawled a larger corpus of 61M words to produce Sindhi-specific word vectors (Ali et al., 2020a).

For contextual embeddings, XLM-R (Conneau et al., 2020) uses 40M tokens of Sindhi text as part of its training collection. Muril-Large (Khanuja et al., 2021) uses the same amount of Sindhi text, but perhaps because it is focused on Indic and Dravidian languages, tends to perform better on downstream tasks.

Prior work aside from the Mazhar-Dootio dataset also addresses lemmatization, including Nathani et al. (2020) and Dootio and Wagan (2017). We use a different lemmatization standard in this work, though; see section 3.2.

This work uses Stanza (Qi et al., 2020) to build and evaluate models. Similar projects which could build a Sindhi annotation pipeline using the data include UDPipe (Straka, 2018) and spaCy (Honnibal et al., 2020).

3 Annotation Process

The UD dataset described here was annotated in partnership with MLtwist, an NLP annotation company. MLtwist recruited two native speakers of Sindhi for annotation, listed here as authors.

For annotation platforms, we used Datasaur for dependencies and UPOS, followed by Kili for features and XPOS. Individual sentences that needed revision were edited via the free tool conllueditor (Heinecke, 2019), with batch edits executed with Semgrex and Ssurgeon (Bauer et al., 2023).

3.1 Incremental Models

Annotating treebanks from the starting point of a silver dataset is a common practice. The English EWT dependency treebank (Silveira et al., 2014) started with LDC constituency trees using a deterministic conversion process (Manning et al., 2014). The English GUM dependency treebank (Zeldes, 2017) uses a parser trained on existing English materials to provide initial silver trees to the annotators. As shown in Mikulová et al. (2022), using silver trees improves the final annotation quality.

Meanwhile, using multilingual representations to improve parsers is a common practice, such as in recent Javanese parsing work (Ghiffari et al., 2024), which used two layers of multilingual representations to build the parser.

This work combines those two concepts to build initial silver trees, even in the setting of very few trees already annotated.

The Muril transformer (Khanuja et al., 2021) contains all of the scheduled languages in India, of which Sindhi is one. Several languages in Muril are also represented in Universal Dependencies, shown in table 1:

Language	Family	Total Trees
Hindi	Indic	16,649
Urdu	Indic	5,130
Marathi	Indic	466
Tamil	Dravidian	600

Table 1: Languages in Muril with sufficient UD data and the size of the largest treebank

---

[2] `https://github.com/AliWazir/SiNER-dataset`

To build silver trees, we train a Stanza POS tagger and dependency parser with a mix of trees from these languages, even without having labeled Sindhi dependency trees available.[3]

To demonstrate the effectiveness of this approach, we train models based on subsets of the available data and score them on the Sindhi test set. Each model trained here includes 1000 Urdu trees, 1000 Hindi trees (Bhat et al., 2017; Palmer et al., 2009), 373 Marathi trees (Ravishankar, 2017), and 400 Tamil trees (Ramasamy and Žabokrtský, 2012).[4] Tamil is an interesting case, as it is Dravidian rather than Indic, but we find there is still some benefit to transfer learning for dependencies. Table 2 shows that even low amounts of gold Sindhi data result in high-quality silver trees, when combined with multilingual training.

Train	Dev	None	Indic	Full
		POS AllTags F1		
0	0	–	82.66	81.94
0	100	–	81.72	82.22
50	100	85.51	87.15	86.88
100	100	89.81	89.39	89.54
200	500	90.65	91.06	90.91
500	500	92.85	92.64	90.90
		Depparse LAS F1		
0	0	–	69.40	72.91
0	100	–	72.02	72.52
50	100	61.54	75.13	76.01
100	100	70.38	78.71	78.74
200	500	76.73	80.87	80.88
500	500	81.33	82.68	82.53

Table 2: Accuracy of crosslingual models. Annotation gets easier when adding more data. "Indic" means the model is trained with additional Indic UD data, and "Full" includes Tamil as well.

3.2 Lemmatization

For the most part, we follow the Dootio and Wagan (2019) standard for lemmatization, with one key difference. In the former, the lemma is chosen to be the root form of the word regardless of POS tag, whereas here the lemma uses the default root form with the same POS tag. We remove inflections, but not derivational endings. For example, in the English dataset EWT (Silveira et al., 2014), *ADJ* words ending with *-ish* are lemmatized to the *-ish* form, whereas the Mazhar Dootio standard would stem the word without the *-ish*.

Form	w/o POS	w/ POS
hawkish	hawk	hawkish
childish	child	childish
sympathize	sympathy	sympathize

The results in Sindhi follow a similar pattern:

Sindhi ADJ	M-D	Ours
اايماندار	ايمان	ايماندار
ٿڌي	ٿڌ	ٿڌو
دولتمند	دولت	دولتمند

Repeated iteration of training a Stanza seq2seq lemmatizer helped speed the lemmatization process, as once 1000 lemmas were manually lemmatized, the lemmatizer model demonstrated an 86% accuracy on unseen lemmas. In contrast, the "identity" lemmatizer, using the word itself, would only be 66% accurate on unseen words.

3.3 Corpus Statistics

After retokenizing and deduplicating the MD dataset, 509 annotated sentences remain. We use this and a collection of 378 sentences of news articles and folk tales to build the test set. Another 386 sentences will comprise the dev set. As of this writing, 4116 total sentences have XPOS and features annotated. Another 1532 have UPOS and dependencies and will be fully annotated by the time of publication.

3.4 Agreement

To measure interannotator agreement, we gave the same batch of 200 silver sentences to both of the UPOS and dependencies annotators. The resulting annotations agreed on 96% of the UPOS (95.4 kappa) and 91.4% of the dependencies.

These numbers have a caveat, though, in that agreement will be much higher when annotating silver trees, as the annotators will be predisposed to choose the silver label in the case of an ambiguous annotation. This phenomenon has been previously studied in work such as Berzak et al. (2016).

4 Linguistically adapting UD for Sindhi

Several constructions in Sindhi are challenging to annotate as part of a Universal Dependencies framework, or are seen only in the context of other low-resource languages.

4.1 Demonstratives as Determiners

Similar to many Indo-Aryan languages, and others such as the Slavic language Czech, Sindhi lacks standalone articles and does not have a well-defined category of determiners. Instead, demonstrative pronouns often function as determiners.

---

[3] With 0 Sindhi trees, the Urdu dev set is used.

[4] The entire training set for Marathi and Tamil, and a random subset of Hindi and Urdu

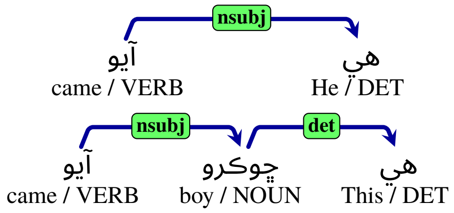
Therefore, Sindhi demonstrative pronouns map to the *DET* part of speech.

This multifunctional use of demonstratives can lead to ambiguity and requires careful annotation in Universal Dependencies relations. For example, هي hī can be a demonstrative (proximal *he*) or a determiner (*this*). The use of هي as both a demonstrative and a determiner can be seen in the following sentences:

1. هي آيو. hī āyo (He came)
2. هي ڇوڪرو آيو. hī chokro āyo (This boy came)

In the first sentence, هي is a demonstrative pronoun used standalone and functions as the nsubj of the sentence. In the second sentence, هي is used as a determiner modifying ڇوڪرو, where ڇوڪرو is the nsubj. When used standalone in a sentence, demonstratives function as pronouns and can appear in the roles of nsubj, obj, iobj, or obl. However, when used before nominals, they function as determiners.

As argued in (Bharati et al., 2008), it can be useful to have a separate part of speech for demonstratives compared to other pronouns. Furthermore, the treebanks for other Indic languages such as Hindi and Urdu establish the convention of always tagging demonstratives *DET*. It is useful to have this distinction even when they function as pronouns, even though this contrasts with English UD, which contextually distinguishes "this" as *PRON* or *DET*, for example. In such cases, the dependency relation demonstrates the use of the word.



4.2 Intensifiers

One common class of words in Sindhi is *intensifiers*. These are words with no direct English translation which put the emphasis of the sentence on the previous word. In English writing, bold or italic text would have a similar effect.

An example sentence from this dataset is:

ان	کي	ئي	ٻاراڻو	ادب	ڪوٺڻ	گهرجي
un	khe	ī	bārāṇo	adab	kōṭhaṇ	ghurje
that	case	EMP	childish	lit.	calling	need

It should be called children's literature

In this sentence, ئي is an intensifier. Figure 1 has an analysis of the phrase ان کي ائي.

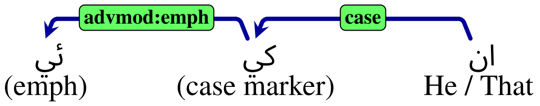


Figure 1: The intensifier ئي turns "he" into "Only he"

In Sindhi, intensifiers are typically labeled with the part-of-speech tag *PART* having *advmod:emph* relation, attached to the modified word.

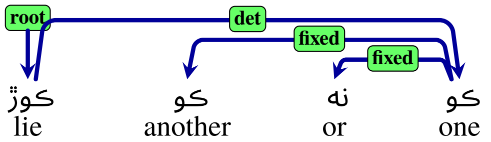
Intensifier words may not always act as intensifiers and can serve other roles based on the context. Notably, the particle نه (na) is usually a negation particle, annotated as *PART* with the *advmod* relation, but it functions as an intensifier when following imperative verbs, as in اچ نه (ach na, "come on"), where it emphasizes the command. However, when نه precedes the verb, as in نه اچ (na ach, "don't come"), it acts as a negation particle. Similarly, ته (ta) is generally a subordinating conjunction, annotated as *SCONJ* with the *mark* relation, but it becomes an intensifier when following specific verb forms, as in اچ ته (ach ta, "come on"), or nominals, as in هو ته (hu ta, "he too" or "he indeed"). These multifunctional particles highlight the need for context-sensitive annotation in Sindhi UD.

4.3 Fixed

In Universal Dependencies (UD), the *fixed* relation connects words within a multiword expression (MWE) where the fixed combination functions as a single grammatical unit with non-compositional meaning. For example, ڪو نه (ko na, "not any") is a two-word fixed expression annotated with the *advmod* relation, representing negation. Individually, ڪو (ko) can serve as a determiner (*det*) or an unspecified dependency (*dep*), while نه (na) acts as a negation particle or intensifier (typically *PART*, see section 4.2). Together, however, these two tokens form a unit with an *advmod* role. Because of their rigidity and role as single grammatical units, we label these *fixed*.

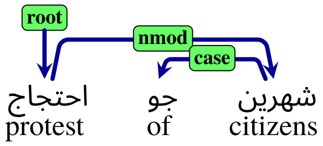
Some examples include:

ڪو نه	not any	advmod
ڪو نه ڪو	one or another	det
کان سواءِ	except	case
نه ئي	nor	advmod



4.4 Postpositions

Most adpositions in Sindhi, tagged *ADP* as per UD standards, are postpositions. An entire adpositional phrase typically precedes the modified head. A simple example in this dataset is “The protest of the citizens”:



Postpositions in Sindhi frequently function as case markers, appearing after nominals and requiring those nominals to take the oblique case form. This oblique case is a fundamental prerequisite for further case marking in the language. Consider the following examples:

هي گاڏيون (hī gāḍiyūn) “These vehicles”

هنن گاڏين تي (hinan gāḍīan te) “On these vehicles”

In the first example, both elements of the adpositional phrase “ هي” (hī, “these”) and “ گاڏيون” (gāḍiyūn, “vehicles”) are in the default nominative (or direct) case, as no postposition is present. In the second example, the postposition “ تي” (te, “on”) triggers the oblique case, causing both the determiner “ هي” (hī) to inflect to “ هنن” (hinan) and the noun “ گاڏيون” (gāḍiyūn) to inflect to “ گاڏين” (gāḍīan). This illustrates how postpositions in Sindhi govern the case of the preceding nominals, shifting them from nominative to oblique forms (accusative case in terms of Universal Dependencies).

Sindhi genitive adpositions (ADPs) differ from other ADPs, such as locative ones, due to their ability to inflect for number, gender, and case. Unlike locative ADPs like “ ۾” (mē, “in”), “ تي” (te, “on”), and “ وٽ” (vat, “near/beside”), which only mark the accusative (oblique) case of preceding nominals and lack further inflectional features, genitive ADPs not only assign case to the nouns they follow but also bear their own inflectional features.

Examples of the distinctions between these features can be found in appendix B.

We use XPOS to separate out two classes of adpositions, genitive and location. Genitive ADP are inflected for case and are featurized for nominative and accusative cases. ADP that represent location are treated as having no further features, as are ADP which are not one of these two subclasses. We do not use *Case=Loc* for location ADP as there is no inflection which indicates the ADP is locative.

4.5 Prepositions

There are some rare cases of prepositions, such as the word سواءِ, “except”, in the sentence analyzed in figure 2:

جنهن سان سواءِ جاهلن جي هر هڪ جو تعلق آهي

4.6 Participles and VerbForm Features

Participles in Sindhi are versatile verb forms that function as adjectives, nouns, and adverbials. They also play a role in tense and aspect formation. Participles in Sindhi include verbal nouns and the present, past, future, and conjunctive participles. Participles are further marked by number, gender, and person inflections, encoded in the features. These features, along with the *VerbForm* feature, distinguish the different participles.

Examples features for the different verb forms described here are in table 6.

Present Participle The present participle in Sindhi is marked by the verbal suffix “ ند” (and), further inflected for number and gender. The following sentence is an example of a present participle paired with an auxiliary verb forming imperfective habitual aspect in present tense where the main verb لک (likh, “write”) is inflected by masculine singular suffix “ ندو” (ando), forming the present participle لکندو (likhando, habitual “writes”).

هو لکندو آهي (hu likhando āhe, “He writes”)

Past Participle The past participle in Sindhi is marked by the suffix “ يل” (yal), which rarely inflects for number (e.g., “ لکيلن” /likhyalan/ for plural in some dialects). In the following sentence, the verb لک (likh, “write”) with the suffix “ يل” (yal) forms the past participle form لکيل (likhyal, “written”), with perfective aspect in passive voice.

هي ڪتاب لکيل آهي hī kitāb likhil āhe “This book is written”

Future Participle The future participle in Sindhi is often an inflected form of the infinitive. For example, the infinitive “ لکڻ” (likhaṇ, “to write”), is marked by number and gender to indicate obligation or futurity. In the following sentence, the verb form لکڻو (likhaṇo, “to write” or “having to write”) is a masculine singular inflection of the above infinitive with an imperfective aspect.

مون کي ڪتاب لکڻو آهي (mūn khē kitāb likhaṇo āhe, “I have to write a book”)

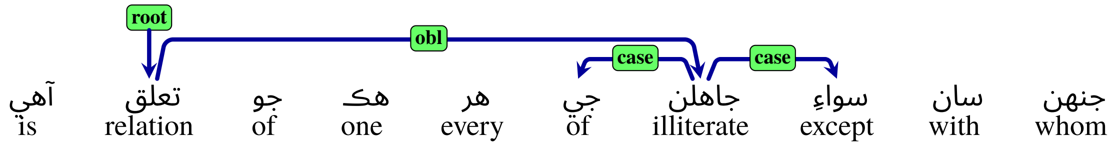


Figure 2: Very rarely, Sindhi has prepositions as well as postpositions
"To which everyone except the illiterate belongs"

Unlike the present and past participles where *VerbForm=Part* is used, we label this *VerbForm=Inf* to avoid ambiguity.

Conjunctive Participle The conjunctive participle in Sindhi typically denotes an action completed before the main verb or clause's action, linking two events without an explicit conjunction. It is formed by inflecting the verb root with the suffix " ي" , as in لکي (likhi) from لک (likh, "write"). However, this form's role, either as a simple verb or a conjunctive participle, depends not only on its morphology but also on its syntactic standing in the sentence. In the following examples, syntactic context such as adverbial modification of the main verb distinguishes two potential roles of لکي (likhi):

A simple past verb: لکي چٺي مون (mūn chithī likhi, "I wrote a letter"). Here, لکي is the root.

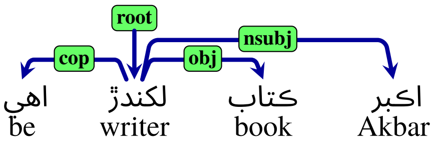
A conjunctive participle: آيس گهر لکي چٺي آئون (āūn chithī likhi ghar āyas, "I came home after writing a letter"). Here, it modifies the root آيس (āyas, "came"), indicating a prior action.

Conjunctive participles in Sindhi do not inflect for number, gender, or person; they are invariable, unlike the present or past participles discussed above. Their aspect is always perfective, reflecting a completed action relative to the main clause.

In Universal Dependencies, the conjunctive participle is mapped to the converb category, tagged with *VerbForm=Conv*. Across UD, a converb is a verb form that adverbially modifies the main verb, often indicating sequence, cause, or manner.

Verbal Noun Verbal nouns in Sindhi are derived from verbs and function as nouns while retaining verbal properties like implying an action. Unlike infinitive forms like لکڻ (likhaṇ, "to write"), which can also act as nouns, Sindhi has a distinct class of verbal nouns marked by the agentive suffix ندڙ (ndar), indicating the doer (e.g., لک (likh, "write") becomes لکندڙ (likhandar, "writer")).

In the sentence اکبر ڪتاب لکندڙ آهي (Akbar kitāb likhandar āhe, "Akbar is the writer of the book"), لکندڙ (likhandar) is a verbal noun, acting as a predicate noun with imperfective aspect (ongoing capacity to write) and singular number. It inflects for number and case, as in لکندڙن کي ٻڌايو ڪتاب (kitāb likhandaran khē buḍhāyo, "Tell the writers of the book"), where لکندڙن (likhandaran) is plural, marked with accusative case via کي (khē, "to").



In the dependency parse of ڪتاب لکندڙ آهي اکبر, there is no *ADP* indicating "of the book", but rather "book" is the *obj* of "writer". Furthermore, unlike English deverbals such as "chaser" in "ambulance chaser", this potential *obj* can have many substitutions, including complete phrases. As "writer" is taking on VERB dependencies, we follow the analysis of Cecchini (2021) and an extensive UD discussion[5] and treat these as *VERB*, with the feature *VerbForm=Vnoun*.

Infinitive Infinitive verbs in Sindhi are formed by the ڻ (-ṇ) suffix inflection. For example, لک (likh) 'write' becomes لکڻ (likhaṇ) "to write". Infinitives always exhibit the imperfective aspect and are marked by the feature *VerbForm=Inf*.

The lack of further inflection distinguishes infinitives from future participles. Although both share the *VerbForm=Inf* and *Aspect=Inf* features, they are distinguished by the remaining features. Unlike infinitives, future participles inflect for number, gender, and person.

Infinitives can appear in various syntactic roles within the Universal Dependencies (UD) structure, including *nsubj* (nominal subject), *obj* (object), *xcomp* (open clausal complement), *ccomp* (clausal complement), and *advcl* (adverbial clause).

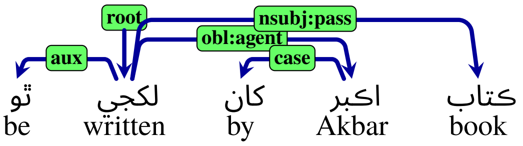
4.7 Passive Voice

Sindhi forms passive voice constructions across various tenses and aspects, primarily utilizing mor-

---

[5] `https://github.com/UniversalDependencies/docs/issues/1125`

phological passive forms and auxiliary verbs. In the present tense with imperfective aspect, passive forms are marked by suffixes such as “ جي” (-je) or “ بو” (-bo), inflected for number and gender, and paired with present tense auxiliaries like “ ٿو” (tho) or “ آهي” (āhe). For example, the active sentence “ اکبر کتاب لکي ٿو” (Akbar kitāb likhe tho, “Akbar writes a book”) becomes “ کتاب اکبر کان لکجي ٿو ” (kitāb Akbar kān likhje tho, “The book is written by Akbar”) for masculine singular. Transition from active to passive voice cause a usual role shift like subject to oblique agent and object to passive subject marked by the UD relations *obl:agent* and *nsubj:pass* respectively. However, the auxiliary is not necessarily tagged with *aux:pass* when the morphological form of the main verb inherently indicates passivity as it serves solely to mark tense and is thus tagged with the general *aux* relation. This can be seen in the following example:



Similarly, “کتاب اکبر کان لکبو آهي” (kitāb Akbar kān likhbo āhe, “The book is written by Akbar”) and “چٺيون لکبيون آهن” (chithiyūn likhbiyūn āhan, “Letters are written”) show passive constructions by using the “ بو” (-bo) suffix, which varies by gender and number (e.g., “ لکبي” /likhbī/ for feminine singular).

As is true in many languages, including English, it is possible to use these constructions for each of past, present, and future tense. For further discussion of how these verbs inflect, see appendix C.

4.8 Non-projectivity

Sindhi is a highly projective language, with less than 6% of the sentences in the dataset having non-projective arcs. There are some cases where this occurs, though. An example non-projective sentence is given in figure 3.

4.9 Pronominal Suffixes

Sindhi employs pronominal suffixes on verbs, nouns, and postpositions to encode possessive, subject, or object roles, as in “ لکيومانس” (likhyomāns, “I wrote to him”). These suffixes blur the line between affixes (bound morphemes) and clitics (semi-independent), posing challenges in UD annotation for tokenization, UPOS tagging, and dependency relations.

To analyze these tokens, two approaches are used across UD datasets.

Single Token Approach Treat the word as one token with rich morphological features capturing the host’s POS (e.g., *VERB*) and suffix roles (e.g., *Person[Subj]*=1, *Person[iObj]*=3).

This can capture the morphology, but it obscures syntactic relations (e.g., ambiguity between object, indirect object, or oblique roles).

Split Token Approach Split suffixes into separate tokens (e.g., *PRON*), each with its own features and relations (e.g., *nsubj*, *iobj*).

This clarifies the syntax via relations. However, it risks ambiguity, as suffixes like “ س” (-s) vary by context (e.g., 3rd singular in مارينس” /mārīns/, “he was beaten”, vs. 1st singular in “ هئس” /hēs/, “I was”).

UD treebanks adopt either approach based on language-specific arguments and conventions, favoring single tokens for bound suffixes but splitting clitic-like forms. Challenges include assigning UPOS (*VERB* vs. *PRON*), defining relations, and resolving ambiguities in suffix meaning.

Here, we choose to split these suffixes using the multi-word token approach, with a compelling reason being that some of the pronouns themselves inflect with features such as *Case*, which are not adequately represented in the single token analysis.

5 Embedding Pretraining

As has been typical in the last decade of NLP work, before building the annotation pipeline, we pretrained static word embeddings, a character model (Akbik et al., 2018), and a transformer using a corpus of raw text. The corpus collected is 372M words, significantly less than the original BERT, which used 3,300M words for English (Devlin et al., 2019), but sufficient for pretraining models for a new language. This is significantly larger than that used in Muril and XLM-R, each with less than 40M.

Source	M Tokens	
Wikipedia	5	
Oscar CC	11	(Abadji et al., 2022)
Books Corpus	2	(Ali et al., 2020a)
Adabi Forums	0.5	
Sindhi LA Journals	0.8	
Sindhi LA Encyclopedia	5	
Various newswire	38	
Sindh Salamt Forums	24	
Sindh Salamt Books	57	
Sangraha	229	(Khan et al., 2024)

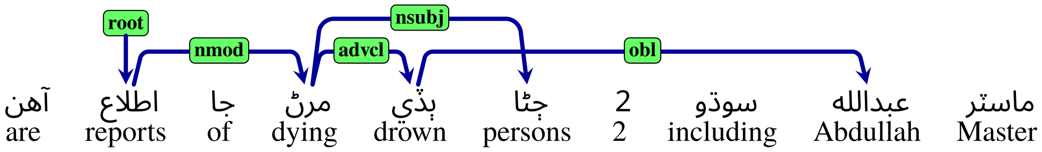


Figure 3: In this non-projective tree, عبدالله attaches to ٻڏي despite the intervening subject “Two people, including Master Abdullah, are reported to have drowned.”

We experimented with GloVe (Pennington et al., 2014), FastText (Grave et al., 2018), and word2vec (Mikolov et al., 2013), the latter two using the Gensim implementation (Rehurek and Sojka, 2011).

Model	NER Entity F1	Tags F1	Depparse LAS
Original	84.25	76.10	80.92
GloVe	86.07	77.90	81.69
Word2Vec	85.22	77.24	81.96
FastText	82.12	77.09	81.75
GloVe w/ CharLM	86.11	80.90	82.84

Table 3: Dev scores for three annotators using static word embeddings, potentially with charlm

Based on these results, we use the GloVe model, to be distributed with the Stanza pipeline for Sindhi. Other pretrained models will be available along with the default Stanza pipeline.

Stanza (Qi et al., 2020) implements a contextual character model. Table 3 also demonstrates the benefits of this model.

We used HuggingFace’s Roberta implementation to finetune Muril and to train two candidate Sindhi transformers from a random initialization. Interestingly, and unfortunately, each of the models produced were less accurate on the three tasks compared with the original Muril-Large (see table 4). Possible explanations include that Muril incorporated an additional machine translation learning objective and a large multilingual corpus. Future work will involve experimenting with multilingual datasets with the larger Sindhi collection or with a more targeted low-resource technique such as MicroBERT (Gessler and Zeldes, 2022).

Model	NER Entity F1	Tags F1	Depparse LAS
Muril	87.06	78.03	84.30
Finetuned	86.78	80.10	83.13
6 layers	82.66	75.59	81.07
12 layers	81.75	76.31	81.77

Table 4: Dev scores for three annotators using transformers

6 Annotation Pipeline

Having annotated a UD dataset and built some pretrained models, we then built a pipeline using Stanza annotation software. For NER, we used the existing SiNER dataset (Ali et al., 2020b).

Lemma and NER scores are high. Surprisingly, feature tagging is presently the lowest, performing below dependency and NER. Continued improvements to the annotations and the embeddings should improve scores.

Task	Emb	Score
Lemma	seq2seq w/ charlm	99.08 Accuracy
UPOS	Muril	98.44 F1
Features	Muril	85.56 F1
Depparse	Muril	91.83 LAS
NER	Muril	88.13 Entity F1

Table 5: Test scores for 5 Sindhi tasks

For a breakdown of individual relations, see Appendix E.

7 Conclusion

This project explores the more challenging structures for analyzing Sindhi grammar using Universal Dependencies. We then analyze roughly 6000 sentences of Sindhi, building a dataset for release in the 2.16 and 2.17 releases of UD. In parallel, we collect a larger corpus of Sindhi text than previously collected, facilitating more accurate static embeddings for the language. Combining each of these tasks with Stanza results in the first end-to-end annotation pipeline for the Sindhi language.

Limitations

Stanza includes three additional annotators, currently not implemented. A coreference dataset would be a useful addition to Sindhi annotation, but was beyond the scope of this work. As noted in section 2, there are sentiment datasets published for Sindhi, but as of this writing none have been made publicly available. There is also a constituency parser, but given the recent trend towards dependency parsing instead of constituency parsing, it is unlikely there will be a gold annotation of this treebank for constituencies.

The dataset utilised in this study has several limitations that may impact the reliability of our findings. The data primarily consisted of text from the Kawish newspaper. Many of the sentences are headlines, resulting in significant syntactical shortcomings, as these brief statements often highlighted named entities without forming complex linguistic structures. Key syntactical elements were largely absent due to newspaper headlines' inherent brevity and formulaic nature.

Furthermore, the dataset showed a significant lack of tense variation, with an overrepresentation of present-tense constructions to convey immediate facts. In contrast, past and future tense forms were either underrepresented or completely absent. This lack of tense diversity hindered capturing important temporal variations essential for comprehensive language modeling, resulting in limited morphological variation in our model.

Additionally, the repeated presence of named entities, geographical references, and thematic content typical in news publications generated potential biases in the distributional properties of the language representation. To improve future iterations of this research, a more diverse Sindhi language corpus should be developed using high-quality sources, including:

- **Sindh Textbook Board**[6]. Provides academic and domain-specific linguistic material.
- **Sindhi Adabi Board**[7]. Offers access to classical and modern Sindhi literature.
- **Encyclopedia Sindhiana**[8]. A rich source for definitional and encyclopedic language patterns.
- **Literary works of various authors**, such as Anwer Pirzado. Contributes complex syntactic structures and culturally rich narratives.

Incorporating these resources will enhance linguistic richness, expand tense usage, reduce redundancy, and increase both lexical and syntactic diversity—ultimately supporting the development of more robust and representative Sindhi language models.

Ethics Statement

As an annotation project for a language primarily spoken in Pakistan and India, with two non-local authors on the author list, the primary ethical concern for this paper would be that of "parachute research", as described in Odeny and Bosurgi (2022) for the medical domain.

To ensure fairness in the process, authorship credits were offered to each of the annotators. Furthermore, Professor Rahman, leading the research group at ISRA, occupies the lead author position in the author list.

As justification for the inclusion of the non-local authors, one provided extensive linguistics advice, whereas the other contributed significant technical time in terms of aligning the treebank with UD requirements, building models, and submitting pull requests and/or issues to conllueditor, Stanza, and Semgrex for features needed to support this project.

In terms of research relevant to the local community, one of the authors of the paper is a native speaker who started the project out of a desire to see this type of annotation pipeline available for Sindhi.

It is our hope that these practices will not only call attention to the strong linguistics group at Isra University, but also facilitate further collaborations between "Global North" research groups and local research groups.

Acknowledgments

We would like to thank the members of the UD community for their efforts advising the annotation process. We also appreciate the valuable feedback from our SyntaxFest reviewers. MLtwist put in a lot of work without which this project would not have been possible. Finally, we thank the licensees of CoreNLP for their continued support of StanfordNLP, which provided funding for this work.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *arXiv e-prints*, arXiv:2201.06642.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–

---

[6] `https://stbb.edu.pk`

[7] `https://sindhiadabiboard.net/library/`

[8] `https://encyclopediasindhiana.org/`

1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Meesum Alam, Francis Tyers, Emily Hanink, and Sandra Kübler. 2024. Universal Dependencies for Saraiki. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 188–197, Torino, Italia. ELRA and ICCL.

Wazir Ali, Jay Kumar, Junyu Lu, and Zenglin Xu. 2020a. Word embedding based new corpus for low-resourced language: Sindhi. *Preprint*, arXiv:1911.12579.

Wazir Ali, Junyu Lu, and Zenglin Xu. 2020b. SiNER: A large dataset for Sindhi named entity recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2953–2961, Marseille, France. European Language Resources Association.

Wazir Ali, Zenglin Xu, and Jay Kumar. 2021. SiPOS: A benchmark dataset for Sindhi part-of-speech tagging. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 22–30, Online. INCOMA Ltd.

Aryaman Arora. 2022. Universal Dependencies for Punjabi. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5705–5711, Marseille, France. European Language Resources Association.

John Bauer, Chloé Kiddon, Eric Yeh, Alex Shan, and Christopher D. Manning. 2023. Semgrex and ssurgeon, searching and manipulating dependency graphs. In *Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023)*, pages 67–73, Washington, D.C. Association for Computational Linguistics.

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746, Berlin, Germany. Association for Computational Linguistics.

Akshar Bharati, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. 2008. Anncorra : Annotating corpora guidelines for pos and chunk annotation for Indian languages. Technical report, Language Technologies Research Center, IIIT Hyderabad.

Riyaz Ahmad Bhat, Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, and 1 others. 2017. The hindi/urdu treebank project. In *Handbook of Linguistic Annotation*. Springer Press.

Flavio Massimiliano Cecchini. 2021. Formae reformandae: for a reorganisation of verb form annotation in Universal Dependencies illustrated by the specific case of Latin. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 1–15, Sofia, Bulgaria. Association for Computational Linguistics.

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. *Preprint*, arXiv:1911.02116.

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

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.

Mazhar Dootio and Asim Wagan. 2017. Automatic stemming and lemmatization process for Sindhi text. *JSSIR*, 6:19–28.

Mazhar Ali Dootio and Asim Imdad Wagan. 2019. Syntactic parsing and supervised analysis of Sindhi text. *Journal of King Saud University – Computer and Information Sciences*, 31(1):105–112.

Luke Gessler and Amir Zeldes. 2022. MicroBERT: Effective training of low-resource monolingual BERTs through parameter reduction and multitask learning. In *Proceedings of the 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pages 86–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Fadli Aulawi Al Ghiffari, Ika Alfina, and Kurniawati Azizah. 2024. Cross-lingual transfer learning for Javanese dependency parsing. *Preprint*, arXiv:2401.12072.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Johannes Heinecke. 2019. ConlluEditor: a fully graphical editor for universal dependencies treebank files. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 87–93, Paris, France. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Mohammed Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad B, Varun G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh Khapra. 2024. IndicLLMSuite: A blueprint for creating pre-training and fine-tuning datasets for Indian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 15831–15879. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. MuRIL: Multilingual representations for Indian languages. *Preprint*, arXiv:2103.10730.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Marie Mikulová, Milan Straka, Jan Štěpánek, Barbora Štěpánková, and Jan Hajic. 2022. Quality and efficiency of manual annotation: Pre-annotation bias. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2909–2918, Marseille, France. European Language Resources Association.

Bharti Nathani, Nisheeth Joshi, and G.N. Purohit. 2020. Design and development of unsupervised stemmer for Sindhi language. *Procedia Computer Science*, 167:1920–1927. International Conference on Computational Intelligence and Data Science.

Beryne Odeny and Raffaella Bosurgi. 2022. Time to end parachute science. *PLOS Medicine*, 19(9):1–3.

Martha Palmer, Rajesh Bhatt, Bhuvana Narasimhan, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. Hindi syntax: Annotating dependency, lexical predicate-argument structure, and phrase structure. In *The 7th International Conference on Natural Language Processing*, pages 14–17.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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*.

Loganathan Ramasamy and Zdeněk Žabokrtský. 2012. Prague dependency style treebank for Tamil. In *Proceedings of Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1888–1894, İstanbul, Turkey.

Vinit Ravishankar. 2017. A Universal Dependencies treebank for Marathi. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 190–200, Prague, Czech Republic.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

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.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

A Features

Table 6 demonstrates the morphological features used for the various verb forms described in section 4.6.

B Postposition Examples

As explained in section 4.4, genitive adpositions in Sindhi inflect to agree with the associated noun, whereas ADPs representing location do not inflect. For example, consider the locative ADPs in these sentences:

- “ مٿي تي ٽوپي” (mathe te topī, “A cap on the head”) the noun “ مٿو” (matho, “head”) shifts to its oblique form “ مٿي” (mathe) before “ تي” (te).
- “ پاڙي ۾ دڪان” (pāṛe mē dukān, “A shop in the neighborhood”) “ پاڙو” (pāṛo, “neighborhood”) becomes “ پاڙي” (pāṛe).

Example	Verb	Verb Type	VerbForm	Aspect	Gender	Number	Other
هو لکندو آهي hu likhando āhe, “He writes”	لکندو	Pres. Part	Part	Imp	Masc	Sing	–
هي ڪتاب لکيل آهي hī kitāb likhil āhe “This book is written”	لکيل	Past Part.	Part	Perf	–	–	Voice=Pass
مون کي ڪتاب لکڻو آهي mūn khē kitāb likhaṇo āhe “I have to write a book”	لکڻو	Fut Part	Inf	Imp	Masc	Sing	–
آئون چٺي لکي گهر آيس āūn chithī likhi ghar āyas “I came home after writing a letter”	لکي	Converb	Conv	Perf	–	–	–
ڪتاب لکندڙن کي ٻڌايو kitāb likhandaran khē buḍhāyo “Tell the writers of the book”	لکندڙن	Verbal Noun	Vnoun	Imp	–	Plur	Case=Acc
ڪتاب لکڻ سٺو آهي kitāb likhaṇ sutho āhe “It is good to write a book”	لکڻ	Infinitive	Inf	Imp	–	–	–

Table 6: Morphological features for example verb phrases

- “ مون وٽ ويھ” (mūn vat vēh, “Sit near me”) “ آئون” (āūn, “I”) shifts to “ مون” (mūn “me”).

In these examples, the locative ADPs mark the accusative/oblique case of the preceding nouns, which inflect accordingly (e.g., “ مٿو” → “ مٿي”), but the ADPs themselves remain uninflected and are typically tagged as locative postpositions (PSPL).

In contrast, genitive ADPs, such as “ جو” (jo, “of”), also govern the accusative/oblique case of preceding nominals but are themselves inflected for number, gender, and case, agreeing with the following noun. Consider these examples:

- “ گهوڙي جو پڇ” (ghoṛe jo puchh, “the tail of the horse”) “ گهوڙو” (ghoṛo, “horse”) becomes “ گهوڙي” (ghoṛe, accusative singular masculine), and “ جو” (jo) is nominative singular masculine, matching “ پڇ” (puchh, “tail”).

- “ گهوڙي جا ڪن” (ghoṛe jā kan, “the ears of the horse”) “ جو” inflects to “ جا” (jā, nominative plural masculine) for “ ڪن” (kan, “ears”).

- “ گهوڙي جون اکيون” (ghoṛe jūn akhiyūn, “the eyes of the horse”) “ جو” becomes “ جون” (jūn, nominative plural feminine) for “ اکيون” (akhiyūn, “eyes”).

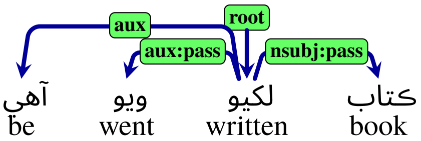
- “ گهوڙي جي پڇ جو وار” (ghoṛe je puchh jo vār, “the hair of the tail of the horse”) “ گهوڙي جي پڇ” (ghoṛe je puchh) uses “ جي” (je, accusative singular masculine) governed by “ پڇ” (puchh, accusative), while “ جو” (jo) is nominative singular masculine for “ وار” (vār, “hair”).

The genitive ADP “ جو” (jo) inflects based on the number and gender of the following noun, defaulting to nominative case in the first three examples. In the last example, its accusative form “ جي” (je) reflects the case of the intermediate noun “ پڇ” (puchh), which is accusative as it is followed by ADP, while the final “ جو” (jo) matches “ وار” (vār). Thus, the genitive ADP’s case aligns with the following noun’s case, accusative if the following noun is accusative, nominative otherwise. The number and gender also agree with the noun.

C Passive Voice Examples

Here, we continue the discussion of the passive voice started in section 4.7.

The present tense perfective aspect in Sindhi employs two distinct patterns for passive voice constructions. The first pattern uses a perfective verb form followed by “ ويو آهي” (viyo āhe, “has been”), where “ ويو” (viyo) serves as an intermediate auxiliary marking passivization, while “ آهي” (āhe) indicates present tense. For example, the active sentence “ اڪبر ڪتاب لکيو آهي” (Akbar kitāb likhyo āhe, “Akbar has written a book”) becomes “ ڪتاب اڪبر کان لکيو ويو آهي” (kitāb Akbar kān likhyo viyo āhe, “The book has been written by Akbar”). In this case, “ ويو” (viyo) is tagged with the UD relation *aux:pass* because it forms the passive voice, while “ آهي” (āhe) retains the *aux* relation as a tense marker. This can be seen in following graph (The obl:agent اڪبر is not shown):



The second pattern involves a perfective morphological passive form, inflected for number and gender, paired with a present tense auxiliary that solely marks tense, not passivization. Here, the auxiliary does not require the *aux:pass* relation, as the main verb's morphology already indicates the passive voice. This can be seen in “ كتاب لكجيا آهن” (kitāb likhjiyā āhan, “Books have been written”), where “ لكجيا” (likhjiyā) is a masculine plural passive form with “ كتاب” (kitāb, “books”) as *nsubj:pass*, and “ آهن” (āhin) is tagged as *aux* for present tense.

In the same way past imperfective passive is formed by the passive verb form combined with past tense auxiliaries (with usual number, gender inflections) like “ لكجي ها” (likhje hā, “would have written”) and “ لكبا هئا” (likhbā huā, “were being written”). The past perfective passive mirrors the present perfective with past tense auxiliaries.

Similarly, Sindhi future passive formations employ various patterns combining perfective or passive verb forms with future tense auxiliaries, such as ويندو” (vendo) or هوندو” (hondo), to indicate passivization.

Despite the variety of patterns, the Universal Dependencies (UD) relation patterns for future tense passive formations—including *nsubj:pass* (passive subject), *obl:agent* (oblique agent), and *aux:pass* (passive auxiliary)—remain consistent with those of the present tense passive constructions discussed earlier. The choice of UD relations depends on the auxiliary's role in marking passivity versus tense, but the syntactic structure mirrors that of the present tense.

D Use of Semgrex and Ssurgeon

As the annotation scheme changed, this project made extensive use of two tools for editing and checking the results of those edits: Semgrex and Ssurgeon (Bauer et al., 2023), and CoNLLUEditor (Heinecke, 2019). Both author groups were quite responsive in adding new features to support the needs of this particular project.

In the early stages of annotation, Future Participles were labeled as VerbForm=FutPart, an annotation that does not exist in current UD. As a first step to update the features to match the annotations described in this paper, we needed to update all existing Inf annotations to match the scheme described in section 4.6. We used figure 4 for this update.

After performing this update, we then searched for other verbs labeled Inf which may not fit the expected annotation patterns using figure 5.

Another example was an edit of a token where the intensifier was incorrectly tokenized as part of the word, which we edited with figure 6.

In order to search for nonprojective arcs, we used a Semgrex expression which searched for an arc going from a node before the current word to after the current word, then an arc going from the current word to a word on either side of that arc using a query shown in figure 7.

Searching for features, searching against features, splitting words, and searching for undirected connections are examples of some of the features added to support this project.

E Dependency Statistics

We present statistics on the dependencies present in the test set, along with the F1 scores for those dependencies from the best Stanza model.

Reln	F1	Total
acl	0.7586	40
acl:relcl	0.4286	5
advcl	0.8387	385
advmod	0.8852	405
advmod:emph	0.9329	142
amod	0.9206	472
appos	0.6667	2
aux	0.9605	436
case	0.9885	1436
cc	0.9366	244
ccomp	0.5185	48
compound	0.8696	711
conj	0.9464	327
cop	0.9564	266
dep	0.7403	62
det	0.9177	161
discourse	0.7619	11
dislocated	0.6667	4
fixed	0.9333	8
flat	0.9655	56
iobj	0.5806	21
mark	0.9119	353
nmod	0.9196	1148
nsubj	0.9136	896
nsubj:pass	0.1111	15
nummod	0.9565	88
obj	0.8565	416
obl	0.8763	790
parataxis	0.4000	3
punct	0.9465	1047
root	0.9763	887
vocative	0.6667	1
xcomp	0.6440	173

```
{word:/^.* ڻ$/;cpos:VERB;morphofeatures:{VerbForm:Inf}}=word
EditNode -node word -morphofeatures Aspect=Imp|VerbForm=Inf
```

Figure 4: An Ssurgeon pattern for updating existing infinitives to match a new feature pattern

```
{morphofeatures:{Aspect!:Imp;VerbForm:Inf}}
```

Figure 5: A Semgrex search for features which do not match the expected Infinitive guidelines

```
{word:/^ يئردنا$/}=split <=edge {}
splitWord -node split -exact ردنا -exact يئ -reln advmod:emph
    -headIndex 0 -name 0=adv,1=emph
editNode -node emph -pos PART -cpos PART -after “ ”
editNode -node adv -pos ADP -cpos ADV -remove morphofeatures
relabelNamedEdge -edge edge -reln advmod
```

Figure 6: A Ssurgeon pattern for splitting an incorrectly attached intensifier, then applying tags and dependencies

```
{} .. {}=later -- ({}=earlier <> {}=later)
    [<> ({}=attach -- {}=later) | <> ({}=attach .. {}=earlier)]
```

Figure 7: A Semgrex search for finding non-projective arcs

Universal Dependencies Treebank for Khoekhoe (KDT)

Kira Tulchynska[1]**, Sylvanus Job**[2]**, Alena Witzlack-Makarevich**[1,3]

[1]Hebrew University of Jerusalem, [2]University of Namibia, [3]University of Cologne,

Correspondence: kira.tulchynska@mail.huji.ac.il

Abstract

This paper reports on the development of the first dependency treebank for Khoekhoe (KDT). Khoekhoe (Khoe-Kwadi, Namibia) is a low-resource language with few linguistic and computational resources available publicly. This treebank consists of 29k words across six texts taken from various registers. It includes a substantial portion of spoken conversational data. These sentences were annotated manually according to the Universal Dependencies framework. In this paper, apart from presenting the strategies that have been followed to create the treebank, we also discussed some challenging morphological features and syntactic constructions found in the corpus and outlined how we have handled them using the current Universal Dependencies specification.

1 Introduction

Universal Dependencies (UD) is a cross-linguistically consistent framework for annotating various aspects of morphosyntax (McDonald et al., 2013; Nivre et al., 2016; de Marneffe et al., 2021). It provides a set of guidelines that enable comparison between treebanks across diverse linguistic typologies.

This paper reports the development of the Khoekhoe Dependency Treebank (KDT) a UD treebank for Khoekhoe, a Khoe-Kwadi language spoken primarily in Namibia. The treebank has been released as UD_Khoekhoe-KDT, an official Universal Dependencies (UD) treebank, available at `https://universaldependencies.org/treebanks/naq_kdt/` under the CC BY-SA 4.0 license. This represents the first UD treebank not only for this language family but also for any of the so-called Khoisan languages, a group of about a dozen languages of Southern Africa characterized by an extensive use of click consonants in their phonology (Güldemann, 2014; Witzlack-Makarevich and Nakagawa, 2019).

The annotation process highlighted several morphosyntactic properties of Khoekhoe that required careful consideration within the UD framework. These include the treatment of mood particles, valency-changing suffixes (e.g. applicative, reflexive), the use of nominalization as a clausal embedding strategy, and reported speech constructions involving quotative particles. In each case, we propose UD-compliant solutions and, where appropriate, introduce new feature values or relation subtypes.

2 Khoekhoe

Khoekhoe (`ISO 639-3: naq`; `Glottocode: nama1264`), also known as Nama-Damara, is a Khoe language of the Khoe-Kwadi family (Güldemann, 2014). It is spoken mainly in central and southern Namibia. Khoekhoe represents a dialect continuum (Haacke, 2018), it includes the northern varieties of Hai‖om and ‖Ākhoe, which differ considerably from the central and southern varieties. This paper is based on the corpus which includes only the latter varieties, which are closer to what is considered as standardized Khoekhoe.

With approximately 245,000 speakers (11.8% of the total population of Namibia), Khoekhoe is the second most spoken language in the country after Oshiwambo (Namibia Statistics Agency, 2011). Although Khoekhoe is well-documented and receives official recognition as a language of instruction in Namibia's educational system (Brenzinger, 2013), and despite being the largest non-Bantu click language in Africa, Haacke and Eiseb (2002) consider it to be an 'endangered language'.

The majority of Khoekhoe speakers are multilingual: many possess at least a functional proficiency in Afrikaans, with English fluency steadily increasing since Namibia's independence (Haacke, 2018, p. 142). Consequently, code-switching and word borrowing from Afrikaans and English to Khoe-

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 119–128
August 27, 2025 ©2025 Association for Computational Linguistics

khoe are very frequent.

3 Corpus

The Khoekhoe corpus used for the treebank currently contains approximately 29k words across six texts. The selection was primarily determined by practical considerations (availability and need of interlinear glossed texts for other research purposes). We balanced the corpus to include texts from multiple modes and genres. Currently, the corpus is skewed towards translated texts. We plan to expand its original spoken and written component in later releases.

One text (5.5k words) is a transcription of a conversation in Khoekhoe between two friends. It was transcribed and translated into English by native speakers, then supplemented with interlinear morphosyntactic glosses in ELAN (Max Planck Institute for Psycholinguistics, 2024). All the sentences containing personal information that could identify the speakers were removed. The remaining sentences were shuffled to ensure further anonymization.

All the other five texts are written. They include translations of the Cairo CICLing Corpus[1] (215 words) and the BivalTyp dataset (Witzlack-Makarevich and Job, 2024) (750 words); subtitles of the films *Bridge of Spies* (2015) (14k words) and a section of *Titanic* (1997) (6k words) translated from English by a native speaker; and chapters from the schoolbook *Khomai da ra I* (1971) with sentences shuffled (2k words). Also these texts were supplemented with interlinear morphosyntactic glosses in ELAN (Max Planck Institute for Psycholinguistics, 2024).

Text	Mode	Text origin	Size
conversation_Windhoek5	Spoken	Original	5.5k
grammar_Cairo	Written	Translation	215
grammar_BivalTyp	Written	Translation	750
film_Bridge	Written	Translation	14k
film_Titanic	Written	Translation	6k
book_Khomai	Written	Original	2k

Table 1: Corpus information

4 Methodology and Overview

The treebank annotations were performed manually by the first author on the basis interlinear morphosyntactic glosses. We plan to cross-check the annotations in the future versions of the treebank. Ambiguous structures were resolved through consensus among all authors.

For the first 60 sentences (330 tokens), a script was created for tokenization and conversion of interlinear morphosyntactic glosses into CoNLL-U format. This annotated sample was then used to train a UDPipe 1 (Straka and Straková, 2017) pipeline. This pipeline was subsequently applied to parse the next batch of sentences. The parsed sentences were manually reviewed, and this process was repeated iteratively until the entire corpus was annotated. UDPipe 1 was selected over UDPipe 2 (Straka, 2018) for this process because it is more user-friendly, and any potential differences in annotation quality would become negligible after manual review.

Annotation guidelines were developed alongside the annotation process, with all newly encountered issues being addressed before proceeding with the next batch.

Of the 17 universal part-of-speech tags recognized by the UD framework, only `SYM` is not currently used in the KDT treebank. Table 2 provides an overview of the frequencies of the universal part-of-speech tags in KDT.

Class	UPOS	Count	%
Open	ADJ	775	2.67
	ADV	1680	5.79
	INTJ	635	2.19
	NOUN	3537	12.19
	PROPN	820	2.83
	VERB	3941	13.59
Closed	ADP	1405	4.84
	AUX	5694	19.63
	CCONJ	652	2.25
	DET	827	2.85
	NUM	191	0.66
	PART	477	1.64
	PRON	3269	11.27
	SCONJ	917	3.16
Other	PUNCT	4144	14.29
	X	44	0.15

Table 2: UPOS tags and their frequencies in KDT

Table 3 presents the features, their values, and counts in KDT, with features not currently recognized by UD marked with †. Some features (e.g. `Assoc`) are introduced here for the first time, while others (e.g. `Voice=Rfl`) have been used in other treebanks. The new features and values, as well as language-specific decisions regarding the feature used with the declarative auxiliary *ge*, are discussed in Section 5.

---

[1] `https://github.com/UniversalDependencies/cairo/tree/master`

Feature	Values	Count	%
Abbr	Yes	21	0.05
Aspect	Imp, Perf, Punct†	1335	2.96
Assoc†	Yes†	17	0.04
Case	Acc, Nom, Voc	7240	16.07
Clusivity	Ex, In	135	0.30
Degree	Dim	86	0.19
Deixis	Contr†, Prox, Remt	860	1.91
Evident	Nfh	8	0.02
ExtPos	ADP, ADV, CCONJ, DET, SCONJ	56	0.12
Foreign	Yes	31	0.07
Gender	Fem, Masc, Neut	6286	13.96
Mood	Ass†, Imp, Ind, Int, Pot, Prh†	2694	5.98
Number	Dual, Plur, Sing	7453	16.55
NumType	Card, Ord	225	0.50
Person	1, 2, 3	7421	16.48
Polarity	Pos, Neg	758	1.68
Poss	Yes	238	0.53
PronType	Dem, Emp, Ind, Int, Neg, Prs, Rel, Tot	4410	9.79
Tense	Fut, Past, Pres, RecPast†	1758	3.90
Typo	Yes	77	0.17
Voice	Act, Appl†, ApplPass†, ApplRefl†, Pass, Rcp, Refl†	3930	8.73

Table 3: Features, values and their frequencies in KDT. † marks features and values introduced in KDT

Finally, the 15 most frequent dependencies in KDT are listed in Table 4. Syntactic analysis and annotation decisions for language-specific structures are discussed in Section 6.

Dependency	Count	%
aux	5299	18.27
punct	4144	14.29
root	3589	12.37
nsubj	2859	9.86
advmod	1714	5.91
case	1455	5.02
obj	1270	4.38
obl	1016	3.50
mark	975	3.36
cc	662	2.28
nmod:poss	660	2.28
det	564	1.94
amod	455	1.57
conj	419	1.44
advcl	386	1.33
...		
other	3541	12.21

Table 4: Dependencies and their frequencies in KDT

5 Morphology

5.1 Associative Plural

Associative plurals are found in most languages across Australia, Asia, and Africa. They are almost entirely absent in Western Europe. This structure conveys the meaning of 'and others associated with it' (Daniel and Moravcsik, 2013).

In Khoekhoe, the associative plural is marked by the suffix *-hâ*, which precedes the person-gender-number suffix. For example, *Jackhân* means 'Jack and company'; *saruhân* (derived from *saru* 'cigarette') means 'cigarettes and other things associated with smoking' (see Hagman 1977, p. 29). The associative plural suffix precedes either a dual or a plural suffix (e.g. the common plural suffix *-n* in the two examples). Thus, it cannot be analyzed as a value of the `Number` feature. Instead, we introduce an `Assoc` feature with the value `Yes` to account for its function.

5.2 Mood Values

The values which are coded as `Mood` in the UD annotation scheme include both the typical mood categories (sometimes referred to as *verbal mood*), such as potential, as well as values which could be better analyzed as *sentence mood*, such as interrogative and imperative (see Portner 2018, pp. 4–5). These two sub-types of mood are not distinguished in UD.

For Khoekhoe we use seven mood values. In addition to the familiar imperative (`Imp`), potential (`Pot`), and interrogative (`Int`) moods, Khoekhoe has four further moods, for which dedicated values had to be introduced to UD. These are the indicative (`Ind`, apprehensive (`App`), assertive (`Ass`), and prohibitive (`Imp`) moods. All grammatical mood categories in Khoekhoe are expressed by grammatical particles and are coded as `AUX` in the UD annotation scheme. The individual mood auxiliaries are in the majority of cases in complementary distribution. With some exceptions, they occur in the clause-second position.

5.2.1 Indicative vs. declarative

In Khoekhoe, declarative sentences can and are often marked with the particle *ge*, as in (1). The function of this particle can be best captured with the sentence mood label *declarative* (Portner 2018, pp. 4–5), as it does not occur in interrogative and imperative sentences. However, we decided against introducing a new `Mood` value and use instead the existing verbal mood value `Ind` in the annotation to enhance the cross-linguistic comparability of the annotations.

(1) *Sara-s* *ge* *‖khai-s-a* *ūhâ.*
Sara-3F.SG.SBJ DECL flu-3F.SG-OBL have
'Sara has the fly.'

5.2.2 Prohibitive

Whereas some languages use their regular sentential negation strategy to express prohibitions directed at second person (also known as negative imperatives), a substantial proportion of world's languages use to this end dedicated negation strategies not found in declaratives (see van der Auwera et al. 2013). In Khoekhoe, the regular negative particles are *tama* in the present and past and *tide* in the future. By contrast, in the prohibitive mood a dedicated particle *tā* is used instead. We introduce the `Mood` value `Prh` to annotate the auxiliary *tā* in this function, as illustrated in (2).

(2) *Tā* *ti* *ôa-b-a* *ǂnau.*
PROH my child-3M.SG-OBL beat
'Do not beat my child.'

5.2.3 Apprehensive

The term *apprehensive construction* is used to refer to constructions which conventionally encode, or pragmatically implicate, that the situation described by the clause is an undesirable possibility (Vuillermet et al., 2025+). As such, apprehensionality is a subtype of modality (AnderBois and Dąbkowski, 2025). In some languages, such as Khoekhoe, apprehensives are part of a grammatical mood paradigm, while in others, they are part of a different syntactic class and can include e.g. markers of negation (Vuillermet et al., 2025+).

We introduce a `Mood` value `App` to annotate the apprehensive mood marker *tā*, as in (3). This is the same particle used to mark the prohibitive (see Section 5.2.2), as is not uncommon cross-linguistically. However, the two contexts are easily distinguished: whereas the prohibitive is used in the imperative construction directed at the second person, which does not have an over expression of subject in the singular, the apprehensive is used in sentences which have overtly expressed subjects, such as *aorob* 'man' in (3).

(3) *Nēsisa* *aoro-b* *ge* *tā* *ti-ta*
now man-3M.SG.SBJ DECL APPR 1SG-1SG.OBL
nî *|gū.*
FUT come_near
'Now the man should not come near me.'

5.2.4 Assertive

Cross-linguistically, there is a range of forms and constructions whose relation to core mood is not currently well understood. Among them Portner (2018, p. 7) lists assertive. The details of the use of the assertive auxiliary in Khoekhoe are still understudied. It seems that speakers uses this form to emphasize the fact that they do not take responsibility or authority over the assertion. Thus, it might be more appropriately characterized as a member of the evidentiality system and not of the mood system, though cross-linguistically the two can share the same slot and be in complementary distribution, as is the case in Khoekhoe (see Aikhenvald 2003).

In Khoekhoe, the assertive marker consist of two parts *kom(o)* … *o*, as in (4). The first part *kom(o)* occurs in the clause-second position, the second element *o* is clause-final. Since the assertive marker is discontnious and does not enclose the whole clause, the structure could not be analyzed as a fixed multi-word expression. Thus, both parts are coded as auxiliaries with `Mood=Ass`.

(4) *‖Nā-n* *komo* *awoxa-n*
DIST-3C.PL.SBJ ASSERT1 ancestor-3C.PL
ǂû-n-a *o.*
food-3C.PL-OBL ASSERT2
'Those are the foods of the ancestors.'

5.3 Voice Values

Khoekhoe has four valency-changing verbal suffixes: passive, reciprocal, reflexive, and applicative (Hagman, 1977, pp. 77–82). Two of these (passive and reciprocal) are already recognized as possible `Voice` values in UD, while the other two (reflexive and applicative) are not.

In UD, reflexivity is recognized as a possible feature of pronouns and determiners and is annotated with `Reflex`. However, in Khoekhoe, as in many other languages of the world (see e.g. 'Feature GB114: Is there a phonologically bound reflexive marker on the verb?' in Skirgård et al. 2023), reflexive events are encoded with a dedicated reflexive voice marked by the verbal suffix *-sen* (Hagman, 1977, pp. 81–82). It indicates that the agent and the patient are coreferential and decreases the syntactic valency of the predicate by one (Zúñiga and Kittilä, 2019, pp. 154–155), as in (5).

(5) *Hui-sen=ta ge.*
help-REFL=1SG DECL
'I help myself.'

This type of reflexive construction is not yet covered by the UD guidelines. To capture this typologically common feature, we use the feature value `Voice=Rfl`, which is already in use in the Abaza, Turkish, and Turkish-German treebanks.

The applicative construction increases the syntactic valency of the verb by one: it introduces a new direct object that corresponds to a non-core argument in the non-applicative voice (Zúñiga and Kittilä, 2019, p. 53). The applicative voice in Khoekhoe is marked by the suffix *-ba* (Hagman, 1977, pp. 78–79), as in (6). To capture this morphosyntactic behavior, we use `Appl` as a value of the `Voice` feature.

(6) *Ti-ta ge ǂhanu-b-a*
1SG-1SG.SBJ DECL government-3M.SG-OBL
sîsen-ba tama hâ.
work-APPL NEG PFV
'I don't **work for** the government.'

Additionally, Khoekhoe verbal morphology allows the applicative voice to combine with the other voice categories (Hagman, 1977, p. 77). This is typologically common, e.g. Zúñiga and Kittilä (2019, p. 75) suggest that in many languages that have both applicative and passive morphological markers, these categories co-occur, as they do in Khoekhoe, as in (7). To account for such combinations, we introduce the following complex values: `ApplPass`, `ApplRcp`, and `ApplRfl`.

(7) *Sîsenao-gu ge kausa tsaugoma-b-a*
worker-3M.PL.SBJ DECL fat ox-3M.SG-OBL
go ǂā-ba-he.
RPST slaughter-APPL-PASS
'One fat ox was slaughtered for the workers.
(lit. The workers were slaughtered one fat ox for.)'

As in some other languages, Khoekhoe has a number of morphologically reflexive, reciprocal, and applicative verbs which are arguably lexicalized. Geniušienė (1987, p. 31) discusses this common issue with respect to reflexives and proposes a distinction between the familiar reversible reflexive verbs and the less studied class of non-reversible reflexive verbs. She suggests the following criteria of reversibility (Geniušienė, 1987, pp. 145–148) to distinguish between the two: (1) morphological reversibility, i.e. a situation when a derived unit is formally related to a base word, morphological non-reversibles are traditionally known as *reflexiva tantum*; (2) syntactic reversibility, viz. a change of reversible reflexive properties according to one of the regular patterns; (3) lexical reversibility, viz. the identity of lexical distribution relative to the corresponding syntactic positions in a non-reflexive construction and related reflexive construction; (4) semantic reversibility, viz. a regular, standard change of the meaning of a reflexive, thus, semantic non-reversible reflexive verbs have the meaning which is related to that of the base non-reflexive way in some idiosyncratic way. Individual morphologically reflexive verbs can be non-reversible according to several of the criteria (2) to (4). Similar observations apply to reciprocal and applicative verbs, see e.g. Peterson (2006, pp. 169–170) and individual contributions in Zuniga and Creissels (2024) on the lexicalization of applicatives.

Though in theory the distinction between lexicalized and productive cases might be clear and there are straightforward examples of lexicalization in Khoekhoe, such as e.g. *mû* 'to see' vs. *mûsen* 'to appear, to look like', in other cases the application of the criteria in (2) to (4) can be admittedly challenging. In KDT we do not annotate for voice unambiguous cases of lexicalized reflexives, reciprocals, and applicatives. The decision is made on a lexeme-by-lexeme basis.

6 Syntax

6.1 Quotative marker

Many African languages have particle-like function words dedicated to marking reported discourse (Güldemann, 2008, pp. 122–124). In Khoekhoe, *ti* is used as the quotative marker. It directly follows the quoted material, as in (8) (Hagman, 1977, pp. 136–138).

(8) *[ǀOwesa=ta a]* ***ti****=n ge*
lazy=1SG.SBJ COP.PRS QUOT=3C.PL.SBJ DECL
ra mî.
IPFV say
'They say [I am lazy].'

The complementizer *!khais* is sometimes used in place of the quotative marker *ti* to express indirect speech, as in (9). This usage of *ti* can therefore be analyzed and annotated as a subordinating conjunction marker (`mark`), as in Figure 1.

(9)

Mî-ba	*tsi=ta*	*go*	*[‖nā*
say-APPL	2M.SG.OBJ=1SG.SBJ	RPST	DIST

‖khami-s-a=ts	*nî*	*ū*	*ǂgao]*
phone-3F.SG-OBL=2M.SG.SBJ	FUT	take	want

!khai-s-a
COMPL-3F.SG-OBL

'I told you that you wanted to take the call.'

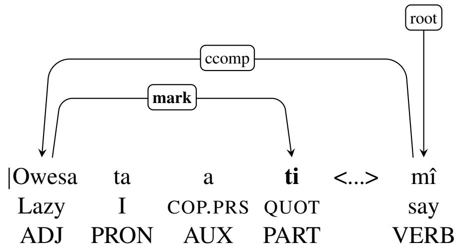


Figure 1: Annotation of (8)

This quotative marker is also used to introduce noun phrases expressing names and labels, as in (10). This usage is annotated with the `case` relation, as in Figure 2. The marker was preliminary tagged as `PART`, but it will be changed to `SCONJ` in the next version of the treebank.

(10)

Nē	*‖gaba-s*	*ge*	*[kompas]*	***ti***
PROX	tool-3F.SG.SBJ	DECL	compass	QUOT

ra	*‖gai-he.*
IPFV	call-PASS

'This tool is called a compass.'

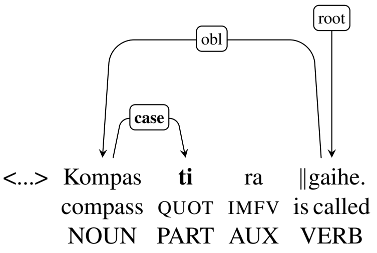


Figure 2: Annotation of (10)

6.2 Nominalization

In Khoekhoe, one way to embed a clause is through nominalization. This process involves the use of the nominalizer enclitic =*s*, which is formally identical to the third-person feminine singular suffix and enclitic. The nominalized clause then functions as a nominal within the sentence, that is, it can take the oblique case suffix, it can be followed by a postposition, and it can functions as a clausal argument, e.g. object (Hagman, 1977, pp. 123–135), as in (11).

(11)

ǂÂihō=ta	*ge*	*ra*	*‖î-n*
remeber=1SG.SBJ	DECL	IPFV	3-3PL.SBJ

gere	*‖nāti*	*mî=s-a*
PST.IPFV	like_that	say=NMLZ-OBL

'I remember them saying like that.'

6.2.1 Analysis of nominalized constructions

Nominalized constructions can be analyzed in two ways: either with the nominalizer as the head of the structure or with it depending on the root of the nominalized clause.

The first approach treats the nominalizer as an *empty noun*, with the nominalized clause acting as a relative clause clause modifier. Under this analysis, the literal translation of the nominalized structure in (11) would be 'one/the thing which is them saying'. This annotation is illustrated in Figure 3.

The second approach treats the nominalizer as a subordination marker dependent on the root of the nominalized clause. In this case, the literal translation of the nominalized structure in (11) would be 'that they said'. This annotation is illustrated in Figure 4.

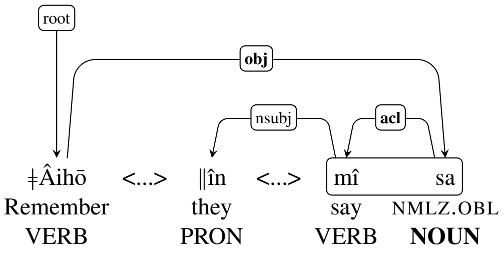


Figure 3: Head-nominalizer analysis of (11)

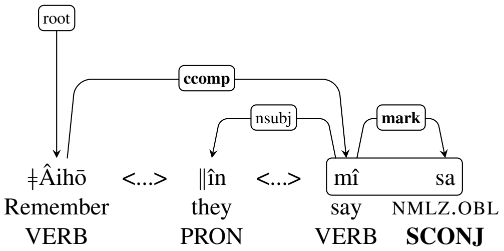


Figure 4: Dependent-nominalizer analysis of (11)

A key advantage of the head-nominalizer analysis is that it highlights the fully nominal function of the nominalized constituent: the root of the structure is a noun, it depends on the main clause's root through a nominal dependency (e.g. `obj` vs. `ccomp`), and it can directly take a `case` dependency from a postposition.

On the other hand, the dependent-nominalizer analysis more directly represents the embedded clause and aligns better with cross-linguistic annotation practices. This approach focuses on the syntactic relation between the subordinate and main clauses, rather than the language-specific realization of it. Given its greater cross-linguistic consistency, we adopt the dependent-nominalizer analysis, treating the nominalizer as a subordination marker.

6.2.2 Nominalization with additive focus

Another usage of the nominalizer =*s* in Khoekhoe is on the constituents modified by the additive focus adverb *tsîn* 'also', when the focused element that precedes the adverb is not a nominal. The nominalized focused element is not necessarily a clause – as in (12) – but can be a constituent of any type, e.g. a postpositional phrase, as in (13). The combination of the nominalizer and the additive focus particle, =*s tsîn*, is analyzed as a fixed multiword expression (and thus annotated with `fixed`), which functions as an emphasizing adverbial (`advmod:emph`, with `ExtPos=ADV` on the first element, *s*). The annotation of the structure is illustrated in Figure 13.

(12) *Tsî=s* *ge* *!hāsara* *te=s*
and=2F.SG.SBJ DECL insult 1SG.OBJ=NMLZ
tsîn-a *go* *dī.*
also-OBL RPST do
'And you've also insulted me.
(Lit. And you've also done the insulting of me.)'

(13) *ǁî-s* *!nâ=s* *tsîn-a=ta* *ge*
3-3F.SG in=NMLZ also-OBL=1SG.SBJ DECL
ǁkhaisa *kurixa* *|gôa-b-a* *ūhâ.*
eight year_old son-3M.SG-OBL have
'Also with (lit. in) her I have an eight-year-old son.'

6.3 Flat Structures

UD already recognizes two subtypes of the `flat` relation: `flat:foreign` and `flat:name`. For Khoekhoe, we introduce three additional subtypes: `flat:num`, `flat:reparandum`, and `flat:title`.

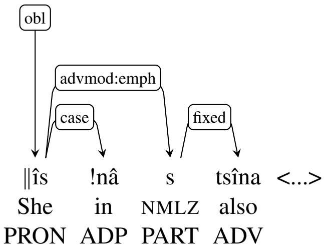


Figure 5: Annotation of =*s tsîn* structure in (13)

6.3.1 Numerals

In Khoekhoe, larger numbers are expressed using separate words for different place values (e.g. *ǁkhaisadīsi haka kurigu* 'eighty-four years'). These multi-word numerals may include a coordinating conjunction (e.g. *|gui|oadīsi khoesekaidīsi* ***tsî*** *hakadīs |gui|aǁî kuri* 'one thousand nine hundred **and** forty-first year'), but otherwise, the structure lacks a clear syntactic head.

The final word in the numeral can be either a cardinal (annotated as `NUM` with `NumType=Card`) or an ordinal (annotated as `ADJ` with `NumType=Ord`). All preceding parts of the numeral are always cardinals and are annotated accordingly.

One approach to annotating such numerals is to use the `compound` relation, as is done in languages such as English and Russian. However, it has two drawbacks. First, the `compound` relation is not used elsewhere in Khoekhoe, making this an isolated and atypical application. Second, it is unclear which word in the multi-word numeral should be treated as the root.

An alternative approach is to use a numeral-specific subtype of the `flat` relation. The `flat:num(ber)` relation is already used for similar cases in Komi Zyrian, Persian, and Vietnamese.

Given these considerations, we adopt an analysis where multi-word numerals are treated as (mostly) flat structures. If a coordinating conjunction is present, the `conj` and `cc` relations are used to connect coordinating elements and the conjunct *tsî* 'and.' Within these elements, numerals are linked to the first word using the `flat:num` relation. Thus, the first word in the structure is treated as a technical root of the numeral.

Depending on whether the numeral expresses a cardinal or an ordinal number (determined by the final word), the first word in the structure attaches to the modified noun using either the `nummod` relation (Figure 6) or the `amod` relation (Figure 7).

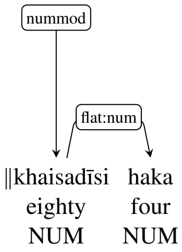


Figure 6: Cardinal multi-word numeral

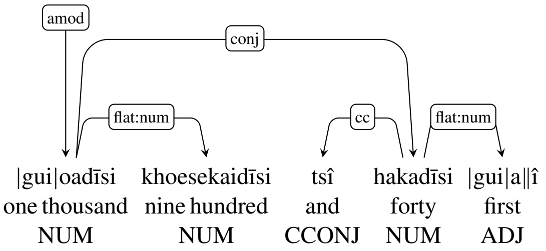

amod
conj
flat:num
cc
flat:num
|gui|oadīsi khoesekaidīsi tsî hakadīsi |gui|a||î
one thousand nine hundred and forty first
NUM NUM CCONJ NUM ADJ

Figure 7: Ordinal multi-word numeral

6.3.2 False start

Part of KDT include various speech disfluencies, such as false starts. In the transcription conventions applied to compile the corpus, false starts are marked with square brackets.

If a false start contains an incomplete word, it is tagged as `X` and left unannotated for morphological features (see Figure 8). Otherwise, if the word is complete, it is fully morphologically annotated (see Figure 9). In all cases, a false start is treated as an overridden disfluency and therefore depends on the repair using the `reparandum` relation.

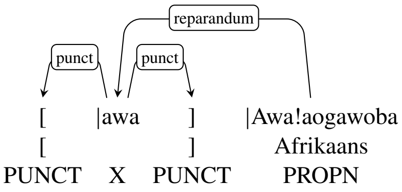


Figure 8: False start with an incomplete word

In many instances, speakers produce multiple false starts of the same word (e.g. *[ti ti] ti zulas* '[my my] my struggle'). One possible approach would be to attach each false start separately to the repair using the `reparandum` relation. However, this would prevent the paired punctuation marks surrounding the false start from attaching to the same word without creating a non-projective structure. To avoid this issue, we instead analyze the entire false start as a flat structure and annotate it using the newly introduced `flat:reparandum` relation, as in Figure 10.

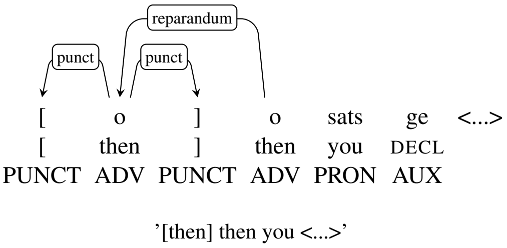


Figure 9: False start with a complete word

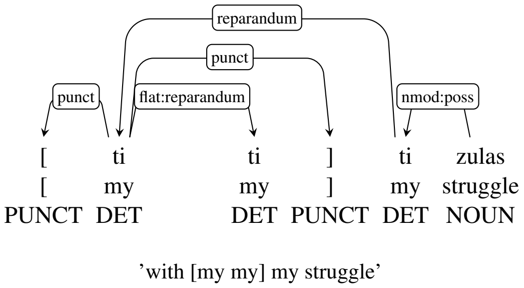


Figure 10: Multiple false starts

6.3.3 Titles

Another common type of flat structure consists of noun phrases that include a title and a proper name, such as *Mr. Smith*, *painter Picasso*, and *brother Sam*. In fact, more than a third (80 out of 218) of the `flat` relations in the Khoekhoe treebank involve a `NOUN` representing a title, profession, or kinship term of a following `PROPN`. Given the frequency of this pattern, we introduce the `flat:title` relation subtype to account for these structures, as in Figure 11.

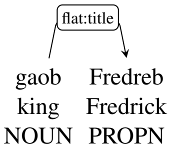


Figure 11: Flat title relation

7 Conclusion

The Khoekhoe treebank presented in this paper is the first case of a Khoisan language added to UD. Khoekhoe has a range of typologically common mood and grammatical voice features so far underrepresented in UD treebanks. We furthermore present several solutions to issues present primarily in corpora of spontaneous spoken language, such as false starts.

Acknowledgments

This research has been funded by Israel Science Foundation, project "Peripheral Khoekhoe varieties: A comprehensive documentation and description" (Personal research grant no. 2892/20); Global Strategy and Partnerships Funding Scheme of the University of Zurich, project "Event Packaging in Language"; the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 281511265 – SFB 1252 "Prominence in Language" at the University of Cologne.

References

Alexandra Y. Aikhenvald. 2003. Evidentiality in typological perspective. In Alexandra Y. Aikhenvald and R. M.W. Dixon, editors, *Studies in evidentiality*, pages 1–33. John Benjamins, Amsterdam.

Scott AnderBois and Maksymilian Dąbkowski. 2025. The semantics and expression of apprehensional modality. *Language and Linguistics Compass*, 19(1):e70002.

Matthias Brenzinger. 2013. The twelve modern Khoisan languages. In Alena Witzlack-Makarevich and Martina Ernszt, editors, *Khoisan languages and linguistics: Proceedings of the 3rd International Symposium, July 6-10, 2008, Riezlern/Kleinwalsertal*, Research in Khoisan Studies. Cologne: Rüdiger Köppe.

Michael Daniel and Edith Moravcsik. 2013. The associative plural (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

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

Emma Geniušienė. 1987. *The typology of reflexives*. Mouton de Gruyter, Berlin.

Tom Güldemann. 2008. *Quotative indexes in African Languages*. Mouton de Gruyter, Berlin.

Tom Güldemann. 2014. 'Khoisan' linguistic classification today. In Tom Güldemann and Anne-Maria Fehn, editors, *Beyond 'Khoisan': Historical relations in the Kalahari Basin*, pages 1–41. John Benjamins, Amsterdam.

Wilfrid H. G. Haacke. 2018. Khoekhoegowab (nama/damara). In Tomasz Kamusella and Finex Ndhlovu, editors, *The social and political history of Southern Africa's languages*, pages 133–158. Palgrave Macmillan, London.

Wilfrid H. G. Haacke and Eliphas Eiseb. 2002. *A Khoekhoegowab dictionary with an English-Khoekhoegowab index*. Gamsberg Macmillan, Windhoek.

Roy Stephen Hagman. 1977. *Nama Hottentot Grammar*. Indiana University publications. Research Center for Language and Semiotic Studies, Indiana University.

The Language Archive Max Planck Institute for Psycholinguistics. 2024. ELAN. Computer program.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Namibia Statistics Agency. 2011. Namibia household income & expenditure survey 2009/2010. Technical report, Namibia Statistics Agency, Windhoek.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 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, Portorož, Slovenia. European Language Resources Association (ELRA).

David A. Peterson. 2006. *Applicative constructions*. Oxford University Press, Oxford.

Paul Portner. 2018. *Mood*. Oxford University Press, Oxford.

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 Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tânia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O. C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175.

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 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, Vancouver, Canada. Association for Computational Linguistics.

Johan van der Auwera, Ludo Lejeune, and Valentin Goussev. 2013. The prohibitive (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Marine Vuillermet, Eva Schultze-Berndt, and Martina Faller. 2025+. Introduction. In *Apprehensional constructions in a cross-linguistic perspective*. Language Science Press, Berlin.

Alena Witzlack-Makarevich and Sylvanus Job. 2024. Bivalent patterns in Khoekhoe. In *BivalTyp: Typological database of bivalent verbs and their encoding frames*.

Alena Witzlack-Makarevich and Hirosi Nakagawa. 2019. Linguistic features and typologies in languages commonly referred to as 'Khoisan'. In Ekkehard Wolff, editor, *The Cambridge handbook of African linguistics*, pages 382–416. Cambridge University Press, Cambridge.

Fernando Zuniga and Denis Creissels, editors. 2024. *Applicative Constructions in the World's Languages*. De Gruyter Mouton, Berlin.

Fernando Zúñiga and Seppo Kittilä. 2019. *Grammatical Voice*. Cambridge University Press, Cambridge.

Parallel Universal Dependencies Treebanks for Turkic Languages

Arofat Akhundjanova[1], **Furkan Akkurt**[2],
Bermet Chontaeva[3], **Soudabeh Eslami**[3], **Çağrı Çöltekin**[3]
[1]Independent Researcher, [2]Boğaziçi University [3]University of Tübingen,
arofat.akhundjanova@gmail.com furkan.akkurt@bogazici.edu.tr
{bermet.chontaeva, soudabeh.eslami}@student.uni-tuebingen.de
cagri.coeltekin@uni-tuebingen.de

Abstract

We introduce a set of fully aligned and manually annotated parallel Universal Dependencies (UD) treebanks for four Turkic languages: Azerbaijani, Kyrgyz, Turkish, and Uzbek. The treebanks were annotated with close collaboration between annotators to ensure the harmonized annotations across all languages. These resources currently consist of 148 strategically selected sentences that illustrate typologically significant morphosyntactic phenomena across these related languages. These parallel treebanks enable systematic comparative studies of Turkic syntax and may be instrumental in cross-lingual NLP applications. All treebanks are available as part of UD v2.16.

1 Introduction

The Universal Dependencies (UD) framework has emerged as the leading standard for cross-linguistically consistent grammatical annotation, covering over 150 languages across more than 300 treebanks (Nivre et al., 2020). However, the coverage and depth of resources for Turkic languages are limited and vary considerably in treebank size and availability across the language family. More critically, the absence of parallel treebanks for these languages limits systematic cross-linguistic comparisons, despite their shared typological characteristics and historical ties. To the best of our knowledge, there are only two parallel treebanking efforts that include a Turkic language (Turkish in both): (1) the PUD treebank, created for the CoNLL 2017 Shared Task on Multilingual Parsing (Zeman et al., 2017), (2) the Atis treebank (Cesur et al., 2024). The PUD treebank consists of 1000 manually translated sentences in 22 languages, including Turkish. The Atis treebank contains a translation of the English ATIS (Airline Travel Information System) corpus to Turkish.

As of UD version 2.16, there are 24 treebanks for 11 Turkic languages (including historical and code-switching varieties, see Table 1 in Appendix A for an overview). As pointed out by multiple earlier studies, annotation inconsistencies are common in Turkic treebanks (Türk et al., 2019; Çöltekin et al., 2023). Some of these issues were addressed in the recent work, such as pronominalized locatives (Washington et al., 2024) and postverbial constructions (Akhundjanova, 2025). However, there remains a notable gap in cross-linguistic research based on parallel corpora for Turkic languages.

This paper addresses this gap by introducing the first fully aligned parallel UD treebanks for four Turkic languages: Azerbaijani, Kyrgyz, Turkish, and Uzbek. These languages represent distinct branches of the Turkic family—Oghuz (Azerbaijani, Turkish), Kipchak (Kyrgyz), and Karluk (Uzbek)—allowing for meaningful cross-branch comparisons. Even though most treebanks introduced in this work were annotated by a single annotator (one of the authors), the treebanks are cross-checked by others to ensure the consistency of annotations as well as their quality.

Our approach combines carefully constructed grammatical examples with linguistic expertise to create resources that highlight both the shared typological features and unique syntactic traits of each language.

The treebanks are available as part of Universal Dependencies v2.16 and support applications in cross-lingual parsing, comparative research, and language education.

2 Languages and Corpora

2.1 Overview of Selected Languages

Turkic languages are agglutinative, head-final, and typically pro-drop (Johanson, 2021). The four languages covered in our study represent three

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 129–136
August 27, 2025 ©2025 Association for Computational Linguistics

Language	UD Treebanks	Size	Parallel?
Azerbaijani	TueCL	Small	Yes
Kazakh	KTB	Small	No
Kyrgyz	TueCL, KTMU	Medium	Yes
Tatar	NMCTT	Small	No
Turkish	Kenet, Penn, Tourism, Atis, GB, FrameNet, IMST, BOUN, PUD, DUDU, Tonqq	Large	Yes (PUD, Atis)
Uyghur	UDT	Medium	No
Uzbek	UDT	Small	No
Yakut	YKTDT	Small	No

Table 1: Status of UD treebanks for Turkic languages as of version 2.15. Treebank sizes are classified based on token counts: small (under 20K tokens), medium (20K–100K tokens), and large (over 100K tokens). The “Parallel?” column indicates whether the language has a treebank which is part of a cross-linguistic parallel corpus. The parallel treebanks for Azerbaijani and Kyrgyz are also part of the current study.

distinct branches of the Turkic family: Azerbaijani and Turkish from the Oghuz branch, Kyrgyz from the Kipchak branch, and Uzbek from the Karluk branch (Johanson, 2022). While they share core typological features, notable differences in morphology and syntax make them well-suited for comparative analysis.

Common challenges in Turkic UD treebanks arise from several language-specific features. Agglutinative morphology leads to long, complex word forms that complicate tokenization and annotation processes. Syntactic analysis is further challenged by complex verbal constructions, such as serial verbs and auxiliary chains. Moreover, lower-resource languages like Azerbaijani, Kyrgyz, and Uzbek lack robust automatic parsing tools, limiting corpus development and comprehensive analysis.

2.2 Source Data

Our parallel treebanks are based on a curated collection of 148 sentences, compiled from multiple sources with linguistic annotation in mind. Specifically, we draw from the Cairo corpus (20 sentences),[1] the UDTW23 corpus (20 sentences),[2] and 97 additional examples illustrating grammatical constructions of interest. We added 9 sentences to the Cairo corpus to capture alternative annotations for pronominal subject omission. Although they appear duplicated, each has a unique sentence ID and reflects a distinct annotation—with and without pronouns. Two new sentences were added to the original UDTW23 corpus for additional analysis of language-specific issues in Kyrgyz. Most of the source sentences originate in Turkish and were manually translated into Azerbaijani, Kyrgyz, and Uzbek, preserving semantic and syntactic alignment across languages. The tokenized data in the Turkish treebank amounts to 904 tokens. Additionally, English translations are provided as metadata to facilitate cross-lingual comparisons. Table 2 shows basic statistics for the parallel treebanks.

Statistic	AZ	KY	TR	UZ
Tokens	912	1048	904	940
Avg. sent. length	6.2	7.1	6.1	6.4
POS tags	15	16	14	15
Dependencies	34	38	37	33
Avg. dep. length	2.3	2.4	2.3	2.4

Table 2: Basic statistics for the parallel treebanks. AZ: Azerbaijani, KY: Kyrgyz, TR: Turkish, UZ: Uzbek.

In translating the data, we prioritized maintaining structural alignment while ensuring idiomatic and grammatical correctness in each target language. This allows for consistent parallelism that supports both syntactic and semantic cross-linguistic analyses.

Sentences were selected based on their value for (cross)linguistic analysis, with a focus on morphosyntactically rich or typologically marked constructions. These include pro-drop sentences, auxiliary chains, postverbal structures, and non-canonical word orders—phenomena that are both theoretically significant and empirically challenging for dependency annotation and parser development.

The treebanks are encoded in Latin script for Azerbaijani, Turkish, and Uzbek. The Kyrgyz corpus is presented in its native Cyrillic script, with transliteration and interlinear glosses provided in the metadata (see Appendix B for an annotation sample).

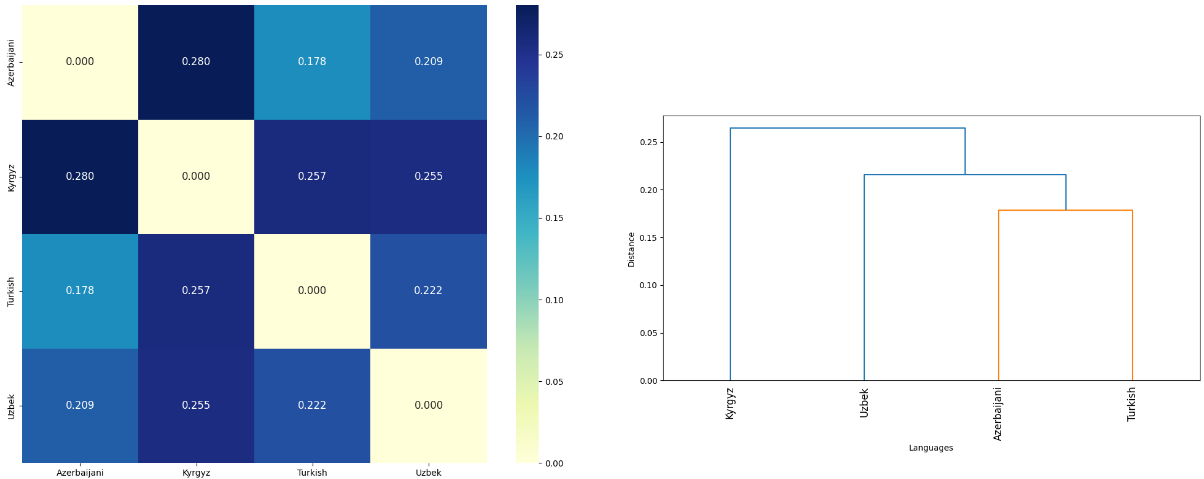
3 Treebank Construction

This section outlines our treebank development process, including the general annotation methodology and language-specific approaches for each treebank. We also provide a quantitative analysis of the parallel corpora to highlight typological similarities and distinctive features of

---

[1] `https://github.com/UniversalDependencies/cairo`

[2] `https://github.com/ud-turkic/udtw23`



(a) Normalized edit distances based on POS sequences.

(b) The dendrogram for language clustering, showing structural similarities among the languages.

Figure 1: Normalized edit distances between languages and language clustering based on these distances.

these languages. Additionally, we discuss some annotation issues and describe our approach to resolving them.

3.1 Annotation Workflow

Our approach integrates automated processing with manual annotation and revision. When feasible, processing tasks were initially automated and then refined by native speakers with relevant expertise. Each treebank underwent thorough manual verification and correction, in line with the UD guidelines.[3] Ambiguous cases were addressed through collaborative discussions with linguists and UD experts.

The development of the four treebanks followed different paths based on existing resources and language-specific requirements. The Azerbaijani and Kyrgyz treebanks, both named *TueCL*, were created prior to the current project (Eslami and Çöltekin, 2024; Chontaeva and Çöltekin, 2024). They were enriched with additional grammar examples and morphological features as part of this project. Turkish-TueCL and Uzbek-TueCL, in contrast, were developed from scratch with distinct approaches.

For Turkish, we first followed two different strategies independently, (1) fully manual annotation and (2) automatic annotation followed by manual correction, after which we manually merged the annotations. For automatic annotations, we used the large language model Claude 3.5 Sonnet (Anthropic, 2025).

Regarding Uzbek, tokenization was automated with the NLTK,[4] but all other annotation layers were performed manually.

3.2 Quantitative Analysis

Using parallel treebanks, we performed a quantitative analysis to highlight structural and typological features of the target languages. It is important to acknowledge that the use of constructed sentences may limit the generalizability of our findings to more natural language use.

Figure 1a presents the normalized edit distances between languages based on POS sequences, while Figure 1b shows the resulting language clustering derived from these distances. The results show that Azerbaijani and Turkish cluster closely with the shortest edit distance, while Kyrgyz shows the longest distances, particularly from Azerbaijani, reflecting less POS alignment and greater structural divergence. Overall, these findings confirm that typological relationships between languages are reflected in our annotations.

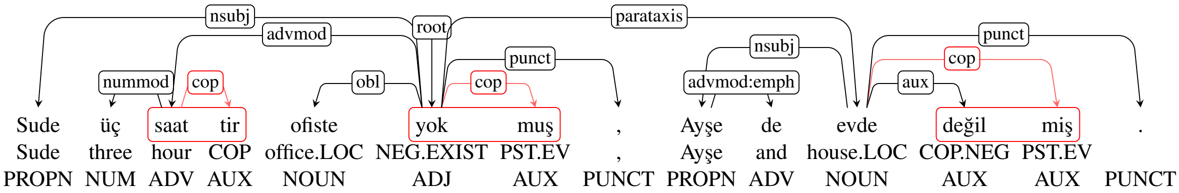
Among the languages studied, Turkish exhibits the shortest dependency lengths (up to 15 tokens), whereas Uzbek displays the longest (up to 21 tokens). This variation may stem from differences in translation styles and structural preferences.

In terms of word order, all languages predominantly adhere to a subject-object-verb (SOV) structure. Relative clauses consistently precede head nouns, conforming to the typical

---

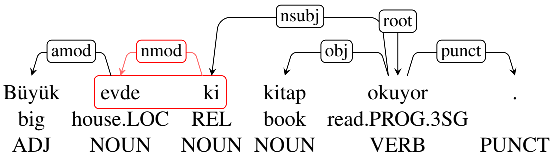
[3] `https://universaldependencies.org/guidelines`

[4] `http://www.nltk.org/api/nltk.tokenize.html`


nsubj
advmod
root
parataxis
nummod
cop
obl
punct
cop
nsubj
advmod:emph
aux
cop
punct
Sude üç saat tir ofiste yok muş , Ayşe de evde değil miş .
Sude three hour COP office.LOC NEG.EXIST PST.EV , Ayşe and house.LOC COP.NEG PST.EV
PROPN NUM ADV AUX NOUN ADJ AUX PUNCT PROPN ADV NOUN AUX AUX PUNCT

‘Sude was not at the office for three hours and Ayşe was not at home.’

Figure 2: Example annotation of several copular constructions in Turkish. Copular elements are treated as separate subtokens from the syntactic head and are highlighted within red frames.


amod
nmod
nsubj
obj
root
punct
Büyük evde ki kitap okuyor .
big house.LOC REL book read.PROG.3SG
ADJ NOUN NOUN NOUN VERB PUNCT

‘The one in the big house is reading (a) book.’

Figure 3: Example annotation of a pronominalized locative in Turkish with a *-ki* construction treated as a separate subtoken from the syntactic head and highlighted within a red frame.

head-final nature of Turkic syntax. In modifier constructions, Turkish often places determiners after adjectives, while the reverse is more common in the other languages, though both patterns are acceptable depending on stylistic variation and pragmatic factors.

3.3 Unique Features

Our cross-linguistic analysis revealed not only the common characteristics mentioned above but also unique features specific to individual languages. One such feature is the flexible positioning of the question particle *mi* in Turkish, which can shift the focus of a sentence depending on its placement. In contrast, the other languages in our study use question markers with a fixed position, consistently appearing at the end of the predicate.

Another feature related to question markers is the tendency in Azerbaijani to form questions without particles, both in written and conversational speech. However, such intonation-based questions are not widely used in the formal contexts of the other languages.

The use of posture and locational verbs, e.g., жат (‘jat’ to lie down), as an auxiliary to mark the progressive aspect is more characteristic and prominent in Kyrgyz. Uzbek also shows similar aspectual chaining with postverbal constructions, but such usage is usually limited to conversational speech. In contrast, in Turkish, such auxiliary constructions are not very productive and progressive aspect is primarily marked by inflectional morphology.

Another feature unique to Kyrgyz is the ability to form compound nouns without a possessive suffix—for example, алма дарактар (alma daraktar, ‘apple trees’)—whereas other languages in this study require a possessive suffix on the head noun (e.g., *olma daraxtlari* in Uzbek).

3.4 Annotation Challenges

Annotating Turkic languages in the UD framework has some challenges discussed in earlier studies (Tyers et al., 2017), some of which have been under active discussion recently (Taguchi, 2022). Here, we briefly note two issues where we adopt the current consensus that resulted from these recent discussions. The first is the copular constructions in Turkic languages, where copula can be realized as an affix attached to a nominal or adjectival predicate, making syntactic and morphological analysis difficult, and may result in inconsistent analyses of similar constructions. We treat copular affixes as separate syntactic units with the `AUX` POS tag and attach them to the main predicate with the cop dependency relation to better reflect their syntactic behavior. Figure 2 presents an example analysis with multiple copular affixes.

Another of the issues is pronominalized locatives with the suffix *-ki* as discussed in Washington et al. (2024). We follow the analysis in that work and treat *-ki* constructions as separate subtokens, as shown in Figure 3. Albeit making its automatic annotation non-trivial, this approach preserves all the linguistic information in both the genitive and locative variants of this construction.

4 Conclusion and Future Work

We presented the first fully aligned parallel Universal Dependencies treebanks for four Turkic languages. This work fills a significant gap in resources for comparative studies of Turkic syntax and provides valuable data for cross-lingual NLP applications.

Our future work includes extending these treebanks with additional texts from various genres and adding more Turkic languages, particularly from less-represented branches of the family. We also plan to conduct more detailed analyses of the morphosyntactic phenomena highlighted in this initial study.

We invite the research community to collaborate on this ongoing project to improve the representation of Turkic languages in language resources and models.

Limitations

Our parallel treebanks have several limitations. First, the relatively small size (148 sentences) may limit their utility for certain applications, particularly data-hungry machine learning tasks. Second, the focus on constructed examples, while valuable for highlighting specific linguistic phenomena, may not fully represent natural language usage. Third, the current version primarily addresses written language and formal registers.

Despite these limitations, these parallel treebanks provide a valuable starting point for cross-linguistic Turkic language studies and demonstrate the feasibility of our annotation approach for future extensions.

Acknowledgments

We thank the Turkic UD working group for fruitful discussions of linguistic issues and annotation approaches. This work was supported by COST Action CA21167 – Universality, diversity and idiosyncrasy in language technology (UniDive).

References

Arofat Akhundjanova. 2025. Harmonizing annotation of Turkic postverbial constructions: A comparative study of UD treebanks. In *Proceedings of the 18th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 10–17, Abu Dhabi, UAE. Association for Computational Linguistics.

Arofat Akhundjanova and Luigi Talamo. 2025. Universal Dependencies treebank for Uzbek. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 1–6, Tallinn, Estonia. University of Tartu Library, Estonia.

Furkan Akkurt, Nursena Teker, Helin Binici, Ahmet Demir, and Konstantinos Sampanis. 2025. UD Turkish-English BUTR. `https://github.com/UniversalDependencies/UD_Turkish_English-BUTR`.

Anthropic. 2025. Claude 3.5 Sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`. Accessed: 2025.

Ibrahim Benli. 2020. UD Kyrgyz KTMU. `https://github.com/UniversalDependencies/UD_Kyrgyz-KTMU`.

Neslihan Cesur, Aslı Kuzgun, Mehmet Kose, and Olcay Taner Yıldız. 2024. Building annotated parallel corpora using the ATIS dataset: Two UD-style treebanks in English and Turkish. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC)@ LREC-COLING 2024*, pages 104–110.

Neslihan Cesur, Aslı Kuzgun, Olcay Taner Yıldız, Büşra Marşan, Neslihan Kara, Bilge Nas Arıcan, Merve Özçelik, and Deniz Baran Aslan. 2022. UD Turkish Penn. `https://github.com/UniversalDependencies/UD_Turkish-Penn`.

Özlem Çetinoğlu and Çağrı Çöltekin. 2022. Two languages, one treebank: Building a Turkish–German code-switching treebank and its challenges. *Language Resources and Evaluation*, pages 1–35.

Bermet Chontaeva and Çağrı Çöltekin. 2024. UD Kyrgyz TueCL. `https://github.com/UniversalDependencies/UD_Kyrgyz-TueCL`.

Çağrı Çöltekin. 2015. A grammar-book treebank of Turkish. In *Proceedings of the 14th Workshop on Treebanks and Linguistic Theories (TLT 14)*, pages 35–49.

Çağrı Çöltekin, A. Seza Doğruöz, and Özlem Çetinoğlu. 2023. Resources for Turkish natural language processing: A critical survey. *Language Resources and Evaluation*, 57(1):449–488.

Mehmet Oguz Derin and Takahiro Harada. 2021. Universal Dependencies for Old Turkish. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 129–141, Sofia, Bulgaria. Association for Computational Linguistics.

Marhaba Eli, Weinila Mushajiang, Tuergen Yibulayin, Kahaerjiang Abiderexiti, and Yan Liu. 2016. Universal dependencies for Uyghur. In *Proceedings of the Third International Workshop on Worldwide*

Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016), pages 44–50, Osaka, Japan. The COLING 2016 Organizing Committee.

Soudabeh Eslami and Çağrı Çöltekin. 2024. UD Azerbaijani TueCL. `https://github.com/UniversalDependencies/UD_Azerbaijani-TueCL`.

Lars Johanson. 2021. The structure of Turkic. In Lars Johanson and Éva Á. Csató, editors, *The Turkic Languages*, pages 26–59. Routledge.

Lars Johanson. 2022. The history of Turkic. In Lars Johanson and Éva Á. Csató, editors, *The Turkic Languages*, pages 83–123. Routledge.

Aslı Kuzgun, Neslihan Cesur, Olcay Taner Yıldız, Oğuzhan Kuyrukçu, Büşra Marşan, Bilge Nas Arıcan, Neslihan Kara, Deniz Baran Aslan, Ezgi Sanıyar, and Cengiz Asmazoğlu. 2023. Turkish tourism treebank of Universal Dependencies 2.13. `https://hdl.handle.net/11234/1-5287`.

Aslı Kuzgun, Neslihan Cesur, Olcay Taner Yıldız, Oğuzhan Kuyrukçu, Arife Betül Yenice, Bilge Nas Arıcan, and Ezgi Sanıyar. 2022. UD Turkish Kenet. `https://github.com/UniversalDependencies/UD_Turkish-Kenet`.

Büşra Marşan, Neslihan Kara, Merve Özçelik, Bilge Nas Arıcan, Neslihan Cesur, Aslı Kuzgun, Ezgi Sanıyar, Oğuzhan Kuyrukçu, and Olcay Yıldız. 2021. Building the Turkish FrameNet. In *Proceedings of the 11th Global Wordnet Conference*, pages 118–125.

Tatiana Merzhevich and Fabrício Ferraz Gerardi. 2022. Introducing YakuToolkit: Yakut treebank and morphological analyzer. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 185–188.

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.

Şaziye Özateş, Tarık Tıraş, Efe Genç, and Esma Bilgin Tasdemir. 2024. Dependency annotation of Ottoman Turkish with multilingual BERT. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 188–196, St. Julians, Malta. Association for Computational Linguistics.

Umut Sulubacak, Memduh Gokirmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal Dependencies for Turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454, Osaka, Japan. The COLING 2016 Organizing Committee.

Chihiro Taguchi. 2022. Consistent grammatical annotation of Turkic languages for more universal Universal Dependencies. In *29th International Conference on Head-Driven Phrase Structure Grammar*.

Chihiro Taguchi, Sei Iwata, and Taro Watanabe. 2022. Universal Dependencies treebank for Tatar: Incorporating intra-word code-switching information. In *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*, pages 95–104, Marseille, France. European Language Resources Association.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Balkız Öztürk Başaran, Tunga Güngör, and Arzucan Özgür. 2019. Improving the annotations in the Turkish universal dependency treebank. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 108–115.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Gözde Berk, Seyyit Talha Bedir, Abdullatif Köksal, Balkız Öztürk Başaran, Tunga Güngör, and Arzucan Özgür. 2021. Resources for Turkish dependency parsing: Introducing the BOUN treebank and the BoAT annotation tool. *Preprint*, arXiv:2002.10416.

Francis Tyers, Jonathan Washington, Çağrı Çöltekin, and Aibek Makazhanov. 2017. An assessment of Universal Dependency annotation guidelines for Turkic languages. In *5th International Conference on Turkic Language Processing (TURKLANG 2017)*, pages 356–377.

Francis M. Tyers and Jonathan N. Washington. 2015. Towards a free/open-source universal-dependency treebank for Kazakh. In *3rd International Conference on Turkic Languages Processing (TurkLang 2015)*, pages 276–289.

Jonathan Washington, Çağrı Çöltekin, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, Gulnura Jumalieva, Aida Kasieva, Aslı Kuzgun, Büşra Marşan, and Chihiro Taguchi. 2024. Strategies for the annotation of pronominalised locatives in Turkic Universal Dependency treebanks. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 207–219, Torino, Italia. ELRA and ICCL.

Enes Yılandiloğlu and Janine Siewert. 2025. DUDU: A treebank for Ottoman Turkish in UD style. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 74–79, Tallinn, Estonia. University of Tartu Library, Estonia.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, and 43 others. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

A Turkic UD Treebanks

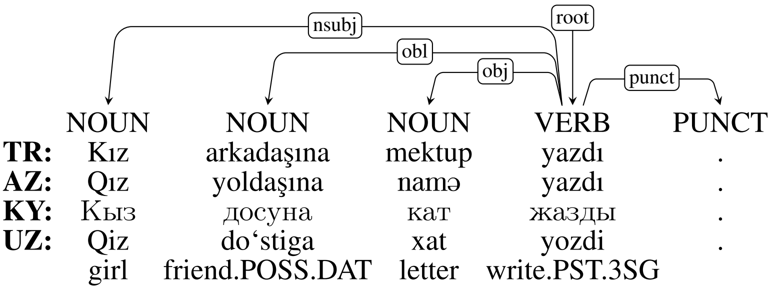
This appendix provides a comprehensive overview of all Turkic language treebanks available in Universal Dependencies version 2.16. These treebanks vary significantly in size, coverage, and domain, highlighting the uneven resource distribution across the Turkic language family. Table 3 presents key statistics for each treebank.

Treebank	Sentences	Tokens	Genre	Additional Information
Azerbaijani-TueCL (Eslami and Çöltekin, 2024)	148	912	grammar	Grammar examples
Kazakh-KTB (Tyers and Washington, 2015)	1078	10 536	news, fiction, wiki	Mixed sources
Kyrgyz-KTMU (Benli, 2020)	2480	23 654	news, fiction	Media texts
Kyrgyz-TueCL (Chontaeva and Çöltekin, 2024)	173	1250	grammar	Grammar examples
Tatar-NMCTT (Taguchi et al., 2022)	148	2280	news, non-fiction	Mixed sources
Turkish-Atis (Cesur et al., 2024)	5432	45 907	news, non-fiction	ATIS domain
Turkish-BOUN (Türk et al., 2021)	9761	125 212	news, non-fiction	Web sources
Turkish-FrameNet (Marşan et al., 2021)	2698	19 223	grammar	FrameNet annotations
Turkish-GB (Çöltekin, 2015)	2880	17 177	grammar	Grammar examples
Turkish-IMST (Sulubacak et al., 2016)	5635	58 096	news, non-fiction	METU-Sabancı corpus
Turkish-Kenet (Kuzgun et al., 2022)	18 687	178 658	grammar	Corpus of example sentences
Turkish-Penn (Cesur et al., 2022)	16 396	183 555	news, non-fiction	Turkish Penn Treebank
Turkish-PUD (Zeman et al., 2017)	1000	16 881	news, wiki	Parallel corpus
Turkish-Tourism (Kuzgun et al., 2023)	19 830	91 152	reviews	Hotel reviews
Turkish-German-SAGT (Çetinoğlu and Çöltekin, 2022)	2184	37 227	spoken	Code-switching corpus
Turkish-English-BUTR (Akkurt et al., 2025)	51	393	spoken	Code-switching corpus
Old Turkish-Tonqq (Derin and Harada, 2021)	20	158	non-fiction	Historical texts
Ottoman Turkish-BOUN (Özateş et al., 2024)	514	8834	fiction, non-fiction	Historical texts
Ottoman Turkish-DUDU (Yılandiloğlu and Siewert, 2025)	1064	10 012	bible, fiction, non-fiction, government, news	Mixed sources
Uyghur-UDT (Eli et al., 2016)	3456	40 236	fiction	Literary texts
Uzbek-UT (Akhundjanova and Talamo, 2025)	500	5850	news, fiction	Mixed sources
Yakut-YKTDT (Merzhevich and Gerardi, 2022)	299	1459	news, non-fiction	Media texts

Table 3: Basic statistics on Turkic UD treebanks available in version 2.16, excluding the Turkish and Uzbek treebanks introduced in this paper.

B Annotation sample of parallel sentences

# sent_id = cairo-1
text[tr] = Kız arkadaşına mektup yazdı.
text[az] = Qız yoldaşına namə yazdı.
# text[kir] = Кыз досуна кат жазды.
# translit[kir] = Qız dosuna qat jazdı.
# text[uz] = Qiz do‘stiga xat yozdi.
# glossing = girl friend-POSS.3SG-DAT letter write-PST.3SG
# text[en] = The girl wrote a letter to her friend.
# issue: obl vs. iobj


nsubj
obl
obj
root
punct
NOUN NOUN NOUN VERB PUNCT
TR: Kız arkadaşına mektup yazdı .
AZ: Qız yoldaşına namə yazdı .
KY: Кыз досуна кат жазды .
UZ: Qiz do‘stiga xat yozdi .
girl friend.POSS.DAT letter write.PST.3SG

‘The girl wrote a letter to her friend.’

Figure 4: Example annotation of parallel sentences in four languages.

Towards better annotation practices for symmetrical voice in Universal Dependencies

Andrew Thomas Dyer
Language Science and Technology
Saarland University
Saarbrücken, Germany
andrew.dyer@uni-saarland.de

Colleen Alena O'Brien
Department of Asian Studies
Palacký University Olomouc
Olomouc, Czechia
obrienca@hawaii.edu

Abstract

Austronesian languages exhibit features that are challenging for Universal Dependencies: most notably, the symmetric voice system, whereby agent, patient, and instrumental arguments (among others) can be the pivot of a transitive structure – complicating the usual assumption that subjects of transitive sentences are semantic agents, and objects semantic patients. To showcase our ideas of how to address the representation of such systems in Universal Dependencies, we introduce a small treebank of sentences from texts and elicitation sessions in Gorontalo, an Austronesian language of Sulawesi (Indonesia), which exhibits a Philippine-type voice system. We discuss the annotation guidelines for this language, and the extensions of the Universal Dependencies guidelines that are needed to accommodate this and other Austronesian languages.

1 Introduction

In Universal Dependencies (Zeman et al., 2024), as in linguistics in general, the representation of morphosyntactic alignment of core arguments is heavily influenced by the grammatical traditions of Indo-European languages – particularly Latin. In such languages there is a clear syntactic subject, which corresponds in an unmarked (active voice) sentence to the semantic agent (Andrews, 2007). Such sentences can be passivized, so that the construction becomes intransitive, the semantic patient is the subject, and the agent is demoted to the status of an oblique, or dropped completely.

Austronesian languages, especially Philippine-type languages, challenge this view by allowing arguments with various semantic roles to become the *pivot* of the sentence: the trigger of agreement, as shown by verbal morphology, and the syntactically privileged argument (Blust, 2013; Chen and Mcdonnell, 2019). The pivot of a transitive sentence can be actors, patients, instruments, locations, and many more. In many of these languages, no voice seems to be more privileged or less marked than any other; nor does the actor voice seem derived from the patient voice nor vice versa. Analysis of these languages is problematic, and there is great debate regarding whether to label them as ergative-absolutive (Aldridge, 2012); nominative-accusative (Chen, 2025); or a separative category (Foley, 2008), among other classifications.

The grammatical relation of subject has been noted to be problematic in several languages (Croft, 2003; Haspelmath and Sims, 2010), and in Austronesian languages other terms such as *pivot* or *focus* are often used instead. Nevertheless, the approach in Austronesian language treebanks so far has been to use the existing labels of UD for Austronesian languages – specifically, by representing the pivot argument of a sentence as the subject, regardless of the argument type.

This is a decent compromise, and it allows the Austronesian voice system to fall into the universal framework of UD. However, there are problems with this approach to annotation in the treebanks that exist so far, particularly for voices other than the actor voice.

In this paper we will:

- Compare different approaches in the existing language treebanks.
- Describe some of the potential problems with the current annotation choices for voices and core arguments in Austronesian languages.
- Suggest some improvements that can be consistently applied across languages.
- Showcase some of these decisions in our treebank for a new language: Gorontalo.

2 Related work

Several Austronesian treebanks of varying sizes already exist in Universal Dependencies, and we use them as precedent for our annotation. Table 1

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 137–142
August 27, 2025 ©2025 Association for Computational Linguistics

shows the currently existing treebanks.

Aside from the Austronesian treebanks themselves, de Marneffe et al. (2021) discuss the representation of symmetric voice in their description of the Universal Dependencies framework. We are not aware of other studies or opinion pieces on this topic.

Language	Treebanks
Indonesian	PUD, GSD, CSUI
Javanese	CSUI
Cebuano	GJA
Tagalog	TRG, Ugnayan

Table 1: A table of currently existing Austronesian language treebanks in Universal Dependencies.

3 Voice and Core Arguments in Universal Dependencies

3.1 General

Most of the well-resourced languages in Universal Dependencies use two main voices: active and passive. In a passive-marked sentence, the syntactic subject instead corresponds to the semantic patient, while the semantic agent becomes an optional oblique argument.

In annotation of active voice sentences, the agent-subject is given the bare dependency label *nsubj*, and the patient-object *obj* or *iobj*. In passive-voice clauses, the passive patient-subject takes the label *nsubj:pass*, showing that it is the subject of a passive construction. The agent is treated as an oblique, but to denote that the oblique is still an agent, the label *obl:agent* is used.

On the morphological level, the feature `Voice` is applied to the VERB or AUX that marks the voice, with the values *Act* or *Pass*. Since the active voice is considered unmarked in the majority of languages, the *Act* value of the `Voice` feature is often not applied, with only passive voice Pass applied when a verb or auxiliary is inflected for passive voice.

3.2 Austronesian languages

Austronesian languages in Universal Dependencies have been annotated to conform with this system as closely as possible. They usually treat the pivot argument as a subject, whether that be an actor, patient, instrument, or any other kind of argument.

Verbs use the `Voice`[1] feature for their voice inflections. Actor voice and patient voice are aliased to active voice and passive voice (*Act* and *Pass* respectively). Three additional `Voice` values are used only in Austronesian languages: *Ifoc* (instrument focus), *Lfoc* (location focus), and *Bfoc* (beneficiary focus). Table 2 shows the representation of these voices in the current Austronesian languages.

Voice	Full name	Used in
Act	Active voice	ind, jav, ceb, tgl
Pass	Passive voice	ind, jav, ceb, tgl
Ifoc	Instrument focus	ceb, tgl
Lfoc	Location focus	ceb, tgl
Bfoc	Beneficiary focus	tgl

Table 2: The distinct voice values contained in the current UD Austronesian languages: Indonesian (ind), Javanese (jav), Cebuano (ceb), and Tagalog (tgl).

The annotation of the syntactic relations between pivot and non-pivot arguments varies by treebank.

Indonesian and Javanese

Indonesian and Javanese in Universal Dependencies have only two voices attested, corresponding to actor and patient.[2] This makes them unproblematic for Universal Dependencies voice representation: actor voice is aliased with active voice; patient with passive. Active voice sentences have the labels of *nsubj* and *obj* respectively. In the Indonesian GSD (McDonald et al., 2013) treebank, patient-subjects are usually given the dependency label *nsubj:pass*, and the actor-object is given the label *obj*. In Indonesian CSUI (Alfina et al., 2020) and Javanese CSUI (Alfina et al., 2023), the labels *nsubj:pass* and *obl:agent* respectively are used, analogously with the handling of passives in European languages.

Cebuano

The Cebuano GJA treebank (Arañes, 2022) treats the pivot of a transitive sentence as the subject, whichever voice it is expressed in, with the other core arguments being treated as objects. Verbs are annotated with the `Voice` feature, but core argument dependencies are annotated with plain labels. Thus, in a patient voice sentence, an actor-object

---

[1] `https://universaldependencies.org/u/feat/Voice.html`

[2] The term *Undergoer* is sometimes preferred for the non-agent voice, as it can perform a broad set of semantic roles.

would just have the label of *obj* without any reference to its semantic role. Likewise in instrument or location voice sentences, other core arguments would have object labels, including patients and actors.

Tagalog

The Tagalog TRG treebank (Samson, 2018) has five voices attested. Pivots are treated as the subject (*nsubj*), and non-pivot arguments as objects (*obj*/*iobj*). However, this treebank often (but inconsistently) uses sub-deprels to differentiate between non-agent subject types, sometimes (but not always) using sub-labels such as *nsubj:pass* (patient), *nsubj:ifoc* (instrumental), etc for nominal subjects. Core non-pivot arguments take object labels, with sub-labels such as *obj:agent* in non-actor voice sentences. However, active voice transitive sentences are treated as unmarked, only using the standard labels without sub-labels.

3.3 Our criticisms

Our biggest point of contention is that the use of Indo-European labels/annotations for Austronesian languages muddles the actual syntax of these languages. We find with the current annotation norm of using active and passive as aliases for actor and patient voice problematic, because these voices are not actually equivalent. Non-pivot actors of a patient-voice sentence are *not* demoted to obliques the way they are in passive voice, but remain core arguments of the verb (Chen, 2025), and there are ongoing debates about the transitivity of actor- and patient-voice clauses.

The labels could be misleading to users familiar with the passive in Indo-European languages, who may expect similar properties in what is labelled as passive in Austronesian languages: namely that the agent of a passive sentence is optional and oblique, whereas in reality patient voice sentences still usually expect an actor to be mentioned as a core argument.

Though some treebanks annotate the non-pivot agent as an object with the labels *obj* or *iobj*, retaining its core status, this is still confusing, because semantic agents are not generally thought of as objects of transitive clauses, nor do they necessarily function as grammatical objects in Philippine-type languages, in the same way syntactic objects do in Indo-European languages.

The annotation in the Tagalog TRG treebank is closer to what we would consider ideal, as it also uses sub-labels to explicitly note the semantic role of the syntactic relations: for example, using *obj:agent* to denote an actor-object. It is only selectively applied, though: bare labels are used for actor voice transitives, implicitly considering this an unmarked voice.

Our other point of contention is the `Voice` feature values in Universal Dependencies. The three extra `Voice` values – *Ifoc*, *Lfoc*, and *Bfoc* – are useful, but the naming is confusing. Focus is a common term in Austronesian linguistics to refer to the morphosyntactic pivot of a clause, but in general linguistics it is also understood as a term in information structure, which is not its actual role (Himmelmann, 2002). For the sake of both universality and clarity, we suggest that it would be better to stick to the term Voice for the purpose of clarity/precision.

These problems are acknowledged by de Marneffe et al. (2021) in their outline of Universal Dependencies linguistic theory, but UD continues to reuse the active and passive labels and voice features for convenience. This is understandable, but it still privileges Indo-European terminology above that of a large proportion of the world's languages, which can have implications for usability of these treebanks in e.g., cross-linguistic typological studies.

4 Annotation

We took these issues into account when designing our annotation for Gorontalo. Our hope is to combine the best practices from the existing treebanks and to make it as clear as possible for users of Universal Dependencies.

Annotation was performed using Arborator (Guibon et al., 2020).

4.1 Gorontalo language

Gorontalo is spoken in Gorontalo Province, northern Sulawesi, Indonesia. It is a member of the Gorontalo-Mongondow branch of the Greater Central Philippine subgroup of Malayo-Polynesian (Usup, 1986). Its word order of *pivot verb non-pivot* is typical of Indonesian languages, but it has a Philippine-type voice system, with at least three attested voices.

Gorontalo has complex verbal morphology (mainly prefixes and infixes) to indicate voice, aspect, mood, intentionality, and more. There are gender-based case markers for proper nouns in all voices (Author et al., To appear), and for non-pivot core arguments in patient voice. The language also has oblique markers, clitics for first and second person, and some derivational morphology.

4.2 The data

At present, the sentences contained within the treebank are elicited sentences from sessions with a native Gorontalo speaker. These sentences were specifically selected to demonstrate the voice and alignment system of Gorontalo, and how this system would be parsed under Universal Dependencies.

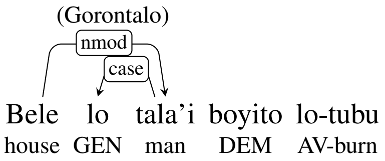
Future additions to the treebank will come from fieldwork data. The data consists of about nine hours of conversations between two or three Gorontalo speakers at a time, with 25 total speakers, in which one person asks the other person questions about their life, local knowledge (regarding topics such as fishing or farming), and traditional beliefs (such as taboos or communication with spirits).

4.3 Annotation decisions

4.3.1 Parts of speech and morphology

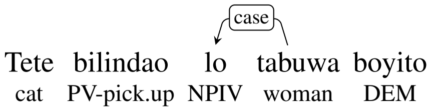
As a Philippine-type language, Gorontalo shares many features with Cebuano and Tagalog, and we have matched much of our morphological annotation with that of the Tagalog TRG corpus in particular.

Pivot and non-pivot markers are adpositions (*ADP*) annotated with the `Case` feature, with values *Nom* (nominative) and *Gen* (genitive) respectively. This follows the precedent in the Tagalog corpus. While the use of *Nom* is motivated by analogy with subjects, *Gen* is more unusual, and deserves some explanation here. In many Philippine-type languages, including Gorontalo, the non-pivot marker doubles as a genitive marker in genitive phrases. Figs. 1 and 2 show examples of this dual usage in Gorontalo.



"The house of the man burned down." (Actor voice)

Figure 1: In this example, the non-pivot marker *lo* is functioning as a genitive marker in a nominal modification.



"The woman picked up the cat." (Patient voice)

Figure 2: In this example, the same marker *lo* is instead functioning as a non-pivot marker.

In the case of a proper name, pivot/non-pivot markers may also encode `Gender` as *Masc* or *Fem*.

Pronouns may also be inflected for pivot or non-pivot status, expressed here with `Case` as *Nom* and *Gen*.

We introduce new values for the `Voice` feature on verbs. Gorontalo has actor-, patient- and instrument-voices[3]. The first of these is unproblematically aliased with *Act* (active), there being reasonable enough overlap between the two concepts. On the other hand, we replace *Pass* (passive) with *Pat* (patient): a value that would be new to UD and would make clear this is patient voice, not passive voice. Finally, we make a single-letter change to the already existing *Ifoc* (instrument focus) to make *Ivoc* (instrument voice), eliminating the ambiguity with information structural terms.

4.3.2 Syntax

As with parts of speech and morphology, we base our syntactic annotation of core arguments on precedent in Tagalog TRG, annotating the pivot NP as a subject and all others as variations of object. We also append sub-labels to specify types of argument. However, we do so for *all* arguments in transitive sentences, clarifying the semantic roles of both subjects and objects in each instance.

The sub-labels we use are also modified. We replace the sub-label *:pass* (passive) with *:patient*, once again making it clear that this is not to be confused with a passive construction. Where in TRG they replicate voice labels such as *:ifoc*, we spell out the word *instrument* for the instrument semantic role expressed by instrumental voice.

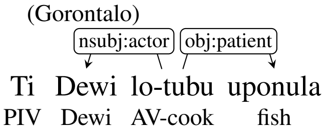
Our motivation for doing this is two-fold. Firstly, we make it easier for any user searching the treebank to query specific syntactic and semantic role combinations, such as *nsubj:instrument* or *obj:patient*. Secondly, we avoid privileging any one voice over another as the unmarked type.

---

[3] And possibly location voice, though this is less certain.

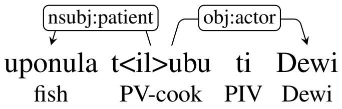
4.4 Annotation examples

We show some examples of annotation in Figs. 3 - 5.



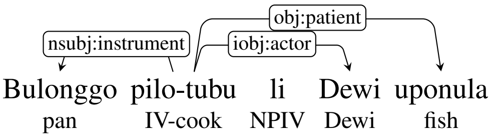
"Dewi cooked the fish." (Actor voice)

Figure 3: A sentence in actor voice in Gorontalo.



"Dewi cooked the fish." (Patient voice)

Figure 4: A sentence in patient voice in Gorontalo. Note infixing of the voice-inflecting morpheme within the verb *tubu* (cook).


obj:patient
nsubj:instrument
iobj:actor
Bulonggo pilo-tubu li Dewi uponula
pan IV-cook NPIV Dewi fish

"Dewi cooked the fish in the pan." (Instrument voice)

Figure 5: A sentence in instrument voice in Gorontalo.

5 Discussion

We have tried to be conservative with the annotations we make, keeping with the conventions of preceding treebanks and restricting the number of additions we need to make to the UD features and relations. However, we could go further in adding more features and labels specifically for Austronesian languages.

For example, we could add *Piv* (pivot) and *Npiv* (non-pivot) respectively as `Case` values for the pivot markers. This would remove the oddness of using *Gen* to describe the role of non-pivot core arguments, while also showing that non-pivots can perform multiple roles. Alternatively, we could keep the *Gen* value for clear instances of nominal modification as in Fig. 1.

6 Conclusion

We have presented a summary of the current state of symmetrical voice and alignment annotation for Austronesian languages, and our criticisms and suggestions for improvements.

In short, we have four main suggestions:

- Replace *Pass* with *Pat* in the `Voice` feature.
- Replace *_foc* with `_voc` in the `Voice` feature.
- Replace the dependency sub-label *pass* with *patient*.
- Obligatorily use semantic role sub-labels for subject and object arguments in transitive sentences.

We make available our preliminary work on a treebank for Gorontalo that demonstrates the changes we would like to see in annotation for Austronesian languages in general[4].

Our next steps are to continue to annotate more data for eventual release in Universal Dependencies. As more data is annotated, we will compare and contrast approaches to more Austronesian linguistic phenomena in UD.

We hope that these proposals will be discussed by the Universal Dependencies community and any contributors of new Austronesian languages, so that Austronesian languages can be included in a way that properly describes them and de-centres Indo-European terminology.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

This work was supported by the Fulbright Commission and National Endowment of the Humanities (USA).

All elicitations are from sessions with Novi Usu, a native speaker of Gorontalo, and a lecturer at Gorontalo State University. who is also collaborating on the treebank.

This collection was made possible by the cooperation of the Gorontalo community of Gorontalo Province, for whom this language is their cultural heritage.

---

[4] `https://github.com/andidyer/UTS_Gorontalo_Sanggala/tree/main`

Ethics

Texts were collected as part of a Fulbright-funded project and were transcribed by Gorontalo speakers studying at the local university. All transcribing assistants were paid per hour of transcription with their hourly rate based on teacher salaries in Gorontalo Province. All elicitations are from sessions with a native speaker of Gorontalo who is also collaborating on the treebank.

Limitations

The Gorontalo language is still being documented and described, and there remain some phenomena in the language that lack analysis, some of which appear in this treebank. As the language undergoes more description, some of the annotation will likely change.

References

Edith Aldridge. 2012. Antipassive and ergativity in tagalog. *Lingua*.

Ika Alfina, Indra Budi, and Heru Suhartanto. 2020. Tree rotations for dependency trees: Converting the head-directionality of noun phrases. *Journal of Computer Science*, 16(11):1585–1597.

Ika Alfina, Arlisa Yuliawati, Dipta Tanaya, Arawinda Dinakaramani, and Daniel Zeman. 2023. A Gold Standard Dataset for Javanese Tokenization, POS Tagging, Morphological Feature Tagging, and Dependency Parsing.

Avery D. Andrews. 2007. *The major functions of the noun phrase*, page 132–223. Cambridge University Press.

Glyd Jun Arañes. 2022. The gja cebuano treebank: Creating a cebuano universal dependencies treebank. Master's thesis, University of Eastern Finland.

Robert Blust. 2013. *The Austronesian languages*. Asia-Pacific Linguistics, School of Culture, History and Language, College of Asia and the Pacific, The Australian National University.

Victoria Chen. 2025. The syntax of philippine-type alignment: Insights from case-marking. *Nat Lang Linguist Theory*.

Victoria Chen and Bradley Mcdonnell. 2019. Western Austronesian Voice. *Annual Review of Linguistics*.

William Croft. 2003. *Typology and Universals*. Cambridge University Press.

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

William Foley. 2008. The place of philippine languages in a typology of voice systems. *Voice and grammatical relations in Austronesian languages*.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5291–5300, Marseille, France. European Language Resources Association.

Martin Haspelmath and Andrea Sims. 2010. *Understanding Morphology (2nd ed.)*. Routledge.

Nikolaus P. Himmelmann. 2002. Voice in western austronesian: an update. *The history and typology of western Austronesian voice systems*.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Stephanie Dawn Samson. 2018. A treebank prototype of tagalog. Master's thesis, University of Tübingen.

Hunggu Tadjuddin Usup. 1986. *Rekonstruksi Proto-Bahasa Gorontalo-Mongondow [Proto-Gorontalo-Mongondow Language Reconstruction]*. Ph.D. thesis, Universitas Indonesia.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Arofat Akhundjanova, Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Matthew Andrews, and 633 others. 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.

Extending Enhanced Universal Dependencies – addressing subjects in pro-drop languages

Magali S. Duran[1], **Elvis A. de Souza**[1],
Maria das Graças V. Nunes[1], **Adriana S. Pagano**[2], **Thiago A. S. Pardo**[1]
[1]Núcleo Interinstitucional de Linguística Computacional, Universidade de São Paulo – Brazil
[2]Faculdade de Letras, Universidade Federal de Minas Gerais – Brazil
`magali.duran@gmail.com, elvis.desouza99@gmail.com,`
`gracan@icmc.usp.br, apagano@ufmg.br, taspardo@icmc.usp.br`

Abstract

Enhanced Universal Dependencies (EUD) serve as a crucial link between syntax and semantics. Beyond basic syntactic dependencies, EUD provide valuable refined logical connections for downstream tasks such as semantic role labeling, coreference resolution, information extraction, and question answering. While the original EUD framework defines six types of relations, this paper introduces an extension designed to address subject propagation in pro-drop languages. This "Extended EUD" proposal increases the number of dependency relations that may be annotated in sentences, improving linguistic representation. Additionally, we report our experiments on a corpus of Portuguese (a pro-drop language), which we make publicly available to the research community.

1 Introduction

Syntax-based approaches for multilingual Natural Language Processing (NLP) have evolved in the last decade mostly due to the growing number of languages that have gold standard treebanks annotated under the Universal Dependencies framework (De Marneffe et al., 2021; Nivre et al., 2016). Currently, there are 296 corpora available on the UD website, but only 42 of them[1] provide Enhanced Universal Dependencies (EUD) in the ninth column of the well-known CoNLL-U format, called *deps*. EUD constitutes a "bridge" between syntax and semantics. In column "deps", basic dependencies are transformed into a more semantic-like representation which associates each token to every head token they modify, regardless of whether this relation is explicit or implicit. For this reason, each token in EUD may have multiple heads. Besides that, EUD includes case markers (dependents of *cc*, *case* and *mark* relations) in the relation name of their heads (*nmod*, *obl*, *acl*, *advcl*), providing helpful clues to their respective semantic roles.

By making explicit logical relations that were previously implicit, EUD increases the number of responses that can be obtained from an annotated corpus and paves the way for downstream applications that need these responses, such as semantic role labeling, coreference resolution, information extraction and question answering.

EUD allows for six types of "enhancements", including the propagation of subjects shared by coordinate clauses (*conj* clauses) and the assignment of "external subjects" for clauses with null subject (*xcomp* clauses). Once a clause with implicit subject receives an enhanced subject, it can share it with clauses dependent on it, creating a propagation chain. This recursive nature of EUD shows that one EUD relation may rely on the product of another EUD relation, and not only on basic syntactic dependencies. As an illustration, Figure 1 shows a sentence with EUD annotation where a *nsubj* is first propagated to a *conj* dependent clause and then to a *xcomp* dependent clause (in red edges).

Considering this recursive property of subject propagation within a sentence, EUD could further leverage syntax, also including propagation of subject for *acl*, *acl:relcl*, *advcl* and *ccomp*. This would be particularly relevant for so-called pro-drop languages, which present high occurrence of subject ellipsis since a verb form encodes Person, Number and sometimes Gender features that indicate who the subject is. If there is at least one explicit subject within a series of clauses in a sentence, such subject can be propagated to the other clauses with implicit subject. As it is possible to have different sequences of clauses in a sentence, the subject propagation path must be clear so as not to interrupt propagation. If, for example, a *ccomp* has an *xcomp* or *conj* as dependents, these clauses will only receive subject propagation once *ccomp* has received subject propagation before them. This

---

[1] According to UD v2.15 homepage: `https://universaldependencies.org`

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 143–152
August 27, 2025 ©2025 Association for Computational Linguistics

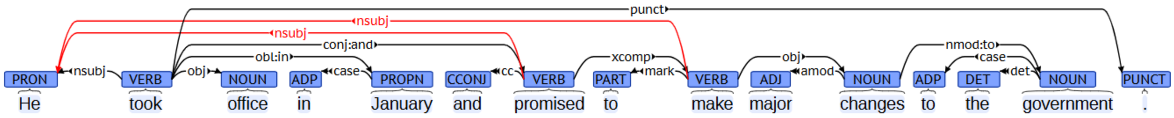


Figure 1: Example of EUD annotation with recursive propagation of subjects for the sentence "He took office in January and promised to make major changes to the government."

may be observed if we compare Figures 2 and 3. Whereas Figure 2 follows the current EUD guidelines and has only one subject, Figure 3, following the proposed extended EUD, has four subjects, as *ccomp* needs to receive subject propagation from its head before propagating it to the *xcomp* and *conj* dependent clauses.

It is important to note that ellipsis of the subject is not exclusive to pro-drop languages and other languages could take advantage of this extension in EUD. In English, for example, it is not uncommon to find an *advcl* with subject ellipsis (in bold in the following examples):

> Following his discharge, he taught English at a Buddhist girls' school **while also taking classes at Kyoto University**.[2]
>
> The meal was extremely overpriced and lacked flavor, **especially for being a special NYE menu**.[3]

However, when there is no explicit subject in a sentence, there is no possible propagation. In these cases, one might consider inserting an empty token to represent the elided subject, but this is a separate task, not related to subject propagation. Moreover, EUD emphasizes that empty tokens are only allowed for elided predicate insertion (predicates that are suppressed in the dependent clause but may be recovered from the matrix clause).

We took on the task of implementing EUD in Portuguese and exploring the full propagation of explicit subjects for dependent clauses that have elided subjects, writing rules based on features typical of verb forms in pro-drop languages. The strategy for doing this was to customize a multilingual rule-based EUD system, pre-annotate the corpus and manually revise the results, making changes to the rules, adding new rules, or fixing the UD gold standard annotation when necessary. This paper reports the finalized enhanced dependency annotation for the Porttinari-base corpus (Duran et al., 2023). A detailed account of the annotation process is not included here, as it has already been partially described by de Souza et al. (2024). More emphasis will be placed on describing the annotation process of the proposed EUD extension. The contributions of this paper are:

- to report the experience of extending subject propagation for other dependent clauses besides those defined in EUD, which we call "*Extended* Enhanced Universal Dependencies" (EEUD), and
- to release a Portuguese corpus annotated with EUD and EEUD (as far as we know, the first one for this language)[4].

The remainder of this paper is organized as follows: Section 2 briefly presents the related work and Section 3 reports the methodology we adopted; the results are reported in Section 4 while Section 5 explores the relation between basic dependencies and EUD; some limitations and final remarks are presented in Sections 6 and 7, respectively.

2 Related work

One of the precursors of UD, the Stanford Dependency model, already included an output containing inferred relations (De Marneffe et al., 2006; De Marneffe and Manning, 2008), intended to provide a semantically richer representation of syntactic relations, making them more readily employable by other applications. The idea was to replicate modifiers for each token they modify and to replicate case markers for each token they introduce. The term "enhanced" first appeared in de Marneffe et al. (2014) referring to a series of transformations: marking external subjects and the external role in relative clauses, renaming dependencies to include

---

[2] Example retrieved from UD English GUM 2.16.

[3] Example retrieved from UD English EWT 2.16.

[4] Both versions of the annotated corpus are available: the one with EUD in UD homepage, and the one with the full subject propagation (EEUD) in our website.

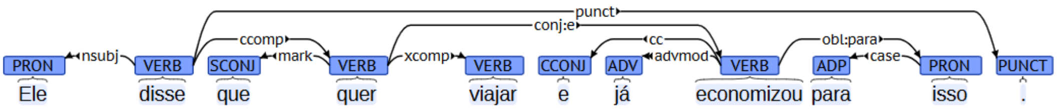


Figure 2: EUD annotation not extended to ccomp in “Ele disse que quer viajar e já economizou para isso.” (lit.: He said that wants to travel and has already saved up for it.)

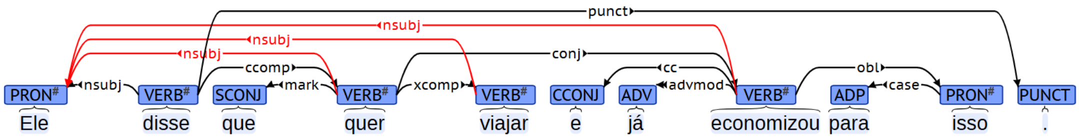


Figure 3: EUD annotation extended to *ccomp* in “Ele disse que quer viajar e já economizou para isso.” (lit.: He said that wants to travel and has already saved up for it.)

case markers, and propagating relations over conjunctions.

Recognizing the relevance of such ideas, Schuster and Manning (2016) developed an English converter to produce an enhanced output according to UD restrictions, that is, expanding basic dependencies without changing them.

Nivre et al. (2018) evaluated two cross-lingual techniques to automatically annotate EUD in UD treebanks: a rule-based English system and a data-driven Finnish system. They used Swedish and Italian corpora to test both systems. The rule-based system performed better at assigning subjects, and the data-driven system performed better at propagating coordinate dependents. Overall, both systems contributed to bootstrap the EUD annotation.

The proposal to develop non-language-dependent systems to annotate EUD inspired two shared tasks held at IWPT: one in 2020 (Bouma et al., 2020) and another in 2021 (Bouma et al., 2021). However, their multilingual dataset did not include Portuguese. EUDs were instantiated in Portuguese by Pagano et al. (2023) and, later, de Souza et al. (2024) developed a Portuguese customization: the rule-based system UDtoEUD[5] (Guillaume and Perrier, 2021), a Graph Rewriting System that uses GREW (Bonfante et al., 2018) to convert basic dependencies into EUD. Droganova and Zeman (2019) clarify that EUD annotation is optional in treebanks, and it is allowed to annotate only one, several or all of the six types of EUD. The UD framework, until now, does not allow the annotation of other types of EUD out of the six types already described in the guidelines[6].

3 Methodology

Revising EUD annotation is an extremely heavy and time-consuming task, which requires even more qualified human resources than those needed to revise basic dependencies. To bootstrap EUD annotation, we customized the UDtoEUD converter to Portuguese (de Souza et al., 2024), after a in-depth linguistic study on enhanced dependencies for the Portuguese language (Duran, 2024).

We chose UDtoEUD because it has all the right qualities for our task: 1) it is based on rules, which enabled us to improve the system to increase the quality of the whole annotation without relying on previously annotated data; 2) it scored above 98% over gold basic UD data for Italian, which is the language closest to Portuguese among those that took part in the shared task; and 3) it allows entering language-specific lexicon: control and raise verbs (to determine the subject of *xcomp*), and adverbs (to determine which adverbs left to the head of *conj* should propagate to the *conj* dependent).

By customizing UDtoEUD, we automatically annotated the six EUD types in Porttinari-base, which has 8,418 sentences. As the converter reached 96.05% ELAS in a gold standard EUD composed of sentences assembled from Porttinari-base, it was not deemed necessary to perform human-revision on a one-by-one basis in this phase.

In the next phase, we improved the converter to include propagation of subjects to other clauses: *ccomp*, *advcl*, *acl* and *acl:relcl*. To find candidates for propagation, we used the following rules:

- the dependent clause should not be head of *nsubj* or *nsubj:pass* or *csubj*, that is, if the

---

[5] `https://gitlab.inria.fr/grew/udtoeud`

[6] `https://universaldependencies.org/u/overview/enhanced-syntax.html`

clause has a subject, the place for the subject is already filled;

- the dependent clause should not be head of *expl:impers*, that is, if there is a mark of impersonalization, there is no place for a subject;
- the dependent clause should not be an impersonal verb (“haver”, “chover”, “anoitecer”, etc.), that is, if a verb is impersonal, it does not have a place for a subject.

Besides that, for *ccomp* and *advcl*, their head should be head of *nsubj* or *nsubj:pass*, i.e., they should have a subject to propagate, regardless of whether it was an explicit subject or a subject resulting from a previous propagation. In other words, if the head does not have a subject to propagate, no propagation is possible. For *acl:relcl* and some types of *acl*, the head of the head should have a subject to propagate.

Another restriction to propagation was that the dependent clause (or its *aux* or *aux:pass* or *cop*) should have the same values for the features Person and Number of the clause (or its *aux* or *aux:pass* or *cop*) from which it would inherit the subject.

As this is a new task, all the cases of propagation (2,147 sentences) and non-propagation (759 sentences) were verified by linguists that are experts in UD, leading to the improvement of the system. In this phase, we detected some non-recurrent errors in basic dependencies, which were promptly corrected. Since we used a strategy of converting basic dependency trees into EUD, EUD automatic annotation only works if the basic dependencies are correctly annotated; hence they are a very good source for checking the basic dependencies logic.

4 Results

In this section, we discuss separately the implementation of EUD and EEUD, but present the increase in relations achieved after each of these phases.

4.1 Phase 1 - EUD identification

As already mentioned, since the UDtoEUD customization for Portuguese achieved an overall ELAS of 96.05% in a gold dataset developed for the task (de Souza et al., 2024), no human-revision was deemed necessary for most of the automatically annotated EUD in Porttinari-base. As expected, the biggest challenge in EUD was elided predicates. This EUD inserts an empty token whenever a token, dependent on a *conj* relation, is head of an *orphan* relation. The relation *orphan* is relatively rare in our corpus (67 occurrences). Inserting an empty token whenever an *orphan* occurs is a trivial task, but deciding where this empty token shall be inserted and naming the relations it establishes with the ex-head and the ex-dependent of the *orphan* relation is not easy for a rule-based system. So we decided to manually revise all the sentences containing *orphan* relations in the corpus, which required a lot of manual editing until the respective sentences reached their final version. For this reason, we excluded the *orphan* relation from the statistics reported in this paper. This issue will be addressed in depth in a forthcoming study. Table 1 shows the ELAS of the other five EUD types before and after customizing UDtoEUD for Portuguese.

Type of EUD	Original	Customized
1 - case assignment	95.17%	98.90%
2 - *xcomp* subjects	92.54%	97.06%
3 - prop. *conj* head	84.13%	96.43%
4 - prop. *conj* dependents	96.67%	96.67%
5 - relative pronoun *ref*	94.23%	94.23%

Table 1: ELAS of UDtoEUD per EUD type before and after customization. Source: de Souza et al. (2024)

Out of these five EUD, the only one that required an amendment in the customized UDtoEUD was the assignment of an external subject to *xcomp*. We found cases of *xcomp* subject corresponding to the *obj*, *obl* or *iobj* of the matrix clause, and sometimes such elements are elided in Portuguese, leading to an *xcomp* whose external subject cannot be determined. To avoid the wrong assignment of the subject of the matrix clause to the *xcomp*, we had to improve the rules, informing a list of verbs that allow elided complements, for example: “permitir” (*to allow*), “deixar” (in the sense of *to allow*) and “mandar” (*to order*). The following example shows an elided *obj*:

> Os leilões permitem [Ø] pagar o animal em até 50 parcelas. (lit.: “Auctions allow [Ø] to pay for the animal in up to 50 installments” meaning “allow anybody to pay”)

4.2 Phase 2 - EEUD identification

Departing from the basic rules cited before, we generated CoNLL-U files containing, for each of the four relations (*acl*, *acl:relcl*, *advcl* and *ccomp*), cases of subject propagation and cases of non-propagation. These files were revised by three expert annotators, who indicated the cases in which propagation succeeded, the cases in which propagation failed (the subject should not be propagated or the subject propagated was not the correct one), the cases in which non-propagation was correct and the cases in which non-propagation was incorrect (the subject should have been propagated). Table 2 shows the result of this revision: 87.98% of the propagated subjects were considered correct and 12.02% were considered incorrect; 73.88% of the non-propagated subjects were considered correct and 26.12% were considered incorrect. The high percentage of subjects which should have been but were not propagated (26.12%) is due to the fact that our rules were initially designed to deal more with verbal predicates than with nominal predicates.

Dealing with each of the four relations separately made it easier to analyze the cause of the errors and come up with solutions to improve the rules. In what follows, we present an example of subject propagation for each focused relation and comment on errors that we found.

acl: The *acl* dependent rarely contains a subject and rarely presents a finite form. Most of the time it is a passive voice construction without an *aux:pass* or an active construction without an *aux*. Its subject is the nominal that is its head in basic dependencies.[7] In the sentence below, the propagated subject is in bold and the *acl* dependent is underlined.

> O último **discurso** gravado de Al-Baghdadi é de novembro de 2016. (lit.: The last **speech** recorded of Al-Baghdadi is from November 2016.)

Two problems were found in the propagation of *acl*: one relating to the annotation of the feature Voice=Pass, and another relating to the propagation of coordinate nominals as subject.

Regarding the first problem, we found inconsistencies in the annotation of past participles, as some of them should have the feature Voice=Pass and did not, and others should not have it and did, thus affecting the type of subject propagated – *nsubj* or *nsubj:pass*. In Portuguese, sometimes a past participle form is not a reduced form of a passive construction. This is the case with intransitive verbs, as shown in the following example, which clearly does not constitute a passive voice sentence. These problems required a complex revision of the Voice=Pass feature in the corpus.

> Outra coisa, no entanto, é a avalanche de 1,5 milhão de **refugiados** chegados em 2015 e 2016. (lit.: Another thing, however, is the avalanche of 1.5 million **refugees** arrived in 2015 and 2016)

The second problem is related to the heuristics used to propagate subjects: when a dependent clause has a subject, it is not a candidate for propagation. This was a problem when two or more coordinate tokens are the subject, causing the system to propagate only the last one. It was solved for all propagation rules, not only for *acl*. A limitation of coordinate tokens as subject, as Schuster and Manning (2016) points out, is that it is not always clear whether the intention is a distributive or collective interpretation. Although we are aware of this, we have chosen to ignore it for the time being, and adopted a distributive interpretation.

acl:relcl: In Portuguese, the *acl:relcl* dependent almost always has an internal subject. When it does not, and the *acl:relcl* clause does not contain a relative pronoun, its subject is in most cases a nominative relative pronoun that is, at the same time, the head of the *acl:relcl* itself, as seen in the next example.

> Quem ouviu julgou que o treinador estava perdido. (lit.: **Who** heard it thought the coach was lost)

In other cases, the subject of *acl:relcl* is the head of the head of *acl:relcl*:

> **Ele** e **Saud** pediram desculpas publicamente pelo que disseram. (lit.: **He** and **Saud** have publicly apologized for what said)

When Person and Number features of the copula verb and the *acl:relcl* dependent present different values, propagation is not performed (Figure 4).

advcl: The *advcl* dependent presents several patterns, depending on its semantic function. We observed good results with temporal and causal *advcl*, but comparative and conformative *advcl* rarely received a propagation. Finally, *advcl* behaves as *xcomp* if its head has an *obj* dependent and such

---

[7] This circularity is a situation that also occurs when the head of an *acl:relcl* substitutes a relative pronoun as subject: "the man who died" turns into "the man died" when the *ref* relation is annotated in EUD.

	PROPAGATED		NOT PROPAGATED			
	correct (TP)	incorrect (FP)	correct (TN)	incorrect (FN)	Total	Accuracy
acl	896	57	29	7	989	93.53%
acl:relcl	221	9	134	59	423	83.92%
advcl	621	151	228	119	1,119	75.87%
ccomp	172	44	186	19	421	85.04%
Total	1,910	261	577	204	2,952	84.25%
(%)	87.98%	12.02%	73.88%	26.12%		

Table 2: Results of EEUD revision.

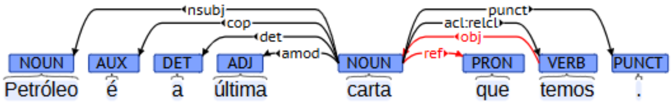


Figure 4: An example in which the subject did not propagate because Person and Number did not match (the AUX “é” is third person singular and the VERB “temos” is first person plural): “Petróleo é a última carta que temos” (lit.: Oil is the last card that have)

obj is the subject to be propagated to the *advcl*. Although we do not have any subrelation for *advcl* at the moment, this could be very productive for EUD purposes. A particularly noteworthy case is that of an *advcl* that has its own subject, while its head does not. Therefore, the subject may be propagated from the dependent to the head, as Figure 5 shows.

ccomp: The *ccomp* dependent is almost always a finite construction (the verb or its auxiliary is a finite form). We analyzed those *ccomp* related to reported speech (which we annotated as *ccomp:speech*) and concluded that subject propagation was not feasible or suitable: almost all present a different Person and Number in the reported speech and the head of *ccomp:speech*.

Some infinitive constructions dependent on *ccomp* seem to be impersonal, and the propagation rules lead to errors, as in the next sentence, where “revitalização” is not the subject of “devolver”.

> A revitalização *implica* em devolver a esses centros a vitalidade. (lit.: Revitalization **implies** to restore vitality to these centers.

Table 3 shows that 146 (31.4%) of the 465 errors found were solved by fine-tuning the rules and 73 (15.7%) were solved by corrections in the basic dependencies annotation. The remaining 246 errors (52.9% of all errors) were solved by manually editing EUD. This does not necessarily mean that there are no patterns in the remaining errors; rather, we were unable to identify them.

	rules	basic	other
acl	11	32	21
acl:relcl	3	5	60
advcl	112	21	137
ccomp	20	15	28
Total	146	73	246
(%)	31.4%	15.7%	52.9%

Table 3: Sources of errors.

We used the resulting gold standard to compute accuracy and precision of the rules before and after fine-tuning (Table 4).

Table 5 specifies the number of subjects propagated by dependency relation (deprel), both in EUD and in EEUD. It is important to note that the application order of the subsets of rules concerning subject propagation is relatively free; however, they should be applied recursively until no additional subject propagation is obtained. In some cases, this means repeating these rules up to four or five times, and there is room to propose ways to optimize this repetition.

Table 6 shows the number of each type of relation in the corpus, in 3 columns: the first one shows basic dependencies, the second one shows EUD and the third one includes EEUD. Sentences with an *orphan* relation were not taken into account for this table due to the difficulty to align the null nodes inserted throughout the three versions

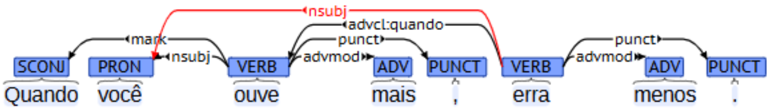


Figure 5: An example in which the subject propagated from the dependent to the head: “Quando você ouve mais, erra menos” (lit.: When you listen more, [you] make fewer mistakes)

	Total	Accuracy (Before)	F1-Score (Before)	Accuracy (After)	F1-Score (After)
acl	858	0.937063	0.966871	0.967366	0.983071
acl:relcl	349	0.787966	0.787356	0.828080	0.831461
advcl	1,877	0.798082	0.797435	0.928077	0.935002
ccomp	441	0.888889	0.887872	0.929705	0.931567

Table 4: Performance of rules before and after fine-tuning, calculated over the gold standard.

	Basic	EUD		EEUD	
subjects					
acl	135	139	(+2.96%)	1,101	(+692.09%)
acl:relcl	1,570	1,621	(+3.25%)	1,792	(+10.55%)
advcl	514	522	(+1.56%)	1,556	(+198.08%)
ccomp	690	706	(+2.32%)	919	(+30.17%)
conj	435	1,278	(+193.79%)	1,385	(+8.37%)
xcomp	-	1,712	-	1,869	(+9.17%)

Table 5: Subjects propagated by deprel in EUD and EEUD.

of the annotation. As may be observed, EEUD contributed to increasing 15.39% the number of *nsubj* and 81.61% of *nsubj:pass*

5 EUD as feedback on basic dependencies annotation

When using a conversion strategy, EUD and EEUD strongly rely on the annotation in other CoNLL-U columns, mainly on Head (dependency head) and Deprel (dependency relation), but also on Upos (universal part-of-speech) and Feat (column that encodes morphological features).

In Portuguese, several verbs have ambiguous forms in the first and third person, mainly in the past imperfect tense. As this feature is used to propagate subjects, annotation errors can lead to EUD errors. When we noticed this type of error, we looked for all the verb forms that allowed more than one value for the Person feature and revised their annotation.

As already mentioned, we also revised the Voice feature of past participles without a dependent *aux:pass*, as they may require Voice=Pass or not. When two predicates have different Voice values, the respective relation of the propagated subject needs to be adjusted from *nsubj* to *aux:pass* (or the inverse), as shown in Figure 6.

6 Limitations

The rules we implement rely heavily on comparing the Person and Number features of the head and the dependent: if they are the same, there is propagation; otherwise, there is not. However, even when these features have the same values, propagation may not work. This problem occurs mainly with Person=3, as Person=1 and Person=2 are normally not ambiguous within the discourse situation.

Two examples illustrate the above problem. In the first example (Figure 7), “o veículo” (the vehicle) is clearly not the subject of “acordou” (woke up), even though the propagation is licensed because the Person and Number rules have been met. In the second example (Figure 8), only world knowledge allows us to detect the error, as the subject of “preferia” (would prefer) is the obl “ex-bailarina” and not “chefe” (boss), which is *nsubj* of the head.

deprel	Basic	EUD		EEUD	
acl	1,635	1,748	(+6.91%)	1,748	-
acl:relcl	1,899	2,039	(+7.37%)	2,039	-
advcl	2,311	2,522	(+9.13%)	2,522	-
advmod	6,107	6,051	(-0.92%)	6,051	-
amod	6,595	6,833	(+3.61%)	6,833	-
appos	1,036	1,153	(+11.29%)	1,153	-
aux	806	831	(+3.10%)	831	-
aux:pass	983	1,008	(+2.54%)	1,008	-
case	22,359	23,196	(+3.74%)	23,196	-
cc	4,182	4,183	(+0.02%)	4,183	-
ccomp	1,102	1,216	(+10.34%)	1,216	-
ccomp:speech	669	811	(+21.23%)	811	-
conj	4,518	4,518	-	4,518	-
cop	2,871	3,043	(+5.99%)	3,043	-
csubj	356	411	(+15.45%)	411	-
csubj:outer	4	4	-	4	-
csubj:pass	1	1	-	1	-
det	23,897	23,881	(-0.07%)	23,881	-
discourse	254	277	(+9.06%)	277	-
dislocated	77	85	(+10.39%)	85	-
expl	562	562	-	562	-
expl:impers	154	154	-	154	-
fixed	1,339	1,339	-	1,339	-
flat	87	87	-	87	-
flat:foreign	68	68	-	68	-
flat:name	3,331	3,332	(+0.03%)	3,332	-
iobj	95	95	-	95	-
list	27	27	-	27	-
mark	4,301	4,520	(+5.09%)	4520	-
nmod	12,534	13,549	(+8.10%)	13,549	-
nsubj	9,480	12,117	(+27.82%)	13,982	(+15.39%)
nsubj:outer	97	97	-	97	-
nsubj:pass	709	968	(+36.53%)	1,758	(+81.61%)
nummod	1,759	1,784	(+1.42%)	1,784	-
obj	7,200	7,650	(+6.25%)	7,650	-
obl	9,058	9,743	(+7.56%)	9,743	-
obl:agent	498	547	(+9.84%)	547	-
parataxis	877	966	(+10.15%)	966	-
punct	22,094	22,094	-	22,094	-
ref	0	1,814	-	1,814	-
reparandum	5	5	-	5	-
root	8,354	8,354	-	8,354	-
vocative	26	29	(+11.54%)	29	-
xcomp	2,330	2,479	(+6.39%)	2,479	-

Table 6: Relations increase, by deprel, due to EUD and EEUD. There are two relations that saw a decrease from Basic to EUD: *advmod* and *det*. The former is due to the relative adverb “onde”, which is labeled as *advmod* dependent on the verb of the relative clause in the basic tree, but is labeled as *ref* dependent on the nominal that is modified by the relative clause in the EUD graph. The latter is due to the relative pronoun “cujo”, which is labeled as *det* dependent on a nominal to its right in the basic tree, but is labeled as *ref* dependent on the nominal to its left in the EUD graph.

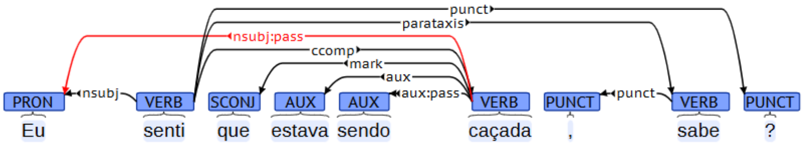

punct
parataxis
nsubj:pass
ccomp
mark
aux
aux:pass
nsubj
punct
PRON VERB SCONJ AUX AUX VERB PUNCT VERB PUNCT
Eu senti que estava sendo caçada , sabe ?

Figure 6: An example of *nsubj* propagated as *nsubj:pass*: “Eu senti que estava sendo caçada, sabe?” (lit.: I felt that was being hunted, you know?

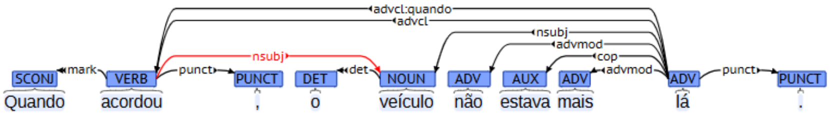

advcl:quando
advcl
nsubj
advmod
cop
advmod
nsubj
mark
punct
det
punct
SCONJ VERB PUNCT DET NOUN ADV AUX ADV ADV PUNCT
Quando acordou , o veículo não estava mais lá .

Figure 7: Error analysis – the subject of the head does not fill the *advcl* subject: “Quando acordou, o veículo não estava mais lá” (lit.: When woke up, the vehicle was not there anymore)

7 Final remarks

We produced two versions of the corpus: one with EUD as described in the UD guidelines (corpus available on the UD website) and another with all enhanced relations, including those extensions we implemented, i.e., EEUD[8].

Working with a rule-based system to annotate enhanced dependencies has given us a deeper understanding of our language. At the same time, when we checked the automatically assigned enhancements, we noticed some inconsistencies in the annotation of the basic dependencies, which were corrected, improving the corpus quality as a whole.

EEUDs are dependency relations more difficult to annotate automatically, as the number of errors produced by the rules was greater than the number of errors in the original EUD, and required many manual corrections. Even so, the task proved to be productive, as we managed to increase the number of propagated subjects by 56%.

Drawing on our two versions of the golden standard corpus (one annotated with EUD and another with EUD + EEUD), as future work, we intend to train automatic EUD and EEUD parsers for Portuguese. We also plan to take a closer look at the enhancement of the *orphan* dependency relation. For this, we intend to produce data augmentation of *orphan* cases, since the sparsity of the data, added to the diversity of patterns, will probably prevent automatic approaches from achieving good performance in EUD annotation.

The rules of the customized version of UDtoEUD are available at Github[9]. We believe the rules we designed to deal with elided subjects may be useful for other languages, especially for the so-called pro-drop languages. Extending EUD to allow any kind of subject propagation would partially solve the inequality in the number of subjects between pro-drop and non-pro-drop languages, remaining cases where there is no subject to propagate. The gold standard corpus may also be useful for those wishing to experiment with a non-rule-based approach.

Acknowledgments

This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI - http://c4ai.inova.usp.br/), with support from São Paulo Research Foundation (FAPESP grant #2019/07665-4) and IBM Corporation. The project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law n. 8.248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44. Adriana Pagano thanks the National Council for Scientific and Technological Development (CNPq 404722/2024-5; 313103/2021-6) and Minas Gerais State Agency for Research and Development (FAPEMIG).

---

[8] `https://sites.google.com/icmc.usp.br/poetisa/resources-and-tools`

[9] `https://github.com/alvelvis/eud-portugues/blob/main/flask/conjunto_regras_porttinari.grs`. The EUD rules can be found as the “eud_portuguese” strategy, while the EEUD rules can be found as the “eud_portuguese_extended” strategy.

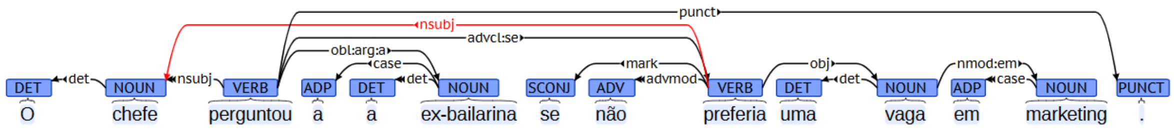


Figure 8: Error analysis – the logical subject of the *ccomp* should be the *obl* of its head (ex-bailarina): "O chefe perguntou à ex-bailarina se não preferia uma vaga em marketing" (lit.: The boss asked the former dancer if wouldn't prefer a job in marketing)

References

Guillaume Bonfante, Bruno Guillaume, and Guy Perrier. 2018. *Application of Graph Rewriting to Natural Language Processing*, volume 1 of *Logic, Linguistics and Computer Science Set*. ISTE Wiley.

Gosse Bouma, Djamé Seddah, and Daniel Zeman. 2020. Overview of the IWPT 2020 shared task on parsing into enhanced Universal Dependencies. In *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, pages 151–161, Online. Association for Computational Linguistics.

Gosse Bouma, Djamé Seddah, and Daniel Zeman. 2021. From raw text to enhanced Universal Dependencies: The parsing shared task at IWPT 2021. 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 146–157, Online. Association for Computational Linguistics.

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, Bill MacCartney, Christopher D Manning, and 1 others. 2006. Generating typed dependency parses from phrase structure parses. In *Lrec*, volume 6, pages 449–454.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8.

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

Elvis A de Souza, Magali S Duran, V Nunes Maria das Graças, Gustavo Sampaio, Giovanna Belasco, and Thiago Pardo. 2024. Automatic annotation of enhanced universal dependencies for brazilian portuguese. In *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 217–226. SBC.

Kira Droganova and Daniel Zeman. 2019. Towards deep universal dependencies. In *Proceedings of the fifth international conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 144–152.

Magali S Duran. 2024. Guidelines for annotating enhanced syntactic dependency relations in portuguese, following the guidelines of the Universal Dependencies (UD) approach (in Portuguese). Technical Report 448, ICMC-USP.

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

Bruno Guillaume and Guy Perrier. 2021. Graph rewriting for enhanced universal dependencies. In *IWPT 2021-17th International Conference on Parsing Technologies*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, and 1 others. 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, Paola Marongiu, Filip Ginter, Jenna Kanerva, Simonetta Montemagni, Sebastian Schuster, and Maria Simi. 2018. Enhancing universal dependency treebanks: A case study. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 102–107.

Adriana Pagano, Magali S Duran, and Thiago Pardo. 2023. Enhanced dependencies para o português brasileiro. In *Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival*, pages 461–470.

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.

Annotating Second Language in Universal Dependencies: a Review of Current Practices and Directions for Harmonized Guidelines

Arianna Masciolini, Aleksandrs Berdicevskis, Maria Irena Szawerna, Elena Volodina
Språkbanken Text, SFS, University of Gothenburg, Sweden
`{arianna.masciolini,aleksandrs.berdicevskis,maria.szawerna,elena.volodina}@gu.se`

Abstract

Universal Dependencies (UD) is gaining popularity as an annotation standard for second language (L2) material. Grammatical errors and other interlanguage phenomena, however, pose significant challenges that official guidelines only address in part. In this paper, we give an overview of current annotation practices and provide some suggestions for harmonizing guidelines for learner corpora.

1 Introduction

Ever since Lee et al. (2017b) proposed to represent learner corpora as parallel dependency treebanks, Universal Dependencies (UD) (de Marneffe et al., 2021) has been gaining popularity as an annotation standard for both written and spoken second language (L2) data. At the time of writing, treebanks have already been released for written Chinese (Lee et al., 2017a), English (Berzak et al., 2016, now retired[1]), Italian (Di Nuovo et al., 2019, 2022) and Korean (Sung and Shin, 2024, 2025), as well as spoken English (Kyle et al., 2022). Two more are in progress, one for written Russian (Rozovskaya, 2024) and one for written Swedish.

When it comes to the annotation of L2 productions, the main advantage of using UD is that it provides a cross-lingually consistent morphosyntactic annotation layer that enables both quantitative and qualitative comparisons between a learner's L1 and L2, between different L2s and, most importantly, between the Target Language (TL) in its standard form and as an L2. In the latter scenario, an especially helpful format is that of parallel, so-called "L1-L2" treebanks, proposed by Lee et al. (2017b) and adopted by the vast majority of the seven aforementioned annotation efforts. In an L1-L2 treebank, each learner sentence is paired with a *correction hypothesis*, i.e. a corrected version of the learner's production based on an expert's interpretation of its intended meaning (if the sentence already adheres to the standard for the TL, the correction hypothesis is identical to the original sentence). As demonstrated by Masciolini (2023), this makes it possible to retrieve and analyze grammatical errors and other divergences between original learner sentences and their corrections via tree queries, without relying on any explicit error labeling scheme.

UD annotation of L2 productions, however, comes with its challenges. Learner texts often deviate from standard language in ways that make them difficult to analyze in a framework designed with standard use of the TL in mind. Relevant phenomena are not limited to grammatical errors themselves, but also include code switching, calques and non-idiomatic expressions. As a result, treebank developers have produced extensive project-specific guidelines, not always consistent with each other.

This study has been carried out as part of an ongoing annotation effort whose aim is to produce a treebank based on the SweLL-gold (Swedish Learner Language) corpus (Volodina et al., 2019), expanding the resources available through the Språkbanken research infrastructure. To ensure that the UD annotation of SweLL is consistent with that of other similar datasets, convenient for treebank users and theoretically motivated, we have reviewed the relevant UD guidelines and compared practices across five L2 treebanks for as many TLs. In this paper, we present our findings and propose directions for harmonizing L2 annotation guidelines. We demonstrate the application of our suggested principles on selected examples from our yet-to-be-released treebank.

2 Challenges of Treebanking L2

As mentioned in the introduction, a number of phenomena typical of — but not exclusive to — learner

---

[1] "Retired" UD treebanks are available for download but not up-to-date with current annotation guidelines and therefore not part of the latest UD release.

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 153–163
August 27, 2025 ©2025 Association for Computational Linguistics

language pose significant challenges for treebank developers. Most prominently, L2 productions are often characterized by the presence of grammatical errors, here understood in a broad sense, i.e. also including issues of orthography and lexical choice. Some of these issues, such as misspellings, are to some extent addressed in the universal UD guidelines (cf. Section 3.1) and do not typically alter the morphosyntactic analysis of the sentences in which they occur. However, it is not uncommon for grammatical errors to create a conflict between observed language use and intended meaning.

As a consequence, much of the discussion that has taken place in the context of L2 treebank development has revolved around trying to strike a balance between two seemingly contradictory approaches: following the principle of *literal reading* (as formulated by Berzak et al. 2016), according to which morphosyntactic analysis should be guided solely by the surface word forms and observed language use, and applying *distributional criteria*, in an attempt to reflect the writer's intentions. While annotating literally is often key to an informative analysis of nonstandard language, some amount of distributional insight is not only indispensable to deal with nonexisting word forms and otherwise unanalyzable constructions, but also intrinsic to the UD annotation scheme at large, independent of what material is analyzed.

We reframe the problem, guided by a similar but not identical question, i.e. to what extent the analysis of a learner sentence should be informed by its assumed intended meaning. While the principle of literal reading encourages not to rely on any particular interpretation of learner productions, it can be difficult to analyze ungrammatical constructions without some degree of speculation as to what the writer is trying to convey. In addition, when a correction hypothesis is available – such as in L1-L2 treebanks – it seems reasonable to try to ensure that the two analyses are consistent with each other, at least for the sake of comparability. Intuitively, the two trees constituting a sentence-correction pair should be "as different as necessary," so as to not hide any information that may help describe any discrepancies more accurately, and "as similar as possible" to facilitate fine-grained comparisons and error retrieval. An example showing the difference between literal and correction-driven annotation is given in Figure 1.

Another challenging situation, also illustrated in Figure 1, occurs when what the most informative annotation results in a category combination invalid in the underlying framework, which is descriptive of the standard use of the TL. This happens not only because of grammatical errors, but also in conjunction with other interlanguage phenomena such as code switching and syntactic calques. Relevant UD guidelines exist for some of these cases and will be discussed in Section 3.2.

3 L2 Phenomena in the Universal Guidelines

In this section, we summarize the general UD guidelines that relate to different aspects of the annotation of learner language. We restrict this discussion to the official guidelines available on the UD website,[2] although several proposals for unified guidelines for spoken data (Dobrovoljc, 2022) and user-generated content (Sanguinetti et al., 2020, 2023) also cover a variety of relevant phenomena.

3.1 Grammatical Errors

The universal UD guidelines comprise a page entirely dedicated to how to annotate typos, which has gradually evolved to also cover various types of strictly grammatical errors.[3] While it is explicitly stated that the recommendations are intended for dealing with sporadic errors rather than annotating learner corpora, they can be seen as a starting point and, as we will show in Section 4, they have been partially adopted by some of the existing treebanks.

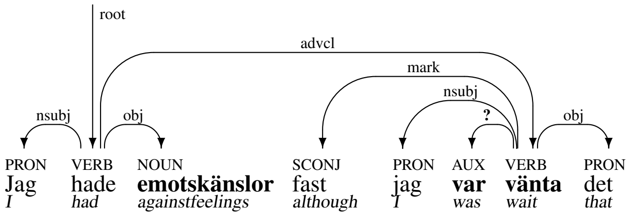
When it comes to minor errors affecting individual tokens, the general idea is to never alter the word form itself, but signal the presence of an error with the feature `Typo=Yes`. Lemmatization should be based on the normalized spelling of the word, but morphological features should always describe the observed form. Additional information, such as the `CorrectForm` and features of the token, may be provided in the `MISC` field, creating a situation in which lemmatization is correction-driven, while morphological tagging follows the literal reading of the word.

This approach is easy to implement for simple **misspellings** and some clear-cut cases of **inappropriate lexical choice**, **incorrect morphological derivation** and **inflection**, but leaves open a wide range of problems, such as the analysis of unrecognizable forms and the syntactic annotation of words

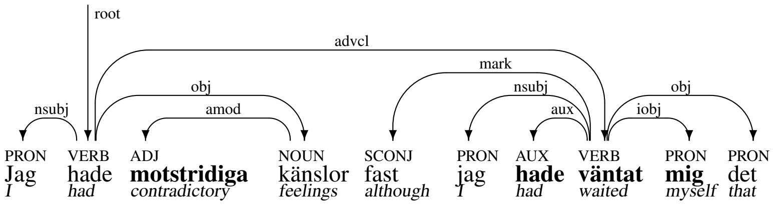
---

[2] `universaldependencies.org/guidelines.html`

[3] `universaldependencies.org/u/overview/typos.html`



(a) L2 sentence.



(b) Correction hypothesis.

Figure 1: Example sentence-correction pair from SweLL. The correction hypothesis suggests that intended meaning is *I had mixed feelings although I had expected that.* The learner, however, uses the past tense of the copular verb *vara* ("to be"), *var* instead of *hade*, the corresponding form of the temporal auxiliary *ha* ("to have"). A **literal** reading of the verb group *var vänta* would use the dependency label `cop`. Conversely, a **correction-driven** approach would follow the correction hypothesis and opt for the relation `aux` – as would happen for the English construction *I was expecting that* – even though the `aux`-*vara* combination goes against the validation rules for Swedish.

that are incorrectly inflected e.g. for case – which is a marker for their syntactic role – or noncanonical in terms of their POS. All of these issues, however, have been addressed in at least some of the existing L2 treebanks.

When it comes to ***hypersegmentation***, i.e. incorrectly split words, annotators are essentially redirected to the well-established guidelines for the dependency relation `goeswith`, to be used in conjunction with the `UPOS` tag `X` and the `Typo` feature. For missing spaces, also referred to as ***hyposegmentation***, on the other hand, the recommendation is to handle them with the `SpaceAfter=No` and `CorrectSpaceAfter=Yes` attributes in `MISC`. **Missing words** are to be treated as ellipsis, whereas for **redundant tokens** it is suggested to follow the guidelines for speech disfluencies, thus using the `reparandum` relation. Issues arising from other **syntactic errors** are acknowledged by the existing guidelines, but not regulated in any way.

3.2 Foreign Expressions and Code Switching

UD provides guidelines for both treebanking extensively code-switched corpora and for annotating sporadic foreign words and expressions in otherwise monolingual treebanks.[4] In the latter case, more relevant for L2 corpora, the annotator may choose between three types of analysis: code-switched, borrowed and foreign.

If the treebanker opts for a *code-switched analysis*, content in a language other than the treebank's primary one is analyzed according to the language-specific guidelines of the former. To allow proper automatic validation, its ISO code is specified in the `Lang` attribute of the `MISC` field of each foreign token. Similarly, their `FEATS` field should contain the feature `Foreign=Yes`.

With a *borrowed analysis*, foreign words and expressions are assumed to have become part of the vocabulary of the main language of the treebank: they should not be marked as `Foreign` nor mandatorily assigned a `Language`, although the donor language may be specified through the `OrigLang` attribute. If the borrowed expression consists of several words, it is reduced to a `flat` structure and its tokens are POS-tagged according to the syntactic role of the expression as a whole.

Finally, a *foreign analysis* implies leaving the third-party language content completely unanalyzed: all relevant tokens are POS-tagged as `X`, marked as `Foreign`, optionally assigned an `OrigLang` and linked toghether using the dependency label `flat`, optionally subtyped as `flat:foreign`.

4 Annotation Practices Across L2 Treebanks

In this section, we provide a comparative analysis of the annotation practices found in different L2 treebanks. Its objective is twofold: on the one

---

[4] `universaldependencies.org/foreign.html`

Language	Name	Parallel	Modality	Sentences	Status	Publications	Annotation manual
Chinese	CFL	(✓)	written	451	released[5]	Lee et al. (2017a)	✓[a]
English	ESL	✓	written	5124	retired	Berzak et al. (2016)	
English	ESLSpok		spoken	2320	released	Kyle et al. (2022)	✓[b]
Italian	Valico	✓	written	398	released	Di Nuovo et al. (2019, 2022)	✓[c]
Korean	KSL		written	7530	released	Sung and Shin (2024, 2025)	
Russian		✓	written	500	in progress	Rozovskaya (2024)	
Swedish	SweLL	✓	written	~5000	in progress	this paper	

Table 1: Overview of UD treebanks of learner language.
[a] `github.com/UniversalDependencies/UD_CFL`; [b] `kristopherkyle.github.io/L2-Annotation-Project/dep_anno_overview.html`; [c] `github.com/ElisaDiNuovo/VALICO-UD_guidelines`

hand, to see to what extent the official guidelines discussed in Section 3 have been adopted in learner corpora; on the other, to investigate whether the annotation of similar phenomena is consistent across projects and what aspects of L2 annotation would benefit from harmonization. Our findings are summarized in Table 2.

To achieve these goals, we systematically review the papers describing the various treebanks, focusing on the sections dedicated to annotation criteria. When available, we integrate this with information with instructions from the treebanks' annotation manuals and complement textual sources with tree queries performed with the STUnD tool for parallel treebanks (Masciolini and Tóth, 2024).

We consider five of the seven annotation efforts mentioned in the introduction and listed in Table 1: the Chinese as a Foreign Language treebank, henceforth **CFL** (Lee et al., 2017a);[5] the English as a Second Language treebank or **ESL** (Berzak et al., 2016); **VALICO**, based on the VALICO corpus of learner Italian (Di Nuovo et al., 2019, 2022), the Korean as a Second Language treebank, **KSL** (Sung and Shin, 2024, 2025) and the **Russian** learner treebank presented in Rozovskaya (2024), not yet released. We exclude the spoken L2 English treebank, ESLSpok, since neither the paper (Kyle et al., 2022) nor its annotation manual focus on L2-specific issues. The insights resulting from this comparison inform the guidelines for our upcoming SweLL-based treebank.

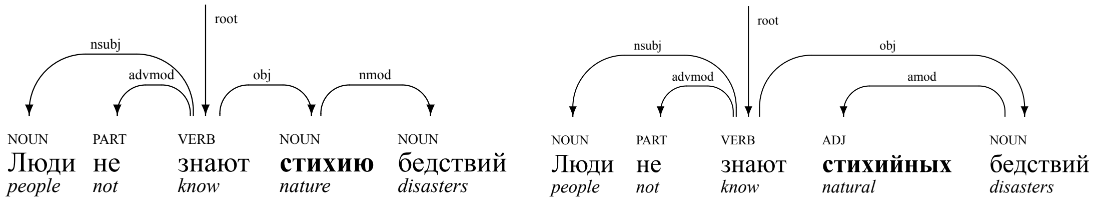
It must be noted that, for some annotation layers, CFL also provides a separate CoNLL-X file with double (literal and distributional) tags. In this paper, we focus on the annotation strategies that were chosen as the "default", i.e. on the values present in the CoNNL-U file distributed as part of the official UD releases. As for ESL, an important caveat is that the version described by Berzak et al. (2016) follows UD version 1 guidelines, which are now largely outdated. Furthermore, it should be kept in mind that such treebank is not annotated for `FEATS` or `MISC`. Similarly, CSL and KSL do not use the `FEATS` fields. Finally, the papers describing the Russian and Korean treebanks only discuss dependency annotation.

4.1 Grammatical Errors

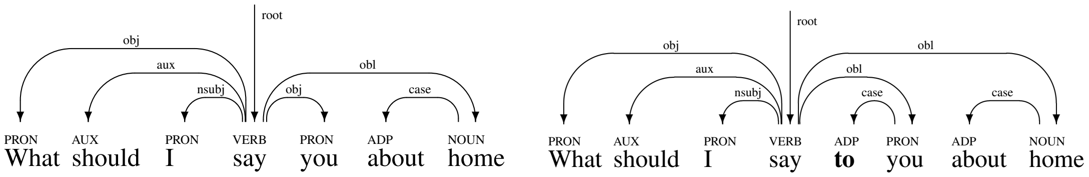
Spelling Errors Misspellings of individual tokens only pose a problem in terms of lemmatization, unless they result in an ambiguous or completely unrecognizable word form. For uncomplicated cases, CFL and KSL, two out of three treebanks providing lemmas, follow the correction-driven approach suggested by the universal guidelines, lemmatizing based on the normalized form. While initially following the same principle, on the other hand, later versions of VALICO adopt a prevalently literal approach to the annotation of misspellings, trying to preserve the errors in the lemmas. When it comes to closed-class words, however, some exceptions are necessary to avoid validation errors. Misspelled auxiliaries, for example, are always assigned normalized lemmas since the language-specific guidelines for Italian only allow a restricted set of lemmas to be associated with the `UPOS` tag `AUX`. In cases of ambiguity or unrecognizable forms, KSL attaches the problematic token to the following one using the dependency label `flat`. None of the other treebanks appears to apply a similar strategy. No treebank, with the possible exception of Russian, uses the feature `Typo` or indicates the `CorrectForm` in the `MISC` field for errors of this kind, even though VALICO does use the former in cases of hypersegmentation (see below).

---

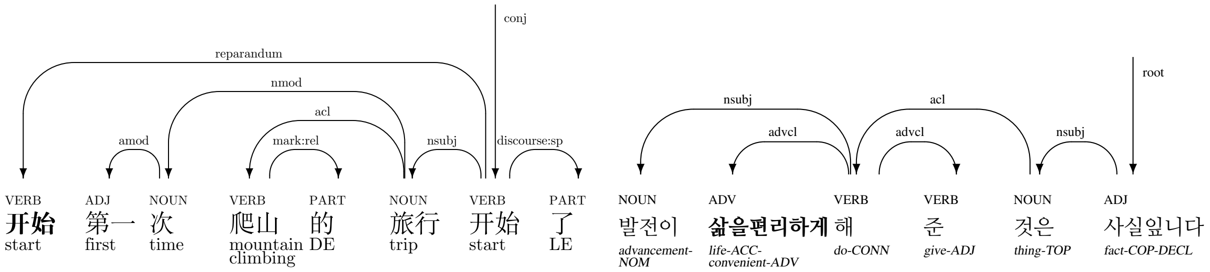
[5] The Chinese treebank was originally designed as a parallel treebank, but only its L2 half has been released.



(a) L2 sentence-correction hypothesis pair from Rozovskaya (2024) (approximate translation: "People don't know natural disasters"). In the learner sentence (on the left), a noun ("nature") is used instead of the corresponding adjective ("natural"). Since the noun is in accusative case, it is annotated as direct object of the main verb, and the following noun ("disasters") becomes its nominal modifier, even though this does not match the semantics of the correction.

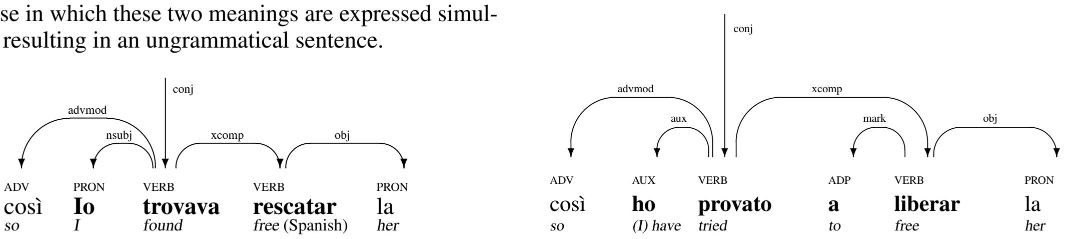


(b) L2 sentence-correction hypothesis pair from the last available version of ESL (updated to UD version 2 guidelines). The learner sentence lacks the preposition *to*. Consequently, *you* is annotated as an additional direct object of *say* rather than as an oblique. An alternative annotation that is also compatible with literal criteria is marking *you* as indirect object.



(c) Example of use of the `reparandum` relation from CFL 2.15. In context, there are two possible translations for this fragment: "the first mountain climbing trip starts" or "(we) start (our) first mountain climbing trip", depending on which of the two occurrences of the verb "start" is considered to be redundant (the dependency tree for the full sentence, with ID `CFL_G_1-3_ori`, can be inspected at `universal.grew.fr/?custom=68509ca3643c1`). Arguably, rather than a self-repair, this is a case in which these two meanings are expressed simultaneously, resulting in an ungrammatical sentence.

(d) Example of hyposegmentation from Sung and Shin (2024), updated to KSL 2.15 (approximate translation: "it is true that advancements have made our lives more convenient"). The second token consists of two syntactic words that would be separated by a space in standard Korean. This tokenization choice goes against the general guidelines discussed in Section 3.1 but makes the learner error visible in the dependency tree.



(e) L2 sentence-correction hypothesis pair from VALICO 2.15 (approximate translation: "so I tried to free her"). The Spanish word *rescatar*, corrected to *liberar* is marked as `Foreign` but lemmatized as following the Spanish (*rescatar*) rather than Italian (*rescatare*) lemmatization rules. Despite various overlapping orthographical, inflectional and lexical errors, the learner construction is still analyzable without any violations of the language-specific guidelines for Italian.

Figure 2: Selection of annotated sentence fragments from the existing L2 treebanks.

Incorrect Lexical Choices In the majority of cases, lexical choices are irrelevant to syntactic structure. However, problems can arise in at least two often overlapping cases: when the incorrect token is a function word and when its POS differs from that of the corresponding correct lexical item. Auxiliaries, for instance, are a closed class. If a lexical verb is used in place of an auxiliary, the annotator must therefore classify it as `VERB`, which in turn affects the dependency structure of the sentence. None of the available guidelines discusses this specific issue, but Di Nuovo et al. (2022) mention choosing the `UPOS` and `DEPREL` of misused closed-class words based on their literal reading. CFL makes abundant use of the generic dep relation type to solve cases in which the literally assigned POS tag is incompatible with any meaningful dependency label, whereas both ESL and the Russian treebank try to avoid it. At least in the case of ESL, this produces some violations of the current validation rules. On the opposite side of the spectrum, VALICO assigns POS tags according to the assumed intended meaning of the sentence.

Incorrect Derivation Issues related to derivational morphology can cause learners to inadvertently produce word forms that belong to a syntactic category other than intended one (cf. Figure 2a). Across treebanks, these cases are handled in the same way as incorrect lexical choices causing a change in POS (see above).

Incorrect Inflection The languages considered in this paper vary widely in terms of richness of their inflectional morphology. The general UD guidelines for Korean, for instance, do not discuss morphological features at all, while the Chinese treebank use them very sparsely. For this reason, neither the description of KSL nor the guidelines for CFL mention any issues of inflectional morphology. In English and Italian, the latter plays a more important role. In ESL, however, this is not discussed in depth since the treebank does not use universal morphological features.[6] In VALICO, on the other hand, inflectional errors are more prominent. Morphological features are assigned based on the observed word form, in accordance with the universal guidelines discussed above, and in the rare cases in which incorrect inflection can be seen as to alter the syntactic structure of the sentence, dependency annotation is consistent with the features. In Russian, the language with the richest morphology out of the ones addressed in this paper, the latter phenomenon is more frequent since the syntactic role of nominals with respect to their heads is marked by their case inflection. Consistently with Berzak et al. (2016) and Di Nuovo et al. (2022), Rozovskaya (2024) always selects the `DEPREL` of nominals based on their morphological suffixes, i.e. according to Berzak et al. (2016)'s definition, following literal criteria. A related phenomenon occurs in Korean, where particles indicate the syntactic role of the words they are attached to. In such cases, KSL guidelines suggest an approach where dependency annotation is guided by the intended use of the affected word, inferred from the context.

Hyper- and Hyposegmentation When it comes to hypersegmentation, the universal guidelines for the `UPOS` and `DEPREL` fields are generally followed, resulting in a widespread use of `goeswith` and `X`. For Korean, however, Sung and Shin (2024) explicitly mention one exception: when an extra space is found between a content word and its subsequent particle(s), the latter are treated as dependents of the former and the relation between them is labelled as `case`. In addition, ESL presents some exceptions when it comes to POS tagging: although the treebank is not entirely consistent in this respect, the tendency is to assign standard `UPOS` tags individually to each half of an incorrectly split word when the two segments are two complete recognizable words. In this sense, the most frequent case is that of incorrectly split compounds. None of the treebanks follows the universal guidelines when it comes to the `MISC` field. As for hyposegmentation, the only two treebanks for which the issue is explicitly discussed adopt two opposite approaches: VALICO follows the universal guidelines, thus splitting the incorrectly glued tokens and annotating them with `SpaceAfter=No` and `CorrectSpaceAfter=Yes`, whereas KSL leaves the tokens fused together, using the final morpheme as a cue to determine the `DEPREL` (cf. Figure 2d).

Missing Words At least to some extent, all five treebanks treat missing words as ellipsis. This works well for handling dependents lacking a head. However, missing function words – which are typically leaves – can alter the literal reading of a sentence in terms of syntactic structure. As a consequence, all treebank-specific annotation guidelines excepts CFL's mention a few cases that require special treatment. The most widespread issue is that of missing prepositions before non-core verb arguments, mentioned by both Berzak et al. (2016), Di Nuovo et al. (2022) and Rozovskaya (2024). VALICO seems to adopt a case-by-case approach, while in the English and Russian treebanks, these cases are consistently treated like incorrect case inflection in Russian: if an oblique is not introduced by a preposition, for example, it automatically becomes an `obj` or `iobj` (cf. Figure 2b).[7] Berzak et al. (2016) consider this to be one of the prime examples of literal – as opposed to distributional – annotation. We, however, argue that the use of

---

[6] `XPOS` tags, however, are assigned literally.

[7] As mentioned above, the version of ESL described in Berzak et al. (2016) actually follows UD version 1 guidelines, where direct objects were labelled `dobj` and there was no distinction between `nmod` and `obl`.

L2 Phenomenon	Guidelines	CFL	ESL	VALICO	KSL	ru
incorrect words	correct `LEMMA`	✓	–	✗*	✓	
	`Typo=Yes`	–	–	✗	–	
hypersegmentation	`goeswith`	–	✓	✓	✓*	
	`UPOS=X`	–	✓*	✓	✓	
	`Typo=Yes`	–	–	✓	–	
hyposegmentation	re-segmentation	–		✓	✗	
	`SpaceAfter=No`	–	–	✓	✗	
missing words	treat as ellipsis		✓	✓		✓
redundant words	`reparandum`	✓				✗
foreign material	(type of analysis)		`F`	`CS`		

(a) Adoption of universal guidelines for different macro-categories of L2 phenomena. A checkmark (✓) means that the treebank follows the general guidelines, whereas a cross (✗) implies the application of project-specific criteria. `F` and `CS` are abbreviation for, respectively, foreign and code-switched analysis. The `MISC` field is not considered here.

Error type	CFL	ESL	VALICO	KSL	ru
spelling errors	✓	✓	✗*	✓	
incorrect POS (derivational or lexical error)	✗	✗	✓		✗
incorrect `Case` inflection or particles	–	✗	✗*	✓	✗
hypersegmentation	–	✓*	✓	✓*	
hyposegmentation	–		✓	✗	
missing function words		✗	✓✗	✓	✗
redundant function words	✗	✗	✗	✗	✗
redundant content words	✓				
missing content words	✗	✗	✗	✗	✗
incorrect word order	✗				

(b) Application of correction-aware approaches to the annotation of different error types. A checkmark (✓) means that the annotation is at least partially correction-driven, whereas a cross (✗) implies the application of purely literal criteria.

Table 2: Overview of annotation practices across L2 treebanks. "–" signifies that the phenomenon and/or CoNNL-U field at hand are irrelevant for the treebank and/or language in question. The presence of any exceptions is signaled by asterisks (*). Cells for phenomena that are relevant but undocumented and/or unattested in the data are left blank.

`obl` and `obj`/`iobj` in this context are the results of two different distributional readings, both based on the context in which the oblique occurs, but only informed by the correction hypothesis in the former case. Similar problems arise for missing clitics (Italian) and particles (Korean). In VALICO, this can alter the dependency tree structure, whereas in KSL content words unmarked by the relevant particle are annotated as if the particle was present.

Redundant Words Across treebanks, the `reparandum` relation, whose use the universal guidelines encourage for redundant words, is attested five times, one in CFL and four in ESL. Only in the first case, however, its usage seems referred to a syntactically redundant word (cf. Figure 2c), while all ESL examples are akin to disfluencies in speech corpora. Guidelines for CFL, ESL, VALICO and the Russian treebank, however, mention a number of interesting subcases. In the English, Italian and Russian data, there are instances of redundant prepositions – a complementary problem to missing prepositions. In all three treebanks, the redundant preposition determines the syntactic role of the nominal it introduces, resulting in either `obl` and `nmod`. Furthermore, there are mentions of redundant clitics (Di Nuovo et al., 2019), aspects markers (Lee et al., 2017a), connectors and markers (Rozovskaya, 2024), all treated literally. Only Di Nuovo et al. (2022) mentions unrecognizable words, annotated with the `UPOS` tag `X` and the dependency label `dep`.

Other Syntactic Errors Other syntactic errors, such as atypical word order, are often not discussed at all in the treebanks' documentation. This is either because they do not affect the analysis of the sentence – which is the case in languages with no strict word order, such as Russian – or because their sparsity and heterogeneity encourages case-by-case decisions. Only the annotation guidelines for Chinese contain a paragraph that specifies how to handle words that appear misplaced: they should be attached to the closest possible head and marked as `dep`, i.e. left unanalyzed.

4.2 Other Interlanguage Phenomena

Surprisingly, only the VALICO documentation mentions interlanguage phenomena other than errors. According to Di Nuovo et al. (2019), **foreign words and expressions** are marked as `Foreign` and lemmatized based on standard rules for Italian. Later on, however, the project switched to a more clearly code-switched analysis using source language lemmas, although their language is not specified in the treebank (Di Nuovo et al., 2022). This, however, does not cause any validation issues since there are no foreign multiwords whose syntax breaks the language-specific validation rules for Italian (cf. Figure 2e). As for the other treebanks, only the feature `Foreign` is attested, exclusively in ESL and paired with the dependency label `flat` and with the POS tag `X` in all but one case, suggesting an approach similar to the foreign analysis in the universal guidelines.

Di Nuovo et al. (2022) also discuss ***syntactic calques***, i.e. phrases from the learner's L1 (or, more generally, from their extra-L2 linguistic repertoire) translated word-for-word to the target language. Arguing that this is the better choice in terms of cross-linguistic comparability, VALICO opts for analyzing the resulting expressions as if they were code-switched material in the source language, unless doing so results in a structure not valid in the Italian-specific UD guidelines.

5 Directions for Harmonized Guidelines

As shown in Table 2, neither the adoption of universal annotation guidelines relevant for L2 phenomena nor the application of correction-aware approaches to the annotation of learner errors is uniform across treebanks. Based on this comparative analysis and on insights from our first-hand experience treebanking SweLL, we propose a set of core principles for the harmonization of annotation guidelines for learner material.

In the following, we distinguish *token-level annotation*, i.e. the assignment of lemmas, morphological features and POS tags to individual words, from *syntactic annotation*, consisting in establishing and labeling head-dependent relations between pairs of tokens. We argue that the two can be guided by different, complementary principles, and that this results in more informative analyses.

5.1 Literal Token-level Annotation

As reported in Section 3.1, the universal UD guidelines call for a literal approach to POS and morphological tagging, but require lemmas to be inferred from the corrected version of the word form at hand. These principles are adopted in most project-specific guidelines, which also seem to agree on the fact that `FEATS` and `UPOS` should be assigned based on the literal criteria in a broader sense. VALICO is exceptional in that even the content of the `LEMMA` field is determined on the basis of the observed word form, e.g. preserving misspellings. We argue that this approach sometimes results in more informative analyses of certain learner errors and should be recommended for all learner treebanks. The application of this principle to a sentence from SweLL is exemplified in Figure 3.

PRON	VERB	ADJ	NOUN
Jag	har	flera	**förslagor**
jag	ha	flera	**förslaga**
I	*have*	*several*	*proposals*

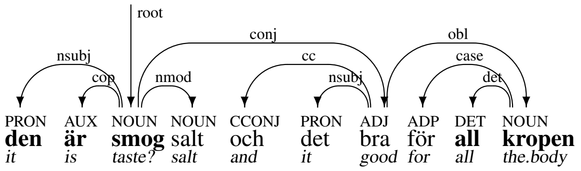
Figure 3: Sentence fragment from SweLL. The learner uses the incorrect plural form of the word *förslag*, inflecting it as if it belonged to a different declension. Lemmatizing as *förslaga* captures this phenomenon.

The `Typo` feature and the `CorrectXXX` attributes are not widely used, perhaps because they are considered redundant for parallel treebanks, where the corrected form, lemma and features of a token can be inferred from the correction hypothesis. Monolingual L2 treebanks, on the other hand, could benefit from their more systematic use. If lemmatization is performed as suggested above, the introduction of a `CorrectLemma` field would make the annotation more complete and could be checked by the UD validator instead of the `LEMMA` field itself to address potential lemma-POS mismatches such as the ones discussed in Section 4. We therefore encourage the use of such fields, which are important for monolingual L2 treebanks and can be largely filled in automatically in parallel datasets.

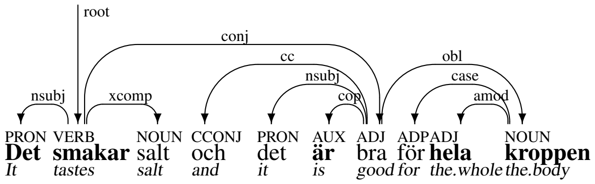
Finally, universal guidelines for hypersegmentation are widely adopted, but this is not the case for hyposegmentation. In the latter case, we strongly support KSL's choice not to re-segment incorrectly merged tokens, as doing so makes learner errors invisible to many UD tools.

5.2 Correction-aware Syntactic Annotation

As discussed in Section 2, we refrain from a discussion of syntactic analysis in terms of the two categories "literal" and "distributional" as we argue that dependency annotation is always, to some extent, distributional. Instead, we focus on the role of corrections in the annotation process and recommend a *correction-aware* approach. This is not to say that dependency relations used in an L2 sentence should follow those found in the annotation


root
conj
obl
nsubj
cc
case
cop
nmod
nsubj
det
PRON AUX NOUN NOUN CCONJ PRON ADJ ADP DET NOUN
den är smog salt och det bra för all kropen
it is taste? salt and it good for all the.body

(a) L2 sentence.


root
conj
cc
obl
nsubj
nsubj
xcomp
cop
case
amod
PRON VERB NOUN CCONJ PRON AUX ADJ ADP ADJ NOUN
Det smakar salt och det är bra för hela kroppen
It tastes salt and it is good for the.whole the.body

(b) Correction hypothesis.

Figure 4: Example sentence pair from SweLL. The L2 production is syntactically unclear, but the correction hypothesizes the intended meaning to be *It tastes salt and it is good for the whole body*. *smog* is therefore interpreted as a misspelling of *smak* ("taste", noun), in a construction that can be literally translated as *it is a taste (of) salt*. Furthermore, the adjective *bra* is considered to be the head of the second conjunct even if the original phrase lacks a copula. The resulting trees share a similar high-level structure but diverge enough to capture all grammatical errors.

of its correction hypothesis – which is in any case not available in non-parallel treebanks – but rather that the annotator should ground syntactic analysis in both the observed language use and the assumed intended meaning. We argue that this is the only informative way to analyze syntactically unclear sentences, which are a relatively common occurrence when the language use deviates from the standard for the TL significantly. In parallel treebanks, this also guarantees the analysis of learner sentences is consistent with that of the corresponding CHs, improving comparability.

Correction-aware syntactic annotation may appear contradictory with what we proposed in Section 5.1. Rather, we see these two principles as complementary, as we illustrate in Figure 4. As shown in Figure 1, however, there are cases in which applying literal criteria at the token level and determining the dependency structure of the sentence based on the distributional properties of the words results in category combinations not allowed by the UD validator. We therefore propose the use of a new relation subtype, `*`, to be used to mark deliberate violations of the UD universal and/or language-specific guidelines in correspondence of an L2 phenomenon that itself goes against the norms of the TL. An example of its intended usage is given in Figure 5.[8]

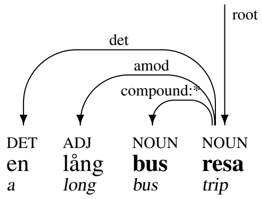


Figure 5: Sentence fragment from SweLL. In standard Swedish, compounds are written as single words (*bus trip*, for instance, translates to *bussresa*). The label `compound` should therefore only be used for borrowed compounds. Since it can be argued that the learner sentence mimics a construction from another language, we propose the use of the label `compound:*` and the POS tag `NOUN` for both *bus* and *resa* as a more informative alternative to `goeswith` and `X`.

5.3 Syntactic Calques as Code-switching

When it comes to foreign content, we recommend VALICO's approach, which allows for more informative annotations than ESL's foreign analysis. Taking inspiration from both VALICO's treatment of syntactic calques and ESL's approach to POS tagging incorrectly split compounds, we propose extending the recommendations for code-switched analysis even to constructions where TL words are combined into a construction reminding of the learner's extra-TL repertoire, with minor adaptations: while using the `Lang` feature may be misleading for tokens that, individually, correspond to words of the primary language of the treebank, the use of `*` would make it clear that the construction is nonstandard without strictly binding it to the validation rules of the calqued language. An example is given in Figure 5.

6 Concluding Remarks

We discussed the challenges L2 poses for UD annotators and gave an overview of the universal guidelines relevant for learner corpora, as well as of the current annotation practices across five of the existing treebanks. We concluded the paper with some suggestions for harmonization, which we hope can be the starting point for a productive discussion, primarily with other members of the UD community working directly with L2 material, but also with annotators of spoken and/or code-switched data, user-generated content and nonstandard language varieties such as pidgins and dialects, all of which share some features with learner language.

---

[8] Similarly, `*` could be used in the example shown in Figure 1 to allow analyzing *var* as an `aux`.

Acknowledgments

This work has been funded by Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (2025–2028; project id 2023-00161) and supported by the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology). In particular, the 2024 UniDive training school in Chișinău provided the first author with excellent opportunities for annotation experiments and preliminary discussions of L2-related issues with other members of the UD community. We also want to thank Xurui Zhang for his invaluable help with interpreting the annotation of the Chinese examples, as well as Dana Dannélls and the SyntaxFest 2025 reviewers for their thoughtful and supportive feedback at different stages of the writing process.

References

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746, Berlin, Germany. Association for Computational Linguistics.

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

Elisa Di Nuovo, Cristina Bosco, Alessandro Mazzei, and Manuela Sanguinetti. 2019. Towards an Italian learner treebank in Universal Dependencies. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR-WS.

Elisa Di Nuovo, Manuela Sanguinetti, Alessandro Mazzei, Elisa Corino, and Cristina Bosco. 2022. VALICO-UD: Treebanking an Italian learner corpus in Universal Dependencies. *IJCoL. Italian Journal of Computational Linguistics*, 8(8-1).

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.

Kristopher Kyle, Masaki Eguchi, Aaron Miller, and Theodore Sither. 2022. A dependency treebank of spoken second language English. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 39–45, Seattle, Washington. Association for Computational Linguistics.

John Lee, Herman Leung, and Keying Li. 2017a. Towards Universal Dependencies for learner Chinese. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 67–71, Gothenburg, Sweden. Association for Computational Linguistics.

John Lee, Keying Li, and Herman Leung. 2017b. L1-L2 parallel dependency treebank as learner corpus. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 44–49, Pisa, Italy. Association for Computational Linguistics.

Arianna Masciolini. 2023. A query engine for L1-L2 parallel dependency treebanks. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 574–587, Tórshavn, Faroe Islands. University of Tartu Library.

Arianna Masciolini and Márton A Tóth. 2024. STUnD: ett Sökverktyg för Tvåspråkiga Universal Dependencies-trädbanker. In *Proceedings of the Huminfra Conference*, pages 95–109, Gothenburg, Sweden.

Alla Rozovskaya. 2024. Universal Dependencies for learner Russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17112–17119, Torino, Italia. ELRA and ICCL.

Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2020. Treebanking user-generated content: A proposal for a unified representation in Universal Dependencies. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5240–5250, Marseille, France. European Language Resources Association.

Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé 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.

Hakyung Sung and Gyu-Ho Shin. 2024. Constructing a dependency treebank for second language learners of Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3747–3758, Torino, Italia. ELRA and ICCL.

Hakyung Sung and Gyu-Ho Shin. 2025. Second language Korean Universal Dependency treebank v1.2: Focus on data augmentation and annotation scheme refinement. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 13–19, Tallinn, Estonia. University of Tartu Library, Estonia.

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, and 1 others. 2019. The SweLL language learner corpus: From design to annotation. *Northern European Journal of Language Technology*, 6:67–104.

Developing a Universal Dependencies Treebank for Alaskan Gwich'in

Matthew Andrews
University of Tübingen
`m.kirkandrews@gmail.com`

Çağrı Çöltekin
University of Tübingen
`cagri.coeltekin@uni-tuebingen.de`

Abstract

This paper presents a Universal Dependencies (UD) treebank of Gwich'in, a severely endangered Athabascan language. The treebank, developed using instructional materials and dictionaries, includes 313 annotated sentences. This paper discusses the methodology used to construct the treebank, the linguistic challenges faced, and the implications of annotating a polysynthetic, morphologically complex language within the Universal Dependencies framework. The treebank was released with UD version 2.15 and available at `https://github.com/UniversalDependencies/UD_Gwichin-TueCL/`.

1 Introduction

We present a treebank of Gwich'in annotated following the Universal Dependencies framework (de Marneffe et al., 2021). To our knowledge, this is the first treebank of the Athabascan language family, a language family spread over a large part of North America. Recent efforts in the Universal Dependencies community have increased coverage of the world's languages substantially. As of the most recent data release, there are 319 treebanks of 179 different languages from 62 different language groups (Zeman et al., 2025). However, excluding the present effort, only one language indigenous to the United States or Canada is included in the database. Despite the growing interest in Universal Dependencies and in the development of tools and resources for low-resource languages among the NLP-community, languages from this linguistically diverse region of the world remain underrepresented among UD treebanks.

Since the early 1990s, linguists have been concerned with the issue of language endangerment. Athabaskanist Michael E. Krauss was one of the first to address this problem in his classic article "The World's Languages in Crisis", in which he compared the loss of linguistic diversity to the loss of biodiversity (Krauss, 1992). In order to draw global attention to the critical situation of language endangerment, the United Nations General Assembly has declared the period between 2022 and 2032 the International Decade of Indigenous Languages (United Nations General Assembly, 2019). Inline with this spirit and in the spirit of the Universal Dependencies project's commitment to linguistic diversity we present a small manually annotated treebank for Gwich'in, a critically endangered language with interesting linguistic properties, including polysynthesis and classificatory verbs.

Following an introduction of the Gwich'in language, this paper will present and discuss the methodology used and some of the linguistic challenges faced in the development of this treebank.

2 Background

Gwich'in is an indigenous language of North America spoken by the Gwich'in people, also know as *Dinjii Zhuh*, who live in what is now the US state of Alaska, Yukon and the Northwest Territory in Canada. Historically, all *Dinjii Zhuh nąįį* could speak their language, *Dinjii Zhuh k'yàa*, but today, no more than 300 of the 4000 *Dinjii Zhuh nąįį* can, most of whom are elders (Mishler and Frank, 2019). Due to colonial pressure on the *Dinjii Zhuh nąįį* and other indigenous peoples to give up their native tongues, parents are no longer able to speak *Dinjii Zhuh ginjìk* to their *tr'iinin nąįį* ('children') at home.

Despite these challenges, there is a vibrant Gwich'in community interested in learning and revitalizing their language. Currently, Gwich'in is taught as a course at the University of Alaska-Fairbanks and in some K-12 schools in Canada. In recent years, some language learning products have also been developed, including books (Gwich'in Social and Cultural Institute, n.d.a), online dictionaries (Gwich'in Social and Cultural In-

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 164–173
August 27, 2025 ©2025 Association for Computational Linguistics

stitute, n.d.b), apps, a language learning platform (Doyon Foundation, n.d.), and Gwich'in language episodes of the Emmy award winning PBS Kids show "Molly of Denali" whose main character Molly is part Gwich'in Athabascan.

Besides these few resources, there is very little documentation and description of the Gwich'in language. The vast majority of data and information on the language consists of scans of handwritten field notes, classroom worksheets, and unannotated audio recordings housed in the Alaska Native Language Archive (Alaska Native Language Archive, n.d.). From the perspective of a computational linguist, these resources in their current form are difficult to use for the development of language technology or a so-called "NLP-toolkit" (Zariquiey et al., 2022).

In addition, few academic works have been published on the linguistic aspects of the Gwich'in language. Publications on neighboring Athabascan languages such as Koyukon and Upper Tanana are much easier to find (Axelrod, 1990; Lovick, 2020, 2023). Moreover, the majority of research on these languages and on Athabascan in general has focused more on the rich morphology of the language, particularly the verbal morphology, than on any other area of linguistic study. To date, the syntactic structure of Athabascan has received far less attention. The present study hopes to address these gaps.

3 Methodology

Text Source Materials referenced for the development of the Gwich'in treebank come from the Alaska Native Language Archive of the Alaska Native Language Center at the University of Alaska-Fairbanks. These materials consisted of scans of previous classroom handouts developed for university language courses (Marlow and Garnett, 1996) as well as English to Gwich'in dictionaries (Peter, 1979; Mueller and Garnett, March 1994). The classroom handouts included verb paradigms and short sentences which were selected for annotation in the Universal Dependencies framework. These materials also provided detailed grammatical descriptions of many linguistic aspects of Gwich'in. Select sentences with interlinear glosses were also chosen from two Master's theses on Gwich'in directionals (Busch, 2000) and classificatory verbs (Bushey, 2021). The treebank focuses exclusively on the Gwich'in variety spoken in Alaska.

Approach At a minimum, sentences in Gwich'in consist of a single morphologically complex verb form. A verb form consists of a verb stem, a subject prefix belonging to one of four classifiers, and zero or more inflectional and derivational prefixes. Verb stems are encoded for tense/aspect, and subject prefixes must agree in person and number as well as with the verb stem's classifier and tense/aspect. Information on a word's part of speech and lemma or dictionary form were annotated as were basic dependency relations between words. English translations, morpheme segmentation, and detailed interlinear glosses were included as well. The annotation tool UD Annotatrix (Tyers et al., 2017) was used to streamline annotations as was a Gwich'in dictionary web application developed by the first author to help determine word lemmas.

Corpus The treebank consists of a total of 313 annotated sentences. Approximately two-thirds of the corpus consists of one- to two-word sentences highlighting several different verb paradigms with a verb in the imperfective, perfective, and future tenses and zero or one noun objects. The remainder of the corpus consists of longer, more complex multi-word sentences covering a majority of part-of speech tags and dependency relations of the Universal Dependencies framework.

Language-specific features[1] include the feature `Classifier` and its values `Ø`, `L`, `Ł`, and `D` (See Appendix A for the descriptions). The name for this feature and its values are consistent with the Athabascan literature. A Gwich'in verb stem takes one of these four classifiers, each of which has its own set of subject prefixes. Compare the verb *yahtsii* 'he or she is making it' which has the verb stem *-tsii* and the subject prefix *-ah-* with the verb *yitsii* 'he or she is caching it' which has the same stem but takes the subject prefix *-i-*. The verb stems are the same, but the third-person subject prefixes are different as these verbs have different classifiers.

The language-specific feature `StemClass` and its values are shown in Appendix A. This feature is used to account for the noun class to which a Gwich'in classificatory verb stem belongs. This and other language-specific challenges are covered in the next section.

A Note on Tone Gwich'in is a tonal language. Low-tone, rising-falling, and falling-rising are

---

[1] Morphological features are currently omitted from the treebank but will be provided in future releases.

Disjunct					Conjunct						Stem
0	1a	1b	2	3	4	5	6	7	8	9	10
postposition object	"null post-position"	adverbial-thematic	iterative	plural	direct object	deictic	adverbial-thematic	mode-aspect	subject	classifier	stem

Table 1: Verb template for Navajo, a Southern Athabascan language.

marked. There is much inconsistency of the marking of tone among the sources referenced in the creation of this treebank. Some sources mark tone, while others do not. Sentences in this treebank are transcribed exactly as they appear in the source text. As such, the same word may appear different elsewhere in the treebank although it is not, one with tone marking and one without.

4 Language-specific Features and Challenges

4.1 Verbs

Gwich'in like all Athabascan/Na-Dene languages is well known for its rich verbal morphology. Gwich'in is a polysynthetic, primarily prefixing, head-final language with the verb stem occupying the last position in a verb and the verb typically occurring as the last element in a clause. An Athabascan verb is composed of a verb stem and a set of prefixes. This is often represented in the Athabascan literature with a verbal template as in the one for Navajo shown in Table 1. There are 12 positions depicted in this template; however, not all positions are necessarily filled, and those that are are filled in a discontinuous manner. The maximum number of occupied slots found in the present treebank is seven.

The exact makeup of a (verb) word cannot be determined from its surface form alone, as the individual morphemes that make up the word are obscured by a number of morphophonological processes (Hale, 2003). This is illustrated for the verb *Giyahąąh'yaa* 'They will see it' in example 1 below.

(1) *gï-* *yï-* *hï-* *nï-* *ah-*
PL 3O FUT THEM 3SG+Ł+FUT
'yaa
see+FUT
'They will see it'

The verb stem *-'yaa* 'will see' is part of the Ł classifier and so takes the subject prefix *-ah-* for third person from the set of subject prefixes for the Ł classifier. Each of the four verb classifiers in Gwich'in has its own distinct set of subject prefixes. The subject prefix is always closest to the verb stem. The next prefix *-nï-* is a thematic prefix and together with the verb stem and classifier forms what is called the verb theme. Then we have *-hï-*, a prefix that indicates future tense or aspect, preceded by *-yï-* to indicate the third person direct object of a third person subject. Third person marking in Gwich'in and Athabascan languages is quite complex and is the subject of a major theoretical question which we will return to below. Finally, the prefix *-gï-* occupies the position for plurality and is used for third person subjects.

Notice that many of these prefixes undergo a vowel change in their surface form. The use of *ï* in these prefixes is to indicate the underlying vowel phoneme /i/ that is subject to vowel harmony with the verb stem and that surfaces as [a] in the presence of [a], [o], and [w]. However, *gï-* does not undergo this change, as plural morphemes belong to the disjunct zone of the verb template as seen in Table 1. All other prefixes belong to the conjunct zone. Only prefixes of the conjunct zone are subject to phonological changes.

The deletion of /n/ of the thematic prefix and the nasalization of adjacent vowels can also be seen in the surface form. This phonological process occurs when the thematic prefix is preceded by another prefix. A second phonological rule is applied to the vowel of the future prefix *-hï-*, where the vowel is deleted when the following vowel is long.

4.1.1 Classificatory Verbs

The noun classification system of Gwich'in and other Athabascan languages is quite elaborate. Bushey (2021) identifies nine classes for noun class membership, so-called 'stem-classes' in the Athabascan tradition. However, in Athabascan languages, noun class membership is not indicated on the noun but rather on the verb.

So-called 'classificatory verbs' are a hallmark of every Athabascan language including Gwich'in. Classificatory verb stems are unique in the sense that they classify a particular noun by its physical characteristics, e.g., shape, as well as describe the movement, handling, or position of the noun.

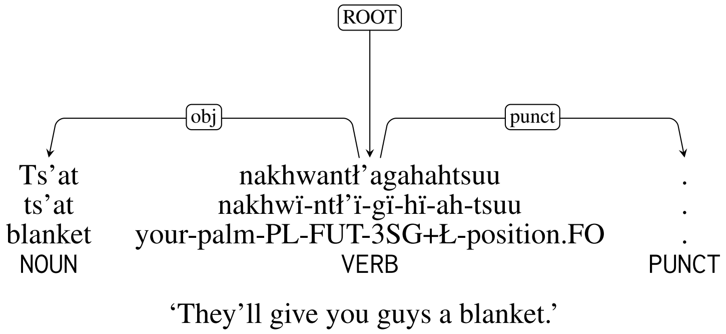

ROOT
obj
punct
Ts'at
ts'at
blanket
NOUN
nakhwantł'agahahtsuu
nakhwï-ntł'ï-gï-hï-ah-tsuu
your-palm-PL-FUT-3SG+Ł-position.FO
VERB
.
.
.
PUNCT
'They'll give you guys a blanket.'

Figure 1: Verb stem *-tsuu* meaning 'give soft, fabric object'.

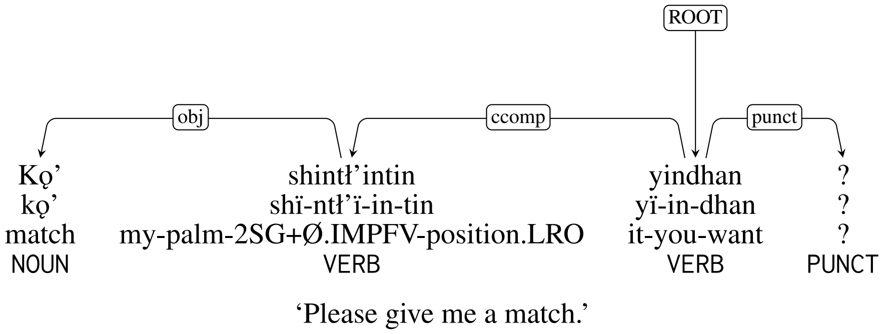


Figure 2: Verb stem *-tin* meaning 'give long, rigid object'.

This phenomenon is somewhat similar to the English verb *roll* which is typically used only with round objects (Lovick, 2023). However, classificatory verbs are much more common and complex in Gwich'in than in English. Take for instance the verb stem *-tsuu* seen in Figure 1 which together with its classifier Ł can be glossed as 'position cloth-like object.' This particular verb may be used with fabric objects (FO) like shirts and blankets. Compare this with the verb stem *-tin* found in Figure 2, which is used with stick-like or long, rigid objects (LRO) such as matches or guns. Notice how both of these verbs are translated with the English word 'give.' Although the complete details of classificatory verbs are beyond the scope of this paper, it is worth noting that nine stem classes have been identified for Gwich'in. Put another way, there are potentially nine different Gwich'in words for the English word 'give.'[2] This is also true for other verbs having to do with the movement, handling, and positioning of objects, e.g., 'fall', 'carry', 'put', etc.

A list of suggested values for the feature `StemClass` to account for the noun classification system of Gwich'in can be found in Appendix A

4.1.2 (Pro)nominal Incorporates

As seen in example 1 above, the verb word functions as a complete sentence, with both subject and object pronominals marked on the verb. The status of these markers has been the subject of considerable theoretical debate. When compared with pro-drop languages like Italian and Spanish, it is often assumed that Gwich'in and other Athabascan languages also feature null arguments.

The Pronominal Argument Hypothesis (PAH) offers an alternative view, arguing that pronominal markers are not simply agreement markers but core arguments of the sentence (Jelinek, 2014; Hale, 2003). Free noun phrases, when they do occur, are treated as adjuncts. This suggests that, unlike in pro-drop languages, there are no null arguments; instead, arguments are marked directly on the verb (Gelderen and Willie, 2012).

In Southern Athabascan languages such as Navajo, both subject and object pronominals must be marked on the verb. This contrasts with Northern Athabascan languages like Gwich'in, where nominal objects are in complementary distribution. Specifically, a nominal object can either be a free noun phrase or expressed as a pronominal marker, but not both at the same time (Gelderen and Willie, 2012). For instance, when *dinjik* 'moose' is present

---

[2] In actuality, only eight of these are attested in the data collected by Bushey (2021). A verb for the stem class of deteriorated objects is unattested in the data.

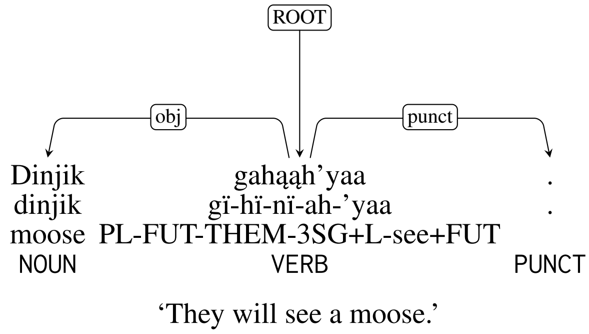


Figure 3: Verb with free noun phrase and no pronominal object marker.

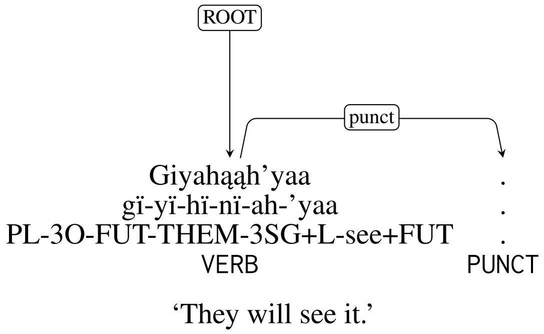


Figure 4: Verb with pronominal object marker and no free noun phrase.

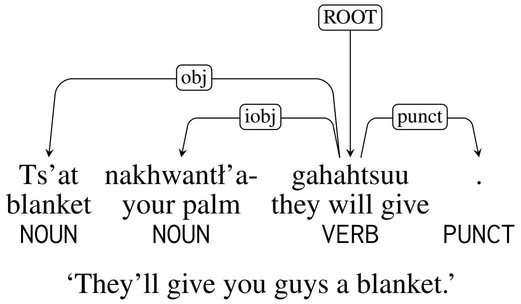

ROOT
obj
iobj
punct
Ts'at nakhwantł'a- gahahtsuu .
blanket your palm they will give
NOUN NOUN VERB PUNCT
'They'll give you guys a blanket.'

Figure 5: An alternative annotation splitting the complex verb (cf. Figure 1).

in Figure 3, the pronominal object marker *-yï-* is absent, whereas in Figure 4, when the free noun phrase is not present, the pronominal marker appears.

This presents challenges for annotation schemes that require core arguments to be lexical (i.e., separate words), as stipulated by the Universal Dependencies (UD) framework. Although UD allows subwords, or 'syntactic words' as units of analysis, word segmentation for Gwich'in is nontrivial. Similar concerns have been raised by others working on Athabascan and other polysynthetic languages, and multiple proposals for segmenting verbs in polysynthetic languages exist (Spence et al., 2017; Park et al., 2021; Tyers and Mishchenkova, 2020). The present treebank does not introduce word segmentation, as illustrated in Figures 3 and 4. However, we include detailed glosses to help future revisions that would segment the complex words following a consensus between the alternatives.

Another distinction between Southern and Northern Athabascan languages is the phenomenon of noun incorporation. It has been suggested that northern Athabascan languages allow a noun to be incorporated as part of the verb word, while Southern Athabascan languages do not. This may be because of the obligatory pronominal object marker on the verb in Southern Athabascan languages (Gelderen and Willie, 2012). The linguistic data used for this treebank does not appear to have incorporated direct objects on the verb; however, given the limited data and Gwich'in proficiency, we do not make a claim to the status of incorporated direct objects in Gwich'in. However, the corpus does contain example sentences with incorporated noun recipients in ditransitive verb themes. This is well attested and documented in neighboring languages such as Upper Tanana (Lovick, 2020, 2023). For instance, the verb in Figures 1 and 2 has the incorporated possessed noun 'palm' to indicate the recipient of the fabric object. A literal translation given for this sentence would be "They will position a blanket at rest in all your hands" (Bushey, 2021). Two separate annotations for this sentence can be seen in Figure 1 and Figure 5, one with and one without splitting the verb.

4.2 Postpositions

Adpositions follow their complement in Gwich'in. There is a three-way distinction in the use of these postpositions depending on whether the object of the postposition is a non-human, human, or areal noun (Marlow and Garnett, 1996). In other words, postpositions must agree with their object. If the object is human, the postposition is inflected for person and number. The pronominal prefixes used for this are identical to the direct object and possessor prefixes. If the object is a noun referring to an area, the postposition takes the areal prefix *gwï-*. Postpositions that modify all other nouns are unmarked. An example of postpositions with non-human, human, and areal objects is shown in Figures 6, 7, and 8 respectively, where the postposition *zhit* ‘inside’ is compared in all three instances.

Objects that are personified in myth and folklore may be treated as human as demonstrated in Figure 9.

When a non-areal object is unspecified in the sentence but understood as in the sentence in Figure 10, the postposition takes the third person object prefix. This can result in an unconventional dependency relation between the predicate of the sentence and the postposition of the unspecified object as is shown in Figure 10. The dependency relation oblique is normally used for a nominal argument or adjunct of a verb, adjective, or other adverb, but here it is used for the adposition of a dropped pronoun object. The syntactic analysis of Figure 7 and the relation between the root verb and its dependent is therefore preserved when the nominal argument is omitted as in Figure 10. Notice the similarity and difference between Figure 7 and Figure 10 when the object of the postposition is dropped.

4.3 Lemmas

All Athabascan verb stems are themselves morphologically complex, which consist of a verb root and aspectual suffixes. The root provides information about the meaning of a verb, while the aspectual suffixes provide information about its tense. In many Athabascan languages, the root and the aspectual suffixes have fused so tightly together that the same verb has different stems for each aspect (Lovick, 2020).

Take for instance the verb sentence *Ch’ąąhtsii* ‘She is picking berries’ and *Ch’įįłtsąįį* ‘She picked berries’ which can be seen in example 2 and example 3 below. This verb takes the classifier Ł subject prefixes and has the stem *-tsii* in the imperfective. In the perfective, the verb has the stem *-tsąįį*. Some verbs, including this one, also take the prefix *dhï-* in the perfective. Because of this, the third-person singular subject prefix surfaces as *-ił* and not *-ah* in the perfective.

(2) *ch’ï-* *nï-* *ah-*
3O.INDF THEM 3SG+Ł
tsii
pick-berry.IMPFV
‘She is picking berries’

(3) *ch’ï-* *nï-* *dhï-* *ił-*
3O.INDF THEM CNJ 3SG+Ł
tsąįį
pick-berry.PFV
‘She picked berries’

However, not all Gwich’in verbs that take the *dhï-* prefix do so in the perfective; some take it in the imperfective. No verbs take it in both.

The future *Ch’ahąąhtsyàa* ‘She will pick berries’ has the stem *-tsyàa*, as seen in example 4.

(4) *ch’ï-* *hï-* *nï-* *ah-*
3O.INDF FUT THEM 3SG+Ł
tsyàa
pick-berry.FUT
‘She will pick berries’

Given the morphological complexity of the verb in Gwich’in described throughout, the task of determining lemmas is not an easy one. Native speaker and teacher Lillian Garnett describes this challenge as follows: “While there are some basic tendencies, there are no rules to predict how the stem will change. All we can do is memorize the past stem (along with the present and future stems) as part of the word” (Marlow and Garnett, 1996).

Previous lexicographic work for Gwich’in uses third-person singular forms for all prefixes for all dictionary entries. Both imperfective and perfective forms are usually given. We follow this tradition of using third-person singular forms for lemmas in this treebank. This includes all parts of speech that require person inflection such as verbs and inalienable nouns. Following the Universal Dependencies framework, we also use one canonical form for each word. However, changes may need to be made to account for the verb stem variation characteristic of Gwich’in and other Athabascan languages described above.

5 Conclusion, Limitations and Future Work

The development of the UD_Gwichin-TueCL treebank is an important step towards expanding the representation of diverse languages in the Universal Dependencies Framework. However, several limitations need to be acknowledged.

First, the corpus is relatively small, with only 313 annotated sentences. This may limit the scope of syntactic and morphological phenomena represented in the treebank. While we attempted to

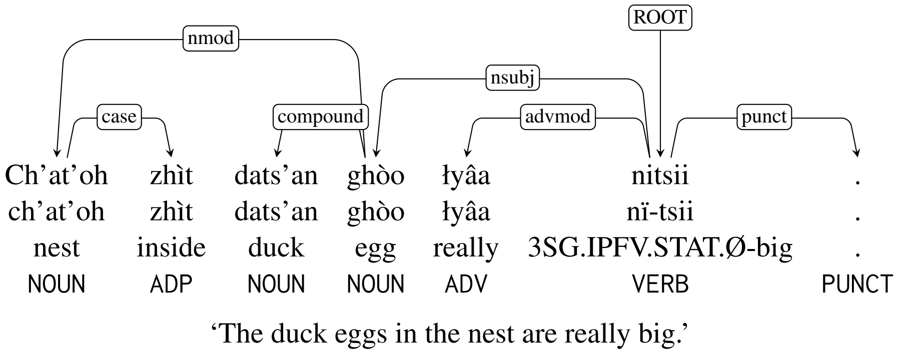

ROOT
nmod
nsubj
case
compound
advmod
punct
Ch’at’oh zhìt dats’an ghòo łyâa nitsii .
ch’at’oh zhìt dats’an ghòo łyâa nï-tsii .
nest inside duck egg really 3SG.IPFV.STAT.Ø-big .
NOUN ADP NOUN NOUN ADV VERB PUNCT
‘The duck eggs in the nest are really big.’

Figure 6: Postposition *zhit* ‘inside’ with inanimate object.

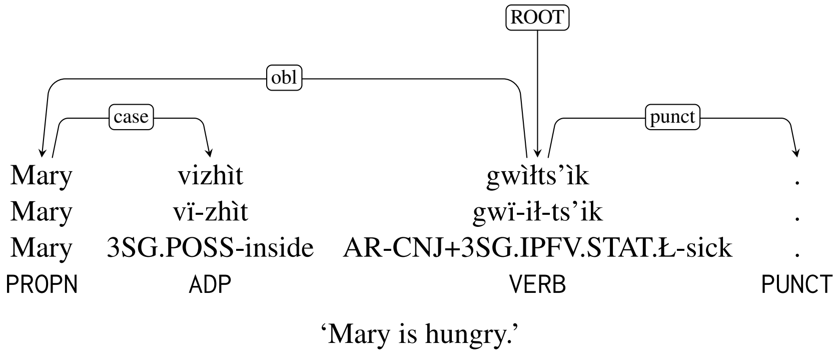

ROOT
obl
case
punct
Mary vizhìt gwìłts’ìk .
Mary vï-zhìt gwï-ił-ts’ik .
Mary 3SG.POSS-inside AR-CNJ+3SG.IPFV.STAT.Ł-sick .
PROPN ADP VERB PUNCT
‘Mary is hungry.’

Figure 7: Postposition *zhit* ‘inside’ with human object.

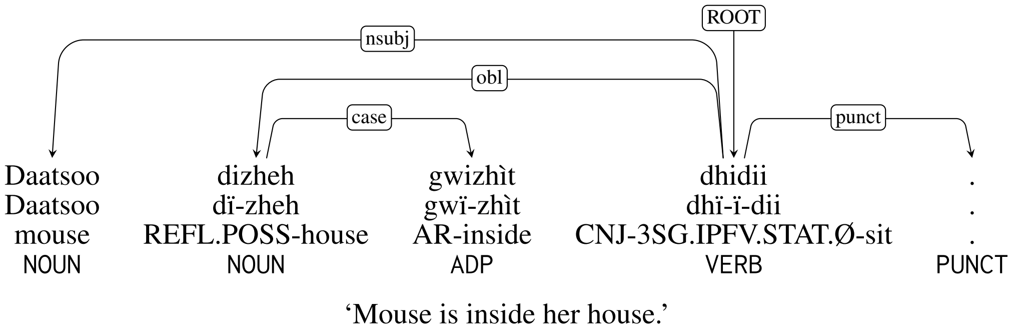

ROOT
nsubj
obl
case
punct
Daatsoo dizheh gwizhìt dhidii .
Daatsoo dï-zheh gwï-zhìt dhï-ï-dii .
mouse REFL.POSS-house AR-inside CNJ-3SG.IPFV.STAT.Ø-sit .
NOUN NOUN ADP VERB PUNCT
‘Mouse is inside her house.’

Figure 8: Postposition *zhit* ‘inside’ with animate (non-human) object.

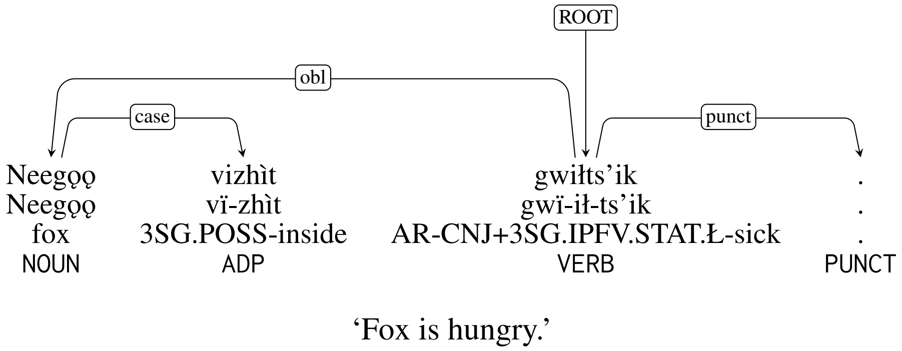


Figure 9: Human-form of postposition used with non-human object because of personification.

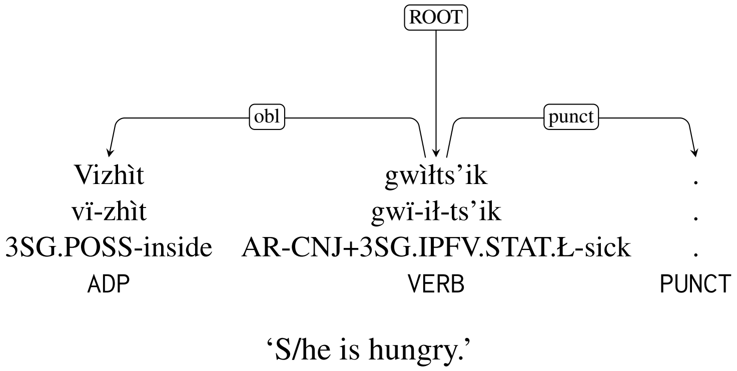


Figure 10: Postposition is promoted due to ‘elided’ noun phrase object.

include a diverse range of phenomena in the treebank, the scarcity of linguistic data and information on the language made this a challenging task. Future expansions of the corpus will benefit from additional data sources including native speakers.

Second, the polysynthetic and morphologically complex nature of the language poses challenges for annotation. In particular, the widely held theoretical position among Athabascanists is to treat pronominal prefixes as core arguments of the verb (Jelinek, 2014; Hale, 2003; Lovick, 2023). Combined with challenges of determining morph(eme) boundaries, this poses some challenges to the UD framework, which requires core arguments to be (syntactic) words. As more North American languages are annotated in UD, it remains to be seen whether the annotation scheme would need major changes to adequately describe languages with verb-internal syntax such as Gwich’in.

Third, the present treebank was annotated by a non-native second language learner of Gwich’in. While every effort was made to ensure accurate annotations, the absence of native speakers in the construction of the treebank may result in inaccuracies and incomplete representations of the language. Furthermore, annotation decisions are those of a single annotator. Therefore, inter-rater agreement scores cannot be obtained at this time. Future work on the language will benefit from collaboration with native speaker consultants as well as additional annotators.

This treebank was developed to contribute to the Universal Dependencies project as well as the study of Athabascan morphology and syntax. We also hope that it will be instrumental for the Gwich’in community in their efforts to develop tools that support language revitalization. It is also hoped that this treebank will not only engender interest in Athabascan languages and North American Indian languages more broadly in the field of Natural Language Processing, but also that Athabascanists and others working with endangered languages will be encouraged to contribute more treebanks to the Universal Dependencies project in the future.

References

Alaska Native Language Archive. n.d. Gwich'in collection - Alaska Native Language Archive (ANLA). Accessed: 15-Oct-2024.

Melissa Axelrod. 1990. Incorporation in Koyukon Athapaskan. *International Journal of American Linguistics*, 56(2):179–195.

John Busch. 2000. Finding your way through a story: Direction terms in Gwich'in narrative. Research paper (M.A. in Anthropology), University of Alaska Fairbanks. Accessed: 15-Oct-2024. 42 pp. in English with some Gwich'in vocabulary.

Scott T Bushey. 2021. *Western Gwich'in Classificatory Verbs*. University of Alaska Fairbanks.

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

Doyon Foundation. n.d. Doyon languages online. Accessed: 15-Oct-2024.

Elly van Gelderen and Mary Willie. 2012. Are there null arguments in Athabascan? *Acta Linguistica Hafniensia*, 44(2):227–245.

Gwich'in Social and Cultural Institute. n.d.a. Gwich'in language store. Accessed: 15-Oct-2024.

Gwich'in Social and Cultural Institute. n.d.b. Gwich'in online dictionary. Accessed: 15-Oct-2024.

Kenneth Hale. 2003. On the significance of Eloise Jelinek's pronominal argument hypothesis. *Formal approaches to function in grammar: In honor of Eloise Jelinek*, pages 11–44.

Eloise Jelinek. 2014. Empty categories, case, and configurationality. In *Pronouns, Presuppositions, and Hierarchies*, pages 7–40. Routledge.

Michael Krauss. 1992. The world's languages in crisis. *Language*, 68(1):4–10.

Olga Lovick. 2020. *A grammar of Upper Tanana, volume 1: Phonology, lexical classes, morphology*. University of Nebraska Press.

Olga Lovick. 2023. *A grammar of Upper Tanana, Volume 2: Semantics, Syntax, Discourse*, volume 2. U of Nebraska Press.

Patrick Marlow and Lillian Garnett. 1996. Beginning Athabaskan Gwich'in ANL142. `https://www.uaf.edu/anla/record.php?identifier=KU985G1996b`. Accessed: October 22, 2023.

C. Mishler and K. Frank. 2019. *Dinjii Vadzaih Dhidlit*. IPI.

Dick Mueller and Lillian Garnett. March 1994. Western Gwich'in topical dictionary. Accessed: 15-Oct-2024.

Hyunji Hayley Park, Lane Schwartz, and Francis Tyers. 2021. Expanding Universal Dependencies for polysynthetic languages: A case of St. Lawrence Island Yupik. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 131–142, Online. Association for Computational Linguistics.

Katherine Peter. 1979. *Dinjii zhuh ginjik nagwan tr'iltsaii: Gwich'in junior dictionary*. Alaska Native Language Center.

Justin Spence, Zoey Liu, Kayla Palakurthy, and Tyler Lee-Wynant. 2017. Syntactic annotation of a Hupa text corpus. *Working papers in Athabaskan languages*, pages 37–53.

Francis Tyers and Karina Mishchenkova. 2020. Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 195–204, Barcelona, Spain (Online). Association for Computational Linguistics.

Francis M. Tyers, Mariya Sheyanova, and Jonathan North Washington. 2017. UD Annotatrix: An annotation tool for Universal Dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 10–17, Prague, Czech Republic.

United Nations General Assembly. 2019. Resolution A/RES/74/135: International Decade of Indigenous Languages 2022-2032. Accessed: 15-Oct-2024.

Roberto Zariquiey, Arturo Oncevay, and Javier Vera. 2022. CLD^2 language documentation meets natural language processing for revitalising endangered languages. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 20–30, Dublin, Ireland. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Jephtey Adolphe, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Arofat Akhundjanova, Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Antonios Anastasopoulos, and 693 others. 2025. Universal dependencies 2.16. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Gwich'in Specific Features and Values

Values:	Ø	L	Ł	D

Table 2: Values for feature `Classifier`.

Value	Description
1	stick-like
2	food
3	cloth-like
4	plural/rope-like
5	animate/dead
6	open container
7	sack of granules, enclosed/sheathed
8	compact
9	deteriorated

Table 3: `StemClass` values key.

Quid verbumst?
Applying a definition of word to Latin in Universal Dependencies

Flavio Massimiliano Cecchini
KU Leuven – Erasmushuis, Blijde-Inkomststraat 21, 3000 Leuven, Belgium
`flaviomassimiliano.cecchini@kuleuven.be`

Abstract

Words, more specifically "syntactic words", are at the centre of a dependency-based approach like Universal Dependencies. Nonetheless, its guidelines do not make explicit how such a word should be defined and identified, and so it happens that different treebanks use different standards to this end. To counter this vagueness, the community has been recently discussing a definition put forward in (Haspelmath, 2023) which is not fully uncontroversial. This contribution is a preliminary case study that tries its hand at concretely applying this definition (except for compounds) to Latin in order to gain more insights about its operability and groundedness. This is helped by the spread of Latin over many treebanks, the presence of good linguistic resources to analyse it, and a linguistic type which is probably not fully considered in (Haspelmath, 2023). On the side, this work shows once more the difficulties of turning theoretical definitions into working directives in the realm of linguistic annotation.

1 Introduction

The notion of (syntactic) word is at the centre of a dependency-grammar approach like that of Universal Dependencies (UD; de Marneffe et al., 2021):[1] quoting from the project's guidelines, "dependency relations hold between *words* [...] the basic units of annotation are *syntactic* words". How this fundamental unit of morphosyntactic analysis can or needs to be defined, however, is still at the centre of debates and contrasting opinions in the community, and it is not directly confronted by the guidelines. In this sense, as part of the COST action "UniDive: Universality, Diversity and Idiosyncrasy in Language Technology" (Savary et al., 2024),[2] task 2.1 of the working group 2 (lexicon-corpus interface) has been devoting itself to the harmonisation of the definition of syntactic word across languages, gathering information through surveys and regular meetings, and presenting some results at the 3rd general meeting in Budapest in 2025. The main starting point for this endeavour has been established in a paper by Haspelmath (2023), where a clear-cut definition of word is given, based among others on his previous works on this topic (Haspelmath, 2017, 2021). The same author admits that his definition might result "unnatural" (Haspelmath, 2023, cf. §4.5), as it does not precisely overlap with how words have been traditionally identified by each respective language. In fact, even inside UD and UniDive's communities, but not limited to them, some approaches are being brought forth to simply "go beyond" the notion of word, cf. (Haspelmath, 2025) or the 1st proposed shared task on morphosyntactic parsing.

The definition of word is a particularly acute issue for Latin in UD, as this language counts many treebanks managed by different teams and originating from different annotation standards. This fact, among other things, affects the notion of wordhood in each of these treebanks, so much so that efforts have already taken place in order to deal with Latin's morphosyntactic harmonisation (Gamba and Zeman, 2023b,a). The present paper wants to offer a more in-depth overview about "what is a word"[3] for what concerns Latin, tapping from the experience of UniDive working group 2 task 2.1's survey, and moving along two main lines: on the one hand, sketching, on the basis of UD-treebank-driven empirical observations, what the characteristics of the "Latin word" are, and, on the other hand, putting to the test Haspelmath's (2023) framework. We believe that

---

[1] `https://universaldependencies.org`

[2] `https://unidive.lisn.upsaclay.fr`

[3] Hence the title *quid verbumst* 'what is a word', with the univerbation *verbumst* for *verbum est*, cf. Section 7.

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 174–185
August 27, 2025 ©2025 Association for Computational Linguistics

such a case study can provide useful insights to its actual operability, both for the harmonisation of Latin treebanks and of all treebanks in general. Latin has the advantage of presenting a good amount of data in UD, of possessing many valid resources for its linguistic analysis, and also of bearing some typological characteristics which we feel are not fully taken into consideration in (Haspelmath, 2023). In this study we only refrain from tackling compounding, as we deem the scope of that phenomenon to be worthy of another entire, dedicated study, which can profit from the definition of "simple" words first.

Further, despite being centered on Latin, this work makes scripts available (see Section 3) which can be directly, or with little adjustments, applied to any other UD treebank, in order to replicate most of the results shown here for other languages. In fact, it is our hope that the scope of this paper[4] be widened and joined by similar case studies so as to create an as cross-linguistic as possible framing of Haspelmath's (2023) definition. UD is a collective endeavour and we contribute to it with the language we know best in its context: Latin.

Section 2 constitutes a brief linguistic profile of Latin, while Section 3 briefly describes the current situation of Latin in UD and the data used in this study; Section 4 comments on Haspelmath's (2023) definition of word; Sections 5 and 6 bring forth the analysis; Section 7 makes some final remarks and Section 8 concludes.

2 Latin: the language

We supply here a concise linguistic profile of Latin in order to make the following discussion easier to follow for those not well acquainted with this language. From the historical point of view, the Latin language emerged from the Latino-Faliscan branch of Italic, as spoken by the tribe of *Latini* (Latins) in *Latium*, roughly corresponding to the modern-day *Lazio* region of Central Italy.[5] As such, it is a typical, as it were, ancient Indo-European language showing an extensive fusional morphology in all major word classes (nominals, modifiers, verbs).[6] Inflection of nominals (including "nouny" adjectives/determiners, cf. Stassen, 2003, §9.1, §15) takes place through suffixes encoding at the same time case, grammatical gender and number, and is traditionally subdivided into five "declensions" for nouns and two "classes" for adjectives according to the similarities between sets of suffixes (concerning a "thematic vowel" or the absence thereof); as discussed in Sections 5 and 6, it is quite rare to find a nominal element without any inflectional affixes. Verbs (including auxiliaries) also inflect mostly fusionally for aspect, mood, number and person of the subject, tense, and voice, and can also take on nominal forms with their own paradigms (cf. Cecchini, 2021 for an overview from the point of view of UD). Verbal inflection is traditionally subdivided into four main "conjugations" (again according to the presence or absence of a "thematic vowel"), but aspectual distinctions are often much more complex from a morphological point of view (see e. g. Pellegrini, 2023), and this might be a consequence of the whole verbal system shifting from being aspect-based to tensed in preliterate times (cf. again Stassen, 2003, §10.2). Word derivation is, just like inflection, predominantly suffixal, especially with regard to changes in word classes (e. g. deverbal nouns etc.), while prefixation seems to be restricted to the expression of more "lexical" categories such as *Aktionsart* (cf. Haverling, 2000 for the verbal system) and degree. This all correlates well with the original dominant verb-final, more precisely SOV, word order of Latin, which however seems to already show the signs of a shift towards the later Romance SVO order in historical times (for example, adpositions are chiefly *pre*positions; cf. Adams, 1976), and can vary noticeably for pragmatic and information-packaging reasons anyway. For finer-grained details on Latin syntax, and much more, we refer to the state-of-the-art, comprehensive, monumental work by Pinkster (2015, 2021).

3 Latin: the data

Latin's presence following UD's annotation formalism is currently[9] articulated over six treebanks in UD proper, and at least two other relevant resources,

---

[4] With some reason defined by a reviewer a "position paper".

[5] Comprehensive references about this topic in general and for the relation of Latin with other Italic languages are (Clackson and Horrocks, 2007), and the Introduction of (de Vaan, 2008).

[6] Under the class of modifiers, what are traditionally called adverbs are problematic in that they are systematically treated as distinct uninflectable lexemes even when they are transparently and productively derived from some base (Cecchini, 2024).

[9] As of v2.15, released in November 2024 (Zeman et al., 2024).

Name/Code	References	Syntactic words	Syntax	`InflClass`
CIRCSE	(Iurescia et al., 2024)	18 968	Yes	Yes
IT-TB	(Cecchini et al., 2018; Passarotti, 2019)	450 517	Yes	Yes
LLCT	(Cecchini et al., 2020a)	242 411	Yes	Partially
Perseus	(Bamman and Crane, 2011)	29 221	Yes	No
PROIEL[7]	(Haug and Jøhndal, 2008; Eckhoff et al., 2018)	205 566	Yes	No
UDante	(Cecchini et al., 2020b)	55 519	Yes	Yes
Sabellicus	(Gamba and Cecchini, 2024)	10 755	Yes	Yes
LASLA[8]	(LAS, 2024)	1 820 405	No	Yes

Table 1: Overview of the annotated resources for Latin in UD or following UD's formalism used in this study. Syntactic words include punctuation marks.

as shown in Table 1 (where the figures for LASLA do not take into account the works already included in CIRCSE). Taken together, they reach 1 002 202 syntactic words only in UD (making Latin one of the most represented languages therein, coming out 9th out of more than 150) and 2 833 362 in total. The language in these annotated corpora, despite its extended chronology (from Antiquity to the Renaissance), is rather homogenous, if not for style and genre, from the morphosyntactic point of view, given how it is mostly oriented towards the Classical variety (cf. Pinkster, 2015, §1.7 for this periodisation) of the 1st c. BCE (Clackson and Horrocks, 2007, chs. V-VI), or at least it does not diverge too much from it (with the exception perhaps of LLCT, cf. Korkiakangas and Passarotti, 2011, §3). Their annotation, though, is not always homogeneous, as in some cases it is natively UD, in others the result of conversions from other annotation standards, but has been harmonised over time (see Gamba and Zeman, 2023b,a); in the case of LASLA, the original annotation does not even include syntax. With respect to our case study (Sections 5 and 6), we are also interested in the annotation of inflectional classes (UD's morpholexical feature `InfClass`), which is not always present.

Some technical remarks for the analyses in later Sections: whenever an analysis is performed, it is always intended to take place on the subcorpus formed by those resources on which it can fully apply (e. g. excluding LASLA if syntax is concerned); all lemmas and forms are normalised (in particular: lowercased, $v > u, j > i$); to avoid data sparsity, parts of speech have been reorganised partly following (Cecchini, 2024), in particular `PROPN` is subsumed under `NOUN`,[10] `NUM` under `DET`,[11] `ADV` redistributed according to the parts of speech of their bases;[12]; `DET`, `NUM` and `PRON` are considered synsemantic (or "functional", or "grammatical") word classes; by lexeme we mean all forms described by a unique couple of lemma plus part of speech.[13] All data cited in this work and the Python scripts to produce them (for Latin, but adaptable to any other UD language) are made available through a sharing platform,[14] mentioned in what follows as "the repository". Some are however shown in Appendix A for ease of reference.

4 Defining the word: MH and UD

Referring to (Haspelmath, 2023) for the complete discussion, we only highlight some main points here. Haspelmath's (2023), henceforth MH, and UD's frameworks use transversal, but partly overlapping and correlated categories. MH's definition is based on three morphological objects which represent themselves word types, i. e. free morphs, clitics, and roots, to which compounds are added as "second-order" formations. These objects are themselves based on the more fundamental notion of morph (Haspelmath, 2020), and on four combining fundamental properties, which we resume in our words as: contentfulness, boundness, selectivity[15]

---

[10] Cf. guidelines: "A proper noun is a noun (or nominal content word) that is the name (or part of the name) of a specific individual, place, or object. [...] Note that `PROPN` is only used for the subclass of nouns".

[11] Cf. guidelines: "Other words functioning as determiners".

[12] See list at `https://github.com/Stormur/OrderlyAdverbs/blob/main/Latin/ADV_omnia.tsv`. Note that we group `REL` with `PRON`, following the most conservative proposal in (Cecchini, 2024, §5.1.5).

[13] This still does not eliminate some further ambiguities, but it is a good enough approximation to identify morphosyntactic patterns.

[14] `https://github.com/Stormur/quidverbumst`

[15] "cannot occur on roots of different root classes."

	Contentful	Bound	Selective	Affixes
Free morph	-	No	No	No
Clitic	-	Yes	No	No
Root	Yes	-	No	-

Table 2: Morphological objects in the definition of word by Haspelmath, 2023, and their properties, if specified.

and presence of affixes. The relation between all these elements is summarised in Table 2. The notion of part of speech, a tenet of UD (de Marneffe et al., 2021, §2.2.2), is not explicitly mentioned by MH, but it is hinted at by his description of a root as "a morph denoting an action, an object or a property", which mirrors the subdivision of words into phrasal types at the core of UD (de Marneffe et al., 2021, §2.1.1); moreover, the notion of contentfulness is parallel to UD's distinction along the grammaticalisation cline between autosemantyc and synsemantic words (de Marneffe et al., 2021, §2.2.1), codified into pairs of part-of-speech tags sharing the same phrasal type, such as the modifiers `ADJ` vs. `DET`. In contrast, the notions of boundness and selectivity are not explicitly considered in UD's formalism. Since we are rooting our study in UD's formalism, we are abiding by its classification of words into parts of speech, and we are interested in seeing how these can be mapped onto MH's word types. In fact, we have to remark that UD's analysis does not reach the level of individual morphs,[16] so that we have to approximate MH's freamework at the level of lexemes, justified by the circumstance that the grouping into lexemes is (usually) centered around commonality of morphs. In MH's scheme, roots are actually generalised by and inferred from what he calls free forms, which possibly include other non contentful (= synsemantic) morphs, not necessarily just affixes. At first, free forms seem to overlap with the notion of sentence in UD, and so with that of main clause with no subordinates. In the following, we will try investigate what these objects look like and to what extent they are applicable for Latin.

5 *Formae liberae*, or Latin free forms

The fact that MH also considers elliptical free forms poses a first serious difficulty for us with respect to their identification in our data: if we assume that any argument of a clause can be "extracted" from it to form an elliptical version thereof, we have to consider all arguments in all sentences, and recursively any subordinated arguments of such arguments, not just whole sentences, as possible free forms. This is problematic to deal with both computationally and definitionally. For this reason, we decide to exclude elliptical constructions and limit our first analysis to free forms appearing as sentences coinciding with explicit simple clauses. Technically, we consider any sentence for which the subtree of its `root`[17] does not contain any edge labelled with a clausal relation[18] (in other words, we exclude sentences with subordinate clauses of any kind), and where all nodes except the `root` belong to synsemantic parts of speech,[19] thus excluding interjections (`INTJ`) as extra-lexical elements, unannotated words (`X`), and ignoring non-lexical elements (`PUNCT`, `SYM`). We ignore `parataxis`, and exclude sentences presenting a co-ordination (relation `conj`) at the `root`, to avoid elliptical structures: ellipsis is a pervasive phenomenon in similar constructions, with no clear preference in Latin for being left- or right-headed,[20] and it is unfortunately not (yet) signalled in UD's basic annotation. For the same reason, we exclude sentences presenting an `orphan` relation, which is the hallmark of ellipsis. We also require the feature `VerbForm` with value `Fin` to appear in the clause,[21] or, to accommodate possible "zero copulas", that there be non-clausal arguments (apart from `orphan`) or predicate modifiers.[22] Finally, words related by the "headless" relations `fixed` and `flat` are considered together as single units.[23] Contrary to MH's definition, and expanding it, we do not

---

[16] Despite efforts to define and operationalise an annotation in this sense, see e. g. (Gamba et al., 2024).

[17] We use the notation for dependency relations to distinguish UD's sense of root from MH's.

[18] These are `csubj`, `ccomp`, `xcomp`, `advcl`, `acl`.

[19] These are `ADP`, `AUX`, `CCONJ`, `DET`, `PART`, `PRON`, and `SCONJ`.

[20] Compare, in the same text in UDante, a left-headed ellipsis in `DVE-89` and a right-headed one in `DVE-313`.

[21] This requirement is partly Latin-specific, as main clauses are always headed by a finite predicate, or at least annotated this way following grammatical tradition.

[22] These are `nsubj`, `obj`, `iobj`, `obl`, `vocative`, `dislocated`, `advmod`, `discourse`.

[23] Though the impact of this is extremely limited in our data.

require the root to be autosemantic, for reasons discussed later.[24]

Out of a total of 60 194 syntactically annotated sentences in our corpus, 50 355 do not present a root co-ordination, and among these we find 1410, the 2.3% of all sentences, distinct (normalised) such free forms, headed by 606 distinct lexemes (for a lexeme/form ratio of 43.0%). Figure 1 (above) in Appendix A shows the distribution of the latter with respect to their parts of speech. As expected, we see that UD's main autosemantic parts of speech `VERB` (1001 types, 350 lexemes), `NOUN` (169, 118) and `ADJ` (142, 94) dominate: in fact, their definitions greatly overlap with those of MH's roots, as noted in (Haspelmath, 2023, p. 287), even if UD does not make semantic distinctions about the concreteness of their meanings, so that these classes will include more morphs in UD than in MH's scheme. We also cannot draw immediate conclusions about the boundness of these roots themselves: they are surely non-selective, in that they are independent from the actual realisations of their arguments and also from their linear order (at least in Latin this being only determined by pragmatic factors, cf. Spevak, 2010), but some of them, surely not all (cf. the single forms discussed below), might actually require some arguments to be expressed, e. g. a direct object. Such an investigation is beyond the scope of this paper, and here we can just note down that boundness emerges as an orthogonal property to UD's parts of speech `VERB`, `NOUN`, `ADJ` and `ADV`. There is a quantitative gap between these and synsemantic parts of speech, which can be partly explained with the more closed nature of the latter. The exiguity of `ADV`, besides their strong relational and metapredicative nature,[25] and of `AUX`, not expected to head a predicate by definition,[26] can also explain their positions after `PRON` and `DET`, this order otherwise being symmetrical to that of `VERB-NOUN-ADJ`.

An even greater divide beckons in the synsemantic field between `PRON`, `DET` and `AUX`, and the other parts of speech not appearing in even one of our free forms:[27] `ADP`, `CCONJ`, `PART` and `SCONJ`. This is also to be expected, because, beyond being synsemantic, these parts of speech are more or less explicitly defined as connectors, as exclusively grammatical elements introducing nominal or clausal arguments, or conveying morphological or pragmatic features (`PART`), so that it is not easy to imagine them as predicate heads. Under this light, we can consider the elements of these classes to be bound to the heads of the phrases they accompany, and so, passing to MH's framework, they are the prime candidates for clitics. In the Latin system, these words are also set apart from all others by a lack of inflectional affixes, that is, they are unique forms in their paradigms, even if this does not necessarily mean that they are bare morphs (see Section 6). But there is an even more striking characteristic that strengthens their identification with clitics: phonologically, many such words present endings which are not shared, or only marginally, by autosemantic classes, remarking their prosodic boundness. It is the case of `ADP` *sub* 'under', `CCONJ` *sed* 'but', `PART` *non* 'not' (Cser, 2020, §2). The interesting consequence is that Latin seems to have morphologically codified the distinction between roots and clitics, with intermediate degrees for non-clitic synsemantic elements (see Section 5.1): the question then arises how common this is typologically. In any case, the identification of roots (if we loosen the request for a concrete meaning) with `VERB`, `NOUN`, `ADJ` and possibly, partly `ADV` (intended *stricto sensu* "true", contentful adverbs such as *saepe* 'often' or *uix* 'with difficulty', and possibly relators like *supra* 'above') on one side, and of clitics with `ADP`, `CCONJ`, `PART` and `SCONJ` on the other side plausibly looks to be universal.

5.1 Not roots nor clitics

The position of the roots' synsemantic counterparts `AUX`, `PRON` and `DET`, in MH's system is less clear. Only to `AUX`, given its extremely limited ap-

---

[24] The results are presented in the repository as the file `freeforms.tsv`.

[25] Cf. (Cecchini, 2024, §4.2). Further, we notice that the two extracted free forms (all in ITTB, `dev-s530`, `dev-s533` and `train-s9004`) are actually faultily segmented subordinates of a previous sentence, and present ellipsis: so e. g. *quia non semper est* '[. . .] since there is not always [any]', where the existential `AUX` *est* should be promoted to head.

[26] We notice indeed that out of the 18 extracted free forms headed by the `AUX` *sum* 'to be', most come from PROIEL and should be annotated otherwise, cf. (Gamba and Zeman, 2023b, p. 11), and there are issues of sentence segmentation and ellipsis. The only viable free form would be *esto* '(so) shall it be' (PROIEL `78649`).

[27] The single occurrence of `ADP` *usque septies* 'up to seven times' (PROIEL `13811`) should actually have *septies*, a multiplicative numeral, as its head instead of *usque* in the current annotation standard. The two occurrences of `PART` are also due to faulty annotations.

pearance in clausal free forms, it comes easy to associate the status of clitic on the road to an affix, at least for Latin (and observed univerbation practices strengthen this claim, cf. Lehmann, 2020 about this process), when looking at its boundness, even though the verbal inflection of this class is a problem if clitics have to be bare morphs. The same argument could actually be repeated for `ADV`, in the sense that its members can only be associated to a predicate, and not head a predicate themselves, making them bound, but again, possibly bearing (derivational) affixes. It has to be admitted, though, that this conclusion is subject to the current annotation practices for `ADV`, which are far from being uniform or coherent (cf. Cecchini, 2024 and Section 3). As for `PRON` and `DET`, strictly speaking, they should not be counted as words in MH's system, as they are no roots, and in general not even free morphs (see Section 6). They are not affixes either, as their distribution is parallel to that of `NOUN` and `ADJ`, in the sense that they are no more selective than those. This grey zone seems to arise from what appears to be a mixing of purely distributional, i. e syntactic (boundness and selectivity) and semantic (contentfulness) criteria in MH's definition of word types, in particular the request of a "concrete" meaning for roots, criteria which however look to be orthogonal to each other. In other words: `AUX`, `PRON` and `DET` do not seem to distribute that much differently than their autosemantic counterparts, but, at least in Latin, they cannot be considered words according to MH's framework simply because their meanings are more abstract and less complex, but they happen to be inflected; at the same time, a good candidate for root, the contentful subset of `ADV`, has a more clitic-like behaviour than them. Section 7 will address this point further, after having a look at the distribution of affixes in Latin.

6 *Liberi morphoi*, or Latin free morphs

In MH's framework, a particular subclass of free forms are free morphs, i. e. free forms consisting of only one morph. Among the clausal free froms discussed in Section 5, only members of `VERB` (and in one case also `AUX`) appear as standalone forms (119 types, 97 lexemes), e. g. *obsecra* 'implore' (PROIEL `50017`), but these forms usually do not consist of single morphs. Given our chosen constraints, we do not observe other standalone representants of non-clitic parts of speech there. However, even if an explicit copula, so at least one extra morph, appears to be the most frequent strategy to from predicates for non-verbs (we record this 78.7% of the times for `NOUN`, 79.6% for `ADJ`, 62.3% for `PRON` and 78.9% for `DET`), zero copula is indeed possible in Latin, cf. (Stassen, 2003, p. 676),[28] as are other kinds of one-word utterances. Thus, in search for free morphs we can shift our attention to sentences consisting of only one syntactic word (barring the presence of alexical `PUNCT` and `SYM`), allowing us to broaden the analysis to our whole Latin corpus; the distribution of such single-word free forms across parts of speech is then also shown in Figure 1 (below) in Appendix A.[29] The picture is very similar to that of clausal free forms: beyond noise in the data, we see that `DET`, `PART` and `AUX` are relatively more represented, but, more interestingly, `INTJ` is now more relevant, with occurrences of e. g. *st*, *hem* and *attatae*. This is surely due to the contribution of LASLA, featuring comedies and more everyday language (also shown by the appearance of vocatives among nouns, 10 out of 115 form types), but it does highlight a class of utterances which cannot be interpreted as predicates, and whose members often have an unclear lexical status. They are in fact unanalysable forms, and we can take them as free morphs; incidentally, an English interjection (*ouch*) is also among MH's examples for this class. However, the key verdict is that, for Latin, free morphs substantially begin and end with interjections.

It is often stated that Latin has a rich morphology, and this can actually be quantified with the help of the available lexical resources. Using LatInfLexi (Pellegrini and Passarotti, 2018; Pellegrini, 2024), we can assess, for what regards `VERB` and `NOUN`, that only a small subclass of `NOUN` lexemes admits bare morphs as parts of their paradigms, e. g. *mel* 'honey', *sōl* 'sun', or *uir* 'man' (all having concrete meanings).[30] These forms can be considered the stems[31] of their respective paradigms, since

---

[28] Even if it is very hard to pinpoint and distinguish from elliptical structures in the current UD's annotation standards.

[29] Complete results are in the file `singleforms.tsv` in the repository.

[30] The complete list is in the file `radicalforms.tsv` in the repository.

[31] Admitting here for semplicity and without loss of generality only continuous stems. Discontinuous stems (cf. Bonami and Beniamine, 2021) would only slightly change the picture: so, for example, for *facio* 'to do, make' we would obtain *f_c*, and then we could posit that its simplest continuous realisation *fac*, the imperative 'do', is a free morph.

all others are obtained by suffixation (which is predominant in Latin), but this does not make them unmarked for grammatical categories: in fact, `Case` (nominative, and also accusative for *mel*), `Number` (singular), and partly `Gender` are completely predictable from their forms.[32] We identify 81 out of 1038 noun lexemes of this kind in LatInfLexi; however, free morphs among these are effectively only a quarter, as many are derived, e. g. *ēruptio* 'a breaking out', an action name form from `VERB` *ērumpo* 'to break forth', itself bearing the preverb *ē* 'out of'.[33] Thus, the magnitude of this phenomenon appears almost irrelevant for Latin, and even more so as we do not find any of similar free forms in our corpus. Two related, parallel questions arise here: whether there exist non-clitic lexemes bearing no inflectional, at most derivational (in MH's terms, "not required") affixes, i. e. uninflectable words; and, even if inflectable, how many underived lexemes, i. e. admitting at most inflectional (in MH's terms, "required") affixes, make up the Latin lexicon.

For the first question, we rely on the annotation of the feature `InflClass` 'inflectional class' in some of the Latin treebanks. From these, we extract all the lexemes which possess forms that are never annotated for an inflectional class.[34] Of the 2175 identified lexemes, we focus on `VERB`, `NOUN`, `ADJ`, their synsemantic counterparts `AUX`, `PRON`, `DET`, and `ADV`. After having excluded abbreviations and symbolic numerals, and filtered out some dubious or simply erroneous annotations through manual inspection,[35] we gather only a handful of contemplatable cases, that is, by their parts of speech:

`NOUN` *māne* 'morning', *here* 'yesterday', *crās* 'tomorrow'

`ADJ` *nēquam* 'worthless', *satis* 'enough', *uolup* 'pleasant'

`PRON` *quandō* 'when', *ubī* and *ibī* 'where', *unde* 'whence', *cūr* 'why', *nīl* 'nothing'

`DET` *ita* 'so', *tam* 'as much', *tot* and *quot* 'as many', *siremps* 'same', and derived forms; cardinal numerals above 3

`ADV` many "true adverbs" such as *ferē* 'approximately'; relators such as *suprā* 'above'

It is interesting to notice some patterns, especially about quantities and time indications, and that modifiers (`ADJ`, `ADV`, `DET`) appear to be slightly more represented than the other types, but the numbers are vanishingly small, and this group by all means constitutes an "inflectionally closed class".

To answer the second question, we turn to Word Formation Latin (WFL, Litta and Passarotti, 2019). From all the 34 277 lexemes found in our whole corpus, we use some heuristics to remove non-lexical elements such as members of `PUNCT`, `SYM`, symbolic numerals, abbreviations, and certainly derived forms such as multiplicative numerals (e. g. *quinquies* 'five times'), or forms expressing a `Degree` (e. g. *ocior* 'swifter', with the *-ior-* comparative affix but no base form with the sole *oc-* stem). We end up focusing on 27 728 lexemes, from which we filter out all those recorded by WFL as the outcome of derivational processes by prefixation, suffixation or compounding, obtaining thus 19 076 candidates. Since WFL cannot cover all words, we manually inspect a random sample of 100 candidates to evaluate its precision,[36] assessing it at 42%. Among unknown, derived words we have *aleo* 'gamester', *mentalis* 'mental', or the compound *tricubitus* 'three cubits long'. Applying this value, we guess that the underived (not necessarily uninflectable) lexemes are ca. 8011, or 23.4% of the total (including noise). Among them, `ADJ` *turpis* 'ugly' (stem *turp-*) or `VERB` *furnus* 'oven' (stem *furn-*). This means that, roughly speaking, almost three quarters of the Latin lexicon are derived from some more basic roots.

This leads us to the perhaps obvious conclusion that the prototypical Latin non-clitic word is inflected, and in most cases includes derivational affixes of some kind, even when it is not inflectable

---

[32] On the topic of predictability in Latin paradigms, even if for verbs, we refer to (Pellegrini, 2023).

[33] We notice that, at least etymologically, the stem variation *rump-*/*rup-* could be explained as the presence of still another affix, an imperfective nasal infix; cf. (Beekes, 2011, §12.1.5).

[34] The complete list is in the files `aclitica_la.tsv`, and with further data in `aclitica_pos_la.tsv`, in the repository.

[35] For example, the `ADJ` *ocior* 'swifter' not marked for the usual `InflClass=IndEurX` of comparatives, or *merito* 'being deserved = deservedly' not marked as a participial form of *mereo* 'to deserve' and instead annotated as an uninflectable `ADV`; on similar "contextual annotations" cf. (Cecchini, 2024, §5.1.1).

[36] The sample is in the file `underived_random100_la_analysis.tsv` in the repository.

(e. g. *nēquam*, with negative particle *ne-*, or *quandō*, with an ablative ending *-ō*). In MH's terms, we expect a Latin morph to be always accompanied by required and/or unrequired affixes, with free morphs occurring, but just as paradigmatic vagaries or not fully lexical elements (interjections).

6.1 Foreign words

We notice that, at this stage of analysis, we are ignoring the relevant category of terms of non-Latin, "foreign", origin, which would otherwise feature prominently among uninflectable words; most of these, not labelled for `InflClass`, are names, e. g. *dāuīd* 'David', but can also be common terms, e. g. *rabbi* 'teacher' (both from Classical Hebrew). However, such words cannot be really considered part of the Latin system as much as they are the remnant of a "failed" morphological integration process into Latin, possibly due to lack of analogy with other inflected words: we do see other foreign names being adapted, e. g. *iōannēs* 'John' (from Classical Hebrew through Ancient Greek) following the so-called third declension (here the inflectional affix is *-ēs*). Uncertainties in their treatment are seen also from the frequent choice of not annotating them in the treebanks (part of speech `X`, no features, `flat` relation), especially when in a non-Latin script (essentially, Greek). This category is thus problematic, because it is not clear when a word ceases to "belong" to one language and starts being part of another. To our ends, though, this is to some extent irrelevant: the uninflectability of these terms is a reaction to a different grammatical system, and not an internal evolution of Latin, at least not synchronically.

7 A good definition (for Latin)?

Trying to identify the morphological objects of MH's definition (see Table 2) for Latin has brought to light some issues.

The most immediate issue is that distinguishing a specific type of word for free morphs, as opposed to roots or clitics possibly appearing with affixes, is totally irrelevant for Latin, and actually for any language like Latin, that is, a language which tends to positively mark through inflectional morphology every term "denoting an action, an object or a property", independently from its semantic complexity (so, be it autosemantic or synsemantic), and where a good part of the lexicon, even uninflectable items, shows transparent extensive derivational processes. Only a fragmentary part of Latin forms can be identified as free morphs, but this does not really bring any insight into the nature of the Latin word. In fact, we get the impression that the inclusion of free morphs in MH's definition might be motivated, on the one hand, by the need to give a place to non completely lexical, difficult to tackle elements such as interjections, and, on the other hand, might be driven by the apparent focus on languages with different, if any, inflectional paradigms than Latin. In particular, we would like to stress the fundamental difference that incurs between a language which always marks some kind of grammatical category such as `Case`, `Number`, and/or `Gender` on the members of entire word classes, as is Latin, and other languages which only mark them in some circumstances and then possibly "additively", as it seems to be the case for English.

Even if we are not considering it in this study, we notice that MH's definition of compound will be directly affected by this parameter: if the members of some word class always occur with required affixes, we cannot expect them not to have some affix even when combining with other non-affixal morphs, and this is the case in Latin with the so-called linking vowel *-i-* (see e. g. Oniga, 1992; Brucale, 2012), so that we have to treat all such cases as instances of multiple, and not single, words. Conversely, we cannot expect such a thing as a linking vowel in languages with no required affixes, and so there we will treat most, if not all, morph combinations as single-word compounds. This, in our opinion, will then create an asymmetry in how and how many compounds are identified as single or multiple words in each language, which is ultimately based on a bias with respect to their respective inflectional types; this in turn, in our opinion, seems to run counter to the establishment of a definition which should apply universally with the same criteria, and which instead would take double standards, as it were. So we argue that it is not so much about deciding to always split forms or always lump them together, but that two cases like Latin *agricultor* 'farmer', from *ager* 'field' and *cultor* 'tender' with linking *-i-*, and *flowerpot*, cited in (Haspelmath, 2023), should both be treated as either one single word or two words, but not as two words the former (Latin) and one word the latter (English), as would now be the consequence of a direct implementation of MH's definition.

Another major issue is that, while the distinction between roots and clitics is partly retraceable in UD's part-of-speech system, so morphosyntactically grounded, and partly reflected in phonological and morphological features, it still seems to not be really related to the presence or absence or affixes (for example, if we admit that `AUX` is clitic, then clitics can be inflected; also, some grammaticalised forms transparently bear derivational affixes), or to the status of boundness. In MH's scheme, what is common to all three basic word types is non-selectivity, and this seems to be more relevant: words could then be tentatively defined as *nuclei* of morphs and affixes which do not always syntactically combine with elements of the same kind. However, on the one hand both boundness and selectivity are not made explicitly relative to linear order or syntactic dependencies in MH's definition; and on the other hand, if we suppose the latter criterion, we can find counterexamples. For example, leaving further detailed investigation about this aspect to future work, we notice that a class such as `ADP`, clitic almost by definition, at least in Latin, always selects a nominal head to depend on, meaning a `NOUN` proper, a verbal form acting as a nominal (`VerbForm=Vnoun`), such as an infinitive (cf. Cecchini, 2021), or a modifier, which in Latin follows nominal morphology. In this sense an `ADP` is selective, but we cannot reasonably treat it as an affix, and this because it is not necessarily adjacent to the nominal in question, while conversely `AUX` is a much stronger candidate for being an affix given its tendency to stay close to its head. Unfortunately, the semantic criterion discriminating roots form other morphs does not come to help here in pinpointing what makes all these elements words, as has been observed in Section 5.1, and should probably be one of the first criteria to be discarded.

8 Conclusion

The conclusion of this preliminary study, at least for what concerns Latin, is that, unfortunately, MH's definition of word is less clear-cut than at first glance when it comes to putting it into practice, and it makes distinctions which do not appear relevant for the Latin system, and we dare to say also for typologically similar languages under the inflectional aspect, as could be Ancient Greek, Russian, and others.[37] In our opinion, some criteria of analysis, such as contentfulness, are already captured by UD's system in an even more systematical approach, and they are anyway (partially) orthogonal to the morphosyntactic ones that we deem have to be at the core of any definition of word, and that are also used in MH's definition. We see a trace, though, that deserves being pursued, and that is just based on selectivity and rigidity of linear order: words, especially in a morphosyntactic context as UD's formalism, can then be those blocks, with a meaningful nucleus and a possible contour of affixes, which can be moved around or separated by other similar blocks inside any possible free form. In this way, a multipartite classification into morphological objects as shown in Table 2, which seems to be more relevant at a language-specific level (e. g. more for English than for Latin), would be a consequence of a more general, and in our eyes useful and polished, definition. A similar definition is probably not new, but would still need to be applied uniformly throughout UD; for Latin, for example, it would have consequences on how to treat preverbs, i. e. verbal prefixes formally identical to adpositions and usually considered to be part of the stem (as seen in Section 6 for *ērumpo* 'to break out' with respect to *rumpo* 'to break'), and would probably go against some of the current traditional practices of annotation (so, for preverbs, it would mean splitting them from their bases). This issue, together with those about univerbation and compounding in Latin, are a material for future work which will be based on this first investigation on the very nature of wordhood in Latin. Beyond Latin, we hope that the scripts for data analysis that we make available with this study will help gather more data on this topic also for other languages, in order to gain a clearer picture of where the definition of word should be headed in a universal framework. At the same time, we are convinced that the particular case of Latin has helped putting a debated and crucial definition under a new perspective.

---

[37] We are well aware that this notion of similarity is impressionistic and that instruments need to be developed to make it quantifiable: this is by the way one of the aims making part of the development of a "tongueprint" for the ERASMOS project, at least for what concerns Latin and Ancient Greek.

Acknowledgments

This work has been performed in the scope and with the support of the ERASMOS ERC-funded project (grant n. 101116087) at KU Leuven. Views and opinions expressed are those of the author only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

2024. LASLA/conllup.

J. N. Adams. 1976. A typological approach to Latin word order. *Indogermanische Forschungen (1976)*, 81:70–99.

David Bamman and Gregory Crane. 2011. The ancient Greek and Latin dependency treebanks. In *Language Technology for Cultural Heritage*, Theory and Applications of Natural Language Processing, pages 79–98, Berlin/Heidelberg, Germany. Springer.

Robert Stephen Paul Beekes. 2011. *Comparative Indo-European Linguistics*, second edition. Number 172 in Not in series. John Benjamins, Amsterdam, the Netherlands; Philadelphia, PA, USA.

Olivier Bonami and Sacha Beniamine. 2021. *Leaving the stem by itself*, pages 81–98. Number 353 in Current Issues in Linguistic Theory. John Benjamins, Amsterdam, the Netherlands.

Luisa Brucale. 2012. Latin compounds. *Probus – International Journal of Latin and Romance Linguistics*, 24(1 – On Romance Compounds):93–117.

Flavio Massimiliano Cecchini. 2021. *Formae reformandae*: for a reorganisation of verb form annotation in Universal Dependencies illustrated by the specific case of Latin. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 1–15, Sofia, Bulgaria. The Association for Computational Linguistics (ACL).

Flavio Massimiliano Cecchini. 2024. Let's Do It Orderly: A Proposal for a Better Taxonomy of Adverbs in Universal Dependencies, and Beyond. *The Prague Bulletin of Mathematical Linguistics*, (121):15–65.

Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. 2020a. A New Latin Treebank for Universal Dependencies: Charters between Ancient Latin and Romance Languages. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, pages 933–942, Marseille, France. European Language Resources Association (ELRA).

Flavio Massimiliano Cecchini, Marco Passarotti, Paola Marongiu, and Daniel Zeman. 2018. Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 27–36, Brussels, Belgium. The Association for Computational Linguistics (ACL).

Flavio Massimiliano Cecchini, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020b. UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020, Bologna, Italy, March 1–3 2021)*, pages 99–105, Turin, Italy. Associazione italiana di linguistica computazionale (AILC), aAccademia University Press.

James Clackson and Geoffrey Horrocks. 2007. *The Blackwell History of the Latin Language*. Blackwell Publishing, Malden, MA, USA.

András Cser. 2020. *The Phonology of Classical Latin*. Number 52 in Publications of the Philological Society. Wiley-Blackwell, Hoboken, NJ, USA.

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

Michiel de Vaan. 2008. *Etymological Dictionary of Latin and the other Italic Languages*. Number 7 in Leiden Indo-European Etymological Dictionary Series. Brill, Leiden, Netherlands; Boston, MA, USA.

Hanne Martine Eckhoff, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Haug, Odd Einar Haugen, and Marius Jøhndal. 2018. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation*, 52(1):29–65.

Federica Gamba and Flavio Massimiliano Cecchini. 2024. Sabellicus.

Federica Gamba, Abishek Stephen, and Zdeněk Žabokrtský. 2024. Universal Feature-based Morphological Trees. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 125–137, Turin, Italy. ELRA and ICCL.

Federica Gamba and Daniel Zeman. 2023a. Latin Morphology through the Centuries: Ensuring Consistency for Better Language Processing. In *Proceedings of the Ancient Language Processing Workshop associated with ehe 14th International Conference on Recent Advances in Natural Language Processing (RANLP 2023)*, pages 59–67, Varna, Bulgaria. Incoma.

Federica Gamba and Daniel Zeman. 2023b. Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 7–16, Washington, D. C., USA. Association for Computational Linguistics (ACL).

Martin Haspelmath. 2017. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 51(s1000 (Jubilee Issue: 50 Years Folia Linguistica)):31–80.

Martin Haspelmath. 2020. The morph as a minimal linguistic form. *Morphology*, 30:117–134.

Martin Haspelmath. 2021. Towards standardization of morphosyntactic terminology for general linguistics. In *Linguistic Categories, Language Description and Linguistic Typology*, number 132 in Typological Studies in Language, pages 35–58, Amsterdam, the Netherlands. John Benjamins.

Martin Haspelmath. 2023. Defining the word. *WORD*, 69(3):283–297.

Martin Haspelmath. 2025. Are "words" important for grammatical dependency analysis? handout. 3rd UniDive general meeting (January 29th, Budapest, Hungary).

Dag Trygve Truslew Haug and Marius Jøhndal. 2008. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34, Marrakesh, Morocco. European Language Resources Association (ELRA).

Gerd Vanja Maria Haverling. 2000. *On Sco-Verbs, Prefixes and Semantic Functions*. Number LXIV in Studia Graeca et Latina Gothoburgensia. Department of Languages and Literatures, University of Gothenburg, Gothenburg, Sweden.

Federica Iurescia, Federica Gamba, Flavio Massimiliano Cecchini, Francesco Mambrini, Giovanni Moretti, Marco Passarotti, and Paolo Ruffolo. 2024. UD_Latin-CIRCSE.

Timo Korkiakangas and Marco Passarotti. 2011. Challenges in Annotating Medieval Latin Charters. *Journal for Language Technology and Computational Linguistics*, 26(2):103–114.

Christian Lehmann. 2020. Univerbation. *Folia Linguistica*, 54(s41-s1-Historica):205–252.

Eleonora Litta and Marco Passarotti. 2019. (When) inflection needs derivation: a word formation lexicon for Latin. In Nigel Holmes, Marijke Ottink, Josine Schrickx, and Maria Selig, editors, *Lemmata Linguistica Latina*, volume I: Words and Sounds, pages 224–239. De Gruyter, Berlin, Boston. Interrogable online at `http://wfl.marginalia.it/`.

Renato Oniga. 1992. Compounding in Latin. *Rivista di Linguistica (currently: Italian Journal of Linguistics)*, 4(1 – The Morphology of Compounding):97–116.

Marco Passarotti. 2019. The Project of the Index Thomisticus Treebank. In Monica Berti, editor, *Digital Classical Philology*, number 10 in Age of Access? Grundfragen der Informationsgesellschaft, pages 299–320. De Gruyter Saur, Berlin, Germany; Boston, MA, USA.

Matteo Pellegrini. 2023. *Paradigm Structure and Predictability in Latin Inflection*. Number 6 in Studies in Morphology. Springer, Cham, Switzerland.

Matteo Pellegrini. 2024. LatInfLexi version 2.0.1. `https://zenodo.org/records/14438647`.

Matteo Pellegrini and Marco Passarotti. 2018. LatInfLexi: an Inflected Lexicon of Latin Verbs. In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018*, pages 324–329, Turin, Italy. Accademia University Press.

Harm Pinkster. 2015. *The Oxford Latin Syntax*, volume 1. Oxford University Press, Oxford, UK.

Harm Pinkster. 2021. *The Oxford Latin Syntax*, volume 2. Oxford University Press, Oxford, UK.

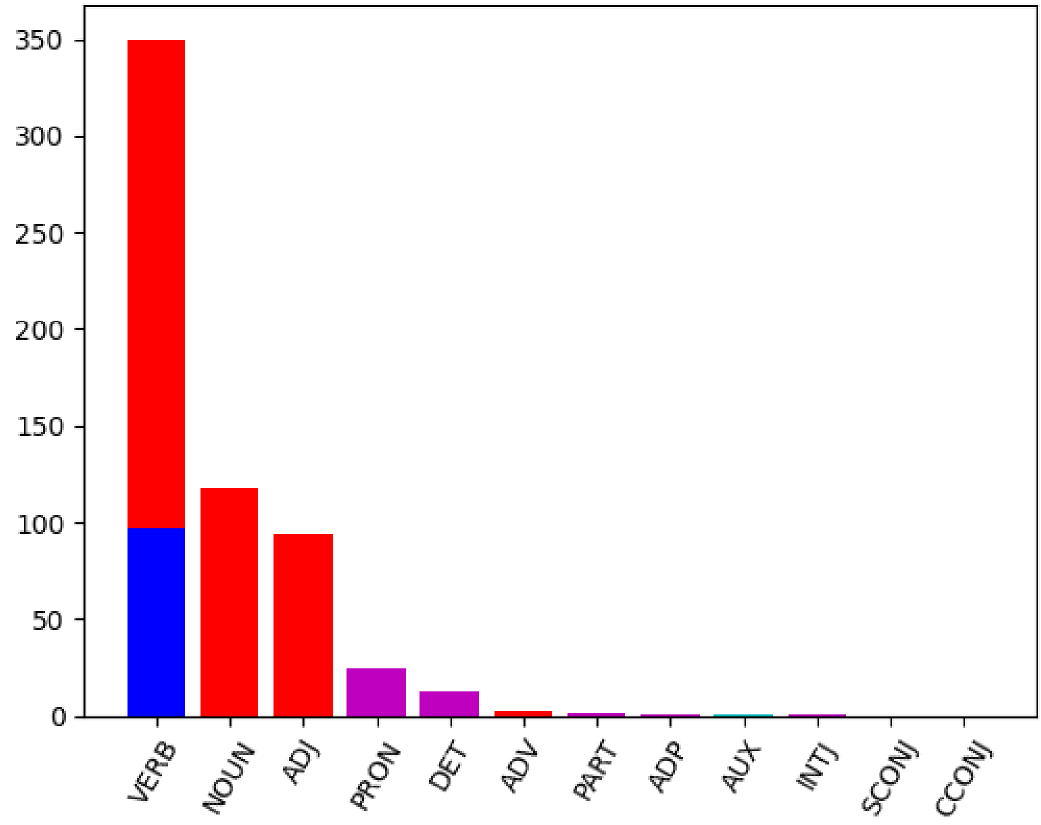
Agata Savary, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz, and Alina Wróblewska. 2024. UniDive: A COST Action on Universality, Diversity and Idiosyncrasy in Language Technology. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC - COLING 2024*, pages 372–382, Turin, Italy. ELRA and ICCL.

Olga Spevak. 2010. *Constituent Order in Classical Latin Prose*. Number 117 in Studies in Language Companion Series. John Benjamin, Amsterdam, the Netherlands.

Leon Stassen. 2003. *Intransitive Predication*. Oxford Studies in Typology and Linguistic Theory. Oxford University Press, Oxford, UK.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Arofat Akhundjanova, Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Matthew Andrews, and 633 others. 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.

A Distributions of free forms



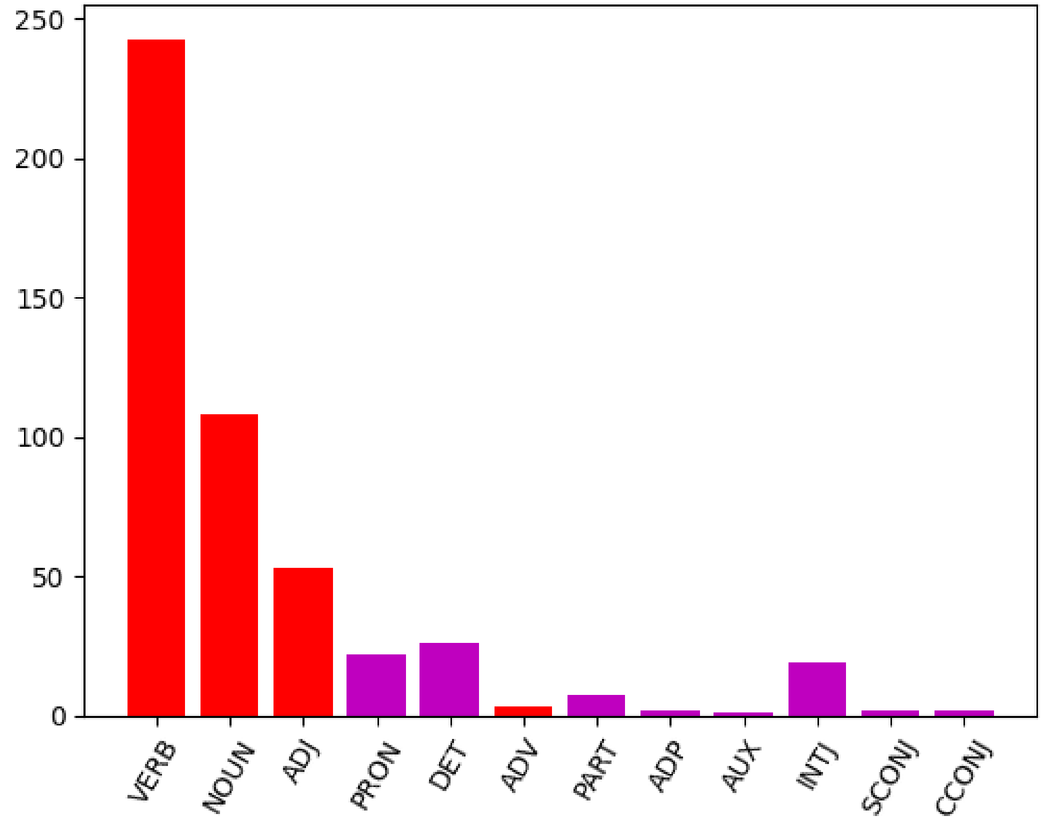


Figure 1: Distribution of parts of speech across our extracted free (above) and single (below) forms. Red is used for autosemantic and magenta for synsemantic classes; blue and cyan give the respective amount of one-word forms among the clausal free forms we define.

ShUD: the First Shanghainese Universal Dependency Treebank

Qizhen Yang
Shanghai World Foreign Language Academy
Shanghai, China
`qzyang.main@gmail.com`

Abstract

This paper introduces **ShUD**[1], the first Universal Dependencies (UD) treebank for Shanghainese, a Wu Chinese variant spoken by approximately 14 million people but severely under-resourced in NLP. The treebank is built through a scalable annotation pipeline that exploits grammatical parallels between Shanghainese and Mandarin. Our pipeline also provides a practical strategy for bootstrapping resources for other Chinese dialects. We documented syntactic phenomena unique to Shanghainese within the UD framework and fine-tuned a dependency parser[2] using our annotated treebank, contributing a foundation to both NLP tool development and cross-linguistic syntactic research.

1 Introduction

Shanghainese, the largest branch of Wu Chinese spoken by about 14 million of the overall 83 million Wu Chinese speakers (Pan et al., 1991; Xie, 2011), remains severely under-resourced in computational linguistics – an issue common among non-Mandarin Chinese varieties. To date, no annotated corpora exist for Shanghainese or any other Wu Chinese variety, in stark contrast to the growing availability of resources for Mandarin.

This lack of data hinders the development of NLP tools and limits linguistic research on Shanghainese. In this paper, we introduce **ShUD**, the first UD treebank for Shanghainese[3]. ShUD contains 983 sentences and 8,584 tokens. We design a sustainable annotation pipeline that leverages grammatical parallelism with Mandarin to improve annotation efficiency and quality. Our approach may extend to other Chinese varieties exhibiting similar syntactic traits. We document Shanghainese-specific constructions within the UD framework (De Marneffe et al., 2021). We also fine-tune a biaffine dependency parser using our data.

2 Related Work

2.1 Chinese Variants in the UD Project

Chinese is underrepresented in the UD project in both volume and variety. Existing treebanks focus almost exclusively on formal Mandarin (e.g., Poiret et al., 2023; Zeman et al., 2017), with rare exceptions for Cantonese (Wong et al., 2017) and Classical Chinese (Yasuoka, 2019; Yasuoka et al., 2022). To date, no other Chinese dialects have been included (Nivre et al., 2020).

2.2 Shanghainese and its NLP Resources

Although Shanghainese lacks annotated corpora, its grammar has been the subject of extensive linguistic study, and several grammar books (Qian, 1997; Zhu, 2006) and dictionaries are available. For instance, the *Shanghainese Dictionary* lists over 20,000 entries (Grayson, 2025). The Wu Chinese Society offers a comprehensive dictionary covering modern and historical Wu Chinese (Wu Chinese Society, 2009), including entries specific to Urban Shanghainese. Wu Chinese expressions are also available on Wiktionary (Wiktionary Contributors, 2024), though many reflect non-Shanghainese varieties.

Recent progress in speech processing has led to the release of a few colloquial Shanghainese corpora. Notably, Magic Data has published two speech corpora for Shanghainese conversation (Magic Data, 2021a,b).

2.3 Spoken Language Treebanks

The UD framework has expanded to include spoken language resources. These encompass di-

---

[1] `https://github.com/UniversalDependencies/UD_Shanghainese-ShUD`

[2] `https://huggingface.co/q1zhen/ShUD`

[3] Shanghainese includes several geographical and historical variants. We focus on Middle and New Period Urban Shanghainese; Old Period Shanghainese, used around a century ago, is no longer spoken.

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 186–193
August 27, 2025 ©2025 Association for Computational Linguistics

verse languages including Slovenian (Dobrovoljc and Nivre, 2016), Cantonese (Wong et al., 2017), etc., with many representing the only available UD resources for low-resource languages (for details, see Dobrovoljc, 2022). These treebanks exhibit considerable variation in transcription approaches, annotation principles, and treatment of speech-specific phenomena such as fillers, disfluencies, and repairs (Dobrovoljc, 2022).

3 Data Source and Features

We use the open-source *Scripted Chinese Shanghai Dialect Daily-use Speech Corpus* (ASR-SCShhiDiaDuSC, hereafter **A.-S.**) as the data source for our treebank. The corpus focuses on daily-use speech, providing an accessible and representative sample of contemporary Shanghainese[4] (Magic Data, 2021b).

The A.-S. corpus contains 4.23 hours of transcribed Shanghainese speech, totalling 4,819 utterances from 10 speakers. While originally created for speech processing, the corpus is well-suited for syntactic annotation. Speakers read Mandarin Chinese prompts aloud in Shanghainese, adapting vocabulary and structure naturally, which is the most common practice for native speakers to read the language from text in everyday life (due to the lack of standardised orthography and formal texts in Shanghainese). Given the high lexical overlap between the two languages in colloquial contexts, such adaptations produce fluent Shanghainese expressions rather than literal translations. Since Shanghainese is primarily used in colloquial contexts, the corpus is particularly well-suited to represent the language, and sentences are typically short. We use only the textual transcriptions for our annotation. An example instance from the dataset is shown below:

Prompt [Mandarin]	这点机会也可能没有。 (zhe dian ji hui ye ke neng mei you)[5]
Transcription [Shanghainese]	搿眼机会啊可能没了。[6] (geh ngae ci ue a khu nen meh leh)[7]
Gloss	This bit chance also possible no.
Translation	*There may not be any chance of this.*

4 Treebank Construction

4.1 Pipeline Overview

Although Shanghainese and Mandarin share many syntactic properties, off-the-shelf UD parsers perform poorly on Shanghainese due to major lexical differences, especially in the use of particles[8]. Leveraging the strong performance of existing Mandarin parsers, we designed a hybrid annotation pipeline: Shanghainese utterances are first manually transliterated into Mandarin vocabulary, then automatically tokenised and parsed by a parser. The original Shanghainese vocabulary is then restored, followed by thorough manual verification and correction at each stage. Each mapping is also saved for future reference to reduce manual workload. Since we are performing dependency annotations (in contrary to constituency), the relations are transferrable between the languages as the they do not contain structural information (De Marneffe and Nivre, 2019). We utilise Stanza (Qi et al., 2020) `v1.10.1`, which's Mandarin Chinese parser is trained upon the GSDSimp treebank[9].

This approach is especially advantageous. It significantly reduces the workload of annotators and boosts efficiency by reusing pre-trained models. The consistency can be improved because fatigue-related errors by annotators can be reduced.

Figure 1 shows the detailed pipeline. Section 6 presents the evaluations on the effects of different steps of the pipeline.

4.2 Inter-annotator Agreement

Annotation was conducted by two annotators fluent in both Shanghainese and Mandarin, each ex-

---

[4] Licensed under Creative Commons BY-NC-ND 4.0. Additional permission for derivative research was granted by Beijing Magic Data Technology Co., Ltd.

[5] In this paper, Mandarin Chinese words will be transcribed in Hanyu Pinyin without tone marks since precise tonal values are not essential in this project.

[6] There are standardised Shanghainese orthographies proposed by some scholars, but they are never widely accepted or used for a limited usage in writing. A common practice (which is also used by the corpus) is to use Mandarin words with similar pronunciations. In this paper, we will present the corpus texts as is and list out as many possible transcriptions as possible in theoretical discussion contexts.

[7] Similarly, existing standardised Romanisation schemes for Shanghainese are also rarely used, especially that pronunciations are rapidly evolving over time due to the influence of Mandarin. In this paper, the pronunciations, not guaranteeing their accuracy, will be mostly taken from Wu Chinese Society's dictionary; precise pronunciations are not essential in this project.

[8] For example, the Shanghainese word 伐 *veh* can function either as a negator (also transcribed as 勿 or 弗, equivalent to Mandarin 不 *bu*) or as a question particle without semantic content. In contrast, in Mandarin, the same character 伐 *fa* primarily means "to cut down [wood]" and is treated as a verb by standard parsers, resulting in incorrect POS tagging when applied to Shanghainese.

[9] `https://universaldependencies.org/treebanks/zh_gsdsimp/index.html`

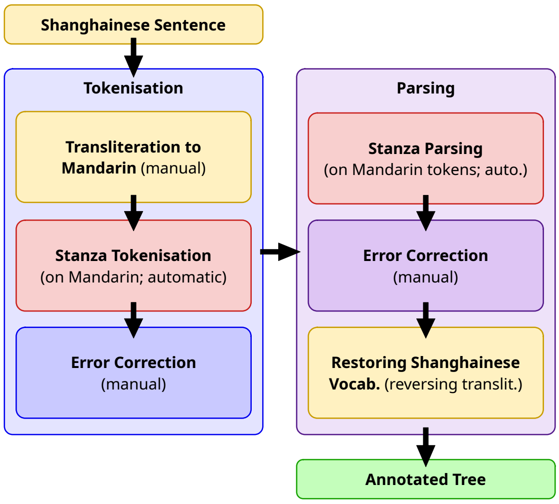

Shanghainese Sentence
Tokenisation
Transliteration to Mandarin (manual)
Stanza Tokenisation (on Mandarin; automatic)
Error Correction (manual)
Parsing
Stanza Parsing (on Mandarin tokens; auto.)
Error Correction (manual)
Restoring Shanghainese Vocab. (reversing translit.)
Annotated Tree

Figure 1: Overview of the annotation pipeline.

tensively trained in UD with a focus on Chinese-specific guidelines.

The first 100 sentences were independently annotated twice. Inter-annotator agreement reached 92.09% for UPOS tagging, 87.82% for Unlabelled Attachment Score (UAS), 81.89% for Labelled Attachment Score (LAS), and 99.09% for tokenisation. Most disagreements arose from differing interpretations of lexical ambiguity in contextually underspecified sentences. An example is provided in Appendix A. These cases are rare and do not substantially affect the overall annotation quality.

5 Annotation Guideline

This section outlines the principles and guidelines followed in the annotation process.

5.1 Text Segmentation and Tokenisation

Our tokenisation scheme follows the principles of the Penn Chinese Treebank (Xue et al., 2005), where a word, defined as one or more characters forming a lexical unit, serves as the basic unit of annotation.

Function words are treated as separate tokens, even when phonologically or morphologically attached to verbs. These include items such as 了 *leh*, marking the perfective aspect, and 勒 *leh*, indicating the continuation of an action.

Following the UD guidelines for Chinese[10], we treat compound words and multi-word expressions in Shanghainese similarly to idiomatic expressions in Mandarin Chinese (e.g., Chengyus). While Mandarin Chengyus are typically fossilised expressions from Classical Chinese, many idiomatic phrases in Shanghainese derive from Wu Chinese or from historical transliterations and reinterpretations of European words. These expressions function as lexicalised units and are not analysable by Shanghainese grammar.

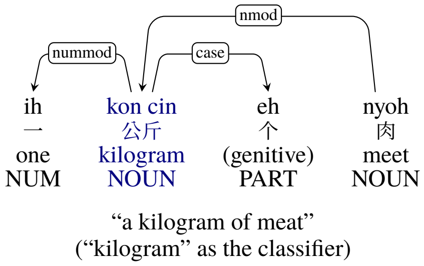


Figure 2: A classifier with genitive 个 *eh*.

If such multi-word expressions are mistakenly segmented during preprocessing, each component is assigned the same part of speech as the full expression and connected using the `goeswith` relation. In the final version of the treebank, they are merged and presented as a single token.

5.2 Linguistically Motivated Guidelines

This section highlights annotation decisions that differ from or are particularly noteworthy relative to the UD for Mandarin.

Nouns. Like Mandarin Chinese, words tagged as `NOUN` include regular nouns, classifiers, temporal nouns, position words, and localisers.

Classifiers can be pre-modified directly by `NUM` and `DET`. They have the feature `NounType=Clf`. In the case of having a numeral or determiner, the classifier is attached to it with a `clf` relation, and the numeral or determiner is then attached to the head noun. However, if the classifier does not come with a numeral or determiner, then the classifier would be the indefinite determiner with the noun as the head.

If there is a genitive 个 *eh* (also transliterated as 呃 or 额) between the classifier and the noun, then the classifier (with the numeral and genitive attached as a phrase) would be a `nmod` dependent of the head noun, as shown in Figure 2.

Temporal nouns, despite typically being the adjunct of verbs, are always tagged as a noun. They would have a `nmod` relation from the verb.

Verbal Polarity. Verbs can be negated by markers such as 勿 *veh* (also 伐, 弗) and 没 *meh*. Negat-

---

[10] `https://universaldependencies.org/zh/`

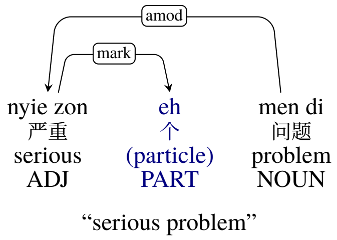


Figure 3: Particle 个 *eh* following an adjective.

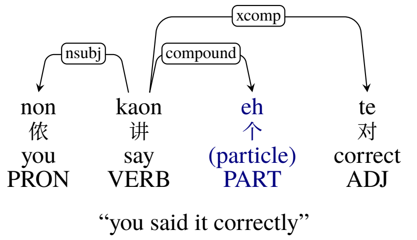


Figure 4: Particle 个 *eh* marking a complement.

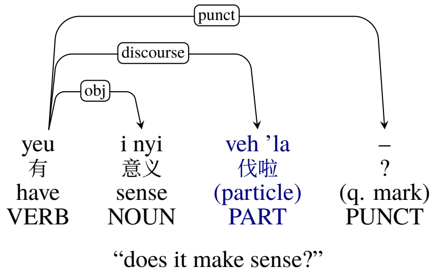


Figure 5: Sentence-final particles.

ors are excluded from lemmatisation, and the token is marked with `Polarity=Neg`.

Particles. The multifunctional particle 个 *eh* (also 呃, 额) corresponds to Mandarin 的, 地, and 得 (*de*), functioning as a genitive, relativise, nominalise, or adverbialiser. It is annotated with the `mark` relation when introducing dependent clauses or modifiers. See Figure 2 for its genitive use and Figure 3 for broader functions.

When 个 *eh* is used in an extent or descriptive construction (corresponding to Mandarin 得 *de*), a `compound` relationship is used, as specified in UD Chinese guidelines (`compound:ext`). In this case, it follows a verb, adjective, or adverb, then followed by a complement part. The complement is treated as a `xcomp` or `ccomp` dependent (depending on its subject) of the configuration’s head. The complement serves as the adverb if it is after a verb; it is similar to the latter clause in English’s “so… that…” construction if it is after an adjective or adverb. Figure 4 shows an example of this usage.

Another group of particles different from Mandarin Chinese are sentence-final particles. Common ones include 伐 *vah*, 了 *leh* (also 嘞), and 啦 *’la*. Combined use is also very common in Shanghainese, especially rhetorical questions, such as 伐啦 *veh ’la* (Myers, 2015). We treat them as a single token; however, if they are syntactically different (e.g., one indicating the end of a sentence and the other marking the question), then they would still be separated. Sentence-final particles are attached to the sentence via a `discourse` relationship. Figure 5 shows this usage.

	Singular	Plural
1st	吾 *ngu*	阿拉 *ah ’la*
2nd	侬 *non*	㑚 *na* (also 拿)
3rd	渠 *yi* (also 伊)	渠拉 *yi ’la* (also 伊拉)

Table 1: Personal pronouns in Shanghainese.

Pronouns. In Middle/New Period Urban Shanghainese, there are no polite forms of pronouns as in Mandarin. Table 1 shows the personal pronouns. Possessive case of the personal pronouns are constructed by appending the genitive particle 个 *eh*.

There are two demonstrative pronouns in Shanghainese. The proximal demonstrative is 搿 *geh* (also 葛; “this/these” or “here”). The distal demonstrative is 埃 *i* (also 伊; “that/those” or “there”). They also have some derived forms, such as 搿搭 *geh teh* “here”, 埃搭 *i teh* “there”, 埃面搭 *i mie teh* “there”, etc. We treat these words as single tokens.

Other rules. Other syntactic features are almost identical with Mandarin Chinese, and we thus primarily reference to Chinese UD guidelines. For words that do not have an exact correspondence in Mandarin, we consider their Mandarin synonyms with the same POS or structurally similar Mandarin constructions to determine the relationships.

6 Statistics and Pipeline Evaluation

Statistics of the treebank. The current treebank contains 983 sentences with 8,584 tokens. An example in ConLL-U format can be found in Appendix B. More sentences will be annotated in the next UD release.

34 relations and 15 UPOS tags are in the treebank. Among all tokens, 3,356 of them (approximately 40%) are mapped during the annotation using our pipeline, and we have collected 374 pairs of Shanghainese-to-Mandarin lexicon correspondences for 296 Shanghainese words. Appendix C shows detailed statistics on the treebank.

Experiment (stage of manual steps applied)	Segmentation		Parsing	
	Tokens	UPOS	UAS	LAS
Raw Sentence	70.69	52.29	30.48	24.97
+Tokenisation	100.00	61.07	51.84	39.15
+Tok. +Lexicons	100.00	92.66	81.54	73.97

Table 2: Evaluation of the effects of each stage of the pipeline, compared with the golden data (segmentation is evaluated using percentage accuracy). The CoNLL 2017 UD evaluation script[11] is used.

Pipeline evaluation. To evaluate the effectiveness of our pipeline, we assess the automatic parses on Shanghainese, produced by Stanza's Mandarin Chinese parser, at different processing stages using the first 100 sentences. The manually corrected annotations are treated as the golden data. Table 2 shows the results.

Feeding the unsegmented Shanghainese sentences directly to Stanza's Mandarin model (*Raw Sentence*) produces poor tokenisations, which in turn drags down UPOS tagging and dependency parsing. Correcting tokenisations (+*Tokenisation*) slightly improves the results; parsing is still directly on Shanghainese tokens by the Mandarin parser, and accuracies are still low. Replacing Shanghainese word forms with their Mandarin equivalents before parsing (+*Tokenisation* +*Lexicons*) yields a large improvement.

Manual tokenisation and lexical mapping substantially improve the accuracy, but the performance of the Stanza parser on Shanghainese remains considerably low, underscoring the need for a dedicated Shanghainese treebank to support the development of more accurate parsers.

Nonetheless, the pipeline still makes use of existing resources and improves annotation efficiency, especially in the early phase. The average annotation efficiency is around 50 sentences per hour, with almost a quarter of the sentences requiring no manual corrections other than lexical mapping.

7 Model Fine-tuning

We fine-tune a graph-based dependency parser on our annotated treebank using SuPar[12]'s implementation (Zhang et al., 2020) of the biaffine-based dependency parser by Dozat and Manning (2018). The model couples a biaffine scorer over head–dependent hidden states with `XLM-RoBERTa-large`, the pretrained multilingual model, as the contextual encoder.

We trained the model over 50 epochs with AdamW at a base learning rate of 5×10^{-5} with 10% warm-up, 20× scheduled decay, and gradient clip of 5. Batches are formed by sentence length bucketing with a maximum of 5,000 tokens per update and a 20-token fixed positional window. The treebank is randomly split into train, dev, and test with ratio of 80% (786 sentences), 10% (98 sentences), and 10% (99 sentences), respectively.

At the final checkpoint, the model reaches UAS of 75.61 and LAS of 64.91. The fine-tuned parser shows strong capacity in learning a robust representational foundation and significantly outperforms Stanza's Mandarin parser without manual correction. However, its capability in generalisation to unseen data is still limited, possibly due to the small size of training data. Appendix D shows more details of training.

8 Conclusion and Future Work

In this paper, we present the first UD treebank for Shanghainese, named **ShUD**. We propose a scalable annotation pipeline that leverages the strong performance of existing parsers and the substantial syntactic overlap between Shanghainese and Mandarin Chinese. We also fine-tuned a dependency parser and achieved considerable parsing accuracy.

Our parser could serve as a primitive foundation for future annotations and further automate the pipeline to reduce reliance on Mandarin parsers.

The treebank is still limited by its size and genre, with current data source solely based on scripted speech in everyday context. In future work, we plan to extend the treebank, by more annotations and possible expansions to social media text, written literature, news podcasts, etc., to develop more accurate models. Universal morphological features, beyond syntactic dependencies, can also be added to better support downstream tasks.

References

Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, pages 1–54.

Marie-Catherine De Marneffe and Joakim Nivre. 2019. Dependency Grammar. *Annual Review of Linguistics*, 5(1):197–218.

---

[11] `https://github.com/ufal/conll2017/blob/master/evaluation_script/conll17_ud_eval.py`

[12] `https://github.com/yzhangcs/parser`

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 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, Portorož, Slovenia. European Language Resources Association (ELRA).

Timothy Dozat and Christopher D. Manning. 2018. Simpler but More Accurate Semantic Dependency Parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 484–490, Melbourne, Australia. Association for Computational Linguistics.

Joel Grayson. 2025. Shanghainese Dictionary.

Magic Data. 2021a. ASR-CShhiDiaCSC: A Chinese Shanghai Dialect Conversational Speech Corpus.

Magic Data. 2021b. ASR-SCShhiDiaDuSC: A Scripted Chinese Shanghai Dialect Daily-use Speech Corpus.

Ethan Myers. 2015. Sentence final particles in Shanghainese: Navigating the left periphery. *Open Access Theses*.

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.

Wuyun Pan, SF Zhengzhang, RJ You, and Lien Chinfa. 1991. An introduction to the Wu dialects. *Journal of Chinese Linguistics Monograph Series*, (3):235–291.

Rafaël Poiret, Tak-Sum Wong, John Lee, Kim Gerdes, and Herman Leung. 2023. Universal Dependencies for Mandarin Chinese. *Language Resources and Evaluation*, pages 1–38.

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*.

Nairong Qian. 1997. 上海话语法 *(Shanghainese Syntax)*. 上海人民出版社 (Shanghai People's Press).

Wiktionary Contributors. 2024. Category:Wu language - Wiktionary, the free dictionary.

Tak-sum Wong, Kim Gerdes, Herman Leung, and John SY Lee. 2017. Quantitative comparative syntax on the Cantonese-Mandarin parallel dependency treebank. In *Proceedings of the fourth international conference on Dependency Linguistics (Depling 2017)*, pages 266–275.

Wu Chinese Society. 2009. 吴音小字典·吴语小词典 (Wu-Language Dictionary / Wu-Language Lexicon).

Yuwei Xie. 2011. Language and Development of City: The Linguistic Triangle of English, Mandarin, and the Shanghai Dialect. *Language*, 1:1–2011.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238.

Koichi Yasuoka. 2019. Universal Dependencies treebank of the four books in Classical Chinese. In *DADH2019: 10th International Conference of Digital Archives and Digital Humanities*, pages 20–28. Digital Archives and Digital Humanities.

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.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, and 43 others. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Yu Zhang, Zhenghua Li, and Zhang Min. 2020. Efficient Second-Order TreeCRF for Neural Dependency Parsing. In *Proceedings of ACL*, pages 3295–3305.

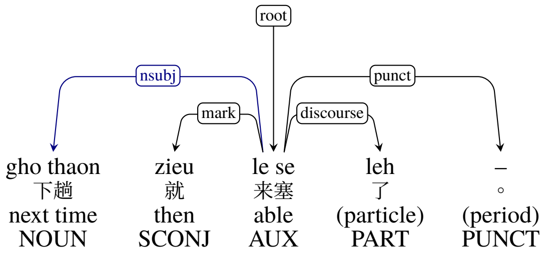
X. Zhu. 2006. *A Grammar of Shanghai Wu*. LINCOM studies in Asian linguistics. LINCOM Europa.

A Ambiguous Example

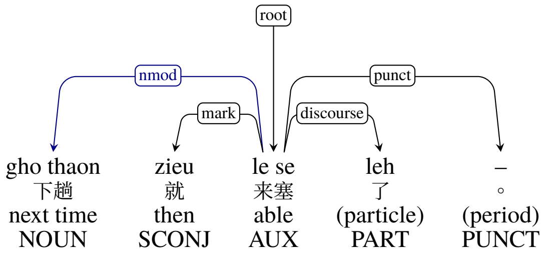
gho thaon	zieu	le se	leh	–
下趟	就	来塞	了	。
next time	then	able	(particle)	(period)
NOUN	SCONJ	AUX	PART	PUNCT

Figure 6: An example of an ambiguous sentence.

Take the sentence in Figure 6 as an example. The tokenisation and UPOS tagging are unanimous. However, when it comes to understanding the sentence, there could be a deviation. The sentence could be treated as a complete sentence, where the word 下趟 *gho thaon* "next time" is treated as the subject; it would have a `nsubj` dependency with the root. On the other hand, in a colloquial context, it could also be interpreted as a sentence omitting its subject, probably "I"; in this case, the word would be a temporal noun and has a `nmod` dependency with the root. This is shown in Figure 7.



"The next time will be okay."



"[I] will be okay next time."

Figure 7: Two possible interpretations that lead to different dependency labels of the ambiguous sentence.

B Example in ShUD Treebank in ConLL-U Format

Figure 8 shows an example.

C Detailed Treebank Statistics

The treebank spans across sentences with lengths from 2 tokens to 18 tokens, with an average of 8.73 tokens per sentence. Figure 10 shows the distribution of sentence lengths.

34 out of UD's 37 dependency relations are found in the treebank, excluding `expl`, `list`, and `fixed`. Figure 9 illustrates the distribution of all relations in our treebank.

In all 17 UPOS tags, 15 of them are found in the treebank, excluding `X` and `SYM`. Figure 11 shows the distribution of UPOS tags.

D Detailed Fine-tuning Results

The model is fine-tuned for 50 epochs. Figure 12, 13 shows the UAS and LAS, respectively, in each epoch. All of them show large generalisation gaps, with the model fitted almost perfectly on training data. They all show an early stopping at around epoch 15. This suggests a typical overfitting behaviour that the model lacks generalisation improvements, highlighting the need for more training data.

Another observation is that unlabelled metrics are generally around 10%–20% better than labelled metrics. The immediate focus could also be on the parser's relation labelling quality.

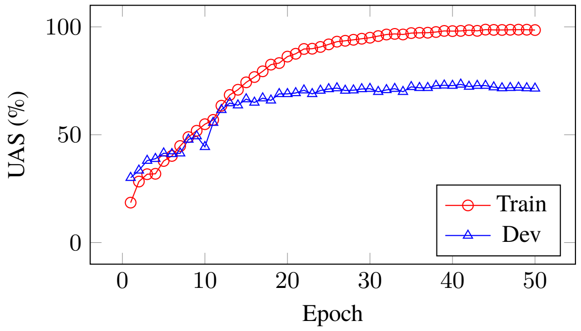


Figure 12: Unlabelled Attachment Score (UAS).

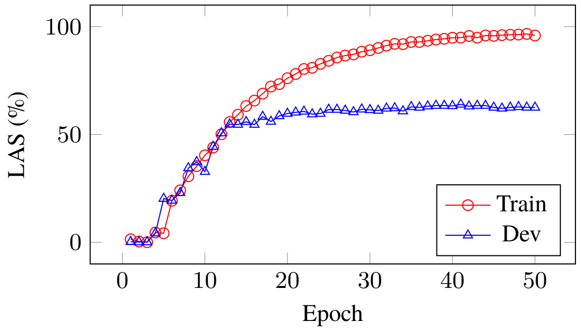


Figure 13: Labelled Attachment Score (LAS).

```
# sent_id = 3
# text = 吾伐会的随便讲吾爱侬呃。
# text_cmn = 我不会随便讲我爱你的。
1  吾      吾      PRON   PRP  Person=1      4  nsubj      _  SpaceAfter=No
2  伐会的  伐会的  AUX    MD   Polarity=Neg  4  aux        _  SpaceAfter=No
3  随便    随便    ADV    RB   _             4  advmod     _  SpaceAfter=No
4  讲      讲      VERB   VV   _             0  root       _  SpaceAfter=No
5  吾      吾      PRON   PRP  Person=1      6  nsubj      _  SpaceAfter=No
6  爱      爱      VERB   VV   _             4  ccomp      _  SpaceAfter=No
7  侬      侬      PRON   PRP  Person=2      6  obj        _  SpaceAfter=No
8  呃      呃      PART   UH   _             4  discourse  _  SpaceAfter=No
9  。      。      PUNCT  .    _             4  punct      _  SpaceAfter=No
```

Figure 8: An example in ShUD treebank in ConLL-U format. Translation: *I won't say I love you casually.*

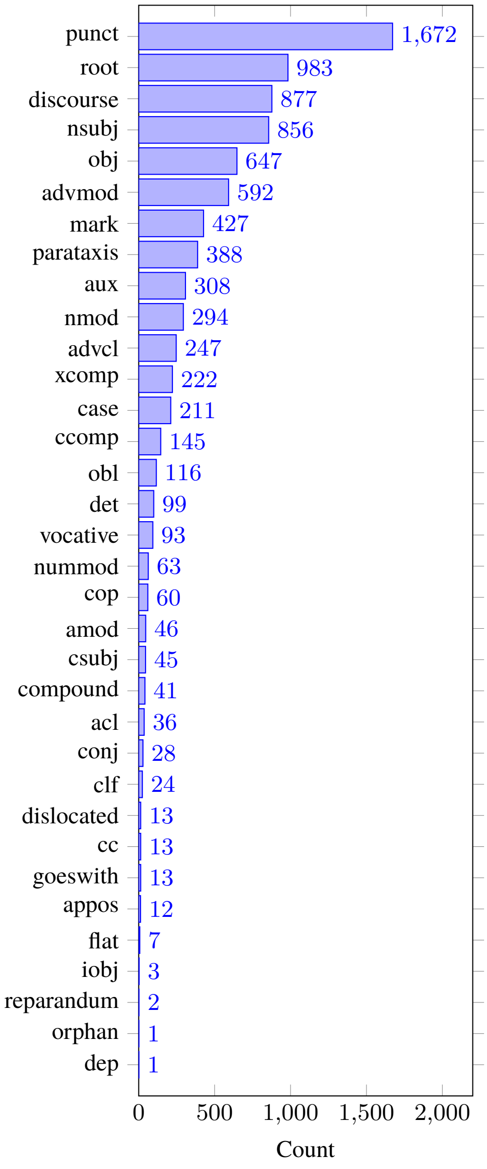


Figure 9: Distribution of relations.

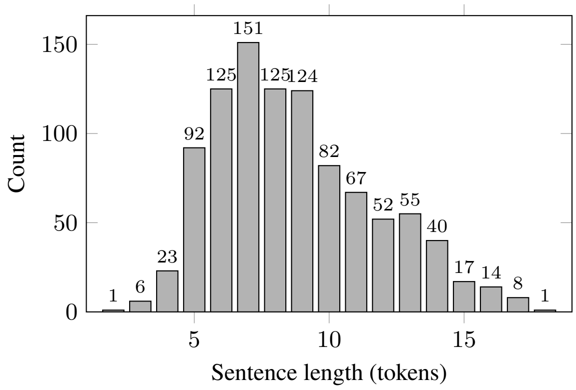


Figure 10: Distribution of sentence lengths.

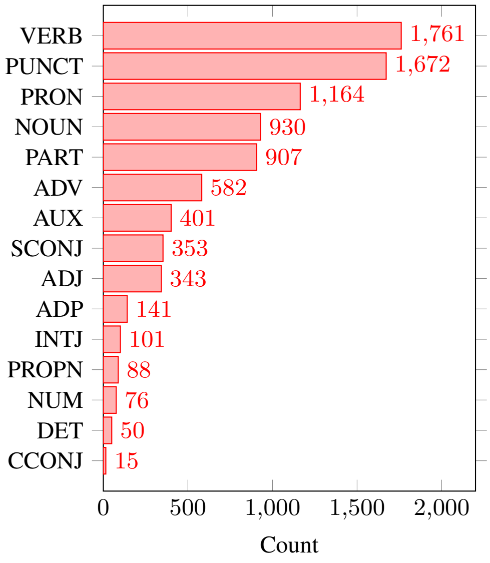


Figure 11: Distribution of UPOS tags.

Author Index

Akhundjanova, Arofat, 129
Akkurt, Furkan, 129
Alena, Witzlack-Makarevich, 119
Anastasopoulos, Antonios, 39
Andrews, Matthew Kirk, 164

Bauer, John, 105
Berdicevskis, Aleksandrs, 153
Bompolas, Stavros, 39
Bosco, Cristina, 80
Bu, Lanni, 52

Cecchini, Flavio Massimiliano, 174
Chierchiello, Elisa, 80
Chiril, Patricia, 80
Chontaeva, Bermet, 129
Coltekin, Cagri, 129, 164
Croft, William, 1

das Graças Volpe Nunes, Maria, 143
de Souza, Elvis A., 143
Diaz Hernandez, Roberto A., 11
Dyer, Andrew Thomas, 137

Eslami, Soudabeh, 129

Findlay, Jamie Yates, 70

Haug, Dag Trygve Truslew, 70

Ivani, Jessica K., 30

Ju, Zhuoxuan, 52

Kira, Tulchynska, 119

Levine, Lauren, 97
Liu, Zoey, 52

Manning, Christopher D., 105
Markantonatou, Stella, 39
Masciolini, Arianna, 153
Min, Junghyun, 97

Nivre, Joakim, 1

O’Brien, Colleen Alena, 137
Oya, Masanori, 22

Pagano, Adriana Silvina, 80, 143
Pardo, Thiago A. S., 143
Pirzada, Shafi, 105

Qureshi, Sarwat, 105

Rahman, Mutee U, 105
Ralli, Angela, 39

Sanches Duran, Magali, 143
Sanjrani, Zubair, 105
Schneider, Nathan, 52
Schroeder, Caroline T., 59
Shah, Sakiina, 105
Shaheer, Muhammad, 105
Speransky, Nina, 59
Sylvanus, Job, 119
Szawerna, Maria Irena, 153

Talpur, Mir Afza Ahmed, 105
Tulchynska, Kira, 30

Volodina, Elena, 153

Wagner, Nicholas E., 59

Yang, Qizhen, 186
Yang, Xiulin, 52

Zeldes, Amir, 59, 97
Zeman, Daniel, 11