



LREC 2026

**The Ninth Workshop on Universal Dependencies
(UDW 2026)**

Workshop Proceedings

Editors
Çağrı Çöltekin and Kaja Dobrovoljc

May 16, 2026

©ELRA Language Resources Association (ELRA), 2026

These proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-62-3

Preface

These proceedings include the program and papers that are presented at the ninth workshop on Universal Dependencies, held in conjunction with LREC on May 16, 2026.

Universal Dependencies (UD) is a framework for cross-linguistically consistent treebank annotation that has so far been applied to over 180 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 bring 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, The Ninth Workshop on Universal Dependencies (UDW 2026), is held in conjunction with LREC 2026 in Mallorca, Spain.

We received 42 submissions of which 29 regular and three non-archival submissions were accepted. As in previous instances of the UD workshop, accepted submissions cover a broad range of papers related to UD, including introduction of new treebanks, discussion of specific linguistic constructions, the applications of UD for certain languages and language families, and the use cases of UD for linguistic research and natural language processing.

We are honored to have two invited speakers at UDW 2026: Marie-Catherine de Marneffe (FNRS – UCLouvain) with a talk on “Diversity in NLP: why, what, where and how”, and Stephen Mayhew (Duolingo) with a talk on “Universal NER: Standing on the Shoulders of Giants”. We would also like to thank the UDW 2026 Program Committee for their timely and meticulous reviews.

We are looking forward to a productive and engaging workshop!

Çağrı Çöltekin, Kaja Dobrovoljc and Joakim Nivre

Organizing Committee

Çağrı Çöltekin, Tübingen University
Kaja Dobrovoljc, University of Ljubljana & Jozef Stefan Institute
Joakim Nivre, Uppsala University

Program Committee

Noor Abo Mokh, Indiana University
Jane Adkins, Dublin City University
Leonel de Alencar, UFC
Giuseppe Attardi, Università di Pisa
David Beck, University of Alberta
Aleksandrs Berdicevskis, University of Gothenburg
Ann Bies, Linguistic Data Consortium, Univ. Of Pennsylvania
Bernd Bohnet, Google, London
Gosse Bouma, University of Groningen
Miriam Butt, University of Konstanz
Giuseppe Celano, University of Leipzig
Xinying Chen, University of Ostrava
Daniel Dakota, Indiana University
Stefania Degaetano-Ortlieb, Saarland University
Kilian Evang, Heinrich Heine University Düsseldorf
Pegah Faghiri, CNRS
Marcos Garcia, University of Santiago de Compostela
Kim Gerdes, Université Paris-Saclay
Carlos Gómez-Rodríguez, Universidade da Coruña,
Bruno Guillaume, LORIA
Eva Hajičová, Charles University, Prague
Dag Haug, University of Oslo
Santiago Herrera, Université Paris Nanterre
Maarten Janssen, Charles University Prague
Mayank Jobanputra, Saarland University
Sylvain Kahane, Université Paris Nanterre
Sandra Kübler, Indiana University
Guy Lapalme, Université de Montréal
Teresa Lynn, Dublin City University
Rob Malouf, San Diego State University
Nicolas Mazziotta, Université de Liège
Alexander Mehler, Goethe-Universität Frankfurt am Main
Wolfgang Menzel, Hamburg University
Aleksandra Miletić, CNRS
Simon Mille, ADAPT, Dublin City University
Jiří Mirovský, Charles University, Prague
Simonetta Montemagni, Istituto di Linguistica Computazionale “A. Zampolli” - CNR
Kaili Müürisep, University of Tartu
Anna Nedoluzhko, Charles University Prague
Ruo Chen Niu, Beijing Language and Culture University
Stephan Oepen, University of Oslo

Petya Osenova, Bulgarian Academy of Sciences
Prokopis Prokopidis, Institute for Language and Speech Processing/Athena RC
Mathilde Regnault, Universität Stuttgart
Tanja Samardžić, University of Zurich
Giuseppe Samo, Beijing Language and Culture University
Nathan Schneider, Georgetown University
Djamé Seddah, Sorbonne University
Achim Stein, University of Stuttgart
Luka Terčon, Faculty of Arts, University of Ljubljana
Giulia Venturi, Istituto di Linguistica Computazionale “A. Zampolli” - CNR
Veronika Vincze, Hungarian Academy of Sciences
Yaqin Wang, Guangdong University of Foreign Studies
Pan Xiaxing, Huaqiao University
Nianwen Xue, Brandeis University, Waltham
Jianwei Yan, Zhejiang University
Zdeněk Žabokrtský, Charles University
Eva Zehentner, University of Zurich
Amir Zeldes, Georgetown University
Daniel Zeman, Charles University
Heike Zinsmeister, University of Hamburg

Table of Contents

<i>Probing the Dynamics of Syntactic Ability Acquisition Throughout LLM Pretraining</i> Hiroshi Matsuda and Masayuki Asahara	1
<i>Which languages are "hot", and which are "cool"? Using Universal Dependencies for large-scale comparisons of subject expression</i> Natalia Levshina	49
<i>Clefting and beyond in Tigrinya</i> Nazareth Amlesom Kifle and Michael Gasser	54
<i>Coconstructions in Spoken Data: UD Annotation Guidelines and First Results</i> Ludovica Pannitto, Kaja Dobrovoljc Zor, Sylvain Kahane, Elena Battaglia, Bruno Guillaume, Caterina Mauri and Eleonora Zucchini	60
<i>Verifying the Menzerath-Altmann law in the verbal domain in 180 languages</i> Pegah Faghiri, Kim Gerdes and Sylvain Kahane	76
<i>Comparing Dependency Distances of Esperanto and Other Languages in a Multi-Lingual Parallel Corpus</i> Masanori Oya	91
<i>Negation of Turkic non-verbal clauses: Analysis and Universal Dependencies Implementation</i> Nikolett Mus, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, Sardana Ivanova, Çağrı Çöltekin, Jonathan N. Washington, Gulnura Dzhumaliev and Aida Kasieva	97
<i>Towards Universal Dependencies for L2 Learners of Modern Greek: Annotation and Challenges</i> Christina Klironomou, Thelka Pasparaki, Arianna Masciolini, Alexandros Tantos, Despoina Ourania Touriki, Konstantinos Tsiotskas and Eleni Tsourilla	103
<i>A Comparative Linguistic Analysis of Ottoman and Modern Turkish through UD Treebanks</i> Enes Yilandiloğlu	109
<i>Word segmentation for UD: a comparison of isiZulu and Sepedi</i> Laurette Marais and Laurette Pretorius	116
<i>CoBra: A Compound Branching Resource for Nominal Triconstituent Compounds in English and German</i> Carmen Schacht, Isabell Landwehr, Diana Davidson, Konrad Grabowski, Magdalena Meiser and Sophia Wiedmann	128
<i>SE Constructions Revisited: Focus on Treebanks for Romance Languages</i> Verginica Barbu Mititelu, Elena Irimia, Adriana S. Pagano, Roxana Ciolanescu and Ioana Buhnica	142
<i>Say "No" to Missing Polarity: A Negation Enrichment of Porttinari UD Treebank</i> Isaac Souza de Miranda Junior, Oto Araújo Vale and Marie-Catherine de Marneffe ...	153
<i>The Grammar Does the Work: Functional vs. Lexical Dependency Length Minimization Across the UD Languages</i> Kim Gerdes	163

<i>Greenberg's Universal 45 in Universal Dependencies: Gender Distinctions and Annotation Challenges</i>	
Antoni Brosa-Rodriguez and M. Dolores Jimenez Lopez	174
<i>A Proposal for a More Universal Annotation of Relative Clauses in Universal Dependencies</i>	
Santiago Herrera and Sylvain Kahane	183
<i>MesoTree: Annotated Linguistic Resources for Quantitative Comparative Linguistic Analysis and NLP in Mesoamerica</i>	
Robert Pugh, Francis Tyers and Robert Henderson	197
<i>Nonprototypical Predication and Nonpredicational Clauses in Universal Dependencies</i>	
Joakim Nivre, William Croft and Andre Coneglian	208
<i>Complex Predicates in Universal Dependencies</i>	
Andre Coneglian, Joakim Nivre and William Croft	218
<i>Bringing Information Structure to Universal Dependencies</i>	
Nikolett Mus, Andrew Dyer, Claudia Corbetta and Sylvain Kahane	229
<i>Introducing Universal Dependencies for Sardinian: the UD ContSar Treebank</i>	
Nicoletta Puddu, Manuela Sanguinetti and Luigi Talamo	242
<i>Exploiting Parallel Aligned Treebanks</i>	
Maarten Janssen	252
<i>Cross-Dialectal Transfer for Low-Resource Arabic: The Tunisian Arabic Dependency Treebank</i>	
Amal Aissaoui	258
<i>From Treebank Metadata to Sentence-Level Genre in Universal Dependencies: A Reproducible, Versioned Resource</i>	
Egon Stemle	268
<i>Towards a Universal Dependency Corpus for Old Saxon (Old Low German)</i>	
Christian Chiarcos and Janine Siewert	277
<i>Gathering valency frames for annotation and batch corrections</i>	
Mathilde Regnault	289
<i>Syntax is the Key to Semantics: Combining Universal Dependencies and Abstract Meaning Representation</i>	
Johannes Heinecke	295
<i>Extending Retag to Conversion Error Detection: A Case Study on SynTagRus Morphology</i>	
Andrei Movsesian and Daniil Timchenko	305
<i>Speech Act Constructions in Universal Dependencies</i>	
Joakim Nivre, William Croft and Andre Coneglian	315

Workshop Program

Saturday May 16, 2026

09:00 Session 1

09:00 *Opening*

09:05 *Keynote 1: Diversity in NLP: why, what, where and how*
Marie-Catherine de Marneffe

10:00 *Probing the Dynamics of Syntactic Ability Acquisition Throughout LLM Pre-training*
Hiroshi Matsuda and Masayuki Asahara

10:15 *Which languages are "hot", and which are "cool"? Using Universal Dependencies for large-scale comparisons of subject expression*
Natalia Levshina

10:30 Coffee break

11:00 Session 2 (poster)

Clefting and beyond in Tigrinya
Nazareth Amlesom Kifle and Michael Gasser

Coconstructions in Spoken Data: UD Annotation Guidelines and First Results
Ludovica Pannitto, Kaja Dobrovoljc Zor, Sylvain Kahane, Elena Battaglia, Bruno Guillaume, Caterina Mauri and Eleonora Zucchini

Verifying the Menzerath-Altmann law in the verbal domain in 180 languages
Pegah Faghiri, Kim Gerdes and Sylvain Kahane

Comparing Dependency Distances of Esperanto and Other Languages in a Multi-Lingual Parallel Corpus
Masanori Oya

Negation of Turkic non-verbal clauses: Analysis and Universal Dependencies Implementation
Nikolett Mus, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, Sardana Ivanova, Çağrı Çöltekin, Jonathan N. Washington, Gulnura Dzhumaliev and Aida Kasieva

Saturday May 16, 2026 (continued)

Towards Universal Dependencies for L2 Learners of Modern Greek: Annotation and Challenges

Christina Klironomou, Thelka Pasparaki, Arianna Masciolini, Alexandros Tantos, Despoina Ourania Touriki, Konstantinos Tsiotskas and Eleni Tsourilla

A Comparative Linguistic Analysis of Ottoman and Modern Turkish through UD Treebanks

Enes Yilandiloğlu

Word segmentation for UD: a comparison of isiZulu and Sepedi

Laurette Marais and Laurette Pretorius

CoBra: A Compound Branching Resource for Nominal Triconstituent Compounds in English and German

Carmen Schacht, Isabell Landwehr, Diana Davidson, Konrad Grabowski, Magdalena Meiser and Sophia Wiedmann

SE Constructions Revisited: Focus on Treebanks for Romance Languages

Verginica Barbu Mititelu, Elena Irimia, Adriana S. Pagano, Roxana Ciolanescu and Ioana Buhnila

Say "No" to Missing Polarity: A Negation Enrichment of Porttinari UD Treebank

Isaac Souza de Miranda Junior, Oto Araújo Vale and Marie-Catherine de Marneffe

12:00 Session 3

The Grammar Does the Work: Functional vs. Lexical Dependency Length Minimization Across the UD Languages

Kim Gerdes

Greenberg's Universal 45 in Universal Dependencies: Gender Distinctions and Annotation Challenges

Antoni Brosa-Rodriguez and M. Dolores Jimenez Lopez

A Proposal for a More Universal Annotation of Relative Clauses in Universal Dependencies

Santiago Herrera and Sylvain Kahane

Saturday May 16, 2026 (continued)

13:00 **Lunch break**

14:00 **Session 4**

14:00 *Keynote 2: Universal NER: Standing on the Shoulders of Giants*
Stephen Mayhew

14:50 *MesoTree: Annotated Linguistic Resources for Quantitative Comparative Linguistic Analysis and NLP in Mesoamerica*
Robert Pugh, Francis Tyers and Robert Henderson

15:10 *Nonprototypical Predication and Nonpredicational Clauses in Universal Dependencies*
Joakim Nivre, William Croft and Andre Coneglian

15:30 *Complex Predicates in Universal Dependencies*
Andre Coneglian, Joakim Nivre and William Croft

15:45 *Bringing Information Structure to Universal Dependencies*
Nikolett Mus, Andrew Dyer, Claudia Corbetta and Sylvain Kahane

16:00 **Coffee break**

16:30 **Session 5 (poster)**

DELTA: A Toolkit for Measuring Linguistic Diversity in Dependency-Parsed Corpora
Louis Estève and Kaja Dobrovoljc

Modelling the Morphology of Verbal Paradigms: A Case Study in the Tokenization of Turkish and Hebrew
Giuseppe Samo and Paola Merlo

Introducing Universal Dependencies for Sardinian: the UD ContSar Treebank
Nicoletta Puddu, Manuela Sanguinetti and Luigi Talamo

Exploiting Parallel Aligned Treebanks
Maarten Janssen

Saturday May 16, 2026 (continued)

Cross-Dialectal Transfer for Low-Resource Arabic: The Tunisian Arabic Dependency Treebank

Amal Aissaoui

From Treebank Metadata to Sentence-Level Genre in Universal Dependencies: A Reproducible, Versioned Resource

Egon Stemle

Towards a Universal Dependency Corpus for Old Saxon (Old Low German)

Christian Chiarcos and Janine Siewert

Gathering valency frames for annotation and batch corrections

Mathilde Regnault

Syntax is the Key to Semantics: Combining Universal Dependencies and Abstract Meaning Representation

Johannes Heinecke

Extending Retag to Conversion Error Detection: A Case Study on SynTagRus Morphology

Andrei Movsesian and Daniil Timchenko

Exploring language relations through syntactic distances and geographic proximity

Juan De Gregorio, Raul Toral and David Sanchez

Speech Act Constructions in Universal Dependencies

Joakim Nivre, William Croft and Andre Coneglian

17:30

Closing: Community discussion

Invited Talks

Universal NER: Standing on the Shoulders of Giants

Stephen Mayhew, Duolingo

The huge success of Universal Dependencies has been an inspiration for similar projects. One such descendant is Universal NER, a large-scale, multilingual, shared-schema named entity recognition project, in the works since 2020. This talk will cover a brief history of the project, the philosophy behind the data, the tagset, and the annotations, as well as details on the data itself, including the latest release, Universal NER v2, presented at LREC 2026.

Diversity in NLP: why, what, where and how

Marie-Catherine de Marneffe, CENTAL - UCLouvain / FNRS

Diversity has become an advocated property of NLP datasets and systems. However, its use remains largely ad hoc, with limited theoretical grounding and definitions and measures that vary across studies. In this talk, I will describe a recent survey with colleagues of 308 diversity-related papers from the ACL Anthology. We propose an NLP-specific framework organized around four perspectives: why diversity matters, what is measured, where it is measured, and how, drawing from insights from other scientific fields where the concept of diversity has been thoroughly conceptualized. I will also look at UD from these perspectives.

Probing the Dynamics of Syntactic Ability Acquisition Throughout LLM Pretraining

Hiroshi Matsuda, Masayuki Asahara

Megagon Labs, Tokyo, Recruit Co., Ltd., National Institute for Japanese Language and Linguistics
Marunouchi 1-9-2, Chiyoda, Tokyo, Japan, Midorichou 10-2, Tachikawa, Tokyo, Japan
hiroshi_matsuda@megagon.ai, masayu-a@ninjal.ac.jp

Abstract

In this research, we introduce LoRA probing, a lightweight approach for observing how core syntactic abilities emerge during LLM pretraining. Leveraging OLMo-2’s public intermediate checkpoints, we trace learning curves across 24 pretraining stages on 33 Universal Dependencies languages by fine-tuning LoRA with step-by-step parsing instructions and a simple tabular output. To fit the relatively short context length of the OLMo-2, we design a compact `2-step-no-form` prompt template and this matches the baseline in average accuracy while halving the context length and substantially increasing throughput, enabling efficient large-scale evaluation. Token Recall surpasses 0.9 within the first 1–2K pretraining steps, indicating that stable output formatting emerges early. Despite OLMo-2-7B’s English-centric pretraining, LAS exceeds 80 points in 29 of 33 languages; however, relations such as *iobj* and *csubj* show delayed onset and instability across many languages. LoRA probing thus provides a practical, reproducible lens on the cross-lingual dynamics of syntactic acquisition during LLM pretraining.

Keywords: syntactic probing, universal dependencies, LoRA fine-tuning, step-by-step prompting

1. Introduction

Large language models (LLMs) have achieved striking gains across a wide range of NLP tasks, yet they continue to display gaps in linguistic competence, especially in settings that stress compositionality, morphology, and long-distance dependencies. A plausible explanation for the steady progress of new model releases is the gradual mitigation of such gaps during pretraining. If so, a natural scientific question arises: *when* and *how* do specific syntactic abilities emerge as pretraining unfolds, and which aspects remain fragile even near the end of training?

To answer this question, we need instruments that can observe the development of syntax without excessively altering the model under study. Prior approaches offer complementary perspectives: *structural probing* attempts to reconstruct syntax from internal representations, while *behavioral evaluation* tests task performance via prompting or fine-tuning. Structural probes illuminate geometry but can drift from actual behavior; behavioral tests reveal competence but may blur whether success stems from base models or from heavy task-specific adaptations. We aim to bridge this gap with a probe that is both *lightweight* and *behaviorally faithful*.

We therefore introduce **LoRA probing**, a simple protocol that fine-tunes only the parameters of Low-Rank Adapter (Hu et al., 2022) while keeping the base model fixed, and then evaluates the base model’s latent syntactic competence via the adapter’s outputs. Concretely, we cast dependency parsing as a step-by-step generation task and read out parses in a tabular format introduced in Mat-

suda et al. (2025). To fit models with relatively short context windows, we design a compact `2-step-no-form` parsing prompt: the model first predicts UPOS tags and then predicts HEAD and DEPREL without reproducing word’s FORM. This yields behaviorally interpretable signal while minimizing prompt length and training cost.

Our empirical setting leverages OLMo-2-7B (Team OLMo et al., 2025), which publicly releases intermediate checkpoints during pretraining. These snapshots allow us to trace learning curves over time: we train LoRA at multiple checkpoints and evaluate on Universal Dependencies (UD) treebanks (Nivre et al., 2020) spanning 33 languages. This multilingual canvas exposes both universal and language-specific dynamics, enabling us to ask whether, for instance, tagging stabilizes earlier than relation labeling, or whether particular relations (e.g., *iobj*, *csubj*) consistently lag across typologically diverse languages. We structure the study around four research questions:

RQ1: Prompt efficiency. Can a compact `2-step-no-form` prompt match or surpass a conventional `3-step` prompt in accuracy while improving throughput, making large-scale evaluation feasible?

RQ2: Onset and stabilization. At what stages of pretraining do token-level formatting discipline emerge (as measured by TOKEN RECALL), and how early do core syntactic signals become reliable?

RQ3: Cross-lingual trajectory. How do

learning curves differ across languages for UPOS and for major dependency relations, and which patterns appear universal versus language-specific?

RQ4: Persistent weakness. Which relations remain difficult or unstable late in pretraining (e.g., *ibj*, *csubj*), and what does this imply about data imbalance, annotation conventions, or inductive biases?

Beyond reporting aggregate scores, we emphasize *dynamics*: the evolution of competence as pretraining progresses. This perspective complements single-snapshot evaluations and helps explain why certain linguistic contexts remain challenging for LLMs even as overall performance improves. By clarifying *which* syntactic abilities appear early, *which* plateau, and *which* lag or fluctuate, LoRA probing offers actionable guidance for model developers (e.g., curriculum design, data selection) and for the broader community seeking the ways to improve linguistic competence in LLMs.

Finally, the remainder of the paper is organized as follows. Section 2 situates our work relative to structural probes and behavioral evaluations. Section 3 details the LoRA probing protocol, the 2-step-no-form prompt, datasets, checkpoints, and evaluation criteria. Section 4 presents cross-lingual learning curves and ablations. Section 5 discusses implications and limitations, and Section 6 concludes.

2. Related Work

A growing body of work investigates how syntactic abilities emerge in language models during *pre-training*, including studies that track the evolution of linguistic competence over time (Liu et al., 2021; Pérez-Mayos et al., 2021; Blevins et al., 2022). In the multilingual setting, Wang et al. (2024) examine the emergence of cross-lingual alignment in BLOOM (BigScience Workshop et al., 2023), highlighting that syntactic behavior can develop unevenly across languages even within the same pre-training regime. These efforts suggest that probing the *trajectory* of syntactic acquisition—rather than only evaluating end points—yields insight into what models learn and when they learn it.

Following the taxonomy of Liu et al. (2021), probing methods for syntactic competence can be grouped into *structural* and *behavioral* approaches. Structural probes attach a lightweight classifier to representations from a pretrained model and test whether syntactic structure can be recovered. They illuminate the geometry of internal states but may overestimate competence if the readout is powerful relative to the underlying signal.

Behavioral evaluations, by contrast, assess a model’s own inference behavior through prompt-

ing or task-specific fine-tuning; they reflect observable competence but can conflate improvements in the base model with heavy adaptation or prompt engineering, and they often require long prompts that limit scalability. Our work builds on the behavioral perspective while addressing these limitations. We introduce a minimally invasive protocol—LoRA probing—which reduces the risk that the probe itself *creates* the competence it measures.

Although a zero- or few-shot inference with an LLM would entirely remove the effect of adding a probe, in dependency parsing it is difficult to assess the parsing potential of intermediate pretraining checkpoints in such settings for the following reasons:

- Head identification accuracy is quite low in zero- or few-shot settings (Tian et al., 2024).
- The definitions of UPOS tags and DEPREL labels are not learned during pretraining and are therefore almost impossible to reproduce in a zero-shot setting.
- Similarly, given the limited context length, learning the definitions of tags and labels in a few-shot setting is also difficult
- For checkpoints from the early stages of pre-training, it is difficult to make the model adhere to the required output format using few-shot instructions alone.

For these reasons, this study adopts a LoRA-based probing approach.

3. Method and Setup

3.1. LoRA probing

We propose LoRA probing, a minimally invasive approach to probe the syntactic abilities of LLMs that combines step-by-step parsing prompts with LoRA-based fine-tuning (Matsuda et al., 2025). Conceptually, LoRA probing sits between structural and behavioral paradigms: like structural probes, it adds only lightweight adapter layers to a pre-trained model; like behavioral probes, it evaluates the model through its own auto-regressive decoding, yielding behaviorally meaningful signals. Crucially, because we evaluate multiple intermediate checkpoints, LoRA probing enables a dynamic view of syntactic acquisition, complementing structural analyses and snapshot behavioral tests with a temporally resolved account of how specific abilities emerge during pretraining.

Our protocol keeps the base model fixed and fine-tunes only low-rank adapter parameters, then casts dependency parsing as a step-by-step generation task whose outputs are read as tabular TSV parses.

We leverage OLMo-2’s public intermediate pre-training checkpoints to study *when* and *which* syntactic abilities emerge. Using Universal Dependencies (UD) treebanks across 33 languages, we analyze which features are acquired—and which remain fragile—over the course of pretraining.

3.2. Parsing prompt

The 3-step parsing prompt proposed by (Matsuda et al., 2025) assumes underlying LLM with a maximum context length of 8,192 tokens. However, OLMo-2’s maximum context length is only 4,096. To accommodate this limitation, we propose a 2-step-no-form parsing prompt: instead of predicting UPOS, HEAD and DEPREL in three separate steps, we first predict UPOS and then jointly predict HEAD and DEPREL. We also remove the FORM field from the output records. This design halves the token budget and the associated computational cost during both training and inference, while preserving the interpretability. An example instance of 2-step-no-form prompt is shown in Figure 1.

As a preliminary study, we compared the parsing accuracy and throughput of the 3-step prompt with the 2-step-no-form prompt in UD English (en_ewt) (Silveira et al., 2014) using gemma-2-9b (Gemma Team et al., 2024). The means and standard deviations of four independent runs are shown in Table 1. The 2-step-no-form matches the 3-step baseline in average accuracy while substantially increasing throughput (training +45%; inference +129%), enabling efficient large-scale evaluation.

In addition, the single-run evaluation results for all 33 languages are presented in Appendix B. The macro-averaged values indicate a clear tendency for the 2-step-no-form prompt to perform comparably to, or better than, the 3-step prompt in all models. Accordingly, we use the 2-step-no-form prompt for all subsequent experiments.

	3-step	2-step-no-form
Accuracy [%]		
TOKEN RECALL	100.0±0.1	100.0±0.0
UPOS (F1)	98.6±0.1	98.4±0.1
UAS (F1)	95.4±0.1	95.2±0.2
LAS (F1)	93.9±0.1	93.5±0.0
Throughput [sentence/sec]		
Training	3.04±0.01	4.40±0.03
Inference	11.6±0.03	24.1±0.13

Table 1: Comparison of accuracy and throughput of gemma-2-9b trained on en_ewt over 4 runs. Single-run evaluation results for all 33 languages are shown in Appendix B.

```
{role: system} You are an English linguist and
specialize in English dependency parsing
based on Universal Dependencies.

{role: user} We will now perform dependency
parsing on English sentence. After splitting
the input sentence into words with indexes as
shown below, execute following two tasks:

- Task 1
Create a TSV with two fields: word index
from 1 to 6 + part of speech.

- Task 2
Add two fields to each row of the output of
Task 1: the dependent word index + the
Universal Dependencies relation label.
However, for the word that is the main
predicate of the sentence, the dependent word
index should be 0.

input sentence:
I'm just speculating now.

words:
I
'm
just
speculating
now
.

{role: assistant}

- Task 1
1 PRON
2 AUX
3 ADV
4 VERB
5 ADV
6 PUNCT

- Task 2
1 PRON 4 nsubj
2 AUX 4 aux
3 ADV 4 advmod
4 VERB 0 root
5 ADV 4 advmod
6 PUNCT 4 punct

<eos>
```

Figure 1: An example of the 2-step-no-form prompt instance. The parts that change from instance to instance are shown in **BOLD**. The **SHADED REGION** in the assistant-role corresponds to the range over which the loss gradient is computed during training, and to the decoded text during inference. At inference time, the span from the system-role up to the assistant-role header is provided as input, and decoding of the subsequent segment continues until <eos> is generated.

3.3. Multilingual capability of OLMo-2-7B

Our primary analysis targets OLMo-2-7B and its publicly released intermediate pretraining checkpoints, which enable stage-wise probing of syntactic acquisition in LLMs. OLMo-2-7B is pretrained on 3.9T English tokens drawn chiefly from the DCLM-baseline corpus (Li et al., 2025). DCLM-baseline retains only documents with a language-identifier estimate of $\geq 99\%$ English, so non-English content is vanishingly small. Given this near-monolingual pretraining mixture, it is important to

	en_ewt	fr_gsd	cs_pdt	ja_gsd	lt_alksnis	33-AVG
gemma-2-9b	93.5	94.6	94.5	93.9	84.9	91.1
Llama-3.1-8B	93.4	94.4	94.2	93.5	81.0	90.3
Qwen2.5-7B	93.1	94.4	94.1	92.9	76.6	89.5
OLMo-2-7B	92.4	92.8	92.1	91.5	64.9	86.4

Table 2: Comparison of LAS (F1; %) among multilingual LLMs and English-centric OLMo-2-7B using 2-step-no-form prompt. The overall results for all 33 languages are shown in Appendix B.

test whether OLMo-2-7B can still be meaningfully evaluated on multilingual syntactic tasks.

As a preliminary evaluation, we verify that OLMo-2-7B achieves reasonable performance on multilingual syntactic evaluation, comparing it with strong multilingual LLM baselines. Table 2 presents the comparison of accuracy metrics¹ of three multilingual LLMs—gemma-2-9b, Llama-3.1-8B (Grattafiori et al., 2024), and Qwen2.5-7B (Qwen et al., 2025)—as well as OLMo-2-7B on the datasets described in Table 3. OLMo-2-7B achieved performance comparable to the multilingual LLMs in English and exhibited unexpectedly strong results in French, Czech, and Japanese, even though these languages are almost absent from its pretraining corpus. In contrast, for Lithuanian, OLMo-2-7B scored 20 LAS points lower than gemma-2-9b². The entire evaluation results for all 33 languages are shown in Appendix B. The macro-average LAS of OLMo-2-7B across 33 languages was 86.4%, a level considered sufficient for comparing dependency parsing accuracies. These findings suggest that OLMo-2-7B has sufficient syntactic ability to support a meaningful evaluation of its multilingual performance. We then use LoRA probing to obtain temporally resolved trajectories of syntactic acquisition across checkpoints.

3.4. Datasets and language families

We evaluate our method on the Universal Dependencies (UD) v2.15 release.³ We selected 33 treebanks listed in Table 3, each representing a distinct language and containing more than 40,000 words in the *training* split. To balance the fine-tuning data volume across languages, we set the number of training epochs so that the product of the training set size (in words) and the number of epochs falls within the range of 300K–600K. In addition, more

¹The metrics definition is described in Section 4.

²The languages in which OLMo-2-7B achieved LAS below 80% were Irish, Lithuanian, Simplified Chinese, and Turkish.

³At the time of submission, UD v2.17 was the latest release; however, all results in this paper are evaluated on UD v2.15 to allow for a fair comparison with baseline.

detailed statistics including all of the UPOS tags and the DEPREL labels for all 33 languages are presented in the Appendix C.

For typological analysis, we group languages by family following the WALS Online (Dryer and Haspelmath, 2013) adopted on the UD website.

3.5. Fine-tuning method

We adopt Low-Rank Adapters (LoRA) for fine-tuning the LLMs, largely following the publicly available implementation and experimental setup⁴ of Matsuda et al. (2025). We modify only two aspects: (i) the number of training epochs, adjusted to balance the training volume across treebanks, and (ii) the learning rate, reduced from 3×10^{-4} to 1×10^{-4} to mitigate overfitting when fine-tuning the intermediate checkpoints of OLMo-2-7B.

3.6. Tokens per step in OLMo-2 pretraining

During the pretraining of OLMo-2, the batch size is set to 1,024 and the context length to 4,096, resulting in approximately 4 million tokens being processed per training step. The X-axis of the learning curves in the experimental results shown in the following sections represents the number of steps; therefore, to convert this to the number of pretrained tokens, multiply by 4 million.

4. Experiments

In this section, we address the question of learning dynamics—specifically, how the syntactic abilities of OLMo-2-7B emerge over the course of pretraining when evaluated at intermediate checkpoints. We use the datasets and protocol described in Section 3 (UD v2.15, 33 languages, 2-step-no-form parsing prompt with LoRA).⁵

Scope. Evaluation covers 33 UD languages and use the official train/dev/test splits without modification. We probe intermediate pretraining checkpoints of OLMo-2-7B to obtain temporally resolved trajectories of syntactic acquisition. At every checkpoint, we perform SFT on LoRA from its initial state.

Tasks and metrics. Following our setup, we cast dependency parsing as step-by-step generation and report UPOS, UAS, and LAS as *F1 score*⁶

⁴<https://github.com/megagonlabs/llmpp/tree/20250612>

⁵All experiments were run on Google Cloud; the aggregate compute for training and inference was approximately 1,600 GPU-hours, measured in H100-80GB.

⁶UPOS: pos tagging accuracy. UAS (Unlabeled Attachment Score): HEAD accuracy. LAS (Labeled Attachment Score: simultaneous accuracy of HEAD and DEPREL. Following standard evaluation tool of UD, we

Family	Language	Treebank	Words in Training-set Total (<i>obj</i> , <i>csbj</i>)	Training Epochs	Training Words	Words in Test-set Total (<i>obj</i> , <i>csbj</i>)
Germanic	Danish	da_ddt (Johannsen et al., 2015)	80,378 (120, 0)	4	321,512	10,023 (15, 0)
	Dutch	nl_alpino (Bouma and van Noord, 2017)	186,027 (509, 400)	2	372,054	11,046 (13, 21)
	English	en_ewt (Silveira et al., 2014)	204,579 (649, 293)	2	409,158	25,094 (71, 25)
	German	de_gsd (McDonald et al., 2013)	263,791 (0, 214)	2	527,582	16,498 (0, 16)
	Norwegian	no_bokmaal*	243,886 (471, 794)	2	487,772	29,966 (51, 103)
	Swedish	sv_talbanken (Nivre and Megyesi, 2007)	66,646 (105, 249)	6	399,876	20,377 (37, 93)
Romance	Catalan	ca_ancora (Taulé et al., 2008)	429,578 (0, 736)	1	429,578	59,610 (0, 101)
	French	fr_gsd (Guillaume et al., 2019)	354,652 (813, 195)	1	354,652	10,018 (34, 12)
	Italian	it_isdt (Bosco et al., 2013)	276,014 (648, 300)	2	552,028	10,417 (20, 3)
	Portuguese	pt_gsd*	255,116 (377, 490)	2	510,232	31,477 (45, 61)
	Romanian	ro_rrt (Mititelu, 2018)	185,125 (1,366, 755)	2	370,250	16,324 (139, 41)
	Spanish	es_ancora (Taulé et al., 2008)	453,039 (0, 988)	1	453,039	53,622 (0, 94)
Slavic	Belarusian	be_hse (Shishkina and Lyashevskaya, 2021)	273,181 (2,269, 828)	2	546,362	15,997 (82, 31)
	Bulgarian	bg_btb (Osenova and Simov, 2004)	124,336 (2,818, 358)	3	373,008	15,724 (336, 43)
	Croatian	hr_set (Agić and Ljubešić, 2015)	152,857 (538, 271)	2	305,714	24,260 (77, 40)
	Czech	cs_pdt (Hajič et al., 2024)	1,173,285 (455, 4,870)	0.35	410,650	173,918 (66, 680)
	Polish	pl_pdb (Wróblewska, 2018)	281,685 (5,311, 185)	2	563,370	33,616 (673, 16)
	Russian	ru_syntagrus (Droganova et al., 2018)	1,204,640 (10,424, 6,743)	0.35	421,624	157,718 (1,425, 897)
	Serbian	sr_set (Batanović et al., 2018)	74,259 (2, 160)	5	371,295	11,421 (0, 24)
	Slovak	sk_snk (Zeman, 2017)	80,628 (80, 132)	4	322,512	12,736 (1, 18)
	Slovenian	sl_ssjs (Dobrovoljc et al., 2017)	215,155 (1,412, 722)	2	430,310	25,442 (145, 78)
	Ukrainian	uk_iu*	92,927 (298, 426)	4	371,708	17,217 (33, 57)
Baltic	Latvian	lv_lvtb (Pretkalniūniedina et al., 2018)	255,954 (6,141, 853)	2	511,908	37,162 (874, 115)
	Lithuanian	lt_alksnis (Bielinskienė et al., 2016)	47,641 (2, 322)	8	381,128	10,846 (0, 64)
Celtic	Irish	ga_idt (Lynn and Foster, 2016)	95,881 (0, 746)	4	383,524	10,109 (0, 95)
Finnic	Estonian	et_edt (Muischnek et al., 2016)	344,581 (0, 1,546)	1	344,581	48,465 (0, 243)
	Finnish	fi_ftb*	127,602 (0, 402)	3	382,806	16,286 (0, 58)
Greek	Greek	el_gdt (Prokopidis and Papageorgiou, 2017)	42,326 (59, 144)	9	380,934	10,672 (9, 40)
Turkic	Turkish	tr_boun (Marşan et al., 2022)	100,713 (184, 467)	4	402,852	12,210 (24, 72)
Afro-Asiatic	Arabic	ar_padt (Hajič et al., 2009)	223,881 (89, 442)	2	447,762	28,264 (6, 42)
Japanese	Japanese	ja_gsd (Asahara et al., 2018)	168,333 (0, 138)	2	336,666	13,034 (0, 12)
Korean	Korean	ko_gsd (Chun et al., 2018)	56,687 (77, 12)	7	396,809	11,677 (15, 5)
Sino-Tibetan	Chinese	zh_gsdsimp*	98,616 (63, 296)	4	394,464	12,012 (6, 37)

Table 3: Treebanks for 33 languages and their language families used in the experiments. These treebanks are distributed under licenses that permit research use; see the LICENSE file in the repository for details. Treebanks marked with an asterisk (*) lack a clear bibliographic reference in their repositories. All statistics are based on Universal Dependencies v2.15 and more detailed statistics including the occurrence count of each UPOS tag and DEPREL label for all 33 languages are presented in the Appendix C.

on test sets, alongside TOKEN RECALL⁷ to quantify output formatting stability.

Reporting. The accuracy figures reflect a *single* fine-tuning and decoding run. When multiple runs are conducted, we report the mean and standard deviation over the stated number of seeds.

4.1. Emergence in output formatting

Figure 2 shows the learning curves of TOKEN RECALL for 2-step-no-form prompt. Across all 33 languages, TOKEN RECALL exceeded 90 points within the first 1,000 to 2,000 steps, confirming that evaluation of the core syntactic parsing capability becomes reliable at an early stage of pretraining.

⁷use F1 score to consider word segmentation correctness.

⁷Word segmentation errors may appear in the output even when the correct segmentation is provided as input, due to the hallucination-like behavior of LLMs. Therefore, we adopt Recall as the evaluation metric, measuring the proportion of words that are correctly generated.

4.2. Emergence in syntactic ability

The results are presented by grouping the languages into four categories: Germanic, Romance, Slavic, and Others.

Since the pretraining data of OLMo-2-7B is predominantly English, the UPOS, UAS, and LAS curves for non-English languages remain below the envelope formed by English throughout the pre-training process. No substantial differences are observed among the Germanic, Romance, and Slavic language groups, whereas several languages in the Others group exhibit a slower initial rise.

4.3. Detailed analysis in each language

We analyze learning curves over the intermediate pretraining steps for 11 major UPOS tags and 13 key DEPREL labels. In the following subsections, we first analyze English as a representative of the Germanic family; we then analyze French as a representative of the Romance family and Czech as a representative of the Slavic family. The learn-

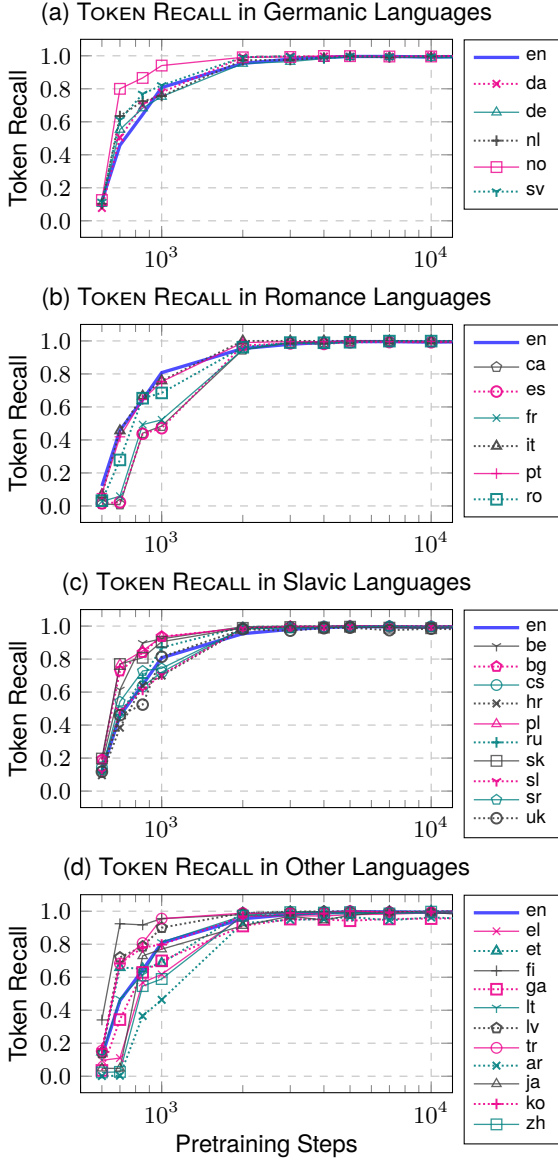


Figure 2: TOKEN RECALL acquisition through pre-training of OLMo-2-7B in 33 languages.

ing curves for all 33 languages are included in Appendix C.

4.3.1. Germanic: English (en_ewt)

The learning curves of UPOS and LAS of en_ewt are shown in Figure 6. For UPOS, learning first progresses on the closed-class categories *AUX*, *CCONJ*, *DET*, and *PRON*, followed by *VERB* and *PART*. Subsequently, the open-class categories *NOUN*, *ADJ*, and *ADV* are acquired, with *PROPN* and *SCONJ* learned in the later stages.

For dependency relations, the model initially learns *root*, *nsubj*, *advmod*, and *amod*, followed by *obj*, *nmod*, *obl*, and *xcomp*. Learning of the clausal relations *ccomp*, *acl*, and *advcl* then proceeds smoothly, whereas *iobj* and *csubj* show de-

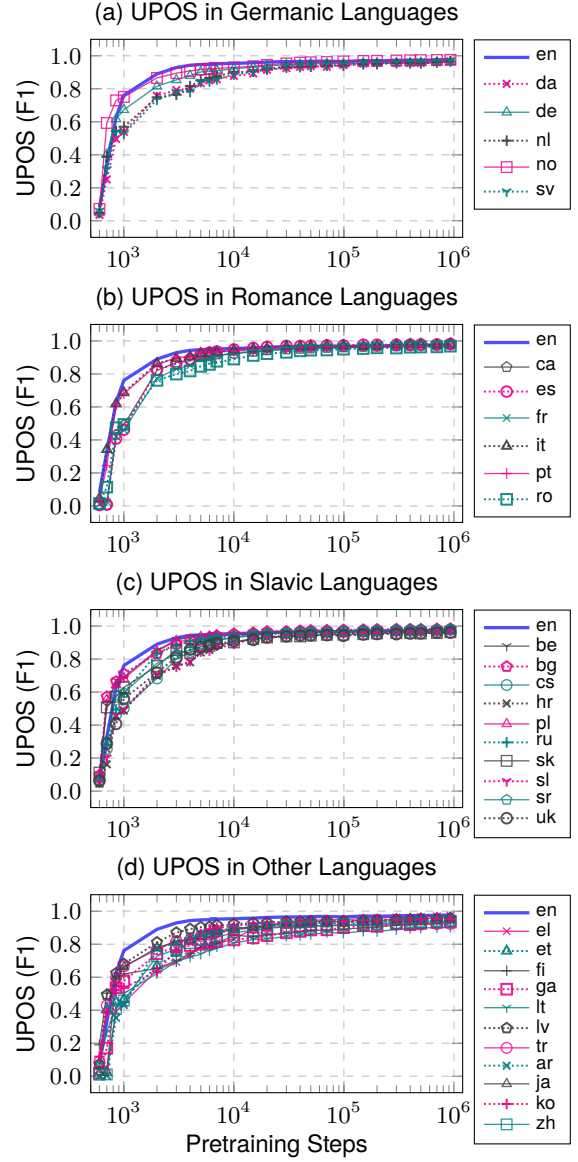


Figure 3: UPOS acquisition through pre-training of OLMo-2-7B in 33 languages.

layed onset and exhibit large fluctuations in their learning curves, indicating under-trained or annotation inconsistency. The unstable tendency on *iobj* is common in 30 out of 33 languages but exceptions are Bulgarian, Latvian, and Russian. In the same way, the unstable tendency on *csubj* is found in 29 out of 33 languages but exceptions are Irish, Norwegian, Russian, and Swedish.

4.3.2. Romance: French (fr_gsd)

The learning curves of UPOS and LAS of fr_gsd are shown in Figure 7. For UPOS, the model first acquires closed-class categories such as *DET* and *CCONJ*, followed by *AUX*, *PRON*, and *PROPN* together with *NOUN*. Subsequently, *VERB* and *ADV* emerge almost simultaneously, followed by *ADJ*.

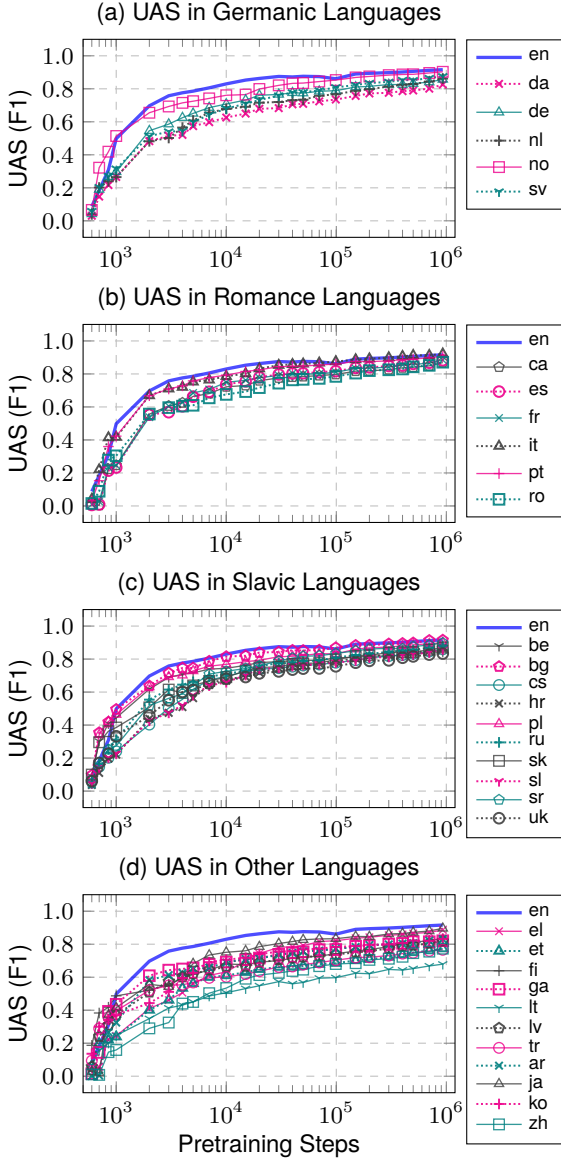


Figure 4: Unlabeled Attachment acquisition through pretraining of OLMo-2-7B in 33 languages.

The learning of *SCONJ* initially rises to around 0.7, after which it remains somewhat unstable but continues to show a gradual upward trend.

For DEPREL, *nmod* and *nsubj* emerge first, followed by *amod*, *advmod*, and *root* with a slight delay, then *obj* and *obl*. After further lag, the model begins to learn clausal relations such as *acl* and *xcomp*, followed later by *advcl* and *ccomp*. The learning of *iobj* fluctuates while gradually improving, similar to the case in English, whereas *csubj* shows little progress throughout training process.

4.3.3. Slavic: Czech (cs_pdt)

The learning curves of UPOS and LAS of cs_pdt are shown in Figure 8. For UPOS, learning first progresses among the closed-class categories ex-

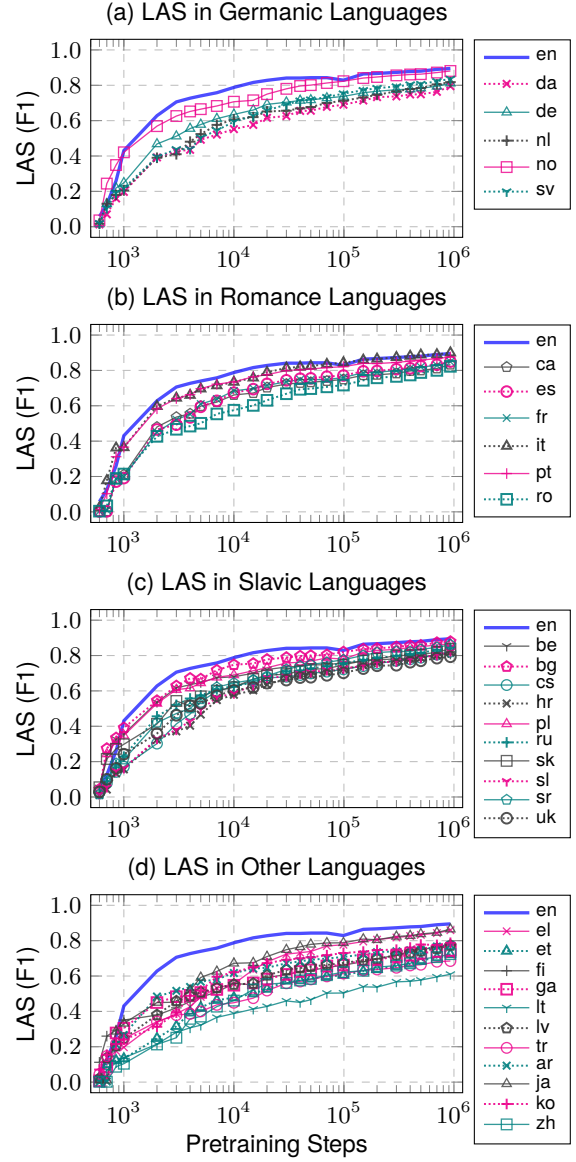


Figure 5: Labeled Attachment acquisition through pretraining of OLMo-2-7B in 33 languages.

cluding *PART*. Within the open-class categories, *PROPN* emerges earlier than *NOUN*, but its growth soon plateaus, after which *NOUN*, *ADJ*, and *VERB* overtake it around 2,000–3,000 training steps. *ADV* rises more slowly than the other open-class categories. *PART* shows the slowest onset and reaches only about 80 points accuracy in the final stage, which is lower than other languages.

For DEPREL, *root*, *obl*, *amod*, and *nmod* increase first up to roughly 2,000 steps. Subsequently, *amod* and *advmod* accelerate, while *obl* decelerates and is followed by *nsubj*. With a considerable delay, *obj*, *acl*, and *xcomp* start to rise synchronously, followed by *ccomp* and *advcl*. *csubj* exhibits the slowest onset but increases steadily compared with other languages. The accuracy of *iobj* remains close to zero throughout pretraining.

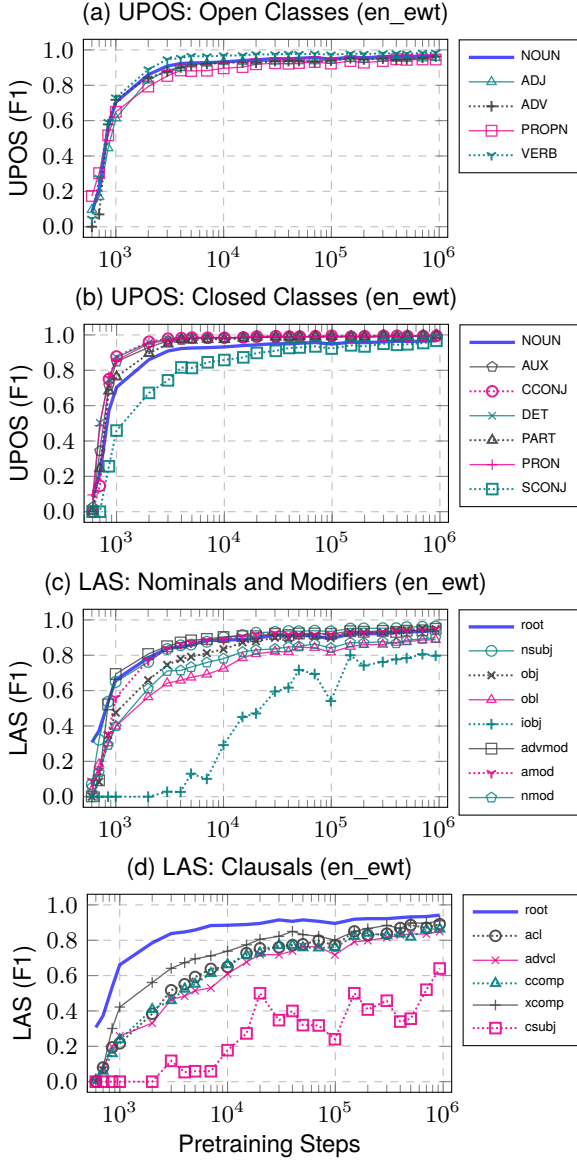


Figure 6: Syntactic acquisition through pretraining of OLMo-2-7B (English: en_ewt).

5. Discussion

The results answer the four research questions posed in Section 1.

For **RQ1**, the compact `2-step-no-form` prompt matches the conventional `3-step` prompt in accuracy while halving the context length and substantially improving throughput, making large-scale multilingual evaluation feasible. This tendency is also supported by the single-run results across 33 languages, where the `2-step-no-form` prompt performs comparably to, or slightly better than, the `3-step` prompt on macro-average.

For **RQ2**, Token Recall exceeds 0.9 within the first 1,000–2,000 pretraining steps in all 33 languages, indicating that output formatting stabilizes early. Core syntactic signals also emerge from an

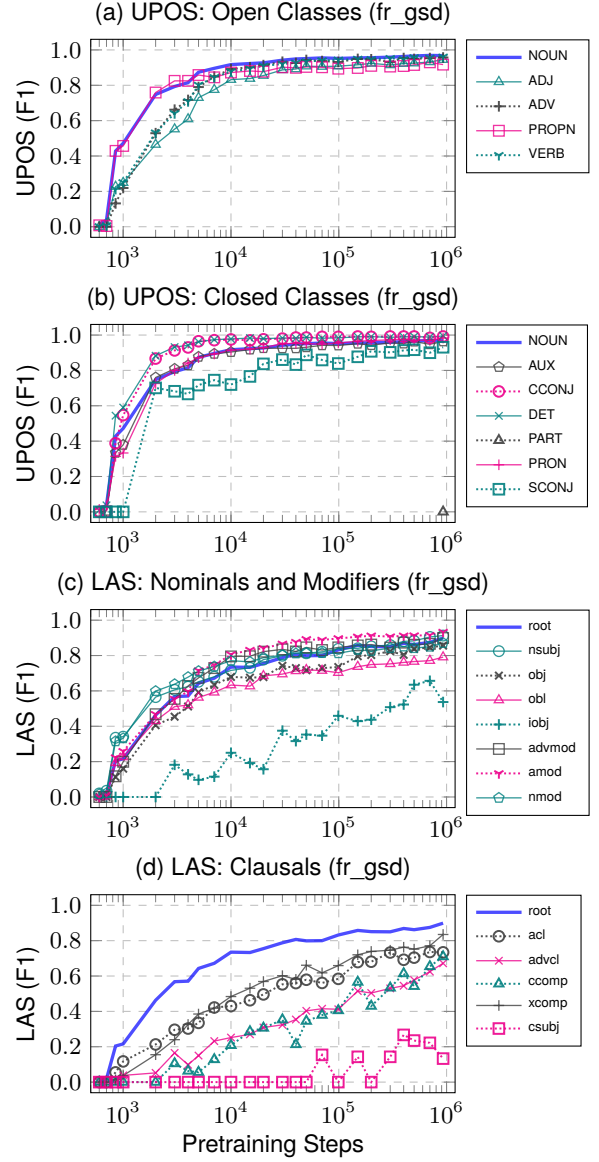


Figure 7: Syntactic acquisition through pretraining of OLMo-2-7B (French: fr_gsd).

early stage: closed-class UPOS categories generally precede open-class ones, and major nominal relations tend to precede clausal relations.

For **RQ3**, the overall learning trajectories are broadly similar across languages. English forms the upper envelope throughout pretraining, consistent with the English-centric pretraining mixture. No substantial differences are observed among the Germanic, Romance, and Slavic groups, whereas several languages in the Others group show a slower initial rise. Even so, OLMo-2-7B achieves LAS above 80 in 29 of 33 languages, suggesting that substantial syntactic ability emerges across typologically diverse languages.

For **RQ4**, *iobj* and *csubj* remain the most difficult relations. Stable *iobj* curves are observed in Bulgarian, Latvian, and Russian, while stable *csubj*

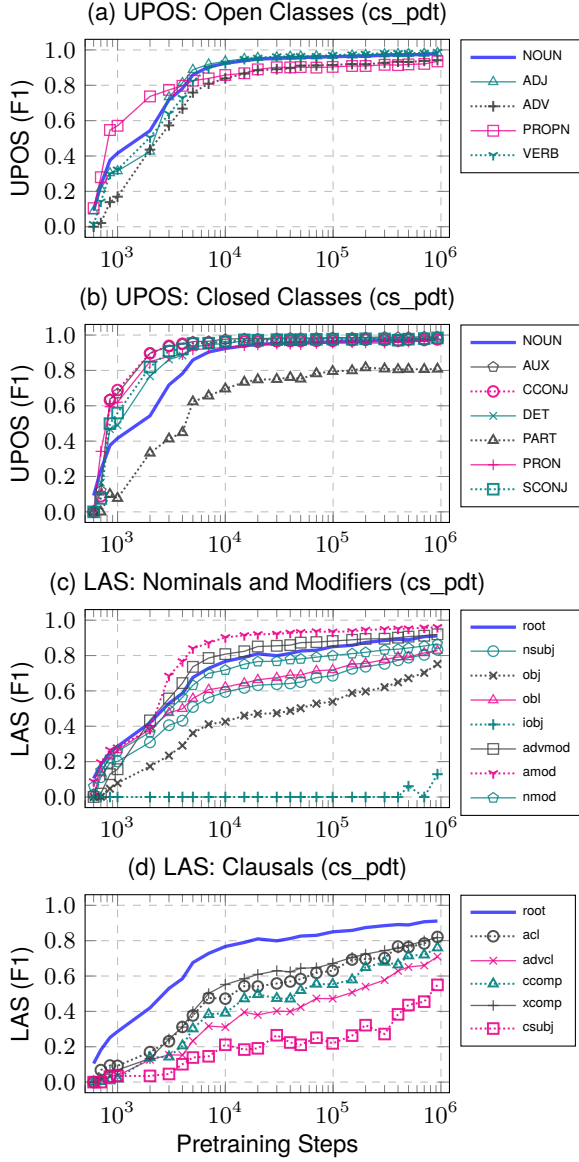


Figure 8: Syntactic acquisition through pretraining of OLMo-2-7B (Czech: cs_pdt).

curves are found in Irish, Norwegian, Russian, and Swedish. Their instability elsewhere is partly explained by low frequency or train-test mismatch, but cases such as Czech and Greek suggest that data scarcity alone is insufficient. We therefore interpret the persistent weakness of these relations as reflecting a combination of data scarcity, annotation inconsistency, and the inherent difficulty of the phenomena. Beyond *iobj* and *csubj*, several other DEPRELs also show unstable learning and persistently low accuracy, which should be investigated in future work.

6. Conclusion

We proposed LoRA probing, a novel approach to investigate how LLMs acquire syntactic abilities dur-

ing pretraining by combining parsing prompts with LoRA fine-tuning. To accommodate OLMo-2’s relatively short context window, we introduced a compact 2-step-no-form parsing prompt, halving the context length while achieving higher throughput with comparable accuracy.

We further applied LoRA fine-tuning with UD treebanks of 33 languages and the 2-step-no-form parsing prompt to OLMo-2-7B’s intermediate checkpoints. The proposed method achieved TOKEN RECALL above 90 points within the first 2,000 pretraining steps across all 33 languages, demonstrating that LoRA probing can reveal syntactic structures from the very early stages of LLM pretraining. Through LoRA probing, we analyzed the development of syntactic ability across 33 languages in intermediate checkpoints of OLMo-2-7B.

We found that the learning of *iobj* and *csubj* relations remained unstable, with low accuracy even in the final stages of pretraining in many languages. These findings suggest that specific dependency relations are inherently harder for LLMs to internalize during pretraining, and that LoRA probing provides a practical means to observe and quantify such cross-lingual differences in syntactic acquisition.

Acknowledgments

This work was conducted as part of a collaborative research project between Recruit Co., Ltd. and the National Institute for Japanese Language and Linguistics. We are grateful to all those involved in the management and support of this project. We would also like to express our sincere gratitude to Yuji Matsumoto of RIKEN AIP for his valuable advice from the early stages of this research. Finally, we thank the anonymous reviewers for their constructive and detailed comments.

Limitations

Effects of test set contamination in language model pretraining

The Universal Dependencies (UD) datasets used in our experiments, including the test sets, are freely and publicly available on the internet. Meanwhile, all four large language models evaluated in this study were pretrained on large-scale web corpora. Consequently, when evaluating UD tasks with these LLMs, there is a potential risk that test set contamination may artificially inflate performance.

According to Matsuda et al. (2025), such contamination may contribute to higher initial accuracy in UPOS tagging; however, no significant effects have been observed for UAS or LAS. In our experiments, because all metrics—TOKEN RECALL, UPOS, UAS,

and LAS—were evaluated using the same base model configurations, any contamination effects are expected to cancel out. Therefore, we believe that the overall conclusions of this study remain unaffected.

Use of an English-centric LLM for multilingual evaluation

In this study, we evaluated multilingual syntactic parsing tasks using `OLMo-2-7B` and its intermediate checkpoints from pretraining. However, it should be noted that the DCLM-baseline dataset, which constitutes the majority of `OLMo-2-7B`'s pretraining data, is composed of texts that passed an English filter, retaining only documents estimated to have a probability of over 99% of being English. Consequently, non-English languages are at an inherent disadvantage in the evaluation setting.

Bibliographical References

- BigScience Workshop, :, Teven Le Scao, et al. 2023. [Bloom: A 176b-parameter open-access multilingual language model](#).
- Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. [Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3575–3590, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Matthew S. Dryer and Martin Haspelmath, editors. 2013. [WALS Online \(v2020.4\)](#). Zenodo.
- Gemma Team, Morgane Riviere, et al. 2024. [Gemma 2: Improving open language models at a practical size](#).
- Aaron Grattafiori et al. 2024. [The llama 3 herd of models](#).
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Jeffrey Li et al. 2025. [Datacomp-lm: In search of the next generation of training sets for language models](#).
- Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. [Probing across time: What does RoBERTa know and when?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Hiroshi Matsuda, Chunpeng Ma, and Masayuki Asahara. 2025. [Step-by-step instructions and a simple tabular output format improve the dependency parsing accuracy of LLMs](#). In *Proceedings of the 18th International Conference on Parsing Technologies (IWPT, SyntaxFest 2025)*, pages 11–19, Ljubljana, Slovenia. Association for Computational Linguistics.
- 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. European Language Resources Association.
- Laura Pérez-Mayos, Miguel Ballesteros, and Leo Wanner. 2021. [How much pretraining data do language models need to learn syntax?](#) In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1571–1582, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Qwen, :, An Yang, et al. 2025. [Qwen2.5 technical report](#).
- Team OLMo, Pete Walsh, et al. 2025. [2 olmo 2 furious](#).
- Yuanhe Tian, Fei Xia, and Yan Song. 2024. [Large language models are no longer shallow parsers](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7131–7142, Bangkok, Thailand. Association for Computational Linguistics.
- Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024. [Probing the emergence of cross-lingual alignment during LLM training](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12159–12173, Bangkok, Thailand. Association for Computational Linguistics.

Language Resource References

- Agić, Željko and Ljubešić, Nikola. 2015. [Universal Dependencies for Croatian \(that work for Serbian, too\)](#). INCOMA Ltd. Shoumen, BULGARIA.

- Asahara, Masayuki and Kanayama, Hiroshi and Tanaka, Takaaki and Miyao, Yusuke and Uematsu, Sumire and Mori, Shinsuke and Matsumoto, Yuji and Omura, Mai and Murawaki, Yugo. 2018. *Universal Dependencies Version 2 for Japanese*. European Language Resources Association (ELRA).
- Batanović, Vuk and Ljubešić, Nikola and Samardžić, Tanja and Erjavec, Tomaž. 2018. *Training corpus SETimes.SR 1.0*. Slovenian language resource repository CLARIN.SI.
- Agnė Bielinskienė and Loïc Boizou and Jolanta Kovalevskaitė and Erika Rimkutė. 2016. *Lithuanian Dependency Treebank ALKSNIŠ*. In: I. Skadiņa and R. Rozis (Eds.): *Human Language Technologies – The Baltic Perspective*. Amsterdam: IOS Press.
- Bosco, Cristina and Montemagni, Simonetta and Simi, Maria. 2013. *Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank*. Association for Computational Linguistics.
- Bouma, Gosse and van Noord, Gertjan. 2017. *Increasing Return on Annotation Investment: The Automatic Construction of a Universal Dependency Treebank for Dutch*. Association for Computational Linguistics.
- Chun, Jayeol and Han, Na-Rae and Hwang, Jena D. and Choi, Jinho D. 2018. *Building Universal Dependency Treebanks in Korean*. European Language Resources Association (ELRA).
- Dobrovoljc, Kaja and Erjavec, Tomaž and Krek, Simon. 2017. *The Universal Dependencies Treebank for Slovenian*. Association for Computational Linguistics.
- Kira Droganova and Olga Lyashevskaya and Daniel Zeman. 2018. *Data Conversion and Consistency of Monolingual Corpora: Russian UD Treebanks*. Linköping University Electronic Press.
- Guillaume, Bruno and de Marneffe, Marie-Catherine and Perrier, Guy. 2019. *Conversion et améliorations de corpus du français annotés en Universal Dependencies [Conversion and Improvement of Universal Dependencies French corpora]*. ATALA (Association pour le Traitement Automatique des Langues).
- Hajič, Jan and Smrž, Otakar and Zemánek, Petr and Pajas, Petr and Šnaidauf, Jan and Beška, Emanuel and Kracmar, Jakub and Hassanová, Kamila. 2009. *Prague Arabic Dependency Treebank 1.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).
- Hajič, Jan and others. 2024. *Prague Dependency Treebank - Consolidated 2.0 (PDT-C 2.0)*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).
- Anders Johannsen and Héctor Martínez Alonso and Barbara Plank. 2015. *Universal Dependencies for Danish*.
- Lynn, Teresa and Foster, Jennifer. 2016. *Universal Dependencies for Irish*. Association pour le Traitement Automatique des Langues. Dépendances universelles de l'irlandais.
- Marşan, Büşra and Akkurt, Salih Furkan and Şen, Muhammet and Gürbüz, Merve and Güngör, Onur and Özateş, Şaziye Betül and Üsküdarlı, Suzan and Özgür, Arzucan and Güngör, Tunga and Öztürk, Balkız. 2022. *Enhancements to the BOUN Treebank Reflecting the Agglutinative Nature of Turkish*.
- McDonald, Ryan and Nivre, Joakim and Quirmbach-Brundage, Yvonne and Goldberg, Yoav and Das, Dipanjan and Ganchev, Kuzman and Hall, Keith and Petrov, Slav and Zhang, Hao and Täckström, Oscar and Bedini, Claudia and Bertomeu Castelló, Núria and Lee, Jungmee. 2013. *Universal Dependency Annotation for Multilingual Parsing*. Association for Computational Linguistics.
- Verginica Barbu Mititelu. 2018. *Modern Syntactic Analysis of Romanian*. Publishing House of "Alexandru Ioan Cuza" University.
- Muischnek, Kadri and Müürisep, Kaili and Puolakainen, Tiina. 2016. *Estonian Dependency Treebank: from Constraint Grammar tagset to Universal Dependencies*. European Language Resources Association (ELRA).
- Nivre, Joakim and Megyesi, Beáta. 2007. *Bootstrapping a Swedish Treebank Using Cross-Corpus Harmonization and Annotation Projection*.
- Petya Osenova and Kiril Simov. 2004. *BTB-TR05: BulTreeBank Stylebook 4 05*.
- Pretkalniundefineda, Lauma and Rituma, Laura and Saulundefinedte, Baiba. 2018. *Deriving Enhanced Universal Dependencies from a Hybrid Dependency-Constituency Treebank*. Springer-Verlag.
- Prokopidis, Prokopis and Papageorgiou, Haris. 2017. *Universal Dependencies for Greek*. Association for Computational Linguistics.
- Shishkina, Yana and Lyashevskaya, Olga. 2021. *Sculpting Enhanced Dependencies for Belarusian*. Springer-Verlag.

Natalia Silveira and Timothy Dozat and Marie-Catherine de Marneffe and Samuel Bowman and Miriam Connor and John Bauer and Christopher D. Manning. 2014. *A Gold Standard Dependency Corpus for English*.

Taulé, Mariona and Martí, M. Antònia and Recasens, Marta. 2008. *AnCorà: Multilevel Annotated Corpora for Catalan and Spanish*. European Language Resources Association (ELRA).

Wróblewska, Alina. 2018. *Extended and Enhanced Polish Dependency Bank in Universal Dependencies Format*. Association for Computational Linguistics.

Zeman, Daniel. 2017. *Slovak Dependency Treebank in Universal Dependencies*.

A. Settings

A.1. Environment and Hyper-parameters

The environment and the hyper-parameters used in the experiments are shown in Table 4 and 5, respectively.

Hardware	
GPU	8 × NVIDIA H100 80 GB
CPU	Intel Xeon 208 core
RAM	1,360 GB
SSD	10 TB
Software	
OS	Ubuntu 22.04
CUDA	12.6
Python	3.12.11
PyTorch	2.7.0
Transformers	4.51.3
TRL	0.18.1
PEFT	0.15.2
vLLM	0.9.1

Table 4: Experimental environment.

Model Comparison Experiments	
Learning Rate	3.0e-4
Epochs	3
Intermediate Checkpoint Experiments	
Learning Rate	1.0e-4
The training epochs are adjusted based on the number of words in the training set to ensure that the amount of training for each language is equivalent.	

Table 5: Hyper-parameters used in LoRA fine-tuning.

A.2. Model Specifications

The specifications of LLMs used in the experiments are shown in Table 6.

B. Comparison between 3-step and 2-step-no-form prompts

The detailed results of the experiment for the comparison between 3-step and 2-step-no-form prompts are shown in Table 7 and 8.

	Model Parameters	Tokenizer Vocab Size	Context Length	Training Tokens	LoRA Trainable Parameters	Supported Languages
gemma-2-9b	9.27B	256k	8,192	8T	27.0M	(multilingual)
Llama-3.1-8B	8.05B	128k	131,072	15T	21.0M	en, de, fr, it, pt, hi, es, th
Qwen2.5-7B	7.64B	152k	131,072	18T	20.2M	zh, en, fr, es, pt, de, it, ru, ja, ko, vi, th, ar, and more
OLMo-2-7B	7.32B	100k	4,096	4T	20.0M	en

Table 6: Model specs.

	Token Recall								UPOS (F1)							
	3-step				2-step-no-form				3-step				2-step-no-form			
	gemma	Llama	Qwen		gemma	Llama	Qwen	OLMo	gemma	Llama	Qwen		gemma	Llama	Qwen	OLMo
ar_padt	99.4	99.4	99.8		100.0	100.0	100.0	100.0	97.0	97.1	97.2		97.3	97.1	97.1	96.6
be_hse	98.4	98.4	98.3		100.0	100.0	100.0	100.0	98.7	98.8	98.6		98.9	98.7	98.7	98.2
bg_btb	99.8	100.0	99.5		100.0	100.0	100.0	100.0	99.3	99.4	99.0		99.4	99.4	99.3	98.9
ca_ancora	99.4	99.6	99.5		100.0	100.0	100.0	100.0	98.9	99.0	99.0		99.2	99.2	99.1	98.9
cs_pdt	99.9	99.6	99.9		100.0	100.0	100.0	100.0	99.3	99.3	99.3		99.4	99.3	99.3	99.1
da_ddt	100.0	99.3	98.0		100.0	100.0	99.8	100.0	98.6	98.1	97.9		98.8	98.4	97.9	97.1
de_gsd	100.0	99.9	99.5		100.0	100.0	100.0	99.8	97.0	97.0	97.2		97.3	97.3	97.2	96.8
el_gdt	99.6	99.3	99.8		100.0	99.7	100.0	100.0	98.2	98.4	97.7		98.3	97.7	97.9	97.2
en_ewt	100.0	99.9	99.4		100.0	99.9	100.0	100.0	98.6	98.3	98.3		98.4	98.4	98.5	98.0
es_ancora	100.0	99.8	99.6		100.0	100.0	100.0	100.0	99.1	99.2	99.1		99.2	99.2	99.1	99.0
et_edt	99.9	99.9	99.8		100.0	100.0	100.0	99.9	98.2	98.1	97.6		98.2	98.0	97.5	96.8
fi_ftb	99.6	99.7	99.9		100.0	100.0	100.0	100.0	98.1	97.8	96.6		98.2	97.8	96.8	96.4
fr_gsd	100.0	100.0	98.9		100.0	100.0	100.0	100.0	98.6	98.7	98.6		98.6	98.6	98.5	98.5
ga_idt	99.0	98.8	99.1		100.0	100.0	100.0	99.8	95.6	96.0	95.8		96.3	96.4	95.8	95.1
hr_set	99.6	99.3	99.2		100.0	100.0	100.0	100.0	98.7	98.5	98.3		98.7	98.7	98.4	97.9
it_isdt	100.0	99.7	99.4		100.0	100.0	100.0	100.0	98.6	98.7	98.4		98.7	98.8	98.8	98.3
ja_gsd	99.2	100.0	99.6		100.0	100.0	99.6	100.0	98.9	98.9	98.8		98.9	98.8	98.4	98.2
ko_gsd	99.8	99.6	99.8		100.0	100.0	100.0	100.0	96.5	96.9	97.0		97.2	96.9	96.9	96.0
lt_alksnis	99.5	98.2	97.2		100.0	100.0	100.0	100.0	97.4	96.3	95.1		97.3	96.2	95.4	91.8
lv_lvrb	100.0	99.7	99.6		100.0	100.0	100.0	100.0	98.3	97.8	97.7		98.3	98.0	97.8	97.3
nl_alpino	100.0	99.7	98.7		100.0	100.0	100.0	100.0	98.1	98.2	97.8		98.4	98.1	98.3	97.5
no_bokmaal	99.9	99.9	99.2		100.0	99.9	100.0	100.0	98.7	98.7	98.5		98.8	98.7	98.6	98.2
pl_pdb	99.9	99.4	99.4		100.0	100.0	100.0	100.0	99.4	99.2	99.2		99.4	99.3	99.3	98.7
pt_gsd	99.9	99.8	99.9		100.0	100.0	100.0	100.0	98.1	97.9	97.9		98.2	98.2	97.8	97.5
ro_rrt	99.8	99.7	99.3		100.0	100.0	100.0	100.0	98.1	98.2	97.9		98.3	98.3	98.1	97.7
ru_syntagrus	100.0	99.8	99.9		100.0	100.0	100.0	100.0	98.8	98.8	99.0		99.0	98.9	98.9	98.7
sk_snk	99.8	99.1	98.9		100.0	100.0	100.0	100.0	98.2	98.0	97.9		98.3	98.0	97.9	96.8
sl_ssj	99.8	99.4	99.6		100.0	99.9	100.0	100.0	98.5	98.8	98.5		98.9	98.9	98.7	98.1
sr_set	99.9	99.7	99.6		100.0	100.0	99.6	100.0	99.1	98.9	98.8		99.0	98.9	98.8	98.4
sv_talbanken	99.8	99.6	99.1		100.0	100.0	100.0	100.0	98.9	98.6	98.3		98.9	98.8	98.5	97.9
tr_boun	99.9	99.0	98.3		100.0	100.0	100.0	100.0	93.6	93.3	92.9		93.4	93.0	92.9	92.3
uk_iu	99.7	98.7	98.1		100.0	99.8	99.9	99.8	98.5	98.2	98.0		98.6	98.0	98.0	96.3
zh_gsdsimp	100.0	100.0	99.5		99.9	99.9	99.9	99.9	97.3	97.1	96.7		97.1	96.9	97.3	95.6
Macro Avg.	99.7	99.5	99.3		100.0	100.0	100.0	100.0	98.1	98.1	97.8		98.3	98.1	97.9	97.3

Table 7: Comparison of Token Recall and UPOS (F1) between 3-step and 2-step-no-form among the four models; gemma-2-9b, Llama-3.1-8B, Qwen2.5-7B, and OLMo-2-7B.

	Unlabeld Attachment Score (F1)								Labeld Attachment Score (F1)							
	3-step				2-step-no-form				3-step				2-step-no-form			
	gemma	Llama	Qwen		gemma	Llama	Qwen	OLMo	gemma	Llama	Qwen		gemma	Llama	Qwen	OLMo
ar_padt	90.4	90.4	90.1		90.7	90.2	89.9	87.2	85.9	86.0	85.6		86.4	85.8	85.7	82.2
be_hse	92.3	92.1	90.7		92.4	91.8	90.6	87.0	90.5	90.4	88.7		90.8	90.0	88.7	84.4
bg_btb	96.4	96.0	95.4		96.2	96.1	95.4	93.8	94.1	93.7	93.0		94.0	93.7	93.0	90.7
ca_ancora	94.5	95.0	94.6		95.3	95.2	94.9	93.4	92.9	93.6	93.1		94.0	93.8	93.4	91.6
cs_pdt	95.1	95.6	95.4		95.8	95.6	95.4	93.9	93.6	94.2	94.0		94.5	94.2	94.1	92.1
da_ddt	90.3	88.8	87.4		90.5	88.8	87.7	85.7	88.7	86.7	85.2		88.7	86.6	85.4	82.9
de_gsd	89.1	90.3	89.9		89.3	89.9	89.7	89.2	85.3	86.6	86.2		85.7	86.1	86.0	85.2
el_gdt	94.7	94.4	93.1		94.2	94.2	92.5	89.7	93.2	92.5	91.0		92.4	92.2	90.3	86.7
en_ewt	95.4	94.8	94.3		95.2	95.1	94.8	94.2	93.9	93.2	92.7		93.5	93.4	93.1	92.4
es_ancora	93.7	94.3	94.1		94.6	94.4	94.0	93.1	92.0	92.8	92.6		93.2	92.9	92.4	91.2
et_edt	92.5	91.3	89.9		92.1	90.9	89.5	86.4	90.4	89.0	87.2		89.9	88.6	86.8	83.3
fi_ftb	95.2	94.0	92.2		95.1	94.1	92.5	90.6	93.5	91.9	89.1		93.4	91.9	89.6	87.2
fr_gsd	95.6	95.9	95.7		96.0	95.8	95.9	94.5	94.0	94.2	94.2		94.6	94.4	94.4	92.8
ga_idt	88.4	87.7	86.3		88.2	88.2	86.6	85.6	83.2	82.3	80.0		82.7	82.7	80.6	78.8
hr_set	94.0	93.7	92.5		94.1	93.6	92.9	90.8	91.1	90.7	89.4		91.2	90.6	89.6	87.1
it_isdt	95.6	95.9	95.2		95.8	96.1	95.7	94.4	94.0	94.5	93.7		94.4	94.7	94.4	92.6
ja_gsd	95.3	94.9	95.0		95.1	94.9	94.3	93.1	94.3	94.0	93.8		93.9	93.5	92.9	91.5
ko_gsd	87.4	89.1	89.3		90.3	89.3	89.3	85.8	83.9	86.0	86.3		87.2	86.1	86.1	81.9
lt_alksnis	88.8	85.3	81.9		88.3	85.0	81.1	71.6	85.6	81.6	77.5		84.9	81.0	76.6	64.9
lv_lvrb	94.1	92.1	91.5		93.8	92.4	91.7	88.3	91.6	89.4	88.5		91.4	89.7	88.8	84.9
nl_alpino	94.0	94.8	94.1		95.3	94.6	94.3	91.8	91.3	92.7	91.8		93.4	92.5	92.1	88.6
no_bokmaal	94.7	95.2	94.2		95.6	95.1	94.6	93.2	93.5	94.0	93.0		94.6	94.0	93.4	91.5
pl_pdb	97.0	96.6	96.2		97.0	96.7	96.4	94.1	95.7	95.1	94.7		95.6	95.2	94.8	91.9
pt_gsd	94.0	93.7	93.4		94.0	93.9	93.5	92.6	92.1	91.6	91.1		92.1	92.0	91.4	90.0
ro_rrt	92.5	94.1	92.6		94.5	93.8	92.8	92.0	88.9	90.7	89.3		91.3	90.6	89.5	88.3
ru_syntagrus	95.2	95.5	95.6		95.8	95.5	95.6	94.7	93.1	93.5	93.7		93.9	93.5	93.6	92.5
sk_snk	96.2	95.3	94.6		95.8	94.8	94.6	89.6	94.9	93.8	92.9		94.4	93.3	92.9	86.8
sl_ssj	94.0	95.0	94.3		95.7	95.3	94.6	91.4	92.4	93.5	92.7		94.4	93.9	93.0	89.1
sr_set	95.1	94.5	93.8		95.2	94.2	94.1	91.6	93.0	92.2	91.3		92.9	91.7	91.4	88.7
sv_talbanken	94.4	93.1	92.6		94.3	93.6	92.8	90.6	92.5	90.9	90.1		92.3	91.4	90.6	87.1
tr_boun	83.0	82.1	80.4		83.2	81.9	80.8	78.0	76.6	75.5	73.5		76.7	75.2	73.9	70.8
uk_iu	94.6	93.6	92.8		94.1	92.8	92.4	86.3	92.8	91.5	90.5		92.2	90.5	89.9	82.8
zh_gsdsimp	89.1	88.2	88.2		88.1	88.6	89.1	83.0	86.7	85.5	85.6		85.5	85.8	86.5	79.4
Macro Avg.	93.1	92.8	92.0		93.4	92.8	92.1	89.6	90.8	90.4	89.5		91.1	90.3	89.5	86.4

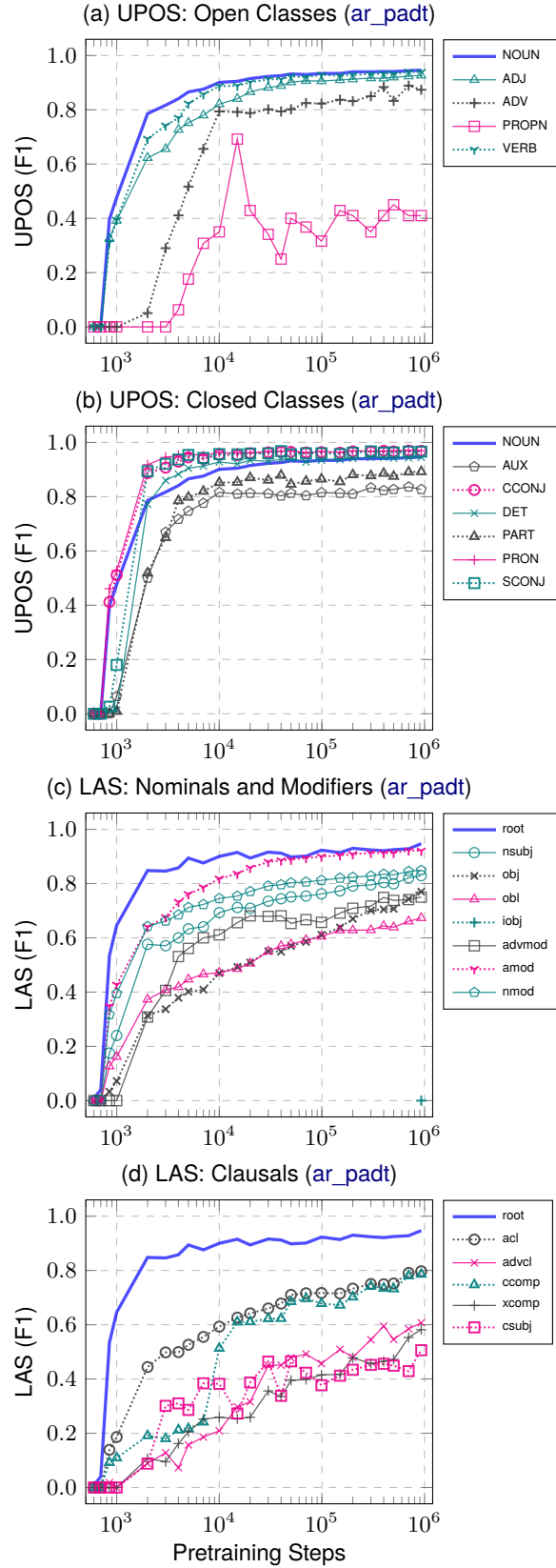
Table 8: Comparison of UAS (F1) and LAS (F1) between 3-step and 2-step-no-form among the four models; gemma-2-9b, Llama-3.1-8B, Qwen2.5-7B, and OLMo-2-7B.

C. Fine-grained Analysis of Language-wise Syntactic Acquisition Process

In this section, we present the results of the analysis of the syntactic acquisition process, organized into subsections for each language. Each subsection includes the learning curves of major UPOS tags and DEPREL labels, as well as statistics of the treebank used in the experiments. The subsections are sorted in alphabetical order by language name. The directories categorized by language family are as follows.

- Germanic
 - Danish ([da_ddt](#)): C.8
 - Dutch ([nl_alpino](#)): C.9
 - English ([en_ewt](#)): C.10
 - German ([de_gsd](#)): C.14
 - Norwegian ([no_bokmaal](#)): C.22
 - Swedish ([sv_talbanken](#)): C.31
- Romance
 - Catalan ([ca_ancora](#)): C.4
 - French ([fr_gsd](#)): C.13
 - Italian ([it_isdt](#)): C.17
 - Portuguese ([pt_gsd](#)): C.24
 - Romanian ([ro_rrt](#)): C.25
 - Spanish ([es_ancora](#)): C.30
- Slavic
 - Belarusian ([be_hse](#)): C.2
 - Bulgarian ([bg_btb](#)): C.3
 - Croatian ([hr_set](#)): C.6
 - Czech ([cs_pdt](#)): C.7
 - Polish ([pl_pdb](#)): C.23
 - Russian ([ru_syntagrus](#)): C.26
 - Serbian ([sr_set](#)): C.27
 - Slovak ([sk_snk](#)): C.28
 - Slovenian ([sl_ssj](#)): C.29
 - Ukrainian ([uk_iu](#)): C.33
- Greek
 - Greek ([el_gdt](#)): C.15
- Finnic
 - Estonian ([et_edt](#)): C.11
 - Finnish ([fi_ftb](#)): C.12
- Celtic
 - Irish ([ga_idt](#)): C.16
- Baltic
 - Latvian ([lv_lvtb](#)): C.20
 - Lithuanian ([lt_alksnis](#)): C.21
- Turkic
 - Turkish ([tr_boun](#)): C.32
- Afro-Asiatic
 - Arabic ([ar_padt](#)): C.1
- Japanese
 - Japanese ([ja_gsd](#)): C.18
- Korean
 - Korean ([ko_gsd](#)): C.19
- Sino-Tibetan
 - Chinese ([zh_gsdsimp](#)): C.5

C.1. Arabic

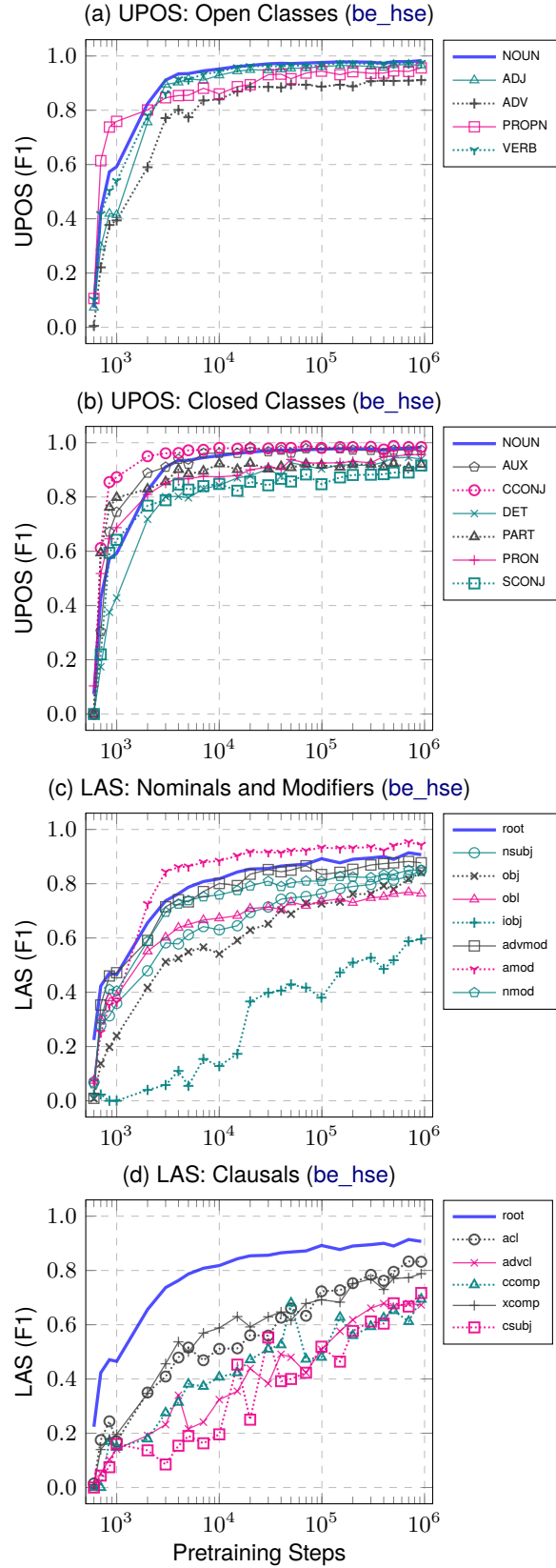


	Train	Dev	Test
Number of sentences	6,075	909	680
Number of words	223,881	30,239	28,264
Number of label occurrences			
ADJ	23,498	2,916	2,937
ADP	33,617	4,410	4,528
ADV	880	108	104
AUX	1,699	258	197
CCONJ	15,803	2,018	1,963
DET	4,648	625	623
INTJ	7	1	0
NOUN	74,546	9,612	9,547
NUM	6,010	969	779
PART	1,709	270	226
PRON	8,533	1,211	1,133
PROPN	187	27	31
PUNCT	17,511	2,882	2,052
SCONJ	4,368	555	534
SYM	329	18	41
VERB	16,789	2,318	2,189
X	13,747	2,041	1,380
<i>acl</i>	4,193	577	567
<i>advcl</i>	1,302	185	208
<i>advmod</i>	2,546	333	337
<i>amod</i>	19,555	2,419	2,447
<i>appos</i>	617	137	68
<i>aux</i>	1,333	214	147
<i>case</i>	31,889	4,171	4,279
<i>cc</i>	11,646	1,349	1,447
<i>ccomp</i>	2,537	356	319
<i>conj</i>	10,640	1,303	1,223
<i>cop</i>	366	44	50
<i>csubj</i>	442	51	42
<i>dep</i>	1,131	331	104
<i>det</i>	1,707	222	221
<i>discourse</i>	1	0	0
<i>dislocated</i>	172	18	13
<i>fixed</i>	2,175	267	263
<i>flat</i>	596	120	29
<i>iobj</i>	89	12	6
<i>mark</i>	6,087	791	764
<i>nmod</i>	50,215	6,501	6,351
<i>nsubj</i>	14,683	2,035	1,890
<i>nummod</i>	2,939	422	382
<i>obj</i>	6,710	882	879
<i>obl</i>	20,786	2,783	2,786
<i>orphan</i>	174	17	5
<i>parataxis</i>	4,275	663	514
<i>punct</i>	17,506	2,882	2,052
<i>root</i>	6,075	909	680
<i>xcomp</i>	1,494	245	191

Table 9: Statistics of *ar_padt*

Figure 9: Syntactic acquisition through pretraining of OLMo-2-7B on Arabic (*ar_padt*).

C.2. Belarusian

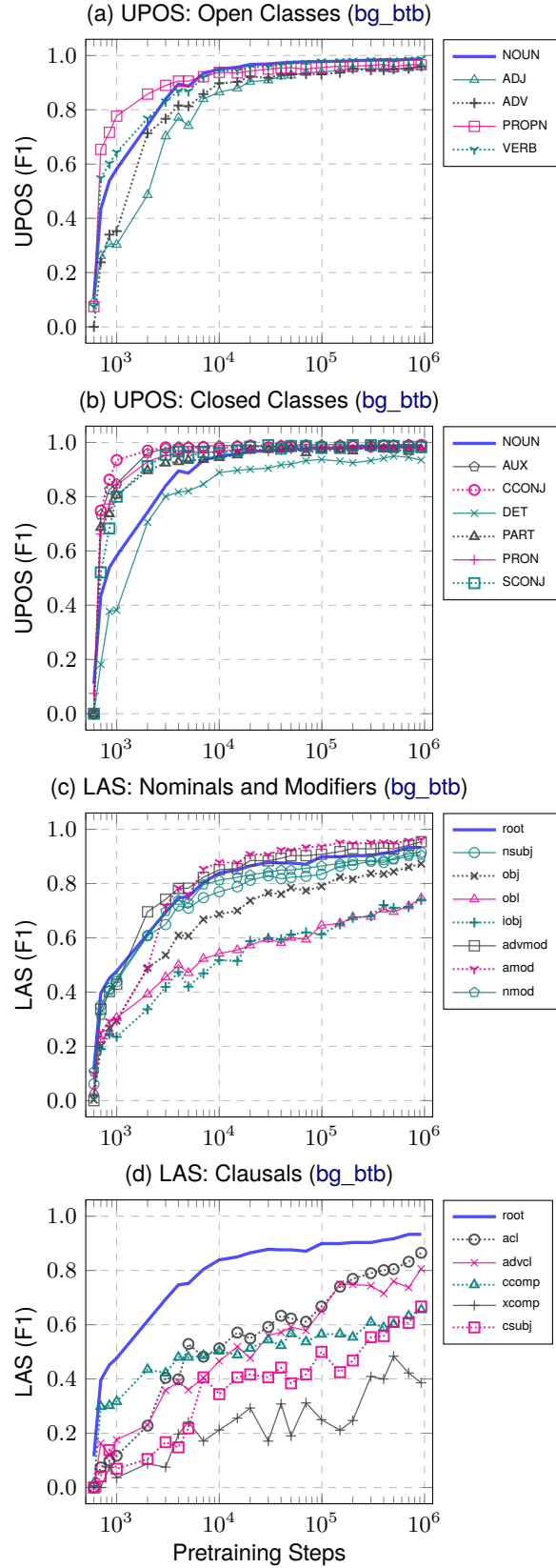


	Train	Dev	Test
Number of sentences	22,852	1,301	1,077
Number of words	273,181	15,931	15,997
Number of label occurrences			
ADJ	23,293	1,630	1,893
ADP	26,423	1,605	1,643
ADV	10,148	495	397
AUX	1,893	101	90
CCONJ	7,922	437	641
DET	6,093	288	349
INTJ	86	3	1
NOUN	63,986	4,080	4,620
NUM	5,326	287	233
PART	4,534	219	134
PRON	9,479	477	367
PROPN	18,323	1,114	946
PUNCT	51,192	3,069	2,754
SCONJ	3,146	161	95
SYM	2,454	88	70
VERB	28,254	1,469	1,369
X	10,629	408	395
<i>acl</i>	3,195	180	239
<i>advcl</i>	1,308	64	56
<i>advmod</i>	12,652	620	426
<i>amod</i>	18,585	1,290	1,510
<i>appos</i>	5,327	318	398
<i>aux</i>	990	62	50
<i>case</i>	26,582	1,596	1,630
<i>cc</i>	7,973	438	628
<i>ccomp</i>	1,296	61	25
<i>compound</i>	537	29	12
<i>conj</i>	11,866	767	984
<i>cop</i>	828	38	35
<i>csubj</i>	828	48	31
<i>dep</i>	5,629	158	175
<i>det</i>	5,212	227	306
<i>discourse</i>	269	13	38
<i>dislocated</i>	14	0	0
<i>expl</i>	196	10	1
<i>fixed</i>	771	80	158
<i>flat</i>	5,582	370	310
<i>goeswith</i>	1	1	0
<i>iobj</i>	2,269	112	82
<i>list</i>	1,704	64	42
<i>mark</i>	2,774	141	82
<i>nmod</i>	24,364	1,598	2,044
<i>nsubj</i>	18,057	1,012	866
<i>nummod</i>	3,219	204	124
<i>obj</i>	10,183	544	476
<i>obl</i>	17,276	1,042	1,011
<i>orphan</i>	219	23	13
<i>parataxis</i>	6,278	290	255
<i>punct</i>	51,174	3,069	2,754
<i>reparandum</i>	5	0	0
<i>root</i>	22,853	1,301	1,077
<i>vocative</i>	155	2	11
<i>xcomp</i>	3,010	159	148

Table 10: Statistics of *be_hse*

Figure 10: Syntactic acquisition through pretraining of OLMo-2-7B on Belarusian (*be_hse*).

C.3. Bulgarian

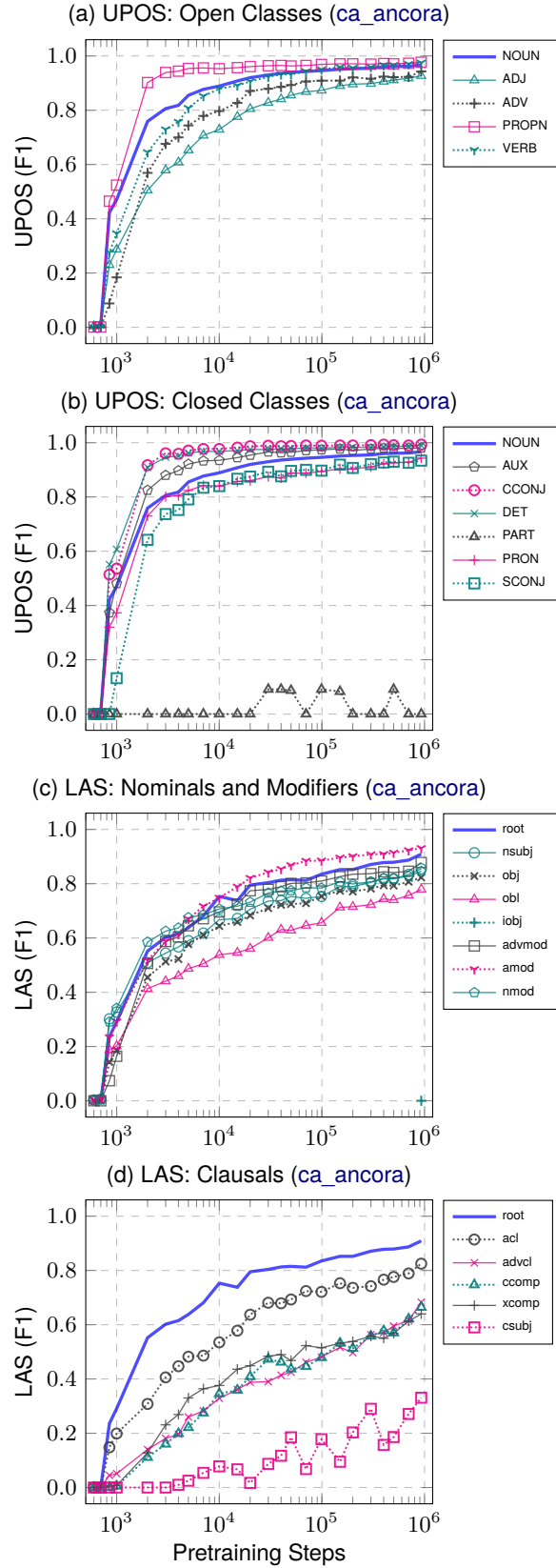


	Train	Dev	Test
Number of sentences	8,907	1,115	1,116
Number of words	124,336	16,089	15,724
Number of label occurrences			
ADJ	10,813	1,401	1,377
ADP	17,572	2,287	2,237
ADV	5,250	637	671
AUX	7,255	963	916
CCONJ	3,885	508	468
DET	1,872	288	273
INTJ	26	1	1
NOUN	27,147	3,519	3,486
NUM	1,668	215	222
PART	1,728	229	210
PRON	8,052	1,061	981
PROPN	6,797	833	805
PUNCT	17,523	2,267	2,269
SCONJ	1,276	174	156
VERB	13,470	1,706	1,652
X	2	0	0
<i>acl</i>	1,282	159	161
<i>advcl</i>	1,276	173	164
<i>advmod</i>	5,691	710	718
<i>amod</i>	9,400	1,240	1,195
<i>appos</i>	28	2	4
<i>aux</i>	4,994	646	585
<i>case</i>	17,210	2,225	2,182
<i>cc</i>	3,891	507	468
<i>ccomp</i>	1,625	298	266
<i>compound</i>	2	0	0
<i>conj</i>	4,415	593	562
<i>cop</i>	1,899	216	270
<i>csubj</i>	358	50	43
<i>det</i>	2,699	401	362
<i>discourse</i>	557	66	68
<i>expl</i>	2,676	389	320
<i>fixed</i>	560	96	83
<i>flat</i>	1,424	136	135
<i>iobj</i>	2,818	399	346
<i>mark</i>	1,460	200	179
<i>nmod</i>	11,374	1,509	1,527
<i>nsubj</i>	9,177	1,167	1,157
<i>nummod</i>	1,385	175	169
<i>obj</i>	5,366	642	713
<i>obl</i>	5,195	661	622
<i>parataxis</i>	426	4	10
<i>punct</i>	17,523	2,267	2,269
<i>root</i>	8,907	1,115	1,116
<i>vocative</i>	55	1	2
<i>xcomp</i>	663	42	28

Table 11: Statistics of *bg_btb*

Figure 11: Syntactic acquisition through pretraining of OLMo-2-7B on Bulgarian (*bg_btb*).

C.4. Catalan

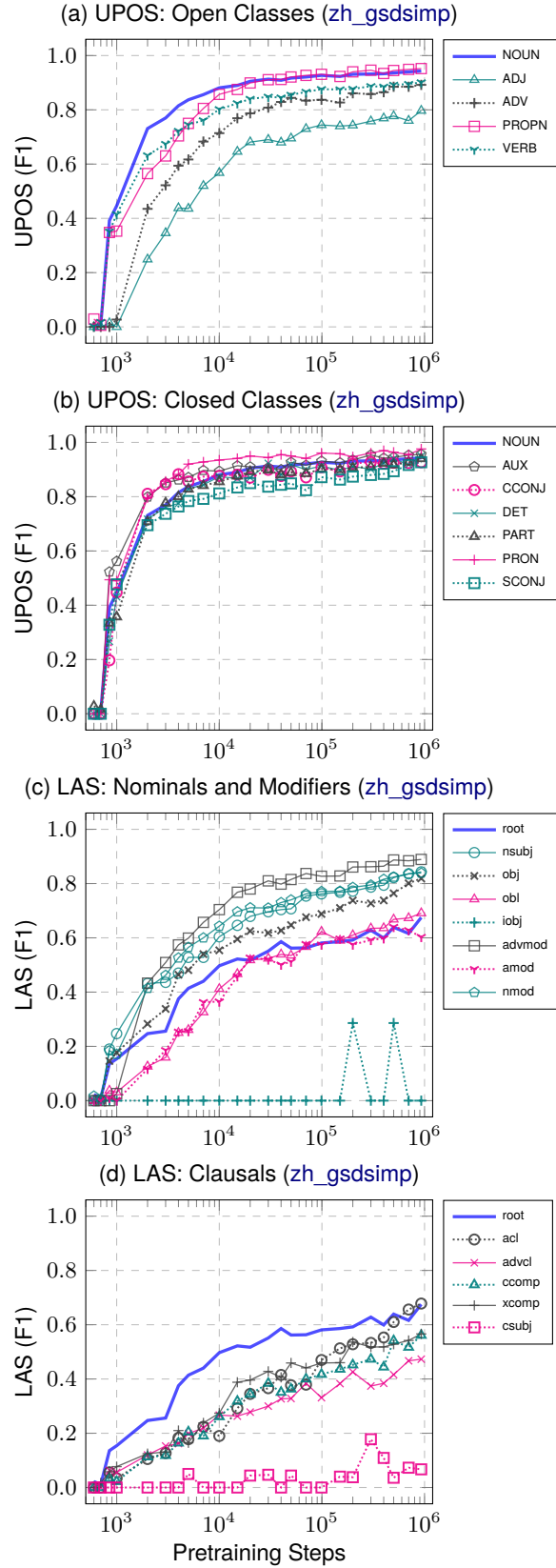


	Train	Dev	Test
Number of sentences	13,123	1,709	1,846
Number of words	429,578	58,073	59,610
Number of label occurrences			
ADJ	23,607	3,218	3,257
ADP	69,040	9,329	9,607
ADV	12,149	1,733	1,600
AUX	17,370	2,382	2,300
CCONJ	11,797	1,647	1,605
DET	68,331	9,369	9,567
INTJ	11	1	3
NOUN	77,539	10,453	10,654
NUM	7,917	917	1,128
PART	113	13	21
PRON	18,498	2,469	2,487
PROPN	36,245	4,782	5,563
PUNCT	44,968	5,961	6,341
SCONJ	8,366	1,185	999
SYM	648	68	105
VERB	32,979	4,546	4,373
<i>acl</i>	7,866	1,081	1,058
<i>advcl</i>	5,381	788	684
<i>advmod</i>	11,961	1,695	1,564
<i>amod</i>	19,615	2,653	2,735
<i>appos</i>	5,960	784	934
<i>aux</i>	13,153	1,837	1,760
<i>case</i>	61,122	8,251	8,560
<i>cc</i>	11,656	1,609	1,612
<i>ccomp</i>	3,370	437	385
<i>compound</i>	2,092	229	331
<i>conj</i>	12,992	1,747	1,786
<i>cop</i>	3,749	463	467
<i>csubj</i>	736	91	101
<i>dep</i>	71	11	7
<i>det</i>	67,526	9,258	9,450
<i>discourse</i>	1	0	0
<i>dislocated</i>	1	0	1
<i>expl</i>	459	54	59
<i>fixed</i>	7,746	1,079	1,047
<i>flat</i>	13,816	1,825	2,373
<i>mark</i>	12,083	1,705	1,466
<i>nmod</i>	32,522	4,304	4,342
<i>nsubj</i>	21,624	2,901	2,962
<i>nummod</i>	4,912	622	737
<i>obj</i>	23,997	3,350	3,209
<i>obl</i>	23,449	3,096	3,304
<i>parataxis</i>	438	69	71
<i>punct</i>	44,968	5,961	6,341
<i>root</i>	13,123	1,709	1,846
<i>vocative</i>	1	0	0
<i>xcomp</i>	3,188	464	418

Table 12: Statistics of *ca_ancora*

Figure 12: Syntactic acquisition through pretraining of OLMo-2-7B on Catalan (*ca_ancora*).

C.5. Chinese

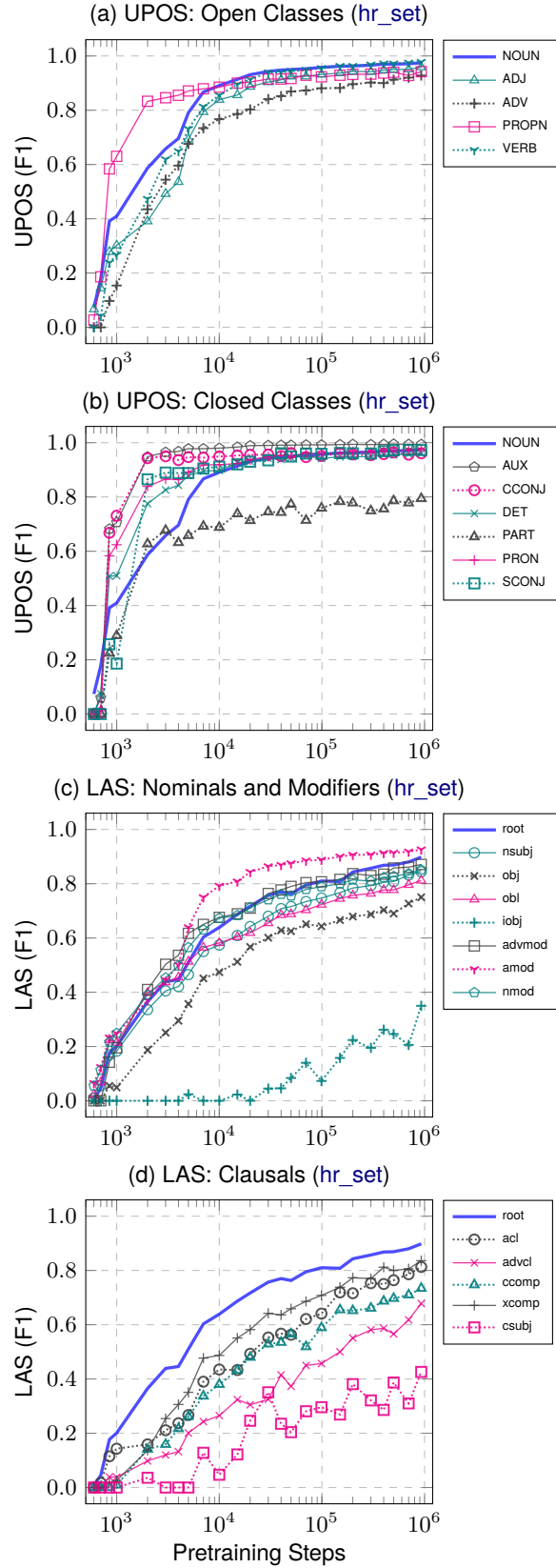


	Train	Dev	Test
Number of sentences	3,997	500	500
Number of words	98,616	12,663	12,012
Number of label occurrences			
ADJ	2,446	312	272
ADP	4,539	555	535
ADV	2,467	322	292
AUX	3,105	365	422
CCONJ	1,385	164	191
DET	1,071	119	139
NOUN	27,097	3,637	3,312
NUM	5,353	654	653
PART	7,847	1,041	994
PRON	1,433	175	168
PROPN	8,678	1,056	1,006
PUNCT	13,627	1,770	1,691
SCONJ	4,137	560	505
SYM	11	3	0
VERB	14,623	1,851	1,745
X	797	79	87
acl	2,034	268	252
advcl	3,232	447	343
advmod	2,552	329	304
amod	1,659	196	190
appos	1,122	128	103
aux	1,812	192	248
case	7,287	919	862
cc	1,391	164	191
ccomp	1,739	218	203
clf	1,814	200	233
compound	4,402	598	590
conj	2,558	329	341
cop	1,285	172	173
csubj	296	47	37
det	1,217	137	155
discourse	158	24	15
dislocated	47	5	10
flat	1,291	142	108
iobj	63	8	6
mark	5,652	749	680
nmod	14,198	1,975	1,653
nsubj	7,736	966	999
nummod	5,006	614	623
obj	6,175	819	754
obl	2,848	347	316
orphan	0	2	0
parataxis	1,829	235	258
punct	13,627	1,770	1,691
reparandum	1	0	0
root	3,997	500	500
vocative	1	0	0
xcomp	1,587	163	174

Table 13: Statistics of zh_gsdsimp

Figure 13: Syntactic acquisition through pretraining of OLMo-2-7B on Chinese (zh_gsdsimp).

C.6. Croatian

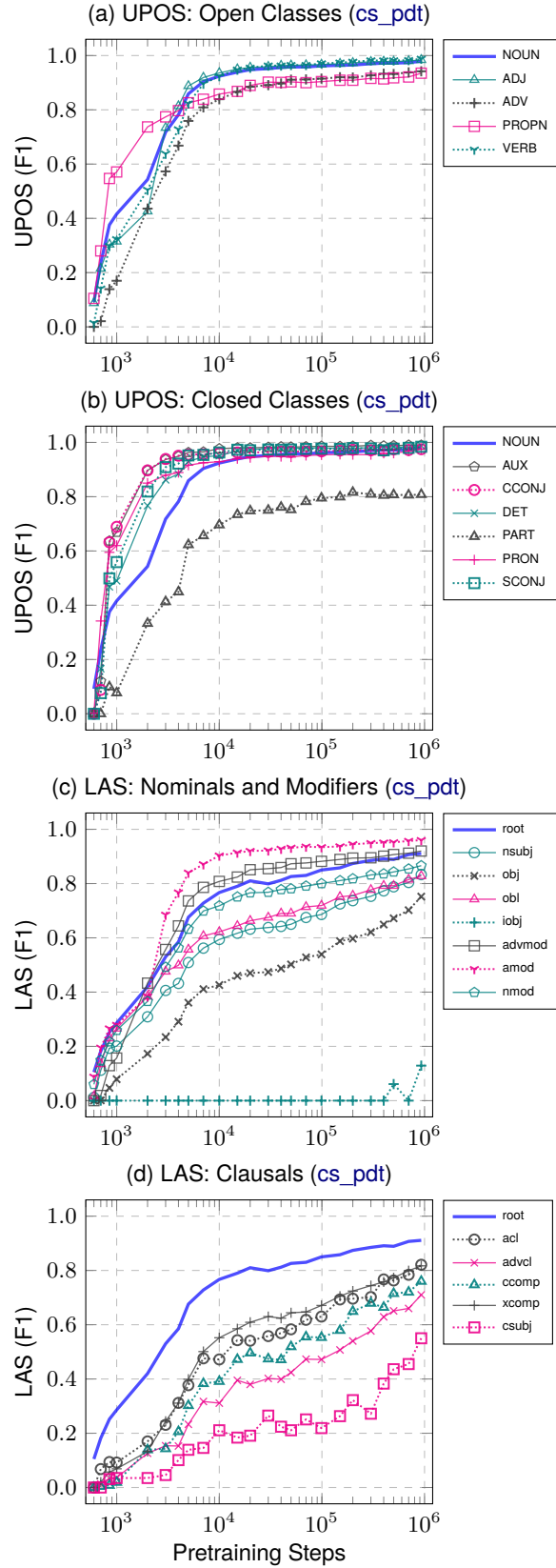


	Train	Dev	Test
Number of sentences	6,914	960	1,136
Number of words	152,857	22,292	24,260
Number of label occurrences			
ADJ	18,515	2,754	2,870
ADP	14,651	2,123	2,315
ADV	6,552	930	946
AUX	9,613	1,437	1,516
CCONJ	6,285	907	950
DET	5,980	839	875
INTJ	12	0	0
NOUN	36,963	5,445	6,169
NUM	2,335	460	353
PART	1,651	226	215
PRON	4,128	564	610
PROPN	9,817	1,413	1,618
PUNCT	18,444	2,684	3,037
SCONJ	3,811	529	589
SYM	94	23	2
VERB	13,382	1,886	2,119
X	624	72	76
<i>acl</i>	2,956	443	452
<i>advcl</i>	1,395	192	198
<i>advmod</i>	5,980	843	830
<i>amod</i>	14,959	2,225	2,355
<i>appos</i>	760	109	130
<i>aux</i>	6,374	1,035	1,037
<i>case</i>	14,846	2,159	2,364
<i>cc</i>	5,899	837	887
<i>ccomp</i>	1,374	178	230
<i>compound</i>	8	1	1
<i>conj</i>	7,333	1,017	1,134
<i>cop</i>	2,750	353	415
<i>csubj</i>	271	49	40
<i>dep</i>	7	1	0
<i>det</i>	2,893	409	429
<i>discourse</i>	1,606	232	208
<i>dislocated</i>	7	0	0
<i>expl</i>	1,795	263	302
<i>fixed</i>	704	91	100
<i>flat</i>	4,182	684	693
<i>iobj</i>	538	67	77
<i>list</i>	23	1	0
<i>mark</i>	3,160	421	471
<i>nmod</i>	14,328	2,032	2,437
<i>nsubj</i>	10,731	1,619	1,725
<i>nummod</i>	2,010	380	304
<i>obj</i>	6,745	983	1,072
<i>obl</i>	9,687	1,427	1,530
<i>orphan</i>	90	39	13
<i>parataxis</i>	1,735	261	300
<i>punct</i>	18,443	2,684	3,037
<i>root</i>	6,914	960	1,136
<i>vocative</i>	20	0	3
<i>xcomp</i>	2,334	297	350

Table 14: Statistics of *hr_set*

Figure 14: Syntactic acquisition through pretraining of OLMo-2-7B on Croatian (*hr_set*).

C.7. Czech

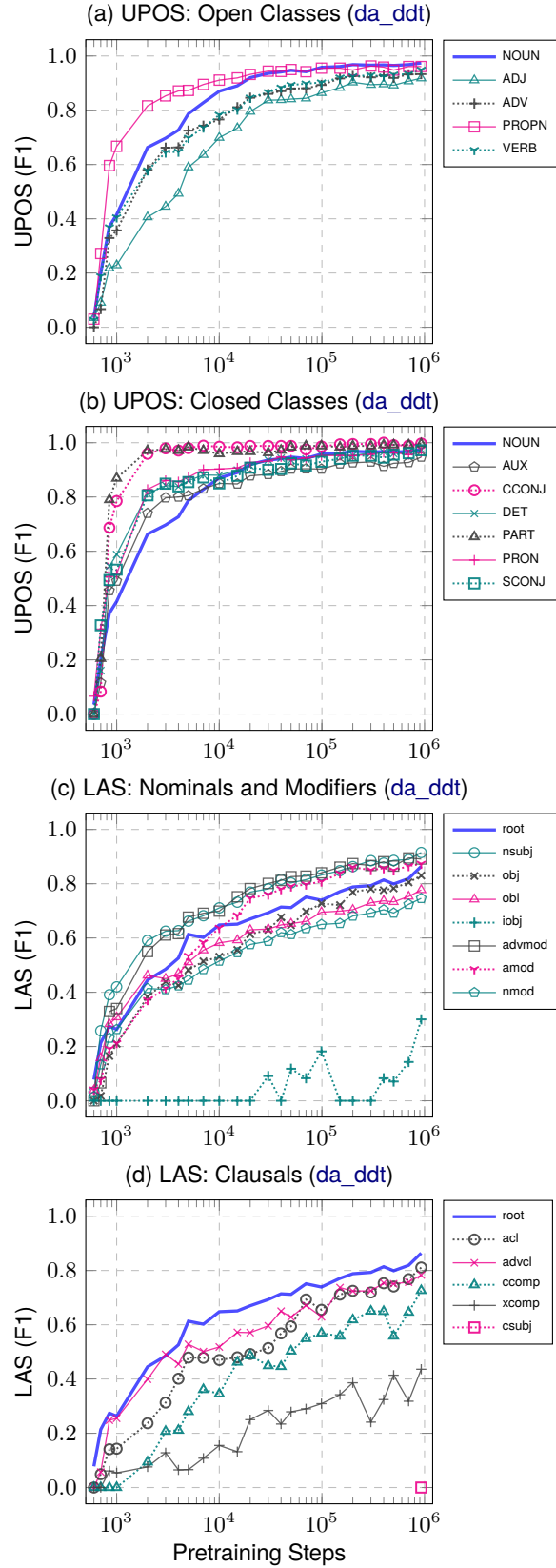


	Train	Dev	Test
Number of sentences	68,491	9,270	10,146
Number of words	1,173,285	159,283	173,918
Number of label occurrences			
ADJ	144,444	19,742	21,025
ADP	113,566	15,420	16,523
ADV	58,687	7,960	8,850
AUX	35,834	5,105	5,648
CCONJ	45,058	6,309	6,721
DET	43,380	6,156	6,642
INTJ	69	5	6
NOUN	292,282	39,508	43,665
NUM	34,897	4,307	4,951
PART	8,696	1,318	1,340
PRON	34,651	4,855	5,241
PROPN	58,837	7,584	8,157
PUNCT	172,565	22,909	25,409
SCONJ	21,800	3,021	3,432
SYM	989	157	222
VERB	100,925	13,999	15,220
X	6,605	928	866
acl	16,875	2,344	2,567
advcl	8,179	1,130	1,346
advmod	63,379	8,765	9,699
amod	122,414	16,662	17,679
appos	6,119	824	827
aux	15,987	2,325	2,477
case	113,328	15,379	16,479
cc	41,616	5,843	6,214
ccomp	8,526	1,147	1,282
compound	3,261	409	344
conj	53,491	7,228	7,722
cop	19,074	2,661	3,055
csubj	4,870	658	680
dep	9,887	1,181	1,572
det	23,797	3,352	3,691
discourse	290	47	45
expl	17,222	2,359	2,505
fixed	3,820	513	564
flat	17,829	2,408	2,486
iobj	455	72	66
mark	21,765	3,036	3,432
nmod	115,178	15,492	17,151
nsubj	73,237	9,901	10,973
nummod	22,349	2,745	3,125
obj	41,524	5,824	6,277
obl	89,975	12,250	13,414
orphan	2,054	306	277
parataxis	1,499	154	233
punct	172,551	22,906	25,402
root	68,491	9,270	10,146
vocative	56	13	12
xcomp	14,187	2,079	2,176

Table 15: Statistics of *cs_pdt*

Figure 15: Syntactic acquisition through pretraining of OLMo-2-7B on Czech (*cs_pdt*).

C.8. Danish

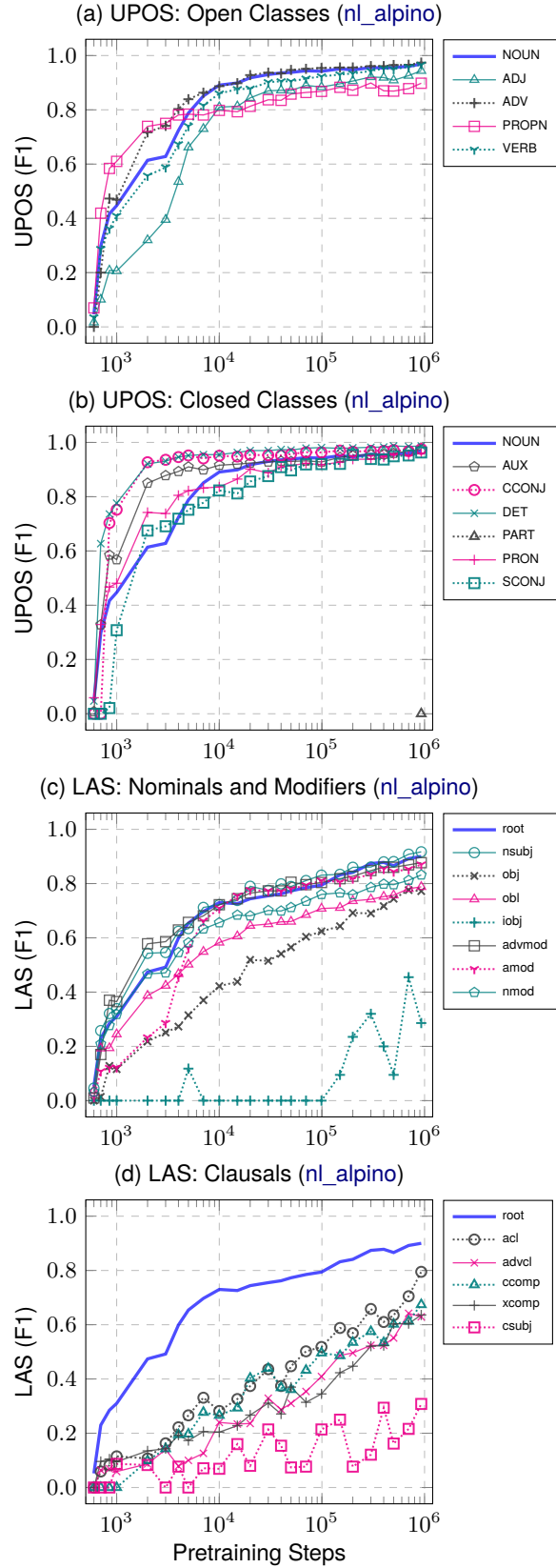


	Train	Dev	Test
Number of sentences	4,383	564	565
Number of words	80,378	10,332	10,023
Number of label occurrences			
ADJ	5,246	675	638
ADP	9,413	1,188	1,145
ADV	6,474	818	778
AUX	3,742	500	488
CCONJ	2,693	367	323
DET	4,396	609	499
INTJ	75	11	8
NOUN	14,956	1,947	1,823
NUM	1,212	140	153
PART	956	137	122
PRON	5,759	720	723
PROPN	3,940	488	550
PUNCT	11,111	1,381	1,444
SCONJ	1,427	180	183
SYM	48	0	6
VERB	8,632	1,148	1,118
X	298	23	22
acl	1,446	187	183
advcl	1,301	173	156
advmod	5,856	728	715
amod	4,618	583	545
appos	288	34	38
aux	2,487	342	325
case	7,947	1,013	953
cc	2,514	346	302
ccomp	461	64	79
compound	321	40	35
conj	2,809	378	346
cop	1,255	158	163
dep	405	50	30
det	4,373	607	496
discourse	25	4	3
dislocated	1	0	1
expl	368	34	39
fixed	390	40	62
flat	1,305	151	188
iobj	120	22	15
list	164	18	17
mark	3,518	461	449
nmod	5,026	616	567
nsubj	7,180	949	945
nummod	1,016	119	113
obj	4,037	521	525
obl	5,063	672	650
orphan	11	0	3
parataxis	2	3	3
punct	11,110	1,379	1,444
reparandum	2	0	0
root	4,383	564	565
vocative	23	5	4
xcomp	553	71	64

Table 16: Statistics of da_ddt

Figure 16: Syntactic acquisition through pretraining of OLMo-2-7B on Danish (da_ddt).

C.9. Dutch

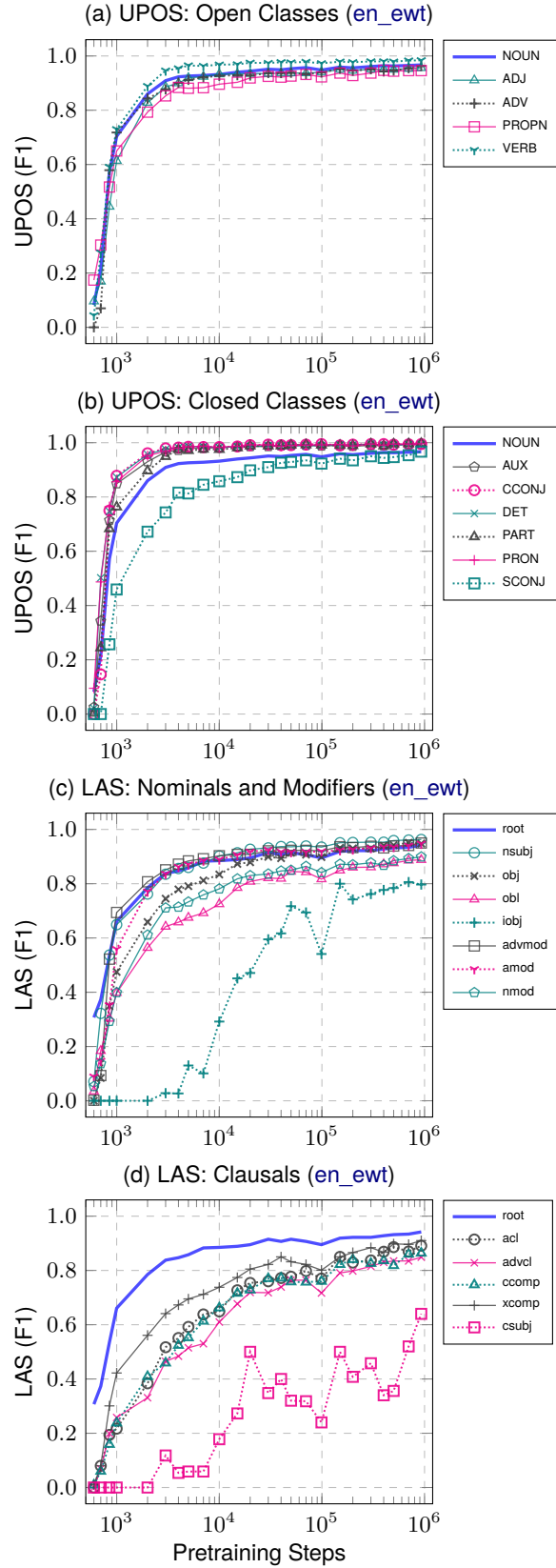


	Train	Dev	Test
Number of sentences	12,289	718	596
Number of words	186,027	11,541	11,046
Number of label occurrences			
ADJ	12,111	761	761
ADP	23,314	1,523	1,463
ADV	11,483	629	601
AUX	8,881	408	527
CCONJ	3,907	253	337
DET	22,561	1,314	1,348
INTJ	94	9	0
NOUN	30,855	1,809	2,090
NUM	3,107	286	156
PRON	12,052	639	451
PROPN	13,609	1,119	660
PUNCT	20,040	1,382	1,191
SCONJ	3,056	163	191
SYM	453	35	57
VERB	19,676	1,152	1,144
X	828	59	69
<i>acl</i>	2,934	114	128
<i>advcl</i>	2,080	109	116
<i>advmod</i>	10,607	607	570
<i>amod</i>	9,025	561	593
<i>appos</i>	2,071	155	115
<i>aux</i>	6,233	278	393
<i>case</i>	19,145	1,267	1,232
<i>cc</i>	3,538	226	326
<i>ccomp</i>	1,224	73	89
<i>compound</i>	1,734	136	103
<i>conj</i>	4,312	271	425
<i>cop</i>	2,640	130	133
<i>csubj</i>	400	14	21
<i>det</i>	22,055	1,289	1,317
<i>expl</i>	810	40	33
<i>fixed</i>	3,033	212	266
<i>flat</i>	4,807	321	230
<i>iobj</i>	509	23	13
<i>mark</i>	5,574	303	318
<i>nmod</i>	11,151	691	768
<i>nsubj</i>	16,148	966	823
<i>nummod</i>	1,657	151	80
<i>obj</i>	6,239	375	274
<i>obl</i>	11,641	856	651
<i>orphan</i>	68	1	3
<i>parataxis</i>	1,345	123	115
<i>punct</i>	20,040	1,382	1,191
<i>root</i>	12,289	718	596
<i>xcomp</i>	2,718	149	124

Table 17: Statistics of *nl_alpino*

Figure 17: Syntactic acquisition through pretraining of OLMo-2-7B on Dutch (*nl_alpino*).

C.10. English

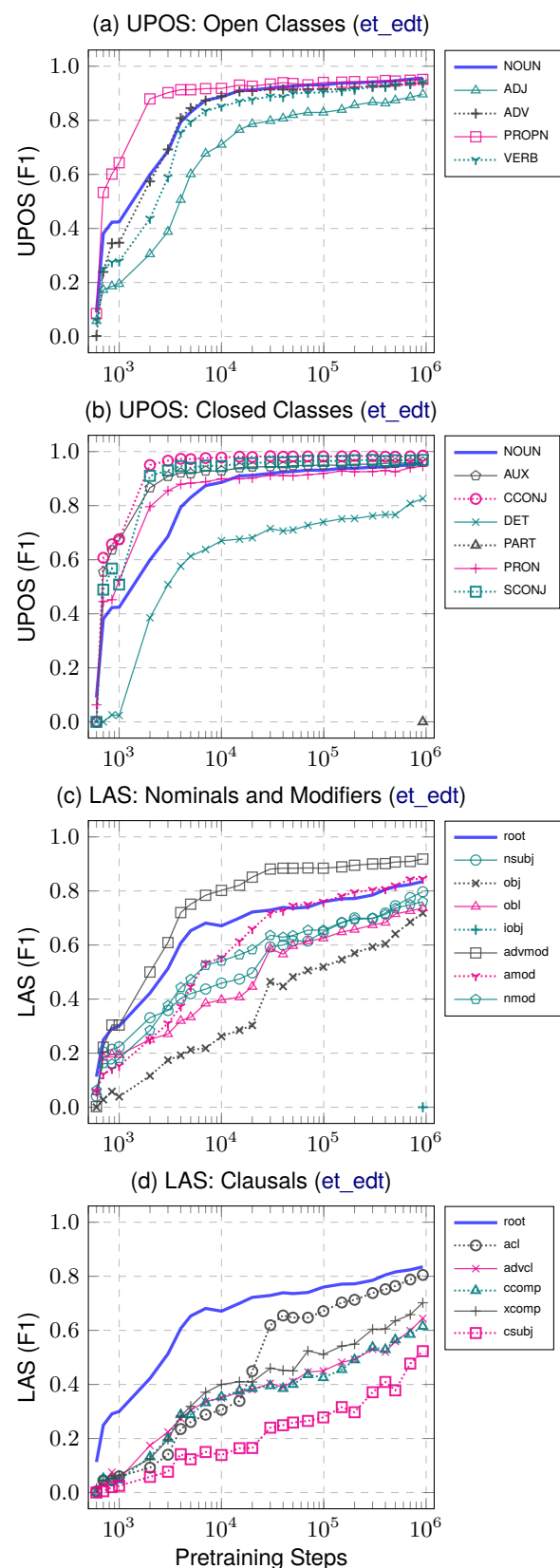


	Train	Dev	Test
Number of sentences	12,544	2,001	2,077
Number of words	204,579	25,149	25,094
Number of label occurrences			
<i>ADJ</i>	13,186	1,873	1,794
<i>ADP</i>	17,739	2,038	2,030
<i>ADV</i>	10,117	1,224	1,183
<i>AUX</i>	12,815	1,567	1,543
<i>CCONJ</i>	6,687	779	736
<i>DET</i>	16,299	1,900	1,896
<i>INTJ</i>	695	115	121
<i>NOUN</i>	34,755	4,212	4,123
<i>NUM</i>	4,127	383	542
<i>PART</i>	5,748	647	649
<i>PRON</i>	18,677	2,225	2,165
<i>PROPN</i>	12,618	1,865	2,076
<i>PUNCT</i>	23,596	3,075	3,096
<i>SCONJ</i>	3,822	397	384
<i>SYM</i>	722	83	109
<i>VERB</i>	22,577	2,707	2,605
<i>X</i>	399	59	42
<i>acl</i>	3,403	380	376
<i>advcl</i>	3,942	372	367
<i>advmod</i>	11,262	1,344	1,315
<i>amod</i>	9,638	1,334	1,253
<i>appos</i>	1,544	172	179
<i>aux</i>	7,859	911	939
<i>case</i>	17,423	2,014	1,969
<i>cc</i>	6,837	798	755
<i>ccomp</i>	2,107	200	223
<i>compound</i>	8,120	990	1,103
<i>conj</i>	7,572	916	861
<i>cop</i>	4,717	623	576
<i>csubj</i>	293	38	25
<i>dep</i>	3	2	0
<i>det</i>	15,953	1,851	1,860
<i>discourse</i>	813	127	126
<i>dislocated</i>	5	1	0
<i>expl</i>	601	87	68
<i>fixed</i>	544	58	64
<i>flat</i>	1,785	266	335
<i>goeswith</i>	130	26	15
<i>iobj</i>	649	74	71
<i>list</i>	444	75	251
<i>mark</i>	7,126	757	752
<i>nmod</i>	11,028	1,250	1,202
<i>nsubj</i>	17,377	2,136	2,074
<i>nummod</i>	2,430	264	254
<i>obj</i>	9,902	1,213	1,153
<i>obl</i>	10,117	1,145	1,165
<i>orphan</i>	26	2	1
<i>parataxis</i>	1,571	249	232
<i>punct</i>	23,556	3,061	3,065
<i>reparandum</i>	35	9	4
<i>root</i>	12,544	2,001	2,077
<i>vocative</i>	139	22	21
<i>xcomp</i>	3,084	381	363

Table 18: Statistics of *en_ewt*

Figure 18: Syntactic acquisition through pretraining of OLMo-2-7B on English (*en_ewt*).

C.11. Estonian

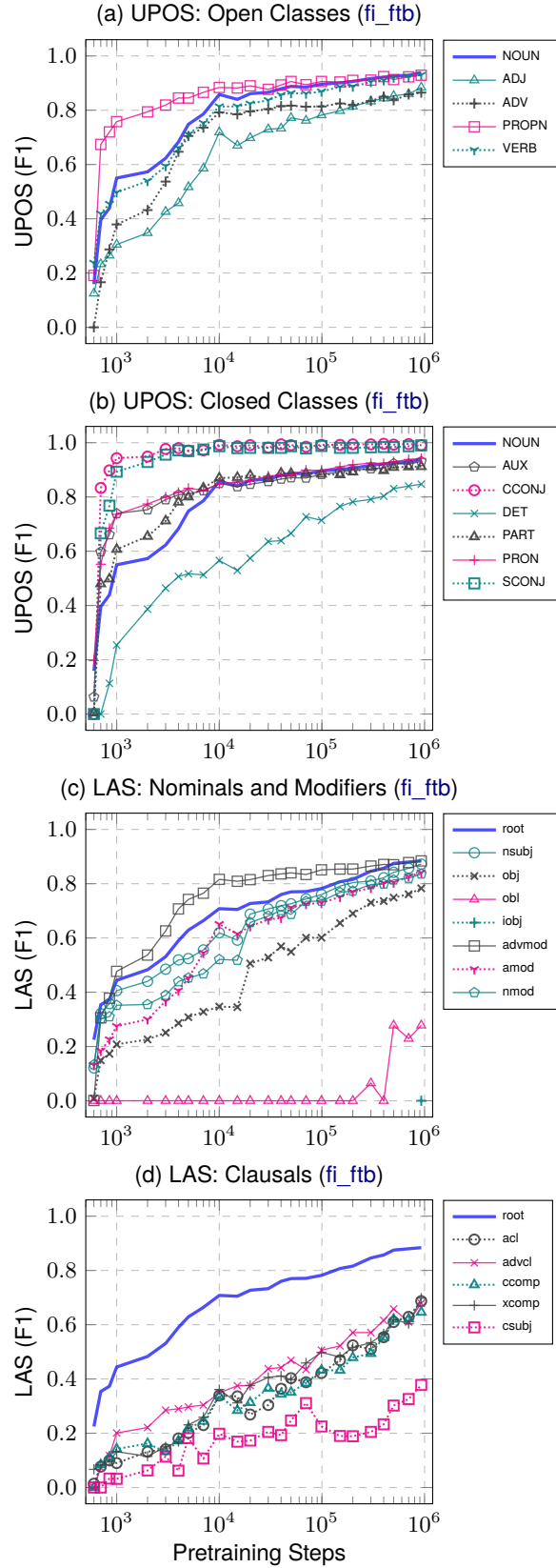


	Train	Dev	Test
Number of sentences	24,601	3,122	3,207
Number of words	344,581	44,742	48,465
Number of label occurrences			
ADJ	28,769	3,999	4,102
ADP	7,485	896	923
ADV	33,058	4,193	4,887
AUX	17,493	2,250	2,583
CCONJ	12,403	1,680	2,008
DET	5,646	688	850
INTJ	256	21	52
NOUN	91,404	11,665	12,616
NUM	7,099	1,077	840
PART	2	0	0
PRON	17,853	2,453	2,500
PROPN	20,712	2,521	3,047
PUNCT	56,395	7,498	7,672
SCONJ	6,737	860	1,112
SYM	599	115	26
VERB	38,015	4,779	5,184
X	655	47	63
acl	9,657	1,227	1,452
advcl	5,381	716	784
advmod	26,103	3,295	3,949
amod	17,373	2,529	2,373
appos	2,871	389	415
aux	9,187	1,154	1,376
case	7,463	894	920
cc	12,765	1,734	2,053
ccomp	3,473	384	478
compound	3,915	485	535
conj	17,311	2,432	2,436
cop	8,301	1,096	1,206
csubj	1,546	193	243
dep	9	0	9
det	5,636	686	844
discourse	239	18	49
fixed	355	56	35
flat	3,815	547	618
goeswith	238	12	21
list	8	2	1
mark	7,671	969	1,268
nmod	30,616	3,903	4,315
nsubj	28,226	3,585	3,764
nummod	4,696	690	555
obj	16,860	2,046	2,439
obl	31,177	3,907	4,138
orphan	479	72	51
parataxis	3,496	513	609
punct	56,395	7,498	7,672
root	24,601	3,122	3,207
vocative	120	18	13
xcomp	4,598	570	637

Table 19: Statistics of et_edt

Figure 19: Syntactic acquisition through pretraining of OLMo-2-7B on Estonian (et_edt).

C.12. Finnish

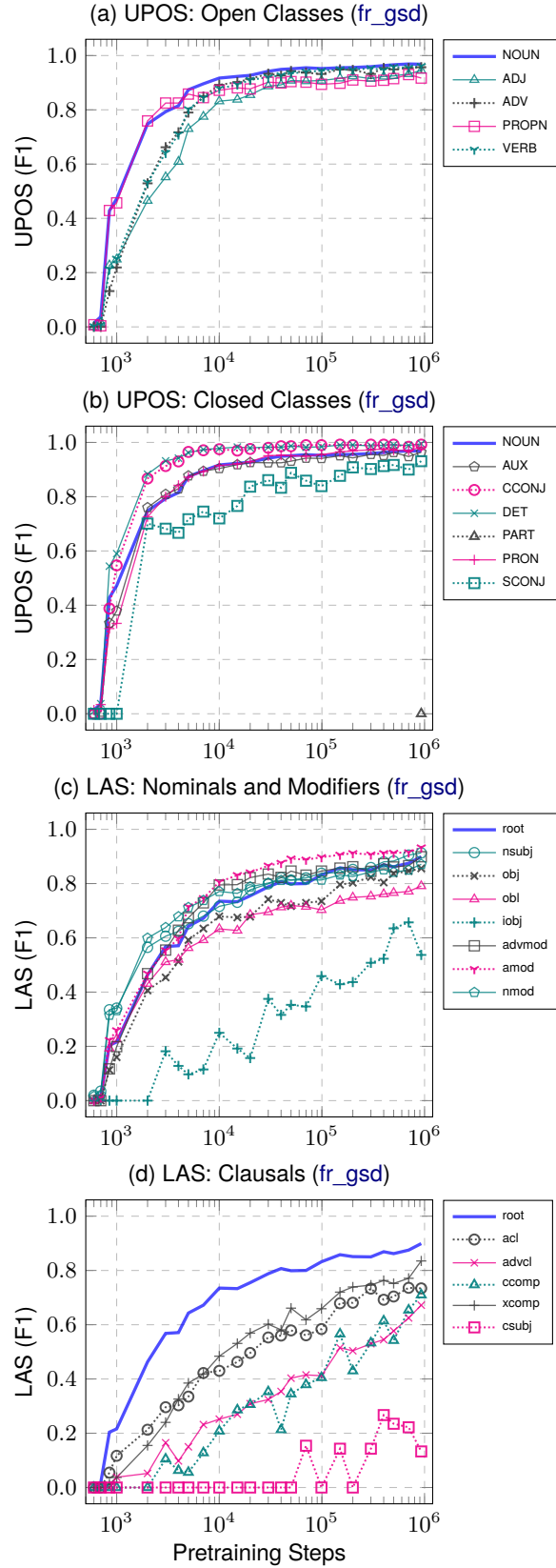


	Train	Dev	Test
Number of sentences	14,981	1,875	1,867
Number of words	127,602	15,724	16,286
Number of label occurrences			
ADJ	8,327	1,056	1,106
ADP	2,203	254	274
ADV	8,089	1,020	1,023
AUX	8,318	1,071	1,017
CCONJ	3,815	453	511
DET	2,940	353	398
INTJ	270	39	32
NOUN	29,650	3,595	3,788
NUM	1,826	223	203
PART	4,536	596	550
PRON	8,818	1,153	1,095
PROPN	5,482	610	680
PUNCT	18,062	2,190	2,313
SCONJ	3,270	431	445
SYM	17	4	1
VERB	21,728	2,639	2,827
X	251	37	23
<i>acl</i>	2,364	291	324
<i>advcl</i>	2,687	340	365
<i>advmod</i>	11,539	1,464	1,412
<i>amod</i>	5,500	695	698
<i>appos</i>	2	0	0
<i>aux</i>	5,456	711	674
<i>case</i>	2,162	253	272
<i>cc</i>	3,774	450	506
<i>ccomp</i>	2,007	244	261
<i>compound</i>	613	59	70
<i>conj</i>	5,230	600	668
<i>cop</i>	2,859	360	343
<i>csubj</i>	402	53	58
<i>dep</i>	229	37	25
<i>det</i>	2,765	333	381
<i>discourse</i>	240	33	35
<i>expl</i>	405	60	59
<i>fixed</i>	425	63	67
<i>flat</i>	682	61	75
<i>goeswith</i>	7	0	0
<i>mark</i>	3,514	469	490
<i>nmod</i>	18,989	2,340	2,396
<i>nsubj</i>	12,595	1,537	1,570
<i>nummod</i>	1,459	162	156
<i>obj</i>	6,861	833	967
<i>obl</i>	143	15	25
<i>orphan</i>	3	0	0
<i>punct</i>	18,062	2,190	2,313
<i>reparandum</i>	17	1	1
<i>root</i>	14,981	1,875	1,867
<i>vocative</i>	119	24	16
<i>xcomp</i>	1,511	171	192

Table 20: Statistics of *fi_ftb*

Figure 20: Syntactic acquisition through pretraining of OLMo-2-7B on Finnish (*fi_ftb*).

C.13. French

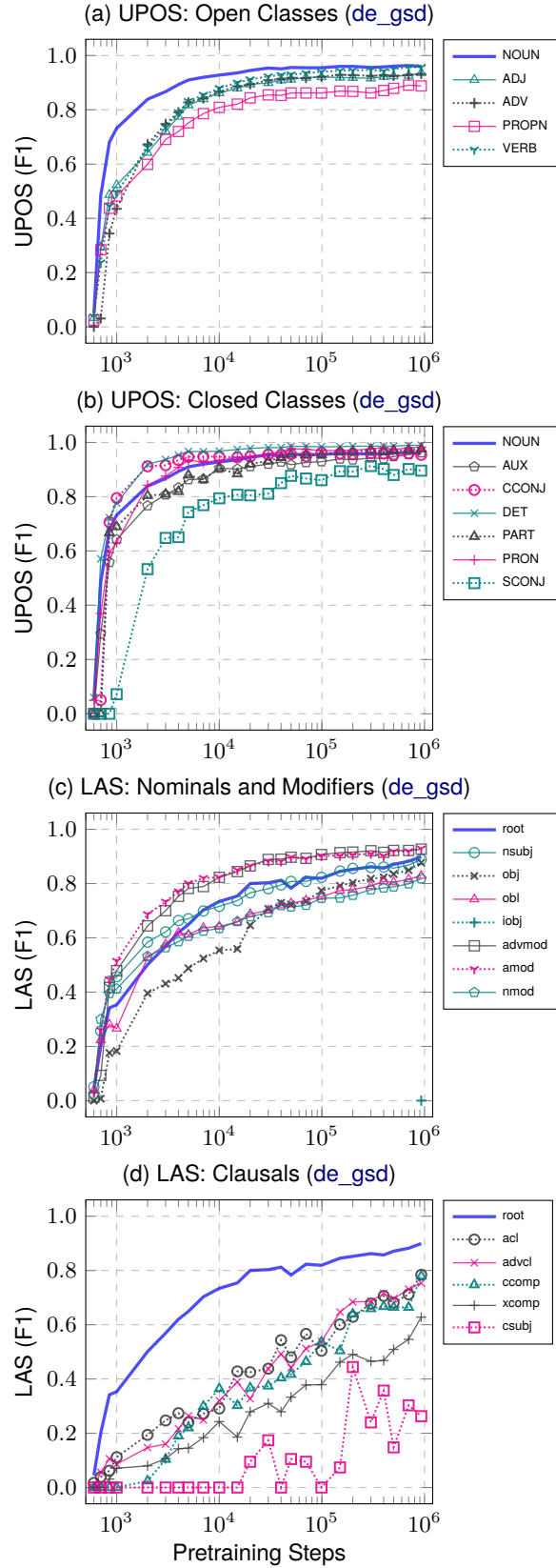


	Train	Dev	Test
Number of sentences	14,450	1,476	416
Number of words	354,652	35,721	10,018
Number of label occurrences			
ADJ	20,999	2,212	609
ADP	56,363	5,729	1,479
ADV	12,430	1,225	487
AUX	11,586	1,136	359
CCONJ	9,303	932	249
DET	54,109	5,506	1,479
INTJ	70	6	9
NOUN	66,624	6,786	1,870
NUM	9,325	968	229
PRON	16,050	1,576	559
PROPN	24,680	2,551	487
PUNCT	39,017	3,802	1,186
SCONJ	2,600	253	128
SYM	616	62	39
VERB	28,189	2,763	821
X	2,691	214	28
<i>acl</i>	7,302	715	178
<i>advcl</i>	3,094	315	117
<i>advmod</i>	12,133	1,188	462
<i>amod</i>	17,692	1,890	485
<i>appos</i>	5,536	574	111
<i>aux</i>	6,651	649	206
<i>case</i>	51,383	5,250	1,309
<i>cc</i>	9,515	955	255
<i>ccomp</i>	1,276	102	55
<i>compound</i>	6	0	0
<i>conj</i>	12,714	1,201	319
<i>cop</i>	4,914	485	153
<i>csubj</i>	195	18	12
<i>dep</i>	44	4	0
<i>det</i>	53,723	5,462	1,459
<i>discourse</i>	39	3	7
<i>dislocated</i>	62	6	8
<i>expl</i>	2,487	260	98
<i>fixed</i>	3,041	323	105
<i>flat</i>	7,201	749	127
<i>goeswith</i>	29	5	3
<i>iobj</i>	813	73	34
<i>mark</i>	5,908	569	244
<i>nmod</i>	31,534	3,210	848
<i>nsubj</i>	20,208	2,000	589
<i>nummod</i>	3,190	339	110
<i>obj</i>	12,047	1,183	368
<i>obl</i>	23,303	2,423	578
<i>orphan</i>	179	17	11
<i>parataxis</i>	981	86	50
<i>punct</i>	39,017	3,802	1,186
<i>root</i>	14,450	1,476	416
<i>vocative</i>	14	1	0
<i>xcomp</i>	3,971	388	115

Table 21: Statistics of *fr_gsd*

Figure 21: Syntactic acquisition through pretraining of OLMo-2-7B on French (*fr_gsd*).

C.14. German

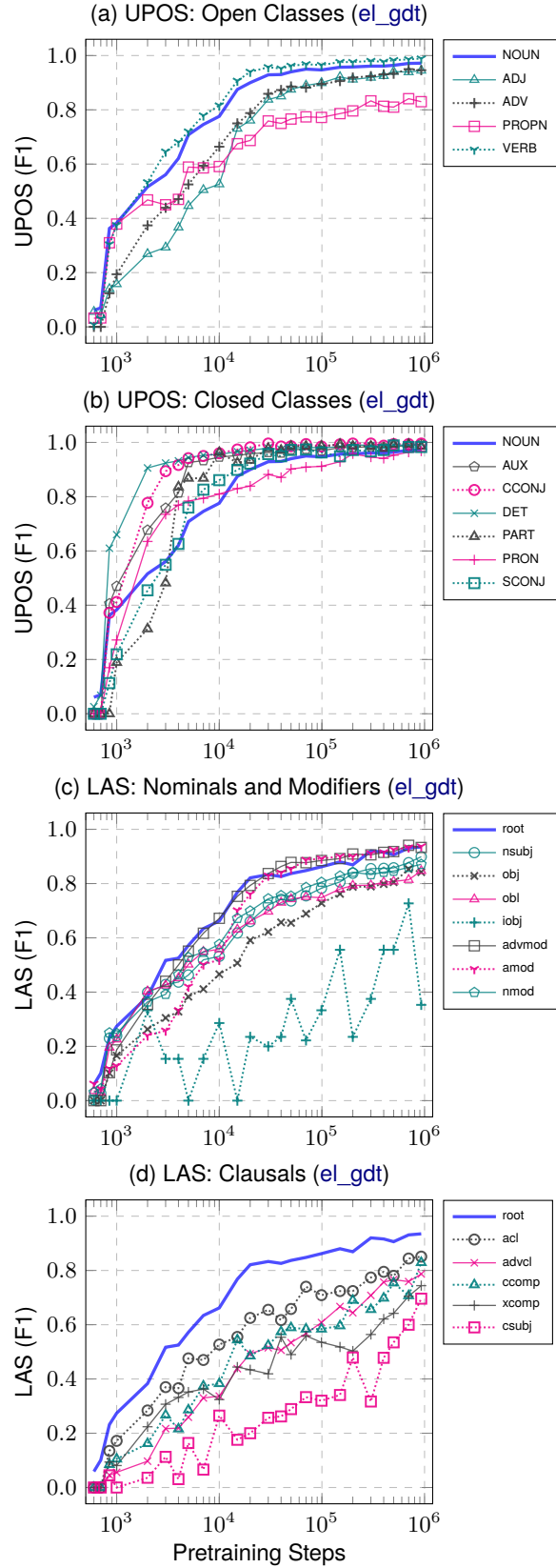


	Train	Dev	Test
Number of sentences	13,814	799	977
Number of words	263,791	12,480	16,498
Number of label occurrences			
ADJ	19,489	1,001	1,249
ADP	29,374	1,125	1,606
ADV	11,900	922	1,062
AUX	9,333	641	691
CCONJ	7,962	417	462
DET	37,305	1,639	2,262
INTJ	3	1	4
NOUN	46,986	2,200	3,111
NUM	6,929	175	233
PART	1,789	150	210
PRON	9,433	707	705
PROPN	28,785	612	1,022
PUNCT	34,453	1,672	2,365
SCONJ	1,451	115	161
SYM	93	3	4
VERB	18,237	1,081	1,326
X	269	19	25
<i>acl</i>	2,475	87	115
<i>advcl</i>	1,273	80	117
<i>advmod</i>	15,082	1,140	1,350
<i>amod</i>	14,790	534	793
<i>appos</i>	6,567	123	240
<i>aux</i>	5,769	428	509
<i>case</i>	28,090	1,070	1,524
<i>cc</i>	7,930	403	449
<i>ccomp</i>	612	110	150
<i>compound</i>	2,537	204	306
<i>conj</i>	10,932	471	524
<i>cop</i>	3,517	210	176
<i>csubj</i>	214	10	16
<i>dep</i>	605	80	206
<i>det</i>	36,095	1,541	2,044
<i>discourse</i>	7	1	3
<i>expl</i>	579	17	16
<i>fixed</i>	108	12	8
<i>flat</i>	7,808	128	223
<i>goeswith</i>	2	0	0
<i>mark</i>	2,671	190	258
<i>nmod</i>	20,237	568	996
<i>nsubj</i>	18,234	1,064	1,251
<i>nummod</i>	2,623	145	147
<i>obj</i>	7,248	502	587
<i>obl</i>	17,339	756	971
<i>orphan</i>	19	0	0
<i>parataxis</i>	483	30	56
<i>punct</i>	34,453	1,672	2,364
<i>reparandum</i>	3	0	0
<i>root</i>	13,814	799	977
<i>vocative</i>	2	2	3
<i>xcomp</i>	1,673	103	119

Table 22: Statistics of *de_gsd*

Figure 22: Syntactic acquisition through pretraining of OLMo-2-7B on German (*de_gsd*).

C.15. Greek

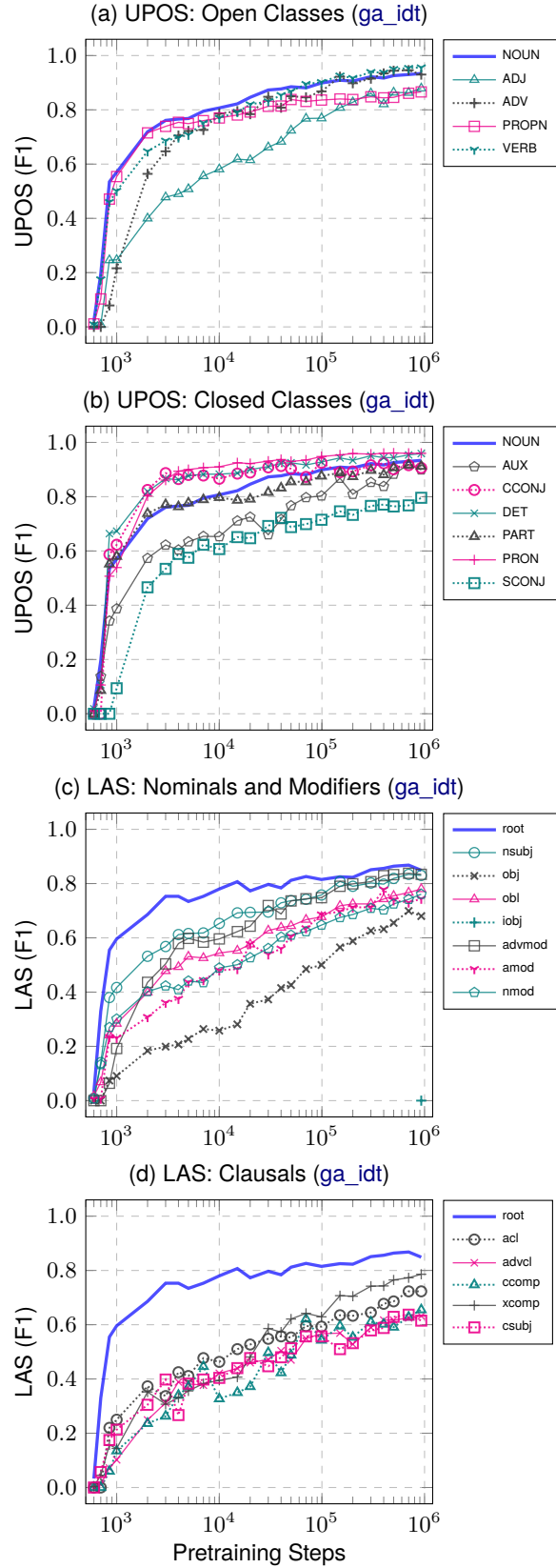


	Train	Dev	Test
Number of sentences	1,662	403	456
Number of words	42,326	10,443	10,672
Number of label occurrences			
ADJ	3,308	918	883
ADP	3,531	918	864
ADV	1,913	376	448
AUX	2,141	373	505
CCONJ	1,363	324	363
DET	7,957	2,146	2,016
NOUN	8,934	2,306	2,304
NUM	484	152	156
PART	312	41	70
PRON	1,584	342	367
PROPN	1,145	370	325
PUNCT	4,187	1,039	1,093
SCONJ	738	126	178
SYM	1	0	0
VERB	4,042	851	973
X	686	161	127
<i>acl</i>	865	228	210
<i>advcl</i>	564	106	142
<i>advmod</i>	1,953	355	474
<i>amod</i>	2,827	824	748
<i>appos</i>	177	49	38
<i>aux</i>	1,759	272	389
<i>case</i>	3,575	936	863
<i>cc</i>	1,345	321	360
<i>ccomp</i>	491	70	98
<i>compound</i>	5	1	16
<i>conj</i>	1,397	364	371
<i>cop</i>	382	101	116
<i>csubj</i>	144	28	40
<i>dep</i>	1	0	0
<i>det</i>	7,953	2,143	2,016
<i>discourse</i>	2	0	0
<i>expl</i>	15	1	4
<i>fixed</i>	67	14	8
<i>flat</i>	461	99	104
<i>iobj</i>	59	2	9
<i>mark</i>	817	143	202
<i>nmod</i>	4,247	1,166	1,038
<i>nsubj</i>	2,358	593	628
<i>nummod</i>	293	83	90
<i>obj</i>	1,642	330	410
<i>obl</i>	2,516	659	615
<i>orphan</i>	27	7	8
<i>parataxis</i>	60	15	15
<i>punct</i>	4,187	1,039	1,093
<i>root</i>	1,662	403	456
<i>vocative</i>	57	7	10
<i>xcomp</i>	418	84	101

Table 23: Statistics of *el_gdt*

Figure 23: Syntactic acquisition through pretraining of OLMo-2-7B on Greek (*el_gdt*).

C.16. Irish

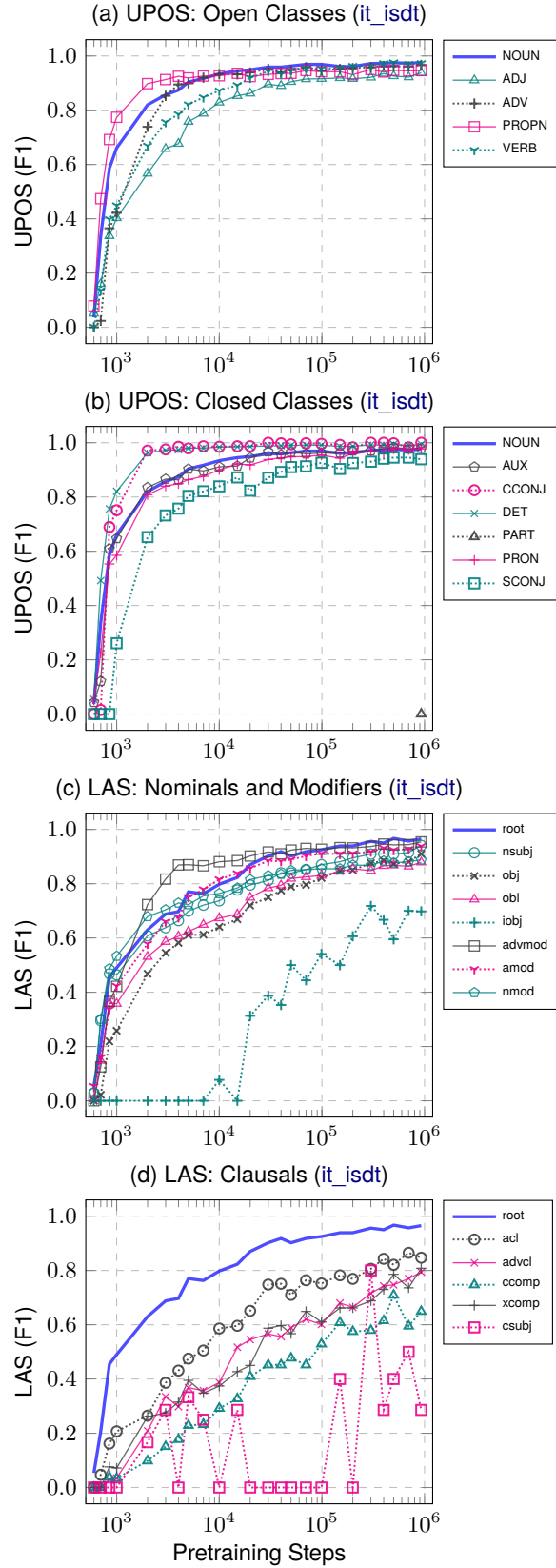


	Train	Dev	Test
Number of sentences	4,005	451	454
Number of words	95,881	10,000	10,109
Number of label occurrences			
ADJ	5,501	531	489
ADP	15,259	1,503	1,514
ADV	1,478	200	208
AUX	1,224	169	164
CCONJ	3,241	292	294
DET	8,549	871	867
INTJ	14	9	7
NOUN	28,141	2,640	2,592
NUM	1,625	140	176
PART	5,635	656	685
PRON	2,759	417	445
PROPN	4,724	487	515
PUNCT	8,993	1,082	1,108
SCONJ	1,304	162	163
SYM	22	1	1
VERB	7,157	793	825
X	255	47	56
<i>acl</i>	1,879	172	212
<i>advcl</i>	1,083	137	146
<i>advmod</i>	1,969	269	269
<i>amod</i>	3,595	323	291
<i>appos</i>	336	58	62
<i>case</i>	13,135	1,221	1,227
<i>cc</i>	3,129	268	246
<i>ccomp</i>	695	83	83
<i>compound</i>	162	16	28
<i>conj</i>	3,948	308	287
<i>cop</i>	1,213	162	163
<i>csubj</i>	764	101	95
<i>det</i>	7,900	798	756
<i>discourse</i>	22	8	7
<i>dislocated</i>	47	0	8
<i>fixed</i>	1,366	153	126
<i>flat</i>	908	202	157
<i>goeswith</i>	3	1	0
<i>list</i>	93	7	66
<i>mark</i>	5,439	627	657
<i>nmod</i>	11,605	1,068	1,151
<i>nsubj</i>	5,310	641	665
<i>nummod</i>	554	44	50
<i>obj</i>	3,825	367	368
<i>obl</i>	9,146	939	891
<i>orphan</i>	9	0	2
<i>parataxis</i>	542	119	95
<i>punct</i>	8,986	1,082	1,105
<i>root</i>	4,005	451	454
<i>vocative</i>	30	14	11
<i>xcomp</i>	4,183	361	431

Table 24: Statistics of *ga_idt*

Figure 24: Syntactic acquisition through pretraining of OLMo-2-7B on Irish (*ga_idt*).

C.17. Italian

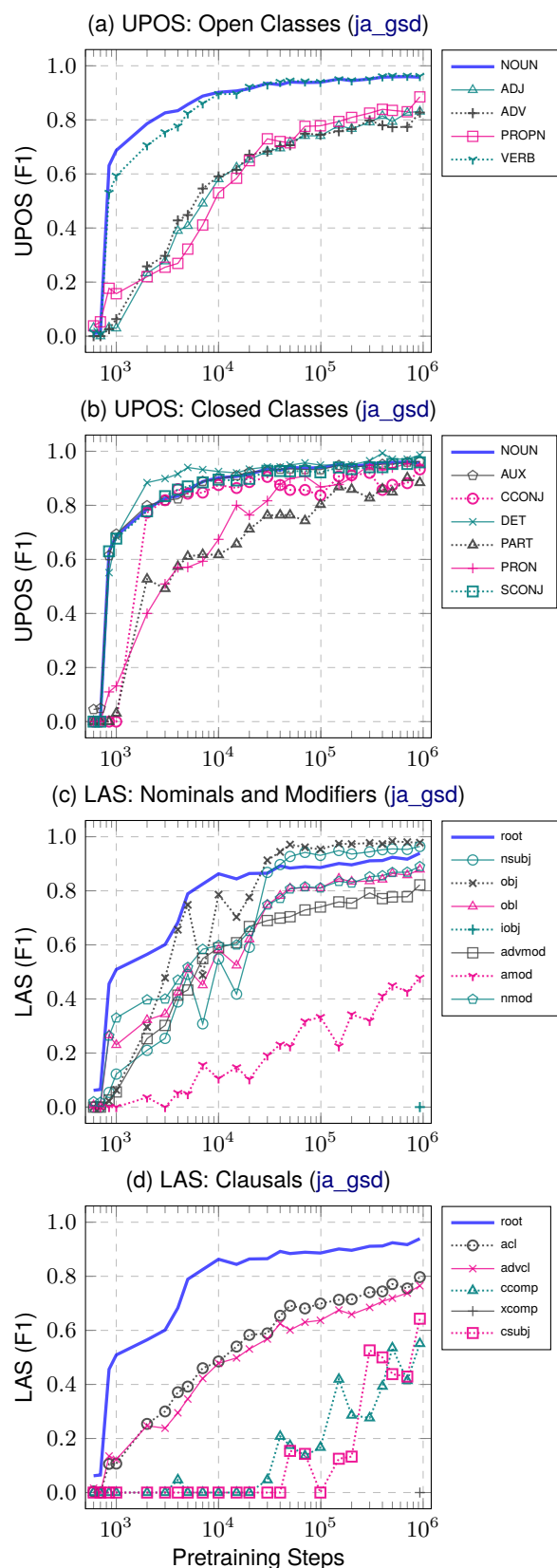


	Train	Dev	Test
Number of sentences	13,121	564	482
Number of words	276,014	11,907	10,417
Number of label occurrences			
ADJ	18,321	778	680
ADP	41,816	1,776	1,638
ADV	10,622	482	406
AUX	10,877	422	405
CCONJ	7,555	323	263
DET	45,035	1,895	1,712
INTJ	62	3	0
NOUN	54,968	2,390	2,070
NUM	4,791	226	171
PART	24	2	0
PRON	10,445	456	411
PROPN	13,676	595	505
PUNCT	31,289	1,413	1,175
SCONJ	2,817	115	100
SYM	93	4	5
VERB	23,364	1,021	863
X	259	6	13
<i>acl</i>	5,653	253	213
<i>advcl</i>	3,444	141	130
<i>advmod</i>	9,662	439	367
<i>amod</i>	15,435	667	586
<i>appos</i>	837	39	40
<i>aux</i>	7,674	296	300
<i>case</i>	38,639	1,643	1,540
<i>cc</i>	7,574	324	263
<i>ccomp</i>	1,349	61	38
<i>compound</i>	709	26	23
<i>conj</i>	9,444	370	311
<i>cop</i>	3,201	126	105
<i>csubj</i>	300	13	3
<i>dep</i>	13	0	1
<i>det</i>	45,020	1,895	1,711
<i>discourse</i>	59	3	0
<i>dislocated</i>	31	0	1
<i>expl</i>	2,757	118	105
<i>fixed</i>	939	41	45
<i>flat</i>	3,887	175	143
<i>goeswith</i>	0	0	1
<i>iobj</i>	648	26	20
<i>mark</i>	5,859	242	190
<i>nmod</i>	22,133	977	842
<i>nsubj</i>	13,986	598	535
<i>nummod</i>	3,263	184	113
<i>obj</i>	9,500	415	336
<i>obl</i>	16,943	726	699
<i>orphan</i>	38	1	1
<i>parataxis</i>	413	31	17
<i>punct</i>	31,289	1,413	1,175
<i>root</i>	13,121	564	482
<i>vocative</i>	96	3	3
<i>xcomp</i>	2,098	97	78

Table 25: Statistics of *it_isdt*

Figure 25: Syntactic acquisition through pretraining of OLMo-2-7B on Italian (*it_isdt*).

C.18. Japanese

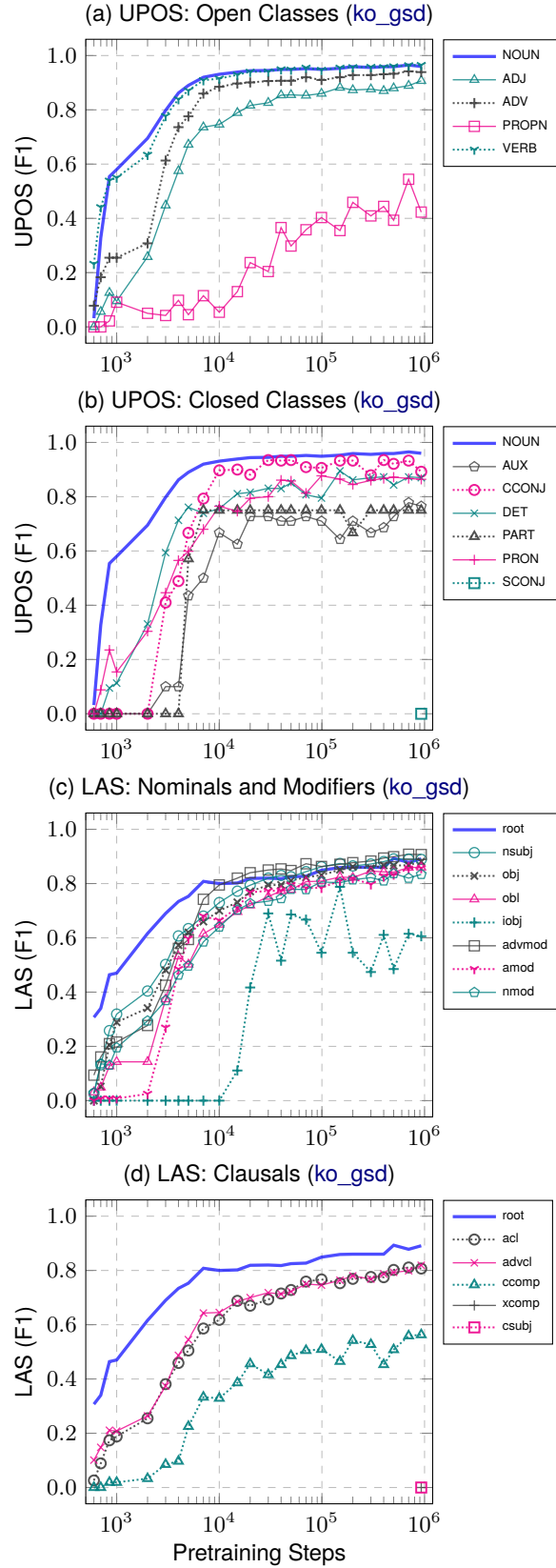


	Train	Dev	Test
Number of sentences	7,050	507	543
Number of words	168,333	12,287	13,034
Number of label occurrences			
ADJ	3,226	264	349
ADP	36,388	2,669	2,807
ADV	1,999	142	223
AUX	18,325	1,349	1,484
CCONJ	727	48	44
DET	863	53	71
INTJ	12	0	1
NOUN	50,730	3,765	3,689
NUM	4,607	305	251
PART	1,061	70	128
PRON	935	75	98
PROPN	6,351	477	313
PUNCT	16,776	1,166	1,291
SCONJ	6,809	507	679
SYM	1,153	78	70
VERB	18,371	1,319	1,536
<i>acl</i>	6,029	456	513
<i>advcl</i>	6,166	435	596
<i>advmod</i>	1,925	140	215
<i>amod</i>	370	37	38
<i>aux</i>	14,962	1,077	1,196
<i>case</i>	35,898	2,632	2,774
<i>cc</i>	727	48	44
<i>ccomp</i>	328	22	40
<i>compound</i>	24,192	1,783	1,514
<i>cop</i>	2,102	172	167
<i>csubj</i>	138	13	12
<i>dep</i>	60	6	10
<i>det</i>	863	53	71
<i>discourse</i>	15	0	1
<i>fixed</i>	7,459	561	603
<i>mark</i>	6,647	500	713
<i>nmod</i>	11,372	846	752
<i>nsubj</i>	7,110	533	593
<i>nummod</i>	2,487	169	144
<i>obj</i>	4,629	331	349
<i>obl</i>	11,028	800	855
<i>punct</i>	16,776	1,166	1,291
<i>root</i>	7,050	507	543

Table 26: Statistics of *ja_gsd*

Figure 26: Syntactic acquisition through pretraining of OLMo-2-7B on Japanese (*ja_gsd*).

C.19. Korean

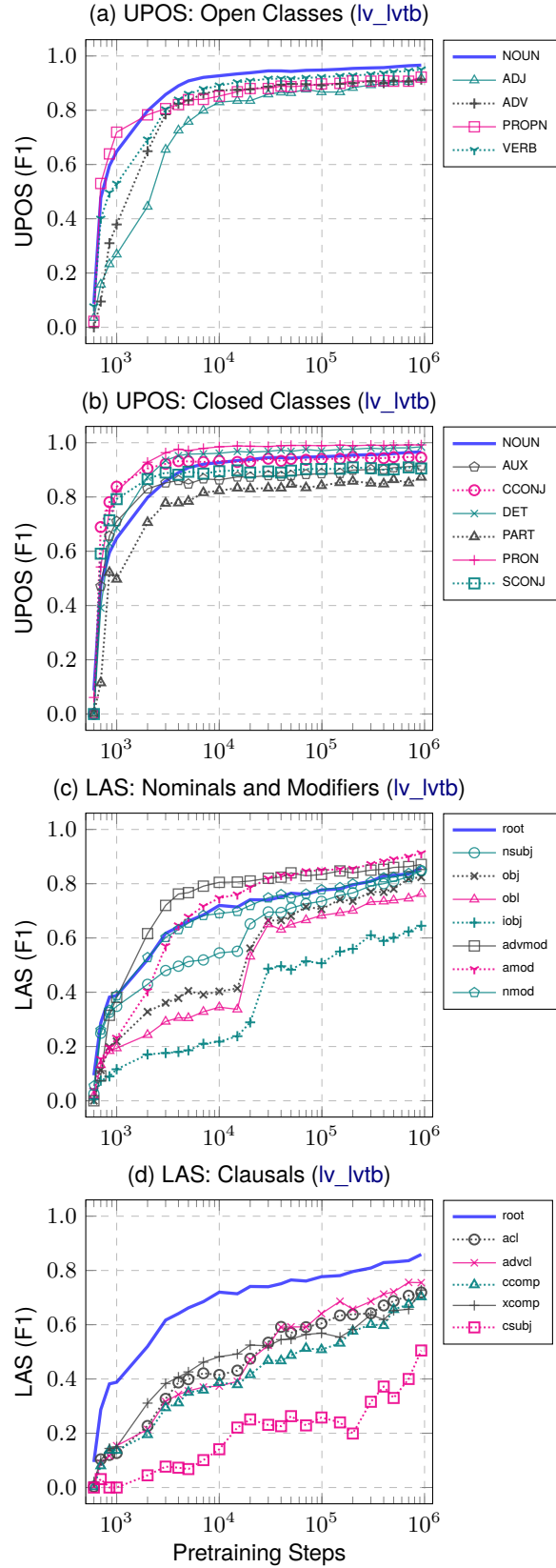


	Train	Dev	Test
Number of sentences	4,400	950	989
Number of words	56,687	11,958	11,677
Number of label occurrences			
ADJ	1,912	405	443
ADP	1,288	253	252
ADV	7,950	1,729	1,683
AUX	87	20	18
CCONJ	168	24	31
DET	393	89	91
INTJ	3	0	0
NOUN	22,915	4,810	4,622
NUM	610	123	115
PART	25	1	5
PRON	480	88	114
PROPN	359	72	64
PUNCT	7,335	1,583	1,493
SYM	197	38	40
VERB	12,955	2,720	2,706
X	10	3	0
<i>acl</i>	2,271	473	459
<i>advcl</i>	3,230	661	664
<i>advmod</i>	4,468	986	1,014
<i>amod</i>	1,114	221	231
<i>appos</i>	1,085	232	193
<i>aux</i>	33	10	4
<i>case</i>	1,162	243	226
<i>cc</i>	168	24	31
<i>ccomp</i>	450	98	103
<i>compound</i>	4	1	2
<i>conj</i>	2,715	598	553
<i>cop</i>	49	10	14
<i>csubj</i>	12	4	5
<i>dep</i>	1,750	365	318
<i>det</i>	513	114	122
<i>discourse</i>	1	0	0
<i>fixed</i>	10	2	1
<i>flat</i>	8,700	1,852	1,690
<i>iobj</i>	77	16	15
<i>list</i>	1	0	0
<i>mark</i>	65	11	10
<i>nmod</i>	2,936	628	526
<i>nsubj</i>	5,771	1,195	1,323
<i>nummod</i>	397	75	83
<i>obj</i>	4,122	810	868
<i>obl</i>	3,847	795	740
<i>parataxis</i>	1	0	0
<i>punct</i>	7,335	1,583	1,493
<i>root</i>	4,400	950	989
<i>xcomp</i>	0	1	0

Table 27: Statistics of *ko_gsd*

Figure 27: Syntactic acquisition through pretraining of OLMo-2-7B on Korean (*ko_gsd*).

C.20. Latvian

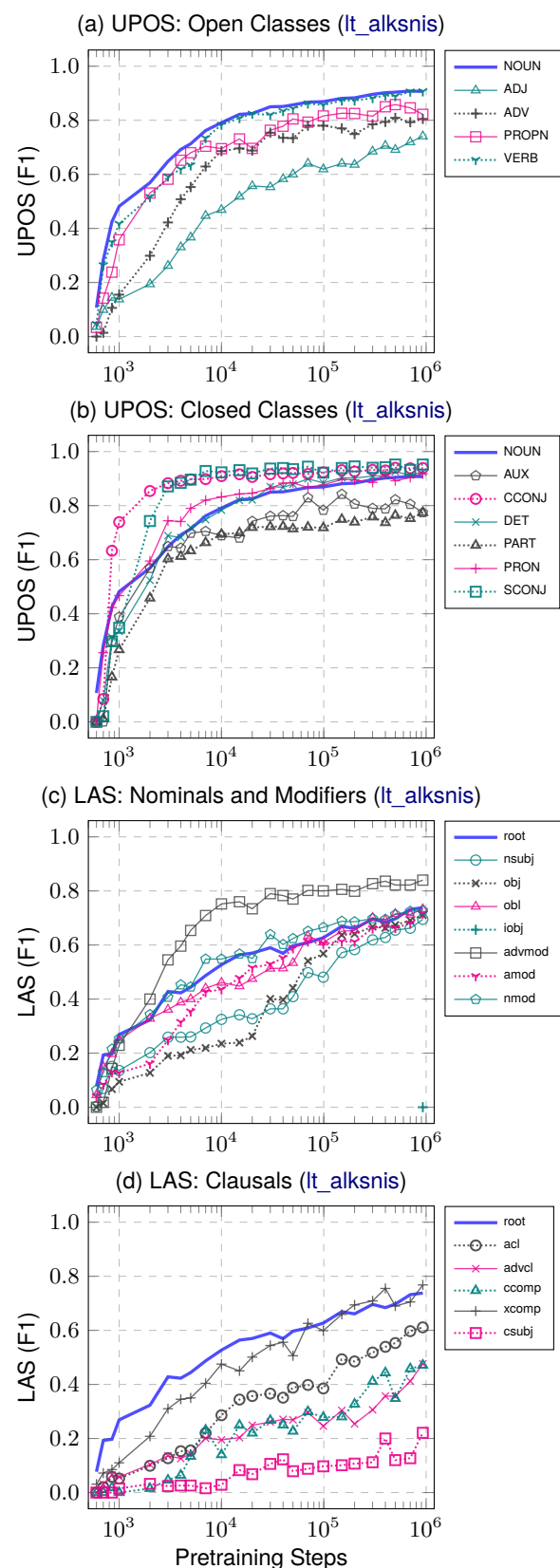


	Train	Dev	Test
Number of sentences	14,887	2,080	2,400
Number of words	255,954	34,348	37,162
Number of label occurrences			
ADJ	13,907	1,762	1,808
ADP	11,432	1,465	1,581
ADV	14,230	1,931	2,197
AUX	7,086	882	950
CCONJ	10,435	1,363	1,482
DET	11,917	1,490	1,765
INTJ	107	46	68
NOUN	69,455	9,143	9,695
NUM	3,343	415	402
PART	5,070	739	938
PRON	8,006	1,144	1,217
PROPN	11,109	1,348	1,521
PUNCT	44,053	6,423	6,831
SCONJ	5,998	743	847
SYM	495	45	47
VERB	38,374	5,282	5,620
X	937	127	193
acl	4,266	545	611
advcl	4,810	635	654
advmod	14,346	1,947	2,187
amod	13,882	1,701	1,834
appos	501	80	67
aux	3,264	420	470
case	11,610	1,490	1,598
cc	10,544	1,379	1,511
ccomp	4,342	561	614
compound	57	7	2
conj	14,635	2,151	2,128
cop	3,779	455	475
csubj	853	103	115
dep	691	79	96
det	5,988	762	916
discourse	3,497	532	684
dislocated	11	2	1
fixed	1,217	156	204
flat	2,474	275	391
goeswith	18	1	3
iobj	6,141	828	874
mark	5,512	666	758
nmod	27,975	3,525	3,837
nsubj	21,561	2,816	3,014
nummod	2,400	326	282
obj	10,929	1,571	1,555
obl	16,338	2,074	2,278
orphan	406	86	100
parataxis	997	141	111
punct	44,051	6,419	6,831
reparandum	10	4	3
root	14,887	2,080	2,400
vocative	71	11	29
xcomp	3,891	520	529

Table 28: Statistics of lv_lvtb

Figure 28: Syntactic acquisition through pretraining of OLMo-2-7B on Latvian (lv_lvtb).

C.21. Lithuanian

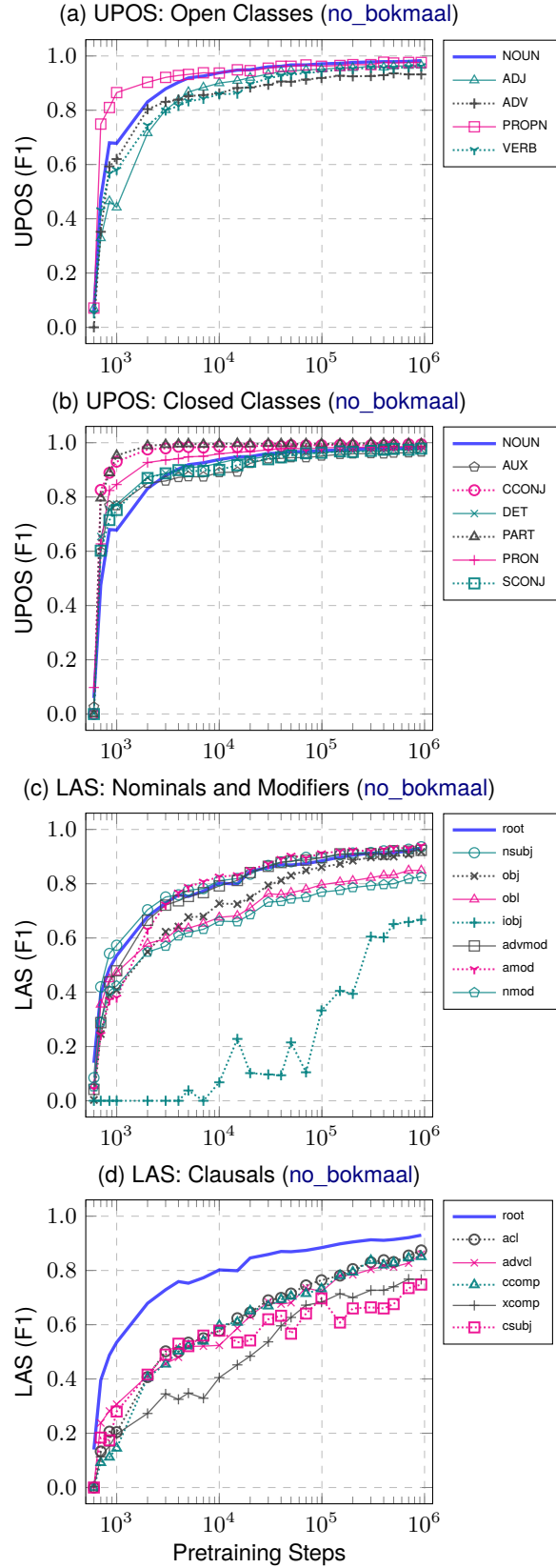


	Train	Dev	Test
Number of sentences	2,341	617	684
Number of words	47,641	11,560	10,846
Number of label occurrences			
ADJ	3,274	768	609
ADP	1,490	402	439
ADV	1,826	432	570
AUX	453	129	102
CCONJ	2,136	471	474
DET	1,181	311	288
INTJ	9	2	8
NOUN	14,933	3,534	2,810
NUM	1,328	202	169
PART	951	246	356
PRON	1,688	283	448
PROPN	983	365	245
PUNCT	8,756	2,081	2,059
SCONJ	917	227	279
SYM	45	9	8
VERB	6,604	1,758	1,818
X	1,067	340	164
<i>acl</i>	1,586	379	297
<i>advcl</i>	851	252	226
<i>advmod</i>	2,569	620	870
<i>amod</i>	2,653	661	481
<i>appos</i>	125	26	43
<i>case</i>	1,476	398	432
<i>cc</i>	2,112	467	470
<i>ccomp</i>	267	88	108
<i>compound</i>	1	0	1
<i>conj</i>	3,650	663	761
<i>cop</i>	453	129	102
<i>csubj</i>	322	81	64
<i>dep</i>	79	7	17
<i>det</i>	664	159	165
<i>discourse</i>	1	0	4
<i>flat</i>	212	117	52
<i>iobj</i>	2	2	0
<i>mark</i>	949	236	289
<i>nmod</i>	7,052	1,663	1,055
<i>nsubj</i>	2,579	703	624
<i>nummod</i>	1,011	153	113
<i>obj</i>	1,742	408	379
<i>obl</i>	5,057	1,328	1,210
<i>orphan</i>	3	0	0
<i>parataxis</i>	402	88	116
<i>punct</i>	8,756	2,081	2,059
<i>root</i>	2,341	617	684
<i>xcomp</i>	726	234	224

Table 29: Statistics of *lt_alksnis*

Figure 29: Syntactic acquisition through pretraining of OLMo-2-7B on Lithuanian (*lt_alksnis*).

C.22. Norwegian

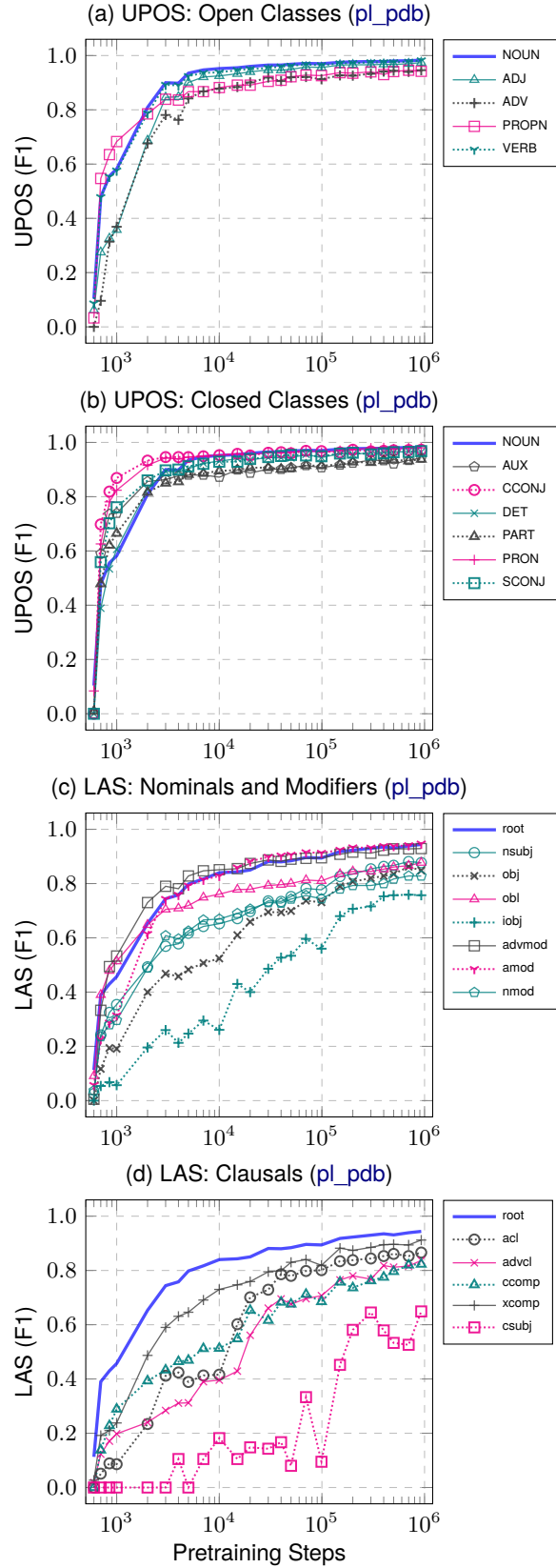


	Train	Dev	Test
Number of sentences	15,696	2,409	1,939
Number of words	243,886	36,369	29,966
Number of label occurrences			
ADJ	21,215	3,145	2,450
ADP	29,753	4,497	3,773
ADV	10,038	1,440	1,220
AUX	13,394	2,064	1,723
CCONJ	8,474	1,307	1,003
DET	11,416	1,671	1,309
INTJ	167	35	22
NOUN	45,008	6,768	5,477
NUM	3,049	486	410
PART	5,514	808	676
PRON	17,817	2,621	2,207
PROPN	14,282	2,175	1,803
PUNCT	29,128	4,352	3,501
SCONJ	7,845	1,074	947
SYM	63	7	11
VERB	26,154	3,868	3,329
X	569	51	105
<i>acl</i>	4,158	619	494
<i>advcl</i>	4,084	565	476
<i>advmod</i>	16,532	2,384	2,051
<i>amod</i>	11,260	1,684	1,257
<i>appos</i>	1,308	174	115
<i>aux</i>	6,996	1,128	954
<i>case</i>	29,238	4,420	3,705
<i>cc</i>	8,462	1,303	1,008
<i>ccomp</i>	2,322	308	315
<i>compound</i>	239	41	24
<i>conj</i>	9,452	1,377	1,138
<i>cop</i>	6,395	936	767
<i>csubj</i>	794	135	103
<i>det</i>	12,474	1,834	1,400
<i>discourse</i>	108	23	11
<i>dislocated</i>	163	33	28
<i>expl</i>	2,552	370	319
<i>flat</i>	5,018	713	646
<i>iobj</i>	471	74	51
<i>mark</i>	11,216	1,603	1,340
<i>nmod</i>	12,761	2,030	1,512
<i>nsubj</i>	21,553	3,172	2,785
<i>nummod</i>	1,831	245	222
<i>obj</i>	10,632	1,583	1,393
<i>obl</i>	14,971	2,221	1,930
<i>orphan</i>	2	1	0
<i>parataxis</i>	187	55	15
<i>punct</i>	29,128	4,352	3,501
<i>reparandum</i>	31	11	1
<i>root</i>	15,696	2,409	1,939
<i>xcomp</i>	3,852	566	466

Table 30: Statistics of *no_bokmaal*

Figure 30: Syntactic acquisition through pretraining of OLMo-2-7B on Norwegian (*no_bokmaal*).

C.23. Polish

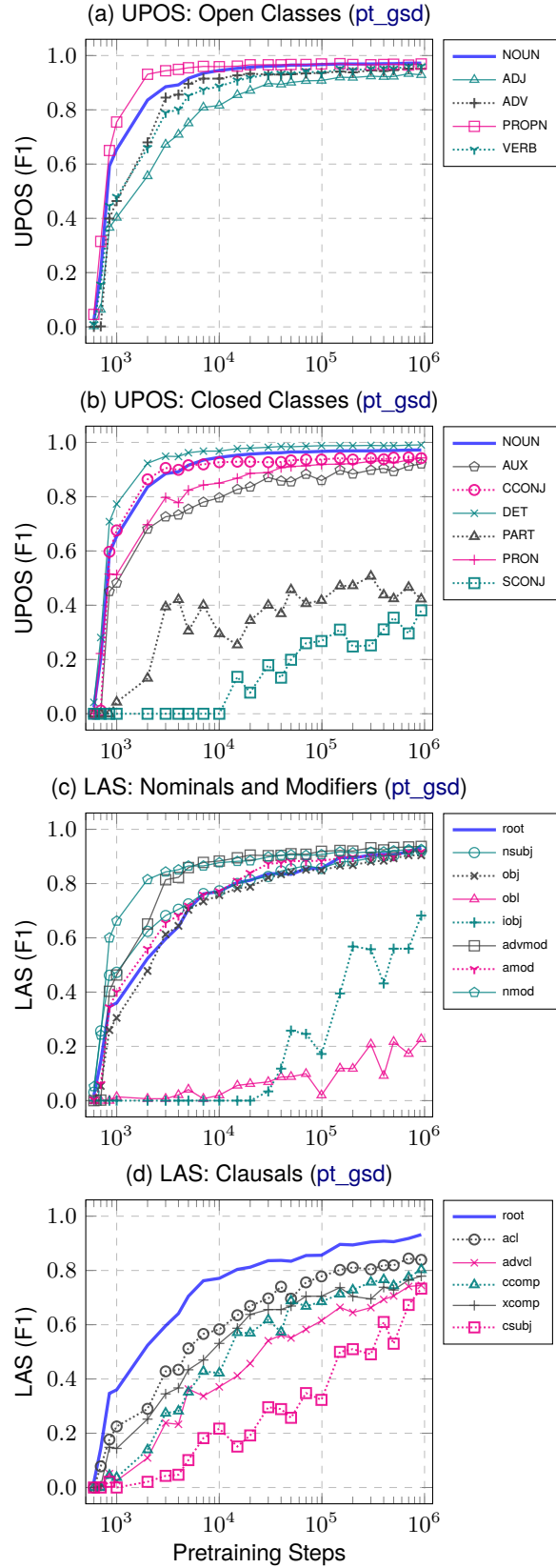


	Train	Dev	Test
Number of sentences	17,722	2,215	2,215
Number of words	281,685	34,677	33,616
Number of label occurrences			
ADJ	29,018	3,545	3,365
ADP	29,890	3,748	3,523
ADV	9,218	1,163	1,097
AUX	7,110	820	845
CCONJ	8,386	1,023	1,033
DET	7,576	920	852
INTJ	139	23	13
NOUN	71,287	8,828	8,519
NUM	2,113	262	258
PART	8,830	1,021	1,059
PRON	13,274	1,620	1,592
PROPN	9,634	1,214	1,152
PUNCT	46,497	5,748	5,628
SCONJ	6,136	706	689
SYM	14	3	4
VERB	31,725	3,922	3,880
X	838	111	107
<i>acl</i>	6,059	782	681
<i>advcl</i>	3,236	374	380
<i>advmod</i>	16,988	2,063	1,994
<i>amod</i>	20,076	2,425	2,375
<i>appos</i>	1,727	232	213
<i>aux</i>	4,352	529	527
<i>case</i>	29,129	3,662	3,440
<i>cc</i>	8,494	1,031	1,059
<i>ccomp</i>	2,735	316	313
<i>conj</i>	12,591	1,534	1,545
<i>cop</i>	2,756	291	318
<i>csubj</i>	185	22	16
<i>dep</i>	10	1	3
<i>det</i>	5,448	664	607
<i>discourse</i>	128	18	12
<i>expl</i>	4,960	600	595
<i>fixed</i>	3,157	381	357
<i>flat</i>	2,081	276	258
<i>iobj</i>	5,311	649	673
<i>list</i>	208	39	26
<i>mark</i>	6,175	711	698
<i>nmod</i>	21,074	2,562	2,454
<i>nsubj</i>	17,077	2,081	2,011
<i>nummod</i>	1,916	241	230
<i>obj</i>	12,242	1,526	1,479
<i>obl</i>	22,628	2,908	2,703
<i>orphan</i>	100	11	7
<i>parataxis</i>	2,313	262	260
<i>punct</i>	46,496	5,748	5,628
<i>root</i>	17,722	2,215	2,215
<i>vocative</i>	218	28	36
<i>xcomp</i>	4,093	495	503

Table 31: Statistics of *pl_pdb*

Figure 31: Syntactic acquisition through pretraining of OLMo-2-7B on Polish (*pl_pdb*).

C.24. Portuguese

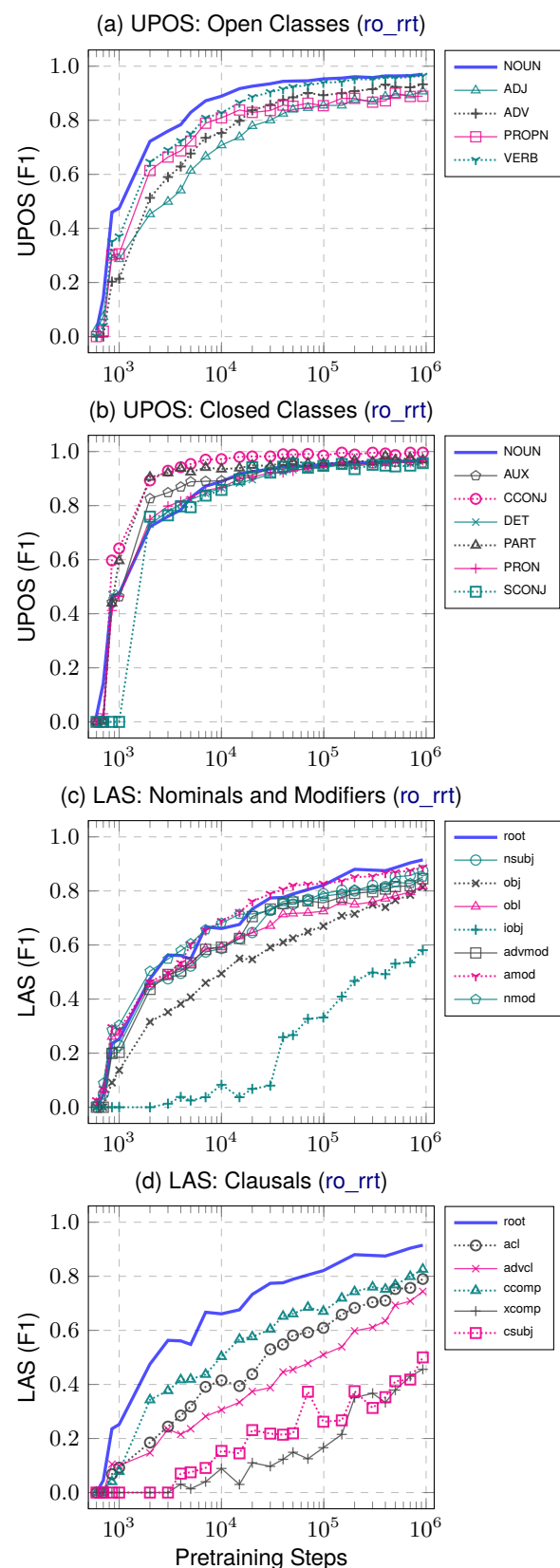


	Train	Dev	Test
Number of sentences	9,616	1,204	1,200
Number of words	255,116	32,073	31,477
Number of label occurrences			
ADJ	12,076	1,486	1,477
ADP	40,980	5,201	5,045
ADV	7,835	985	950
AUX	5,588	687	661
CCONJ	8,314	1,086	1,026
DET	37,999	4,907	4,696
NOUN	45,237	5,769	5,584
NUM	6,824	832	856
PART	459	60	46
PRON	6,214	763	743
PROPN	25,908	3,143	3,228
PUNCT	33,234	4,114	4,074
SCONJ	975	90	120
SYM	813	98	107
VERB	22,356	2,812	2,805
X	304	40	59
acl	4,948	665	630
advcl	2,143	243	284
advmod	7,167	884	875
amod	12,367	1,550	1,502
appos	4,880	579	550
aux	3,786	471	454
case	38,458	4,862	4,727
cc	6,187	815	760
ccomp	1,838	244	219
compound	18	2	1
conj	8,253	1,046	1,045
cop	1,779	214	203
csubj	490	50	61
dep	195	20	16
det	37,918	4,902	4,685
discourse	2	0	1
expl	611	65	60
fixed	940	107	115
flat	5,948	676	770
iobj	377	56	45
mark	5,312	671	677
nmod	37,002	4,834	4,583
nsubj	13,188	1,661	1,659
nummod	3,777	452	483
obj	9,046	1,162	1,127
obl	2,420	189	270
parataxis	670	74	70
punct	33,234	4,114	4,074
root	9,616	1,204	1,200
xcomp	2,546	261	331

Table 32: Statistics of pt_gsd

Figure 32: Syntactic acquisition through pretraining of OLMo-2-7B on Portuguese (pt_gsd).

C.25. Romanian

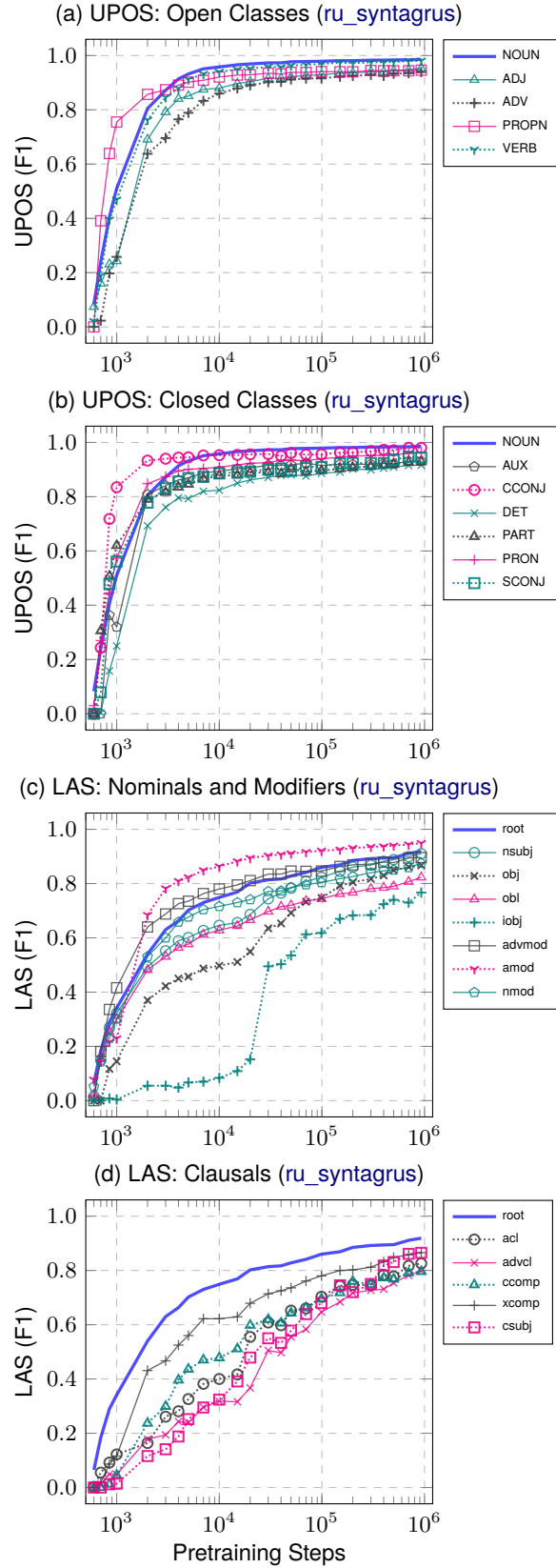


	Train	Dev	Test
Number of sentences	8,043	752	729
Number of words	185,125	17,073	16,324
Number of label occurrences			
ADJ	12,855	1,271	1,172
ADP	26,329	2,393	2,333
ADV	7,554	662	650
AUX	7,289	658	618
CCONJ	5,940	519	471
DET	10,196	931	898
INTJ	32	3	6
NOUN	45,990	4,226	4,042
NUM	4,675	418	456
PART	4,149	378	358
PRON	10,014	931	862
PROPN	4,919	511	455
PUNCT	23,691	2,223	2,083
SCONJ	1,896	151	154
VERB	19,466	1,784	1,749
X	130	14	17
acl	4,168	345	356
advcl	2,880	256	231
advmod	8,491	776	704
amod	11,093	1,118	1,040
appos	951	100	87
aux	5,414	490	478
case	23,007	2,064	2,072
cc	5,647	492	464
ccomp	2,115	192	224
compound	55	3	3
conj	7,417	660	614
cop	1,828	167	135
csubj	755	68	41
dep	32	2	3
det	10,090	914	879
discourse	16	1	3
expl	3,968	377	325
fixed	5,268	536	475
flat	1,237	156	91
goeswith	67	4	11
iobj	1,366	140	139
list	9	0	1
mark	5,480	457	454
nmod	17,054	1,582	1,532
nsubj	10,204	939	894
nummod	3,565	326	352
obj	6,430	588	562
obl	12,355	1,124	1,128
orphan	61	3	5
parataxis	1,337	133	99
punct	23,691	2,223	2,083
reparandum	1	0	0
root	8,043	752	729
vocative	54	3	4
xcomp	976	82	106

Table 33: Statistics of ro_rrt

Figure 33: Syntactic acquisition through pretraining of OLMo-2-7B on Romanian (ro_rrt).

C.26. Russian

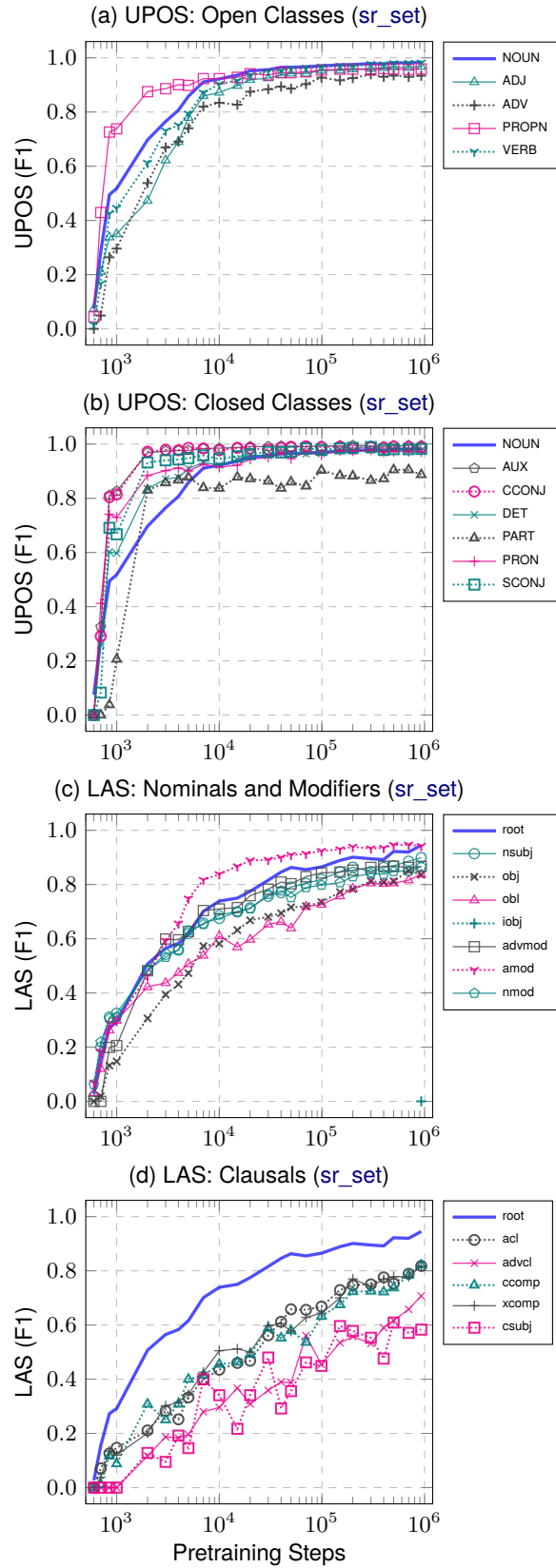


	Train	Dev	Test
Number of sentences	69,630	8,906	8,800
Number of words	1,204,640	153,325	157,718
Number of label occurrences			
<i>ADJ</i>	116,158	15,103	14,471
<i>ADP</i>	112,193	13,717	15,062
<i>ADV</i>	60,719	7,783	8,085
<i>AUX</i>	10,819	1,390	1,518
<i>CCONJ</i>	43,636	5,672	5,736
<i>DET</i>	32,780	4,265	4,094
<i>INTJ</i>	183	24	23
<i>NOUN</i>	287,221	36,238	36,568
<i>NUM</i>	15,169	1,734	2,528
<i>PART</i>	39,634	5,125	4,921
<i>PRON</i>	57,486	7,444	8,015
<i>PROPN</i>	44,399	5,473	5,883
<i>PUNCT</i>	222,074	29,186	29,463
<i>SCONJ</i>	22,668	2,865	2,992
<i>SYM</i>	1,052	62	165
<i>VERB</i>	137,570	17,110	18,146
<i>X</i>	879	134	48
<i>acl</i>	20,368	2,594	2,593
<i>advcl</i>	12,362	1,542	1,810
<i>advmod</i>	77,228	9,737	9,860
<i>amod</i>	95,863	12,319	11,954
<i>appos</i>	9,579	1,114	1,371
<i>aux</i>	5,236	583	788
<i>case</i>	110,899	13,647	14,943
<i>cc</i>	44,657	5,832	5,869
<i>ccomp</i>	6,726	803	865
<i>compound</i>	1,386	164	141
<i>conj</i>	56,740	8,034	7,640
<i>cop</i>	5,555	805	729
<i>csubj</i>	6,743	755	897
<i>dep</i>	4	0	1
<i>det</i>	30,241	3,930	3,752
<i>discourse</i>	979	127	116
<i>dislocated</i>	12	0	0
<i>expl</i>	801	107	100
<i>fixed</i>	11,125	1,354	1,445
<i>flat</i>	8,812	1,276	1,374
<i>iobj</i>	10,424	1,374	1,425
<i>list</i>	11	1	0
<i>mark</i>	21,281	2,601	2,804
<i>nmod</i>	99,066	12,409	12,179
<i>nsubj</i>	89,215	10,985	11,546
<i>nummod</i>	11,229	1,135	1,622
<i>obj</i>	39,513	4,890	5,124
<i>obl</i>	97,544	12,304	13,404
<i>orphan</i>	1,106	179	171
<i>parataxis</i>	24,841	3,125	3,285
<i>punct</i>	222,074	29,186	29,463
<i>root</i>	69,630	8,906	8,800
<i>vocative</i>	166	16	32
<i>xcomp</i>	13,224	1,491	1,615

Table 34: Statistics of *ru_syntagrus*

Figure 34: Syntactic acquisition through pretraining of OLMo-2-7B on Russian (*ru_syntagrus*).

C.27. Serbian

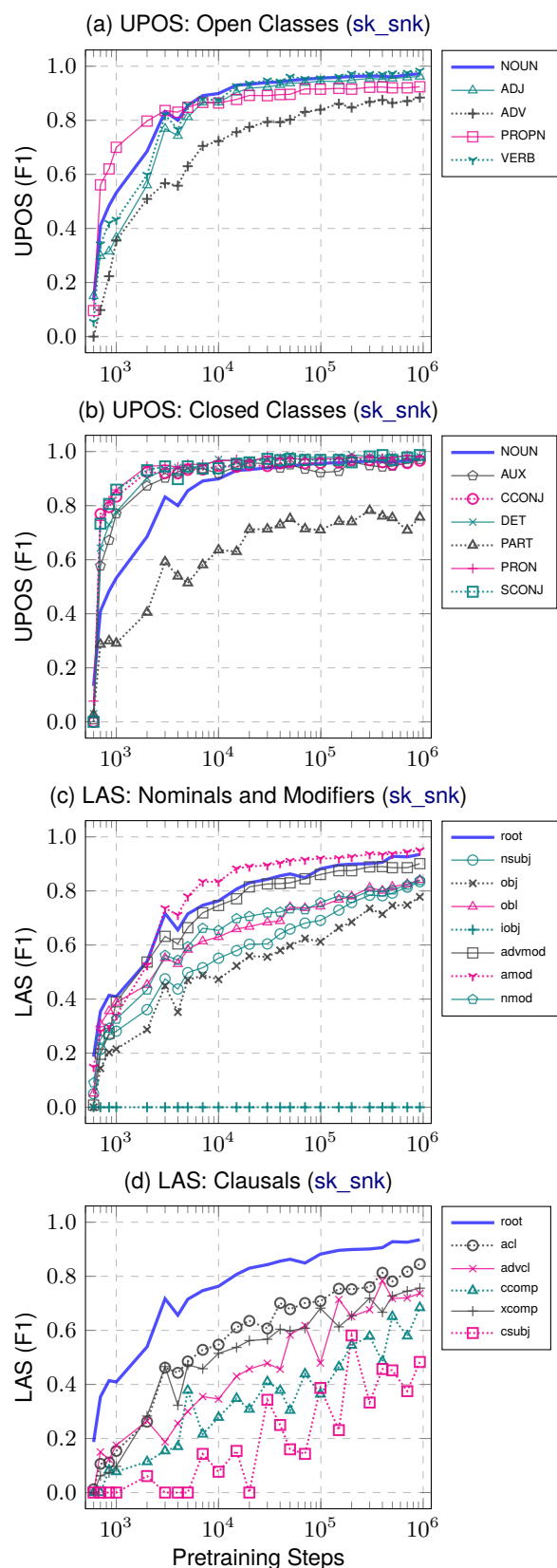


	Train	Dev	Test
Number of sentences	3,328	536	520
Number of words	74,259	11,993	11,421
Number of label occurrences			
ADJ	8,835	1,464	1,362
ADP	7,130	1,124	1,122
ADV	2,543	388	393
AUX	4,667	817	719
CCONJ	2,541	410	380
DET	2,848	433	358
INTJ	3	1	0
NOUN	18,103	2,853	2,862
NUM	944	179	155
PART	461	74	51
PRON	1,843	306	254
PROPN	5,622	860	930
PUNCT	9,351	1,574	1,418
SCONJ	2,729	422	415
SYM	1	2	1
VERB	6,406	1,036	971
X	232	50	30
acl	1,571	240	202
advcl	564	106	79
advmod	2,213	339	318
amod	7,102	1,144	1,105
appos	563	86	78
aux	3,350	594	535
case	7,233	1,131	1,138
cc	2,349	378	343
ccomp	1,012	139	163
compound	2	1	0
conj	2,920	471	432
cop	1,185	205	166
csubj	160	24	24
dep	2	0	0
det	1,476	204	198
discourse	390	57	53
expl	1,076	193	155
fixed	357	70	59
flat	2,152	368	357
iobj	2	0	0
list	6	5	0
mark	2,426	374	372
nmod	6,598	962	1,121
nsubj	5,714	961	859
nummod	808	147	130
obj	2,651	401	395
obl	5,919	959	925
orphan	35	6	4
parataxis	912	176	135
punct	9,351	1,574	1,418
root	3,328	536	520
vocative	1	0	0
xcomp	831	142	137

Table 35: Statistics of sr_set

Figure 35: Syntactic acquisition through pretraining of OLMo-2-7B on Serbian (sr_set).

C.28. Slovak

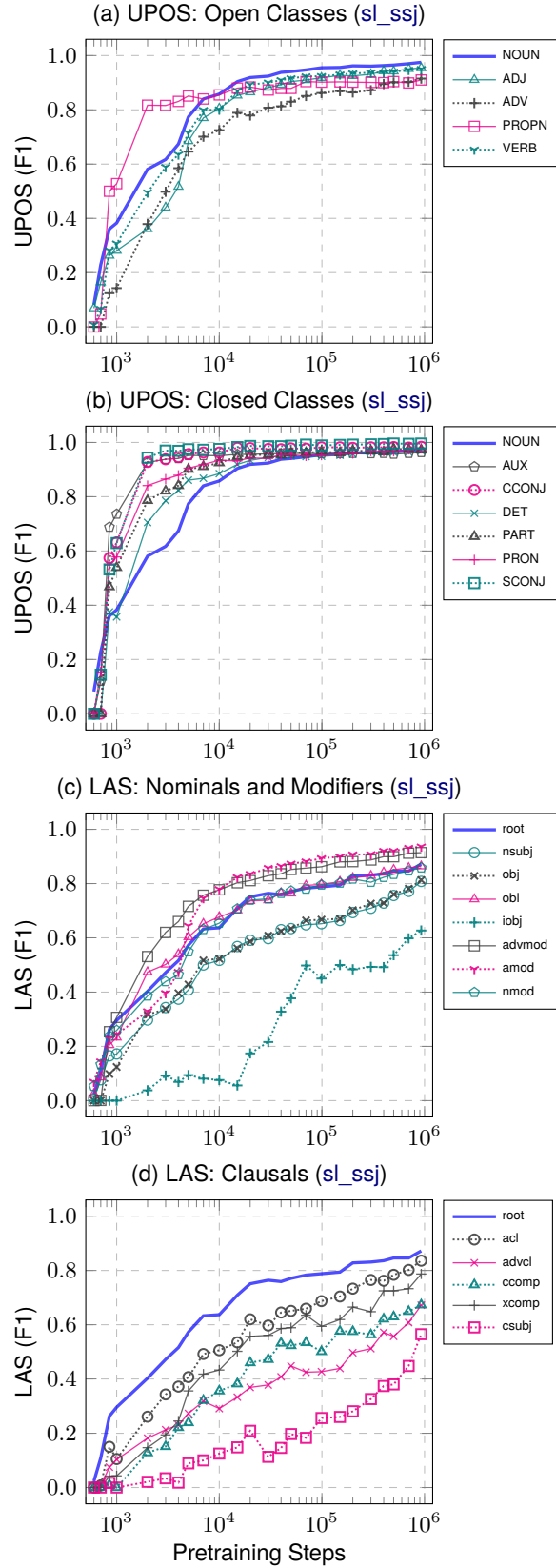


	Train	Dev	Test
Number of sentences	8,483	1,060	1,061
Number of words	80,628	12,733	12,736
Number of label occurrences			
ADJ	6,321	1,635	1,518
ADP	6,209	1,154	1,249
ADV	3,641	394	410
AUX	3,197	286	240
CCONJ	2,527	401	448
DET	3,446	507	451
INTJ	66	0	0
NOUN	15,269	3,219	3,173
NUM	661	442	470
PART	1,591	179	146
PRON	5,718	353	364
PROPN	2,916	827	1,087
PUNCT	15,532	1,809	1,789
SCONJ	1,541	213	111
SYM	2	16	5
VERB	11,757	1,187	1,124
X	234	111	151
<i>acl</i>	685	147	152
<i>advcl</i>	729	77	33
<i>advmod</i>	4,746	545	507
<i>amod</i>	5,232	1,429	1,359
<i>appos</i>	162	65	68
<i>aux</i>	2,033	78	28
<i>case</i>	6,204	1,153	1,254
<i>cc</i>	2,467	390	446
<i>ccomp</i>	1,182	62	17
<i>compound</i>	5	0	0
<i>conj</i>	3,004	507	593
<i>cop</i>	1,163	208	212
<i>csubj</i>	132	35	18
<i>dep</i>	500	53	52
<i>det</i>	1,991	324	288
<i>discourse</i>	45	1	0
<i>expl</i>	2,346	232	266
<i>fixed</i>	129	28	32
<i>flat</i>	31	24	44
<i>iobj</i>	80	7	1
<i>mark</i>	1,621	216	113
<i>nmod</i>	4,188	1,343	1,426
<i>nsubj</i>	5,622	902	818
<i>nummod</i>	406	311	355
<i>obj</i>	4,237	559	514
<i>obl</i>	6,565	1,031	1,143
<i>orphan</i>	60	8	9
<i>parataxis</i>	39	8	18
<i>punct</i>	15,532	1,809	1,789
<i>root</i>	8,483	1,060	1,061
<i>vocative</i>	34	0	0
<i>xcomp</i>	975	121	120

Table 36: Statistics of *sk_snk*

Figure 36: Syntactic acquisition through pretraining of OLMo-2-7B on Slovak (*sk_snk*).

C.29. Slovenian

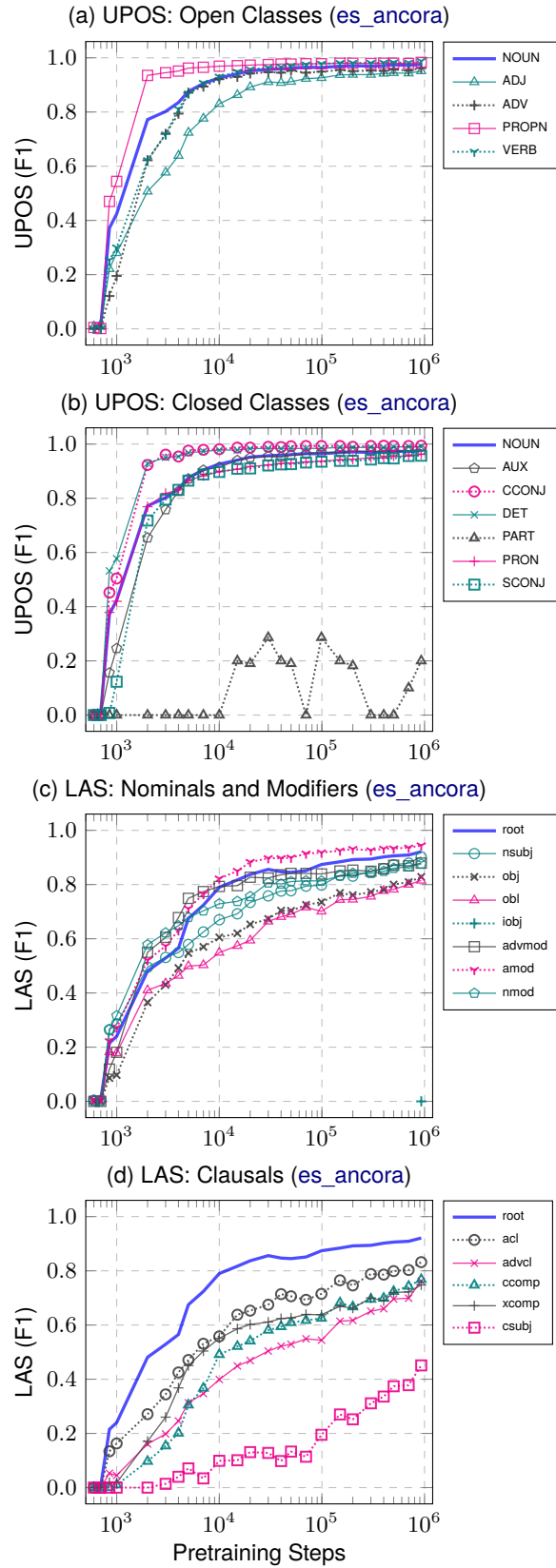


	Train	Dev	Test
Number of sentences	10,903	1,250	1,282
Number of words	215,155	26,500	25,442
Number of label occurrences			
ADJ	22,726	2,919	2,781
ADP	19,351	2,535	2,349
ADV	9,511	1,167	1,125
AUX	14,028	1,594	1,705
CCONJ	8,960	1,146	1,131
DET	7,530	867	955
INTJ	43	1	1
NOUN	45,586	5,790	5,489
NUM	4,625	562	398
PART	6,036	788	718
PRON	7,371	860	875
PROPN	8,269	1,075	895
PUNCT	32,110	3,745	3,623
SCONJ	7,171	905	898
SYM	171	7	21
VERB	19,816	2,376	2,400
X	1,851	163	78
<i>acl</i>	3,329	469	437
<i>advcl</i>	1,970	240	246
<i>advmod</i>	16,288	2,021	1,929
<i>amod</i>	17,627	2,316	2,165
<i>appos</i>	1,497	171	164
<i>aux</i>	9,773	1,081	1,162
<i>case</i>	19,810	2,592	2,415
<i>cc</i>	7,368	949	940
<i>ccomp</i>	1,624	195	203
<i>conj</i>	9,293	1,134	1,108
<i>cop</i>	4,241	512	542
<i>csubj</i>	722	93	78
<i>dep</i>	170	6	12
<i>det</i>	4,732	533	617
<i>discourse</i>	152	15	15
<i>dislocated</i>	10	0	0
<i>expl</i>	2,998	350	361
<i>fixed</i>	987	117	95
<i>flat</i>	2,291	248	156
<i>goeswith</i>	1	0	0
<i>iobj</i>	1,412	135	145
<i>list</i>	583	82	16
<i>mark</i>	6,417	827	804
<i>nmod</i>	15,887	2,182	1,887
<i>nsubj</i>	10,589	1,273	1,312
<i>nummod</i>	3,532	457	311
<i>obj</i>	9,191	1,134	1,085
<i>obl</i>	14,049	1,735	1,722
<i>orphan</i>	797	101	83
<i>parataxis</i>	3,152	349	325
<i>punct</i>	32,108	3,745	3,623
<i>root</i>	10,903	1,250	1,282
<i>vocative</i>	64	5	1
<i>xcomp</i>	1,588	183	201

Table 37: Statistics of *sl_ss*

Figure 37: Syntactic acquisition through pretraining of OLMo-2-7B on Slovenian (*sl_ss*).

C.30. Spanish

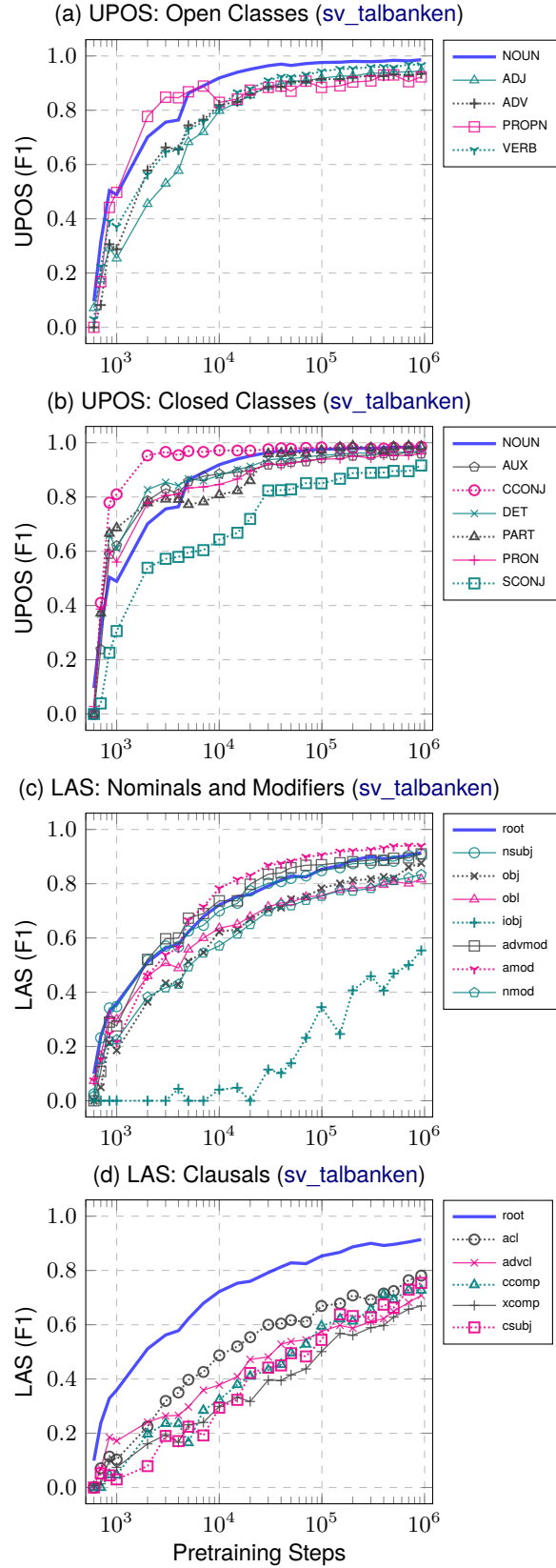


	Train	Dev	Test
Number of sentences	14,287	1,654	1,721
Number of words	453,039	53,476	53,622
Number of label occurrences			
ADJ	29,418	3,541	3,468
ADP	71,074	8,523	8,332
ADV	14,878	1,710	1,763
AUX	11,061	1,200	1,304
CCONJ	12,211	1,454	1,439
DET	68,315	8,061	8,055
INTJ	151	10	16
NOUN	81,411	9,593	9,531
NUM	7,277	920	977
PART	122	14	18
PRON	20,397	2,370	2,417
PROPN	34,259	4,037	4,094
PUNCT	52,988	6,303	6,334
SCONJ	10,128	1,136	1,210
SYM	358	41	28
VERB	38,987	4,563	4,636
X	4	0	0
<i>acl</i>	8,580	1,054	983
<i>advcl</i>	7,326	811	839
<i>advmod</i>	15,617	1,794	1,850
<i>amod</i>	23,784	2,873	2,822
<i>appos</i>	6,096	717	704
<i>aux</i>	5,889	668	718
<i>case</i>	61,994	7,418	7,302
<i>cc</i>	12,354	1,466	1,455
<i>ccomp</i>	4,950	533	610
<i>compound</i>	1,279	156	137
<i>conj</i>	13,161	1,559	1,560
<i>cop</i>	4,560	468	519
<i>csubj</i>	988	92	94
<i>dep</i>	238	9	24
<i>det</i>	68,456	8,095	8,110
<i>discourse</i>	4	0	0
<i>dislocated</i>	3	0	0
<i>expl</i>	4,933	573	589
<i>fixed</i>	6,661	795	711
<i>flat</i>	11,670	1,331	1,435
<i>list</i>	5	0	16
<i>mark</i>	15,700	1,811	1,839
<i>nmod</i>	31,882	3,864	3,661
<i>nsubj</i>	24,023	2,741	2,877
<i>nummod</i>	4,806	603	596
<i>obj</i>	19,966	2,371	2,356
<i>obl</i>	26,409	3,157	3,227
<i>orphan</i>	1	0	0
<i>parataxis</i>	538	84	80
<i>punct</i>	52,981	6,301	6,331
<i>root</i>	14,287	1,654	1,721
<i>vocative</i>	1	0	0
<i>xcomp</i>	3,897	478	456

Table 38: Statistics of *es_ancora*

Figure 38: Syntactic acquisition through pretraining of OLMo-2-7B on Spanish (*es_ancora*).

C.31. Swedish

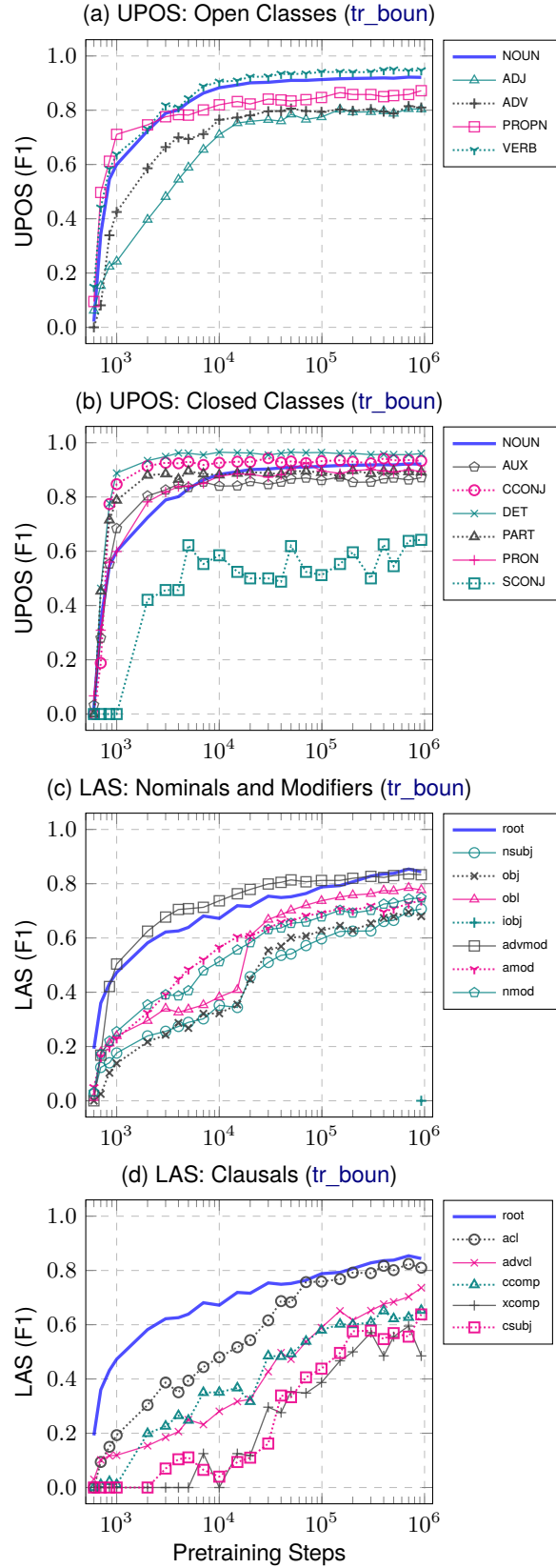


	Train	Dev	Test
Number of sentences	4,303	504	1,219
Number of words	66,646	9,797	20,377
Number of label occurrences			
ADJ	5,876	936	1,859
ADP	7,860	1,080	2,299
ADV	4,628	809	1,504
AUX	2,851	410	1,021
CCONJ	2,456	456	791
DET	3,375	499	1,003
INTJ	21	11	3
NOUN	16,096	2,190	4,711
NUM	1,324	59	357
PART	1,222	246	406
PRON	4,385	839	1,449
PROPN	1,081	36	243
PUNCT	7,303	962	2,104
SCONJ	1,368	243	491
SYM	75	10	1
VERB	6,725	1,011	2,134
X	0	0	1
<i>acl</i>	1,323	246	427
<i>advcl</i>	1,136	207	406
<i>advmod</i>	4,407	797	1,466
<i>amod</i>	4,251	630	1,288
<i>appos</i>	361	35	103
<i>aux</i>	1,713	237	614
<i>case</i>	7,318	978	2,121
<i>cc</i>	2,434	448	771
<i>ccomp</i>	358	47	108
<i>compound</i>	589	65	166
<i>conj</i>	3,054	484	898
<i>cop</i>	1,137	173	406
<i>csubj</i>	249	32	93
<i>dep</i>	1	0	0
<i>det</i>	3,363	501	1,008
<i>discourse</i>	12	8	2
<i>dislocated</i>	88	17	19
<i>expl</i>	381	43	94
<i>fixed</i>	104	23	26
<i>flat</i>	122	11	26
<i>goeswith</i>	0	0	1
<i>iobj</i>	105	21	37
<i>list</i>	5	0	0
<i>mark</i>	2,559	495	884
<i>nmod</i>	4,237	621	1,181
<i>nsubj</i>	6,105	873	1,972
<i>nummod</i>	1,009	49	281
<i>obj</i>	2,863	479	930
<i>obl</i>	4,778	613	1,386
<i>orphan</i>	25	2	8
<i>parataxis</i>	288	51	79
<i>punct</i>	7,303	962	2,104
<i>root</i>	4,303	504	1,219
<i>vocative</i>	0	1	0
<i>xcomp</i>	665	144	253

Table 39: Statistics of *sv_talbanken*

Figure 39: Syntactic acquisition through pretraining of OLMo-2-7B on Swedish (*sv_talbanken*).

C.32. Turkish

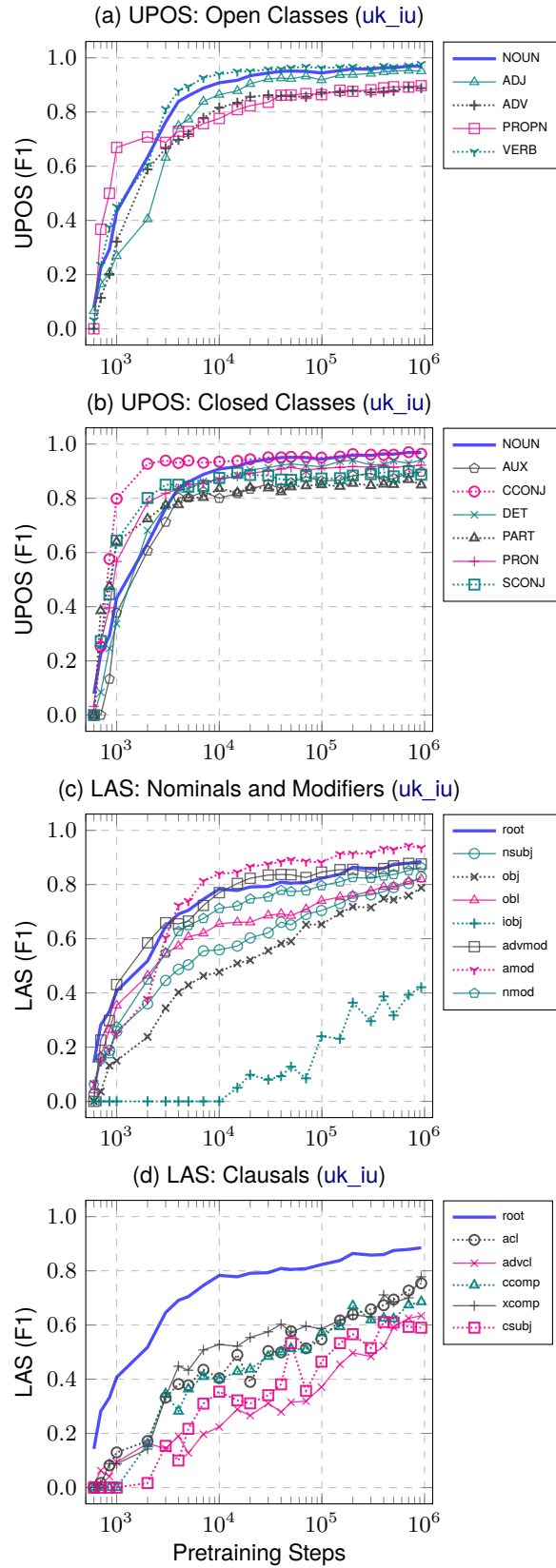


	Train	Dev	Test
Number of sentences	7,803	979	979
Number of words	100,713	12,289	12,210
Number of label occurrences			
ADJ	6,022	807	681
ADP	1,984	249	264
ADV	4,362	570	479
AUX	3,187	325	240
CCONJ	2,881	452	337
DET	4,048	505	546
INTJ	213	18	22
NOUN	30,686	3,713	3,952
NUM	2,105	225	276
PART	1,891	46	162
PRON	2,877	384	321
PROPN	6,438	766	677
PUNCT	16,715	2,043	2,028
SCONJ	179	15	26
VERB	17,121	2,171	2,199
X	4	0	0
<i>acl</i>	2,791	341	351
<i>advcl</i>	2,434	309	314
<i>advmod</i>	5,248	609	587
<i>amod</i>	5,650	759	766
<i>appos</i>	121	34	37
<i>aux</i>	425	26	11
<i>case</i>	2,214	271	294
<i>cc</i>	2,452	326	273
<i>ccomp</i>	1,247	174	175
<i>clf</i>	0	13	16
<i>compound</i>	3,101	366	385
<i>conj</i>	5,725	689	673
<i>cop</i>	2,455	262	192
<i>csubj</i>	467	43	72
<i>dep</i>	535	38	21
<i>det</i>	3,997	483	524
<i>discourse</i>	707	69	83
<i>dislocated</i>	17	3	3
<i>fixed</i>	56	13	16
<i>flat</i>	1,617	173	181
<i>iobj</i>	184	20	24
<i>list</i>	52	2	2
<i>mark</i>	179	9	14
<i>nmod</i>	10,138	1,173	1,170
<i>nsubj</i>	6,729	832	827
<i>nummod</i>	1,248	143	163
<i>obj</i>	5,914	752	740
<i>obl</i>	9,963	1,279	1,243
<i>orphan</i>	34	5	2
<i>parataxis</i>	207	22	23
<i>punct</i>	16,691	2,039	2,026
<i>root</i>	7,803	979	979
<i>vocative</i>	88	10	9
<i>xcomp</i>	224	23	14

Table 40: Statistics of *tr_boun*

Figure 40: Syntactic acquisition through pretraining of OLMo-2-7B on Turkish (*tr_boun*).

C.33. Ukrainian



	Train	Dev	Test
Number of sentences	5,521	673	898
Number of words	92,927	12,606	17,217
Number of label occurrences			
ADJ	8,831	1,258	1,976
ADP	8,313	1,157	1,588
ADV	5,138	635	718
AUX	856	82	120
CCONJ	3,722	446	630
DET	3,581	461	630
INTJ	109	4	10
NOUN	21,571	3,263	4,537
NUM	1,164	211	390
PART	2,876	292	374
PRON	4,151	404	519
PROPN	2,484	406	633
PUNCT	17,702	2,423	3,130
SCONJ	1,838	245	261
SYM	83	8	17
VERB	10,133	1,209	1,559
X	375	102	125
<i>acl</i>	1,215	220	205
<i>advcl</i>	1,315	139	164
<i>advmod</i>	5,227	617	752
<i>amod</i>	7,065	1,022	1,659
<i>appos</i>	636	97	128
<i>aux</i>	208	20	27
<i>case</i>	8,302	1,155	1,584
<i>cc</i>	3,748	451	630
<i>ccomp</i>	475	55	89
<i>compound</i>	350	77	101
<i>conj</i>	4,877	566	890
<i>cop</i>	648	62	93
<i>csubj</i>	426	70	57
<i>det</i>	2,589	301	494
<i>discourse</i>	1,911	183	227
<i>dislocated</i>	11	7	1
<i>expl</i>	94	9	13
<i>fixed</i>	218	31	39
<i>flat</i>	1,430	289	441
<i>goeswith</i>	2	1	2
<i>iobj</i>	298	15	33
<i>list</i>	13	3	9
<i>mark</i>	1,809	241	265
<i>nmod</i>	6,403	1,211	1,902
<i>nsubj</i>	6,012	793	917
<i>nummod</i>	735	111	185
<i>obj</i>	4,422	589	701
<i>obl</i>	6,609	849	1,172
<i>orphan</i>	208	15	33
<i>parataxis</i>	1,282	156	225
<i>punct</i>	17,702	2,423	3,130
<i>reparandum</i>	1	0	0
<i>root</i>	5,521	673	898
<i>vocative</i>	95	1	6
<i>xcomp</i>	1,070	154	145

Table 41: Statistics of *uk_iu*

Figure 41: Syntactic acquisition through pretraining of OLMo-2-7B on Ukrainian (*uk_iu*).

Which Languages are “Hot”, and which are “Cool”?

Using Universal Dependencies for Large-scale Comparisons of Subject Expression

Natalia Levshina

Radboud University
Nijmegen, The Netherlands
natalia.levshina@ru.nl

Abstract

This study uses Universal Dependencies to investigate subject omission across fifty-six news corpora and twenty geographic varieties of English. Building on McLuhan’s “hot–cool” distinction, Hall’s LC–HC continuum, and Bisang’s notion of overt vs. hidden complexity, it tests whether subject omission rates reflect degrees of contextual reliance. The results broadly support these theories: low-context, “hot” languages, such as German, Dutch and Swedish, show low omission rates, while high-context, “cool” languages, such as Japanese, Korean and Chinese, show higher rates. English behaves as a “hot” language but exhibits internal variation across varieties, with Southeast Asian varieties exhibiting more omission than African ones. The study provides large-scale quantitative evidence while highlighting the need for further theoretical and methodological refinement, particularly regarding the role of word order and agreement.

Keywords: “hot” and “cool” communication, high- and low-context cultures, subject, varieties of English

1. “Hot” and “Cool” Languages

The well-known distinction between “hot” and “cool” media, introduced by Marshall McLuhan (1964), captures differences in the degree of sensory detail provided by a medium and the level of audience participation required to complete the message. For example, films can be considered “hot” in comparison to books, as the former supply more sensory input and place fewer interpretative demands on the recipient.

A related and influential distinction is that by between high-context (HC) and low-context (LC) communication, proposed by E. T. Hall (1976). In HC communication, much of the information is inferred from context rather than stated explicitly, whereas LC communication relies on the explicit encoding of information with less shared background assumed. Although different versions of the scale exist (see discussion of this and other problems in Cardon 2008; Kittler et al., 2011), countries such as Germany and the USA are typically classified as LC cultures, while China and Japan are often situated close to the HC end of the spectrum (see Figure 1).



Figure 1: A continuum of high- and low-context cultures in Rösch & Segler (1987), based on Hall (1976).

Similar conceptual distinctions have also been proposed in linguistic typology. For example, Bisang (2009) differentiates between overt and

hidden complexity, contrasting obligatory explicit encoding of grammatical meanings with their inference from context and pragmatic principles. Likewise, Tsao (1977) distinguishes between “sentence-oriented” languages (e.g., English) and “discourse-oriented” ones (e.g., Chinese). Similar to McLuhan’s distinction, Ross (1982) suggested that languages can also be placed along a “hot–cool” continuum, particularly with regard to how anaphoric referents are expressed. For example, English, in which the speaker provides information about the referents with the help of anaphoric pronouns, as in (1), would be close to the “hot” pole of the continuum.

(1) *She is writing an article.*

In contrast, Japanese would be close to the “cool” pole because the recovery of the anaphoric referents in zero anaphora contexts requires more effort on the addressee’s part. Consider (2):

(2) *Kinoo ryoori o shita.*
Yesterday cooking OBJ do.PST
‘Yesterday, (someone) cooked’.

Other languages can be located in-between. For instance, in Russian, pronominal subjects can be either overt or implicit:

(3) *(Ona) pišet statju.*
3SG.NOM.F write.IPFV.PRES.3SG
article.SG.ACC
‘(She) is writing an article.’

The use and omission of referents are guided by pragmatic principles (Horn, 1984; Huang, 1984), whereby more accessible (i.e., recoverable, non-contrastive, topic-continuous) referents tend to be omitted more often than less accessible ones

(Levshina, 2022). Sometimes the information about the referent can be restored from the verb agreement, as in (3), but this is not always the case. For example, Japanese verbs, as in (2), have no person, number or gender agreement markers, but the use of personal pronouns is very restricted (McGloin et al., 2014: Ch. 8).

This study exploits the Universal Dependencies framework for large-scale cross-linguistic and cross-varietal comparison. The first case study examines subject expression across 56 news corpora in different languages. The second focuses on blogs in twenty varieties of English, a language typically characterized as “hot.” The study addresses two main questions:

- 1) Do cross-linguistic patterns of subject omission support existing theoretical accounts?
- 2) Are all English varieties equally “hot,” and if not, do differences correlate with sociolinguistic or cultural factors?

2. Subjects in 56 News Corpora

The data for this case study are drawn from 56 news corpora in the Leipzig Corpus Collection (Goldhahn et al., 2012). The corpora consist of sentences from diverse online news sources. Where possible, I selected corpora representing specific varieties:

- German: a general (“unspecified”) corpus (deu) and data from Switzerland (deu-ch).
- Portuguese: a general corpus (por), alongside data from Portugal (por-pt), Brazil (por-br) and Macao (por-mo).
- Arabic: a general corpus (ara) and data from Saudi Arabia (ara-sa).

The sentences were annotated with UD information using the default language models in Stanza (Qi et al., 2020). Using information about the dependencies and morphological features, I extracted potentially finite verbal predicates (UPOS = VERB or AUX) with the following dependency labels: *acl*, *acl:relcl*, *advcl*, *ccomp*, *csubj*, *parataxis* and *root*. Imperative forms were excluded because they normally do not allow for overt subjects. Every predicate was checked for having an overt subject as its dependent (*nsubj*, *nsubj:pass* or *csubj*).

The proportions of predicates without overt subjects are displayed in Figure 2. The distribution reflects the general outline of the LC–HC continuum (cf. Rösch and Segler, 1987; Mordon, 1999; Meyer, 2014) and the “hot–cool” distinction. Typical low-context (LC) varieties of German cluster at the lower end of the scale, alongside Afrikaans, Dutch, and the North Germanic languages, all showing low rates of subject omission (with Icelandic closer to the high-context pole). At the opposite end, prototypical high-context (HC) languages, such as Korean, Japanese, and Chinese, occupy the top positions.

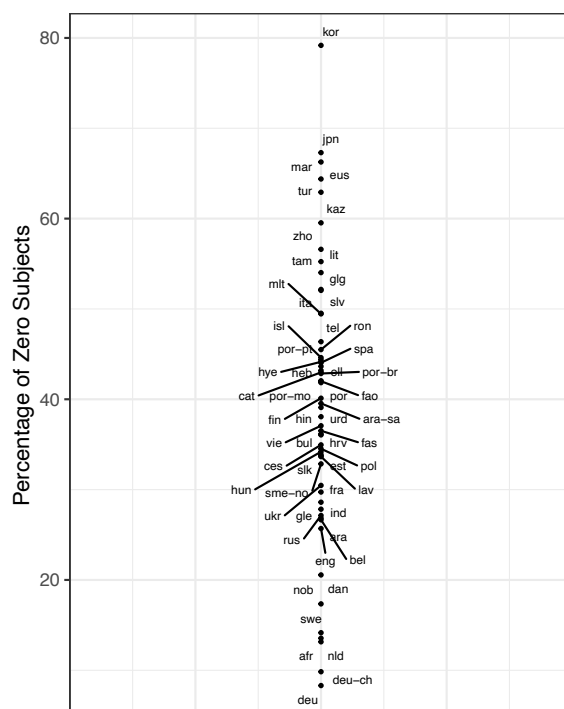


Figure 2: Percentages of predicates without overt subject in 56 online news corpora.

Typical LC German varieties are located at the bottom followed by Afrikaans, Dutch and some North Germanic languages with low subject omission rates (although Icelandic and Faroese are closer to the HC pole). In contrast, prototypical HC languages, such as Korean, Japanese and Chinese are located at the top.

Romance languages display variation: French patterns more closely with English toward the LC end, while the other Romance languages lie nearer the HC pole. Among Slavic languages, East Slavic varieties (Belarusian, Russian, Ukrainian) are relatively LC-oriented, Slovene is closest to the HC end, and Bulgarian, Croatian, Czech, Polish, and Slovak fall in between.

Regional varieties of German and Portuguese show little internal divergence, whereas Arabic exhibits greater variation, with Saudi Arabic positioning closer to the HC end than the general Arabic corpus.

Although the results broadly align with the LC–HC continuum, other factors are likely involved. In particular, verb-final languages tend to realize fewer arguments than VO languages, often by omitting the subject (Levshina, 2025). This behaviour may reflect pressures to manage processing costs and avoid long dependencies. The study also does not address subject indexing on the verb (e.g., subject clitics), which may be in complementary distribution with full forms. Because agreement systems and their representation in UD vary across languages, this issue is left for future research.

3. Subjects in Geographic Varieties of English

The second case study compares the rates of explicit and implicit subjects in varieties of English spoken in twenty countries: Australia, Bangladesh, Canada, Ghana, Great Britain, Hong Kong, India, Ireland, Jamaica, Kenya, Malaysia, New Zealand, Nigeria, Pakistan, the Philippines, Singapore, South Africa, Sri Lanka, Tanzania and the USA. The texts are blogs from the Global Corpus of Web-based English (Davies, 2013).

There is substantial research in variational linguistics on the use and omission of pronominal subjects in English. For example, the electronic World Atlas of Varieties of English, or eWAVE (Kortmann et al., 2020), describes the omission of referential and dummy pronouns as subjects (Features 43 and 44, respectively). The data relevant for the countries in this case study is shown in Table 1. According to eWAVE, traditional L1 varieties (e.g., most British dialects), normally have overt subjects, although exceptions are found in Newfoundland and Scottish Englishes, in which referential subject pro-drop is possible, being neither pervasive nor extremely rare. Subject omission is common in indigenized L2 varieties that originated in Asia (e.g., Indian, Pakistani, Sri Lankan and Malaysian Englishes), but not in those with African origin (e.g., spoken in Nigeria and Jamaica). A potential explanation may have to do with language contact: West Africa is an area in which languages requiring overt subject pronouns are particularly common (Dryer, 2013).

Omission of subject pronouns	Referential pronouns	Dummy pronouns
Absent	Australia, US (colloquial), Ghana, Jamaica, Nigeria	Ghana, Hong Kong, Jamaica, New Zealand, Nigeria
Extremely rare	Philippines	Ireland, Pakistan, Philippines
Neither pervasive nor extremely rare	Ireland, Pakistan, Sri Lanka, (white) South Africa	Sri Lanka
Pervasive or obligatory	Singapore (colloquial), Hong Kong, India, Malaysia	Singapore (colloquial), India, Malaysia

Table 1: Omission of subject pronouns in English varieties according to eWAVE.

There is no consensus about the position of the twenty English-speaking countries on the LC–HC continuum. Also, not all the countries have been investigated. As an illustration, Figure 3 displays the classifications from Meyer (2014) and Morden (1999). Note that the positions are only approximate, representing rankings, rather than absolute values. According to Morden (1999), New Zealanders and (white) South Africans represent the most LC cultures in this list, whereas Meyer (2014) regards the USA as the most LC culture. At the same time, there is some convergence. For example, African cultures are the most HC in both classifications, followed by the Indian subcontinent and then by Southeast Asia. In many classifications, the British are considered more HC than Americans, as in Rösch and Segler (1987) based on Hall (1976).

Although English is a “hot” language that is not regarded as favouring pro-drop, it is interesting to test if we can detect cross-lectal differences that can be correlated with the LC–HC continua. The data from the variationist literature partly contradicts data from cross-cultural comparisons. Most notably, the African cultures are considered HC, but the corresponding English varieties are reported to use more explicit subjects.

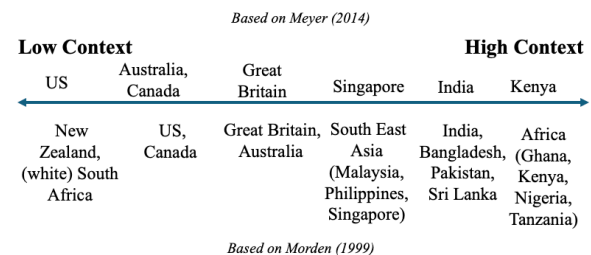


Figure 3: Ranking of cultures according to Meyer (2014) (top) and Morden (1999) (bottom).

To investigate subject omission rates, I sampled randomly 2,000 sentences for every variety from the Global Corpus of Web-based English (blogs) and annotated them with UD using the default Stanza neural model for English, performing the same analyses as the ones described in Section 2. In addition to that, all examples without subject were manually checked to avoid errors due to poor editing or sentence segmentation. Fixed expressions such as *thank you* and *see you* and boilerplate expressions (*The entry was posted on...*) were excluded.

After the semi-automatic cleaning procedure, the dataset contained 45,610 instances of finite predicates. Only 329 of them, or 0.7%, had no overt subject. Figure 4 displays the percentages in each variety. All of them are very low, which means that the web-based varieties of English are “hot”.

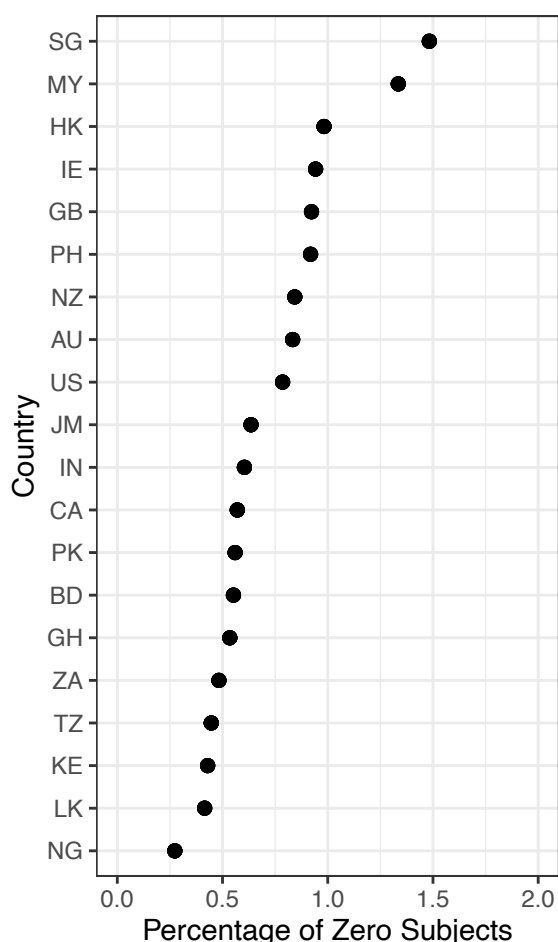


Figure 4: Percentage of zero subjects of finite non-imperative predicates in GloWbE blogs.

At the same time, we observe some variability, which to some extent matches the predictions based on eWAVE: Singapore and Malaysia have the highest proportion of zero subjects, while Nigeria has the lowest, as well as the other African countries in the sample. Although there are some unexpected findings, e.g., relatively low omission rates for South Asian varieties, a statistical analysis of standardized residuals (Agresti, 2002) reveals that only Singapore, Malaysia and Nigeria have observed frequencies that significantly ($|res| > 2$) deviate from expected values based on marginal frequencies.

A few examples of predicates without overt subjects are provided below, with the predicate tokens in bold. Very often, the first person subject (*I*) is omitted:

(4) *me was a bright young student, full of hope and confidence. hoping to step up into the world and prove my worth. **studied** my arse off, wanting to prove the world that me is good, me can do it. **told** meself that i have time to relax when me is done studying. **wasted** all my youth, the best of my years into mugging those damn tomes. they might as well be tombstones for all*

*they are worth. yeah **got** the grade, **got** a place where i hoped to get.* (GloWbE Malaysia)

(5) *Just **got** back from SMTown Live In Singapore concert.* (GloWbE Singapore)

(6) *Kinda **love** the 100% sarcastic tone of RPS, tho.* (GloWbE US)

(7) ***Agree** that Australia's university system has powerful linkages with Asian regions.* (GloWbE Great Britain)

The subject is also omitted in questions addressed to the 2nd person:

(8) ***Remember** mom brownies?* (GloWbE Canada)

Highly accessible abstract 3rd person referents (some event or situations) can be omitted, too:

(9) ***Happens** when you have 6 pups together in a pen with pee and poop all over the place.* (GloWbE Australia)

(10) ***Means** the industry is growing* (GloWbE Canada)

Among other predicates used without overt subject are *looks (like)*, *sounds (great)*, *reminds (you of)*..., which express the writer's opinions and impressions. Examples of dummy pronouns without any referent (cf. Feature 44 in eWAVE) occur, too, e.g., *seems/turns out (that) P*, as in the example below, but they are relatively infrequent.

(11) *Ahhh ... **turns** out these people were just as confused as us.* (NZ)

It must be emphasized that the frequencies of finite predicates without subject are extremely low, which suggests that grammatical constraints in this register largely override communicative, cultural or any other factors.

Note that it can be reasonable to exclude lexical or clausal subjects from the analysis, as most previous research focused on the omission of pronouns (pro-drop). When proportions of zero subjects are calculated with nouns and clauses excluded, the cross-country differences remain virtually identical (Pearson's $r = 0.99$).

4. Conclusions and Outlook

This study is a first attempt, to the best of my knowledge, to explore the rates of subject omission in a large number of languages and language varieties. The potential of Universal Dependencies is used here to find quantitative evidence automatically in large corpora.

The results lend support to previous claims about "hot", "medium" and "cool" languages and the LC-HC continuum and expand the typology to many new language and varieties. Although the view of English as a "hot" language with predominantly overt subjects is supported by the data, we also find interesting quantitative variation, with

Singapore and Malaysian English having higher rates of zero subjects than Nigerian English and other African varieties, in line with eWAVE.

Explaining these results is challenging, because it is difficult to disentangle linguistic and cultural-communicative factors. In particular, word order and certain types of verb agreement can play a role, as well as language contact. However, the study provides quantitative descriptive data that can be useful for linguistic and sociolinguistic typology.

The present study has several limitations. First, the web-based data are noisy, and the reliability of automatic UD annotations would benefit from more systematic manual validation. The problem of defining *subject* as a comparative category also deserves more attention (de Marneffe et al., 2014). Future work should extend the analysis to additional text types, particularly spoken interaction, in order to get a fuller range of contextual reliance in communicative practice.

5. Bibliographical References

- Agresti, A. (2002). *Categorical Data Analysis*. Wiley, Hoboken, NJ.
- Bisang, W. (2009). On the evolution of complexity—Sometimes less is more in East and mainland Southeast Asia. In G. Sampson, D. Gil and P. Trudgill (Eds.), *Language complexity as an evolving variable*. Oxford University Press, Oxford, pp. 34–49.
- Cardon, P. (2008). A critique of Hall's contexting model: A metaanalysis of literature on intercultural business and technical communication. *Journal of Business and Technical Communication*, 22(4): 399–428.
- Hall, E. T. (1976). *Beyond Culture*. Anchor Press/Doubleday, Garden City, NY.
- Horn, L. (1984). A new taxonomy for pragmatic inference: Q-based and R-based implicature. In Schiffrin, D. (Ed.), *Meaning, Form, and Use in Context: Linguistic Applications*. Georgetown University Press, Washington, pp. 11–42.
- Huang, C.-T. J. (1984). On the Distribution and Reference of Empty Pronouns. *Linguistic Inquiry*, 15(4):531–574.
- Kittler, M.G., Rygl, D. and Mackinnon, A. (2011). Beyond culture or beyond control? Reviewing the use of Hall's high/low-context concept. *International Journal of Cross Cultural Management*, 11(1): 63–82. <https://doi.org/10.1177/1470595811398797>
- Levshina, N. (2022). *Communicative Efficiency: Language Structure and Use*. Cambridge University Press, Cambridge.
- Levshina, N. (2025). The paradox of SOV: A case for token-based typology. *Glottometrics*, 59: 1–23. https://doi.org/10.53482/2025_59_425
- McGloin, N., Endo Hudson, M., Nazikian, F. and Kakegawa, T. (2014). *Modern Japanese Grammar: A Practical Guide*. Routledge,

London.

- McLuhan, M. (1964). *Understanding Media: The Extensions of Man*. McGraw-Hill, New York.
- de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C.D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 4585–4592, European Language Resources Association (ELRA), Reykjavik, Iceland.
- Meyer, E. (2014). *The Culture Map: Decoding how people think, lead, and get things done across cultures*. PublicMatters, New York.
- Morden, T. (1999). Models of National Culture – A Management Review. *Cross Cultural Management*, 6(1):19–44.
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J. and Manning, C.D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics (ACL) System Demonstrations*. 2020. <https://arxiv.org/abs/2003.07082>
- Ross, J.R. (1982). Pronoun-deleting processes in German. Paper presented at the annual meeting of the Linguistic Society of America, San Diego, California.
- Rösch M. and Segler, K.G. (1987). Communication with the Japanese. *Management International Review*, 27: 56–67.
- Tsao, F. (1977). *A Functional Study of Topic in Chinese: The First Step toward Discourse Analysis* [Doctoral dissertation]. USC.
- ## 6. Language Resource References
- Davies, M. 2013. Corpus of Global Web-Based English. Available online at <https://www.english-corpora.org/glowbel/>.
- Dryer, M.S. (2013). Expression of Pronominal Subjects. In Dryer, M.S. & Haspelmath, M. (Eds.), *WALS Online (v2020.4)* [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.13950591>. Available online at <http://wals.info/chapter/101>
- Goldhahn, D., Eckart, Th. and Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 759–765. ELRA. Available online at <https://wortschatz.uni-leipzig.de/en/download/>.
- Kortmann, B., Lunkenheimer, K. and Ehret, K. (Eds.) 2020. *The Electronic World Atlas of Varieties of English*. Zenodo. DOI: 10.5281/zenodo.3712132. Available online at <http://ewave-atlas.org>.

Clefting and beyond in Tigrinya

Nazareth Amlesom Kifle, Michael Gasser

Østfold University College, Indiana University
nazareth.a.kifle@hiof.no, gasser@iu.edu

Abstract

This paper aims to give a preliminary analysis of focus marking constructions in Tigrinya within the Universal Dependency (UD) framework. We identify three types of constructions that we consider clefting. All three involve a copula placed after the focus element, the UD root, and a subordinate verb form representing the presupposed information. Tigrinya has three possibilities for this subordinate verb, which we annotate as *csubj*, *advel*, and *xcomp*. In each, we use the subrelation *cleft* to indicate the common structure and function of the different cleft types on a par with the use of subrelations *pass* for passive in different languages.

Keywords: Tigrinya, cleft construction, focus, copula, UD annotation, relative clause

1. Introduction

Cleft sentences are very common in Tigrinya (Palmer, 1962), a property the language shares with other members of the Ethio-Eritrean Semitic (EES) family (Appleyard, 1989; Fulass, 1972; Kapeliuk, 2005). Clefting is an information structure manipulating device that divides the proposition in the sentence into *presupposition* and *focus* elements (Lambrecht, 2001, 489). Syntactically speaking, the structure results when an element in the matrix clause, the *clefted constituent*, is extracted for emphasis and put in a prominent position with respect to the other elements of the clause, the *cleft clause*.

- (1) ሐሳብ እያ አስቲር
Hisab ḡiyy-a ḡaster
math IMPR.BE-3FS Aster
ዝምሃረት
zi-mihar-ət
REL-teach:PFV-s3FS

‘It is math that Aster taught.’

In this example, we see the typical features of cleft sentences: (1) the use of the copula adjacent to the clefted constituent (following it in this SOV language), and (2) a sentence-final subordinate clause, the cleft clause.

As we shall see, however, Tigrinya cleft sentences go beyond the type shown in (1), including variants characterized by other forms of subordination in the presupposition clause and a copula with multiple functions. These types are apparently not found in other languages in the EES family and, as far as we know, have not been studied systematically before. In this paper, we present a unified analysis of the different clefting constructions in the language and propose UD relations that reflect the syntactic differences and a subrelation to reflect the syntactic and functional similarities.

2. Overview of morphosyntax

In Tigrinya, the unmarked word order is SOV. However, the SOV order can vary for purposes of information structuring, for example, in cleft sentences. Tigrinya is a pro-drop language, a property which is apparent in cleft sentences, which lack the dummy pronoun of languages such as English. Finite verbs in Tigrinya feature obligatory subject agreement and object agreement for definite objects (Kifle, 2011, 2007). Subjects are unmarked for case, whereas objects are differentially marked through the prepositional case *ጉ- ni-* based on definiteness.

As in other Semitic languages, aspectual categories are distinguished by templates and by different sets of subject agreement affixes. Tigrinya has a present-tense copula, እያ ገiyyu, which agrees with its subject and may take nominals, adjectives, and adverbials as predicates.

The heads of subordinate clauses may be finite or non-finite. Relative clause verbs, which we shall refer to as *relativized verbs*, take finite agreement affixes and perfective or imperfective aspect. Relativized verbs are distinguished by a relativizing prefix which has two allomorphs, ዝ- *zi-* and እ- *ḡi-*. The allomorph እ-ገi- is used with some derived stems and with subject markers that employ the prefix ጉ- *ti-* or ጉ-*ni-* (Kifle, 2011, 38). Relative clauses precede their head nouns (2a).

- (2) a. እቲ አስቲር
ገit-i ḡaster
the-3MS Aster
እትምህሮ
ገi-tti-mihir-o
REL-s3FS-teach:IMPR-o3MS
ዓውደ እትምህሮቲ
ṣawde timhirti
subject

‘the subject that Aster teaches’

- b. አስቴር እትምህር
ʔaster ʔi-tti-mihir-o
 Aster REL-s3FS-teach:IMPR-o3MS
 ‘what/the thing that Aster teaches’

The relative marker is both a relativizer (2a) and, when it appears without a head nominal, a nominalizer (2b). In this example, the relativized verb bears agreement values for both the object and the subject and has the interpretation ‘the one (3MS) who/which’.

Another verb form relevant for this paper is the perfective form. Tigrinya has two perfective forms which Kifle (2011, 41) identifies as simple perfective and historic perfective. The simple perfective has two functions in the language, as a main verb and as a subordinate converb (3).

- (3) አስቴር ምሂራ ከይዳ
ʔaster mihir-a kayd-a
 Aster teach:PFV-s3FS go:PFV-s3FS
 ‘Aster left after having taught.’

The converb expresses an event simultaneous with or preceding that expressed by a main verb (Amha and Dimmendaal, 2006).

The simple perfective form is incompatible with the relative marker **zi-mihir-a*; for relativized verbs, the second form, *zi-mihar-ət* ‘that she taught’, the historic perfective is used.

Also relevant for our purposes is a further subordinate form, the “purposive,” consisting of a finite imperfective verb marked with the conjunctive prefix *h-* *ki-*, meaning ‘in order to V’ (4).

- (4) አስቴር ሕሳብ ክትምህር
aster Hisab ki-t-mihir
 Aster math PURP-s3FS-teach:IMPV
 ደልዖ
dəly-a
 want:PFV-s3FS
 ‘Aster wants to teach math.’

The purposive clause and the matrix clause share a subject, as it can be seen from the subject agreement, the 3FS, which is associated both with the main verb and purposive verb.

In combination with the copula, the purposive form expresses future tense (Kifle, 2011, 45), as in (5). The copula agrees with the subject of the purposive verb.

- (5) ጽባሕ አስቴር ሕሳብ
s'baH aster Hisab
 tomorrow Aster math
 ክትምህር እዩ
ki-t-mihir ʔiyy-a
 PURP-s3FS-teach:IMPV IMPR.BE-3FS

‘Tomorrow Aster will teach math.’/ Lit. Aster is to teach math.

3. Clefting in Tigrinya

3.1. Clefting with relativized verbs

In cross-linguistic perspective, the syntactic configuration of a cleft sentence can be described as a complex sentence that has the form of a specificational sentence with a subordinate clause, the cleft clause, and a focused element, the clefted constituent, in a subject-predicate-like relation (Malcher, 2023). According to Creissels (2021, 7), if a language has an equative predication construction and a participant nominalization construction, then it is always possible to combine these structures “to make explicit the exclusive identification of a participant in a presupposed event.”

Tigrinya has cleft constructions which contrast with copula predication sentences in the way different elements are focused and linked by a copula to a presupposition (Gragg, 1974; Kifle, 2011, 64) (6).

- (6) አስቴር እዩ ሕሳብ
ʔaster ʔiyy-a Hisab
 Aster IMPR.BE-3FS math
 ዝምህረት
zi-mihar-ət
 REL-teach:PFV-s3FS
 ‘It is Aster who taught math.’

In (6), we have a relativized verb form, and a copula which agrees with this verb—and hence its subject—in person, number, and gender. The clefted constituent appears right before the copula and the cleft clause follows it. This is the most common clefting type. In the TLMD corpus (Gaim et al., 2021), 141 sentences of this type are found within the first 10,000 sentences.¹ We will call such sentences *canonical cleft sentences*.

¹We applied the morphological analyzer HornMorpho (Gasser, 2025) to the unannotated TLMD corpus, permitting us to isolate copulas and the relevant subordinate verb categories and thus count instances of the cleft types.

It is also possible for the relativized cleft clause to appear sentence-initially, though this is less common.

- (7) ሕሳብ ዝምሃረት አስቲር
Hisab zi-mihar-at ?aster
 math REL-teach:PFV-s3FS Aster
 እያ
 ?iyy-a
 IMPR.BE-3FS

‘It is Aster who taught math.’ / ‘The one who taught math is Aster.’

In Tigrinya, any lexical constituent can be focused in clefted sentences. In the corpus data we find examples of subjects, objects, obliques, adjuncts, and adverbial modifiers. Sentence (8a) illustrates this for an oblique constituent and sentence (8b) for an adverbial modifier.

- (8) a. ትማሊ እያ አስቲር
timali ?iyy-a ?aster
 yesterday IMPR.BE-3FS/3MS Aster
 ሕሳብ ዝምሃረት
Hisab zi-mihar-at
 math REL-teach:PFV-s3FS

‘It is yesterday that Aster taught math.’

- b. ብወኒ እያ
bi-wəni ?iyy-a
 with-enthusiasm IMPR.BE-3FS/3MS
 አስቲር ዝምሃረት
?aster zi-mihar-at
 Aster REL-teach:PFV-s3FS

‘It is with enthusiasm that Aster taught.’

3.2. Clefting without relativized verbs

Tigrinya also has focus constructions that resemble the canonical clefted sentence in which a copula is preceded by the clefted constituent and followed by non-relativized verb forms. We will refer to these as *non-relative cleft sentences*. Such cleft structures resemble the types Khan (2019, 4) identified as reduced clefts, where the relativizer that marks the subordinate clause or the copula is omitted.

The first type, which we will refer to as *converb cleft* (9), has been identified as a cleft sentence in some studies (Zellou, 2010; Bulakh, 2019; Gebregziabher and Duguine, 2024). There are 102 instances of converb cleft sentences in the first 10,000 sentences of the TLMD corpus.

- (9) ሕሳብ እያ አስቲር ምሃራ
Hisab ?iyy-a ?aster mihir-a
 math IMPR.BE-3FS Aster teach:PFV-s3FS
 ‘It is math that Aster taught.’

The converb cleft involves a copula whose function is to mark focus (Zellou, 2010, 112), as in canonical cleft sentences, and a perfective form of the verb which is identified as a converb (Bulakh, 2019, 186), though, as we have seen, this form of the verb is ambiguous and can be used as a main clause perfective verb (Kifle, 2011, 44) as well (10).

- (10) አስቲር ሕሳብ ምሃራ
?aster Hisab mihir-a
 math IMPR.BE-3FS Aster
 ‘Aster taught math.’

The converb cleft seems to be related to sentences in which a copula is introduced to add emphasis to the whole proposition expressed by the main verb (Kifle, 2011, 258). Cacchioli (2024, 25) calls this kind of emphasis “verum focus”.

- (11) a. አስቲር ሕሳብ ምሃራ
?aster Hisab mihir-a
 Aster math teach:PFV-s3FS
 እያ
 ?iyy-a
 IMPR.BE-3FS

‘Aster DID teach math.’

- b. አስቲር ሕሳብ ትምህር
aster Hisab ti-mihir
 Aster math s3FS-teach:IMPV
 እያ
 ?iyy-a
 IMPR.BE-3FS

‘Aster DOES teach math.’

Preceding the copula, both perfective (11a) and imperfective (11b) verb forms are possible. However, interestingly, only the perfective form is possible in the converb cleft (12).

- (12) እኔ እያ ሕሳብ
?ana ?iyy-a Hisab
 nom.Pro.1s IMPR.BE-1S math
 *እምህር
 *?i-mihir
 s1S-teach:IMPV

‘It is me *teach math.’

There is a syntactic property that distinguishes the converb clause from the relative clause in cleft sentences. Although both are allowed following the copula, as in (13a), only the relative clause is allowed in sentence initial position (13b). That is, unlike the relativized verb, the converb cannot function as a nominalization.

- (13) a. አስቴር እያ ሕሳብ
 Aster ?iyy-a Hisab
 Aster IMPR.BE-3FS math
 ዝምሃረ/ምሃራ
 zi-mihar-at/mihir-a
 teach:REL/PFV-s3FS/PFV-s3FS
 ‘It is Aster who taught math.’
- b. ሕሳብ ዝምሃረት/ *ምሃራ
 Hisab zi-mihar-at/*mihir-a
 math teach:REL/PFV.s3FS/PFV-s3FS
 አስቴር እያ
 Aster ?iyy-a
 Aster IMPR.BE-3FS
 ‘The one who taught math is Aster.’

The second type of non-relative cleft is possible with the purposive form expressing future, which was discussed above (5). When this form appears in cleft sentences, the copula, which is already a component of the future tense construction, is not duplicated, and the verb appears in its non-relativized purposive form. We shall refer to such sentences as *future cleft sentences*. There are 10 instances of future cleft sentences among the first 10,000 sentences of the TLMD corpus.

- (14) ጽባሕ እያ
 s’baH ?iyy-a
 tomorrow IMPR.BE-3FS
 ክትምህር
 ki-t-mihir
 PURP-s3FS-teach:IMPV
 ‘It is tomorrow that she will teach.’

4. UD annotation

Cleft sentences, in languages that have them, consist of a nonverbal clause, containing the clefted constituent, and a subordinate verb, the cleft clause, following the nonverbal clause. As in other nonverbal clauses, the root in UD annotation is the clefted constituent (rather than the copula) (de Marneffe et al., 2021, 273), and a dependency joins the root to the cleft clause. The relation on this dependency varies with the language; there are at least seven possibilities, and in at least one

language, Polish, two different relations are used for the cleft clause dependency. In 14 languages, the dependency has *cleft* as a subrelation. For English, the relation is *advcl:relcl*, as illustrated in Figure 1.

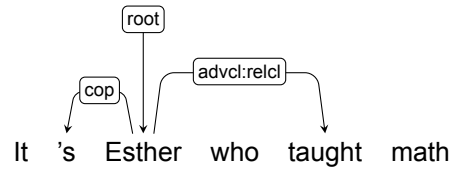


Figure 1: Annotation of English cleft sentences.

For Tigrinya, we have seen that there are three sentence types that, on functional and syntactic grounds, can be called *cleft*. Although the clefted clause is subordinate in each type, morphosyntactically its verb takes different forms, motivating the use of more than one relation for these clauses. However, we would like at the same time to indicate that these variants share the same focusing function, that all of the cleft sentences can be seen as related to non-cleft sentences that are simpler syntactically and/or have canonical word order. We propose to do this using the subrelation *cleft*, as in other languages such as Polish, French, and Irish.

Let’s consider first *canonical cleft sentences*, in which the clefted clause has a relativized verb. As we have seen, relativized verbs in the language function as relative clauses, but these can appear without head nouns, in which case the verb functions as a nominalization, as in Example (2b). In this form it takes a *csubj* relation because the cleft clause syntactically functions as a clausal subject linked to the cleft element through the copula similar to copula predication sentences.

The less common variant of the canonical cleft sentence has the relative clause in the usual subject position.

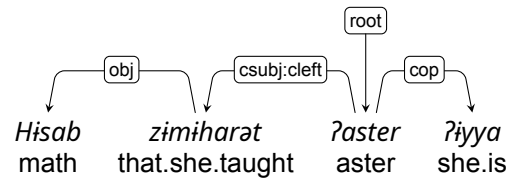


Figure 2: Dependencies in (7), ‘it is Aster who taught math’.

The more frequent pattern for the *canonical cleft sentences* differs only in word order; the clefted clause is sentence final.

Now consider the *converb cleft* pattern. Here the sentence-final constituent, rather than being a relativized verb, takes the form of a converb. Else-

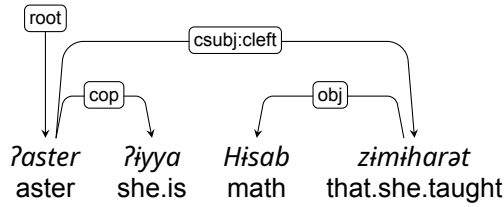


Figure 3: Dependencies in (6), ‘it is Aster who taught math’.

where we annotate converbs as adverbial clauses (Figure 4).

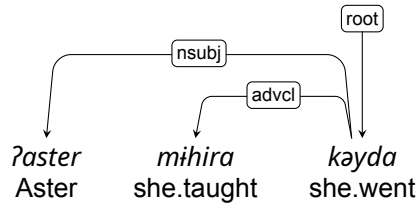


Figure 4: Dependencies in (3), ‘Aster left, after having taught’.

For *converb cleft* sentences, we again make use of the *advcl* relation but add the *cleft* subrelation, as before, to indicate that the sentence is related to a non-cleft sentence (Figure 5). Note that *csubj* would be inappropriate for the *converb cleft* because, as we have seen (13b), the converb lacks the nominalizing function of relativized verbs.

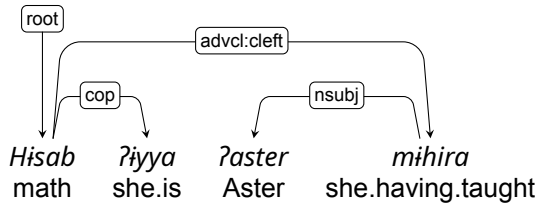


Figure 5: Dependencies in (9), ‘it is math that Aster taught’.

Finally, consider *future cleft sentences*. Here the clefted clause contains a subordinate verb with a purposive prefix conjunction. We analyze these sentences in terms of their relationship to unclefted sentences with the purposive future, sentences in which the purposive verb is the root. In this case, clefting simply moves the purposive verb form, possibly along with one or more arguments, after the copula, leaving the focused element immediately before the copula.

To determine how to annotate such sentences, we again look at how the subordinate verb form is used in non-clefted sentences. Besides its use in the purposive future, these purposive verbs function as *xcomp* arguments of verbs such as **ጸለጽ**

ḍalaya ‘want’ (4). Consequently we annotate the verb in the cleft clause in future cleft sentences with *xcomp:cleft* (Figure 6).

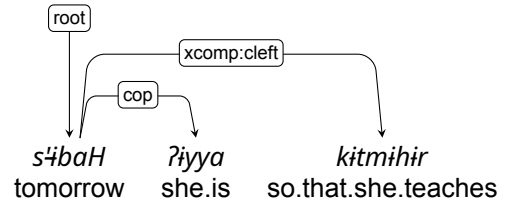


Figure 6: Dependencies in (14), ‘it is tomorrow that she will teach’.

5. Conclusion

In this preliminary study, we have described various types of focus marking constructions in Tigrinya and shown how such constructions can be annotated in UD in a unified manner. Focus marking through clefting is claimed to be widespread in the language, and any type of constituent can be clefted. We have identified canonical clefting types, where the verb in the cleft clause is relativized, and non-relative clefting, where the verb in the cleft clause is either a converb/perfective or a purposive verb form. We argue that the non-relative cleft types are instances of focus marking on a par with canonical cleft constructions. In both types, the copula is the main feature enabling focus and there is some sort of dependent verb, a converb or purposive verb, in the residual part of the sentence.

As with the Polish treebank, we propose different relations for the cleft categories: *csubj*, *advcl*, and *xcomp* for the canonical cleft, converb cleft and purposive future, respectively, and assign the *cleft* subrelation to each to reflect the syntactic and functional similarities. For the canonical clefts, we opt for *csubj:cleft*, rather than *acl:cleft* (as in Swedish), because in Tigrinya the relative verb is a nominalized form.

As part of our study, we would like also to compare focus marking in predication sentences as these also use word order variation to create focus by placing the focus element before the copula and the presupposed constituent after it. In the perfective, predication focus sentences may involve copula doubling where the focusing copula is in the present while the predication copula is in the perfective aspect. We leave these issues for future research.

References

- Azeb Amha and Gerrit J Dimmendaal. 2006. Converbs in an African perspective. *Catching language: The standing challenge of grammar writing*, pages 393–440.
- David L Appleyard. 1989. The relative verb in focus constructions: an Ethiopian areal feature. *Journal of Semitic Studies*, 34(2):291–305.
- Maria Bulakh. 2019. Tigrinya 1. In *The Semitic Languages*, pages 174–201. Routledge.
- Gioia Cacchioli. 2024. A note on Verum Focus in Tigrinya. In *Language Use and Linguistic Structure. Proceedings of the Olomouc Linguistic Colloquium 2023*, volume 10, page 24. Palacký University Olomouc.
- Denis Creissels. 2021. Remarks on the grammaticalization of identificational clefts. *Faits de Langues*, 52(1):13–31.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*.
- Hailu Fulass. 1972. [On Amharic relative clauses](#). *Bulletin of the School of Oriental and African Studies, University of London*, 35(3):497–513.
- Fitsum Gaim, Wonsuk Yang, and Jong C. Park. 2021. [Tlmd: Tigrinya language modeling dataset](#).
- Michael Gasser. 2025. [HornMorpho](#).
- Keffyalew Gebregziabher and Maia Duguine. 2024. Typology of Tigrinya WH-interrogatives. *ACAL in SoCAL*, page 263.
- Gene B Gragg. 1974. Cleft sentences in Tigrinya. *Journal of African Languages*, 11(3):74–89.
- Olga Kapeliuk. 2005. [Topicalization in amharic and its degrees](#). *Scrinium*, 1(1):354 – 366.
- Geoffrey Khan. 2019. Copulas, cleft sentences and focus markers in biblical hebrew. In *Ancient Texts and Modern Readers*, pages 14–62. Brill.
- Nazareth Amlesom Kifle. 2007. Differential object marking and topicality in Tigrinya. In *Proceedings of the LFG*, volume 7, pages 5–25.
- Nazareth Amleson Kifle. 2011. *Tigrinya Applicatives in Lexical Functional Grammar*. Ph.D. thesis, University of Bergen.
- Knud Lambrecht. 2001. A framework for the analysis of cleft constructions. *Linguistics*, 39(3).
- Kurt Malcher. 2023. *Clefts. A cross-linguistic investigation*. Ph.D. thesis, Universität zu Köln.
- Frank Robert Palmer. 1962. [Relative clauses in Tigrinya](#). *Journal of Semitic Studies*, 7(1):36–43.
- Georgia Zellou. 2010. Tigrinya fronted copula constructions: Focus and evidence of speaker marking information relevant to hearer expectations. *of: Rice Working Papers in Linguistics*, 2:106–21.

Coconstructions in Spoken Data: UD Annotation Guidelines and First Results

Ludovica Pannitto¹, Sylvain Kahane², Kaja Dobrovoljc³,
Elena Battaglia¹, Bruno Guillaume⁴, Caterina Mauri¹, Eleonora Zucchini⁵

¹ University of Bologna, Bologna - Italy;

² Université Paris Nanterre, Modyco, Paris - France;

³ University of Ljubljana and Jozef Stefan Institute, Ljubljana - Slovenia;

⁴ Université de Lorraine, CNRS, Inria, LORIA, Nancy - France;

⁵ Masaryk University, Brno - Czech Republic

Abstract

The paper proposes annotation guidelines for syntactic dependencies that span across speaker turns — including collaborative coconstructions proper, wh-question answers, and backchannels — in spoken language treebanks within the Universal Dependencies framework. Two representations are proposed: a speaker-based representation following the segmentation into speech turns, and a dependency-based representation with dependencies across speech turns. New propositions are also put forward to distinguish between reformulations and repairs, and to promote elements in unfinished phrases.

Keywords: spoken language, coconstruction, backchannel, reformulation, syntactic treebank, French, Italian, Slovenian

1. Introduction

Spoken language is intrinsically temporal, incremental, and interactive (e.g., Hopper 1992; Auer 2024; Deppermann and Günthner 2015): unlike written text, it unfolds in real time and is jointly produced by multiple participants cooperating in the construction of syntactic structure. Linearized transcriptions, however, necessarily flatten this temporal and multi-party dimension.

Universal Dependencies (UD, de Marneffe et al. 2021), despite its success as a cross-linguistic annotation framework, inherits a broadly shared assumption in syntactic theory: that syntax operates within the boundaries of a sentence. While this assumption is largely unproblematic for written language (although sentence boundaries detection is not yet considered a completely solved task, e.g. Rudrapal et al. 2015; Read et al. 2012; Sprugnoli and Redaelli 2025), it raises difficulties for spoken interaction, where segmentation into sentence-like units is neither trivial nor theoretically neutral.

Two additional issues complicate the picture. First, segmenting the speech flow is not equivalent to segmenting speaker turns: different segmentation strategies may be adopted depending on whether the focus is on individual speakers' production or on jointly constructed syntactic units. Second, syntax can be viewed both as a cognitive mechanism for building grammatical structures and as an interactional resource for coordinating meaning between speakers. These perspectives stem from different research traditions, yet an an-

notation framework such as UD should not commit exclusively to either one.

We introduce a novel proposal for annotation guidelines that can enable resource creators and users to account for syntactic dependencies holding among speech produced by different speakers. In Section 2, we argue that spoken data require an explicit treatment of syntactic relations that cross turn boundaries. We refer to these relations as *coconstructions*. We provide a typology of such constructions in Section 3, and we propose annotation guidelines that preserve information from both speaker-based and dependency-based perspectives, while remaining fully compatible with the UD formalism, in Section 4. Section 5 presents the conversion between the speaker-based and the dependency-based representation. A quick overview of the implementation of our annotation scheme in three treebanks from French, Italian, and Slovenian is provided in Section 6.

2. The Units of Spoken UD Treebanks

No clear agreement emerges in the literature about what a *sentence* (as in, maximal unit of segmentation) in spoken language is (Sacks et al., 1974; Pietrandrea et al., 2014; Mettouchi and Vanhove, 2020) and, despite sentences being the main building blocks of UD resources, no explicit definition of what should count as a sentence is ever given in the current annotation guidelines.

Syntactic annotation frameworks typically operate on segmentation units that resemble sentences but can vary greatly in their definitions. This mis-

match becomes particularly evident when syntactic dependencies span across speaker turns, overlap with other speakers' contributions, or are resumed after interruptions. Pietrandrea et al. (2014) recall three commonly applied criteria employed for the definition of sentences, namely the locutionary criterion, that attributes sentences to the locutionary responsibility of a given speaker; the graphic or prosodic criterion, that defines sentences based on punctuation markers or major prosodic breaks; and the syntactic criterion, that focuses on sentences as maximal structurally autonomous units, linked to the surrounding context by discursive relations.

While for standard written data the three criteria are likely to converge to the same analysis, this is not the case for spoken language and other varieties such as user-generated content (Sanguinetti et al., 2023). Earlier attempts at building spoken treebanks annotated for constituent syntax have implemented a turn-based approach. For instance, in the Penn Treebank-style annotation of the Switchboard corpus, turns by different speakers are parsed as standalone trees (Bies et al., 1995); in the extended Switchboard Dialog Act Corpus - SwDA (Jurafsky et al., 1997), by representing a discourse layer on top of a syntactic one, only some forms of co-construction, for instance collaborative completions, receive a tag. Another example is the Verbmobil corpus, where the primary domain of syntactic analysis and annotation is "the dialog turn, (...) an uninterrupted contribution by one dialog participant" (Hinrichs et al., 2000, 108); while this approach can account for some phenomena of spoken language such as disfluencies (see Meteer et al. 1995), it privileges single-speaker syntax and obscures collaborative syntax, which cannot be properly searched within trees. Dependency frameworks are, as a matter of fact, more apt to represent relations holding across turn boundaries. However, among existing spoken language treebanks in UD, no uniform segmentation criterion is applied, probably as a result of pre-existing design choices in the source data (Dobrovoljc, 2022).

We argue that two different units of analysis need to be considered in the case of spoken data:

- **speaker-dependent maximal units**, loosely related to the notion of *turn* and that anchor segmentation to the speech production of each speaker participating in the conversation, as it is usually done;
- **structurally-autonomous maximal units**, based on the notion of government (Tesnière, 1959, 2015; Blanche-Benveniste, 1987), defined by Kahane and Pietrandrea (2012) as the property by which words impose constraints on other words, specifically constraints on (i) their syntactic category, (ii) their morphologi-

cal or syntactic marking, and (iii) their linear position.

An alternative segmentation strategy, adopted for instance in Rhapsodie, is based on the notion of illocutionary unit, defined through prosodic and pragmatic criteria. Such units capture the organization of speech in terms of illocutionary completeness and prosodic packaging, and therefore reflect an important dimension of spoken interaction. However, because they are grounded in prosodic-pragmatic segmentation, they do not necessarily align with domains of syntactic dependency. As a result, using illocutionary units as primary units for syntactic annotation would make the representation of syntax dependent on a specific analysis of prosodic and illocutionary organization, which is not always available, consistently applicable, or comparable across corpora and languages. By contrast, speaker-dependent maximal units provide a more neutral and operational point of departure: they are directly anchored in the observable distribution of speech across participants, preserve information about production and turn organization, and can be systematically enriched to recover syntactic dependencies that extend across unit boundaries, without conflating speaker organization, prosodic packaging, and syntactic structure into a single segmentation criterion.

In the remainder of this paper, we refer to speaker-dependent maximal units as *maximal units* and to structurally-autonomous maximal units as *rectional units*. Currently, spoken treebanks in UD are organized in speaker-dependent units, and only Rhapsodie (see Section 6.1) and KIParla Forest (see Section 6.2) provide special features for cases of coconstructions.

3. A Typology of Coconstructions

We consider as *coconstructions* those syntactic relations (without restrictions on dependency type) that hold between utterances produced (i) by different speakers or (ii) by the same speaker after an interruption by another participant. In interactive scenarios, speakers may jointly contribute to the realization of a single syntactic structure, sometimes in a tightly coordinated fashion and sometimes with partial or even complete overlap (Lerner, 1991; Ono and Thompson, 1996; Sacks, 1992; Helasvuoto, 2004; Calabria, 2026). From a dependency-based perspective, this implies that syntactic relations cannot always be confined to a single speaker turn.

Our notion of coconstruction is grounded in a dependency-oriented view of syntax and is compatible with the concept of rectional unit (*unité rectionnelle*) introduced by Benzitoun et al. (2010); Kahane and Pietrandrea (2012), described in Section 2: this is defined independently of speaker

turns or prosodic segmentation and it captures syntactic governance relations wherever they occur in the speech flow. This makes the notion particularly well suited to spoken interaction, where the realization of a single rectional domain may be distributed across multiple speakers or interrupted by other contributions. Coconstructions can thus be understood as cases in which a single rectional unit is jointly realized across turns, by one or more speakers.

On this basis, we distinguish three major types of coconstructions, reflecting different interactional and syntactic configurations: *co-construction proper*, *wh-question constructions*, and *backchanneling*.

3.1. Coconstruction Proper

Coconstruction proper covers cases in which multiple contributions target the same dependency structure by completing, duplicating, or extending dependencies. Such contributions can be produced (i) by different speakers or (ii) by the same speaker after an interruption by another speaker, i.e. in non-contiguous utterances, which formally results in the same pattern as (i)¹. In the remainder of the article we will refer to Speaker 1 and Speaker 2 as the producers of the first and second unit which are focus of the discussion. Here and after, in the event of an interruption, Speaker 2 coincides with Speaker 1. All examples are reproduced preserving the original transcription conventions of the corpus.

The key distinction within coconstructions proper concerns whether Speaker 2 realizes a dependency that is already projected by Speaker 1 (case A), or whether Speaker 2 adds a new dependency to the structure initiated by Speaker 1 (case B).

A. Speaker 2 realizes an already projected dependency. When Speaker 2 realizes a dependency projected by Speaker 1, we distinguish three subcases reflecting the status of that dependency at the moment of realization: unrealized, in progress, or already realized.

A.1 Completion: Speaker 1 leaves a syntactic dependency unrealized, and Speaker 2 provides the missing material, *completing* the structure initiated by Speaker 1, as in Example (1);

- (1) English (Ono and Thompson 1996, p. 72)
L: his position is **pretty uh**
A: ... **stable**.

¹We do not treat the case of contiguous maximal units produced by the same speaker, as we hypothesize that these should be merged in a single unit in case they can be linked through syntactic dependency, thus yielding a single maximal rectional unit.

A.2 Parallel realization: The same syntactic dependency is realized at the same time by both speakers. Speaker 1 projects and eventually completes the structure, often after hesitation or a false start, while Speaker 2 simultaneously or near-simultaneously fills the same syntactic slot, as in Example (2), see also Figure 3 in Appendix A. This configuration frequently results in overlap;

- (2) Italian (KIParla², BOD2018)
BO140: mh: vabbè [ci **stiamo**] **insie:me:**
well we stay together there
esatto: c'è il divano comodo,
right there's a comfortable sofa
BO118: [**state insie:me:**]
stay together

A.3 Subsequent realization: Speaker 1 completes their own syntactic structure, and Speaker 2 subsequently selects a dependency in Speaker 1's structure and realizes it again, either by repeating or replacing material in the same syntactic slot (as in Examples (3) and (4)).

- (3) French (Rhapsodie, D2001)³
\$L2 "eh bien" je crois que je ne me suis pas
well I think I haven't
conduit d'une façon conforme à ce qu'on
behaved as one would
attend "euh" { { **d'une jeune fille d'abord**
expect from a young girl first
| ^ **et d'une femme ensuite** } } //+
and from a woman afterwards
\$L1 { | **d'une jeune bourgeoise** } ?//+
from a young bourgeois girl?
\$L2 { | **"disons" d'une jeune bourgeoise** } //
let's say from a young bourgeois girl
(4) Italian (KIParla, KPS021)
PKP126: quindi l- la linea **tra finzione e realtà**
so the line between fiction and reality
cioè tra verità non verità
that is between true and false
non ho ancora capito [dove sta]
I still haven't understood where it lies
PKP125: [**più tra de]tto e non detto**
more between said and unsaid

²(Mauri et al., 2019). In all examples from KIParla, square brackets indicate overlaps between speakers' productions.

³Curly brackets indicate conjuncts in a stacking construction. See Section 4.2. The sign // indicates the end of an illocutionary unit. The sign //+ indicated that there is a link to the next unit.

B. Speaker 2 adds a new dependency. Speaker 2 realizes a syntactic dependency that is *not* already projected by Speaker 1, thereby *extending* the syntactic structure initiated by Speaker 1 (Examples (5), (6), (9))

- (5) English (Ono and Thompson 1996, p. 81)

M: they must know each other.

H: ... **very well**.

- (6) Italian (KIParla, BOA3017)

BO139: quando si parlano sopra
when they speak one over the other
 devi mettere delle parentesi quadre
you have to put square parentheses
 intorno **alle parole**
around words

BO146: ah okay

oh okay

BO139: **che si sovrappongono**

that overlap

BO136: **come nei sottotitoli**

like in subtitles

3.2. Wh-questions

Answers to *wh*-questions are analysed as syntactically dependent on the question, and thus as part of the same rectional unit, as already proposed in the guidelines of Rhapsodie.

- (7) Italian (KIParla, PTA007)

TOR001: **dove** vai ad arrampicare?

where do you climb?

TOI007: **al bi side**⁴ vicino alla colletta

at the bi side close to colletta

We therefore consider answers to *wh*-questions as cases of coconstructions, insofar as they realize syntactic material licensed by the interrogative structure of the preceding turn (see an example in Figure 6). By contrast, we do not treat answers to polar questions as coconstructions, since their response items do not depend syntactically on any element of the question nor fill a pre-existing syntactic slot (ex., cf. Example (7) with A: *do you climb at the bi side?* B: *yes*).

3.3. Backchannelling

Backchannels are short productions uttered by one participant while another speaker holds the floor. They may be verbal or paraverbal (e.g. *mhm* as in Example (8), *yes* as in Example (9), *right*) and

mainly serve to signal attention or alignment (Ward and Tsukahara, 2000; Mereu et al., 2024; Furkó, 2020; Khan, 2023). At their core, these verbal reactions show that the speaker has heard the contribution of the partner, often adding that it has been understood and accepted.

To be recognized as backchannels, such utterances (i) must be addressed to the content of the other speaker's contribution; (ii) must not be required or expected by the preceding turn (e.g. answers to *wh*- and polar questions are expected and required, so they cannot be considered backchannels, see (7)); (iii) must not require a reaction from the main speaker.

- (8) Italian (KIParla, BOA3017)

BO145: ma perché mamma c'ha dei

but because my mom has
 pregiudizi nei miei confronti
prejudices against me

BO139: **mhmh**

mhmh

BO145: e poi daniela non devi avere

and then daniela you shouldn't have
 pregiudizi su di me
prejudice against me

- (9) Italian, (KIParla, BOD2018)

BO118: sì sì ma anch'io però era proprio

yeah yeah same here but I really just
 l'esigenza di stare da sola
needed to be alone

BO140: **sì** di stare da sola di non

yeah to be alone not
 parlare
to talk

Backchannels differ from co-construction proper and from *wh*-questions coconstructions, in that they do not contribute syntactic material to the rectional unit projected by the main speaker. Nevertheless, they are tightly integrated into the temporal unfolding of speech and must be accounted for in the syntactic analysis of spoken interaction. Jucker and Smith (1996) argue that the distinction between a "discourse marker" and a "backchannel" is often more about conversational role than linguistic function and both belong to a broader category of elements used to negotiate "common ground".

We are aware that the syntactic status of backchannels is not uncontroversial. What counts as "syntax", however, is itself a matter of theoretical tradition: different frameworks draw the boundary between syntax, pragmatics, and interaction differently, and UD itself already encodes relations — such as *discourse* — that would not be considered syntactic under a narrow structural definition. More importantly, even if a backchannel is not itself

⁴'bi side' is the quasi-phonetic rendering of the name of a climbing gym, 'Bside'.

a syntactic dependent, it actively shapes the syntactic structure being built by the main speaker: it may prompt continuation, signal that repair or elaboration is needed, or mark the boundaries within which the syntactic unit unfolds. Ignoring backchannels in the annotation of spoken interaction thus risks misrepresenting the conditions under which syntactic structure is produced and interpreted. Including them in the `MISC` field — without altering any dependency relation — is a minimal and non-disruptive way of making this interactional layer visible to downstream users of the treebank.

4. Proposed Guidelines

To explicitly represent coconstructed configurations in UD annotation, we introduce two features encoded in the `MISC` field: `Coconstruct` and `Backchannel`. These features do not replace dependency relations, but provide a lightweight mechanism to signal linkage between speaker-based maximal units involved in coconstruction phenomena.

All features introduced in this paper are encoded in the `MISC` field, keeping the proposal fully backward-compatible: the core dependency structure, POS tags, morphological features, and tokenization remain untouched. The `MISC` field is already used across treebanks for typographic information (`SpaceAfter=No`), transliteration, named entity spans, and discourse-level features; our proposal follows this established practice.

As a general principle, both `Coconstruct` and `Backchannel` are always assigned to the element that comes second in time: the feature is assigned to the responding token and its value identifies the eliciting or projecting token in the earlier utterance. This also enables multiple later tokens to point to the same earlier anchor.

Formally, `Coconstruct` encodes a backward pointer of the form `<deprel>::<sent_id>::<tok_id>`, while `Backchannel`, which already uniquely identifies a dependency relation (i.e., `discourse:backchannel`), encodes a backward pointer of the form `<sent_id>::<tok_id>` (see Figure 1).

The `Coconstruct` feature is used in two main configurations, which differ in the syntactic status of the dependency slot targeted by the responding token:

Completion of an open dependency slot: the responding token realizes a dependency that was projected but not realized in the preceding unit.

Stacking on an already realized dependency slot: the responding token provides an alternative

realization of a dependency that is already present in the preceding unit.

4.1. Completion of an Open Dependency Slot

A configuration qualifies as completion of an open dependency if all of the following conditions are met:

- Speaker 1 produces a token that syntactically licenses a dependent (although not necessarily requiring it), but the dependent is not realized before Speaker 2’s intervention.
- Speaker 1 does not complete the same dependency slot before or concurrently with Speaker 2’s contribution.

When these requirements are met, the head of the coconstructed dependency is the token in Speaker 1’s utterance that projects the unrealized dependency. Head identification follows standard UD principles: the head is the lexical item that selects the dependent.

This configuration applies both to cases A.1 and B described in Section 3. In cases of A.1 Speaker 1 might produce an element that projects a dependency but remains syntactically unfinished, because its canonical dependent is not realized before turn transfer. This situation is particularly frequent with adpositions (more specifically, prepositions), complementizers, and similar selecting elements. From a UD perspective, such cases raise a problem: the projected dependent is missing, but dependency annotation cannot be postponed or left underspecified. To address this, we adopt the promotion mechanism used in UD for ellipsis. When a projected dependency remains unrealized in Speaker 1’s utterance, the unfinished element is promoted to fill the syntactic position that would normally be occupied by its missing complement. This allows the dependency structure to remain well-formed while preserving information about the element’s incomplete status.

We introduce two additional `MISC` features to encode this mechanism: `Scrap=Yes` to mark any element (or subtree) whose syntactic projection remains unfinished and `Promotion=<deprel>` to indicate the dependency relation the element would have had before promotion, i.e. the relation it would bear to its governor if it were present.

- (10) Italian, (KIParla, BOA3017)
 BO146: ((ride)) vabbè te ti ci ti ci devo **fare**
((laughs)) well you I have to do you
pe::r
by
 BO139: ((ride)) **mestiere**
((laughs)) trade

For instance, in Example (10), a noun is missing after the preposition *per*. The oblique prepositional phrase is unfinished, so the *obl* dependency relation has to go on the preposition. The preposition, which should have been the *case* of the missing noun, has then been promoted. Since this promotion follows a *case* relation, the preposition receives the feature *Promotion=case*, as well as a feature *Scrap=Yes* indicating that the phrase it heads is unfinished (see first line in Figure 2).

Unfinished syntactic structures may arise in spoken interaction independently of whether co-construction is eventually realized. The use of *Scrap=Yes* therefore does not presuppose that the missing material will be supplied later: it encodes structural incompleteness as such, regardless of subsequent interactional developments. *Promotion* is essentially used in cooccurrence with *Scrap=Yes* when there is an unfinished phrase without its expected head.⁵ As we show in Section 5, the feature *Promotion* plays an important role in the conversion to the dependency-based representation.

The promotion of an element to a given position introduces an additional problem, because a dependent of a promoted word can be either a true dependent of the word itself or of its missing head. We mark this by means of a feature *Head*, with possible values *Word* and *Position*, to solve this problem. In Example (11), *nice* will be promoted as the *obj* of *has* and it receives a feature *Promotion=amod*, *a* is *det* of *nice* with a feature *Head=Position*, and *very* is *advmod* of *nice* with a feature *Head=Word*. In appendix, the full annotation of Example (11) is given (Figure 4) together with a Slovenian example (Figure 5) taken from the corpus.

- (11) A: she has a very nice
B: attitude

4.2. Stacking on an Already Realized Dependency Slot

A configuration qualifies as stacking on an already realized dependency slot if all of the following conditions are met:

- Speaker 1 realizes a complete dependency structure, including the dependent that fills the relevant syntactic slot;

⁵*Promotion* could, for instance, be introduced in case of gapping, as in (i) where *department* is promoted as root of the second conjunct.

(i) Mary called the police and Peter the fire department.

The feature *Promotion=obj* could be an elegant way to indicate that *department* is *obj* of an ellipsed verb.

- Speaker 2 produces material that is syntactically comparable to the dependent already realized by Speaker 1 and could occupy the same dependency position.

In stacking configurations, Speaker 2 does not complete an open dependency, but rather produces material that targets the same syntactic slot as an existing dependent. From a production perspective, Speaker 2's contribution is not structurally autonomous: its interpretation requires reference to Speaker 1's syntactic structure.

Stacking can however be instantiated by various deprels.

The most straightforward case is that of simple repetitions or repairs without any communicative intention, for which we recommend the *reparandum* relation. These are cases such as *des, des* or *ce dont, ce dont* in Example (13), or cases where a word is unfinished as *il sotto~* 'the sub~' in Example (12) (see also Figure 7): the first speaker has difficulties to produce the word *sottotitolatore* 'subtitled', which is then correctly produced by the second speaker. *reparandum* is a right-headed relation in UD: in order to mark a coconstruction on the latter item of the relation, we introduce the special deprel *repair* which is converted to *reparandum* in dependency-view.

- (12) Italian (KIParla, BOA3017)
BO147: infatti te potresti fare **il sotto~**
indeed, you could be the subti~
quello che fa i sottotitoli
the one that does subtitles
BO145: **sottotitolatore**
subtitled

If instead the first conjunct is complete and the second conjunct is different, we consider that we have a reformulation. Example (4) is a typical case: The first speaker already reformulates *tra finzione e realtà* 'between fiction and reality' by *tra verità non verità* 'between true and false' and then the second speaker proposes a second reformulation *più tra detto e non detto* 'more between said and unsaid'.

We propose to introduce a new relation for reformulation, which we call *conj:reform*. Contrary to coordinated conjuncts, where conjuncts have different referents, in a reformulation, the conjuncts are different denotations of the same referent (Kahane and Pietrandrea, 2012). We extend the relation *conj:reform* to other cases where two conjuncts target the same referent.

The first case is confirmation. In Example (3), speaker \$L2 produces *d'une jeune fille d'abord et d'une femme ensuite* 'from a young girl first and from a woman afterwards' and \$L1 proposes the reformulation *d'une jeune bourgeoise* 'from a young

bourgeois girl' (with an interrogative prosody). Then \$L2 confirms this proposition by repeating it (with the addition of the discourse marker *disons* 'let's say'). The second case is the answer to a wh-question as in Example (7), where the interrogative pronoun *dove* 'where' is instantiated by the answer *al bi side* 'at the bi side'. A third case is correction. We will illustrate it by the monologue in Example (13). In this Example, the speaker repeats three times *des Français* 'of the French', so there is no reformulation strictly speaking, but the second occurrence questions the choice of the proposed denotation, while the third, preceded by *pas seulement* 'not only', refutes it.

In all cases, $\langle \text{tok\_id} \rangle$ points to the head of the first denotation.

- (13) French (Rhapsodie, D0004)
 \$L1: mais ce dont, ce dont vous parlez, c'est
but what, what you're talking about is
la crise générale des, des Français, oui,
the general crisis of the French, yes,
enfin, des Français, pas simplement
well, of the French, not just
des Français, hein, des, de l'humanité
of the French, but of the, of humanity
et de la lecture.
and reading.

5. Automatic Conversion between Speaker-based and Dependency-based Representation

The two proposed annotation views (speaker-based and dependency-based) rely on different maximal segmentation principles. As UD and SUD annotation schemas strongly rely on the notion of 'sentence' for segmentation, we propose providing the two views separately. The speaker-based view, properly enriched with the features introduced in Section 4, can be converted into the dependency-based view, which relies on the definition of 'sentence' as 'rectional unit', enabling coconstructions to be realised as regular syntactic dependencies.

In the Figure 1, the first line is the speaker-based view and the last line is the dependency-based view. An intermediate richer structure is introduced in which: new dependency relations are identified with a suffix */attach* and they are drawn in green. The richer structure helps to understand the conversion process and adds the possibility to query these modified dependencies specifically. Note also that metadata relative to the speaker, which would usually be placed at sentence level, is now moved to tokens, to take into account the fact that a dependency-based sentence can include tokens uttered by different speakers.

As explained above, there may be other dependency relations needing to be restored in dependency-based view. The new features *Promotion* and *Head* are used to reconstruct the expected dependency structure. The Figure 2 shows an Italian example where the oblique prepositional phrase *per mestiere* is built by two different speakers. In this case, we mark, on *per* that the structure is unfinished (*Scrap=Yes*) and that it has been promoted (*Promotion=case*). The second line in the figure show the intermediate (graph) structure where both dependencies on *per* are drawn: in red with suffix *sb* for the promoted relation and in green with suffix *attach*, the expected UD relation dependency-based view. The last line gives the UD regular annotation of the dependency-based view. Similar examples are given in appendix for English (Figure 4) and for Slovenian (Figure 5).

Note that when the promoted token has its own dependencies, they may have to be lifted up in the final dependency-based structure. The figure 4 gives such an example on a built English example *this is pretty* in which *pretty* is a separate maximal unit in speaker-based view.

6. Coconstructions in Existing UD Treebanks

6.1. French (Rhapsodie)

Coconstructions and backchannels were already annotated in the first version of Rhapsodie (Benzi-toun et al., 2010; Lacheret et al., 2014), using the delimiter *//+*, as in Example (3). When the treebank has been converted into UD, they were lost, and then reintroduced with a feature *AttachTo* for the address of the governor and *Rel* for the relation (Kahane et al., 2021). For backchannels, the value of *Rel* was *discourse*. Backchannels in Rhapsodie currently include *yes-no* answers to questions. Instead of *conj:reform* and *reparandum*, seven different labels were first considered (Kahane and Pietrandrea, 2012), but due to a quite poor inter-annotator agreement, they were merged into a unique relation, i.e., *conj:dicto*.

Among the 3,209 trees in the speaker-based version of Rhapsodie, 229 are annotated as backchannels, but there are also 93 interjections as roots that are potential backchannels and need to be revised. The treebanks also contain 35 coconstructions: 24 *conj:reform* and 11 modifiers, including two relative clauses.

6.2. Italian (KIParlaForest)

KIParla Forest (Pannitto et al., 2025) was first introduced in UD v2.17, based on the broader KIParla Corpus of Spoken Italian (Mauri et al., 2019; Ballarè

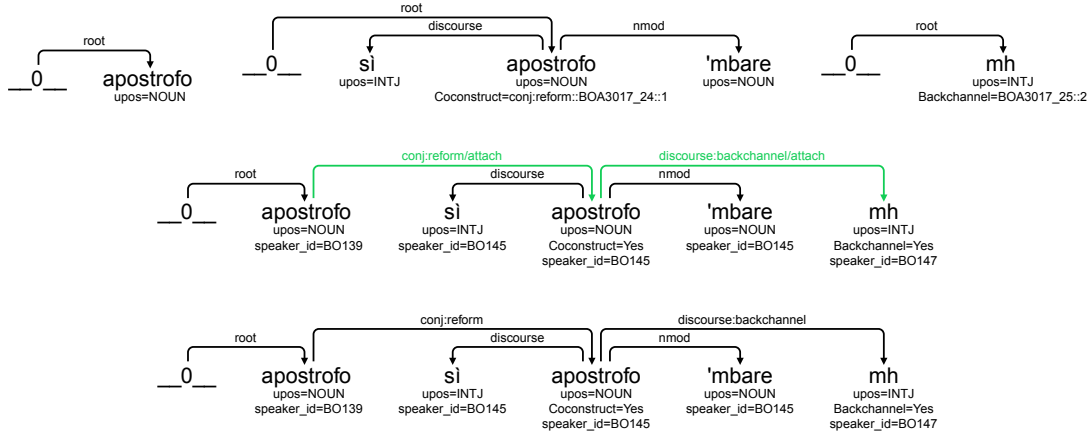


Figure 1: An Italian example of a `conj:reform` coconstruction and a backchannel. The example is taken from KIParla, BOA3017. Tr. en. ‘apostrophe’, ‘yes, apostrophe ‘mbare’, ‘mh’. ‘mbare’ is a dialectal form for *compare*, ‘mate’.

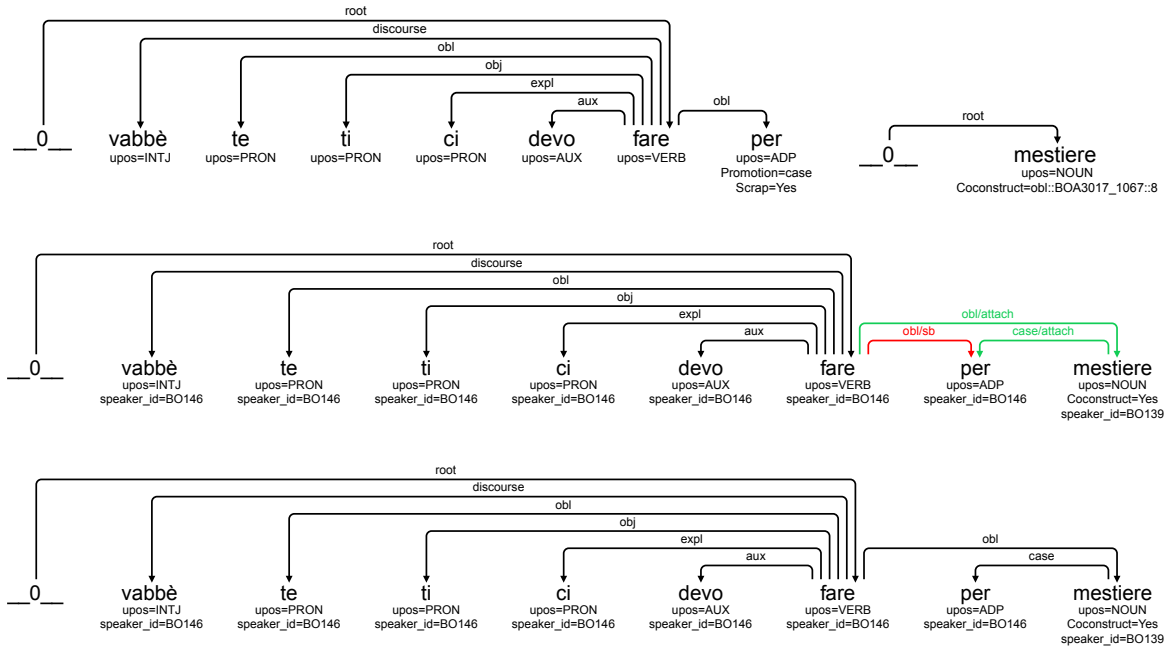


Figure 2: Italian example (10) with promotion (NB: the sentence is shortened, for space reasons). The first line shows speaker-based annotation, the second line shows the intermediate graph representation of the dependency-based view and line 3 shows final UD well-formed tree for the dependency-based view.

and Mauri, 2020).

The treebank currently hosts 2 conversations from the original corpus, an interview (BOD2018, comprising 2 speakers) and a free conversation (BOA3017, comprising 4 speakers), for a total of 9,135 tokens and 1,007 sentences.

Backchannels are annotated in 134 sentences: most of these are sentences composed by one single token that is frequently an INTJ such as *ah*, *eh*, *mh* or *ok* (67 cases), an ADV as *sì* ‘yes’, *no*, *certo* ‘certainly’, *ecco* ‘that’s it’, *esatto* ‘exactly’ (51 cases). Another significant group is that of

exact repetitions (4 cases) Also exclamations such as *porca troia* ‘holy shit’ are treated as cases of backchannels, and longer sequences such as *ah me l’hai raccontato* ‘right you told me’.

Currently the treebank contains 70 annotations of coconstructions. Most cases (38) concern conjuncts, either proper `conj` or `conj:reform` relations. `obl` and `acl:*` also constitute a big group (13 cases), as they capture the prototypical case of expansion of the dependency tree during interaction.

6.3. Slovenian (SST)

In the original compilation of the Spoken Slovenian Treebank (SST, Dobrovoljc and Nivre 2016; Dobrovoljc 2025), coconstructions were not treated explicitly. The treebank followed the speaker-based utterance segmentation of the GOS corpus (Verdonik et al., 2024), which encompasses 6,121 utterances and 98,393 tokens, and syntactic annotation was applied independently within each utterance. To evaluate the feasibility of the proposed guidelines and estimate the extent of coconstructions in Slovenian spoken data, we therefore began by targeting two configurations: completion of an unrealized dependency (Type A.1 in Section 3.1) and backchannelling (Section 3.3).

Backchannels were identified semi-automatically by extracting consecutive utterance pairs by different speakers where the first was not interrogative and the second contained one or more backchannel candidates (based on a preexisting list derived from items already marked as `discourse`, `INTJ` and/or `PART`). This procedure yielded 396 cases of backchannelling (69 different types), mostly single words (e.g. *ja* 'yes', *mhm* 'mm-hm', *aha* 'I see', *aja* 'oh, right', *dobro* 'okay'), but also longer sequences (e.g. *tako, tako, tako, tako* 'right, right, right, right').

For coconstruction proper (type A.1), we manually inspected utterance pairs where the first did not end in final punctuation, indicating potential incompleteness, and manually identified 17 cases of completed projected dependencies, including argument insertion (Speaker 1: *če atom odda elektrone, nastane...* 'if an atom gives away electrons, there arises...' – Speaker 2: *...kation* '...a cation') and sentence-level completion (Speaker 1: *kdor ti bo reklamo poslal...* 'whoever sends you an advertisement...' – Speaker 2: *...zaračuna* '...charges money').

These results provide an initial empirical basis for studying coconstructions in Slovenian and a retrieval procedure that can be refined to capture additional types, or extended to other UD treebanks where such phenomena have not yet been explicitly annotated.

7. Conclusion

UD has been first developed on the basis of written corpora with the aim of developing parsers (McDonald et al., 2013). The collection now includes spoken data for languages with a written tradition (as the three languages considered in this paper), as well as languages with only a spoken tradition (including sign languages). Due to the lack of clear guidelines for spoken data, important discrepancies exist among treebanks for spoken data (Dobrovoljc, 2022). This work is part of a common effort to propose guidelines and unify the annotation of

different treebanks. The first step in the syntactic analysis of spoken data is segmentation into minimal units (words or morphemes) and maximal units. The latter question is the main topic of this paper and the question of coconstructions is one of the very first problems to be solved. We hope to have convinced the reader of the interest in the question, as well as the relevancy of our propositions.

While the framework has been developed with spoken interaction in mind, its underlying mechanisms are more broadly applicable. Two extensions are worth noting. First, the proposed annotation scheme is potentially relevant for computer-mediated communication (CMC) data — such as instant messaging, or platforms like Slack or Discord — where users frequently fragment a single syntactic unit across multiple messages by sending partial utterances. These *split utterances* are structurally analogous to unfinished phrases in speech, and the `Coconstruct`, `Scrap`, and `Promotion` features could be applied to capture cross-message syntactic dependencies. A distinctive challenge of CMC, however, is the presence of *lag*: responses may not appear chronologically adjacent to the message they depend on, due to interleaved threads or parallel conversations. The pointer-based design of our features (encoding the identity of the eliciting token rather than relying on adjacency) makes the framework robust to such non-linear environments. Second, even within spoken data, the mechanism may be useful for cases of same-speaker continuation across deliberately maintained turn boundaries — for instance, when a speaker resumes a syntactic structure after a long backchannel sequence and the annotator wishes to preserve segmentation boundaries for prosodic or discourse reasons. In both cases, the features introduced here provide a lightweight but expressive mechanism for encoding syntactic continuity without collapsing the original segmentation.

We submit this proposal to the UD community hoping that it will be adopted as a standard extension for spoken and interactive treebanks, contributing to a more linguistically adequate treatment of multi-party and mediated language data.

Acknowledgements

This work received support from the CA21167 COST action UniDive, funded by the European Union via the COST (European Cooperation in Science and Technology), and from the ARIS research programme P6-0411.

References

- Peter Auer. 2024. [Online-Syntax](#). In Peter Auer, editor, *Online-Syntax: Eine Einführung in die Analyse gesprochener Sprache*, pages 1–19. Springer, Berlin, Heidelberg.
- Silvia Ballarè and Caterina Mauri. 2020. La creazione del corpus kiplara: criteri metodologici e prospettive future. *Rivista Italiana di Dialettologia*, 44:53–69.
- Christophe Benzitoun, Anne Dister, Kim Gerdes, Sylvain Kahane, Paola Pietrandrea, and Frédéric Sabio. 2010. tu veux couper là faut dire pourquoi-propositions pour une segmentation syntaxique du français parlé. In *Congrès Mondial de linguistique française (CMLF)*, pages 2075–2090.
- Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredennick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. 1995. Bracketing guidelines for treebank ii style penn treebank project. Technical report, University of Pennsylvania.
- Claire Blanche-Benveniste. 1987. *Pronom et syntaxe: l'approche pronominale et son application au français*. Peeters Publishers.
- Virginia Calabria. 2026. [The Grammar of Collaborative Turns: Temporality, Projection, Emergence and Clause Combination in Italian Talk-in-Interaction](#). De Gruyter Mouton.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. [Universal Dependencies](#). *Computational Linguistics*, 47(2):255–308.
- Arnulf Deppermann and Susanne Günthner. 2015. *Temporality in interaction*, volume 27. John Benjamins Publishing Company.
- 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. 2025. Treebanking spoken slovenian: New data, models, and lessons learned. *Prispev. Novejso Zgod.*, 65(3).
- 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).
- Péter B. Furkó. 2020. *Discourse Markers and beyond: Descriptive and Critical Perspectives on Discourse-Pragmatic Devices across Genres and Languages*. Palgrave Macmillan.
- Marja-Liisa Helasvuo. 2004. Shared syntax: the grammar of co-constructions. *Journal of Pragmatics*, 36(8):1315–1336.
- Erhard W. Hinrichs, Julia Bartels, Yasuhiro Kawata, Valia Kordoni, and Heike Telljohann. 2000. The verbmobil treebanks. In *KONVENS 2000 / Sprachkommunikation, Vorträge Der Gemeinsamen Veranstaltung 5. Konferenz Zur Verarbeitung Natürlicher Sprache (KONVENS), 6. ITG-Fachtagung "Sprachkommunikation"*, page 107–112, DEU. VDE-Verlag GmbH.
- Paul J Hopper. 1992. Times of the sign: Discourse, temporality and recent linguistics. *Time & Society*, 1(2):223–238.
- Andreas H. Jucker and Sara W. Smith. 1996. Explicit and implicit ways of enhancing common ground in conversations. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 6(1):1–18.
- Daniel Jurafsky, Elizabeth Shriberg, and Debra Bisasca. 1997. Switchboard swbd-damsl shallow discourse-function annotation: Coders manual. Technical report, Institute of Cognitive Science, University of Colorado, Boulder, CO, USA.
- Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021. [Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal](#). In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 35–47, Sofia, Bulgaria. Association for Computational Linguistics.
- Sylvain Kahane and Paola Pietrandrea. 2012. [La typologie des entassements en français](#). *SHS Web of Conferences*, 1:1809–1828.
- Shah Zeb Khan. 2023. The role of discourse markers, deixis, hedges, and backchannels in enhancing communication and emotional support. *Review of Law and Social Sciences*, 1(3):31–40.
- Anne Lacheret, Sylvain Kahane, Julie Beliao, Anne Dister, Kim Gerdes, Jean-Philippe Goldman, Nicolas Obin, Paola Pietrandrea, and Atanas Tchobanov. 2014. Rhapsodie: A prosodic-syntactic treebank for spoken french. In *Language Resources and Evaluation Conference (LREC)*.
- Gene H. Lerner. 1991. [On the syntax of sentences-in-progress](#). *Language In Society*, 20(3):441–458.

- Caterina Mauri, Silvia Ballarè, Eugenio Gorla, Massimo Cerruti, and Francesco Suriano. 2019. KIParla Corpus: A New Resource for Spoken Italian. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*.
- Ryan McDonald, Joakim Nivre, Yvonne Quirnbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 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.
- Daniela Mereu, Francesco Cangemi, and Martine Grice. 2024. Backchannels are not always very short utterances. the case of italian multi-unit backchannels. *Journal of Pragmatics*, 228:1–16.
- Marie Meteer, Ann Taylor, Robert MacIntyre, and Ruth Iyer. 1995. Disfluency annotation stylebook for the switchboard corpus. Technical report, Linguistic Data Consortium.
- Amina Mettouchi and Martine Vanhove. 2020. Prosodic segmentation and grammatical analysis in cross-linguistic corpora. In *42nd Annual Conference of the German Linguistic Society (DGfS), Workshop Corpus-based typology: spoken language from a cross-linguistic perspective*.
- Tsuyoshi Ono and Sandra A. Thompson. 1996. Interaction and syntax in the structure of conversational discourse: collaboration, overlap, and syntactic dissociation. In *Computational and conversational discourse*, pages 67–96, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Ludovica Pannitto, Eleonora Zucchini, Silvia Ballarè, Cristina Bosco, Caterina Mauri, and Manuela Sanguinetti. 2025. [Introducing KIParla forest: seeds for a UD annotation of interactional syntax](#). In *Proceedings of the Eighth International Conference on Dependency Linguistics (depling, Syntaxfest 2025)*, pages 54–73, Ljubljana, Slovenia. Association for Computational Linguistics.
- Paola Pietrandrea, Sylvain Kahane, Anne Lacheret-Dujour, and Frédéric Sabio. 2014. [The notion of sentence and other discourse units in corpus annotation](#). In Tommaso Raso and Heliana Mello, editors, *Spoken Corpora and Linguistic Studies*, pages 331–364.
- Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg. 2012. [Sentence boundary detection: a long solved problem?](#) In *Proceedings of COLING 2012: Posters*, pages 985–994, Mumbai, India. The COLING 2012 Organizing Committee.
- Dwijen Rudrapal, Anupam Jamatia, Kunal Chakma, Amitava Das, and Björn Gambäck. 2015. [Sentence boundary detection for social media text](#). In *Proceedings of the 12th International Conference on Natural Language Processing*, pages 254–260, Trivandrum, India. NLP Association of India.
- Harvey Sacks. 1992. *Lectures on Conversation*. Blackwell, Oxford.
- Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.
- 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.
- Rachele Sprugnoli and Arianna Redaelli. 2025. [Annotating manzoni: Challenges in the annotation of lemmas, POS and features in “I promessi sposi”](#). In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 1084–1092, Cagliari, Italy. CEUR Workshop Proceedings.
- Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Klincksieck.
- Lucien Tesnière. 2015. *Elements of structural syntax*. John Benjamins Publishing Company.
- Darinka Verdonik, Kaja Dobrovoljc, Tomaž Erjavec, and Nikola Ljubešić. 2024. [Gos 2: A new reference corpus of spoken Slovenian](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7825–7830, Torino, Italia. ELRA and ICCL.
- Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8):1177–1207.

A. Annotation Examples

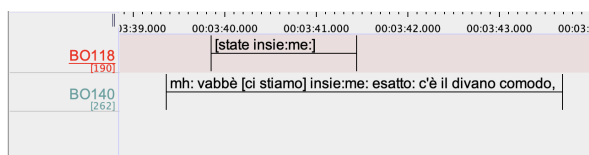


Figure 3: Original transcription of Example (10) in the KIParla corpus.

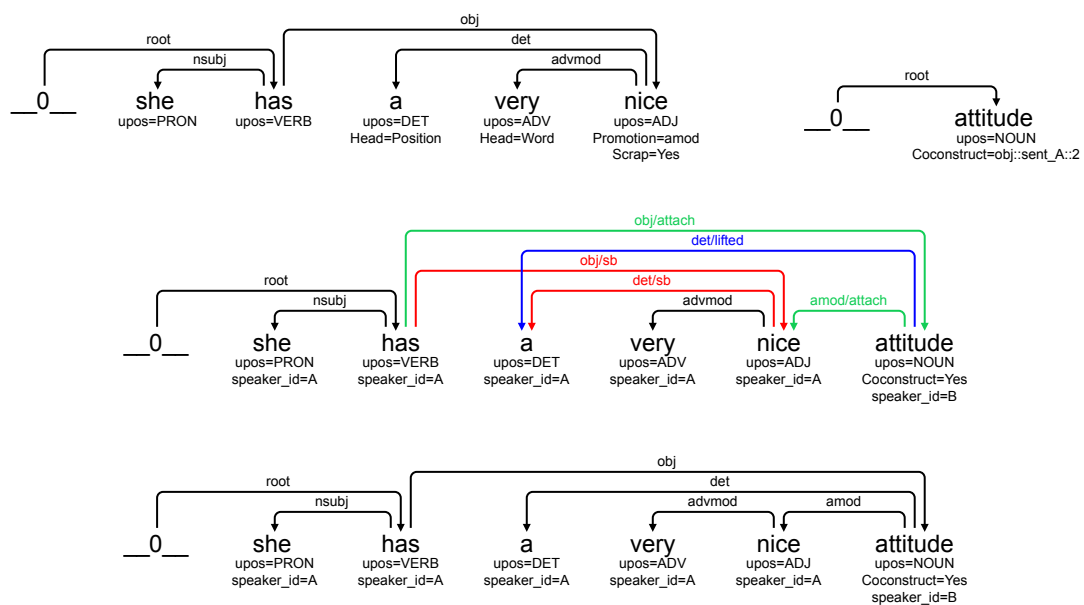


Figure 4: English example with lifting.
 Line 1: speaker-based annotation;
 Line 2: intermediate representation of the dependency-based view;
 Line 3: final UD tree for the dependency-based view

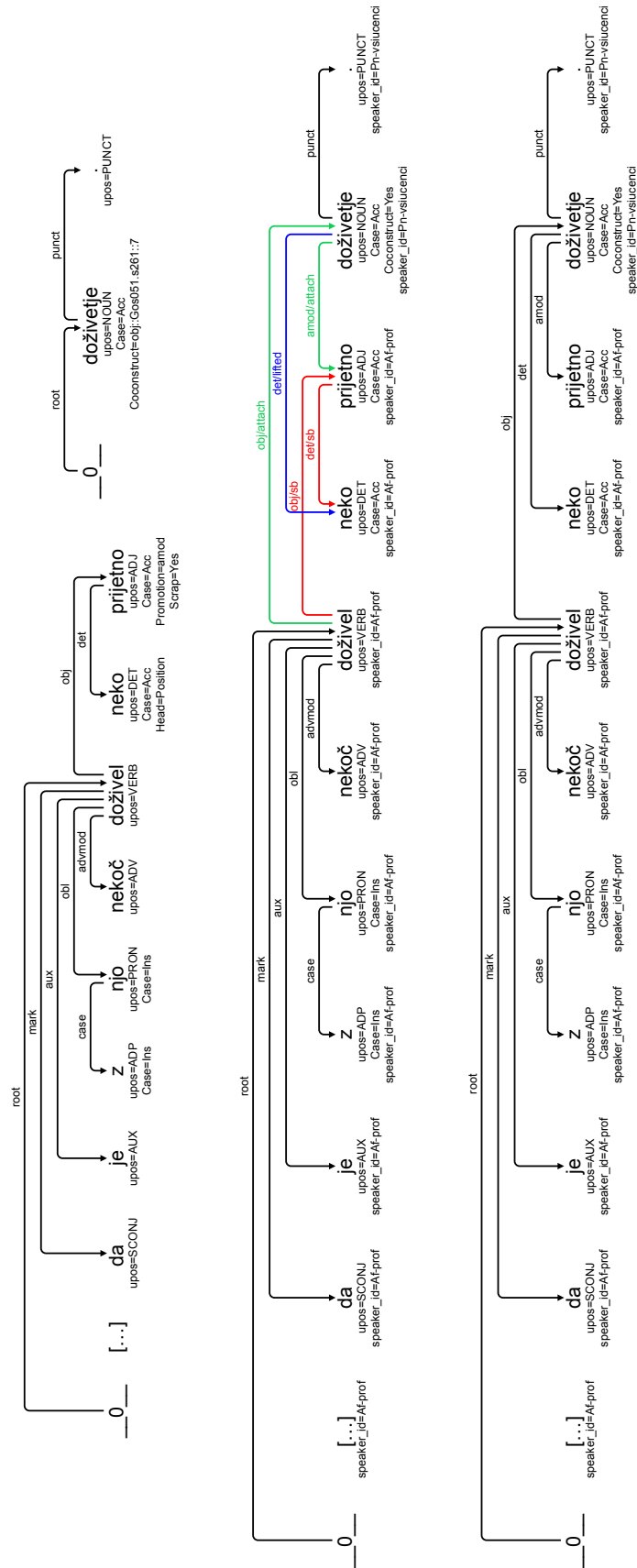


Figure 5: Slovenian example of promotion
 '[...] that he once had with her a pleasant ... experience.'

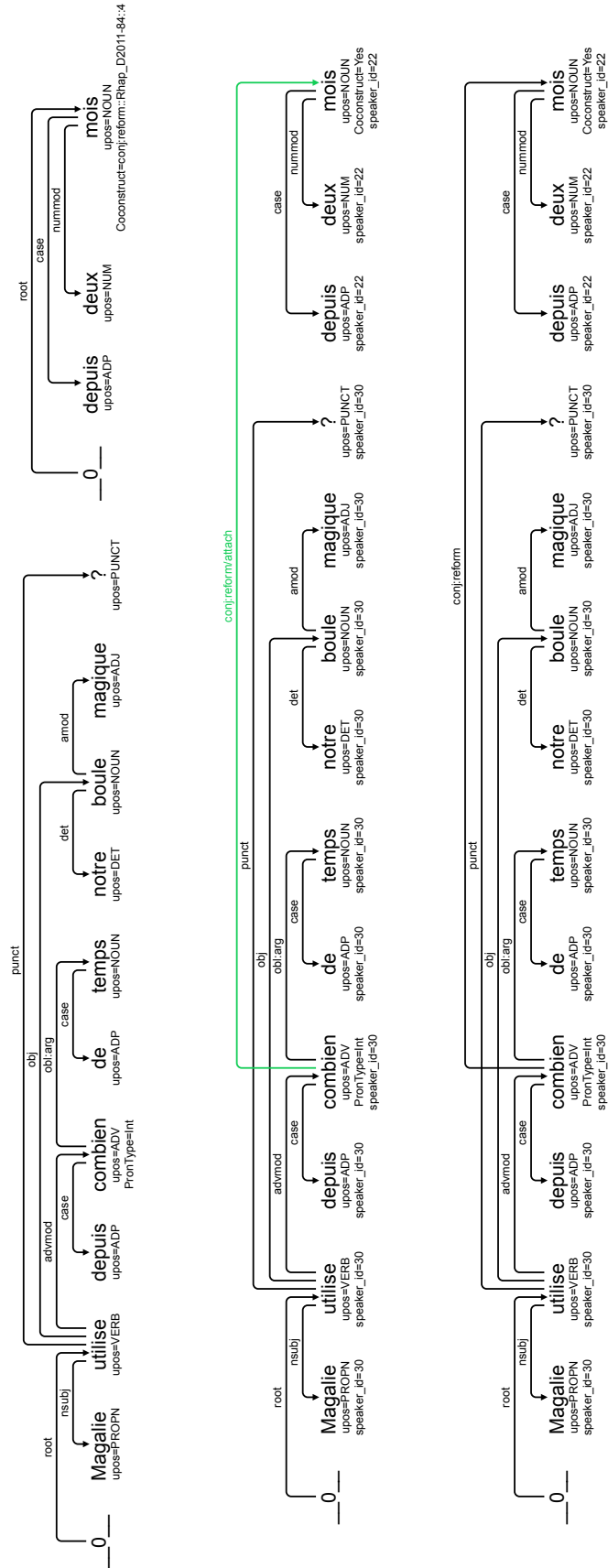


Figure 6: French example with a question
 'How long has Magalie been using our magic ball?', 'for two months'

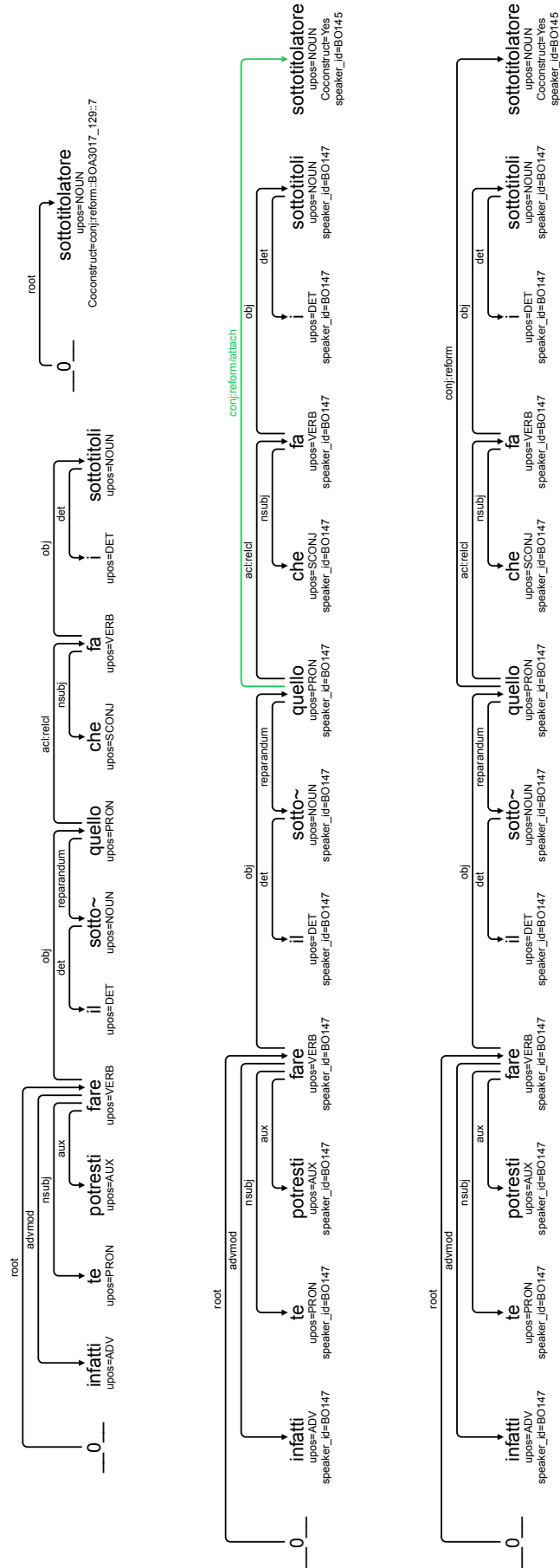


Figure 7: Italian repair example, see Example (12) for translation.

Verifying the Menzerath-Altmann Law in the Verbal Domain in 180 Languages

Pegah Faghiri^{1,2}, Kim Gerdes², Sylvain Kahane^{1,3}

¹Paris Nanterre University, Modyco, CNRS ²Paris-Saclay University, LISN, CNRS

³IUF - Institut Universitaire de France

pfaghiri@parisnanterre.fr, kim.gerdes@universite-paris-saclay.fr, skahane@parisnanterre.fr

Abstract

We present a large-scale evaluation of the Menzerath-Altmann law (MAL) in the verbal domain across 180 languages, using the Universal Dependencies (UD) treebank collection (v2.17). MAL predicts that as the number of constituents of a linguistic unit increases, their average size decreases. We propose a metric to estimate the MAL effect across corpora of widely varying sizes and define threshold-based categories to classify languages along a MAL preference cline. Crucially, we analyse the preverbal and postverbal domains separately, in addition to the standard bilateral MAL, and control for potential sampling bias by comparing results across language families (Indo-European vs. non-Indo-European) and syntactic types (VO, OV and no dominant order). Our results confirm MAL as a typologically widespread preference but not an absolute universal: several languages display a trivial or even opposite (anti-MAL) tendency. Furthermore, we uncover a significant asymmetry between the two sides of the verb: the MAL effect is stronger in the postverbal domain, while anti-MAL is stronger in the preverbal domain. VO languages tend to show a stronger MAL preference postverbally, whereas OV languages do so preverbally. These findings challenge the widespread assumption that length-based ordering constraints apply symmetrically on both sides of the verb and contribute new cross-linguistic evidence to the debate on the interaction between dependency length minimization and constituent size.

Keywords: Menzerath-Altmann law, quantitative typology, Universal Dependencies, verbal valency, constituent length, word order

1. Introduction

The current version of the Universal Dependencies (UD) collection of syntactic treebanks allows us to evaluate universal properties across a set of 180 languages (v2.17, Zeman et al., 2025). In this paper, we investigate the typological validity of the so-called Menzerath-Altmann law (MAL), which concerns the nature of complexity in linguistic units in terms of the relationship between the size of a given unit and the size of its components (Menzerath, 1928, 1954; Altmann, 1980). Roughly speaking, the greater the number of components of a given unit, the shorter their respective length. In the verbal domain, for instance, this implies that the length of verbal arguments decreases as the valency, or number of verbal dependents, increases (Mačutek et al., 2017). We also aim to propose a reliable metric to evaluate this hypothesis in the unbalanced UD collection, which contains annotated corpora of significantly different sizes—including the so-called LOL languages (“Literate, Official, and with Lots of users”), with millions of tokens, but also a number of corpora from less-studied languages with fewer than a thousand tokens. Furthermore, in order to make typologically relevant claims based on an *ad hoc* language sample that was not designed for such purposes, we compare

independent subsets of the collection separately to ensure that our conclusions are replicable beyond the sample as a whole and are not mere artefacts of sampling biases. First, with respect to language families, we consider the two main groups that are fairly equally represented in the collection, that is, Indo-European (IE) and non-Indo-European (non-IE) languages, separately. Second, we control for syntactic language types, namely OV, VO, and no dominant order (NDO), which we define using a token-based approach.¹

Another important contribution of our study is that we examine preverbal and postverbal domains separately. That is, in addition to our general (or combined) MAL metrics, we calculate two separate sets of metrics for left/preverbal and right/postverbal arguments, respectively LMAL and RMAL. It is important to bear in mind that constituent size is known to obey other constraints. In particular there is the well-studied Dependency Length Minimization (DLM), which is one of the most widespread accounts of length-based word order preferences

¹We defined the three types based on a VO score calculated for each language, as the ratio of the *obj* relations with a nominal object occurring after the verb. We labelled VO languages with a score greater than 0.67 (that is, languages in which the VO order is at least twice more widespread than OV), and OV languages with a score less than 0.33.

Authors are listed in alphabetical order.

across languages (Futrell et al., 2015; Ferrer i Cancho, 2004). According to DLM, natural languages tend to favour orders that minimize dependency lengths. This results in a short-before-long preference (SbL) in the postverbal domain (typical of VO languages) and the reverse, that is a long-before-short preference (LbS), in the preverbal domain (typical of OV languages). In other words, the left and right sides of the verb are assumed to behave as mirror images, showing the same DLM effect but in opposite directions (Hawkins, 2007). Unlike DLM, however, MAL has a direction-free definition. Consequently, assuming that a direction-free length-based effect should apply similarly on each side of the verb, MAL is expected to hold equally on both sides. Relatedly, Chen et al. (2022) showed that the combined effect of MAL and HCS on the postverbal side is a cross-linguistic universal, but left the preverbal domain as an open question.

Our results show that MAL is a clearly widespread typological preference across the languages of the world. It is, however, not an absolute universal principle. Importantly, we find that while MAL is a strong cross-linguistic statistical preference, there are not only languages in which this preference is trivial, but also languages that display the opposite tendency. Furthermore, our results reveal a clear difference between the pre- and postverbal domains. On the one hand, the MAL preference is stronger in the postverbal domain; on the other hand, the anti-MAL preference is stronger in the preverbal domain. Importantly, our findings show that in the postverbal domain, VO languages are more likely to exhibit a strong RMAL preference, while, conversely, OV languages are more likely to exhibit a strong LMAL preference in the preverbal domain.

In the following, we first define MAL and briefly present the previous studies on which we build (Section 2). Next, we present our data extraction method using the UD annotation scheme (Section 3) and the metrics we use in this study (Section 4 and Section 5), while motivating our methodological choices. We then present our results (Section 6), including a closer look at individual languages with unexpected behaviour (Section 6.3), before concluding the paper (Section 7). The results for MAL, LMAL, and RMAL on the 180 languages of UD are given in Table 5 in Appendix A. An interactive version of all results is available at <https://typometrics.elizia.net/menzerath/>; the full analysis code is released under a CC BY 4.0 licence at <https://github.com/typometrics/UDW26-Menzerath>.

2. The Menzerath-Altmann law in the verbal domain

MAL predicts a negative correlation between the average length of constituents and their total number in any given domain. MAL has been previously studied in the verbal domain using the UD annotation scheme in Czech (Mačutek et al., 2017), as well as in version 2.3 of the UD collection covering 76 languages (Tanaka-Ishii, 2021). Chen et al. (2022) linked MAL to the Heavy Constituent Shift (HCS) phenomenon on the postverbal side and showed that the co-effect of MAL and HCS constitutes a very regular universal across 80 languages of SUD 2.7. Our study builds on the latter while proposing a number of substantial improvements in order to draw more reliable and typologically valid conclusions. First, we study MAL in the preverbal and postverbal domains separately, in addition to bilateral MAL. Second, our data extraction is more fine-grained than in Tanaka-Ishii’s study (see Section 3). Third, we use the latest version of UD, which, with 180 languages, offers a greater diversity of languages from around the world. Finally, we analyse the data while taking into account both language families and language types, that is, VO, OV, or no dominant order (NDO).

3. Extracting the constituents of verbal constructions in UD

We assume a certain familiarity with the UD annotation scheme in this paper and simply recall that the syntactic structure of all sentences is encoded as a dependency tree. Here we focus on verbal constructions, that is, a verb and all the constituents that depend on it. We consider all lexical verbs and have decided not to restrict our study to main verbs (see (1b)) for at least two reasons: first, including all lexical verbs can double or triple the number of constructions, which is not negligible when the corpus is small; second, in some sentences, the most interesting verb in terms of construction may be subordinated because the main verb is merely a modal or an auxiliary.

All dependents of a verb (in the tree) cannot be considered constituents (of the verbal construction we are interested in). The following are excluded in our data extraction: punctuation signs (attached by the `punct` relation in UD), discourse markers (`discourse` relation; see *alas* in (1a)), juxtaposed clauses or parentheses (`parataxis` relation), conjuncts in a coordination (`conj` relation; (1c)), and coordinating conjunctions (`cc` relation; see *but* in (1a)). The `vocative` relation, which concerns addressees to a participant (*Guys, take it easy!*), is excluded but dislocated arguments (*These guys, I don’t trust them.*) are kept in, especially because in

some languages the boundary between governed units and dislocated units is difficult to draw. The *aux* relation (see *do* in (1a)), is excluded, considering that auxiliaries are part of the verb form; compounds (see *out* in (1d)) are excluded as well. Markers of subordination such as subordinating conjunctions and adpositions (*mark* and *case* relations; see *when* in (1b)), which are analysed as dependents in UD but are generally considered as heads of their constructions, are also discarded.²

Examples of English verbal constructions are given in (1) for illustration; the verb is underlined and the constituents are shown between square brackets; filtered elements are given in italics.

- (1) a. *But* [I] , *alas* , *do* [not] know [how to see sheep] [through the walls of boxes]. *en_littleprince*
- b. This means that *when* [you] move [a series or category field] [to the filter area and back], previously hidden items are again hidden. *en_lines*
- c. [You] tell [me] [what they look like] *and I'll tell you what they are.* *en_childes*
- d. [One minister] [reportedly] handed out [100 dollar 'gifts'] [to journalists attending a press conference for Allawi], [a practice that brings back bad memories to many Iraqis]. *en_ewt*

Another possibility taken into account is that a constituent may be split into two contiguous parts, for example in cases of extraction or extraposition. These cases are treated as two separate constituents. For instance, in (2a), the relative pronoun *what* is extracted and separated from its governor *do*, and in (2b), the relative clause *that go well with the Santorum cocktail* is extraposed and separated from its governor *dishes*.

- (2) a. [What] *do* [you] like [to do] ? *en_ewt*
- b. [What culinary dishes] *would* [you] recommend [that go well with the Santorum cocktail]? *en_gum*

For each language *L* and each $n \geq 1$, we extracted all the verbal constructions with n constituents. We then calculated the size of each constituent, excluding punctuation. In other words, we kept all relations except *punct*; if a constituent contains a coordination or a discourse marker, there is no reason to remove them.

We call $MAL_n(L)$ the mean size of the constituents in all verbal constructions with n constituents. Note that we compute $MAL_n(L)$ for a

given language *L* only when we have at least 100 constructions with n constituents. We return to this point in Section 4. We define $LMAL_n(L)$ and $RMAL_n(L)$ in the same way. For $LMAL$, all constructions with n preverbal constituents (i.e. on the left) were selected, regardless of the other side of the verb. Likewise for $RMAL$, with all constructions having n postverbal constituents (i.e. on the right).

4. Defining the MAL effect

We call $\lambda MAL(L)$ the function that maps $MAL_n(L)$ to each n . It has been shown by Altmann (1980) that a function such as $\lambda MAL(L)$ should follow a power law. In other words, $MAL_n(L)$ is expected to be roughly proportional to $n^{-\beta}$, where β is a positive value (generally smaller than 1), which we call the *MAL effect* for *L*. Note that if $MAL_n(L) \approx n^{-\beta}$, then $\log(MAL_n(L)) \approx -\beta \log(n)$. In other words, $-\beta$ is the slope of the regression line in the log-log representation of $\lambda MAL(L)$. The greater β is, the faster the decrease and the stronger the MAL effect. Figure 1 gives the value of β (1.171) as well as the coefficient R^2 (0.892), which shows that $\lambda MAL(\text{German})$ is close to a power law and that MAL holds for German.

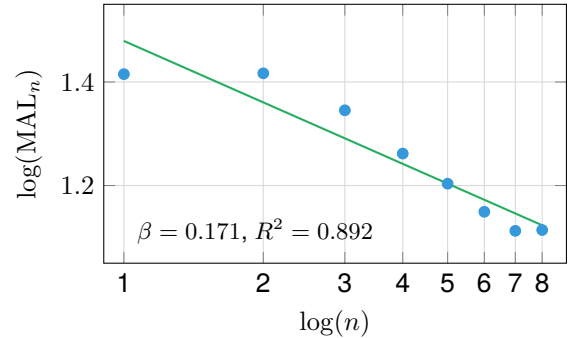


Figure 1: $\lambda MAL(\text{German})$.

Two methodological clarifications are in order at this point. Firstly, there is the question of the statistical reliability of the estimated slope. Indeed, our study faces a methodological challenge in calculating the MAL effect as defined above for our sample, the UD collection, which includes corpora of different sizes, varying from fewer than 1k to more than 1000k tokens. As a result, our samples of extracted tokens, on the basis of which we aim to calculate one MAL effect per language, vary in size. This is a problem because we are estimating coefficients via linear regression modelling and the statistical reliability of these estimates depends on the sample size. In other words, our samples do not offer the same statistical power for the fitted models. To accommodate this issue, we apply a threshold to guarantee a baseline statistical power

²In her study, Tanaka-Ishii (2021) did not exclude any dependent, except the punctuation.

across samples. We have opted for a fairly high threshold (a minimum of 100 tokens per configuration) in order to maximize the reliability of our MAL metrics, at the cost of losing a number of languages with smaller corpora. Given that we are evaluating the null hypothesis that MAL is a universal preference, our rationale is to favour minimizing the risk of incorrectly detecting counterexamples, at the cost of mislabelling borderline cases.

Secondly, there is the choice of which values of n to include or exclude from our calculations. Tanaka-Ishii (2021) showed that, in her data, the values for $n = 1$ and $n = 2$ did not always align, and, accordingly, she took into account only the values for $n \geq 3$ when computing β . We do not follow her on this methodological decision, as we believe that the problem in her study may, at least partly, be due to the fact that the data were not filtered - that is, contrary to our data extraction, in her study all verbal dependents were taken into account without excluding any relations. Moreover, since our data is filtered, for many small corpora there may not be enough occurrences to calculate MAL_n for $n \geq 4$. This is even more problematic when we look at $LMAL_n$ and $RMAL_n$. In other words, excluding MAL_n for $n \leq 2$ is more costly and less straightforward in our study. We therefore considered all options before making our own choice.

We call $\beta(j \rightarrow k)$ the slope of λMAL for $j \leq n \leq k$. If $k = j + 1$, $\beta(j \rightarrow k)$ is simply the slope between the points $(\log(MAL_j), \log(j))$ and $(\log(MAL_k), \log(k))$. In Figure 2, we compare $\beta(1 \rightarrow 2)$ with $\beta(2 \rightarrow \infty)$. The considerable dispersion suggests that the value for MAL_1 is indeed poorly aligned with the other values. The cases where MAL_1 is aligned correspond to points on the diagonal $y = x$. Nevertheless, we observe that the value of MAL_1 can sometimes fall below the regression line for $n \geq 2$ (as is the case for German, Fig. 1) and sometimes above it (see Arabic, Fig. 3, or Czech, Fig. 4).

Given these observations, we decided to take $\beta(1 \rightarrow \infty)$ as the metric for evaluating the MAL effect.³ There are 131 languages in our sample for which a value for $\beta(1 \rightarrow \infty)$ is calculated for λMAL , and respectively 124 and 103 for $\lambda LMAL$ and $\lambda RMAL$. These values mostly fall between -0.3 and 0.3 .⁴ Given the range of values for $\beta(1 \rightarrow \infty)$ in

³We should nevertheless mention an additional issue with $n = 1$. While the number of configurations with n constituents decreases for $n \geq 2$, only 20 of the 131 languages that have at least 100 configurations for $n = 2$ and $n = 3$ have more data for $n = 1$ than for $n = 2$. Among them, 12 languages have fewer than 100 configurations with one constituent ($n = 1$), and we must take $\beta(2 \rightarrow \infty)$ to evaluate the MAL effect.

⁴For λMAL , no $\beta(1 \rightarrow \infty)$ values are below -0.3 and only 9 languages have a $\beta(1 \rightarrow \infty)$ greater than 0.3 . The

our data, we defined the following three categories. Here again, our aim was to minimize the risk of incorrectly detecting counterexamples, which resulted in a comparatively large borderline (or grey) zone.

- Language L with $\beta(1 \rightarrow \infty) > 0.1$ is considered to have a *MAL effect*;
- Language L with $\beta(1 \rightarrow \infty) < -0.1$ is considered to have an *Anti-MAL effect*.
- Language L with $-0.1 \leq \beta(1 \rightarrow \infty) \leq 0.1$ is considered to be in the grey zone;

The majority of languages in our sample (62 languages) fall into the first category, that is, they show a MAL effect. However, there are also 10 languages that show an Anti-MAL effect. The remaining 59 languages are in the grey zone.

5. MAL compliance ratio

In addition to the MAL effect, we define a simple regression-free measure, which we call the MAL compliance ratio. This metric gives the proportion of pairs where MAL_n decreases (for instance, $MAL_{n+1} \leq MAL_n$).

We calculated an average compliance ratio for MAL, LMAL, and RMAL for all languages with a corresponding value for $\beta(1 \rightarrow \infty)$, and defined the following three MAL compliance types:

- Language L with a MAL compliance ratio > 0.67 is considered to have high MAL compliance;
- Language L with a MAL compliance ratio < 0.33 is considered to have low MAL compliance.
- Language L with $0.33 \leq \text{MAL compliance ratio} \leq 0.67$ is considered to have medium MAL compliance;

6. Results

6.1. MAL effect in the UD collection

Out of the 186 languages available in our data, we calculated a $\beta(1 \rightarrow \infty)$ for 131. In this sample, we identify 62 languages with a MAL preference, 10 with an Anti-MAL preference, and 59 in the grey zone. Importantly, these three categories are fairly comparably represented across language families as well as across the main syntactic language types (Table 1). The only noticeable exception is that

situation is similar for $\lambda LMAL$. For $\lambda RMAL$, 48 languages have a value greater than 0.3 , but for the sake of comparison we keep the same categorization.

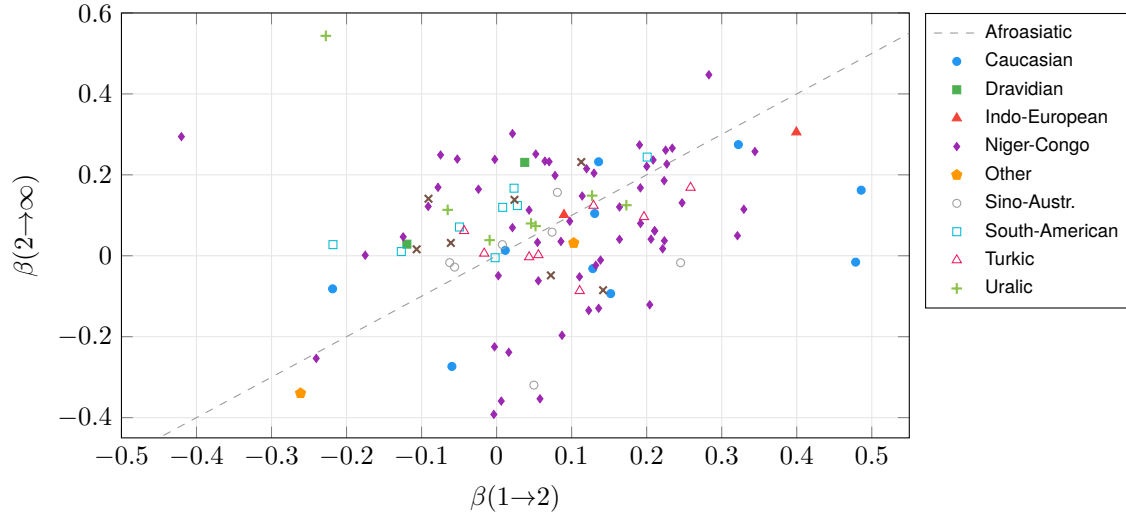


Figure 2: $\beta(1 \rightarrow 2)$ vs $\beta(2 \rightarrow \infty)$

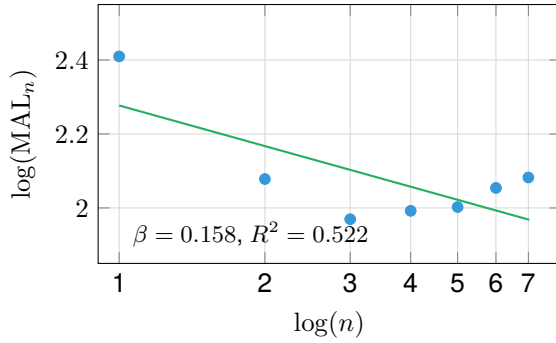


Figure 3: $\lambda\text{MAL}(\text{Arabic})$.

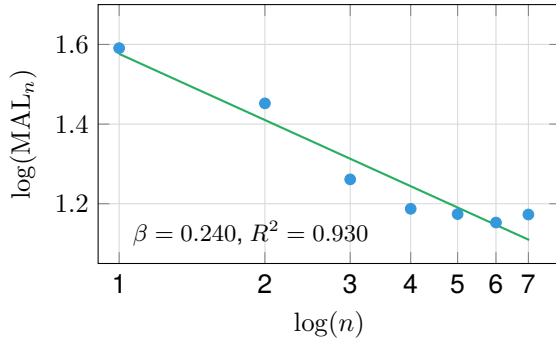


Figure 4: $\lambda\text{MAL}(\text{Czech})$.

languages with no dominant order (NDO) are significantly more likely to be in the grey zone (and less likely to have a MAL preference) than VO or OV languages.⁵

⁵More precisely, the difference is significant between NDO and VO languages for the grey zone (respectively, 65% vs. 36%, Fisher's Exact test $p < 0.05$), as well as for the MAL type (25% vs. 56%, Fisher's Exact test $p < 0.05$).

Family / MAL type	IE	Non-IE	Total
Anti-MAL	6	4	10
Gray zone	22	37	59
MAL	38	24	62
Total	66	65	131

VO/OV / MAL type	NDO	OV	VO	Total
Anti-MAL	2	2	6	10
Gray zone	13	20	26	59
MAL	5	16	41	62
Total	20	38	73	131

Table 1: MAL type per language family and VO/OV type.

For RMAL (postverbal domain) and LMAL (preverbal domain), we calculated $\beta(1 \rightarrow \infty)$ for 124 and 103 languages respectively, out of the total of 186 languages (Table 2 and Table 3). Interestingly, we find a clear difference between the pre- and postverbal domains. On the one hand, while we identify 62 languages (47%) showing a MAL effect, there are 81 languages (79%) showing a MAL effect in the postverbal domain (an RMAL effect), compared to only 39 languages (31%) showing a MAL effect in the preverbal domain (an LMAL effect). In other words, 79% of languages show a MAL effect in the postverbal domain, compared to 31% in the preverbal domain, while the combined MAL effect is found in 47% of languages. On the other hand, while we identify 10 Anti-MAL languages, there are 29 Anti-LMAL languages and 7 Anti-RMAL languages. In other words, 23% of languages show an Anti-MAL

effect in the preverbal domain, compared to 7% in the postverbal domain, while 8% show a bilateral Anti-MAL effect.

Family RMAL type	IE	Non-IE	Total
Anti-MAL	4	3	7
Gray zone	5	10	15
MAL	51	30	81
Total	60	43	103

Family LMAL type	IE	Non-IE	Total
Anti-MAL	15	14	29
Gray zone	28	28	56
MAL	20	19	39
Total	63	61	124

Table 2: RMAL and LMAL per language family.

We can therefore speak of both an RMAL bias and an Anti-LMAL bias—in other words, a preference for MAL in the postverbal domain and for Anti-MAL in the preverbal domain. Importantly, in our data, as with the combined MAL, the three LMAL and RMAL types are also fairly well distributed across language families (Table 2). However, the same is not true for syntactic language types (Table 3). Nevertheless, the differences we observe in LMAL and RMAL types for OV and VO languages show a familiar pattern and are not surprising.

VO/OV RMAL type	NDO	OV	VO	Total
Anti-MAL	3	2	2	7
Gray zone	2	4	9	15
MAL	14	10	57	81
Total	19	16	68	103

VO/OV LMAL type	NDO	OV	VO	Total
Anti-MAL	2	4	23	29
Gray zone	12	19	25	56
MAL	4	15	20	39
Total	18	38	68	124

Table 3: RMAL and LMAL per VO/OV type.

In the postverbal domain, we find that VO languages show a clearer bias towards RMAL than OV languages. In our sample of 103 languages with enough data on the postverbal domain, we have 68 VO, 16 OV, and 19 NDO languages. Out of the 81 languages showing an RMAL effect, that is, with a significant positive RMAL effect in the

postverbal domain, 57 are VO and 10 are OV languages. That is, 83% of our VO languages show an RMAL effect compared to 63% of our OV languages. Moreover, the Anti-RMAL preference is particularly rare among VO languages (2 out of 68 languages, 3%, compared to 2 out of 16, 13%, for OV languages).

Conversely, in the preverbal domain, we observe a larger proportion of OV languages with an LMAL preference than VO languages: 39% (15 out of 38) of OV languages show an LMAL effect, compared to 29% (20 out of 68) of VO languages. Importantly, the Anti-LMAL preference is much more frequent in VO languages than in OV languages: 34% (23 out of 68) of VO languages show an Anti-LMAL effect compared to 11% (4 out of 38) of OV languages.

In other words, on the one hand, VO languages are significantly more likely to show an RMAL preference than OV languages (83% vs. 63%), while also being less likely to show an Anti-RMAL preference (3% vs. 13%). On the other hand, OV languages are more likely to show an LMAL preference than VO languages (39% vs. 29%) and are also less likely to show an Anti-LMAL preference (11% vs. 34%). This latter observation is the most interesting, since the difference between the two language types is the largest. We observe that VO languages are roughly three times more likely to exhibit the Anti-LMAL preference than OV languages and this difference is statistically significant (Fisher’s Exact test, $p < 0.01$).

These differences are indeed not surprising, given that an asymmetry in length-based effects between head-final/OV and head-initial/VO languages, for directional preferences such as DLM, is well established in the literature. They are also in line with previous observations on the difference between the effect of phrasal length in pre- and postverbal domains (Faghiri and Samvelian, 2020). We return to this point in our concluding remarks.

6.2. MAL compliance in the UD collection

Out of our 131 languages with a MAL effect, 79 show high MAL compliance, 23 show low MAL compliance, and 29 show medium MAL compliance (Table 4). The mean MAL compliance ratios in these three groups are 0.90 (high), 0.13 (low), and 0.50 (middle), respectively. This classification is less strict (and hence less precise), resulting in a smaller number of borderline languages and a greater number of languages at the opposite end of the cline with low MAL compliance, compared to Anti-MAL languages.

In particular, we observe that NDO languages with a notably high number of languages in the grey zone (65%), compared to VO and OV languages (36% and 53%, respectively), based on the MAL effect, show a more balanced distribution

with MAL compliance types: 30% are labeled middle, 45% high and 25% low. Borderline languages are still more frequent in NDO than in VO and OV languages, but the distribution of the three MAL compliance types is closer in NDO to what we observe for VO and OV languages, respectively, 19% (middle), 64.5% (high) and 16.5% (low), and 24% (middle), 60% (high) and 16% (low).

Importantly, the average compliance ratio values suggest that both MAL and Anti-MAL languages in our sample are clearly located at the edges of the cline. Based on these observations, we can expect the number of Anti-MAL languages to increase with a less strict measure and/or classification of the MAL effect.

Language family	MAL compliance			Total
	high	middle	low	
IE	38	16	12	66
Non-IE	41	13	11	65
Total	79	29	23	131

VO/OV type	MAL compliance			Total
	high	middle	low	
NDO	9	6	5	20
OV	23	9	6	38
VO	47	14	12	73
Total	79	29	23	131

Table 4: MAL compliance per language family and VO/OV type.

Note that we observe a similar pattern for RMAL and LMAL, which we do not present here for reasons of space.

6.3. Zooming into languages with unexpected behaviour

It is not possible to present the results for all the languages that contradict a universal MAL hypothesis. All the results are presented in Table 5 in Appendix A. Here, we present some notable cases of Anti-MAL languages, as well as some cases of languages with inconsistent behaviour in the pre- and/or postverbal domains.

One interesting example of an Anti-MAL language is Old East Slavic, which not only has an Anti-MAL effect (that is, the mean length of constituents increases with the number of constituents), but moreover displays an Anti-MAL effect that grows stronger with the number of constituents (Figure 5).

Starting with the preverbal domain, we find 29 Anti-LMAL languages, of which 17 are RMAL – and they are mostly in the grey zone when bilateral MAL is considered. Two of these languages are note-

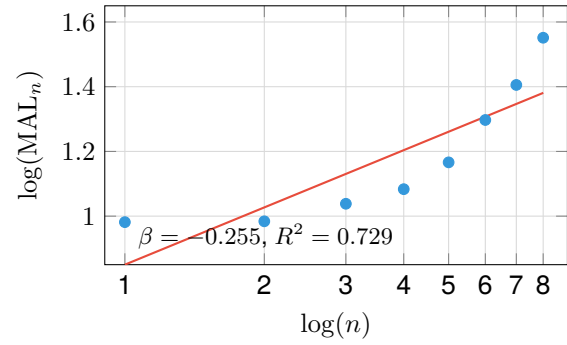


Figure 5: $\lambda_{MAL}(\text{OldEastSlavic})$.

worthy: Occitan (Romance, France) is MAL, while Naija (English-based pidgincreole, Nigeria) is Anti-MAL. Naija is also remarkable in that it is the most strongly Anti-LMAL language in our sample, following a near-perfect power law (Figures 6 and 7).

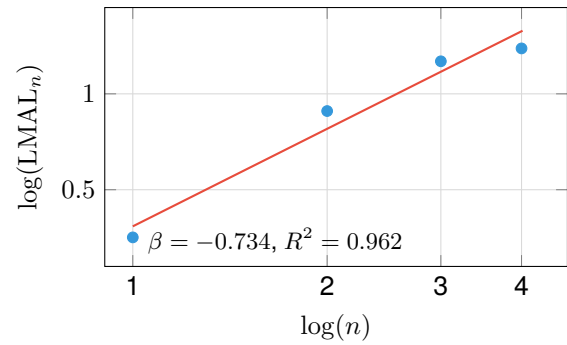


Figure 6: $\lambda_{LMAL}(\text{Naija})$.

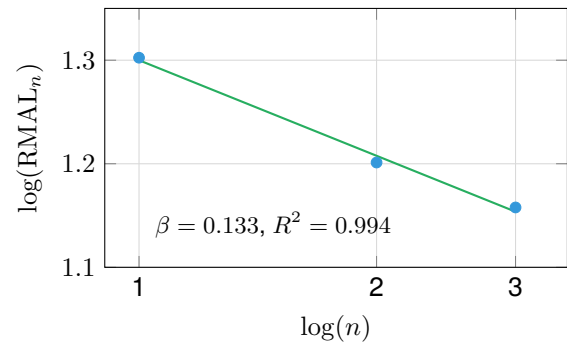


Figure 7: $\lambda_{RMAL}(\text{Naija})$.

As for the postverbal domain, we find the following seven Anti-RMAL languages: Khoekhoe, Bambara, Egyptian, Old East Slavic, Western Armenian, Gothic, and Latin. It is worth noting that among them, four are corpora of ancient languages: Egyptian, Old East Slavic, Gothic, and Latin. Three are both Anti-MAL and Anti-LMAL as well (Bambara, Old East Slavic, and Latin), the others are in the grey zone (for MAL, LMAL or both). In other

words, none of the seven languages shows a MAL or an LMAL preference.

Exploring the genetic and/or areal similarities and differences among these individual languages is beyond the scope of this paper, but it constitutes a necessary avenue for future research in understanding length-based effects across languages and linguistic domains.

7. Conclusion

In this paper, we presented an exhaustive token-based evaluation of MAL in the verbal domain across the UD collection of 180 languages. We proposed a classification based on a metric that we defined in order to estimate the MAL effect for any given language satisfying a minimum threshold of occurrences for relevant configurations. We obtained a reliable classification of 131 languages, well distributed between IE and non-IE languages as well as between VO and OV languages. This classification allowed us to confirm MAL as a typologically widespread preference across the languages of the world, but importantly, it also showed that MAL is not an absolute universal principle. Not only did we find a number of languages in which MAL is absent or trivial, but we also found languages on the opposite end of the cline, displaying a clear MAL dispreference.

The second contribution of our study is the finding that, contrary to what has been previously assumed, MAL does not apply equally on both sides of the verb. MAL is stronger on the right side than on the left side, while Anti-MAL is stronger on the left side. This extends the findings of [Chen et al. \(2022\)](#), who showed the co-effect of MAL and HCS to be universal on the postverbal side, by revealing that the picture is more complex on the preverbal side. This is a crucial finding, as it implies that length-based effects are not necessarily symmetric in the pre- and postverbal domains, contrary to the widespread assumption. Not only do our results show a clear difference between the pre- and postverbal domains, but they also show that VO and OV languages have different preference profiles in these domains. VO languages are more likely to show a strong MAL preference in the postverbal domain, while OV languages are more likely to show a strong MAL preference in the preverbal domain. Importantly, similar observations on the differences between pre- and postverbal domains have previously been made for length-based ordering preferences in individual languages such as Persian ([Faghiri and Samvelian, 2020](#)) or Mandarin ([Yao, 2018](#)). Moreover, there are alternative psycholinguistic production-based accounts of mirror-image word order preferences (SbL and LbS) in VO and OV languages that posit differences between

the two domains. The latter claim that verbs exert strong influence over the constituents that follow them and, accordingly, assume that formal constraints are stricter in the postverbal domain than in the preverbal domain ([Stallings et al., 1998](#)). Our study sheds new light on this ongoing debate.

All detailed results and interactive visualisations are available at <https://typometrics.elizia.net/menzerath/>. The full analysis code is released under a CC BY 4.0 licence at <https://github.com/typometrics/UDW26-Menzerath>.

8. Bibliographical References

- Gabriel Altmann. 1980. Prolegomena to menzerath’s law. In *Glottometrika 2*, pages 1–10, Bochum. Brockmeyer.
- Xinying Chen, Kim Gerdes, Sylvain Kahane, and Marine Courtin. 2022. The co-effect of Menzerath-Altmann law and heavy constituent shift in natural languages.
- Pegah Faghiri and Pollet Samvelian. 2020. [Word order preferences and the effect of phrasal length in SOV languages: evidence from sentence production in Persian](#). *Glossa: a journal of general linguistics*, 5(1):86.
- Ramon Ferrer i Cancho. 2004. [Euclidean distance between syntactically linked words](#). *Phys. Rev. E*, 70:056135.
- Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. [Large-scale evidence of dependency length minimization in 37 languages](#). *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.
- John A. Hawkins. 2007. [Processing typology and why psychologists need to know about it](#). *New Ideas in Psychology*, 25(2):87–107. Modern Approaches to Language.
- Ján Mačutek, Radek Čech, and Jiří Milička. 2017. Menzerath-altmann law in syntactic dependency structure. In *Proceedings of the fourth international conference on dependency linguistics (Depling)*, pages 100–107.
- Paul Menzerath. 1928. Über einige phonetische probleme. In *Actes du premier Congrès international de linguistes*, pages 104–105. Sijthoff Leiden.
- Paul Menzerath. 1954. *Die architektonik des deutschen wortschatzes*. Dümmler: Bonn, Germany.

Lynne M. Stallings, Maryellen C. MacDonald, and Padraig G. O'Seaghdha. 1998. [Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-np shift](#). *Journal of Memory and Language*, 39(3):392–417.

Kumiko Tanaka-Ishii. 2021. Menzerath's law in the syntax of languages compared with random sentences. *Entropy*, 23(6):661.

Yao Yao. 2018. [NP weight effects in word order variation in Mandarin Chinese](#). *Lingua Sinica*, 4(1):5.

9. Language Resource References

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Jephthé Adolphe, Noëmi Aeppli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Arofat Akhundjanova, Furkan Akkurt, Gabrielë Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Antonios Anastasopoulos, Erik Andersen, Kirk Andrews, Matthew Andrews, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arican, Hórunn Arnardóttir, Wirote Aroonmanakun, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Hiwa Asadpour, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, Elena Badmaeva, Jana Bajorat, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Bryan Khelven da Silva Barbosa, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shabnam Behzad, Nathanaël Beiner, Juan Belieni, Alevtina Bémová, Kepa Bengoetxea, İbrahim Benli, Yifat Ben Moshe, Marie Benzerrak, Aleksandrs Berdicevskis, Ansu Berg, Gözde Berk, Delphine Bernhard, Astrid Berntsson Ingelstam, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Esma Fatıma Bilgin Taşdemir, Helin Binici, Kristín Bjarnadóttir, Verena Blaschke, Rogier Blokland, Nina Böbel, Victoria Bobicev, Loïc Boizou, Stavros Bompolas, Johnatan Bonilla, Emanuel Borges Völker, Lars Borin, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, António Branco, Myriam Bras, Théo Brillet, Kristina Brokaitė, Lanni Bu, Eva Buráňová, Aljoscha Burchardt,

Carmen Cabeza, Natalia Cáceres Arandia, Olesea Caftanatov, Marisa Campos, Marie Candito, Caterina Maria Cappello, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Anila Çepani, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Claudine Chamoreau, Shweta Chauhan, Yifei Chen, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Bermet Chontaeva, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Esther Cocco, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Claudia Corbetta, Daniela Corbetta, Francisco Costa, Marine Courtin, Benoît Crabbé, Mihaela Cristescu, Vladimir Cvetkoski, Netanel Dahan, Ingerid Løynning Dale, Sabrina D'Alì, Philemon Daniel, Khensa Daoudi, Bijayalaxmi Dash, Satya Ranjan Dash, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Ahmet Demir, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Roberto Antonio Díaz Hernández, Carly Dickerson, Ariani Di Felippo, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Hoa Do, Kaja Dobrovoljc, Caroline Döhmer, Adrian Doyle, Timothy Dozat, Kira Droganova, Magali Sanches Duran, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Roald Eiselen, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Louise Esher, Soudabeh Eslami, Farah Essaidi, Aline Etienne, Wograiné Evelyn, Sidney Facundes, Richárd Farkas, Ján Faryad, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Barbara Francioni, Theodorus Fransen, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Edith Galy, Federica Gamba, Marcos Garcia, José María García-Miguel, Moa Gärdenfors, Tanja Gaustad, Efe Eren Genç, Fabrício Ferraz Gerardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Gili Goldin, Xavier Gómez Guinovart, Berta González Saavedra, Mathieu Goux, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Mario Guglielmetti, Bruno Guillaume, Kirian Guiller, Céline Guillot-Barbance, Tunga Güngör, Vladimir Gurevich, Nizar Habash, Henrik Hafsteinsson, Michael Hahn, Jan Hajič, Jan Hajič jr., Eva Hajičová, Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudi-stira Hanifmuti, Takahiro Harada, Sam Hard-

wick, Kim Harris, Naïma Hassert, Dag Haug, Jiří Havelka, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Diana Hoefels, Barbara Hoff, Petter Hohle, Nick Howell, Yidi Huang, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Inessa Iliadou, Anton Karl Ingason, Radu Ion, Elena Irimia, Olájiḍé Ishola, Artan Islamaj, Kaoru Ito, Federica Iurescia, Jessica K. Ivani, Sandra Jagodzińska, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Ratanon Jiamsundutsadee, Katharine Jiang, Sylvanus Job, Mayank Jobanputra, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Zhuoxuan Ju, Markus Juutinen, Hüner Kaşıkara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóḡa, Jiří Kárník, Andre Kåsen, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Václava Kettnerová, Lilit Kharatyan, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Petr Kocharov, Arne Köhn, Abdullatif Köksal, Veronika Kolářová, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Nelda Kote, Natalia Kotsyba, Barbara Kovačić, Jolanta Kovalevskaitė, Emmanuelle Kowner, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Lucie Kučová, Adrian Kuqi, Elmurod Kuriyozov, Oğuzhan Kuyrukçu, Asli Kuzgun, Sookyung Kwak, Kris Kyle, Kābi Laan, Veronika Laippala, Lorenzo Lambertino, Israel Landau, Tatiana Lando, Septina Dian Larasati, Pierre Larrivé, Kusum Lata, Alexei Lavrentiev, John Lee, Phng Lê Hồng, Alessandro Lenci, Wei Qi Leong, Saran Lertpradit, Herman Leung, Lori Levin, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Zoey Liu, Nikola Ljubešić, Irina Lobzhanidze, Olga Loginova, Markéta Lopatková, Lucelene Lopes, Edita Luftiu, Arsenii Lukashevskiy, Stefano Lusito, Anne-Marie Lutgen, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Punyanuch Maitreenukul, Ilya Makarchuk, Aibek Makazhanov, Francesco Mambri, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Măărăduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Cláudia Martins, Arianna Masciolini, Jan Mašek, Sanatbek Matlatipov, Hiroshi Matsuda, Yuji Matsumoto, Caterina Mauri, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Maitrey Mehta, Pierre André Ménard, Gustavo Mendonça, Hilla Merhav, Tatiana Merzhevich, Paul Meurer, Niko Miekka, Marie Mikulová, Emilia Mi-

lano, Aleksandra Miletić, Aaron Miller, Junghyun Min, Yael Minerbi, Jiří Mírovský, Karina Mischenkova, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Biswakalpita Mohapatra, AmirHossein Mojiri Froushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskiy, Katerina Mouzou, Kadri Muischnek, Robert Munro, Yugo Murawaki, Nikolett Mus, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhle, Minoo Nassajian, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Běrzkalne, Manuela Nevaci, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Victor Norrman, Alireza Nourian, Michal Novák, Maria das Graças Volpe Nunes, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adédayo Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Annika Ott, Lilja Øvrelid, Masanori Oya, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Petr Pajas, Thomas Palakapilly, Alessio Palmero Aprosio, Jarmila Panevová, Ludovica Pannitto, Anastasia Panova, Thiago Alexandre Salgueiro Pardo, Shantipriya Parida, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Oggi Peeters, Angelika Peljak-Łapińska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Sílvia Pereira, Genel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Eva Pettersson, Andrea Peverelli, Jason Phelan, Claudel Pierre-Louis, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Rodrigo Pintucci, Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Alistair Plum, Thierry Poibeau, Charin Polpanumas, Larisa Ponomareva, Martin Popel, Clamença Poujade, Rangga Prangwedana Prangwedana, Lauma Pretkalniņa, Rigardt Pretorius, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Christoph Purschke, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Ella Rabinovich, Alexandre Rademaker, Mutee-u Rahman, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Joana Ramos, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arij Riabi, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Norton Trevisan Roman, Mykhailo Romanenko, Natalia Romanova,

Rudolf Rosa, Valentin Roşca, Paulette Roulon, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Paolo Ruffolo, Kristján Rúnars-son, Rozana Rushiti, Attapol T. Rutherford, Shoval Sadde, Pegah Safari, Aleksí Sahala, Kalyanamalini Sahoo, Saraswati Sahoo, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Kon-stantinos Sampanis, Stephanie Samson, Xu-
lia Sánchez-Rodríguez, Manuela Sanguinetti, Ezgi Saniyar, Dage Särg, Marta Sartor, Albina Sarymsakova, Mitsuya Sasaki, Baiba Saulīte, Agata Savary, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Em-
manuel Schang, Robert Schikowski, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Sven Sell-
mer, Mojgan Seraji, Magda Ševčíková, Petr Sgall, Syeda Shahzadi, Mo Shen, Atsuko Shi-
mada, Gyu-Ho Shin, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Jean Sibille, Janine Siewert, Einar Freyr Sigurðs-
son, João Silva, Aline Silveira, Natalia Silveira, Sara Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Sí-
monarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Dolores Sollberger, Barbara Sonnenhauser, Shafi Sourov, Nina Speransky, Rachele Sprugnoli, Panyut Sriwirote, Vivian Sta-
mou, Steinhórf Steingrímsson, Antonio Stella, Jan Štěpánek, Barbora Štěpánková, Abishek Stephen, Milan Straka, Omer Strass, Emmett Strickland, Jana Strnadová, Sara Stymne, Alane Suhr, Yogi Lesmana Sulestio, Umut Suluba-
cak, Jack Sun, Hakyung Sung, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Maria Irena Szawerna, Chihiro Taguchi, Dima Taji, Luigi Talamo, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Nursena Teker, Samson Tella, Isabelle Tellier, Marinella Testori, Santhawat Thanyawong, Guil-
laume Thomas, Tarık Emre Tıraş, William Chan-
dra Tjhi, Thea Tollersrud, Kamil Tomaszek, Sara Tonelli, Liisi Torga, Lucas Toribio, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsar-
faty, Kira Tulchynska, Utku Türk, Francis Tyers, Sveinbjörn Hórdarson, Vilhjálmur Hósteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Ure-
šová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Socrates Vakirtzian, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Helena Vaz, Uliana Vedenina, Giu-
lia Venturi, Marianne Vergez-Couret, Annemarie Verkerk, Barbora Vidová Hladká, Eric Villemonte de la Clergerie, Veronika Vincze, Anishka Vis-
samsetty, Natalia Vlasova, Eleni Vligouridou, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail

Walsh, John Wang, Jonathan North Washington, Leonie Weissweiler, Maximilian Wendt, Paul Wid-
mer, Aleksandra Wieczorek, Shira Wigderson, Sri Hartati Wijono, Vanessa Berwanger Wille, Seyi Williams, Miriam Winkler, Shuly Wintner, Mats Wirén, Christian Wittern, Alena Witzlack-
Makarevich, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Qishen Wu, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunx-
iao Yan, Qizhen Yang, Xiulin Yang, Koichi Ya-
suoka, Marat M. Yavrumyan, Arife Betül Yenice, Enes Yilandiloğlu, Olcay Taner Yıldız, Zhuo-
ran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, Annie Zhang, Larry Zhang, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, Rayan Ziane, Artūrs Znotiņš, and Eleonora Zucchini. 2025. [Universal dependen-
cies 2.17](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

A. MAL results for all languages

Table 5: MAL, LMAL and RMAL mini log-log plots and $\beta(1 \rightarrow \infty)$ for all languages. Green = MAL ($\beta > 0.1$), red = Anti-MAL ($\beta < -0.1$), orange = grey zone.

Language	Family	Type	MAL	β	LMAL	β	RMAL	β
Abkhaz	Caucasia	OV		-0.07		+0.07	—	—
Afrikaans	IE	OV		+0.18		+0.19		+0.55
Akkadian	Afroasia	OV		+0.05		+0.09	—	—
Albanian	IE	VO		-0.05		+0.12		+0.47
Amharic	Afroasia	OV		+0.07		+0.30		-0.00
AncientGreek	IE	mix		+0.05		-0.06		+0.01
AncientHebrew	Afroasia	VO		+0.03		-0.13		+0.03
Arabic	Afroasia	VO		+0.16		-0.12		+0.10
Arhuaco	South-Am	OV		+0.13		+0.25	—	—
Armenian	IE	mix		+0.06		-0.05		+0.15
Bambara	Nig-Co.	OV		-0.29		-0.51		-0.20
Basque	Other	OV		+0.06		+0.18		+0.11
Bavarian	IE	mix		+0.13		+0.09		+0.75
Beja	Afroasia	OV		+0.38		+0.40	—	—
Belarusian	IE	VO		+0.18		+0.09		+0.34
Bhojpuri	IE	OV		-0.04		+0.09	—	—
Borôro	South-Am	OV		+0.13		+0.23		+0.10
Breton	IE	VO		+0.12	—	—		+0.20
Buglere	South-Am	OV		-0.01		-0.05	—	—
Bulgarian	IE	VO		+0.17		+0.17		+0.33
Buryat	Other	OV		+0.16		+0.19	—	—
Cantonese	Sino-Aus	VO		-0.05		-0.24		+0.39
CappadocianGreek	IE	VO	—	—	—	—		+0.24
Catalan	IE	VO		+0.26		+0.25		+0.36
Chhintange	Sino-Aus	OV		+0.21		+0.19	—	—
Chinese	Sino-Aus	VO		-0.04		-0.15		+0.53
Chukot	Other	mix		-0.05		+0.01	—	—
ClassArmenian	IE	mix		+0.00		-0.07		+0.01
ClassChinese	Sino-Aus	VO		+0.07		+0.04		+0.01
Coptic	Afroasia	VO		-0.15		-0.49		-0.03
Croatian	IE	VO		+0.27		+0.15		+0.35
Czech	IE	VO		+0.24		+0.22		+0.42
Danish	IE	VO		+0.13		+0.26		+0.31
Dutch	IE	mix		+0.09		+0.15		+0.49

continued on next page

Language	Family	Type	MAL	β	LMAL	β	RMAL	β
Egyptian	Afroasia	VO		+0.08	—	—		-0.20
English	IE	VO		+0.06		-0.30		+0.19
Erzya	Uralic	mix		-0.02		-0.00		+0.27
Estonian	Uralic	VO		+0.21		+0.18		+0.45
Faroese	IE	VO		-0.08		-0.01		+0.02
Finnish	Uralic	VO		+0.11		+0.06		+0.38
French	IE	VO		+0.25		+0.01		+0.32
FrisianDutch	IE	mix		+0.33	—	—	—	—
Gaelic	IE	VO		+0.35	—	—		+0.24
Galician	IE	VO		+0.39		+0.24		+0.50
Gbaya	Nig-Co.	VO	—	—	—	—		+0.17
Georgian	Caucasia	mix		+0.15		-0.04		+0.44
German	IE	mix		+0.17		+0.20		+0.54
Gheg	IE	VO		+0.11		-0.04		+0.31
Gothic	IE	VO		-0.12		-0.07		-0.12
Greek	IE	VO		+0.21		-0.04		+0.32
Guajajára	South-Am	VO		+0.08	—	—		+0.18
Haitian	IE	VO		+0.19		-0.29		+0.30
Hausa	Afroasia	VO		-0.16		-0.40		+0.07
Hebrew	Afroasia	VO		+0.19		+0.20		+0.32
HighlandNahuatl	South-Am	VO		+0.23		+0.10		+0.43
Hindi	IE	OV		+0.10		+0.03		+1.03
Hungarian	Uralic	mix		+0.05		+0.08		+0.24
Icelandic	IE	VO		-0.03		-0.17		+0.22
Indonesian	Sino-Aus	VO		+0.11		-0.07		+0.31
Irish	IE	VO		+0.05		-0.31		+0.15
Italian	IE	VO		+0.15		+0.00		+0.22
Japanese	Other	OV		+0.03		+0.02	—	—
Javanese	Sino-Aus	VO		+0.22		-0.00		+0.21
Karo	South-Am	OV		+0.05		+0.07	—	—
Kazakh	Turkic	OV		+0.06		+0.07	—	—
Khoekhoe	Other	OV		-0.04		+0.07		-0.37
Kiche	South-Am	VO		+0.05	—	—		+0.21
Komi	Uralic	mix		-0.04		+0.01		+0.20
Korean	Other	OV		+0.13		+0.12		-0.06
Kurmanji	IE	OV		+0.14		-0.13	—	—
Kyrgyz	Turkic	OV		+0.15		+0.15	—	—
Latin	IE	mix		-0.23		-0.14		-0.10
Latvian	IE	VO		+0.11		-0.02		+0.30

continued on next page

Language	Family	Type	MAL	β	LMAL	β	RMAL	β
Ligurian	IE	VO		-0.07		+0.05		+0.55
Lithuanian	IE	VO		+0.15		-0.08		+0.38
LowSaxon	IE	mix		+0.10		+0.14		+0.51
Maghrebi	Afroasia	VO		+0.12		+0.15		+0.17
Maltese	Afroasia	VO		+0.30		-0.07		+0.41
Manx	IE	VO	—	—	—	—		+0.39
Marathi	IE	OV		+0.12		+0.10	—	—
MbyáGuaraní	South-Am	mix		-0.01		+0.00	—	—
MiddleFrench	IE	VO		+0.12		-0.29		+0.15
Moksha	Uralic	VO		+0.07		+0.06	—	—
Naija	IE	VO		-0.25		-0.73		+0.13
Nhengatu	South-Am	VO		+0.03		+0.14		+0.10
NorthSami	Uralic	VO		+0.03		-0.00		+0.19
Norwegian	IE	VO		+0.22		+0.18		+0.39
Occitan	IE	VO		+0.21		-0.15		+0.33
OldChurchSlavonic	IE	VO		-0.13		-0.17		-0.08
OldEastSlavic	IE	mix		-0.26		-0.21		-0.16
OldFrench	IE	VO		-0.05		-0.18		+0.15
OldProvençal	IE	VO		-0.20		-0.44		-0.01
OttomanTurkish	Turkic	OV		+0.06		+0.03	—	—
Persian	IE	OV		-0.01		-0.09		+0.68
Pesh	Other	OV		+0.10		-0.04	—	—
Polish	IE	VO		+0.14		+0.12		+0.26
Pomak	IE	VO		+0.05		+0.10		+0.19
Portuguese	IE	VO		+0.13		+0.13		+0.17
Romanian	IE	VO		+0.09		-0.11		+0.21
Russian	IE	VO		+0.11		+0.07		+0.32
Sanskrit	IE	OV		-0.02		+0.03		+0.42
Serbian	IE	VO		+0.17		+0.23		+0.37
Sicilian	IE	VO		+0.32		-0.08		+0.36
Sindhi	IE	OV		-0.05		+0.05		+0.84
SkoltSami	Uralic	mix		+0.03		-0.10	—	—
Slovak	IE	VO		+0.14		+0.24		+0.34
Slovenian	IE	VO		+0.23		+0.06		+0.42
Spanish	IE	VO		+0.21		+0.22		+0.26
Swedish	IE	VO		+0.11		+0.10		+0.35
SwissGerman	IE	OV		-0.00		+0.17		+1.01
Tamil	Dravid.	OV		+0.35		+0.38	—	—
Tangkhul	Sino-Aus	OV		+0.11		+0.09	—	—

continued on next page

Language	Family	Type	MAL	β	LMAL	β	RMAL	β
Tatar	Turkic	OV		-0.08		-0.13	—	—
Teko	South-Am	mix		+0.10		+0.13	—	—
Telugu	Dravid.	OV		+0.09		+0.09	—	—
Thai	Sino-Aus	VO		+0.09		-0.22		+0.30
Turkish	Turkic	OV		+0.14		+0.14		-0.06
TurkishGerman	Turkic	OV		+0.05		+0.13		+0.40
Ukrainian	IE	VO		+0.19		-0.04		+0.32
UpperSorbian	IE	VO		+0.50		+0.50		+0.72
Urdu	IE	OV		+0.14		+0.05		+0.76
Uyghur	Turkic	OV		+0.02		+0.04	—	—
Uzbek	Turkic	OV		+0.14		+0.16	—	—
Vietnamese	Sino-Aus	VO		+0.01		-0.20		+0.24
Welsh	IE	VO		+0.63		-0.02		+0.70
WestArmenian	IE	mix		-0.03		-0.10		+0.17
WestNahuatl	South-Am	VO		+0.04		+0.03		-0.13
Wolof	Nig-Co.	VO		+0.06		-0.32		+0.22
Wu	Sino-Aus	VO		-0.00		-0.05	—	—
Xibe	Other	OV		-0.13		-0.13	—	—
Yiddish	IE	VO		+0.04		-0.10		+0.29
Yoruba	Nig-Co.	VO		+0.34		-0.21		+0.49
Yupik	Other	mix		-0.05	—	—	—	—
Zaar	Afroasia	VO		+0.01		-0.28		+0.35

Comparing Dependency Distances of Esperanto and Other Languages in a Multi-Lingual Parallel Corpus

Masanori Oya

Meiji University

Tokyo

masanori_oya2019@meiji.ac.jp

Abstract

This study attempts to integrate Esperanto into the research of dependency distance, based on a small-scale multi-lingual parallel corpus of Manifesto de Prago, annotated with UD relations, to contribute to the development of research on Esperanto as a natural language. The mean dependency distance and the distribution of dependency distances of Esperanto are not significantly different from the majority of the 21 languages in the corpus, thus partially supporting the claim that Esperanto is no less natural than other natural languages.

Keywords: Esperanto, Dependency Distance, Multi-Lingual Parallel Corpus

1. Introduction

This study attempts to integrate Esperanto into the research of *Dependency Distance*, based on a small-scale multi-lingual parallel corpus, which has been constructed for this unique purpose.

Despite the origin of Esperanto as a constructed language (Zamenhof, 1887), several researchers today investigated its linguistics characteristics as one of the natural languages (Gledhill, 2000; Goodall, 2023; Liu, 2011; Koutny, 2015, among others) which can be the topic of theoretical linguistics, because it shares some linguistic characteristics with other natural languages. In this context, Oya (2025) has introduced a small-scale *Universal Dependencies* (UD) treebank based on the Esperanto version of *Manifesto de Prago* (Fettes, 1996) (MdP), an official document published online, advocating the status of Esperanto as an international auxiliary language. That treebank is intended to contribute to the development of research on Esperanto as a natural language.

If Esperanto is to be considered as one of the natural languages as argued by some researchers mentioned above, then we need to demonstrate that (1) it does not show any extraordinary trait which cannot be found in other natural languages, and that (2) it shares some common characteristics with other natural languages. In order to do so, we need to employ some metrics to compare and contrast Esperanto to other natural languages so that such investigations can be as objective and replicable as possible. In this context, this study focuses on the Dependency Distance of Esperanto and other languages as such a metric, based on the text data provided in MdP as a multi-lingual parallel corpus.

This article is organized as follows: Section 2 summarizes the concept of Dependency Distance,

with a focus on the idea of *Mean Dependency Distance* (Liu, 2007, 2008, etc.), followed by the rationale of this study. Section 3 describes how the raw text data available online were parsed in the format of UD for this study, and how these data are analyzed in terms of their Dependency Distances. Section 4 reports the results of the analysis, which are further discussed from several viewpoints in Section 5, and Section 6 concludes this article.

2. Dependency Distance

Dependency Distance (DD) is the number of words from a word in a sentence to the word on which it depends (Hudson, 1995; Liu, 2007, 2008, among others). Liu (2008) states the definition of DD as follows:

“let $W_1...W_i...W_n$ be a word string. For any dependency relation between the words W_a and W_b , if W_a is a governor and W_b is its dependent, then the *dependency distance* (DD) between them can be defined as the difference $a - b$; ... in measuring dependency distance the relevant measure is the absolute value of dependency distance.”

For example, in a sentence *Sarah read 30 books for her Ph.D dissertation*, the noun *Sarah* depends on the verb *read*, and its DD is one; the numeral *30* depends on the noun *book*, and its DD is one; the noun *books* depends on the verb *read*, and its DD is two, etc.

DD has attracted considerable scholarly attention as a key metric of syntactic complexity. A number of studies further contend that DD captures certain universal properties of natural languages (Ouyang and Jiang, 2018; Ouyang et al., 2022;

Wang and Liu, 2017; Yan and Li, 2019, among others). Moreover, cross-linguistic research has proposed a general preference for shorter dependency distances, attributed to constraints on short-term memory. This is an effect commonly described as Dependency Distance Minimization (Gibson, 2000; Gildea and Temperley, 2010; Jiang and Ouyang, 2018; Niu et al., 2023; Temperley, 2007; Temperley, 2008, among others).

In this context of research on dependency distance and its minimization, mean dependency distance (MDD) is defined as follows (Liu, 2008), where n means the word count of a sentence and DD_i means the dependency distance between the i -th word and its head:

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

Liu (2008) assumed “that the greater the MDD of a sentence, the more difficult the sentence” and “the greater the MDD of a text, the more difficult the text (or language)”, and conducted a corpus-based research of the MDDs of 20 languages (Esperanto is not included), and revealed that all MDDs of these languages fall below 3.662.

Oya (2024) also conducted a research on multi-lingual parallel corpus to calculate the DDs of a variety of languages (Esperanto is not included) to examine whether the distributions of DDs are different across different languages. The results show that the majority of the language pairs show similar distributions of DDs, though their MDDs are diverse within a certain threshold. This suggests that, while MDDs may reflect language-dependent characteristics across different languages within a certain threshold, the distributions of DDs are stable across them.

This study is an extension of Oya (2024), employing MdP mentioned above as a multi-lingual parallel corpus. MdP is actually provided not only in Esperanto, but in a variety of languages, and therefore it can serve as a multi-lingual parallel corpus when annotated in a principled way, such as UD. At present, the Esperanto version of MdP has been annotated with UD (Oya, 2025) as mentioned above. This study introduces UD annotation on MdP in 22 languages, including Esperanto, and examines whether the MDD and the distribution of DDs in Esperanto are similar to those of these 21 languages. If it is found the case, then that will partially support the claim that Esperanto share a certain trait with other natural languages as far as the distribution of DDs is concerned.

This study is also an extension of Liu (2011), which conducted analyses of lexical and syntactic features of an Esperanto book (Zamenhof, 1907). Along with other lexical and syntactic properties which strongly suggest that Esperanto is a struc-

turally natural language, it revealed that the MDD of the book is 3.85, which is longer than MDDs of other languages mentioned in Liu (2008). If we follow the logic of Liu (2008), that is, the greater the MDD of a text, the more difficult the text (or language), this result could lead us to conclude that Esperanto is more difficult than other languages, and it would be in contrast to the claim that Esperanto is no less natural than other languages, as far as relative difficulty of a language (represented by a longer MDD) is considered to represent its relative lack of naturalness. However, the comparison of MDDs across Liu (2008) and Liu (2011) is not sound because they analyze different corpus data, and therefore the contents of the text are not controlled for syntactic analyses across them.

This insight on the necessity of controlling the content of the text data motivates to investigate the MDD of Esperanto along with those of other languages using a multi-lingual parallel corpus as the primary data, in which the texts share the same content across different languages, and therefore we can focus on the difference or similarity of their syntactic properties.

Against the background described above, this study attempts to answer the following question: Are the MDD and the distribution of DDs in Esperanto similar to those of the other languages in a multi-lingual parallel corpus?

3. This study

3.1. Data

The corpus data used in this study is Esperanto MdP with UD annotations, and its multi-lingual versions available online (<https://lingvo.org/prago>). It is a manifesto which describes seven principles of Esperanto movement to advocate it as an international auxiliary language, published in the 81st World Esperanto Congress in Prague, Czech Republic in 1996. It is translated into 50 languages.

For example, the sentence (1) below is quoted from MdP followed by its English translation, its UD tree is Figure 1, and its UD analysis is shown in Table 1:

- (1) *Ni asertas, ke lingva malegaleco sekvigas komunikan malegalecon je ĉiuj niveloj, inkluzive de la internacia nivelo.*

(We maintain that language inequality gives rise to communicative inequality at all levels, including the international level.)

3.2. Analysis

The Esperanto version of MdP has been manually annotated with UD tags in the format of

# sent_id = prago-009									
# text = Ni asertas ke lingva malegaleco sekvigas komunikan malegalecon je ĉiuj niveloj, inkluzive de la internacia nivelo.									
# text_en = We maintain that language inequality gives rise to communicative inequality at all levels, including the international level.									
1	Ni	ni	PRON	—	Case=Nom Number=Plur Person=1 PronType=Prs	2	nsubj		
2	asertas	aserti	VERB	—	Mood=Ind Tense=Pres VerbForm=Fin	0	root		
3	,	,	PUNCT	—		7	punct		
4	ke	ke	SCONJ	—		7	mark		
5	lingva	lingva	ADJ	—	Degree=Pos	6	amod		
6	malegaleco	malegaleco	NOUN	—	Case=Nom Number=Sing	7	nsubj		
7	sekvigas	sekvigi	VERB	—	Mood=Ind Tense=Pres VerbForm=Fin	2	ccomp		
8	komunikan	komunika	ADJ	—	Case=Acc Degree=Pos Number=Sing	9	amod		
9	malegalecon	malegaleco	NOUN	—	Case=Acc Number=Sing	7	obj		
10	je	je	ADP	—		12	case		
11	ĉiujn	ĉiu	DET	—	Case=Acc Number=Plur PronType=Tot	12	det		
12	nivelojn	nivelo	NOUN	—	Case=Acc Number=Plur	7	obl		
13	,	,	PUNCT	—		14	punct		
14	inkluzive	inkluzive	ADV	—		7	advmod		
15	de	de	ADP	—		18	case		
16	la	la	DET	—	Definite=Def PronType=Art	18	det		
17	internacia	internacia	ADJ	—	Case=Nom Degree=Pos Number=Sing	18	amod		
18	nivelo	nivelo	NOUN	—	Case=Nom Number=Sing	14	nmod		
19	.	.	PUNCT	—		2	punct		

Table 1: The UD annotations of the sentence (1); the fields *DEPS* and *MISC* are excluded

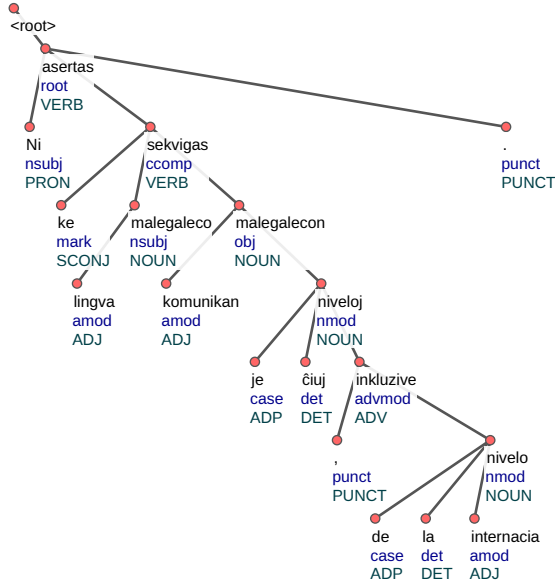


Figure 1: The UD tree for the sentence (1)

CoNNL-U and made public at the Web page of Universal Dependencies (Oya, 2025). Among these 50 languages in MdP, this study has chosen 20 languages, and parsed them using an original Python script which uses *SpaCy* for dependency parsing (I parsed the Hungarian version using *HuSpaCy*). The parsed output of these 20 languages has been made public at Github (<https://github.com/MasanoriOya/Manifesto-de-Prago>).

The DDs of the dependency relations in each language are extracted from the treebank, as the absolute value of the difference between the governor and its dependent, and its MDD is calculated using the formula proposed by Liu (2008). This process is repeated for all the 21 languages in the parallel corpus. Then these MDDs are compared and contrasted with each other. The differences

of the distribution of these DDs are statistically examined using a Kruskal-Wallis test. This test is chosen because the distribution of the DDs of a language is not normal. This means that we cannot use an ANOVA, which presupposes the normal distributions of the data.

4. Results

The descriptive statistics of the DDs of the 21 languages in MdP is shown in Table 2:

Multi-lingual comparisons of MDDs show us variances across these languages, which do not necessarily reflect their typological categorizations. In other words, languages which belong to the same branch are not necessarily close to each other in terms of their MDDs. The MDD of Esperanto is 3.409, which falls in the middle of the average dependency distances of these 21 languages, and it is also shorter than the result of Liu (2011) (3.85).

A Kruskal-Wallis test showed that the DDs significantly differed across these 21 languages, ($H(20) = 185.27, p < .01$). A post-hoc Steel-Dwass test revealed that the distribution of the DDs of Esperanto was significantly different from those of Catalan, English, Japanese and Spanish, but no significant difference was found between the distribution of DDs of Esperanto and those of other 16 languages.

5. Discussions

The result suggests that Esperanto is not conspicuously different from the majority of the languages in the MdP, as far as the MDD and the distribution of DDs is concerned. This fact may constitute a partial yet significant support of the claim that Esperanto is no less natural than other natural languages. This insight is available only from the viewpoint of Universal Dependencies and the DDs

	eo	ca	da	de	el	en
mean	3.41	3.29	2.99	3.72	3.08	2.85
median	2	2	2	1	2	1
mode	1	1	1	1	1	1
SD	4.60	4.99	4.31	5.39	4.57	4.36
token	839	1006	853	890	924	907

	es	fi	fr	hr	hu	it
mean	3.10	2.98	3.42	3.41	3.81	3.67
median	2	1	2	2	2	2
mode	1	1	1	1	1	1
SD	4.88	4.25	4.94	4.47	5.32	6.05
token	973	686	1055	785	913	1025

	ja	lt	nl	pl	pt	ro
mean	3.92	3.37	4.14	3.26	3.52	3.56
median	2	2	2	2	2	2
mode	1	1	1	1	1	1
SD	8.47	4.43	5.59	4.54	5.48	6.70
token	1432	737	926	794	963	864

	ru	sl	sv
mean	3.46	3.23	3.65
median	2	2	2
mode	1	1	1
SD	4.61	4.37	5.34
token	786	781	899

Table 2: The descriptive statistics of the dependency distances of the 21 languages: eo, Esperanto; ca, Catalan; da, Danish; de, German; el, Greek; en, English; es, Spanish; fi, Finnish; fr, French; hr, Croatian; hu, Hungarian; it, Italian; ja, Japanese; lt, Lithuanian; nl, Dutch; pl, Polish; pt, Portuguese; ro, Romanian; ru, Russian; sl, Slovenian; sv, Swedish

of different languages. It will be interesting to investigate whether Esperanto was actually less natural than other natural languages in the beginning yet it has come to have a characteristics which is similar to natural languages at present. For example, the longer MDD of Esperanto in the result of Liu (2011) may reflect the fact that its text data was Zamenhof (1907), which was published more than 100 years ago, while the result of this study is based on the parallel texts of MdP published in 1990s. This issue can be addressed by investigating the chronological change of Esperanto, based on larger-scale treebanks of Esperanto, which are not yet available today. Therefore, we need to construct them for that purpose along with others. This is one of the questions to be answered in future research.

Another issue of interest related to Esperanto is the difference between Esperanto and other constructed languages, such as Ido or Interlingua. If it is found that these are less natural than Esperanto in terms of the distribution of DDs, then it may serve as a partial support that Esperanto has come to be used more widely than these other constructed languages due to its naturalness. If, on the other

hand, it is found that they are as natural as other languages like Esperanto, then it may support the claim that DDs represent a certain universal characteristics of language as we know it, whether it is natural or constructed. Again, we need large-scale corpora of these constructed languages for comparing and contrasting them to Esperanto, which is another question to be addressed in future.

Along with these issues related to Esperanto, this study collaterally found that some languages in MdP show significant differences from others. For example, the distribution of DDs of English is significantly different from those of Croatian, Dutch, Finnish, Hungarian, Italian, and Swedish, and the distribution of DDs of Dutch is significantly different from those of Catalan, English, Japanese, Polish, Portuguese, and Romanian. We need to explain why they are different from others, and that explanation may be applied to Esperanto, thus contributing to the understanding of why Esperanto is different from some natural languages as mentioned above.

Overall, this study only focuses on the distributions of DDs of Esperanto and other natural languages in a small-scale multi-lingual parallel corpus, and we need to address its limitations, such that (1) we need to employ other metrics for sentence complexity, and (2) we need larger-scale multi-lingual parallel corpus. As for (1), there are several metrics for it, such as the average sentence lengths of languages or the centrality of dependency structures. As for (2), *Parallel Universal Dependencies* (PUD) would be a good candidate; by constructing an Esperanto part of PUD and employing it for further research, we will deepen our understanding of the characteristics of this language which can be both unique and universal, just like other natural languages.

6. Conclusion

This study attempted to integrate Esperanto into the research of dependency distance, based on a small-scale multi-lingual parallel corpus of Manifesto de Prago with UD relations, to contribute to the development of research on Esperanto as a natural language. The distribution of dependency distances of Esperanto is not significantly different from the majority of the 21 languages in the corpus, thus partially supporting the claim that Esperanto is no less natural than other natural languages.

7. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP24K04089.

8. Bibliographical References

- Yude Bi and Hua Tan. 2024. Language transfer in L2 academic writing: a dependency grammar approach. *Frontiers in Psychology*, 15:1–14.
- Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.
- Yu Fang and Haitao Liu. 2018. What factors are associated with dependency distances to ensure easy comprehension? a case study of ba sentences in mandarin chinese. *Language Sciences*, 67:33–45.
- Mark Fettes. 1996. *Manifesto de Prago*. Universala Esperanta Asocio.
- Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence for dependency length minimization in 37 languages. *Proceedings of Natural Academy of Science*, 112(33):10336–10341.
- Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.
- Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In W. O’Neil A.P. Marantz, A.P. Miyashita, editor, *Image, language, brain: Papers from the first mind articulation project symposium*, pages 95–126. MIT Press.
- Daniel Gildea and David Temperley. 2007. Optimizing grammars for minimum dependency length. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 184–191.
- Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.
- Chris Gledhill. 2000. *The Grammar of Esperanto: A Corpus-based description*. Lincom.
- Grant Goodall. 2023. Esperanto kaj lingvistiko: cent jaroj da (mal)amikeco. *Esperantologio / Esperanto Studies*, 4(12).
- Daniel Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistics input for sentential complexity. *Cognitive Science*, 29(2):261–290.
- Richard Hudson. 1995. [Measuring syntactic difficulty](#).
- Jingyan Jiang and Haitao Liu. 2015. The effects of sentence length on dependency distance, dependency direction and the implications—based on a parallel english-chinese dependency treebank. *Language Sciences*, 50:93–104.
- Jingyang Jiang and Jinghui Ouyang. 2018. Minimization and probability distribution of dependency distance in the process of second language acquisition. In Jingyang Jiang and Haitao Liu, editors, *Quantitative Analysis of Dependency Structures*, pages 167–190. De Gruyter Mouton, Berlin, Boston.
- Saeko Komori, Masatoshi Sugiura, and Wenping Li. 2026. Changes in syntactic complexity indices with the language development of japanese as a second language: A longitudinal japanese learner corpus study. *Journal of Quantitative Linguistics*, 33(1):64–79.
- Ilona Koutny. 2015. A typological description of esperanto as a natural language. *Język. Komunikacja. Informacja*, 10:43–62.
- Wenping Li and Jianwei Yan. 2021. Probability distribution of dependency distance based on a treebank of japanese efl learners’ interlanguage. *Journal of Quantitative Linguistics*, 28(2):172–186.
- Haitao Liu. 2007. Probability distribution of dependency distance. *Glottometrics*, 15(1):1–12.
- Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.
- Haitao Liu. 2009. Probability distribution of dependencies based on chinese dependency treebank. *Journal of Quantitative Linguistics*, 16(3):256–273.
- Haitao Liu. 2011. Quantitative analysis of Zamenhof’s *Esenco kaj estonteco*. *Language Problems & Language Planning*, 35(1):57–81.
- Haitao Liu, Chunshan Xu, and Junyin Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193.
- 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*, pages 102–108.
- Ruochen Niu, Yaquin Wang, and Haitao Liu. 2023. The cross-linguistic variations in dependency distance minimization and its potential explanations.

- In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pages 4034–4043, Marseille, France.
- Jinghui Ouyang and Jingyang Jiang. 2018. Can the probability distribution of dependency distance measure language proficiency of second language learners? *Journal of Quantitative Linguistics*, 25(4):295–313.
- Jinghui Ouyang, Jingyang Jiang, and Haitao Liu. 2022. Dependency distance measures in assessing L2 writing proficiency. *Assessing Writing*, 51:100603.
- Masanori Oya. 2024. Cross-linguistic variances of dependency distances in multi-lingual parallel corpus. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 1166–1171.
- Masanori Oya. 2025. UD treebanks for esperanto as a natural language. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 22–29.
- Mikael Parkvall. 2000. How European is Esperanto?: A typological study. *Language Problems and Language Planning*, 34(1):63–79.
- David Temperley. 2007. Minimization of dependency length in written english. *Cognition*, 105(2):300–333.
- David Temperley. 2008. Dependency length minimization in natural and artificial languages. *Journal of Quantitative Linguistics*, 15(3):256–282.
- Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Klincksieck, Paris.
- Yaqin Wang and Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Science*, 59:135–147.
- Hengbin Yan and Yanghui Li. 2019. Investigating dependency distance across l2 modalities and proficiency levels. *Open Linguistics*, 5(1):601–614.
- Lingyu Yi and Zhongqing He. 2026. Syntactic complexity in knowledge construction across disciplines: Evidence from dependency distance. *Journal of English for Academic Purposes*, 79:1475–1585.
- Ludoviko Lazaro Zamenhof. 1887. *Dr. Esperanto's International Tongue*. Chaim Kelter.
- Ludoviko Lazaro Zamenhof. 1907. *Esenco kaj Estonteco de la Ideo de Lingvo Internacia*. Hachette.
- 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 Gökırmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 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.

Negation of Turkic non-verbal clauses: Analysis and Universal Dependencies Implementation

Nikolett Mus¹, Furkan Akkurt², Bermet Chontaeva³, Soudabeh Eslami³,
Sardana Ivanova⁴, Çağrı Çöltekin³, Jonathan Washington⁵,
Gulnura Dzhumaliev⁶, Aida Kasieva⁶

¹ELTE Research Centre for Linguistics, ²Boğaziçi University, ³University of Tübingen,

⁴Independent Researcher, ⁵Swarthmore College, ⁶Kyrgyz-Turkish Manas University

mus.nikolett@nytud.elte.hu, furkan.akkurt@bogazici.edu.tr,

bermet.chontaeva@student.uni-tuebingen.de, soudabeh.eslami@student.uni-tuebingen.de,

sardana.n.ivanova@gmail.com, cagri.coeltekin@uni-tuebingen.de,

jonathan.washington@swarthmore.edu, gulnur.jumaliev@manas.edu.kg, aida.kasieva@manas.edu.kg

Abstract

The paper examines the grammatical behavior of the negative element used to negate predicates in non-verbal clauses in three Turkic languages: Azerbaijani, Kyrgyz, and Turkish. We focus on its interaction with verbal copulas, subject agreement, and the distribution of agreement suffixes, as well as its position within the predicate phrase. The study draws on both previously described corpus data and newly collected examples. Across all three languages, agreement features are realised on the negative element only in the absence of an overt copula. The agreement morphology involved is identical to that found with nominal, adjectival, and adverbial predicates. In all the languages examined, the negative element remains within the predicate phrase; thus, its position is syntactically constrained. At the same time, we observe differences among the languages in the degree to which the position of the negator is fixed within the predicate. In Turkish and Azerbaijani, regardless of which element of the nominal predicate it negates, the negator invariably follows the predicate. In Kyrgyz, by contrast, it consistently appears immediately after the element it negates within the predicate. These patterns suggest that the negative element behaves syntactically as a phrasal operator associated with non-verbal predicates. For annotation purposes, we therefore propose analysing the negative element as a negation modifier, assigning it the POS tag `ADV` and the dependency relation `advmod:neg`.

Keywords: non-verbal negation, non-verbal negation element, Turkic languages, Universal Dependencies (UD)

1. Introduction

This paper investigates a negative element used in the negation of non-verbal clauses in several Turkic languages, examining its morphosyntactic behavior, distribution, and interpretive properties.

The empirical focus is on Kyrgyz, Southern Azerbaijani, and Turkish, which represent two distinct branches of the Turkic family – Oghuz (Azerbaijani and Turkish) and Kipchak (Kyrgyz) – thereby enabling cross-branch comparison. Although the analysis is developed on the basis of these languages, it may extend to other Turkic languages in which this type of negation is attested.

At the same time, some Turkic languages – such as Sakha (Lena Turkic) – employ a different strategy for non-verbal negation, namely a copular verb marked with a negation morpheme, and lack an element comparable to the one examined here.

Our primary goal is to provide a detailed descriptive account of this construction, laying the groundwork for a unified analysis that can inform and improve Universal Dependencies (UD) treebanks (De Marneffe et al., 2021; Akhundjanova et al., 2025).

While existing UD analyses often treat this element as an auxiliary or a copular verb, our observations indicate that such a characterization is

not adequate.¹ We show that an overt verbal copula can co-occur with the negative element, which argues against analyzing it as a copula or auxiliary head. In contexts where no overt copula is present, agreement morphology surfaces on the negative element itself. The agreement suffixes are formally identical to those used with nominal, adjectival, and adverbial predicates, suggesting that in copula-less environments the negative element occupies a predicate-related position within the clause.

Patterns of combinability with copulas, agreement behavior, and positional restrictions indicate that the negative element functions as a phrasal operator, taking scope over the constituent it negates. Accordingly, within the UD framework, we propose that it should be assigned the part-of-speech tag `ADV` and the dependency relation `advmod:neg`, providing a linguistically motivated basis for more accurate annotation in Turkic treebanks.

2. Background

In Turkic languages, multiple types of negative elements and structures exist, often occurring in complementary distribution. Some negative elements

¹All treebanks examined are part of UD v2.17.

appear exclusively with verbs, such as the Turkish verbal negation suffix *-ma*; e.g. *oku-ma-dı* ‘(s/he) did not read’ (Kornfilt, 2013), while others are restricted to clauses with non-verbal predicates, such as *değil* in Turkish (Lee, 1996; Robbeets and Saveyev, 2020; Kornfilt, 2013). However, recent studies suggest that the distinction is not entirely clear-cut, with some overlap between these structures and/or elements (Seydi, 2020). In the present study, we set aside these overlaps to focus specifically on the negative element used in the negation of non-verbal clauses. This construction is illustrated by means of minimal pairs (affirmative vs. negative) in Azerbaijani (1a)–(1b),² Kyrgyz (2a)–(2b), and Turkish (3a)–(3b).³

- (1) a. Deniz döktürdür.
Deniz doctor-COP;[3SG]
‘Deniz is a doctor.’
b. Deniz döktür **dəyil**.
Deniz doctor NEG[3SG]
‘Deniz is not a doctor.’
- (2) a. Дениз доктор.
Deniz doktor.
Deniz doctor[3SG]
‘Deniz is a doctor.’
b. Дениз доктор **эмес**.
Deniz doktor **emes**.
Deniz doctor NEG[3SG]
‘Deniz is not a doctor.’
- (3) a. Deniz doktor.
Deniz doctor[3SG]
‘Deniz is a doctor.’
b. Deniz doktor **değil**.
Deniz doctor NEG[3SG]
‘Deniz is not a doctor.’

As already mentioned, this element is not attested in all Turkic languages; for example, in Sakha (Lena Turkic) the same meaning appears to be expressed by attaching a negative morpheme to the verbal copula (Ubryatova et al., 1982; Sleptsov, 2018), as illustrated in (4).⁴

- (4) Дениз доктор буол-батах.
Deniz doktor **buol-batax**.
Deniz doctor be-NEG;PRES[3SG]
‘Deniz is not a doctor.’

What is particularly striking about the negative element attested in Azerbaijani, Kyrgyz, and Turkish is that it appears, at least superficially, to behave like a verb, or more specifically, like a verbal copula, since it can bear subject agreement suffixes, see (5). In Turkic languages, subject agreement is canonically realized on the finite predicate in clauses.

- (5) a. Мән дөктүр **dəyil-əm**.
1SG doctor NEG-1SG
‘I am not a doctor.’
b. Мен доктор **эмесмин**.
Men doktor **emes-min**.
1SG doctor NEG-1SG
‘I am not a doctor.’
c. Ben doktor **değil-im**.
1SG doctor NEG-1SG
‘I am not a doctor.’

It therefore appears consistent with previous analyses, including within frameworks such as Universal Dependencies (UD), that this negative element is often analyzed as an auxiliary, i.e. a copula in these sentences (see, for instance, Turkish BOUN (Marşan et al., 2022) or Turkish TueCL (Akkurt and Çöltekin, 2025) UD Treebanks.) Such an analysis is illustrated in Figure 1.

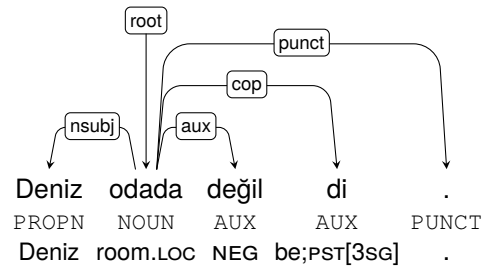


Figure 1: An existing UD analysis of *değil* as AUX with *aux* relation. The meaning of the sentence is: ‘Deniz was not in the room.’

However, we do not adopt this interpretation, as our observations reveal grammatical behaviors that are incompatible with either an auxiliary or a copular analysis, including agreement patterns and positional restrictions.

3. Grammatical behavior of the negative element

The starting point of our argumentation is that the omission of the copula in negative constructions is

²Although the Azerbaijani examples are drawn from the Southern (Iranian) variety, they are transcribed using the Latin alphabet according to the orthographic conventions of North Azerbaijani.

³In addition to its use in non-verbal predication, the negative element discussed here is also attested in constituent negation in these languages (Kornfilt, 2013); this use falls outside the scope of the present study.

⁴Historically this is a negative copula form, just like *emes*, but it also presents some challenges for synchronic analysis as such. Further work is required to understand whether it might also constitute a negative particle of the type being discussed in this paper.

not obligatory but is constrained by general properties of non-verbal predication. In particular, in the present tense, non-verbal predicates do not require an overt copula, and the negative element appears without one accordingly. By contrast, in contexts where the non-negated counterpart requires an overt copula, the copula is also realized in the negative construction; for example, it must be expressed in the past tense with the negative element, as shown in (6).⁵

- (6) a. Ben doktor **değil** **(i-di-m)*.
1SG doctor NEG COP-PST-1SG
'I was not a doctor.'⁶
- b. Мен доктор **эмес** **(эле-м)*.
Men doktor **emes** **(ele-m)*.
1SG doctor NEG COP;PST-1SG
'I was not a doctor.'

While the occurrence of two copular elements within a single predication domain is not entirely inconceivable, such cases render the behavior of the negative element more obvious and provide tentative evidence against analyzing it as a copula in this construction.

A further argument against analyzing the negative element as a copular verb comes from agreement patterns. As illustrated in (5) and (6), subject agreement appears on the negative element only in the absence of an overt copula. For convenience, we repeat the Turkish examples in (7). In example (7a), the negative element bears first person singular agreement in the absence of an overt copula. Example (7b) illustrates the corresponding structure in which the negative element co-occurs with an overt copula, and the agreement suffix is realized on the copular verb rather than on the negative element. Together, these examples highlight the complementary distribution of agreement marking between the negative element and the overt copula in non-verbal predication contexts.

- (7) a. Ben doktor **değil-im**.
1SG doctor NEG-1SG
'I am not a doctor.'
- b. Ben doktor **değil i-di-m**.
1SG doctor NEG COP-PST-1SG
'I was not a doctor.'

This distribution shows that agreement morphology is realized on the negative element in environments where the copula is absent. Importantly, this does not imply that the negative element functions

as a copular verb. Rather, the data are consistent with the interpretation that the negative element is part of the predicate phrase, serving as the locus for the morphological expression of person and number features when no overt copula is present.

Second, the subject agreement suffixes realized on the negative element are morphologically identical to those appear on nouns, adjectives, and adverbs when they function as predicates in the present tense, i.e. in the absence of an overt copula. By contrast, the verbal copula employs a distinct set of subject agreement morphemes. This contrast is clearly illustrated by the Kyrgyz data (c.f. [Kasieva et al., 2023](#)), which show that the agreement suffix attached to the negative element in (8a) patterns with the nominal/adverbial predicate agreement type in (8b), and differs from that realized on the copula in (8c), thereby making the distinction between the two paradigms explicit.

- (8) a. Мен доктор **эмесмин**.
Men doktor **emes-min**.
1SG doctor NEG-1SG
'I am not a doctor.'
- b. Мен доктормун / парктамын.
Men doktor-**mun** / park-ta-**min**.
1SG doctor-1SG / park-LOC-1SG
'I am a doctor/at the park.'
- c. Мен паркта элем.
Men park-ta **ele-m**.
1SG park-LOC COP;PST-1SG
'I was at the park.'

The morphological similarity of the agreement markers on the negative element to those found on nominal, adjectival, and adverbial predicates in copula-less clauses further distinguishes it from the verbal copula, which uses a distinct set of agreement morphemes.⁷ This mismatch provides independent evidence against analyzing the negative element as a copular or auxiliary element.

Overall, the data support viewing the negative element as a polarity-sensitive element operating within the predicate domain rather than as a verbal or copular head. Its verbal-like agreement morphology arises from its structural position: agreement surfaces on the negative element only when the copula is absent, without requiring a specific representation of a null copula.

⁵As the Azerbaijani example is parallel to the Turkish one, it is not illustrated separately here.

⁶The notation **(X)* indicates that the element in parentheses is obligatory; its omission renders the sentence ungrammatical.

⁷An alternative interpretation would be to analyze the negative element as instantiating the predicate head in copula-less clauses, given that it bears agreement morphology identical to that found on nominal predicates. While compatible with the morphological facts, pursuing this analysis would require a broader reconsideration of the structure of non-verbal predication in these languages.

Finally, in Turkish and Azerbaijani, the negative element occupies a fixed position within the predicate phrase, intervening between the nominal predicate and the verbal copula. While its surface position is invariant, the interpretation of the negation may vary: it can be understood as targeting the entire predicate or a specific constituent within it, such as a modifier. For instance, in a sentence like “I was not a good doctor” in (9), the negative element may be interpreted as negating the entire predicate (the individual was not a doctor at all) or only the modifier (the individual was a doctor, but not a good one). Since the position of the negative element does not vary, disambiguation relies on context or on explicit linguistic devices (e.g. *Ben iyi bir doktor değil, iyi bir hemşireydim* ‘I was not a good doctor, but a good nurse’). Its consistent adjacency to the copula and integration within the predicate phrase suggest that it is structurally part of the predicate and functions as a polarity-sensitive element, rather than as a freely mobile or independent auxiliary.

- (9) Ben iyi bir doktor **değil**
 1SG good INDEF doctor NEG
 i-di-m.
 COP-PST-1SG
 ‘I was not a good doctor.’

In contrast, in Kyrgyz the position of the negative element within the predicate phrase is less clearly fixed. It appears to be structurally flexible in that it can follow different constituents within the predicate phrase, possibly depending on which part is being negated. At the same time, there is a tendency for it to occur immediately after the constituent in its scope. This pattern suggests that the negative element can take scope over internal modifiers of non-verbal predicates: in (10a), it follows the adjective and is interpreted as negating it, whereas in (10b), it follows the entire nominal predicate and is interpreted as negating the predicate as a whole.

- (10) a. Мен жакшы **эмес** доктор
 Men jakşı **emes** doktor
 1SG good NEG doctor
 элем.
 ele-m.
 COP;PST-1SG
 ‘I was not a good doctor.’ (I was a doctor, but not a good one.)⁸
 b. Мен жакшы доктор **эмес**
 Men jakşı doktor **emes**
 1SG good doctor NEG

⁸While (10a) is grammatically acceptable, the adjective-negating interpretation may be reinforced by prosodic stress on the adjective or by the addition of the focus particle *деле* ‘even, indeed’.

элем.
 ele-m.
 COP;PST-1SG
 ‘I was not a good doctor.’ (I was not a doctor at all.)

In sum, in Azerbaijani, Kyrgyz, and Turkish, the negative element does not behave as a copula or an auxiliary. In contexts where the non-verbal predicate in the non-negated clause does not require an overt copula, no copula appears in the negative construction. Agreement morphology surfaces on the negative element only in the absence of the copula, and the forms of these markers are identical to those found on nominal, adjectival, and adverbial predicates. The negative element occupies a fixed position within the predicate phrase, consistently following the non-verbal predicate. Together, these patterns indicate that, across all three languages, the negative element is structurally integrated within the predicate domain and functions as the locus for agreement realization.

4. Proposed analysis

Building on the grammatical and positional patterns outlined above, we now turn to the analysis of the negative element. Across the constructions illustrated in Azerbaijani, Kyrgyz, and Turkish, the negative element behaves as an operator: it takes scope over the constituent it negates and requires adjacency to establish a clear scope relationship. Its position immediately preceding the copula, together with the fact that agreement features appear on the negative element only when the copula is absent, further distinguishes it from a true copula or an auxiliary. Taken together, these observations motivate a descriptive analysis in which the negative element is treated as a phrasal operator associated with non-verbal predicates rather than as a copula or an auxiliary.

Within the framework of Universal Dependencies, this interpretation can be captured by assigning the negative element the part-of-speech tag *ADV* and the dependency relation *advmod:neg*.⁹ This analysis accommodates its operator-like behavior, adjacency requirements, and interaction with covert copulas, providing a linguistically informed and typologically consistent annotation strategy for Turkic treebanks. A closely related analysis is already attested in the UD Turkish–German SAGT treebank (Çetinoğlu and Çöltekin, 2023), where *değil*

⁹Note that analyzing the negative element as an adverb (*ADV*) is not contradicted by the presence of agreement markers on it: in Turkic languages, adverbs can take agreement morphology when functioning as predicates (Kornfilt, 2013), and this agreement is associated with the predicate as a whole rather than the adverb itself.

is consistently annotated with the relation `advmod` and the POS tag `PART` (Çetinoğlu and Çöltekin, 2023). Our proposal differs only in the choice of POS tag (`ADV` rather than `PART`) and in the use of the subtype `advmod:neg` to explicitly mark the negation function. By distinguishing this element from the copula, our approach ensures that UD treebanks can more accurately reflect the syntactic and semantic properties of negation in non-verbal clauses across Turkic languages. The proposed analysis is illustrated in Figure 2, showing UD analyses of the Turkish examples in (3b) and (5c), respectively. Additionally, the Kyrgyz examples in (10a) and (10b) are illustrated in Figure 3.

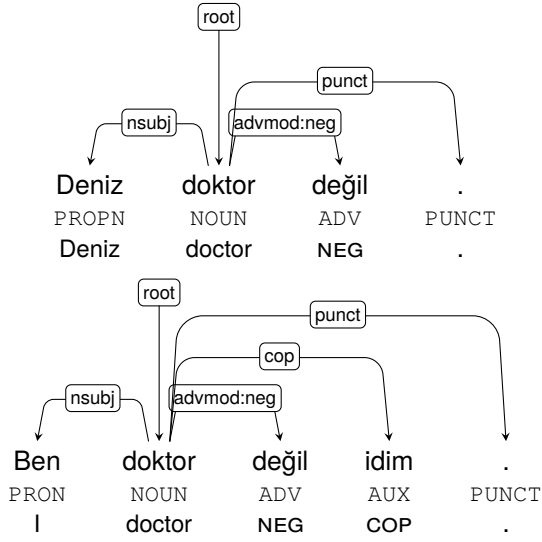


Figure 2: Proposed UD analyses of the Turkish examples in (3b) and (5c), respectively, with *değil* analyzed as `ADV` and `advmod:neg`.

When the negative element bears copular suffixes, as in Turkish *değildim* (*değil* + *i-di-m*), the bound copula and agreement morphology should be segmented into a separate token following the tokenization guidelines for Turkic copula constructions proposed by Coltekin et al. (2026).

5. Conclusions

This study has examined the behavior of the negative element in non-verbal predicates across Azerbaijani, Kyrgyz, and Turkish, focusing on its syntactic position, interaction with the copula, and agreement patterns. In Turkish and Azerbaijani, the negative element occupies a fixed position, intervening between the nominal part of the predicate and the copula, whereas in Kyrgyz it shows greater positional flexibility, consistently appearing immediately after the constituent it negates. Across all three languages, agreement appears on the negative el-

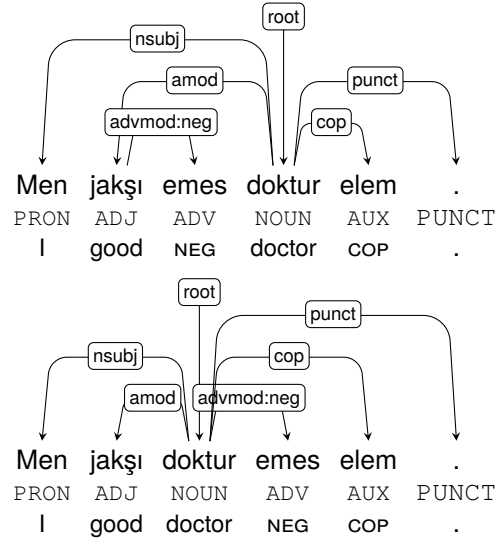


Figure 3: Proposed UD analyses of the Kyrgyz examples in (10a) and (10b), respectively, with *emes* analyzed as `ADV` and `advmod:neg`. In the first tree, *emes* attaches to the adjective; in the second, to the nominal predicate.

ement only when the copula is absent, highlighting the distinction between the element itself and the copula.

These patterns consequently support an analysis in which the negative element is treated as a phrasal operator associated with non-verbal predicates rather than as a copula or auxiliary. Its fixed or constrained positioning, adjacency requirements, and interaction with agreement provide clear evidence of its structural integration and functional role within the clause.

Acknowledgements

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).

Bibliographical References

Arofat Akhundjanova, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, and Cagri Coltekin. 2025. [Parallel Universal Dependencies treebanks for Turkic languages](#). In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 129–136, Ljubljana, Slovenia. Association for Computational Linguistics.

Özlem Çetinoğlu and Çağrı Çöltekin. 2023. [Two languages, one treebank: building a Turkish–German code-switching treebank and its challenges](#). *Language Resources and Evaluation*, 57(2):545–579.

Cagri Coltekin, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, Sardana Ivanova, Gulnura Dzhumalieva, Aida Kasieva, Nikolett Mus, and Jonathan Washington. 2026. [Tokenisation of Turkic copula constructions in Universal Dependencies](#). In *Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026)*, pages 172–178, Rabat, Morocco. 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.

Aida Kasieva, Gulnura Dzhumalieva, Anna Thompson, Murat Jumashev, Bermet Chontaeva, and Jonathan Washington. 2023. Issues of Kyrgyz syntactic annotation within the Universal Dependencies framework. In *Proceedings of the XI International Conference on Computer Processing of Turkic Languages “TurkLang 2023”*, pages 189–215, Bukhara, Uzbekistan.

Jaklin Kornfilt. 2013. *Turkish*. Routledge.

S.N. Lee. 1996. [A Grammar of Iranian Azerbaijani](#). Cognitive science research papers. University of Sussex.

Martine Robbeets and Alexander Savelyev. 2020. *The Oxford guide to the Transeurasian languages*. Oxford University Press.

Muberra Seydi. 2020. Existential negation in Turkish: A functional approach. *Journal of Language and Linguistic Studies*, 16(2):647–660.

Petr Sleptsov. 2018. Саха тылын быһаарыылаах улахан тылдьыта: Большой толковый словарь якутского языка. [*Large explanatory dictionary of the Yakut language: in 15 volumes*]. Новосибирск: Наука.

E.I. Ubryatova, E.I. Korkina, L.N. Kharitonov, and N.E. Petrov, editors. 1982. Грамматика современного якутского литературного языка: Фонетика и морфология [*Grammar of the modern Yakut literary language: Phonetics and morphology*]. Москва: Наука.

Language Resource References

Akkurt, Furkan and Çöltekin, Çağrı. 2025. [Universal Dependencies Turkish TueCL Treebank](#). Universal Dependencies Consortium.

Çetinoğlu, Özlem and Çöltekin, Çağrı. 2023. [Universal Dependencies Turkish–German SAGT Treebank](#). Universal Dependencies Consortium.

Marşan, Büşra and Akkurt, Furkan and Türk, Utku and Atmaca, Furkan and Özateş, Şaziye Betül and Berk, Gözde and Bedir, Seyyit Talha and Köksal, Abdullatif and Öztürk Başaran, Balkız and Güngör, Tunga and Özgür, Arzucan. 2022. [Universal Dependencies Turkish BOUN Treebank](#). Universal Dependencies Consortium.

Towards Universal Dependencies for L2 Learners of Modern Greek: Annotation and Challenges

Christina Klironomou¹, Thelka Pasparaki²,
Arianna Masciolini¹, Alexandros Tantos³,
Despoina Ourania Touriki³, Konstantinos Tsiotskas³, Eleni Tsourilla³

¹University of Gothenburg, ²University of Crete, ³University of Thessaloniki
guskilrch@student.gu.se, philp1011@philology.uoc.gr,
arianna.masciolini@gu.se, alextantos@lit.auth.gr
dtouriki@lit.auth.gr, ktsiotskas@gmail.com, etsourib@lit.auth.gr

Abstract

This paper focuses on annotating the Greek Learner Corpus in Universal Dependencies (UD). It presents the annotation process, development of guidelines and evaluation of the attempted annotation of two annotators. This work is part of a larger annotation project which aims to compile a sizeable learner treebank that can be used to promote research on second language acquisition and its automatic processing.

Keywords: Universal Dependencies, Modern Greek, L2

1. Introduction

Universal Dependencies (UD) (de Marneffe et al., 2021) is a well-established framework for the cross-linguistically consistent annotation of morphosyntax, aiming to promote multilingual grammatical parsing and typological research. Since its first release in 2015, when only a handful of European languages were represented in the corpus, the UD annotation scheme has been used to create hundreds of resources for over 180 languages from all over the world, including multiple minority languages, dialects, and nonstandard varieties.

In a 2017 paper, Lee et al. proposed parallel UD treebanks as a format for L2 corpora, arguing that morphosyntactically annotated learner productions contrasted with teacher corrections are functionally equivalent to – and at the same time richer than – explicitly error-annotated corpora. In the following years, this idea has grown in popularity, leading to the development of nine such UD treebanks (Lee et al., 2017a; Berzak et al., 2016; Pulido et al., 2025; Di Nuovo et al., 2022; Sung and Shin, 2024; Hana and Hladká, 2018; Kyle et al., 2022; Rozovskaya, 2024; Volodina et al., 2025).

The aim of this work is twofold. From a computational perspective, gold-annotated learner productions allow for the inclusion of nonstandard text in parser development, thereby improving the robustness and overall performance of the models. From the perspective of second language acquisition (SLA) research, the treebank can constitute a helpful resource for cross-lingual interlanguage studies, shifting the focus from individual errors to L2 grammar as a whole.

Annotating learner corpora is a challenging un-

dertaking, as standard annotation guidelines are typically designed for well-formed, native-like language data and do not adequately account for the systematic ungrammaticality and non-standard structures found in learner texts. Furthermore, existing learner treebanks do not follow a unified approach to the treatment of learner errors (Masciolini et al., 2025). This lack of convergence underscores the need for universal annotation guidelines tailored for learner corpora, as well as for at least some language-specific extensions that capture the idiosyncrasies of the learner variety of the language under investigation.

In the present paper, we present ongoing work on a new treebank based on version 2 of the Greek Learner Corpus (GLCII) (Tantos et al., 2023), consisting of written data produced by learners of Modern Greek.¹ For the purposes of this project, we annotated 50 texts (671 sentences) produced by L2 learners of Modern Greek across different CEFR levels along with their reference counterparts (658 sentences), extracted from the development set of the GCLII corpus. In addition, we developed treebank-internal annotation guidelines to ensure consistency across annotations and to systematically address language- and treebank-specific phenomena, particularly those arising from learner errors.

2. The Greek Learner Corpus

GLCII, the largest and freely available error-annotated Learner Corpus of Greek as a second

¹The treebank is and its first version is planned to be released as part of UD 2.18.

Level	Texts
A1	6
A2	7
B1	13
B2	15
C1	7
C2	2

Table 1: Texts per CEFR level in the subset of GLCII currently annotated in UD.

L1	Texts
Russian	15
Serbian, Turkish	5
Albanian, Chinese, English	4
Arabic, Italian	3
Polish	2
Abkhazian, Armenian, French, Georgian, Slovak	1

Table 2: Texts per first language in the subset of GLCII currently annotated in UD.

language (L2), was developed within the Latent Aspects in L2 Acquisition (LAL2A) research program, funded by the Hellenic Foundation for Research and Innovation and designed to address limitations identified in earlier Greek learner corpora (Tantos et al., 2023). GLCII is a growing corpus, comprising 1,101 authentic raw written texts (422,360 tokens) and 318 spoken interviews produced by adult L2 Greek learners. The corpus represents a wide range of sociolinguistic and cultural backgrounds, which are extensively documented through rich metadata. The majority of the learner data was collected in classroom instruction settings. Furthermore, GLCII includes a control sub-corpus of Greek L1 data consisting of 42 written texts (66,645 tokens), which is particularly suitable for systematic comparisons of L2-L1.

GLCII is error-annotated for five basic grammatical categories: Agreement, Voice, Gender, Case and Aspect. Each category is associated with a dedicated tag system designed to capture specific error types.² Overall, GLCII is grounded in contemporary learner corpus practices and constitutes a valuable resource for both L2 research and teaching purposes.

3. Annotation process

As mentioned above, we annotated 671 sentences and their reference counterparts (658) from 50 learner texts, extracted from the development set of the GLCII corpus.³ The distribution of CEFR proficiency levels and L1s for this annotated subset is displayed in Table 1 and 2 respectively.

This effort is the work of two annotators, both native speakers of Modern Greek and currently master-level students of Linguistics, with previous annotation experience and basic knowledge of UD.

²Further information about the annotation scheme is available through the official GLC Gateway interaction platform, which offers documentation and access to selected portions of the corpus (glc.lit.auth.gr/app/GLC_Gateway)

³The total number of sentences varies slightly depending on the annotators' segmentation criteria.

However, this work is part of a larger project that will lead to the expansion of both the language specific guidelines, discussed in Section 4, and the treebank itself. The complete annotation group consists of six native speakers of Modern Greek.

Syntactic relations offer insights about the learners' processing of L2 structures and were therefore identified as a crucial annotation layer. For that reason, the dependency analysis of the original sentences was performed entirely manually, thus preventing any automatic annotation bias. Conversely, tagging and lemmatization were performed automatically with the *greek-gdt-ud-2.17* UD Pipe 2 (Straka, 2018) model and manually validated. The same model was used to pre-annotate corrections, in this case without excluding any CoNLL-U fields.

Given the particular nature of the learner data, the quality of the automatic pre-annotation was overall satisfactory, and undoubtedly saved us a large amount of time that would be spent, e.g., annotating morphological features from scratch. With that said, the parser had serious problems segmenting sentences when the punctuation was nonstandard or when full texts were typed in capital letters (in the latter case, tokenization was also problematic, especially when punctuation was involved), thus a lot of post-processing and decisions had to be made. In addition, misspelled or unstressed words were often lemmatized incorrectly. The feature and MISC columns' values attributed by the parser were for the majority of the times correct, but often enough had to be modified and/or complemented with extra values in correspondence to grammatical errors. It must be noted that learner data pose additional challenges as features cannot always be assigned with certainty when the words are malformed or the sentence is ambiguous. For example, when a learner sentence contains structures such as *ήρθα αύριο* [=‘I came tomorrow’], we do not know – and cannot always infer from the context – if the learner meant ‘I came yesterday’ or ‘I will come tomorrow’. Having a reference text can often suggest a way out for the annotator, since a lot of choices for resolving such ambiguities are taken while constructing

a reference text, but this is not always the case.

Furthermore, Modern Greek presents significant challenges due to widespread syncretism, particularly within the nominal domain where identical morphological forms can map to multiple grammatical values. This leads to two primary types of error ambiguity according to (Tantos and Amvrazis, 2022): classification-level ambiguity, where it is unclear whether an error stems from a failure in agreement or incorrect gender assignment to a noun, and identification-level ambiguity, where a single malformed token could represent multiple non-target cases or genders. During this initial stage, each annotator had to keep the automatic feature annotation or modify it according to their intuition, with discrepancies settled during a conflict resolution process, and ambiguities were resolved by selecting one of the possibly suitable feature values.

To address these challenges without imposing biased or deterministic interpretations, we are considering the possibility of incorporating the descriptive, theory-neutral approach proposed by (Tantos and Amvrazis, 2022) in future versions of the treebank. Specifically, this approach utilizes disjunctive tags to handle identification-level issues and represents disjunctive relations between multiple plausible error domains. Such a strategy would allow the treebank to capture the inherent complexity of the learner’s interlanguage while significantly improving inter-annotator agreement (IAA) by reducing subjectivity.

4. Internal guidelines

4.1. Segmentation

Original texts are segmented into sentences based on their punctuation unless there are very strong signs that the learner forgot or misplaced the period (e.g. another sentence beginning with a capitalized word is following, but the period is missing). The annotators tried to keep the text segmentation as faithful to the learner’s judgment as possible, sometimes diverging from the choice made in the reference text. To address the potential sentence misalignment between learner and reference text, a three-level naming system for sentence IDs was developed: If, for instance, the second sentence of the sixth text in the reference file (6.2) corresponds to two sentences in the original text, in the learner file, the sentence id will be 6.2.1 for the first sentence and 6.2.2 for the second. This allows obtaining a fully sentence-parallel treebank, similar to the ones available for Chinese (Lee et al., 2017a), English (Berzak et al., 2016), Italian (Di Nuovo et al., 2022), Russian (Rozovskaya, 2024) and Swedish (Volodina et al., 2025).

When dealing with under-segmented words where the merged tokens are assumed to be a result of mistyping, tokenization and annotation is done as if they were correctly separated, indicating that a space is missing in the MISC field of the first token. In *το ιδιο ισχυει καιγια την νυχτερίδα* [=‘the same applies also for the bat’], for instance, we assume that the learner knows that *και* (‘and, also’) and *για* (‘for’) are separate words and annotate them as separate tokens. When the under-segmentation is not clearly a result of mistyping, the under-segmented tokens are treated as one and analyzed based on what is considered as the head of the merged tokens.

Following the KSL treebank (Sung and Shin, 2024), when dealing with over-segmented tokens, if the resulting tokens are not existing words in the language and no reasonable morphosyntactic annotation can be drawn, the first token is given a dependency relation, the following tokens depend on the first with the *goeswith* relation, and POS tag X and no features are assigned to the latter. In addition, *CorrectSpaceAfter=No* is placed in the MISC column of the following token.

Ellipsis points (‘...’) are annotated as one token, but multiple punctuation marks that do not form a unit (i.e. their meaning does not change when they are combined, e.g. ‘!!’, ‘;;’, ‘..’) are treated as individual tokens.

4.2. Derivational morphology

Following the VALICO treebank (Di Nuovo et al., 2022), lemmas in the annotation of the learner texts keep the spelling and stress of the tokens as they were typed by the learner. The features are also descriptive of the observed form, independent of the learner’s intended meaning, unless it causes problems in producing a syntactic annotation. POS tags and syntactic relations are also assigned based on the form of the given lemma. Thus, in the sentence *Αν και ήθελα να πάω όχι στην Αλβανός αλλά στην Ελλάδα* [=‘Even though I wanted to go not to Albania(n) but to Greece’], the word *Αλβανός* (‘Albanian’, lit. person of Albanian origin) is annotated as a noun:

PART	ADP	DET	NOUN	CCONJ	ADP	DET	PROPN
όχι	σ	την	Αλβανός	αλλά	σ	την	Ελλάδα
not	in	the	Albanian	but	in	the	Greece
			Αλβανός				
			Case=Nom				
			Gender=Masc				
			Number=Sing				

A fully interpretative annotation would lemmatize according to the hypothesized intended form *Αλβανία* (‘Albania’), which would be tagged as a proper noun with the features *Case=Acc | Gender=Fem | Number=Sing*.

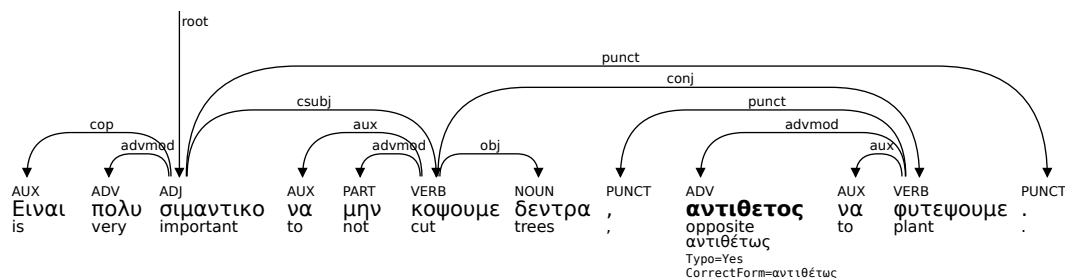


Figure 1: Application for the guidelines for misspellings. The adjectival form *αντιθετος* is used instead of the corresponding adverb, but since this is the result of a phonetic misspelling, the lemma, UPOS tag and DEPREL are assigned based on the target form *αντιθέτως*. The presence of a misspelling and the target form itself are indicated in the FORM and MISC fields. Another spelling error is present in the above example sentence (*σιμαντικο* instead of *σημαντικό*), but the wrong choice of /i/ does not change the morphosyntactic analysis of the token or the sentence.

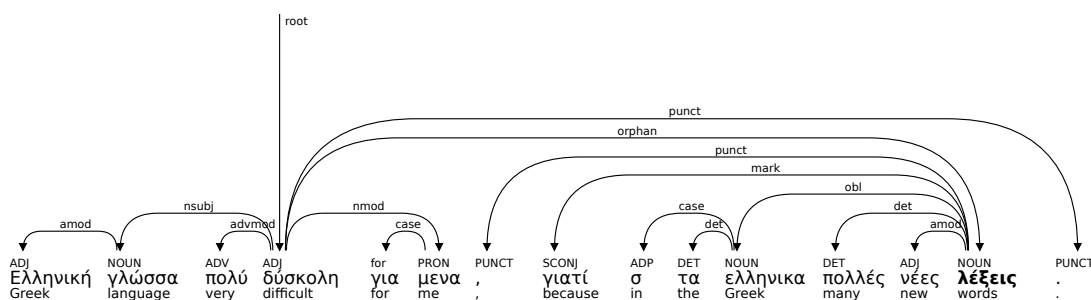


Figure 2: Application of the guidelines for ellipsis. The sentence, which translates approximately to ‘(the) Greek language (is) very difficult for me, because in Greek (there are) many new words’, lacks a determiner and an auxiliary in the main clause, but these do not pose a problem as they would be leaves in the UD tree. In the subordinate, however, *λέξεις* appears to be the subject of an (omitted) existential verb (e.g. *υπάρχουν*, ‘there exist’). It is therefore attached to the root and labeled *orphan*.

4.3. Orthography

When the spelling or stress point to a UPOS tag that does not permit any kind of reasonable syntactic annotation but it is obvious that the learner misspelled, ‘*Τypo=Yes*’ is added to the feature column and the lemma and features correspond to the token that is assumed that the learner meant. The intended token is also specified to the MISC column using *CorrectForm*=[...]. For example, in the sentence *Είναι πολύ σιμαντικο να μην κοψουμε δεντρα, αντιθετος να φυτεψουμε* [=‘It is very important not to cut trees, converse(ly) to plant] we suggest that the learner used the adjective *αντιθετος* (‘opposite’) instead of the adverb *αντιθέτως* (‘on the contrary’) not because they were unaware of the target form, but rather as a result of a phonetic misspelling that is common even among L1 speakers. We therefore annotate on the basis of the target form (cf. Figure 1).

Spelling and stress errors are not marked with the current guidelines in the L1 CoNLL-U file, but can be found by comparing the UD annotation of the learner file with the corrected (L1) counterpart. Note also that a script that automatically extracts such errors from comparing UD annotated files is currently developed by authors of this paper.

For words transliterated from other languages, we apply the guidelines for borrowed analysis of code-switched material,⁴ i.e. assign no features and specify the source language in the MISC column. For instance, the word *Haouai* – a transliteration of the name ‘Hawaii’, used instead of its proper Greek translation *Χαβάη* – is annotated with *OrigLang=en*.

4.4. Syntax

In our annotated subset of texts, written mostly by B2 learners within the CEFR framework, most of the non-trivial cases of syntactic annotation concern missing or redundant tokens. When it comes to missing tokens, we follow the guidelines for promotion by head elision and ellipsis of the CLF and Russian learner treebanks, as described in Mascioli et al. (2025) (cf. Figure 2).

Repeated tokens (not for emphatic reasons), in case they do not serve any semantic purpose or cannot be assigned any syntactic function, are assigned the *reparandum* relation to their following counterparts that can be analyzed syntactically

⁴universaldependencies.org/foreign.html#option-2-borrowed-analysis

	LEMMA	UPOS	UAS	LAS
Originals	97.5%	97%	94.1%	91.32%
References	98.95%	98%	95.26%	92.99%

Table 3: IAA on original learner texts and corrected references.

(the presumed ‘repair’). This comes as an extension to the typical use of *reparandum*, where it is used to indicate disfluencies overridden in a speech repair.

5. Inter-Annotator Agreement

The manual syntactic annotation of the original learner sentences proved less challenging than expected. This can also be seen from the extent of the inter-annotator agreement (IAA) (cf. Table 3) and is perhaps also related to the relatively high level of proficiency of the learners (cf. Section 1). Having a reference text, to which we had chosen to remain faithful, was also helpful sometimes, giving us a way out of structures that posed ambiguity (semantic or syntactic) to their interpretation. However, the shortage of learner-specific/error-specific guidelines posed additional requirements of effort and attention to the annotation process.

We conducted a three-level quantitative evaluation of the annotation:

- Raw agreement on lemma assignment (counts of identical lemmas / total number of tokens).
- IAA on POS tags, measured with Cohen’s kappa coefficient, a metric that also accounts for potential cases where the two annotators selected the same label by chance.
- IAA on dependencies, using UAS (Unlabeled Attachment Score) and LAS (Labeled Attachment Score).

The scores acquired are comparable to the ones of other learner treebanks (cf. for example Berzak et al. 2016; Lee et al. 2017a; Di Nuovo et al. 2022; Sung and Shin 2024). For the case of the dependency relations on the reference texts, as the normalized data are less ambiguous and the parser annotation bias is present, agreement scores were higher. There was no dependency that distinguished particularly for causing annotator disagreement.

6. Limitations

The present work is part of an ongoing annotation process, thus not all choices followed are definitive and the size of the treebank is still relatively small,

especially compared to that of the source corpus (the annotation covers 4,5% of the whole corpus). We are currently working on extending the treebank to include more pairs of learner and reference texts, and enhancing our annotation technique in a way to better handle the ambiguity issues discussed in Section 3. This will allow to derive error labels compatible with the original GLCII guidelines from the UD annotations themselves.

7. Conclusions

This study presented the annotation and evaluation of a Greek learner treebank within the UD framework. The annotation process required the adaptation of existing UD guidelines and led to the development of new ones for non-standard linguistic data, addressing phenomena characteristic of learner language. Inter-annotator agreement scores indicate a satisfactory level of consistency, suggesting that the proposed guidelines can be applied reliably despite the variability inherent in learner production. Nevertheless, certain constructions remain challenging, highlighting the need for further refinement of annotation criteria and additional data. Overall, the resulting dataset, scheduled for a first release as part of 2.18, provides systematically categorized and linguistically analyzed material, suitable for research on SLA and pedagogical applications. The present work is part of a broader annotation effort focusing on GLCII annotation for UD. Future work will focus on assessing its utility in downstream tasks, particularly dependency parser training and fine-tuning.

8. Acknowledgments

This work has been supported by Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (2025–2028; project id 2023-00161). In addition, we would like to thank Onassis Foundation, who supports the main author of this work under the Scholarship Program for Greek students 2025-2026.

9. Bibliographical 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, 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).
- Jirka Hana and Barbora Hladká. 2018. Universal dependencies and non-native Czech. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, Oslo University, Norway. LiU Electronic Press.
- 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, Aleksandrs Berdicevskis, Maria Irena Szawerna, and Elena Volodina. 2025. [Annotating second language in Universal Dependencies: a review of current practices and directions for harmonized guidelines](#). In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 153–163, Ljubljana, Slovenia. Association for Computational Linguistics.
- Emiliana Pulido, Robert Pugh, and Zoey Liu. 2025. [I speak for the árboles: Developing a dependency treebank for Spanish L2 and heritage speakers](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 814–822, Vienna, Austria. Association for Computational Linguistics.
- 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.
- 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.
- Alexandros Tantos and Nikolaos Amvrazis. 2022. [Classification and identification level ambiguity in error annotation](#). *Applied Corpus Linguistics*, 2(3):100035.
- Alexandros Tantos, Nikolaos Amvrazis, and Eleni Drakonaki. 2023. [Greek Learner Corpus II \(GLCII\): Design and development of an online corpus for L2 Greek](#). *Journal of Applied Linguistics*, 36.
- Elena Volodina, Arianna Masciolini, Beáta Megyesi, Julia Prentice, Lisa Rudebeck, Gunlög Sundberg, and Mats Wirén. 2025. [SweLL with pride: How to put a learner corpus to good use](#). In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis, and Elena Volodina, editors, *Huminfra Handbook: Empowering digital and experimental humanities*, number 59 in NEALT Proceedings Series, pages 251–306. University of Tartu Library.

10. Language Resource References

- 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.
- Alexandros Tantos, Nikolaos Amvrazis, and Eleni Drakonaki. 2023. [Greek Learner Corpus II \(GLCII\): Design and development of an online corpus for L2 Greek](#). *Journal of Applied Linguistics*, 36.

A Comparative Linguistic Analysis of Ottoman and Modern Turkish through UD Treebanks

Enes Yilandiloğlu

University of Helsinki
Metsätalo, Unioninkatu 40, 00170 Helsinki, Finland
enes.yilandiloglu@helsinki.fi

Abstract

While the linguistic shifts between Ottoman and modern Turkish are well-documented qualitatively, quantitative analyses remain scarce. This study addresses this by conducting a comparative computational analysis using two Universal Dependencies treebanks: OTA-DUDU for Ottoman Turkish and TR-BOUN for modern Turkish. By employing descriptive statistics and a log-likelihood ratio test, we demonstrate the change and quantify the magnitude of diachronic variation. The analysis yields three primary statistical findings. First, our data reveals a 77% compliance rate with labial vowel harmony for suffixes, while this value is 98% in modern Turkish. This discrepancy can be explained by the presence of rounding in Ottoman Turkish, which disappears in modern Turkish. On the other hand, the compliance rate of palatal vowel harmony is quite high for both languages, 96% for Ottoman Turkish and 99% for modern Turkish. Second, some suffixes, such as the converb $-(y)lp^1$ and the dative infinitive $-mAyA$, changed by reducing their allomorphs in modern Turkish. Third, we demonstrate that Arabic and Persian pluralization rules, which constituted 28% of plural nouns in Ottoman Turkish, lost their pluralizing function in modern Turkish, although the words remain with singular meaning.

Keywords: Universal Dependencies, Ottoman Turkish, modern Turkish, linguistic variation

1. Introduction

The transition from Ottoman Turkish to modern Turkish involved not only an alphabet change but also significant lexical, phonological, and morphological shifts. Although such changes were documented in previous qualitative studies, the use of quantitative methods remained limited. This study differentiates itself by utilizing annotated Universal Dependencies (UD) treebanks, OTA-DUDU (Yilandiloğlu and Siewert, 2025) for Ottoman Turkish and TR-BOUN (Marşan et al., 2022) for modern Turkish, to conduct a comparative analysis. It demonstrates that such changes can be traced through treebanks to follow how Turkish evolved throughout time. Thus, it analyzes historical linguistic changes in terms of vowel harmony in suffixes attached to nouns, the change in certain suffixes, and pluralization rules. First, Ottoman Turkish exhibited a tendency toward rounding in vowels (labialization). In contrast, modern Turkish enforces stricter labial vowel harmony for grammaticality, transforming Ottoman forms like *insānuñ* (human-2SG.POSS) into *insanın* (human-2SG.POSS). Despite these shifts, certain exceptions persist in modern Turkish due to loanwords like *bahsi* (topic-ACC) and *halim* (condition-1SG.POSS). Second, the changes in the suffixes, $-(y)lp$ and $-mAyA$ were further investigated, revealing the obsolescence of certain allomorphs. Third, it shows how pluralization takes place in Ottoman and modern Turkish.

2. Background

Turkish is a language in the branch of the Altaic language family (Lewis, 1977, v). Ottoman Turkish refers to the historical version of Turkish (see Demir and Yılmaz, 2002 for a discussion on naming Ottoman Turkish and Kerslake, 2015, 180-1 for problems of terminology).

The vowel harmony in modern Turkish is a phonological process to determine the vowels in a word except the first vowel (Göksel and Kerslake, 2005, 21). Vowel harmony in both Ottoman Turkish and modern Turkish has two subsystems: palatal harmony and labial harmony. Palatal harmony (2-way alternation) is related to the selection of low vowels in suffixes: back vowels (a, ı, o, u) trigger $-a$, while front vowels (e, i, ö, ü) require $-e$. Labial harmony (4-way alternation) is for the selection of high vowels in suffixes based on both backness and rounding: back unrounded vowels (a, ı) trigger $-ı$, front unrounded (e, i) trigger $-i$, back rounded (o, u) trigger $-u$, and front rounded (ö, ü) trigger $-ü$ (see Clements and Sezer 1982; Lewis, 1977, 19-22; Özçelik 2024, 128-157 for detailed explanation of vowel harmony in Turkish). Yet the main difference between Ottoman Turkish and modern Turkish lies in the labial harmony. Especially during the early stages, Ottoman Turkish allowed the forms like *gelüp* (come-CVB) or *vilāyetinüñ* (province-3SG.POSS-n-GEN) in its suffixes where the suffix can become rounded, even though the palatal harmony is maintained. Modern Turkish, conversely, is stricter in such cases and requires labial harmony for

¹ In this paper, A refers to low/unrounded vowels for palatal harmony (a and e), I refers to high vowels for labial harmony (ı, i, u, and ü), B refers to b and p, and G refers to g and ğ.

grammaticality, changing *gelüp* (come-CVB) to *gelip* (come-CVB) and *vilâyetinüñ* (province-3SG.POSS-n-GEN) to *vilâyetinin* (province-3SG.POSS-n-GEN). However, a few exceptions remain in modern Turkish due to loanwords, such as *bahsi* (topic-ACC) and *halim* (condition-1SG.POSS) (Göksel and Kerslake, 2005, 24).

Studies (Demir and Yılmaz, 2002, 853; Timurtaş, 1976, 332, 334) argue in Ottoman Turkish and Old Anatolian Turkish, a period encompassing Ottoman Turkish between the 14th and the 15th centuries (Demir and Yılmaz, 2002, 853), vowel harmony had a tendency towards rounding in vowels (labialization), and labialization is observed in the vowels of many suffixes. Similarly, Stachowski (2013, 551) analyses the labiality of *-lık* suffix, and points out the presence of “dislabial” vowels in Ottoman Turkish. It was also stated that labial vowel harmony became prevalent in Ottoman Turkish before modern Turkish after the 17th century (Develi, 1995; Gümüşkılıç, 1997; Kerslake, 2015, 185). On the other hand, palatal harmony was well-established in Ottoman Turkish (Demir and Yılmaz, 2002, 861). Additionally, Timurtaş (1976, 341) considers *b/p* change as another difference between Old Anatolian Turkish and modern Turkish while Kerslake (2015, 185-6) argues in modern Turkish, stops in word-final position have become unvoiced (devoiced). The converbial suffix *-(y)lp* was also realized as *-(y)lbAn* which was not encountered in Turkish grammar (Timurtaş, 1976, 336). The converb *-(y)lp* was also used to link clauses in Ottoman Turkish sentences (Kerslake, 2015, 198).

Another suffix *-mAK* followed by the dative suffix is realized as *-maya* and *-meye* in modern Turkish, instead of *-mağa* and *-meğe* which are the older forms (Göksel and Kerslake, 2005, 87). Alpay (2015, 12) also emphasizes the use of soft *ğ* (*ğ*) in such circumstances as an old variation. However, the old variation was still documented in Turkish (Lewis, 1977, 84).

Lastly, pluralization in Ottoman Turkish consists of Turkish, Arabic, and Persian grammatical rules. While modern Turkish pluralization relies on the agglutinative suffix *-lAr* (*-lar* or *-ler* depending on vowel harmony), Ottoman Turkish utilized pluralization rules from Arabic and Persian. From Arabic, words were pluralized by using suffixes, such as *-în*, *-ün*, and *-ât*, and by changing the internal structure of a singular word, known as broken plurals (*cem-i mükesser*) for instance, from *fikr* (idea) to *efkâr* (idea.PL). From Persian, on the other hand, the suffixes *-ân* and *-hâ* were also used. It should be noted that both Arabic and Persian words can be pluralized through the Turkish suffix, *-lAr*. In contrast, modern Turkish simplified this complex system by standardizing the use of *-lAr* suffix for all nouns.

Various studies in the domain of historical linguistics analyzed language variation across time with the support of computational methods (e.g., Nevalainen et al., 2011 for English and Montes et al., 2024 for Spanish). Such studies also utilized parsed treebanks for the data such as Eckhoff et al. (2020). When it comes to the change from Ottoman Turkish to modern Turkish, studies such as Cristaldi (2025) and Petrova and Kuleshova (2020) used dictionaries as the main data source. Other studies leveraged primary sources for the comparison (e.g., Yağmur, 2025). This study differentiates itself from others in the Turkish domain by utilizing annotated treebanks as the data and leveraging computational methods to analyze the change.

3. Data and Methods

This study utilized two treebanks in Universal Dependencies, namely OTA-DUDU for Ottoman Turkish, the largest treebank for the language and TR-BOUN for modern Turkish. OTA-DUDU was chosen as it covers various genres from the 14th to the 20th centuries, though the data is not equally distributed across the centuries, while TR-BOUN was selected since its annotation guidelines align with OTA-DUDU the most among ten Turkish treebanks. The statistics of the OTA-DUDU and TR-BOUN can be seen in Table 1.

	Total tokens	Total words	Total sentences
OTA-DUDU	16,728	17,125	1,782
TR-BOUN	121,835	125,212	9,761

Table 1: Overall statistics of the treebanks

As shown in Table 1, TR-BOUN is considerably larger than OTA-DUDU. We did not downsample TR-BOUN to match the Ottoman corpus size because the G^2 log-likelihood test handles unequal corpus sizes. Absolute frequency comparisons, however, are not normalized and should be interpreted with this imbalance in mind.

Subsequently, we normalized each token by removing diacritics for specific characters (e.g., *ñ* → *n*, *ğ* → *g*). This step was essential as the target treebanks utilize different transliteration schemes, namely Ottoman Turkish transliteration chart (Birnbäum, 1967) and modern Turkish alphabet (Lewis, 1977, 11). This study analyzes three phenomena: suffix variation, vowel harmony compliance, and pluralization patterns. The variations can be seen in Table 2.

Feature	Old variation	Incoming variation
Converb	-(y)lB(An)	-(y)lp

Dative infinitive	-mAGA	-mAyA
Vowel harmony-palatal during suffixation	Strict harmony (Front/Back) 2-way	Strict harmony (Front/Back) 2-way
Vowel harmony-labial during suffixation	Partial Inconsistent (Generalized Rounding) /	Standardized / Strict (4-way harmony)
Pluralization	Turkish/Arabic/Persian pluralization rules	Only Turkish plural rule

Table 2: Variations

The method consists of statistical analysis of the treebanks in a comparative way. In addition to applying the techniques in descriptive statistics such as absolute frequency, this study also leverages log-likelihood ratio test (G^2), a method proposed by Dunning (1993) and previously used in computational and historical linguistics (Rayson et al. 2004 and Lijffijt et al. 2012, also see Lijffijt, 2013 for potential issues in this method). The method is resilient to data imbalance. Log-likelihood ratio test was calculated via the formula below.

$$G^2 = 2 \sum O_i \ln \left(\frac{O_i}{E_i} \right)$$

In this formula, O_i denotes the observed frequency of a feature in corpus i (OTA-DUDU or TR-BOUN), and E_i refers to the expected frequency under the null hypothesis of no difference between corpora. The summation is performed over all four cells of the contingency table (feature present/absent $\times$ Ottoman/modern Turkish).

The analysis was conducted on token counts normalized via log-likelihood, which is robust to large corpus size asymmetries. This approach assesses not only the statistical significance but also the magnitude and direction of change. By doing so, it aims to present what changes and remains in the languages in a quantifiable way.

4. Results

This section presents the results of this study in three subsections: suffix variation analysis, vowel harmony compliance, and pluralization patterns.

4.1 Suffix Variation

To analyze the data, we then explored linguistic variation by two suffixes. These are (i) verb forms ending in $-(y)IB(An)$ in Ottoman Turkish vs. $-(y)Ip$ in modern Turkish and (ii) verbal nouns ending in

$-mAGA$ ($-mağa/-mege$) versus the common version in modern Turkish $-mAyA$ ($-maya/meye$), known as dative infinitive. The first suffix derives adverbs from verbs, the second derives nouns from verbs. They were detected using a rule-based approach: UPOS is VERB, and the suffixes end with either $-(y)IB(An)$ or $-mAGA/-mAyA$. Here the aim was not to extract pairs but to acquire all instances in each treebank.

Table 3 provides the frequency distribution for each variation across treebanks.

	Ottoman Turkish	Turkish
VERB-mAGA	28	0
VERB-mAyA	16	648
VERB-(y)IBAn	0	0
VERB-(y)Ib	7	0
VERB-(y)Ip	453	558

Table 3: Frequency table for linguistic variations

Table 3 demonstrates the obsolescence of $-mAGA$ and $-(y)Ib$ in modern Turkish while they were employed in Ottoman Turkish. It also shows $(y)IBAn$ was completely absent in both treebanks.

4.2 Vowel Harmony Compliance

We analyzed the shift through vowel harmony compliance on two different subsets. The first subset consists of all nouns with suffixes in both treebanks, yielding 3,456 Ottoman Turkish tokens and 21,500 Turkish tokens. This subset provides the overall result; however, it has words that do not appear in the other treebank which results in an imbalance in the lexical coverage. The second subset includes identical noun pairs. The subset was prepared by using a rule-based string-matching approach where the normalized form starts with the normalized lemma. With this approach, we aimed to reduce the impact of lexical variation. We identified 226 shared lemmas across both corpora, which gives 217 identical lemma-feature combinations. This resulted in 429 Ottoman tokens and 1,866 Turkish tokens. The pairs were then examined to quantify historical shift in terms of vowel harmony. Hereafter, the first subset will be referred to as *all nouns* and the second one will be called *matched pairs*.

By utilizing these two subsets, tokens were selected based on whether their suffix contains the relevant vowel type for each harmony. For palatal harmony, only tokens whose suffix contains a low vowel (a or e) were analyzed, as these are the vowels that alternate under palatal harmony. For labial harmony, only tokens whose suffix contains a high vowel (ı, i, u, or ü) were analyzed, as these are the vowels that alternate under labial harmony. Tokens whose suffix lacks

the relevant vowel type are ignored for each analysis, as they provide no evidence for or against the harmony rule. This filtering is based on a rule-based approach to detect the noun suffixes first, excluding the characters except the lemma and identifying the words that are eligible for each vowel harmony analysis.

Table 4 presents the rates for vowel harmony compliance. It shows that palatal vowel harmony compliance is high in both treebanks, while labial harmony shows a prominent difference between Ottoman Turkish (77.06%) and modern Turkish (98.42%) when all nouns are considered. A similar view can be observed while only identical pairs are analyzed. While palatal vowel harmony is 100%, labial vowel harmony in Ottoman Turkish is lower than labial vowel harmony in modern Turkish. An example for the change in labial harmony is *öfine* (front-3SG.POSS-DAT) in Ottoman Turkish and *önüne* (front-3SG.POSS-DAT) in modern Turkish.

Type of Vowel Harmony	Ottoman Turkish – all nouns (%)	Modern Turkish – all nouns (%)	Ottoman Turkish – matched pairs (%)	Modern Turkish – matched pairs (%)
Palatal	96.04 (n=1,941)	98.75 (n=14,141)	100 (n=313)	100 (1,460)
Labial	77.06 (n=2,568)	98.42 (n=13,959)	81 (n=321)	99.85 (1,317)

Table 4: Vowel harmony compliance rates

Additionally, we calculated the use of variations via log-likelihood ratio, which can be seen in Table 5.

Feature	G ²	p-value
Dative Infinitive (-mAGA vs. mAyA)	154	<0.001
Converb (- (y)IB(An) vs. - (y)Ip)	12	<0.001
Palatal Vowel Harmony – all nouns	61	<0.001
Labial Vowel Harmony – all nouns	1,431	<0.001

Table 5: Log-likelihood ratios

Table 5 presents statistically significant results for all linguistic variations. Firstly, it demonstrates a statistically significant change in terms of the dative infinitive, -mAyA, in Turkish. Secondly, it provides a statistically significant shift to the converbial suffix -(y)IB ending with *p*. Lastly, it shows statistically significant result for palatal and labial vowel harmony. Moreover, the G² values indicate the magnitude of difference between the two corpora: higher values represent greater divergence. Notably, labial vowel harmony shows the highest G² value (1,431), indicating that this feature exhibits the most substantial difference between Ottoman and Modern Turkish among all variables. However, palatal vowel harmony shows a considerably lower G² value (61), suggesting less divergence compared to labial vowel harmony. Among suffix variations, the dative infinitive (-mAğA/-mAyA) shows higher divergence (G²=154) than the converb suffix (G²=12).

Further analysis reveals that while vowel harmony, in general, became more restricted, some irregularities persist in modern Turkish, such as *halinde* (state-3SG.POSS-n-LOC) and *saatiyle* (time-3SG.POSS-y-INS) which prevents the vowel harmony compliance rate from reaching 100%.

4.3 Pluralization Patterns

Lastly, we categorized plural nouns based on how they are pluralized: the standard Turkish suffix -lAr, Arabic and Persian pluralization rules. We extracted tokens whose UPOS tag is NOUN with the morphological feature Number=Plur. Tokens containing -lar or -ler in the suffix (after the lemma) were classified as Turkish plurals. Tokens lacking these suffixes, despite being tagged as plural, were classified as Arabic or Persian plurals if they also ended with specific foreign suffixes (e.g., -āt, -ān, -īn, -hā) or contained the Gender=Fem feature, an indicator of Arabic broken plurals, annotated exclusively in the Ottoman Turkish treebank. Table 6 provides the distribution of the patterns across both corpora. In total, 224 cases out of 800 plural nouns were pluralized with either Arabic or Persian grammar rules, corresponding to 28% of all plural nouns. However, the dominant pluralization rule was still Turkish, -lAr, by 72%, which shows the significance of Turkish morphology in the Ottoman Turkish language.

Pluralization Type	Ottoman Turkish	Modern Turkish
Turkish -lAr	576	6,535
Arabic/Persian plurals	224	0

Table 6. Observed pluralization patterns

In modern Turkish, Arabic and Persian plurals have lost their function. Nevertheless, plural

words such as *eṣnā* (moment.PL) in Ottoman Turkish can be still found in TR-BOUN as *eṣna* (moment), although they are annotated as singular because their meaning does not hold any plurality anymore.

Finally, we identified five instances where a noun was first pluralized with an Arabic pluralization rule (broken plurals), and then with the Turkish pluralization suffix, *-lar*, meaning that the word was pluralized twice. The instances of double pluralization were *evlād u nisvān u emvālleri* (their child.PL, woman.PL, and good.PL-PL-3PL.POSS), *evbāṣ-lar* (ruffian.PL-PL), and *mevācib-ler* (salary.PL-PL).

5. Discussion

This study quantitatively examines the transition from Ottoman to modern Turkish. Our quantitative findings align with the previous qualitative studies of Ottoman Turkish language. Timurtaş (1976) previously argued that Ottoman Turkish exhibited a certain tendency toward labialization, particularly in suffixes. Our data on compliance in vowel harmony confirms this claim as labial vowel harmony compliance rate in Ottoman Turkish was significantly lower than in modern Turkish. This aligns with Stachowski's (2013, 550) observation of "labially disharmonized vowel sequences" in Ottoman Turkish. While Stachowski noted these forms in grammar books, our study quantifies their prevalence in actual usage and shows that the transition to modern Turkish involved the elimination of these rounded forms in favor of the strict labial harmony described by Göksel and Kerslake (2005). Moreover, the discrepancy for the vowel harmony compliance rates can be explained by the lack of front and back vowels for some suffixes in Ottoman Turkish, like first person singular possessive suffix and genitive case suffix, as shown by Demir and Yılmaz (2002, 862). Although previous studies stated the gap in vowel harmony between Ottoman Turkish and modern Turkish, this study quantifies this gap. Thus, the study both reaffirms previous studies and demonstrates the magnitude of this gap at this scale.

Additionally, Table 4 demonstrates that palatal vowel harmony was well-established in both Ottoman Turkish and modern Turkish. A close analysis uncovers that the disharmonic instances are mainly loanwords such as *ḥakikatde* (truth-LOC) in Ottoman Turkish and *ḥakikate* (truth-DAT) in modern Turkish. The presence of persistent disharmony aligns with Göksel and Kerslake (2005) and Khalilzadeh (2010, 149). We frame these not as modern anomalies, but as remnants of Ottoman phonological constraints that resisted the general tendency in the language towards a more solid vowel harmony.

The presence of nouns that were first pluralized with Arabic pluralization rule, and then the Turkish

pluralization suffix, *-lar* provides quantitative evidence for the mechanism of reanalysis. Following Harris and Campbell (1995, 61-119), these Arabic plurals were reanalyzed by Ottoman Turkish authors as singular stems. Consequently, the mechanism of extension was triggered to align these forms with the productive agglutinative grammar.

6. Conclusion

To sum up, this study computationally compares Ottoman and modern Turkish to quantify the morphophonological transition. By utilizing annotated treebanks in UD, we demonstrated the diachronic shift in language use, particularly for vowel harmony in noun suffixes as well as the verb suffixes, the converb *-(y)Ip* and the dative infinitive *-mAyA*, and pluralization. The method goes beyond descriptive statistics and applies log-likelihood ratio test to quantify the shift further. The statistical analysis revealed that while modern Turkish achieved 98% compliance with labial vowel harmony, Ottoman Turkish complied with it by 77%. Additionally, it exhibits the obsolescence of certain allomorphs in modern Turkish. Furthermore, the analysis of plural nouns both demonstrated the disappearance of the function of Arabic and Persian pluralization rules in modern Turkish and the presence of double pluralization in Ottoman Turkish through Turkish and Arabic grammar. Future work will expand this analysis in a larger and more granular dataset in terms of period, genre, and author gender, and deploy logistic regression analysis to find the features such as period, genre, or surrounding phonological environment, that predict the observed patterns.

7. Limitations

A primary challenge in this study appears in interpreting the Ottoman Turkish data. Since Ottoman Turkish was used across six centuries, the changes in it should be acknowledged. However, due to the lack of metadata on when a given sentence was written in UD for Ottoman Turkish, the study diachronically flattens Ottoman Turkish and treats Ottoman Turkish as a monolith. Furthermore, the methodology for data collection and genres of the treebanks vary, which may affect how well they represent the language. For instance, although *-(y)lBAn* does not appear in Ottoman Turkish treebank, this does not mean the form was absent in the language at all but points to an important limitation of the data on representativeness. A simple search on canonical Ottoman Turkish sources suggests that the suffix was present in Ottoman Turkish across centuries such as *oluban* (be-CVB) in *Adlî Dîvânî* from 16th century (Sultân Bâyezîd-i Sâni, 2018, 20) and *açuban* (open-CVB) in *Mecmûa-l Gazeliyyât* from 17th century (Konevî, 2017, 191). Thus, such

limitations should be considered while interpreting the data.

8. Bibliographical References

- Alpay, N. (2015). *Türkçe Sorunları Kılavuzu*. Metis Yayınları, 5th edition.
- Birnbaum, Eleazar. (1967). The Transliteration of Ottoman Turkish for Library and General Purposes. *Journal of the American Oriental Society*, 87(2):122–156.
- Clements, G. and Sezer, E. (1982). Vowel and Consonant Disharmony in Turkish. In H. van der Hulst & N. Smith (Eds.), *The Structure of Phonological Representations (Part II)*. Berlin, Boston: De Gruyter, pp. 213–256. <https://doi.org/10.1515/9783112423325-007>
- Cristaldi, M. P. E. (2025). Between Linguistic Standardization and Nationalism: Descriptions and Grammars of the Turkish Language from the Late Ottoman to the Early Turkish Republican Era. In É. Aussant, & F. Simon (Eds.), *Documenter Et décrire Les Langues d'Asie : Histoire Et épistémologie*. Paris: Société d'histoire et d'épistémologie des sciences du langage, pp. 241–266. <https://doi.org/10.4000/138n7>
- Demir, N. and Yılmaz, E. (2002). Ottoman Turkish. *The Turks*, 3: 853–67.
- Develi, H. (1995). *Evliya Çelebi Seyahatnâmesine göre 17. yüzyıl Osmanlı Türkçesinde ses benzeşmeleri ve uyumlar* [Sound assimilations and harmonies in 17th century Ottoman Turkish according to Evliya Çelebi's Seyahatnâme]. Türk Dil Kurumu.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):51–74.
- Eckhoff, H. M., Luraghi, S., and Passarotti, M. (Eds.). (2020). *Diachronic treebanks for historical linguistics*. John Benjamins Publishing Company. <https://doi.org/10.1075/bct.113>
- Gümüşkılıç, M. (1997). *18. yüzyıl Türkçesinde ses hadiseleri* [Sound events in 18th century Turkish] [Unpublished doctoral dissertation]. Istanbul University, Faculty of Letters.
- Harris, A.C. and Campbell, L. (1995). *Historical Syntax in Cross-Linguistic Perspective*. Cambridge University Press.
- Kerslake, C. (2015). Ottoman Turkish. In L. Johanson & É. Á. Csató (Eds.), *The Turkic Languages* (2nd ed., pp. 179–202). Routledge. <https://doi.org/10.4324/9781003243809>
- Khalizadeh, A. (2010). Vowel Harmony in Turkish. *Karadeniz Araştırmaları*, 6(24):141–150.
- Konevi, N. b. M. (2017). *Mecmûa-i gazeliyyât* (M. Özdemir, Ed.). T.C. Kültür ve Turizm Bakanlığı, Istanbul.
- Lewis, G. L. (1977). *Turkish* (14th ed.). New York: David McKay Company Inc.
- Lijffijt, J. (2013). *Computational methods for comparison and exploration of event sequences* (Aalto University publication series Doctoral Dissertations 205/2013) [Doctoral dissertation, Aalto University]. <http://urn.fi/URN:ISBN:978-952-60-5475-9>
- Lijffijt, J., Säily, T., and Nevalainen, T. (2012). CEEcing the baseline: lexical stability and significant change in a historical corpus. In *Variation, Contacts and Change in English*, 10. https://varieng.helsinki.fi/series/volumes/10/lijffijt_saily_nevalainen/ (accessed 25 December 2025)
- Göksel, A. and Kerslake, C. (2005). *Turkish: A comprehensive grammar*. Routledge.
- Montes, T., Manrique-Gómez, L., and Manrique, R. (2024). Historical Ink: Semantic Shift Detection for 19th Century Spanish. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 29–41, Bangkok, Thailand, August. Association for Computational Linguistics.
- Nevalainen, T., Raumolin-Brunberg, H., and Mannila, H. (2011). The diffusion of language change in real time: Progressive and conservative individuals and the time depth of change. *Language Variation and Change* 23(1):1–43.
- Özcelik, Ö. (2024). Vowel harmony. In *The Phonology of Turkish*. Oxford: Oxford Academic.
- Petrova, Y. A. and Kuleshova, K. A. (2020). Cultural and historical evolution of the Turkish language under the influence and borrowing of European languages. *Nauchnyj al'manakh stran Prichernomor'ya [Science Almanac of Black Sea Region Countries]*, 21(1):65–70. <https://doi.org/10.23947/2414-1143-2020-21-1-65-70>
- Rayson, P., Damon B., and Brian F. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In *Le poids des mots: Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data (JADT 2004)*, pages 926–36.
- Sultân Bâyezîd-i Sâni. (2018). *Adlî divânı (Divân-ı Sultân Bâyezîd-i Sâni)* (Y. Bayram, Ed.). T.C. Kültür ve Turizm Bakanlığı, Istanbul.
- Stachowski, M. (2013). On the evolution of Ottoman Turkish vowel harmony and the (il)labiality of the suffix +lık. In B. Gül, F. Ağca & F. Gökçe (Eds.), *Bengü Bitig*. Ankara: Öncü Kitap, pp. 551–558.
- Timurtaş, F. (1976). Küçük Eski Anadolu Türkçesi Grameri. *Türkiyat Mecmuası*, 18, pp. 331–368.
- Yağmur, Ö. (2025). Ottoman to modern Turkish: Persian izafet and its colloquial variations. *TEKE (Türkçe Eğitimi ve Kültürü Dergisi)*, 14(3):821–45. <https://doi.org/10.7884/teke.1639691>

9. Language Resource References

- Marşan, B., Akkurt, S. F., Şen, M., Gürbüz, M., Güngör, O., Özateş, Ş. B., Üsküdarlı, S., Özgür, A., Güngör, T., and Öztürk, B. (2022). *Enhancements to the BOUN Treebank reflecting the agglutinative nature of Turkish*. arXiv.
<https://doi.org/10.48550/arXiv.2207.11782>
- Yıldılıoğlu, E. and Siewert, J. (2025). DUDU: A treebank for Ottoman Turkish in UD style. In Š. A. Holdt, N. Ilinykh, B. Scalvini, M. Bruton, I. N. Debess, & C. M. Tudor (Eds.), *The Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025)*, pages 74–79, Tallinn, Estonia, March. University of Tartu Library.
<https://aclanthology.org/2025.resourceful-1.16/>

Word Segmentation for UD: a Comparison of IsiZulu and Sepedi

Laurette Marais¹, Laurette Pretorius²

¹NLP Research Group, CSIR, Pretoria, South Africa

²Division of Computer Science, Mathematical Sciences, Stellenbosch University, South Africa

¹laurette@loerie.org.za, ²lpretorius@sun.ac.za

Abstract

The Southern Bantu language family contains languages with so-called conjunctive orthographies and disjunctive orthographies. In languages with conjunctive orthographies, such as isiZulu, orthographic words correspond to linguistic words, whereas in languages with disjunctive orthographies, prefix morphemes of verbs and other predicates are written as disjunct, orthographic words. When developing Universal Dependencies treebanks, the basic principle is to consider syntactic (linguistic) words, but for languages with agglutinating morphology, it has been argued that this reduces the informativeness of the treebank. In this paper we investigate this claim by analysing and measuring the effects of annotating universal dependencies on the basis of orthographic words on two morphosyntactically parallel treebanks for isiZulu and Sepedi.

Keywords: disjunctive orthography, conjunctive orthography, token, syntactic word, annotation, UD treebank

1. Introduction

The question of *wordhood* (What is a word?) is central to describing and annotating the grammatical structure (morphology and syntax) of any language, including Southern Bantu languages¹ such as Sepedi and isiZulu. By distinguishing between an *orthographic word*² and a *linguistic word*³, the two writing systems prevalent in the Southern Bantu languages may be described as *disjunctive*, where a single linguistic word may be represented by a number of orthographic words, and *conjunctive*, where a single linguistic word is represented by a single orthographic word (Kosch, 2006, pp.2-4).

Sepedi has a disjunctive orthography and isiZulu a conjunctive one. Moreover, these languages exhibit a rich agglutinative morphology and are closely related in terms of grammatical structure. We illustrate this by means of the example in Table 1:

In the context of the Universal Dependencies⁴ (UD) framework, an orthographic word is usually referred to as a *token* and a linguistic word as a *syntactic word*.

In principle, UD annotation is based on a lexicalist view of syntax, which means that dependency relations hold between words and the basic units of annotation are syntactic words. However, the

Language	Sentence POS	# Words (orth,ling)
English	I see them PRON VERB PRON	3,3
isiZulu	<i>ngiyababona</i> VERB	1,1
Sepedi	<i>ke a ba bona</i> VERB	4,1

Table 1: Orthographic words vs. linguistic words.

example in Table 1 raises the question of marking boundaries of syntactic words in UD. It is noted⁵ that “[t]here are also situations where multiple surface tokens act as one (morpho)syntactic word. The standard solution is that these tokens are kept as independent nodes in the dependency structure and they are connected using specialized relations (compound, flat, fixed, goeswith)”.

However, in the case of Sepedi, these specialized relations are less reflective of the morphosyntactic structure of multi-token words such as in Table 1. For example, *ke* (I) and *ba* (them) play the role of the subject and object of the sentence, and is therefore more accurately annotated with the syntactic relations *nsubj* and *obj*, respectively.

It remains an open question which annotation strategy in UD is best for morphologically rich languages (see also (Goldman et al., 2025)). The Southern Bantu languages provide a unique angle from which to investigate this question due to their divergent orthographies while being highly similar

¹These languages include the Nguni group of languages, viz. isiZulu, isiXhosa, Siswati and isiNdebele, and the Sotho-Tswana group, viz. Sepedi, Setswana and Southern Sotho.

²Defined by Kosch (2006) as “a sound [grapheme] or a sequence of sounds separated from other sounds or sequences of sounds by means of spaces”.

³Defined by Kosch (2006) as “a unit which has its own independent meaning”.

⁴<https://universaldependencies.org/>

⁵https://universaldependencies.org/workgroups/newdoc/word_segmentation.html

morphosyntactically.

Gaustad et al. (2024) annotated a Setswana corpus of 20 sentences by using syntactic relations to avoid reducing “the granularity and informativeness of the treebank”. Setswana belongs to the same language group as Sepedi and they share a disjunctive orthography.

Given this decision, the question then for a language like isiZulu (also closely related but with a conjunctive orthography) is whether some kind of sub-word tokenisation would be beneficial to combat the same loss of informativeness. However, morphological surface tokenisation for isiZulu is non-trivial due to widespread morphophonological alternation (Pretorius and Bosch, 2003). How then to approach this question?

In this paper, we exploit the high structural similarity between Sepedi and isiZulu as well as their respective, standardised orthographies to investigate and to some degree quantify the effect of tokenising at the sub-syntactic-word level.

Since neither language is yet represented within the UD project, we develop and utilise two parallel treebanks of 94 sentences each as the basis for our experiment, annotated at the token level.

In Section 2 we introduce our parallel treebank and describe how it was designed to provide a truly parallel setup via the Grammatical Framework formalism.

Sections 3 and 4 attempt to shed light on the question of informativeness of a treebank from an intrinsic and extrinsic perspective, respectively. Section 3 considers the differences between the treebanks from a linguistic and complexity point of view, while Section 4 shows how these differences affect downstream NLP applications, with our example being a comparison between UDPipe parsers trained on the two treebanks.

We provide some conclusions in Section 5 and note directions for future work.

2. Designing Parallel Treebanks

One way of approaching our question about informativeness might be to compare an isiZulu treebank annotated on orthographic words (tokens) to an isiZulu treebank annotated on some systematic subword segmentation. (Or, conversely a Sepedi treebank annotated on tokens compared to a Sepedi treebank where tokens are grouped together into linguistic words.) However, as mentioned above, morphological surface segmentation is not a trivial task for isiZulu. Hence, it is not always obvious where morphological surface boundaries should be inserted into a token. Moreover, it is not always clear what granularity of segmentation is best – should all morpheme boundaries be marked or does it make sense to consider some sequences

of morphemes together, such as ignoring the morpheme boundary between the pre-prefix and basic prefix of nouns?

Instead of speculating about alternatives for a specific language, we exploit the fact that isiZulu and Sepedi are morphosyntactically highly similar, but have different orthographies. These standardised orthographies provide a natural, linguistically informed basis for considering the question. Our experiment is performed on a treebank designed to be highly parallel in terms of the morphosyntax of the sentences.

2.1. GF Resource Grammars for IsiZulu and Sepedi

Grammatical Framework (GF) is a formalism, programming language and software ecosystem that enables multilingual grammar engineering. One of the key features of GF is its distinction between abstract syntax and concrete syntax, where an abstract syntax tree (AST) is typically a language independent representation of a sentence, with one or more corresponding concrete syntax trees (CSTs) that describe how the AST is realised, or linearised, in different natural languages.

The GF Resource Grammar Library (RGL) (Ranta, 2009) consists of an abstract syntax core of linguistic functions that attempt to model universal syntax functions of languages, such as predication, determination and modification. The concrete implementation of this for a specific language is called a GF Resource Grammar.

GF Resource Grammars (RGs) for isiZulu and Sepedi have been developed⁶. In addition to implementing concrete syntax linearisation functions for syntax functions in the existing RGL, a significant number of Bantu-language specific abstract syntax functions were defined in so-called extension modules. These functions were first developed for use by the isiZulu RG, and were refined when the Sepedi RG was developed. These two RGs therefore share an abstract syntax, and due to the close similarity between the languages, their concrete syntaxes share many similarities reflecting the shared characteristics of nominal classification and concordial agreement (Faaß et al., 2012).

For example, consider the two GF trees in Figure 1 representing the sentence ‘He praises you’ in both languages. The two trees not only show the identical abstract syntax tree shared by the two sentences, but also the striking similarities in how the sentences are linearised in the two languages.

⁶Available at <https://github.com/LauretteM/gf-rgl>

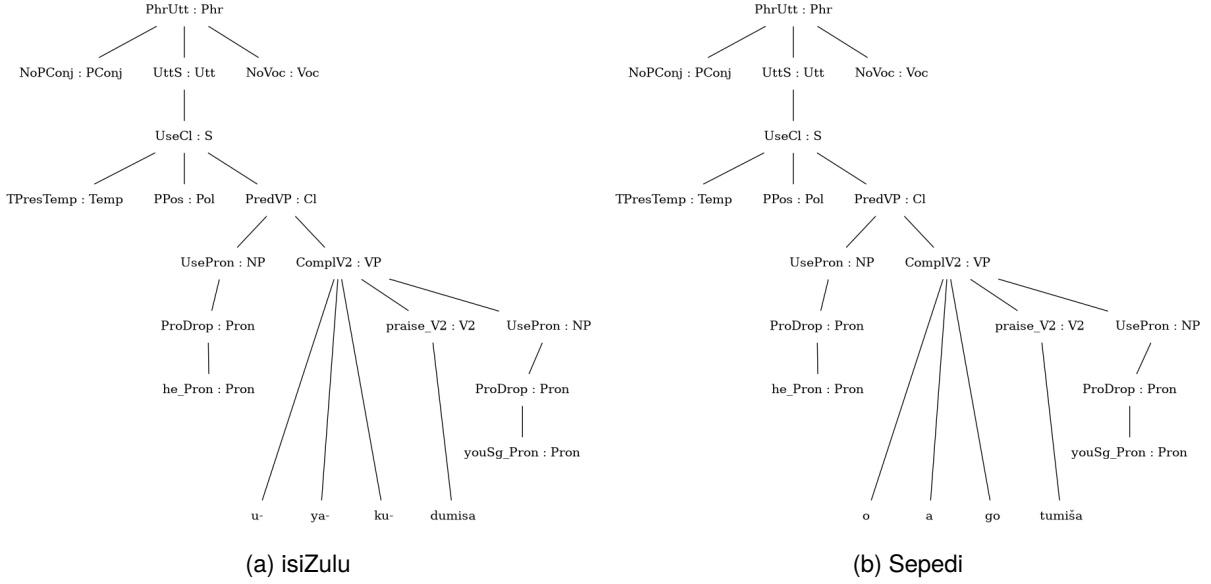


Figure 1: GF ASTs for the sentence ‘He praises you.’

2.2. A Parallel GF Treebank

Our starting point for setting up our experiment was a GF treebank of 100 isiZulu sentences. These sentences were captured from an isiZulu textbook (Taljaard and Bosch, 1988), and a large variety of morphosyntactic constructions are represented in the treebank. The original purpose of the treebank was to serve as a regression test treebank during development of the isiZulu Resource Grammar. In previous work (Marais et al., 2024), the abstract syntax trees of the treebank also served as a regression test during development of a Siswati Resource Grammar, where the trees would be linearised in Siswati using the Siswati RG, and compared to the translations of the original isiZulu sentences in a parallel Siswati textbook by the same authors.

To obtain a parallel isiZulu-Sepedi treebank, i.e. a multilingual GF treebank, we followed a similar approach by linearising the abstract syntax trees using the Sepedi RG to obtain Sepedi sentences. These generated sentences were then manually evaluated. Since isiZulu and Sepedi belong to different language groups within the Southern Bantu family, it was expected that some abstract syntax trees would not linearise to valid Sepedi sentences. Indeed, this was the case for 6 of the 100 sentences, primarily due to lexical mismatches between the languages and the fact that Sepedi only has one basic past tense, while isiZulu has two.

In any case, 94 of the abstract syntax trees originally obtained via isiZulu sentences did linearise into correct Sepedi sentences, and these 94 pairs of sentences that share an abstract syntax tree form the basis of our experiment.

The significance of the shared abstract syntax tree is that it allows us to define in which way the

sentences are identical and therefore in what sense the resulting isiZulu and Sepedi treebanks are parallel. Equivalence at the level of abstract syntax trees can be considered as equivalence of the deep structure, while the high degree of similarity between the implementations of the two RGs contributes to near-equivalence of the surface structure.

2.3. Parallel UD Treebanks

Before we discuss our annotation approach, we briefly return to two aspects of Bantu grammar that are important when discussing UD annotation.

Firstly, the grammatical (morphological and syntactic) structure is based on two systems: nominal classification (the noun class system) and concordial agreement (the system of concords). In most Bantu languages, each noun belongs to one of between 15 and 18 noun classes (Marten, 2021). IsiZulu has 17 noun classes and Sepedi has 18. A concord is a structural element (agreement marker/morpheme) which formally marks the relationship between a noun and other words in a sentence, which must be observed in all parts of the utterance which are linked to the noun. Word categories such as verbs, pronouns, adjectives, relatives, possessives etc. are brought into concordial (i.e. grammatical) agreement by means of these concords⁷. It should be clear that such information is essential for useful UD annotation. Therefore, we comprehensively include noun class and agreement features in our annotations.

Secondly, pronominalisation in the languages under discussion is, in general, not a process of *sub-*

⁷Detailed expositions may be found in (Poulos and Msimang, 1998) and (Poulos and Louwrens, 1994).

stitution, as in English⁸, but a process of *deletion* (Louwrens, 1991, pp.91-95). In particular, the subject and object concord are not *primarily* pronominal forms, but agreement morphemes which mark the syntactic relationship between subject and object nouns and verbs. Only when the noun is deleted, does the concord acquire the *secondary* status of being a pronominal form due to the co-referential relationship which exists between the concord and the deleted noun.

For example, *monna o a sepela* (the man walks) becomes *o a sepela* (he walks) by *deleting* the noun from the sentence. There is no substitution taking place and the subject concord *o* now acquires a pronominal function.

For this reason, we follow Gaustad et al. (2024) by annotating Sepedi subject and object concords with the UPOS `PRON`, even in sentences where they do not perform a pronominal function. Our reason for this is that the UPOS tagset offers no better alternative. The only other possible option is `PART`, which is already used for various other morphemes. Using `PRON` throughout is consistent.

The two treebanks were annotated manually by adding dependency relations and UPOS (Table 2), using as point of departure the approach proposed by Gaustad et al. (2024). We also list the morphological information and features that we employ (Table 3).

Ann. types	Labels
Relations	advmod, conj, cop, det, nmod, nsubj, obj, obl, root, vocative, xcomp, case, compound, expl, mark
Subrelations	det:emph, obl:lmod, nmod:poss, advmod:lmod, obl:tmod
UPOS	ADJ, ADV, AUX, NOUN, PRON, VERB, DET, PART

Table 2: IsiZulu and Sepedi: UD annotation labels for relations and UPOS (**only Sepedi**).

A closer look at feature annotation shows that in the Setswana treebank the only universal feature (and value) that was employed is `Polarity=Neg`. The only miscellaneous feature (and values) was `NounClass` and `Bantu1, ..., Bantu17`.

We decided to separate annotation of inherent agreement information (using universal features `Person` and `Number` and miscellaneous feature `NounClass`, in line with Gaustad et al. (2024)) from concordial agreement information (new miscellaneous features `Agreement` and `AgreementObj`).

⁸For example, the noun in the sentence ‘The man walks.’ is pronominalised to ‘He walks.’ by *substituting* the noun with the pronoun.

Feature	Value
Universal	
Case	Com, Loc
Mood	Imp, Sub
Number	Sing, Plur
Person	1, 2
Polarity	Neg
Tense	Past, Pres
VerbForm	Inf
Miscellaneous	
Agreement	1ps, 2ps, 1pp, 2pp Bantu1, ..., Bantu17, Bantu18
AgreementObj	1ps, 2ps, 1pp, 2pp Bantu1, ..., Bantu17, Bantu18
NounClass	Bantu1, ..., Bantu17, Bantu18

Table 3: IsiZulu and Sepedi: UD annotation labels for features and their values (**only Sepedi**).

Examples showing the reasoning for this are given in Appendix 8.1.

In terms of dependency relations and UPOS, the annotation for the Sepedi and isiZulu treebanks followed as closely as possible the approach of the Setswana treebank (Gaustad et al., 2024), as summarised in Appendix 8.2. However, regarding morphological features, our annotation was more extensive, as shown in Table 3⁹.

3. Intrinsic Comparison

We start our comparison of the intrinsic differences between the treebanks by noting some token statistics, shown in Table 4. The low type-token ratio (TTR) of Sepedi relative to isiZulu is expected, given that frequently occurring morphemes occur as separate tokens, whereas these same morphemes contribute to unique tokens (types) in the isiZulu treebank. We also note that for Sepedi the average tokens per sentence (which is equal to the average relations per sentence) is higher than that for isiZulu, which is to be expected.

In comparing the two treebanks, we follow two approaches. We first highlight the way in which these differences manifest themselves from a linguistic point of view by walking through a number of example sentences.

Then we consider the issue from a quantitative perspective, in which we measure the entropy – and as such the informativeness – of the dependency relations. This serves to confirm and to some

⁹In the Bantu languages the third person is made up of a large variety of noun classes where each noun class has its own “he/she/it” and “they” (Louwrens, 1994, p.198).

Metric	isiZulu	Sepedi
Sentences	94	94
Tokens	221	352
Types	167	158
Type-Token Ratio	0.76	0.45
Avg Tokens/Sentence	2.35	3.74

Table 4: Token statistics for isiZulu and Sepedi treebanks.

<i>u-</i> SC1	<i>ya-</i> LongPres	<i>ku-</i> OC2ps	<i>dumisa</i> VStem
<i>o</i> SC1	<i>a</i> LongPres	<i>go</i> OC2ps	<i>tumiša</i> VStem

Table 5: Comparative morphological analysis of ‘He praises you’.

degree quantify the intuition that annotating on the basis of the orthography (i.e. tokens) of Sepedi and isiZulu would lead the Sepedi treebank to be more informative.

3.1. Linguistic Considerations

Consider the trees shown in Figure 2, which are the isiZulu and Sepedi renderings for ‘He praises you’.

It is immediately obvious that the isiZulu tree encodes significantly less information in its dependency relations than the Sepedi tree, since the isiZulu tree consists of a single node. This is typical of the way the orthography affects how verbs are annotated. However, as shown in Table 5, the morphosyntactic structure of the two sentences is identical.

To mark the morphological features of the sentences, we use the custom `NounClass`, `Agreement` and `AgreementObj` features.

For the Sepedi sentence, the `PRON` element marked by `nsubj` is annotated with `NounClass=Bantul` and the `PRON` element marked by `obj` is annotated with `Person=2|Number=Plur`. This reflects the fact that pronouns typically supply agreement information in a sentence.

For the isiZulu sentence, we can only annotate the verb, and hence we annotate it using `Agreement=Bantul` and `AgreementObj=2ps`. This is necessary since two agreement values are present in the verb, and none of them are inherent features of the verb *to praise*.

We may simply note, at this point, that our decision to follow Gaustad et al. (2024) in assigning `PRON` to subject and object morphemes now causes the morphological annotation for Sepedi and isiZulu to diverge significantly on near identical

morpheme sequences.

More importantly, essential information is lost in the case of isiZulu, since the treebank does not communicate which morphs are present in the verb and have given rise to the morphological annotation. Clearly, a language model trained on the Sepedi sentence would conceivably be better able to generalise to the new sentence ‘He praises him (the boy)’ than a model trained on the isiZulu sentence. The new Sepedi sentence would be *o a o tumiša*, while the isiZulu sentence would be the token *uyamdu-misa*.

The inability in the isiZulu treebank to isolate the morphemes that function as pronouns in the sentence has some other strange side-effects. Consider the sentence in Figure 3. We again give a (slightly simplified) morphological analysis in Table 6, showing the high degree of similarity between the sentences.

Here, in the Sepedi sentence, the subject concord *ke* (‘I’) functions as the unmarked subject of the sentence, while the addition of *nna* has a determinative and emphatic function. This can easily be accommodated in the Sepedi treebank as shown in Figure 3. However, the isiZulu sentence mirrors the morpheme sequence of the Sepedi sentence exactly, and *mina* performs the exact same function in the isiZulu sentence as *nna* in the Sepedi. An annotation that captures this leads to the tree in Figure 3, where the relation between *mina* and the verb *ngihamba* is marked as `det : emph`. At best, this strikes us as linguistically awkward.

3.2. Entropy with Respect to Dependency Relations

We turn now to comparing the information contained in the dependency relations of the two treebanks. Specifically, we want to know how informative dependency relation annotations are for each of the tokenisation strategies.

To measure informativeness, we calculate the Shannon entropy over dependency relations for each treebank. This measures on average how surprising, or stated otherwise, informative, dependency relations are within each treebank.

Shannon entropy H over dependency relations X is defined as

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

where $X = \{x_1, x_2, \dots, x_n\}$ is the set of dependency relations for a given treebank, and $p(x_i)$ is the relative frequency of x_i .

The resulting number for H represents the average number of bits needed to encode a dependency relation, and hence the higher the number,

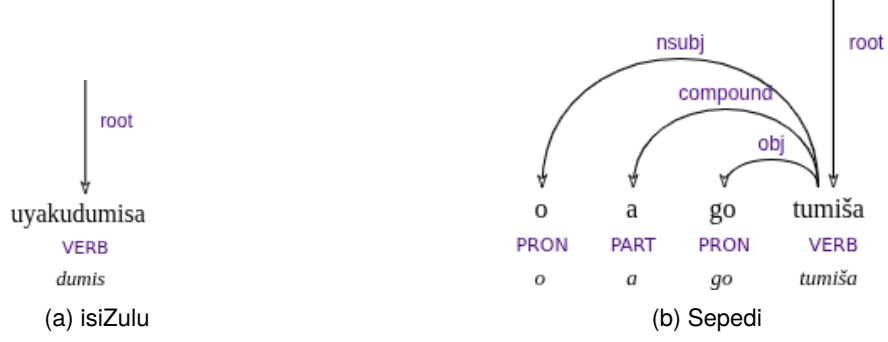


Figure 2: Trees for the sentence *He praises you*.

<i>mina</i>	<i>ngi-</i>	<i>hamba</i>	<i>ne-</i>	<i>ntombi</i>
Pron1ps	SC1ps	VStem	AdvPre	N9
<i>nna</i>	<i>ke</i>	<i>sepela</i>	<i>le</i>	<i>mosetsana</i>
Pron1ps	SC1ps	VStem	AdvPre	N1

Table 6: Comparative morphological analysis of ‘I myself walk with a girl’.

the more informative, on average, dependency relation annotations are within the treebank.

Finally, to get a sense of how informative the dependency relation annotations are per sentence, we can multiply H with the average number of relations per sentence.

The results are given Table 7 and clearly show that, per sentence on average, the Sepedi treebank is more than twice as informative as the isiZulu treebank.

4. Extrinsic Comparison

Given the significant difference in informativeness of the annotation approaches for the two languages, the next question to consider is what practical effects this may have in using the treebanks for downstream tasks. One such task is to train a parser on the treebanks.

UDPipe is an open source trainable pipeline for training dependency parsers on treebanks in the CoNLL-U format (Straka et al., 2016). In order to investigate the two different annotation approaches on training dependency parsers, we run the UDPipe training script on both treebanks, following a 10-fold cross-validation approach to mitigate the small sizes of the treebanks.

Since the Sepedi treebank is significantly more informative, annotating a Sepedi sentence is a harder task than annotating an isiZulu sentence. On the other hand, the sentences in the parallel treebanks are essentially equivalent, and hence a learning algorithm may benefit from a more informative annotation scheme.

Our experiment consists of training four different models on each treebank. We train two base-

line models in which both a tagger and parser are trained, and two models where the gold standard UPOS tags are provided and only a parser is trained. For both the baseline and gold UPOS models, we train one version each that includes the morphological feature annotations, and one model each in which the features are removed before training. Since much of our annotations were added to the miscellaneous column (which is ignored by UDPipe), for the models trained on data that includes morphological features, we merge the miscellaneous features with the universal features before training.

The results for labeled attachment scores (LAS) and unlabeled attachment scores (UAS) are given in Table 8. Table 9 shows the performance of the tagger trained on the baseline models.

The first thing we notice is that the unlabeled attachment score (UAS) is always better for Sepedi than for isiZulu. In addition, the labeled attachment scores (LAS) provide some useful insights into the relative informativeness of the treebanks, visualised in Figure 4.

For the Sepedi, model performance improves in an intuitive way. Performance on the baseline model is better when the features are supplied. If we compare the baseline model with features to the gold UPOS model without features, we only see a small improvement, suggesting that the UPOS tagger trained when features are present produces a tagger that is about as useful as having gold standard UPOS tags (without features). Measuring the UPOS tagger directly, we indeed find that it has an accuracy of 88.39% (Table 9). The biggest gain comes from providing gold standard morphological features in addition to gold standard UPOS

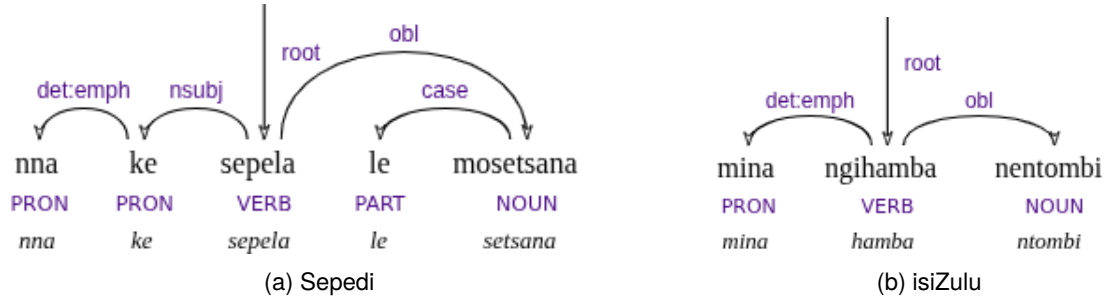


Figure 3: Trees for the sentence ‘I myself walk with a girl’.

Measure	isiZulu	Sepedi
Relations/sentence	2.351	3.744
Avg bits/relation	2.477	3.222
Avg bits/sentence	5.824	12.067

Table 7: A comparison of the informativeness of dependency relation annotations for both treebanks.

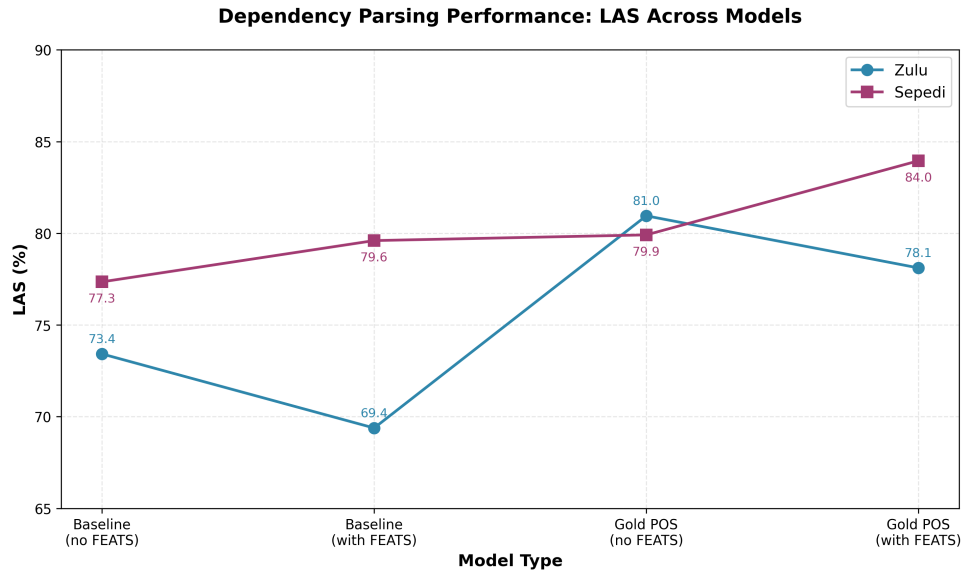


Figure 4: Labeled attachment score across model types.

Model	isiZulu	Sepedi
<i>LAS (%)</i>		
Baseline (no FEATS)	73.42	77.35
Baseline (with FEATS)	69.38	79.60
Gold UPOS (no FEATS)	80.95	79.91
Gold UPOS (with FEATS)	78.11	83.95
<i>UAS (%)</i>		
Baseline (no FEATS)	83.77	89.99
Baseline (with FEATS)	77.41	93.53
Gold UPOS (no FEATS)	89.96	93.00
Gold UPOS (with FEATS)	88.83	95.61

Table 8: Dependency parsing performance (LAS and UAS) across model configurations for isiZulu and Sepedi.

Metric	isiZulu	Sepedi
<i>Baseline (no FEATS)</i>		
POS Accuracy	74.63	88.39
FEATS F1	—	—
<i>Baseline (with FEATS)</i>		
POS Accuracy	71.22	85.59
FEATS F1	21.86	44.78

Table 9: POS accuracy and FEATS F1 scores for baseline models on isiZulu and Sepedi.

tags. This shows that UDPipe is not learning to produce all relevant features and is helped when they are provided. This is confirmed by the measured F1 score of the predicted features, which is only

44.78%.

The picture for isiZulu is very different. Apart from the gold UPOS model without features (where the isiZulu model is slightly better), isiZulu models consistently perform worse than their Sepedi counterparts. What is also notable is that model performance degrades whenever features are added. Indeed, we see the baseline model's tagger struggling to learn the features, with an F1 score of 21.86%. The fact that adding features degrades performance could suggest overfitting or a search space that is too large given the small amount of training data. The latter is very likely on such a small dataset as ours, given the well-known problem of data sparsity when it comes to a conjunctively written, agglutinative language such as isiZulu. This has been shown for sequence-to-sequence models to be mitigated by performing morphological segmentation (Mkhwanazi and Marais, 2024). The superior performance of the Sepedi model, which is in effect morphologically segmented to a large degree, confirms this.

5. Conclusion

We have presented parallel UD treebanks for isiZulu and Sepedi that have their origin in a multilingual GF treebank¹⁰. This means that each pair of sentences in the treebanks are linearisations of the same GF abstract syntax tree. Given the high linguistic similarity between the two languages, we have what is in effect treebanks that are near-identical from a morphosyntactic point of view. This has allowed us to compare the informativeness of the two treebanks when they are annotated as UD treebanks *at the token level*.

Our intrinsic comparison consisted of considering linguistic effects of the two annotation strategies and measuring the relative entropy of the relations in the treebanks. The latter showed that Sepedi token annotations are twice as informative as isiZulu token annotations. Our extrinsic comparison centred around the performance of dependency parsers trained on the treebanks, where Sepedi parsers outperform isiZulu parsers in a variety of model configurations.

From these comparisons, it is clear that there are significant advantages in annotating a Sepedi treebank at the token (and hence sub-syntactic) level, while annotation of an isiZulu treebank may share the same benefits if morphological surface segmentation of tokens was performed.

We relied on the standardised orthography of Sepedi to determine sub-syntactic elements for annotation. Given the non-triviality of morphological surface segmentation for isiZulu, future work might

involve determining what systematic for segmentation would best serve isiZulu within the UD framework. However, our results indicate that a segmentation strategy that mirrors the standardised Sepedi orthography would provide significant increases in informativeness.

6. Future work

A full contextualisation of this comparative analysis of two Bantu languages in terms of similar work for other languages (see, for example, (Goldman et al., 2025)) is beyond the scope of this paper, but forms part of important future work.

In this paper we performed an analysis on a small, high quality dataset. Although the sentences in our corpus are relatively short, the investigation centred around the interaction of the morphology and orthography of isiZulu and Sepedi. As such, our analysis would generalise to any corpus containing longer sentences. Future work might entail a similar analysis of a larger corpus.

7. Bibliographical References

- Gertrud Faaß, Sonja Bosch, and Elsabé Taljard. 2012. Towards a Part-of-Speech Ontology: Encoding Morphemic Units of Two South African Bantu Languages. *Nordic Journal of African Studies*, 21(3):23–23.
- Tanja Gaustad, Ansu Berg, Rigardt Pretorius, and Roald Eiselen. 2024. The first Universal Dependency Treebank for Tswana: Tswana-Popapolelo. In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages@LREC-COLING 2024*, pages 55–65.
- Omer Goldman, Leonie Weissweiler, and Reut Tsarfaty. 2025. Workshop of The UniDive 2025 Shared Task on Multilingual Morpho-Syntactic Parsing: Proceedings of the Workshop.
- Inge M. Kosch. 2006. *Topics in morphology in the African language context*. Unisa Press.
- Louis J. Louwrens. 1991. *Aspects of Northern Sotho grammar*. Via Afrika Ltd.
- Louis J. Louwrens. 1994. *Dictionary of Northern Sotho grammatical terms*. Via Afrika, Pretoria, South Africa.
- Laurette Marais, Laurette Pretorius, and Lionel Clive Posthumus. 2024. Bootstrapping syntactic resources from isiZulu to Siswati. In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages@LREC-COLING 2024*, pages 77–85.

¹⁰The data is available at <https://github.com/LauretteM/gf-bantu-resources>.

- Lutz Marten. 2021. [Noun Classes and Plurality in Bantu Languages](#). In *The Oxford Handbook of Grammatical Number*. Oxford University Press.
- Sthembiso Mkhwanazi and Laurette Marais. 2024. Generation of segmented isiZulu text. *Journal of the Digital Humanities Association of Southern Africa*, 5(1).
- George Poulos and Louis J. Louwrens. 1994. *A linguistic analysis of Northern Sotho*. Via Afrika Ltd.
- George Poulos and Christian T. Msimang. 1998. *A linguistic analysis of Zulu*. Via Afrika Ltd.
- Laurette Pretorius and Sonja E Bosch. 2003. Computational aids for Zulu natural language processing. *Southern African Linguistics and Applied Language Studies*, 21(4):267–282.
- Aarne Ranta. 2009. The GF resource grammar library. *Linguistic Issues in Language Technology*, 2.
- Milan Straka, Jan Hajic, and Jana Straková. 2016. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.
- P.C. Taljaard and S.E. Bosch. 1988. *Handbook of isiZulu*. J.L. Van Schaik, Pretoria.

<i>Sepedi</i>			<i>isiZulu</i>		
Token	UPOS	Features	Token	UPOS	Features
yo	PRON	Agreement=Bantu1	le	PRON	Agreement=Bantu9
monna	NOUN	NounClass=Bantu1	ndoda	NOUN	NounClass=Bantu9
ga	ADV	Polarity=Neg	ayinamfazi	NOUN	Polarity=Neg
a	PRON	Agreement=Bantu1			Agreement=Bantu9
na	VERB	—			NounClass=Bantu1
mogatša	NOUN	NounClass=Bantu1			

Table 10: Morphological features in the sentence ‘This man has no wife’ in Sepedi and isiZulu.

8. Appendices

8.1. Appendix A: Morphological Annotation Examples

In Table 10, we give an example that contains both inherent morphological features and concordial features of tokens. In the subject noun phrases of the sentences (*yo monna* and *le ndoda*), the nouns are annotated using the `NounClass` feature, indicating that it is an inherent feature, and the demonstrative pronouns preceding them are annotated using the `Agreement` feature, indicating that the feature is concordial.

In the predicates of the sentences (*ga a na mogatša* and *ayinamfazi*), the annotation reflects the different orthographies, with each token in the Sepedi annotated with a single feature, and the single token in isiZulu annotated with multiple features.

In this case, a distinction between inherent and concordial features may not be strictly necessary for the Sepedi, since each token has a single feature annotation. However, the morphologically complex token *ayinamfazi*, which includes a subject agreement morpheme (*-yi-*), as well as the copulative base noun (*-mfazi* belonging to noun class 1), requires that we annotate two kinds of agreement information.

In addition, utilising two sets of features `NounClass` and `Agreement` reflects accurately in the Sepedi which tokens provide the agreement information, and which tokens reflect the agreement information of other elements of the sentence (see Figure 5).

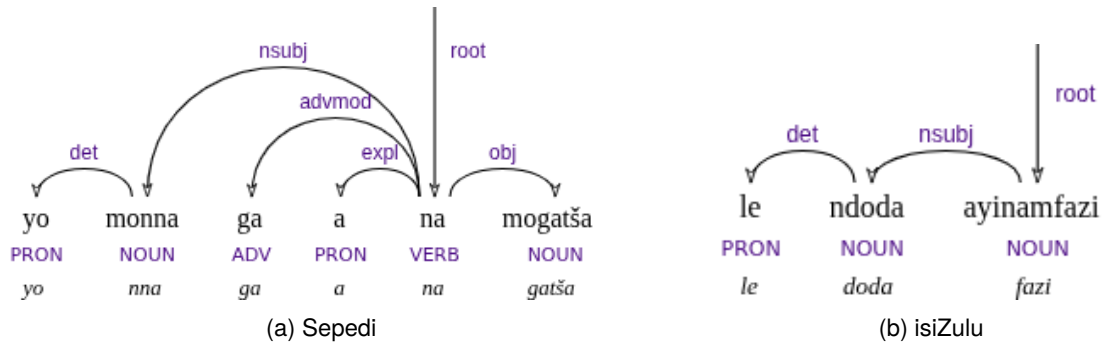


Figure 5: Trees for the sentence ‘This man has no wife’ in Sepedi and isiZulu.

Table 11 provides another example of how inherent and concordial features are used in both languages (see Figure 2 for their trees).

In the Sepedi sentence, the subject and object agreement morphemes (*o* and *go*) are annotated using inherent features `Person`, `Number` and `NounClass`. However, this annotation does not make sense for isiZulu. Although the subject and object agreement morphemes (*u-* and *ku-*) fulfill exactly the same role in the sentences, it makes no linguistic sense to annotate a verb token as though it has inherent agreement features, and two different inherent agreement features to boot. Our schema for morphological annotation therefore diverges for the two languages on this point, but we feel this strikes the correct balance between consistency across closely related languages and linguistic integrity for each language on its own terms.

8.2. Appendix B: Summary of Annotations

Table 12 lists the *morphemes* and *word categories* that occur as *tokens* in the Sepedi UD treebank, together with the dependency relations and UPOS with which they are annotated. Similarly, Table 13

<i>Sepedi</i>		
Token	UPOS	Features
o	PRON	NounClass=Bantu1
a	PART	Tense=Pres
go	PRON	Person=2 Number=Sing
tumiša	VERB	

<i>isiZulu</i>		
Token	UPOS	Features
uyakudumisa	VERB	Agreement=Bantu1 Tense=Pres AgreementObj=2ps

Table 11: Morphological features in the sentence ‘He praises you’ in Sepedi and isiZulu.

Token	Relation	UPOS
noun	nsubj, obj, nmod:poss, obj:lmod, obl, root, vocative, obl:tmod	NOUN
absolute pronoun	det:emph, root	PRON
demonstrative	det, obj	PRON
possessive pronoun	nmod:poss	PRON
quantitative pronoun	det	PRON
adjective stem	root (when copulative complement)	ADJ
verb stem	root	VERB
infinitive verb stem	xcomp	VERB
subject agreement morpheme	expl or nsubj (when pronominalised)	PRON
subject agreement morpheme class 15 (<i>go</i>)	compound	PRON
object agreement morpheme	expl or obj (when pronominalised)	PRON
negative morpheme	advmod	ADV
potential morpheme	compound	PART
progressive morpheme	compound	PART
present tense morpheme	compound	PART
future tense morpheme	compound	PART
adjective agreement morpheme	mark	PART
possessive agreement morpheme	case	PART
associative, instrumental morpheme	case	PART
auxiliary verb stem	aux	AUX
identifying copulative	cop	AUX
descriptive copulative	cop	AUX
associative copulative	root	VERB
infinitive verb (<i>go</i>)	mark	PART

Table 12: Sepedi: Common tokens, their dependency relations and UPOS.

lists the *word categories*¹¹ that occur as *tokens* in the isiZulu UD treebank, together with the dependency relations and UPOS with which they are annotated. We follow the word categorisation of Poulos and Louwrens (1994) and Poulos and Msimang (1998).

¹¹Due to the conjunctive orthography of isiZulu there are no morphemes that can occur as tokens.

Token	Relation	UPOS
noun	nsubj, obj, nmod:poss, obj:lmod, obl, root, vocative	NOUN
absolute pronoun	det:emph, root	PRON
demonstrative	det, obj	PRON
possessive pronoun	nmod:poss	PRON
quantitative pronoun	det	PRON
verb	root, nsubj, obj	VERB
infinite verb	xcomp	VERB
adjective	root (when copulative complement)	ADJ
adverb	advmod, obl	ADV
auxiliary verb	root	ADV
auxiliary verb	cop	AUX

Table 13: IsiZulu: Common tokens, their dependency relations and UPOS.

CoBra: A Compound Branching Resource for Nominal Triconstituent Compounds in English and German

Carmen Schacht^{* ‡}, Isabell Landwehr^{† ‡}, Diana Davidson[†],
Konrad Grabowski^{*}, Magdalena Meiser[†], Sophia Wiedmann[†]

^{*}Ruhr University Bochum
Department of Linguistics
carmen.schacht@ruhr-uni-bochum.de

[†]Saarland University
Department of Language Science and Technology
isabell.landwehr@uni-saarland.de

[‡]These authors contributed equally to this work and share first authorship

Abstract

We present CoBra, a resource containing triconstituent nominal compounds in English and German. This addresses an understudied aspect of compound processing, since research and resources in psycholinguistics and NLP have mostly focused on two-constituent compounds. In addition, our resource covers both general and scientific language, allowing for a register-informed perspective on compounds. It provides syntactic and semantic annotation of compound structure, in particular of the branching direction (i.e. the internal embedding structure, the **Compound Branching**) and the semantic relationship between constituents. Annotations are implemented using extensions of Universal Dependencies (UD) labels. To explore applications of our new resource, we also conduct a pilot study investigating the relationship between semantic transparency and branching direction. Our results indicate that there is indeed a correlation. Overall, our resource contributes to gaining a more detailed understanding of the structure and processing of morphologically complex words within the UD framework.

Keywords: compounds, semantic relationship, semantic transparency

1. Introduction

Compounding, i.e. the process of creating a new lexical item by combining two or more individual elements, is a basic word formation strategy in many languages. Combining the lexemes *rattle* and *snake*, for instance, creates the new lexical item *rattlesnake*. Adding an additional lexeme, e.g. as in *rattlesnake habitat*, creates another new lexical item and so on. Compounds have been the focus of extensive research in psycholinguistics and NLP. In psycholinguistics, studies are often concerned with the question whether compounds are decomposed during processing (e.g. Andrews et al., 2004; Hyönä et al., 2020) or with the implicit semantic relationships between compound constituents (Benjamin and Schmidtke, 2023). Compounds have also received considerable attention in NLP, with tasks ranging from determining compound constituents (Henrich and Hinrichs, 2011) to compositionality prediction (Schulte im Walde et al., 2013).

Researchers in compound processing have created and benefited from large datasets (e.g. Gagné et al., 2019; Schmidtke et al., 2021). However, there are two gaps in current research. First, most studies and datasets have focused on compounds consisting of two nominal constituents. This leaves other compound types understudied, even though

their specific properties are also of interest for existing theories: For compounds consisting of three constituents, for instance, identifying the compound branching, i.e. the compound's internal embedding structure (compare *[rattlesnake] habitat* with *clay [flower pot]*), is an important aspect of their processing. Their ambiguous branching structure needs to be correctly identified for comprehension. Second, existing datasets describe compound structure either superficially or inconsistently across languages. Regarding Universal Dependencies (UD; de Marneffe et al., 2021) annotation, this has been previously pointed out by Svoboda and Ševčíková (2024). The annotation of compounds in UD also touches upon the broader issue of the distinction between syntax and morphology and the limits of token-based representations (see e.g. Gerdes and Kahane, 2016; Savary et al., 2023; Goldman et al., 2025).

To address these issues, we have created CoBra, a **Compound Branching** resource. Using corpus data for English and German and covering two registers (general and scientific language), we annotated 295 compounds with semantic and syntactic information. The final annotated data are provided as .conllu files containing UD-label extensions for compound structure and as .csv tables. To facilitate the annotation process, we have also developed the

CoBra Annotator, a language-agnostic annotation tool for the flexible annotation of multi-constituent compounds of various spelling variations. We also present a small pilot study investigating the correlation between semantic transparency and branching direction.

2. Related Work

2.1. Working Definition of Compound

In this paper, we consider nominal triconstituent compounds and use a relatively broad definition of the term compound: We regard all nouns modified by two other nouns as triconstituent compounds. This approach is supported by Bauer (1998, 2019): Since it is not possible to draw a clear and consistent distinction between compounds and syntactic phrases, he proposes to treat them as variants of a single construction. This view is also similar to Levi (1978)’s concept of complex nominals.

2.2. Compounds in Psycholinguistics and NLP

Psycholinguistic research in compounding has focused on some main aspects: An important area of research is the question whether compounds are decomposed during processing, understood in a holistic way or in a combination of both (Juhasz et al., 2003; Andrews et al., 2004; Kuperman et al., 2009; Baayen et al., 2010; Hyönä et al., 2020; Libben et al., 2020). Additionally, researchers have investigated how language users understand and memorize novel (i.e. not previously encountered) compounds (Meßmer et al., 2021). Another focus has been on the semantic relationship between compound constituents, since structurally very similar compounds can possess very different internal semantic relationships: The compound *chocolate cake*, for instance, describes a ‘cake made of chocolate’, while the compound *chocolate factory* describes a ‘factory for producing chocolate’. Choosing the appropriate relation is therefore a crucial task in compound understanding. Some accounts assume competition between possible relations, with greater relation entropy slowing down processing (Benjamin and Schmidtke, 2023).

Much of this research has focused on nominal two-constituent compounds, which is the prototypical compound type in Germanic languages (as observed in studies of compound acquisition in children, see e.g. Korecky-Kröll et al., 2017). However, there are also some studies on the processing and production of nominal triconstituent compounds: An early study by Geer et al. (1972) investigated the role of language user characteristics on their ability to successfully paraphrase and memorize

triconstituent compounds. Krott et al. (2004) studied the role of interfixes in compound processing, for German and Dutch. For English, Kösling and Plag (2009) analyzed the influence of branching direction on stress patterns, while Schebesta and Kunter (2022) focused on constituent durations.

In computational linguistics, research on compounds has also mostly focused on nominal two-constituent compounds. Here, tasks include determining compound constituents (Fritzing and Fraser, 2010; Henrich and Hinrichs, 2011) or identifying the compound head for coreference resolution (Tuggener, 2016). More recently, studies have been concerned with modeling semantic roles (Marelli et al., 2017; Ormerod et al., 2024), semantic transparency (Günther et al., 2020) or, similarly, predicting a compound’s compositionality (Schulte im Walde et al., 2013; Miletić and Schulte im Walde, 2023).

2.3. Previous Compound Resources

Various datasets have been created for the study of compound processing, both in psycholinguistics and NLP: A large-scale dataset for English is LADEC (Gagné et al., 2019), a database of two-constituent compounds containing compositionality ratings as well as compound and constituent properties such as frequency or family size. *CompLex* (Schmidtke et al., 2021) additionally contains eye-tracking data for two-constituent compounds, apart from linguistic characteristics and participant data. Other datasets have focused on novel noun-noun compounds (Chen et al., 2023) or variables like ratings of familiarity, imageability or sensory experience (Juhasz et al., 2015). For German, current resources include *GhoSt-NN* (Schulte im Walde et al., 2016), which contains noun-noun compounds annotated with information on compound and constituent frequency, constituent productivity and ambiguity, semantic relations between constituents and compositionality ratings. Again, these resources focus on two-constituent compounds.

2.4. Compounds in Universal Dependencies

In recent years, the annotation of morphologically complex lexical units in Universal Dependencies has received considerable interest: Examples include a proposal for the representation of multi-word expressions (Savary et al., 2023) and UniDive, a shared task on morpho-syntactic parsing, in which existing treebanks were extended as to model both morphology and syntax (Goldman et al., 2025).

Our focus is on compounds, which are currently inconsistently represented in Universal Dependencies (UD), as pointed out by Svoboda and

Ševčíková (2024). According to their survey study of five languages, three main variants of compound spelling are currently represented in Universal Dependencies (UD): closed, open compounds and hyphenated compounds.

Closed Compounds Exemplified by ‘waterfall’ in Svoboda and Ševčíková (2024) are compounds spelled as a singular string without any hyphenation or spaces. In UD, these are treated as a single token, regardless of their internal composition. As a result, any internal relations between the constituents are not captured in the dependency structure, since only the combined morpho-syntactic features of the entire form are represented.

Open Compounds Open compounds such as ‘apple pie’ (Svoboda and Ševčíková, 2024), are written as separate orthographic units, with each constituent realized as an individual token and assigned an individual token ID. In these cases, the internal structure of the compound is explicitly represented in the dependency tree: The compound head is annotated as the head of the construction, and the modifier receives the dependency label *compound* or an extended variant of this label. In contrast to closed compounds, this approach captures the internal syntactic complexity of the compound, since the relations between constituents are explicitly encoded in the head-dependent structure.

Hyphenated Compounds The hyphenated spelling variant, for example ‘father-in-law’ in Svoboda and Ševčíková (2024), is treated similarly to open compounds in UD. The hyphen is tokenized as a separate token of the type *punct*, while the individual lexical constituents are treated as separate units. As with open compounds, the compound head is annotated as the root of the substructure and the modifier is attached as a dependent. The UD relations are likewise marked as *compound* or one of the current variants. These constructions are thus represented as multi-token structures with explicitly annotated internal syntactic relations.

For languages like German, which are particularly productive in the use of closed compounds (see the data comparison in Svoboda and Ševčíková (2024)), the current use of the UD label *compound* does not capture the syntactic structure of those compounds. The internal structure remains unrepresented due to their treatment as single-unit tokens. Svoboda and Ševčíková (2024) therefore propose to make the internal structure of closed compounds accessible to syntactic annotation, while remaining compatible with existing UD principles.

In the present study, we follow Svoboda and Ševčíková (2024) in several regards. First, we annotate closed compounds as complex structures that receive syntactic structures of their own. Second, we make use of the proposed *:nmod* label extension for our embedded recursive compound structures. This applies to the two internal nominal constituents, while leaving the relation of the root as is. In order to do so, we split closed compounds into their individual constituents, an approach also suggested in other previous work (Savary et al., 2023). Each of the newly created tokens receives a UD relation and a head ID to mark the internal structure of the compound. This allows the compound’s internal organization, i.e. the branching structure, to be explicitly represented.

However, we diverge from Svoboda and Ševčíková (2024) regarding the technical realization of the split. While they propose splitting closed compounds into individual tokens by adding a plus sign to the respective modifier and then creating as many individual tokens as there are constituents, we instead propose to treat them as multi-word tokens (MWTs) in CoNLL-U, analogous to contractions or clitics (see [sta](#)). Multi-word tokens are intended exactly for cases in which syntactic boundaries are not necessarily concordant with spacing between orthographic units, which is the case for closed compounds. We realize this by inserting a token span, as currently used for MWTs in CoNLL-U, containing the original closed spelling of the compound. This span is followed by the individual constituents, each receiving the appropriate token ID corresponding to the span, as well as individual UD relations (specifically the *compound:nmod*-variant proposed by Svoboda and Ševčíková (2024)) and head IDs. Interfixes remain with the respective modifier in the FORM column of CoNLL-U, while each constituent is additionally annotated with a lemma. In this way, the orthographic integrity of the closed compound is preserved, while its internal syntactic structure is made fully explicit and accessible within the UD representation.

3. Building CoBra

3.1. Data Sources

As our data sources for our annotation, we used several corpora. We cover both general and scientific language, in order to provide a register-informed perspective on compounding. Scientific language was chosen since compounds are very typical structures in this register (Degaetano-Ortlieb, 2021). For general language, we used TIGER (Brants et al., 2004) for German and the British National Corpus (Consortium, 2007) for English. For the scientific register, we used the Much-

More Springer bilingual (English and German) corpus of medical abstracts ([muc](#)). We chose these corpora due to their open access availability, allowing us to make our dataset publicly available on the OSF platform ([osf](#)) in an annotation-only form. To access the full-text data, the respective licensing agreements of the individual corpora have to be accepted via the according distribution platform. As a first step, all corpus files were parsed and annotated with UD-labels using the Python library *stanza* ([Qi et al., 2020](#)). For each language, a random selection of the parsed files was then reviewed by two annotators, who manually identified nominal tri-constituent compounds. Triconstituent compounds were identified in 190 files (101 files for English and 89 files for German), which were subsequently annotated.

3.2. Annotation Scheme and Workflow

In the annotation process, the following criteria were used for the identification of compounds:

- The candidate noun phrase consisted of three lexical elements. Compounds with more constituents were not considered, even if they contained an embedded triconstituent compounds.
- Each lexical element was listed as a noun in a reference dictionary. For English the *Oxford English Dictionary* ([Oxf](#)) was consulted for general language and *Merriam Webster Medical* ([mer](#)) for scientific language. For German, *Digitales Wörterbuch der deutschen Sprache* ([dwd](#)) and *Duden* ([dud](#)) were consulted for general language and *DocCheck Flexikon* ([doc](#)) for scientific language. Proper names or foreign language elements were only considered if they were listed in the dictionary.
- All spelling variants were considered, i.e. written as one word, separated by hyphen(s), separated by space(s).
- In some cases, constituents could theoretically be a noun or another part of speech. For English, the dependency annotation and the part of speech of their first dictionary entry were consulted to decide on a part of speech.¹ For German, the paraphrase of the compound was considered to make a decision.²

¹Consider e.g. *human knee joint*, in which the first element could theoretically be an adjective or a noun. During parsing, it was annotated as an adjective. In addition, it is primarily listed as an adjective in the dictionary, which means that it was not considered.

²In the case of e.g. *Laufschuhschrank* ('running shoe cabinet'), the first element *Lauf* could be a verb stem or a noun. However, since the paraphrase of *Laufschuh*

- In order to ensure relatively clear and consistent annotations of semantic relationships, the compound needed to be transparent and have a reasonable paraphrase in modern-day language. Opaque compounds (e.g. *hamstring lesion*, with *hamstring* being the opaque compound part) were therefore not considered. Borderline cases were disregarded as well.

Four annotators with high language proficiency and a background in linguistics implemented the annotations: two annotators for English (both native speakers) and two for German (one native speaker, one C1-level speaker). The annotators extracted and annotated the compounds in our custom annotation tool (see Section 3.3). Annotations were performed in a sentence context to ensure a contextual basis for disambiguation decisions. Compounds which were already spelled as separate words were immediately annotated with syntactic role and semantic relationship information. Compounds which were originally spelled as one word with hyphens were first split (which included an update of the token IDs of the compound constituents and the following sentence context) and then annotated. Annotations were implemented as extensions of the UD format: For the syntactic role annotation, constituents acting as modifiers were annotated with the dependency relation *compound:nmod*. In addition, the ID of their head was updated. Consider this example: In a left-branching compound such as *notch width index* with the token IDs 1, 2 and 3, both *notch* and *width* were assigned the dependency relation *compound:nmod*. However, *notch* was assigned the head ID 2, while *width* was assigned the head ID 3. For the annotation of the semantic relationships, the *misc* column of the .conllu format was used. We used the classification proposed by [Benjamin and Schmidtke \(2023\)](#), which builds on earlier classification schemes (refer to Table 4 in the appendix for a complete overview of the semantic relationship labels). Each modifier was annotated with the most appropriate class describing its relationship to the head (e.g. *semRel:HforM*, i.e. 'head for modifier'). Consider the semantic relationships for a left-branching compound such as *addiction research unit*:

- *addiction*: *semRel:HaboutM* 'research about addiction'
- *research*: *semRel:HforM* 'unit for research about addiction'

Each annotator was assigned a set of files. After the annotation was completed, a second annotator

would be *Schuh zum Laufen* ('shoe for running') and not *Schuh für einen Lauf* ('shoe for a run'), it was considered a verb stem and included in the annotation.

reviewed the annotation and corrected it if necessary. This procedure has been shown to maximize the efficiency of the annotation process while maintaining output quality (Schacht et al., 2025). It was therefore chosen over a procedure based on inter-annotator agreement. In a few selected, highly ambiguous cases, a third person with a linguistic background was consulted.

3.3. Annotation Tool: CoBra Annotator

The annotation of the branching structure of multi-constituent compounds and their semantic relation is not trivial, especially due to the varied spelling possibilities. This causes the need for a certain degree of flexibility in a potential annotation tool. Tools like Arborator-Grew (Guibon et al., 2020) or INCEPTION (Klie et al., 2018) are capable of character-based annotations, which would be especially relevant for single-word variants of compounds, and additionally offer online cooperation of multiple annotators on the same data. However, the creation and integration of a new multi-constituent compound span into an existing .conllu structure is not readily available. As a result, these tools do not accommodate the flexible annotation of varied compound data.

To account for these issues, we developed CoBra Annotator, which allows for a custom annotation of multi-constituent compounds in a standardized data format (.conllu). The tool provides a lightweight, Python-based GUI for annotating and editing multi-constituent compounds in CoNLL-U format on a local machine.

The tool is implemented using the base Python (Van Rossum and Drake, 2009) library *tkinter* (Lundh, 1999), both licensed under the Python Software Foundation License Version 2. It comes with a GUI allowing for flexible and accurate annotations of compound structure. The tool offers various annotation settings depending on the type of compound that has to be annotated: You can choose to create a new span or select the existing span mode (should the compound already be tokenized). In the case of a hyphenated compound with all constituents and hyphens split into individual tokens, a dedicated hyphenated setting can be selected.

4. Evaluation and Analysis

We present an overview of our resource and a brief pilot study on the correlation between semantic transparency and branching direction.

4.1. Overview of the Resource

We annotated 295 compounds in total, 53 for Scientific English, 85 for General English, 47 for General German and 110 for Scientific German. The

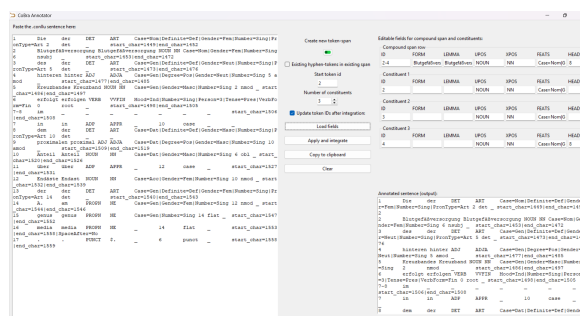


Figure 1: Exemplary annotation mask for the creation of a new compound span for the compound *Blutgefäßversorgung* (Engl. ‘blood vessel supply’).

most common semantic relationship for English was HforM, which was observed 84 times. For German, the most common relationship was MhasH, with 75 observations. Table 1 contains the compound counts for each branching direction per language and register as released in the extracted file. A complete overview of compounds per register and semantic relationship is displayed in Table 2 for English and Table 3 for German according to the automatic extraction of semantic relation frequency.

	En-Sci	En-Gen	Ger-Sci	Ger-Gen
right	2	35	4	11
left	51	50	106	36
Total	53	85	110	47

Table 1: Overview of branching direction per language and register.

4.2. Pilot Study: Predicting the Branching Structure of German Compounds

As a first application of our new resource, we conduct a small pilot study in which we investigate the predictive strength of semantic transparency on a compound’s branching structure. We conduct this pilot study with the German three-constituent nominal compounds from our dataset.

4.2.1. Semantic Transparency and Branching

While previous research has largely focused on binary compound structures, the increased structural complexity of three-constituent compounds offers a promising case for examining how semantic structure interacts with hierarchical organization. We orient our choice of semantic transparency measures loosely on Günther et al. (2020), who investigate transparency in two-constituent nominal

Relation	Scientific	General
HaboutM	5	19
HbyM	7	7
HcausedbyM	1	4
HderivedfromM	19	17
HduringM	3	5
HforM	29	55
HhasM	2	3
HisM	6	6
HlocationisM	14	16
HmadebyM	-	2
HmadeofM	-	3
HmakesM	2	2
HusedbyM	-	2
HusesM	4	1
MforH	-	2
MhasH	13	26
MlocationisH	1	-

Table 2: Observations of semantic relation types by register in English Compounds. Note that for each compound, two semantic relations are counted.

compounds. In their framework, semantic transparency is quantified by means of vector-based similarity measures between the compound and its constituents.

We extend this assumption to the internal nodes of multi-constituent compounds. Additionally, we focus on the predictiveness of transparency features regarding the branching structure of the compounds. More specifically, we test whether the internal node corresponding to the correct branching structure is semantically closer to the whole-word vector than the internal node implied by the alternative branching. This assumption is guided by [Levi \(1978\)](#), who argues that semantic coherence is a precondition for structural plausibility, and by [Selkirk \(1982\)](#), who suggests that compounds show hierarchical internal structures whose constituents must correspond to semantic constituents. We hypothesize that the preferred branching structure is the one whose internal node is semantically more coherent with the compound meaning.

4.2.2. Vector Representation of the Compounds

For the semantic representations, we use *fastText* embeddings, as they are especially suitable in cases of out-of-vocabulary (OOV) words, which we have to assume given our dataset. Since *fastText* operates on the level of character n-grams, many three-part compounds will have usable vector representations even if they were not observed as whole forms in the training data. This works especially

Relation	Scientific	General
HaboutM	11	12
HbyM	12	1
HcausedbyM	1	-
HcausesM	2	-
HderivedfromM	28	1
HduringM	1	-
HforM	6	29
HfromM	1	-
HhasM	7	7
HisM	5	1
HlocationisM	38	2
HmadeofM	1	-
HmakesM	10	4
HusesM	8	9
MaboutH	2	13
MderivedfromH	5	-
MfromH	1	-
MhasH	63	12
MisH	15	-
MlocationisH	2	3
MusesH	1	-

Table 3: Observations of semantic relation types by register in German Compounds. Note that for each compound, two semantic relations are counted.

well for German morphology, where compounding frequently leads to low-frequency or novel forms.

Queried vs. Composed Representations In a preliminary analysis, we compare different representations of the internal nodes. There are two ways of creating a compound vector: Querying the embedded compound on its own, resulting in a vector such as *ab* vs. representing it as a composed vector, derived additively from two individually queried constituent vectors (i.e. $a + b$). The latter follows the additive compositional approach motivated by [Reddy et al. \(2011\)](#).

We evaluated whether the composed or the queried representation provides a stronger approximation of the internal nodes with respect to the whole-word meaning. To this end, we compared the cosine similarity of each internal node (AB and BC) to the whole compound across both representation-forms (composed or queried representation) using paired t-tests. This was done regardless of gold-branching, as we first only tested validity of the form of representation. Our results (details are included in Section A of the Appendix) indicated that the queried representation yielded a substantially higher mean similarity to the whole word. We therefore we conducted all subsequent analyses using the queried representation.

Finally, we base our metrics on the relatedness measures proposed by Günther et al. (2020): They argue that these measures reflect access to the lexicalized ‘whole-word meaning’ in tasks such as semantic rating experiments.

4.2.3. Operationalization of the Metrics

We define three metrics that quantify different aspects of semantic transparency and its relation to branching structure. All measures are calculated for the competing parses of a given three-constituent compound ABC, comparing the left-branching structure [AB]C and the right-branching structure A[BC]. Similarity is always calculated as the cosine distance between two vector representations.

Semantic Coherence Our first metric captures the semantic coherence between the embedded compound (defined by a given branching structure) and the whole-word meaning. Specifically, we compare $\text{sim}(\text{AB}, \text{ABC})$, i.e. the cosine similarity of AB and ABC, to $\text{sim}(\text{BC}, \text{ABC})$, the cosine similarity of BC and ABC. This reflects the contribution of the respective embedded compound to the whole-word meaning. The underlying idea is that the preferred branching structure should be the one whose embedded constituent is more distributionally similar to the compound as a whole. While the head is typically the rightmost constituent in both English and German and therefore determines the overall compound category, the sub-structure containing the head (i.e. BC) is not necessarily most similar to the compound as a whole: The compound *trade union leader*, for instance, describes a type of leader. Due to its left-branching structure, however, we would expect a greater distributional similarity between *trade union* and *trade union leader* than between *union leader* and *trade union leader*. To determine the preference, we compute a delta coherence measure (distance of the two competing structures), defined as:

$$\Delta_{coh} = \text{sim}(\text{AB}, \text{ABC}) - \text{sim}(\text{BC}, \text{ABC}) \quad (1)$$

A positive value indicates an advantage for the left-branching structure, whereas a negative value indicates an advantage for the right-branching structure.

Similarity between Modifier and Head The second metric focuses on the relation between competing modifier structures and their respective heads. Here, we compare $\text{sim}(\text{AB}, \text{C})$ and $\text{sim}(\text{A}, \text{BC})$. This reflects the degree to which the embedded modifier structure aligns semantically with the head in the respective branching variant. As above, we

compute a delta value to calculate a signed relative competition score between the different variants:

$$\Delta_{head} = \text{sim}(\text{AB}, \text{C}) - \text{sim}(\text{A}, \text{BC}) \quad (2)$$

This score indicates which modifier–head structure exhibits greater semantic compatibility and thus greater structural plausibility.

Transparency Asymmetry The third metric addresses the question of whether the embedded compound (e.g. AB) contributes more to the whole-word meaning than the sum of its individual parts. We define transparency asymmetry (TA) as:

$$TA = \frac{\text{sim}(\text{AB}, \text{ABC})}{(\text{sim}(\text{A}, \text{ABC}) + \text{sim}(\text{B}, \text{ABC}))} \quad (3)$$

$$TA = \frac{\text{sim}(\text{BC}, \text{ABC})}{(\text{sim}(\text{B}, \text{ABC}) + \text{sim}(\text{C}, \text{ABC}))} \quad (4)$$

This measure extends the coherence metric by asking whether the embedded compound functions merely as a transparent combination of its constituents. Alternatively, it can be semantically emergent, carrying more meaning for the whole word than its individual constituents considered separately. This would be the case if the similarity was greater in composed substructures. Again, we compute a delta value to calculate a signed relative competition score between the two branching options. This allows us to compare the relative degree of semantic asymmetry across competing internal structures.

To evaluate the predictive contribution of these metrics, we compare the quality of several regression models based on the Aikake Information Criterion (AIC; Akaike, 1998). In addition, we perform likelihood ratio tests to determine whether the inclusion of a given semantic transparency metric significantly improves model fit. This approach allows us to quantify the extent to which our distributional transparency metrics contribute to explaining our annotated gold branching structures from the CoBra resource.

4.2.4. Results

Semantic Coherence We examined whether the gold internal structure is generally more similar to the whole-word representation than the competing internal structure. A paired t-test revealed a significant difference ($t = -4.388$, $p < 0.001$), indicating that the gold internal structure exhibits significantly higher similarity to the compound as a whole than the alternative structure. This directly supports the semantic coherence hypothesis: The

branching structure corresponds more closely to the compound's overall semantic representation than the competing structure.

Similarity between Modifier and Head To assess head-alignment coherence, we compared similarity values of alignment between the compound's first-level head and its modifier of the gold-branching and the competitor. Results were highly significant ($t = 7.408$, $p < 0.001$). This indicates that the gold branching structure displays stronger head-alignment coherence than the competing structure. It suggests that the branching-structure organization of German three-constituent nominal compounds is not only represented in overall semantic similarity to the whole word, but also in internal head-modifier alignment patterns.

Transparency Asymmetry Regarding the metric of transparency asymmetry, we first tested the reliability of the metric: We examine whether the embedded structure of the gold variant reliably surpasses the contribution of its summed constituents (the additive baseline). This would be indicated by a TA mean greater than zero. A one-sample t -test on the transparency asymmetry values for the gold structure revealed a very large positive mean (mean = 1.213), which was highly significant ($t = 54.267$, $p < 0.001$). This indicates that the gold internal structures are consistently more than the sum of their parts. Additionally, this result underlines the previous choice of the queried representation for the embedded compound structure.

Following the analysis of TA-reliability, we tested whether TA is stronger for the gold branching than for the competing branching. The mean TA value for the gold structure (mean = 1.213) exceeded that of the competing structure (mean = 1.018), with a mean difference of 0.195. This difference was statistically significant ($t = 6.79$, $p < 0.001$). Thus, TA not only exists within the gold structure but is significantly stronger than in the competing variant, indicating that it is structurally selective and supports the gold branching.

Predictiveness of the Metrics As a final analysis, we evaluated how well the different metrics predict the gold branching using delta accuracy. This measures whether the directional competition score correctly identifies the gold structure. General coherence achieved a delta accuracy of 0.33, indicating limited predictive power. In contrast, transparency asymmetry achieved a noticeably higher delta accuracy of 0.717 and head-alignment coherence achieved a comparable accuracy of 0.711. While general coherence between the branching constituent and the whole-word provides evidence

for structural coherence at the group level (significance tests), transparency asymmetry and head-alignment coherence suggest stronger predictive power for identification of the gold branching.

Model Comparison To investigate the contribution of the different transparency metrics to the prediction of gold branching among each other, we conducted various nested-model comparisons using logistic regression models with a binomial link. All metric delta values were z -standardized. Gold branching was transformed as a binary ($AB = 1$) and model fit was evaluated using likelihood-ratio tests and AIC comparisons.

We first compared a baseline intercept-only model to a model including general semantic coherence). The likelihood-ratio test did not turn out significant ($LR = 2.216$, $df = 1$, $p = 0.137$), indicating that semantic coherence alone does not significantly improve prediction of gold branching over the baseline model. We then added TA to the semantic coherence model, which improved model fit ($LR = 7.800$, $df = 1$, $p = 0.005$). TA thus seems to provide significant additional explanatory power. Subsequently, we added head-alignment coherence to the second model. This extension significantly improved model fit ($LR = 5.206$, $df = 1$, $p = 0.023$), showing that head-alignment coherence contributes noticeably. Finally, we tested a full interaction model including all two- and three-way interactions among the metrics. The likelihood-ratio comparison between the additive three-predictor model and the interaction model was not significant ($LR = 5.355$, $df = 4$, $p = 0.253$). These results suggest that adding an interaction does not significantly improve model fit, indicating that the metrics operate largely additively. Thus, semantic coherence alone does not seem to be sufficient to predict branching, but TA and head-alignment coherence contribute significantly.

Model comparison based on AIC values supports the likelihood-ratio results: The largest reduction of AIC was achieved for the model with semantic coherence, TA and head-alignment coherence.

We can therefore conclude that general semantic coherence alone is not necessarily sufficient to predict structure. TA, on the other hand, forms the strongest predictor, with head-alignment coherence providing additional explanatory power. Since the best model was the additive model including all three metrics without interactions, we can assume that these semantic pressures operate mostly independently in shaping transparency of the German three-constituent nominal compounds.

5. Discussion

Our resource contributes to the study of compounds, more precisely of triconstituent nominal compounds. Covering two distinct registers, it proposes a way of annotating syntactic and semantic structure within morphologically complex words. However, it also highlights some open questions regarding compounds. One of these is which structures we actually consider as compounds and which as syntactic phrases. We have presented a working definition, which includes all nouns modified by other nouns and therefore makes no distinction between the two. But can this approach be applied to other compound types, such as adjective-noun compounds? In German, lexicalized adjective-noun compounds are easily identifiable due to their spelling as one word and their lack of (adjectival) inflection (consider e.g. *Blaumeise*, 'blue tit'). In English, orthography could theoretically also be used to make a distinction between lexicalized adjective-noun compounds and syntactic phrases in which an adjective modifies a noun (as in *blackbird* vs. *blue tit*). Still, it is unclear if orthographic conventions always accurately reflect the degree of lexicalization (since e.g. *blue tit* is an established term). One approach might be the one taken in the UniDive shared task on morpho-syntactic parsing (Goldman et al., 2025), in which morphology and syntax are not separated but considered jointly. This would also abolish the need for a distinction between compounds (as structures derived from morphological processes) and syntactic phrases.

Another question concerns opaque vs. transparent compounds: Should they be treated in the same way or differently? Since we were interested in semantic relationships, we chose to consider only transparent ones. This is in line with research indicating that lexicalized compounds, particularly short ones, are actually processed differently than longer and non-lexicalized compounds (Hyönä et al., 2020). An issue here is that semantic transparency or compositionality of compounds is typically regarded not as a binary feature, but as a continuum (see e.g. Reddy et al., 2011).

Finally, our choice to represent closed compounds as multi-word tokens, an approach also suggested by Savary et al. (2023), naturally has implications for tokenization standards. In order to systematically annotate these compounds as multi-word expressions, they would have to be split during preprocessing.

Our annotation efforts relied heavily on manual work by human annotators. As of now, there is unfortunately no reliable way to automate the annotation of semantic roles. Still, research on compounds, both in psycholinguistics and NLP, relies

on high-quality datasets.

The results of our pilot study yielded interesting results: In German compounds, TA was the strongest predictor of branching structure. Adding semantic coherence and head-alignment coherence yielded the best-fitting model. These results indicate that there are transparency differences within compounds and that these are predictive of branching structure, even if only compounds with an overall high degree of semantic transparency are considered. Future research will have to test how this interacts with other variables, such as internal semantic relationships, and if these results also hold for English.

Overall, many avenues for future research remain in order to better understand the processing of triconstituent compounds and to appropriately model their semantics.

6. Conclusion

Our resource addresses a gap in research on compound processing, which has been traditionally focused on two-constituent compounds. By creating CoBra, a dataset of triconstituent nominal compounds, we take a step into the deeper study of other compound constructions. We have focused on semantic relations between constituents and compound branching structure as areas which are particularly interesting for triconstituent compound processing. In addition, we contribute to current efforts and discussions on the integration of morphological and syntactic information within the UD framework. In particular, our choice of treating closed compounds as multi-word tokens is in line with other work proposing alternatives to a strict reliance on token-based representations in UD (Savary et al., 2023; Goldman et al., 2025).

7. Ethics Statement

This paper does not contain any studies with human participants performed by any of the authors.

8. Limitations

A significant limitation of our dataset is its size: The annotation process, particularly for semantic relations, is very time-consuming and relies heavily on manual work by human annotators. Due to resource constraints, we were therefore only able to annotate a relatively small number of compounds. The dataset could be extended in the future.

Furthermore, our choice of splitting closed compounds into separate tokens works well for English and German, and most likely other Germanic languages. In future research, its application for other

language families with different compounding patterns would need to be investigated as well.

Regarding our annotation of semantic relations, it is debatable whether we have selected the ideal classification scheme. The classification used for CoBra was chosen because it is relatively new and builds on previous classification schemes. However, there is currently no consensus on the best practice for annotating semantic relationships within compounds.

Regarding our pilot study and the predictiveness of semantic transparency measures, our results are limited to compounds with a relatively high degree of semantic transparency. For opaque compounds, different measures would need to be tested (e.g. based on co-occurrence frequency, mutual information or syntactic features).

We used out-of-the-box models for our vector representations of the compounds. These might not have captured all nuances in the distribution of whole compounds or their constituent elements. Additionally, the corpora used for compound extraction might, due to their age, have been used in the training of the vector representations. Applying our method to newer data or novel compound creations would yield more insights into the generalizability of our results. In this case, the training of custom models might be necessary. However, a critical precondition would be the availability of high-quality training data.

Finally, we did not directly test the reliability of our gold annotation or the validity of our semantic transparency measures. As a future project, we therefore plan to validate some of our results in a behavioral rating study.

9. Acknowledgements

The authors thank Stefania Degaetano-Ortlieb and three anonymous reviewers for their constructive feedback on a previous version of this paper. This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) –Project-ID 232722074 – SFB 1102 Information Density and Linguistic Encoding.

10. Bibliographical References

- Multi-Word Token (MWT) Expansion — stanfordnlp.github.io. <https://stanfordnlp.github.io/stanza/mwt.html>. [Accessed 16-02-2026].
- Open Science Framework (OSF).
- Hirotougu Akaike. 1998. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotougu Akaike*, pages 199–213. Springer.
- Sally Andrews, Brett Miller, and Keith Rayner. 2004. [Eye movements and morphological segmentation of compound words: There is a mouse in mousetrap](#). *European Journal of Cognitive Psychology*, 16(1-2):285–311.
- Harald Baayen, Victor Kuperman, and Raymond Bertram. 2010. [Frequency effects in compound processing](#). In *Cross-disciplinary Issues in Compounding*, pages 257–270. John Benjamins Publishing Company.
- Laurie Bauer. 1998. [When is a sequence of nouns a compound in English?](#) *English Language and Linguistics*, 2(1):65–86.
- Laurie Bauer. 2019. [Compounds and multi-word expressions in English](#). In Barbara Schlücker, editor, *Complex Lexical Units: Compounds and Multi-Word Expressions*, pages 45–68. De Gruyter.
- Shaina Benjamin and Daniel Schmidtke. 2023. [Conceptual combination during novel and existing compound word reading in context: A self-paced reading study](#). *Memory and Cognition*, 51(5):1170–1197.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. [Universal Dependencies](#). *Computational Linguistics*, 47(2):255–308.
- Stefania Degaetano-Ortlieb. 2021. Measuring informativity: The rise of compounds as informationally dense structures in 20th-century scientific English. In *Corpus-based Approaches to Register Variation*, pages 291–312. John Benjamins Publishing Company.
- Fabienne Fritzinger and Alexander Fraser. 2010. How to avoid burning ducks: Combining linguistic analysis and corpus statistics for German compound processing. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics (MATR)*, pages 224–234.
- Sandra E. Geer, Henry Gleitman, and Lila Gleitman. 1972. Paraphrasing and remembering compound words. *Journal of Verbal Learning and Verbal Behavior*, 11:348–355.
- Kim Gerdes and Sylvain Kahane. 2016. [Dependency annotation choices: Assessing theoretical and practical issues of Universal Dependencies](#). In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 131–140, Berlin, Germany. Association for Computational Linguistics.

- Omer Goldman, Leonie Weissweiler, Kutay Acar, Diego Alves, Anna Baczkowska, Gulsen Eryigit, Lenka Krippnerová, Adriana Pagano, Tanja Samardžić, Luigi Talamo, Alina Wróblewska, Daniel Zeman, Joakim Nivre, and Reut Tsarfay. 2025. [Findings of the UniDive 2025 shared task on multilingual morpho-syntactic parsing](#). In *Proceedings of The UniDive 2025 Shared Task on Multilingual Morpho-Syntactic Parsing*, pages 1–18, Ljubljana, Slovenia. Association for Computational Linguistics.
- Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. [When collaborative tree-bank curation meets graph grammars](#). In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5293–5302, Marseille, France. European Language Resources Association.
- Fritz Günther, Marco Marelli, and Jens Bölte. 2020. Semantic transparency effects in german compounds: A large dataset and multiple-task investigation. *Behavior Research Methods*, 52(3):1208–1224.
- Verena Henrich and Erhard Hinrichs. 2011. Determining immediate constituents of compounds in GermaNet. In *Proceedings of Recent Advances in Natural Language Processing*, pages 420–426.
- Jukka Hyönä, Alexander Pollatsek, Minna Koski, and Henri Olkonien. 2020. [An eye-tracking study of reading long and short novel and lexicalized compound words](#). *Journal of Eye Movement Research*, 13(4):1–18.
- Barbara J. Juhasz, Matthew S. Starr, Albrecht W. Inhoff, and Lars Placke. 2003. The effects of morphology on the processing of compound words: Evidence from naming, lexical decisions and eye fixations. *British Journal of Psychology*, 94:223–244.
- Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. [The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation](#). In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.
- Katharina Korecky-Kröll, Sabine Sommer-Lolei, and Wolfgang U. Dressler. 2017. Emergence and early development of German compounds. In *Nominal Compound Acquisition*, pages 19–37. John Benjamins Publishing Company.
- Kristina Kösling and Ingo Plag. 2009. [Does branching direction determine prominence assignment? an empirical investigation of triconstituent compounds in English](#). *Corpus Linguistics and Linguistic Theory*, 5(2):201–239.
- Andrea Krott, Gary Libben, Gonia Jarema, Wolfgang Dressler, Robert Schreuder, and Harald Baayen. 2004. Probability in the grammar of German and Dutch: Interfixation in triconstituent compounds. *Language and Speech*, 47(1):83–106.
- Victor Kuperman, Robert Schreuder, Raymond Bertram, and Harald Baayen. 2009. [Reading polymorphemic Dutch compounds: toward a multiple route model of lexical processing](#). *Journal of Experimental Psychology: Human Perception and Performance*, 35(3):876–895.
- Judith N. Levi. 1978. *The Syntax and Semantics of Complex Nominals*. Academic Press.
- Gary Libben, Christina L. Gagné, and Wolfgang U. Dressler. 2020. [The representation and processing of compound words](#). In Vito Pirrelli, Ingo Plag, and Wolfgang U. Dressler, editors, *Word Knowledge and Word Usage: A Cross-Disciplinary Guide to the Mental Lexicon*, pages 336–352. De Gruyter Mouton, Berlin, Boston.
- Fredrik Lundh. 1999. An introduction to tkinter. www.pythonware.com/library/tkinter/introduction/index.htm.
- Marco Marelli, Christina L. Gagné, and Thomas L. Spalding. 2017. Compounding as abstract operation in semantic space: Investigating relational effects through a large-scale, data-driven computational model. *Cognition*, 166:207–224.
- Julia A. Meßmer, Regine Bader, and Axel Mecklinger. 2021. The more you know: Schema-congruency supports associative encoding of novel compound words. Evidence from event-related potentials. *Brain and Cognition*, 155:105813.
- Filip Miletić and Sabine Schulte im Walde. 2023. A systematic search for compound semantics in pretrained BERT architectures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1499–1512.
- Mark Ormerod, Jesús Martínez del Rincón, and Barry Devereux. 2024. [How is a “kitchen chair” like a “farm horse”? Exploring the representation of noun-noun compound semantics in Transformer-based language models](#). *Computational Linguistics*, 50(1):49–81.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza:

A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. [An empirical study on compositionality in compound nouns](#). In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

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).

Carmen Schacht, Tobias Nischk, Oleksandra Yazdanfar, and Stefanie Dipper. 2025. [Cheap annotation of complex information: A study on the annotation of information status in German TEDx talks](#). In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 297–307, Vienna, Austria. Association for Computational Linguistics.

Annika Schebesta and Gero Kunter. 2022. Constituent durations in English NNN compounds: A case of strategic speaker behavior? *Journal of Phonetics*, 94:101164.

Sabine Schulte im Walde, Stefan Müller, and Stefan Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 255–265.

Elisabeth O. Selkirk. 1982. *The Syntax of Words: Volume 7*. Linguistic Inquiry Monographs. MIT Press, London, England.

Emil Svoboda and Magda Ševčíková. 2024. [Compounds in Universal Dependencies: A survey in five European languages](#). In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 88–99, St. Julian's, Malta. Association for Computational Linguistics.

Don Tuggener. 2016. *Incremental Coreference Resolution for German*. Ph.D. thesis, University of Zurich.

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

11. Language Resource References

[Digitales Wörterbuch der deutschen Sprache \(DWDS, online\)](#).

[DocCheck Flexikon \(online\)](#).

[Duden \(online\)](#).

[Merriam-Webster Medical Dictionary \(online\)](#).

MuchMore Springer Bilingual Corpus. <https://muchmore.dfki.de/resources1.htm>. [Accessed 02-06-2025].

[Oxford English Dictionary \(online\)](#).

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Phoebe Chen, David Poeppel, and Arianna Zuanazzi. 2023. A dataset of 108 novel noun-noun compound words with active and passive interpretation. *Journal of Open Psychology Data*, 11(1).

BNC Consortium. 2007. [The British National Corpus, XML Edition](#).

Christina L. Gagné, Thomas L. Spalding, and Daniel Schmidtke. 2019. LADEC: The large database of English compounds. *Behavioral Research Methods*, 51(5):2152–2179.

Barbara J. Juhasz, Yun-Hsuan Lai, and Michelle L. Woodcock. 2015. A database of 629 English compound words: ratings of familiarity, lexeme meaning dominance, semantic transparency, age of acquisition, imageability, and sensory experience. *Behavior Research Methods*, 47(4):1004–1019.

Daniel Schmidtke, Julie A Van Dyke, and Victor Kuperman. 2021. [CompLex: An eye-movement database of compound word reading in English](#). *Behavioral Research Methods*, 53(1):59–77.

Sabine Schulte im Walde, Anna Hättty, Stefan Bott, and Nana Khvtisavishvili. 2016. GhoSt-NN: A representative gold standard of German noun-noun compounds. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2285–2292.

A. Appendix

Sem. Rel. Label	Meaning
HaboutM	head about modifier
HbyM	head by modifier
HcausedbyM	head caused by modifier
HderivedfromM	head derived from modifier
HduringM	head during modifier
HforM	head for modifier
HhasM	head has modifier
HisM	head is modifier
HlocationisM	head location is modifier
HmadebyM	head made by modifier
HmadeofM	head made of modifier
HmakesM	head makes modifier
HusedbyM	head used by modifier
HusesM	head uses modifier
MforH	modifier for head
MfromH	modifier from head
MhasH	modifier has head
MisH	modifier is head
MlocationisH	modifier location is head
MusesH	modifier uses head

Table 4: Overview of semantic relationship labels used for annotation (based on [Benjamin and Schmidtke, 2023](#)).

Variant	En-Sci	En-Gen	Ger-Sci	Ger-Gen
Open	41	77	-	-
Closed	-	-	110	42
Hyphenated	-	-	-	5
Mixed	12	8	-	-

Table 5: Overview of spelling variants annotated in CoBra.

	AB Node	BC Node	Comparison
Mean Comp.	0.3705	0.4165	0.3935
Mean Queried	0.6372	0.7020	0.6695
Mean Diff. (Comp. - Queried)	- 0.2667	-0.2854	-0.2761
Cohen's d	-2.136	-1.939	-2.021
<i>t</i> -value	-27.526	-24.985	-36.829
<i>p</i> -value	<0.001	<0.001	<0.001

Table 6: T-test comparing cosine similarity of internal node and whole compound for both representation types (queried vs. composed).

	Est.	Std. Err.	<i>z</i>	<i>p</i>
Intercept	0.903	0.025	35.967	<0.001
SC	-0.381	0.144	-2.646	<0.01
TA	0.364	0.143	2.539	<0.05
MHS	0.049	0.027	1.819	0.069

Table 7: Regression model summary for the maximal model, including the variables semantic coherence (SC), transparency asymmetry (TA) and modifier-head similarity (MHS).

Romance Reflexive Constructions Revisited

Verginica Barbu Mititelu¹, Elena Irimia¹, Adriana Pagano²,
Ioana Buhnila³, Roxana Ciolăneanu^{4,5}

¹Romanian Academy Research Institute for Artificial Intelligence, Bucharest, Romania;

²Federal University of Minas Gerais, Brazil;

³Centre for Data Science in Humanities, Chosun University, South Korea

⁴Centro de Linguística da Universidade de Lisboa, Portugal;

⁵Institute of the Romanian Language, Bucharest, Romania

Abstract

We analyze the current annotation of reflexive constructions, i.e. verbal constructions marked by a reflexive clitic, across five Romance languages (French, Italian, Portuguese, Romanian and Spanish) in several Universal Dependencies treebanks (version 2.17). We discuss the morphologic, syntactic and semantic characteristics of such constructions in each of the languages considered, both from a theoretical perspective and from that of existing annotation. To address inconsistencies in the data and strengthen Universal Dependencies as a scaffold for the automatic conversion of morphosyntactic annotation into semantic representations (Uniform Meaning Representation), we propose a clear distinction between argumental and non-argumental uses of the reflexive clitic, and outline systematic ways to implement this distinction in the annotation guidelines. We also examine how some of the reported inconsistencies can be handled in the treebanks under study and discuss the extent to which these practices can be extended to other treebanks, within the same or across different languages.

Keywords: reflexive constructions, Romance languages, uniform meaning representation, syntactic and semantic arguments

1. Introduction

In this paper we focus on reflexive constructions, by which we refer to a family of verbal constructions marked by reflexive clitics, with forms varying across Romance languages, and with meaning encompassing reflexive, passive, impersonal, middle, and lexically specified uses.

The contributions of this paper are the following: (i) at a theoretical level, we document the types of reflexive constructions that need to be distinguished for annotation in Universal Dependencies (UD) (de Marneffe et al., 2021) treebanks; (ii) we report on the current state of the annotation of these reflexive constructions in UD version 2.17 treebanks for the languages under consideration (see below); (iii) we propose annotation guidelines for such constructions against the presented theoretical aspects; (iv) we consider possible ways of improving the uniformity and consistency of the annotation of such constructions in UD treebanks.

The perspective on reflexive constructions we adopt here is not only morpho-syntactic, but also semantic, in that it envisages automatic conversion of UD trees into their Uniform Meaning Representation (UMR) (Bonn et al., 2023).

UD annotation of reflexive clitics is highly relevant to UMR conversion because UMR relies on syntactic structure to determine semantic roles and argument identity. In UD, reflexive clitics may be annotated either as core arguments (obj, iobj, obl:arg) or as non-argumental markers (expl, expl:pass, expl:impers, expl:pv) (see Section 5). This distinc-

tion directly affects whether a reflexive construction is interpreted in UMR as involving role coreference (e.g., the same participant filling two semantic roles) or as a valency-changing or voice-related operation (e.g., passive, impersonal, anticausative). If UD annotation conflates these uses or applies them inconsistently across treebanks, automatic UMR conversion may incorrectly introduce or omit semantic participants. Therefore, consistent and linguistically grounded UD treatment of reflexive clitics is crucial for ensuring accurate semantic representation, especially in multilingual UMR pipelines.

The languages under study here are French, Italian, Portuguese (mainly the Brazilian variant, with occasional references to European¹ one as well), Romanian, and Spanish. They are all Romance

¹For European Portuguese, CINTIL (Branco et al., 2011) (Mariana Avelãs and others, 2022) is the only available treebank. With approximately 18,000 sentences (around 450,000 tokens), it is an important resource and has been widely used in syntactic and morphosyntactic studies of European Portuguese. It was therefore relevant to assess whether it could be incorporated into our cross-treebank comparison. However, CINTIL could not be productively queried for the purposes of the present study. Although it tokenizes pronominal forms, its annotation scheme does not distinguish clitic from non-clitic pronouns, nor does it encode reflexivity through a dedicated morphological feature such as *Reflex*=Yes. In addition, it does not employ the UD relation *expl* or its subtypes (e.g. *expl:impers*, *expl:pass*, *expl:pv*). As a result, reflexive and non-reflexive uses of pronominal forms cannot be reliably separated, and neither reflexive-object queries nor dependency-based filtering of expletive con-

languages, each with its own evolution that has contributed to shaping its current characteristics. Reflexive constructions existed also in Latin, their common ancestor, and have been preserved in the contemporary languages. For each language, we selected one or more UD treebanks (see Section 3) to analyze the current annotation of their reflexive constructions.

In Section 2 we comment on limited body of related work on reflexive constructions in UD treebanks, in Section 3 we briefly present the treebanks we work with and our approach. The characteristics we consider relevant to the analysis of this phenomenon are compared across the five languages in Section 4, which examines the cross-linguistic morphological properties of reflexive clitics, and in Section 5, which addresses their annotation in treebanks at the syntactic and semantic levels. A discussion of the current annotation of reflexives in the UD treebanks considered here can be found in Section 6. We present some considerations relevant to a more consistent annotation of reflexives in Section 7, before concluding the paper.

2. Related Work

Despite the relevance of reflexive constructions for both syntactic analysis and downstream semantic representation, related work on their treatment in Universal Dependencies (UD) treebanks remains limited. Our paper comes as a next step after [Marković and Zeman \(2018\)](#), who showed the unsatisfactory annotation of reflexives in UD treebanks. They examine how reflexive markers are annotated in UD 2.2, with a particular focus on Slavic languages. The paper evaluates whether current UD guidelines are sufficiently clear and applied consistently across languages.

The authors show that reflexives have multiple functions – true reflexive, reciprocal, inherent/pronominal verb marker, passive marker, impersonal marker, and anticausative – and that their annotation in UD is highly inconsistent across treebanks. Problems include uneven use of the feature `Reflex=Yes`, inconsistent distinction between argumental uses (`obj`, `iobj`, `obl`) and non-argumental uses (`expl`, `expl:pass`, `expl:impers`, `expl:pv`), and divergent treatments of inherently reflexive verbs. Some treebanks overgeneralize a single label (e.g., treating all reflexives as `expl` or `compound`), while others fail to distinguish passive and impersonal constructions.

structions can be formulated in a way comparable to the other treebanks considered here. European Portuguese is therefore included only for contextual and comparative purposes, not as a fully queryable dataset in the empirical analysis.

The authors argue that UD should, as an essential first step, distinguish true reflexive arguments from non-argumental reflexive constructions, and they advocate for clearer guidelines, more consistent use of subtypes of `expl`, and harmonization across languages. They conclude that improving reflexive annotation is crucial for cross-linguistic research and for maintaining UD’s goal of typological consistency.

[Duran et al. \(2025\)](#) present a corpus-based study of the clitic *se* in Brazilian Portuguese based on 732 occurrences extracted from the UD-annotated Porttinar-base treebank. Their goal is to develop a more consistent and computationally useful annotation scheme for NLP. The analysis shows that traditional grammatical classifications (reflexive, reciprocal, middle, synthetic passive, impersonal, etc.) are difficult to apply consistently in contemporary Brazilian Portuguese, especially the distinction between synthetic passive and impersonal constructions.

To address this, the authors propose a streamlined annotation framework aligned with UD. They argue for simplifying the treatment of *se* in UD by: (i) clearly distinguishing argumental *se* (`obj/iobj`) from non-argumental uses; (ii) unifying impersonal and synthetic passive constructions under a consistent impersonalization analysis; and (iii) refining the annotation of pronominal and middle constructions. The proposal aims to increase annotation consistency and improve downstream NLP tasks such as information extraction and semantic role labeling.

Both [Marković and Zeman \(2018\)](#) and [Duran et al. \(2025\)](#) address the annotation of reflexive markers in UD, but from complementary perspectives. The former take a cross-linguistic and guideline-oriented approach, showing that reflexive annotation in UD is inconsistent across languages and advocating for clearer distinctions between argumental and non-argumental uses, as well as more systematic use of `expl` subtypes. In contrast, the latter focus specifically on Brazilian Portuguese, demonstrating empirically that traditional fine-grained grammatical classifications are difficult to apply consistently in corpus annotation and proposing a streamlined UD-oriented solution that prioritizes syntactic behavior and annotation reliability. Together, the two works highlight the tension between typological granularity and the need for practical consistency in UD reflexive annotation.

Issues in UD annotation, including those related to reflexive constructions, can be revisited with downstream applications in mind, given recent work highlighting the potential of UD as a scaffold for semantic annotation, particularly in multilingual settings. [Gamba et al. \(2025\)](#) propose a method for bootstrapping partial UMR graphs from UD parses, showing that UD’s syntactic structures can be sys-

tematically exploited to derive initial semantic representations that reduce manual annotation effort. Their approach leverages UD relations to construct partial UMR graphs that preserve event structure and core argument configurations, which can then be completed and refined by human annotators. This work demonstrates that, despite UD not being a semantic formalism, its cross-linguistic consistency and wide language coverage make it a practical starting point for scalable semantic annotation.

Building on [Gamba et al. \(2025\)](#), our study similarly exploits UD annotations to guide the identification of reflexive and expletive constructions across languages, while making explicit the limitations imposed by treebank-specific annotation choices. Reflexive constructions such as the one in Figure 1 illustrate how syntactic information encoded in UD can be exploited in the construction of semantic representations such as UMR. In the Portuguese sentence *A menina se olhou no espelho*. (“The girl looked at herself in the mirror.”), UD can encode reflexivity through annotating the clitic pronoun (*se*) as an argumental object (obj), with the feature *Reflex=Yes*, which will be interpreted as coreferential with the subject *A menina*. In UMR, this syntactic configuration naturally maps onto a representation in which a single participant is assigned both actor and undergoer roles in the event, with reflexivity expressed via role identity rather than by introducing an additional argument. This type of mapping exemplifies how UD structures can provide a principled basis for bootstrapping semantic representations, as proposed in recent UD-to-UMR conversion work ([Gamba et al., 2025](#)).

3. Methodology

Our methodology involved four main steps: the selection of the languages considered, the selection of the treebanks examined, a cross-linguistic analysis of the relevant properties of these languages, and a cross-treebank analysis of the annotation of reflexive clitics. Within this framework, the identification of reflexive and reciprocal objects was carried out in a way that accommodates cross-linguistic and cross-treebank variation in UD annotation practices.

The selection of the Romance languages considered in this study was guided by the availability of treebanks. For each language selected on this basis, we then identified the following treebanks for analysis:

- French – (i) GSD ([Guillaume et al., 2019](#)) ([Marie-Catherine de Marneffe and others, 2015](#)), which represents contemporary written French spanning multiple genres (blog, news, reviews, wiki), and contains 16,342 sentences (400,387 tokens); (ii) Sequoia ([Candito](#)

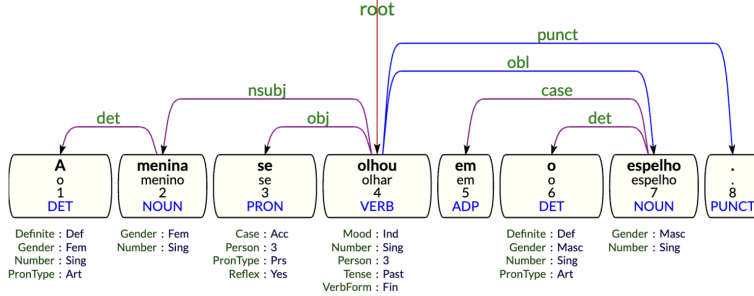
and [Seddah, 2012](#)) ([Marie Candito and others, 2017](#)), a treebank of written French composed of 3,099 sentences (70,567 tokens) from various sources: French Europarl, Wikipédia Fr, Newspaper *Est Républicain*, and European Medicines Agency. Both treebanks are regularly updated.

- Italian – (i) VIT (Venice Italian Treebank) ([Delmonte et al., 2007](#); [Alfieri and Tamburini, 2016](#)) ([Rodolfo Delmonte, 2007](#)), represents contemporary written Italian across multiple genres and contains approximately 10,000 sentences (about 250,000 tokens); and (ii) ISDT (Italian Stanford Dependency Treebank) ([Bosco et al., 2012](#); [Sanguinetti and Bosco, 2014](#)) ([Bosco et al., 2015](#)), composed primarily of newspaper text, includes approximately 13,000 sentences (about 300,000 tokens). Together, they provide a broad coverage of modern Italian usage.
- Brazilian Portuguese – (i) Porttinari ([Duran et al., 2023](#)) ([Magali Sanches Duran and others, 2017](#)) contains approximately 5,000 sentences (about 150,000 tokens); and (ii) Petrogold ([de Souza and Freitas, 2022](#)) ([de Souza and Freitas, 2022](#)) comprises approximately 3,000 sentences (about 80,000 tokens), providing complementary coverage of syntactic phenomena. Both represent contemporary Brazilian Portuguese and include newspaper and academic texts, respectively.
- Romanian – the Romanian Reference Treebank ([Barbu Mititelu, 2018](#)) ([Verginica Barbu Mititelu and others, 2015](#)) (RRT) was chosen because it represents contemporary written Romanian and includes texts from several genres (literature, medicine, law, news, science, academic writing, Wikipedia, etc.); the corpus contains 9,524 sentences (218,522 tokens).
- Spanish – the AnCora treebank ([Taulé et al., 2008](#)) ([Héctor Martínez Alonso and Daniel Zeman, 2016](#)), was used, a large manually annotated corpus of contemporary written Spanish covering multiple genres; it contains approximately 17,000 sentences (about 500,000 tokens) and is widely adopted in UD-based research.

The cross-linguistic analysis of the relevant properties of these languages consisted in comparing the main morphological properties of reflexive clitics across the languages considered, in order to establish the dimensions of variation relevant to the study. Syntactic and semantic aspects were then addressed separately through the cross-treebank analysis.

The cross-linguistic analysis of the relevant properties of these languages consisted in comparing

::id ex1
::snt A menina se olhou no espelho.
::gloss
A menina se olh-ou em o espelh-o
DEF.F girl.F REFL look-PST.3SG in DEF.M mirror-M
'The girl looked at herself in the mirror.'



(e1 / olhar
:actor (p / menina
:ref-person 3rd
:ref-number Singular
:ref-gender Feminine)
:undergoer p
:location (x2 / espelho))

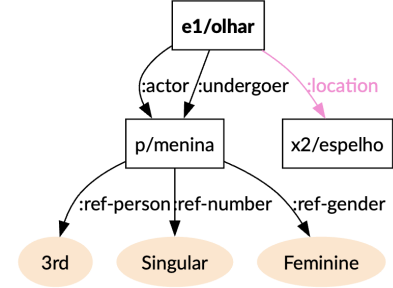


Figure 1: UD and UMR representations of sentence illustrating a reflexive construction with clitic *se* annotated as object in UD and mapped in UMR to a coreferential undergoer participant.

the main morphological properties of reflexive clitics across the languages considered, in order to establish the dimensions of variation relevant to the study. Syntactic and semantic aspects were then addressed separately through the cross-treebank analysis.

To identify reflexive and reciprocal objects, we adopted a strategy-based approach designed to accommodate cross-linguistic and cross-treebank variation in UD annotation practices. On the basis of this approach, we developed a set of language-specific query patterns, presented in Section 5. Direct and indirect objects were identified through agreement in person and number (for the first and second persons) or in number only (for the third person) with the head verb. Since first- and second-person reflexive objects are morphologically indistinguishable from non-reflexive object clitics in Romance languages, reflexivity cannot be identified solely on the basis of clitic form. Instead, it must be detected through form-based queries combined with a person and number agreement heuristic, operationalized by matching the person and number features of the object clitic with those encoded on the governing verb. This is illustrated by the Spanish contrast between *Me pregunté por qué pasó eso* ('1sg.obj ask-PST.1sg why happen-PST.3sg that', 'I asked myself why that happened'), which is retrievable as reflexive because the person and number features of the object clitic (*me*) match those of the governing verb (*pregunté*), and *Te pregunté por qué pasó eso* ('2sg.obj ask-PST.1sg why happen-PST.3sg that', 'I asked you why that happened'), which is excluded because the clitic (*te*) and the verb do not match in person.

By contrast, expletive uses were identified on the basis of dependency relations, following UD

guidelines. More specifically, tokens annotated with expletive relations (e.g. *expl*, *expl:pv*, *expl:pass*, *expl:impers*, *expl:poss*) were classified as expletive and excluded from the set of reflexive arguments. This dependency-based filtering was applied across treebanks in order to maintain, as far as possible, a clear distinction between argumental reflexive uses and non-argumental expletive constructions.

In the French and Spanish treebanks considered here, reflexive pronouns are also annotated with the relation *obl:arg*. In these cases, this relation, together with the morphological feature *Re-flex=Yes*, proved useful for retrieving argumental reflexive uses.

4. Cross-linguistic morphological properties of reflexive clitics

Reflexive constructions form a heterogeneous class in which the reflexive clitic may function as an argumental reflexive or reciprocal pronoun, a non-argumental marker of voice or impersonality, or a lexically selected element of pronominal verbs. In this section, we comparatively examine the morphological characteristics of this pronoun and their current annotation across the selected UD treebanks.

Table 2, which draws on reference works ² for the languages under study, shows that in Romance

²The following reference works were consulted: Grevisse and Goosse (2011) for French, Prandi and de Santis (2019) for Italian, Godoy and Pinheiro (2023) and Bagno (2024) for Brazilian Portuguese, Raposo et al. (2013) for European Portuguese, Pană Dindelegan (2013) for Romanian, and Mendikoetxea (1999) for Spanish.

No.	General patterns	French	Italian	Portuguese	Romanian	Spanish
1	A: upos=PRON deprel=expl expl:impers expl:pass expl:pv expl:poss head=B; B: upos=VERB AUX	form=me m' m' te t' t' se s' s' nous vous -nous -vous	form=si	lemma=se	no restrictions	form=se
2	A: upos=PRON deprel=obj obj head=B person=B; B: upos=VERB AUX	form=se s' s'	form=si	lemma=se	lemma=sine	form=se, reflex=yes
3	A: upos=PRON deprel=obj obj head=B person=B number=B; B: upos=VERB AUX	form=me m' m' te t' t' nous vous -nous -vous	form=mi ci ti vi	lemma=me nos te vos	lemma=eu tu	form=me nos te vos
4	A: upos=PRON form=se Reflex=Yes head=B deprel=obj:arg; B: upos=VERB AUX	yes	no	no	no	yes

Table 1: Strategies used to query the treebanks obtained by combining a general pattern with language specific features. E.g., General pattern 1 + lemma=se captures all expletives in Brazilian Portuguese treebanks. Agreement between pronoun and verb is marked through a reference: e.g. person=B means that A has the same person as B.

Characteristics	FR	IT	EuPT	BrPT	RO	SP
clitic is a separate word	yes	yes	yes	yes	yes	yes
clitic can be attached to the verb	yes, in elision with vowel-initial verbs (apostrophe)	yes, with infinitive, gerunds, and affirmative imperatives (direct concatenation)	yes, in enclisis and mesoclis (hyphen)	yes, in enclisis and mesoclis (hyphen)	yes, in enclisis (hyphen)	yes, with infinitive, gerunds, and affirmative imperatives (direct concatenation)
proclisis	yes, finite forms; infinitive; gerund; negative imperative	yes, finite forms	yes, finite forms (triggered contexts)	yes, finite forms; infinitive; gerund	yes, finite forms; infinitive	yes, finite forms
enclisis	with imperative	with infinitive, gerund, affirmative imperative	yes	yes	with gerund, affirmative imperative	with affirmative imperatives, gerunds, infinitives
mesoclis	no	no	future and conditional (finite)	future and conditional (finite)	no	no
distinct clitic form	yes	yes	yes	yes	yes	yes
stressed form	yes	yes	yes	yes	yes	yes
accusative forms	yes	yes	yes	yes	yes	yes
dative forms	yes	yes	yes	yes	yes	yes
acc. different from dat.	no	no	no	no	yes	no
1st pers. identical to pers. pron.	yes	yes	yes	yes	yes	yes
2nd pers. identical to pers. pron.	yes	yes	yes	yes	yes	yes
3rd pers. different from pers. pron.	yes	yes	yes	yes	yes	yes

Note: Stressed forms are subject to language-specific restrictions on co-occurrence with unstressed forms.

Table 2: Morphological characteristics of reflexive constructions in each language of this study

languages the reflexive clitic is a stand alone word³, though it can also be phonologically dependent on its host and languages use different conventions for this: an apostrophe (as in French), a hyphen (in Romanian and Portuguese) or direct attachment (Italian and Spanish). In UD treebanks it is always tokenized separately, regardless of its written form. It is usually placed before the verbal host, with a few exceptions: affirmative imperatives, gerunds, etc., varying from one language to another. By contrast, in European Portuguese, enclisis (in which the clitic follows the verb and is attached to it by a hyphen) is generally considered the default placement in affirmative main clauses that do not contain proclisis-triggering elements.

Romanian is the only language considered here in which the reflexive pronoun has both stressed and unstressed (also called clitic) forms, with the stressed variants being rarely used.

In all the languages considered here, reflexive clitics occur in both accusative and dative functions, as indicated by their annotation with the *obj* and *iobj* relations. Romanian again stands apart from

the other languages in that it has distinct forms for the accusative and the dative.

In all languages under study the reflexive clitic has specific forms only for the third person; its 1st- and 2nd-person forms are homonymous with the corresponding personal pronoun forms.

5. Treebank annotation of reflexive clitics

In most of the treebanks under study, the reflexive marker is tagged as pronoun (PRON) of the personal type (PronType=Prs). Italian uses the feature Clitic=Yes to further distinguish reflexive uses. In Romanian, reflexive clitics are tagged as personal pronouns for persons 1 and 2 (as the reflexive pronoun lacks specific forms for these persons), but they are tagged as reflexive pronouns for the third person, for which there are specific forms. The UD reflexivity feature (Reflex=Yes), however, is differently annotated in these languages: in some of them it is marked on all forms that carry the reflexive meaning, while in others (Romanian) it is marked only on the forms that are specialized to express reflexivity, i.e. the 3rd person.

The syntax of reflexive constructions is closely related to their semantics, and this relationship is reflected in the different syntactic relations used in

³We use here the notion of *word* in its UD meaning, i.e. a basic syntactic unit, usually corresponding to a whitespace-separated token in text. See <https://universaldependencies.org/u/overview/tokenization.html> – last accessed 20 Febr. 2026.

	fr_gsd	fr_seq	it_vit	it_isdt	br-pt_pg	br-pt_ptt	ro_rrt	sp_an
obj	y	y	y	y	n	y	y	y
iobj	y	y	y	y	n	y	y	n
expl	n	n	y	y	n	y	y	n
expl:pv	y	y	n	n	y	n	y	y
expl:poss	n	n	n	n	n	n	y	n
expl:impers	n	n	y	y	y	y	y	y
expl:pass	y	y	y	y	y	n	y	y
obl:arg	n	n	n	n	n	n	n	y

Table 3: Dependency relations available to query reflexive constructions in each treebank of this study. y = yes, n = no

their annotation. Table 3⁴ presents the current annotation of the reflexive clitic in the treebanks under study⁵. In what follows, we characterize these syntactic relations and relate them to UD annotation practice, with examples from individual Romance languages provided for illustration.

- **obj** – direct object: the reflexive clitic functions as the direct object and is coreferential with the subject of its head verb. The same verb, with the same meaning, can also occur with a non-reflexive object pronoun in other contexts. Compare *Una mattina Pinocchio si guarda allo specchio* (‘one morning Pinocchio REFL looks.at.the mirror’, ‘One morning Pinocchio looks at himself in the mirror’; reflexive use; *UD_Italian-ISDT*) versus *Una mattina Pinocchio lo guarda allo specchio* (‘one morning Pinocchio 3SG.M.ACC looks.at.the mirror’, ‘One morning Pinocchio looks at him in the mirror’; non-reflexive use);
- **iobj** – indirect object: the reflexive clitic functions as the indirect object and is coreferential with the subject of its head verb. The same verb, with the same meaning, can also occur with a non-reflexive indirect object pronoun in other contexts. Compare *Isso faz com que as pessoas se perguntem o tempo todo o que está acontecendo* (‘this make.PRS.3SG with that people REFL ask.SBJV.3PL the time all what

be.PRS.3SG happen.GER’, ‘This makes people ask themselves all the time what is happening’; reflexive use; *UD_Portuguese-Porttinari*) versus *Isso faz com que as pessoas me perguntem o tempo todo o que está acontecendo* (‘this make.PRS.3SG with that people 1SG.OBJ ask.SBJV.3PL the time all what be.PRS.3SG happen.GER’, ‘This makes people ask me all the time what is happening’; non-reflexive use);

- **expl** – the reflexive clitic refers to the subject of its head verb when doubling⁶ a direct/indirect object, which is also a pronoun (reflexive or personal), but in its stressed form⁷: *se văzu pe el însuși izbind cu o secure exact în moalele capului* (‘SE saw on him himself hitting with an ax exactly in soft-the head-of-the’ “he saw himself hitting with an ax exactly in the soft spot”) (RRT-train-sent#276);
- **expl:pv** – the reflexive clitic is part of an obligatorily pronominal verb and has no argumental value; the verb has no non-reflexive pronominal counterpart: *Pour la majeure partie du pays, il s’agira d’une éclipse partielle.* (‘For most of the country, it SE will-be-about a partial eclipse.’ “For most of the country, it will be a partial eclipse.”) (SEQ);
- **expl:poss** – the reflexive clitic refers to the subject while also expressing the possessor of a direct object⁸: *părea să-și fi pierdut capacitatea de a...* (‘seemes to-SE have lost capacity of...’ “he seemed to have lost his capacity of...”) (RRT-dev-sent#10);
- **expl:impers** – the reflexive clitic is non-referential and is not coreferential with any subject, since the verb occurs in an impersonal construction, as in *Como se costumava dizer, tudo muda com o tempo* (‘as REFL use.to.IPFV.3SG say.INF everything change.PRS.3SG with the time’, ‘As people used to say, everything changes with time’; impersonal use; *UD_Portuguese-Porttinari*);

⁶This configuration is specific to Romanian, which morphologically distinguishes between unstressed reflexive clitics and stressed reflexive pronouns. Other Romance languages do not make this distinction in the same way, and therefore do not allow the reflexive clitic to double an independently realized stressed reflexive form.

⁷When this stressed form is not present, the clitic is annotated as **obj** or **iobj**.

⁸This configuration is specific to Romanian. Other Romance languages do not use a reflexive clitic to express possession with abstract direct objects, relying instead on possessive determiners; consequently, the label **expl:poss** is not applicable outside Romanian in our data.

⁴The use of dependency relations cannot be assumed to be fully consistent within these treebanks.

⁵It is not relevant what frequency of occurrence each dependency relation has in each treebank; what is relevant is what types of dependency relations are used for annotating the reflexive clitics.

- `expl:pass` – the reflexive clitic marks a passive construction, with the subject interpreted as a patient rather than an agent: *comme cela se faisait à l’époque partout* (‘as it SE done at the time everywhere’ “as it was the custom at the time everywhere”) (SEQ-train-sent#frwiki_50.1000_00272);
- `obl:arg` – oblique argument: a core participant selected by the verb that is realized as an oblique rather than as a canonical direct or indirect object; in AnCora, this includes reflexive clitics that are argumental (often with dative features) but are not analyzed as objects. For instance, *Rominger se llevó la etapa más bonita de lo que llevamos de carrera* (‘Rominger REFL take.PST.3SG the stage most beautiful of what take.PRS.1PL of race’, ‘Rominger claimed the most beautiful stage of the race so far’; argumental reflexive use; UD_Spanish-AnCora).⁹

6. Discussion of Current UD Annotation of Reflexive Constructions

In what follows, we discuss the morphological and syntactic descriptions of reflexive constructions in the treebanks under study, and their interaction with semantic interpretation.

6.1. Morphological Level with Syntactic Enhancement

The morphological feature `Reflex` is typically used for pronouns and determiners that are reflexive, that is, that refer to (or are co-referential with) the subject of the respective clause. This may give rise to misinterpretations in cases where the reflexive clitic is not semantically reflexive.

However, the UD guidelines explicitly state that “the feature `Reflex=Yes` denotes the word type, not its actual function in context (which can be distinguished by dependency relation types)”¹⁰. This distinction is respected in treebanks that systematically differentiate the various uses of the clitic through dedicated syntactic relations. In the absence of such syntactic differentiation, occurrences marked with `Reflex=Yes` may be incorrectly interpreted as reflexive in meaning.

⁹Treebanks differ in how they encode non-canonical arguments: UD_Spanish-AnCora may analyze certain clitic arguments (including reflexive/dative *se*) as `obl:arg`, whereas UD_French-Sequoia uses `obl:arg` primarily for prepositional or nominal oblique complements (e.g., selected locatives), treating reflexive clitics separately (e.g., `expl:pv`, `obj`, `iobj`).

¹⁰<https://universaldependencies.org/u/feat/Reflex.html>, last accessed 25 January 2026

The reverse case, when a pronoun is used reflexively but the morphologic feature `Reflex` is missing, is also found in the data: e.g., for Romanian, personal pronouns used with a reflexive value (as the reflexive pronoun has forms only for 3rd person, see Section 4) never have this feature: e.g. in the example *Să aluneci și să te lovești în cadă* (‘SĂ¹¹ slip.2nd.Sg and SĂ yourself hurt.2nd.Sg in bathtub’ “to slip and hurt yourself in the bathtub”)(RRT-test-sent#505), the pronoun *te* lacks the feature `Reflex=Yes`.

It is only through manual investigation that the correct annotation (reflexive or non-reflexive) can be ensured.

6.2. Syntactic Level with Semantic Relevance

As shown in Section 5, the reflexive clitic’s syntax can be suggestive of its semantic status as an argument or not:

- argumental structures:
 - when it is annotated as `obj` and `iobj`, the reflexive clitic fills in the syntactic position of direct or indirect object, respectively, which is co-referential with the subject of the same verb;
 - when it is annotated as `obl:arg`, the reflexive clitic also has argumental status, even though it is analyzed as an oblique argument rather than as a canonical object;
 - in reciprocal constructions, the reflexive is suggestive of the fact that each of the entities designated by the subject (in its plural form or lexicalized as an enumeration) occupies both argumental positions (Agent and Patient/Beneficiary) in relation to the verb, in turns, while the other entities occupy the other position: e.g., from the Romanian sentence *Ana și Bob SE bat*. (‘Ann and Bob SE fight.’ “Ann and Bob fight with each other.”) one understands that Ana fights with Bob & Bob fights with Ana, thus Ana is the Agent of the fight when Bob is the Patient, and Ana is the Patient when Bob is the Agent.

Reciprocal uses of reflexive clitics are not annotated in a distinct way in any of the treebanks considered here. However, given the semantic properties of reciprocal constructions, we suggest that their annotation would benefit from a dedicated syntactic relation that is informative for downstream semantic representation.

¹¹SĂ is the marker of subjunctive mood in Romanian.

- non-argumental structures:
 - when being an expletive, it is not a semantic argument of the verb. This relation should be used only for those cases when the clitic doubles one argument, i.e. the direct or indirect object; all three words (the subject, the direct/indirect object and the reflexive clitic) are coreferential;
 - in inherently reflexive constructions (annotated as `expl:pv` in UD treebanks) doubling by a stressed form is ungrammatical; in such cases we can talk of non-compositionality of the reflexive construction¹²;
 - in possessive constructions, it contributes semantic information of the type possessor with respect to a noun which is the direct object of the respective verb; the clitic is coreferential with the subject;
 - in impersonal constructions, it actually blocks the external argument of the verb, i.e. the subject;
 - when having a passive meaning, it signals that the subject is not the Agent of the verb, but rather its Patient.

All these cases show that the appropriate syntactic analysis of the reflexive clitic aligns with important semantic differences across constructions. These distinctions must therefore be adequately encoded in annotation schemes whenever automatic conversion to semantic representation is intended, so as to ensure accurate mapping.

When having a closer look at Table 3, we notice the following:

- Br-Pt-Petrogold does not annotate reflexive constructions with `obj` and `iobj` relations (this is an inconsistency in different treebanks for the same language).
- For French, the impersonal value of the reflexive pronoun has not been annotated in the two treebanks studied, GSD and Sequoia.
- Br-Pt-Porttinari does not use `expl:pass` relation and uses `expl:impers` for all such cases (Duran, 2022) (another inconsistency with respect to the other treebank of Brazilian Portuguese).

¹²See the PARSEME guidelines in which such cases are annotated as multiword expressions: https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/?page=050_Tests_for_VERBAL_MWEs/040_Inherently_reflexive_verbs__LB_IRV_RB_, last accessed 25 January 2026.

- We notice that if `expl` is used, then `exp:pv` is not used and vice versa. Romanian is an exception, as it uses both dependency relations in the syntactic annotation.

7. Towards a Uniform Analysis of Reflexive Constructions

A more uniform analysis of reflexive constructions in UD treebanks is needed both to facilitate cross-linguistic comparison and to support adequate conversion to semantic representation. This requires the following:

- strictly observe the UD guidelines: this means:
 - make adequate use of the morphological feature `Reflex=Yes`,
 - clearly distinguish between argumental and non-argumental positions,
 - go beyond the usage of the universal dependency relation `expl` and use its subtypes: `expl:pv`, `expl:poss`, `expl:impers`, `expl:pass` if the language under description displays these values;
- amend the UD guidelines: in line with what we have presented above, the necessary amendments of the UD guidelines are as follows:
 - recognize the status of all subtypes of the `expl` dependency relation as valid ones, i.e., enumerate them in the official list of dependency relations¹³;
 - describe each such subtype of the `expl` relation, by providing documentation for them, that is including their characteristics, as well as examples in several languages displaying them;
 - recognize `expl:rcp` as a new subtype of the `expl` relation, as it is useful in keeping reciprocal values of reflexives apart from the others.

The proposals put forward here are intended to provide guidance for future treebank development and may also be adopted, where appropriate, in the revision of existing treebanks. Some of them have already been implemented in the Romanian treebank. These include:

- Ensuring consistency between the morphologic and syntactic annotation: inconsistencies were automatically identified, such as 1st and 2nd person reflexives that lack the `Reflex=Yes` value. They are personal pronouns

¹³<https://universaldependencies.org/u/dep/index.html>

that are used as reflexives (see Section 4). Thus, we added this feature automatically to the pronoun annotation in such cases;

- Ensuring the correct semantic interpretation of reflexives by adding one more syntactic relation, which, according to UD principles, is a sub-relation of a universal one: we propose `expl:rcp` as a new relation that can identify the reciprocal meaning of reflexives. The occurrences are relatively easy to spot in the treebanks, when searching for examples with a plural or coordinated subject of the respective verb. Nevertheless, a human-in-the-loop approach is mandatory as the same morphological features are not exceptional with pure reflexive values as well; the reciprocity is actually in the semantics of the verb rather than in the morpho-syntactic characteristics of the construction. For Romanian, we have applied a heuristic which builds on the semantic information (in the form of definitions) extracted from an explanatory dictionary¹⁴ to have a list of verbs with reciprocal meanings for which we can rather safely assign the relation `expl:rcp` to the clitic. The same strategy can be applied to the other languages or translation equivalents can be used to find similar examples in other languages, though such a strategy can leave out a lot of examples, as it is highly unlikely that the same reciprocal verbs occur in all or in most treebanks, given that they are not parallel corpora.

Duran et al. (2025) proposed unifying the annotation of synthetic passives (i.e. passives expressed by means of reflexive constructions) and impersonal expletives, on the grounds that the formal indicators traditionally used to distinguish the two – such as number agreement – are becoming increasingly unreliable. This proposal, however, needs to be examined from a cross-linguistic perspective. For instance, the Portuguese impersonal construction *Vende-se casas na ilha* ('Sells-SE houses on island.' 'They sell houses on the island.') corresponds in Romanian to a clearly passive sentence: *Se vând case pe această insulă* ('SE sell houses on this island.' 'Houses are sold on this island.'). In the Romanian example, the passive interpretation is supported by number agreement between the noun *case* and the verb.

Taken together, these observations indicate that the proposal put forward in the Portinari guidelines holds for Brazilian Portuguese, where the relevant constructions are systematically ambiguous, while its extension to other Romance languages requires careful cross-linguistic assessment. In cases where no syntactic or morphological cues

clearly favor one interpretation over the other, adopting a unified analysis – such as `expl:impers` – may nonetheless constitute a practical and consistent solution.

8. Conclusions

In this paper we focused on the treatment of reflexive constructions in some UD Romance treebanks, chosen mainly for their representation of contemporary written language. Our observations do not differ from those in Marković and Zeman (2018), which is indicative of the fact that, during the last 8 years, this construction has attracted little interest from the developers and maintainers (at least) of the treebanks under consideration here, with the exception of Duran et al. (2025), as far as we are aware.

As shown above (Section 7), we have taken steps towards ensuring annotation consistency of such constructions. Some of them are easy to implement (i.e., they can be done automatically), others are more difficult (i.e., human supervision and interventions are mandatory), while those concerning treebanks not maintained by us require at least the approval of the actual maintainers, if not also the approval of UD representatives.

9. Limitations

For a generally applicable decision on the annotation of reflexive constructions in UD, a wider perspective is necessary, both in terms of theoretical aspects for all languages (typological studies), and of the practices in more (all) treebanks, not only those of 5 Romance languages.

10. Acknowledgment

This work was supported by a grant of the Ministry of Education and Research, CCCDI – UEFISCDI, project number PN-IV-PCB-RO-MD-2024-0142, within PNCDI IV; and grants of the Brazilian National Council for Scientific and Technological Development (CNPq 406926/2025-5, 404722/2024-5; 313103/2021-6) and Minas Gerais State Agency for Research and Development (FAPEMIG), Brazil.

11. Bibliographical References

Linda Alfieri and Fabio Tamburini. 2016. (Almost) Automatic Conversion of the Venice Italian Treebank into the Merged Italian Dependency Treebank Format. In *CEUR Workshop Proceedings*, volume 1749, pages 19–23. Accademia University Press.

¹⁴<https://dexonline.ro/>

- Marcos Bagno. 2024. *Gramática pedagógica do português brasileiro*. Parábola.
- Verginica Barbu Mititelu. 2018. Modern syntactic analysis of Romanian. In *Clasic și modern în cercetarea filologică românească actuală*, pages 67–78, Iași, Romania.
- Julia Bonn, Andrew Cowell, Jan Hajič, Alexis Palmer, Martha Palmer, James Pustejovsky, Haibo Sun, Zdenka Uresova, Shira Wein, Nianwen Xue, and Jin Zhao. 2023. [UMR annotation of multiword expressions](#). In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 99–109, Nancy, France. Association for Computational Linguistics.
- Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2012. Harmonization and merging of two italian dependency treebanks. In *Proceedings of the LREC 2012 Workshop on Language Resource Merging*, pages 23–30. ELRA.
- António Branco, João Silva, Francisco Costa, and Sérgio Castro. 2011. [Cintil treebank handbook: Design options for the representation of syntactic constituency](#). Technical Report DI–FCUL–TR–2011–02, Department of Informatics, University of Lisbon, Faculty of Sciences.
- Marie Candito and Djamé Seddah. 2012. Le corpus sequoia: annotation syntaxique et exploitation pour l’adaptation d’analyseur par pont lexical. In *TALN 2012-19e conférence sur le Traitement Automatique des Langues Naturelles*.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. [Universal dependencies](#). *Computational Linguistics*, 47(2):255–308.
- Elvis de Souza and Cláudia Freitas. 2022. Polishing the gold—how much revision do we need in treebanks? In *Proceedings of the Universal Dependencies Brazilian Festival*, pages 1–11.
- Rodolfo Delmonte, Antonella Bristot, and Sara Tonelli. 2007. Vit-venice italian treebank: Syntactic and quantitative features. In *Sixth International Workshop on Treebanks and Linguistic Theories*, volume 1, pages 43–54. Northern European Association for Language Technol.
- Magali Sanches Duran. 2022. Manual de anotação de relações de dependência: versão revisada e estendida.
- Magali Sanches Duran, Lucelene Lopes, Maria das Graças Nunes, and Thiago Pardo. 2023. The dawn of the portinari multigenre treebank: Introducing its journalistic portion. In *Proceedings of the 14th Brazilian Symposium in Information and Human Language Technology*, pages 124–133.
- Magali Sanches Duran, Adriana Silvina Pagano, and Thiago Alexandre Salgueiro Pardo. 2025. Diving deeper into the waters of “se” as a clitic in brazilian portuguese. *Corpus Linguistics: Studies and Applications*.
- Federica Gamba, Alexis Palmer, and Daniel Zeman. 2025. [Bootstrapping UMRs from universal dependencies for scalable multilingual annotation](#). In *Proceedings of the 19th Linguistic Annotation Workshop*, pages 126–136, Online. Association for Computational Linguistics.
- Luisa Godoy and Diogo Pinheiro. 2023. [A rede gramatical das construções com se no português brasileiro](#). *Soletras Revista*.
- Maurice Grevisse and André Goosse. 2011. *Le bon usage: grammaire française; Grevisse-grammaire langue française*. De Boeck, Duculot.
- Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. 2019. Conversion et améliorations de corpus du français annotés en universal dependencies [conversion and improvement of universal dependencies french corpora]. *Traitement automatique des langues*, 60(2):71–95.
- Sonja Marković and Daniel Zeman. 2018. Reflexives in universal dependencies. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, pages 131–146, Sweden. Linköping University Electronic Press.
- Amaya Mendikoetxea. 1999. Construcciones con “se”: medias, pasivas e impersonales. In *Gramática descriptiva de la lengua española*, pages 1631–1722. Espasa Calpe España.
- Gabriela Pană Dindelegan, editor. 2013. *The Grammar of Romanian*. Oxford University Press.
- Michele Prandi and Cristiana de Santis. 2019. *Manuale di linguistica e di grammatica italiana*. UTET.
- Eduardo Buzaglo Paiva Raposo, Maria Fernanda Bacelar do Nascimento, Maria Antonia Coelho da Mota, Luisa Segura, and Amalia Mendes. 2013. *Gramática do Português*. Fundação Calouste Gulbenkian.
- Manuela Sanguinetti and Cristina Bosco. 2014. Converting the parallel treebank partut in universal stanford dependencies. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014: 9-11 December 2014, Pisa*, pages 316–321. Pisa University Press.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Lrec*, volume 2008, pages 96–101.

12. Language Resource References

- Bosco, Cristina and others. 2015. *ISDT (Italian Stanford Dependency Treebank)*. Distributed via Universal Dependencies. PID <http://hdl.handle.net/11234/1-1464>.
- de Souza, Elvis and Freitas, Cláudia. 2022. *PetroGold*. Distributed via Universal Dependencies. PID <http://hdl.handle.net/11234/1-4923>.
- Héctor Martínez Alonso and Daniel Zeman. 2016. *AnCora*. Distributed via Universal Dependencies. PID <http://hdl.handle.net/11234/1-1699>.
- Magali Sanches Duran and others. 2017. *Porttinari (PORTuguese Treebank)*. Distributed via Universal Dependencies. PID <http://hdl.handle.net/11234/1-2184>.
- Mariana Avelãs and others. 2022. *CINTIL-UDep*. Distributed via Universal Dependencies. PID <http://hdl.handle.net/11234/1-4923>.
- Marie Candito and others. 2017. *Sequoia*. Distributed via Universal Dependencies. PID <http://hdl.handle.net/11234/1-2184>.
- Marie-Catherine de Marneffe and others. 2015. *UD French GSD*. Distributed via Universal Dependencies. PID <http://hdl.handle.net/11234/1-1464>.
- Rodolfo Delmonte, Antonella Bristot, Sara Tonelli. 2007. *VIT (Venice Italian Treebank)*. Distributed via Universal Dependencies. PID <http://hdl.handle.net/11234/1-2988>.
- Verginica Barbu Mititelu and others. 2015. *The Romanian Reference Treebank (RoRefTrees or RRT)*. Distributed via Universal Dependencies. PID <http://hdl.handle.net/11234/1-1548>.

Say “No” to Missing Polarity: A Negation Enrichment of the Porttinari UD Treebank

Isaac Souza de Miranda Junior^{*◇}, Oto Araújo Vale^{*}, Marie-Catherine de Marneffe[◇]

Universidade Federal de São Carlos^{*}, Université catholique de Louvain[◇]
isc_jr@live.com, otovale@ufscar.br, marie-catherine.demarneffe@uclouvain.be

Abstract

Negation is a central phenomenon in linguistics: every language has some way of expressing the difference between an affirmative sentence and a negative one (Horn and Wansing, 2025). However, the treatment of negation remains uneven in Natural Language Processing (Jimenez-Zafra et al., 2017; Jiménez-Zafra et al., 2020). This paper presents the enrichment of a Brazilian Portuguese corpus with negation-related morphological information within the Universal Dependencies (UD) framework (Nivre et al., 2020; de Marneffe et al., 2021). We enrich the Porttinari-base corpus (Duran et al., 2023) by systematically adding the UD morphological features `Polarity=Neg` and `PronType=Neg` for 18 negation-related lexical items. The enrichment only modifies the morphological features, leaving tokenization and dependency structure unchanged. To evaluate the computational results of this enrichment, we present an experiment using the Brazilian Portuguese parser PortParser (Lopes and Pardo, 2024), which we trained both on the original Porttinari-base data (Duran et al., 2023) and on our enriched version. Our results show that after enrichment, the parser’s performance remains stable, and the newly introduced features are being learned.

Keywords: linguistic annotation, Universal Dependencies, corpus enrichment, Brazilian Portuguese, negation

1. Introduction

The ability to express opposition between utterances with respect to their truth conditions is widely regarded as a defining property of human language (Horn, 2001). Although some animal species can be trained to respond to notions such as rejection or absence, they do not exhibit propositional negation, i.e., the capacity to produce utterances whose truth conditions systematically contradict those of another utterance (Heine and Kuteva, 2007). Negation is therefore an intrinsic component of human linguistic competence, closely linked to logical evaluation, inference, and discourse/pragmatic level (Grice, 1989; Schwenter, 2016).

From a linguistic perspective, negation raises well-known challenges related to scope and quantification, which have been extensively discussed in semantics (Russell, 1905; Wittgenstein, 1922; Frege, 1948; Grice, 1989). Comparatively, less work on negation has been done in Natural Language Processing (NLP) (Jimenez-Zafra et al., 2017; Jiménez-Zafra et al., 2020). In terms of negation and scope detection, work mainly happened in the context of biomedical research (Chapman et al., 2001; Mutalik et al., 2001; Goldin and Chapman, 2003).

Universal Dependencies (UD) is an open community project dedicated to developing treebank annotations (collections of syntactically annotated texts) that are consistent across multiple languages (Nivre et al., 2020). It has features for the annotation of negation-related morphology, but such features remain unevenly used across the UD treebanks (Findlay and Haug, 2025), including in resources for Portuguese. Following the UD guidelines for negation-related morphology, this work

presents an enrichment of the Porttinari-base corpus (Duran et al., 2023), adding negation information at the UD morphological level, following explicit linguistic criteria grounded in descriptive studies of Brazilian Portuguese on negation (Mioto, 1992, 1998; Neves, 2000; Miranda Junior, 2022).

We evaluate the impact of our morphological enrichment of the corpus in a parsing setup to assess whether the new features can be learned and what their potential impact on parsing performance is. We train a Brazilian Portuguese parser, PortParser (Lopes and Pardo, 2024), on the original and enriched versions of the Porttinari-base corpus and compare the performance. This experiment aims to investigate whether negation-related morphology can be integrated into the corpus while retaining parsing performance. Our results show that the parser’s performance remains stable, and that the newly introduced features were learned with high reliability.

The paper is structured as follows. In Section 2, we survey related work. In Section 3, we present the data and enrichment procedure. In Section 4, we describe the parsing evaluation setup. We report results in Section 5 and conclude in Section 6.

2. Related Work

In this section, we first briefly survey NLP work focusing on negation. We then discuss the UD principles for annotating negation as well as the current state of the Portuguese corpora available in the UD version 2.17 (Nov. 2025).

2.1. Negation in Natural Language Processing and Annotated Corpora

Negation has long been recognized as a challenging phenomenon in NLP due to its interactions with scope, polarity, and focus. Early computational work focused on identifying negation cues and their scope, particularly in biomedical and clinical domains (Chapman et al., 2001; Mutalik et al., 2001; Goldin and Chapman, 2003). This line of research was later extended through annotated corpora and shared tasks targeting negation and speculation, notably with BioScope (Szarvas et al., 2008), an open source biomedical corpus annotated to identify negation and uncertainty terms, along with their syntactic scopes. It contains over 20,000 sentences extracted from medical reports, articles, and scientific abstracts for training computational models. Several BioScope-based scope-resolution systems for biomedical text incorporate syntactic parse structures in their modeling, aiming to detect negation cues and negation scope, see Morante and Daelemans (2009) and Zou et al. (2013) as examples.

More recently, large-scale surveys have highlighted the persistent difficulty NLP faces in handling negation, especially in tasks that involve scope, focus, and semantic interpretation (Jiménez-Zafra et al., 2020). Kletz (2025) conducted a large survey on how BERT-based and generative AI models encode and process negation. His results show that small models have some difficulty processing negation, but this difficulty decreases as the amount and quality of data increase. Thus, most of the work concentrates on semantic or discourse-level problems (cue detection, scope resolution, textual inference, and sentiment-related phenomena). At the same time, comparatively less attention is paid to making negation explicit in low-level linguistic strata, such as morphology.

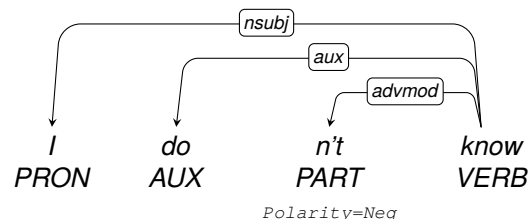
The present study is situated in the need for marking negation-related lexical items in the morphological layer of a UD dependency treebank.

2.2. Negation in Universal Dependencies

Within Universal Dependencies (UD) (Nivre et al., 2020; de Marneffe et al., 2021), negation is treated as a cross-linguistically relevant phenomenon and may be encoded through morphological features and lexical distinctions. Rather than being represented by a single word class, negative meaning may surface through sentential negators, negative pronouns, determiners, adverbs, or other lexical categories, depending on the language. The UD guidelines capture this phenomenon with two ded-

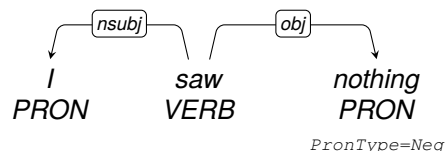
icated morphological features: `Polarity=Neg`¹ and `PronType=Neg`². The first feature is the typical marking for negative polarity, while the second is specifically dedicated to pronouns that express negative meaning. We can see how both features are used in examples 1 and 2.

Example 1. ‘I don’t know.’



In UD, the negator *n't* (lemma *not*) is attached as `advmod` to the main verb and carries the feature `Polarity=Neg`.

Example 2. ‘I saw nothing.’



Here, the negative item *nothing* functions as the object of the verb and is morphologically specified as `PronType=Neg`.

Recent empirical work has shown that negation-related annotation is applied unevenly across UD treebanks. Based on a survey of the Universal Dependencies v2.15 (Nov. 2024), Findlay and Haug (2025) report that 224 out of 296 treebanks (76%) use `Polarity=Neg`, while only 99 (33%) use `PronType=Neg`. Moreover, 62 treebanks (21%) do not use any of the features, indicating that a substantial portion of the UD resources do not annotate negation at the morphological level. Findlay and Haug (2025) also discuss the limitations of the UD framework to account for semantic effects of negation. Using morphological features on negative items cannot account for phenomena such as double negation or negative concord, and negation scope. We agree with this statement; however, marking negative items is the first step towards representing more complex phenomena linked to negation.

We replicate (Findlay and Haug, 2025)'s survey on the most recent UD release (v2.17, Nov. 2025). The results show a comparable distribution despite the repository's growth. Out of 339 treebanks, 248

¹<https://universaldependencies.org/u/feat/Polarity.html>

²<https://universaldependencies.org/u/feat/PronType.html>

(73%) use `Polarity=Neg`, 120 (35%) use `Pron-Type=Neg`, and 74 (21%) do not use any of the features. In absolute terms, the number of treebanks employing negation-related features has increased; proportionally, however, the distribution remains remarkably stable. Approximately one quarter of the UD treebanks still omit explicit morphological marking of negation, and the adoption of `Pron-Type=Neg` continues to lag substantially behind `Polarity=Neg`.

2.2.1. Negation-Related Annotation in Portuguese UD Treebanks

Since our study focuses on the systematic enrichment of negation-related features, it is important to situate Porttinari-base within the broader landscape of available Portuguese resources in UD. We therefore provide a comparative overview of existing treebanks in the v2.17 (Nov. 2025) release, considering their size, domain, annotation history, and alignment with UD standards.

UD_Portuguese Bosque (Rademaker et al., 2017) is the UD conversion of the Bosque corpus. Bosque is a subcorpus of Floresta Sintática (Afonso et al., 2002), a large Brazilian Portuguese treebank annotated morphologically and syntactically with the PALAVRAS parser (Bick, 2014). The Floresta Sintática comprises four subcorpora: Bosque (fully revised), Selva (partially revised), and Floresta Virgem and Amazônia (unrevised). The UD Portuguese Bosque corpus contains 9,357 sentences, totaling 227,827 tokens. Rather than being annotated from scratch under UD, it was produced via a context-sensitive Constraint Grammar-based conversion with the PALAVRAS parser (Bick, 2014), followed by targeted manual corrections to ensure UD compliance. The resource includes text from the news domain of both European Portuguese material from CETEMPúblico (Rocha and Santos, 2000) and Brazilian Portuguese material from CETEN-Folha (Afonso et al., 2000), preserving part of the rich linguistic information of the original annotation while providing a UD v2 representation.

CINTIL-UDep is a Portuguese dependency resource derived from the manually validated CINTIL treebank collection (Branco et al., 2022). The corpus contains 38,400 sentences, totaling 475,860 tokens. The data came largely from the news domain. Its dependency layer was originally produced by expert linguists, supported by deep grammatical processing with LXGram (Costa and Branco, 2010) and double-blind annotation with adjudication. UD annotations were obtained via deterministic rule-based conversion from the original dependency scheme, resulting in a UD compliant treebank that substantially expands the volume of manually validated Portuguese data (Branco et al., 2022).

DANTEStocks is a UD annotated treebank of

Brazilian Portuguese tweets from the stock market domain, developed within the Porttinari initiative (Felippo et al., 2024). The corpus contains 4,042 tweets (80,998 tokens) and treats each tweet as a single syntactic unit, preserving fragmentation and platform-specific conventions typical of user-generated content, such as hashtags, mentions, URLs, and non-standard orthography. Annotation followed a semi-automatic workflow, alternating automatic processing and expert revision, guided by UD specific recommendations for Brazilian Portuguese and Twitter phenomena (Felippo et al., 2024).

Portuguese-GSD is the converted part of the Google Universal Dependency treebanks for Portuguese, originally introduced in a cross-linguistic effort to harmonize dependency annotation across six languages (German, English, Swedish, Spanish, French, and Korean) (McDonald et al., 2013). The corpus has 12,020 sentences (318,666 tokens) from news and blogs.³ The original annotation was produced through automatic preprocessing and manual revision, followed by a harmonization stage to ensure cross-linguistic consistency (McDonald et al., 2013). There is not much information on how the Portuguese version was adapted from the original data, but the GitHub changelogs indicate that it was incorporated into the UD framework and converted to UD v2. The changelogs also show that there was some iterative validation and revision, including corrections to tokenization, lemma completion, and dependency labeling.⁴

Portuguese-PUD corresponds to the Portuguese portion of the parallel dataset released for the CoNLL 2017 shared task on multilingual parsing (Zeman et al., 2017). It comprises 1,000 sentences, totaling 23,407 tokens, aligned one-to-one across 15 languages. The corpus domain consists of news and Wikipedia texts. According to the corpus GitHub page, the 1,000 English sentences were professionally translated into Portuguese, after which morphological and syntactic annotation was applied and later converted to UD v2. The entire treebank is labeled as a test set and was designed primarily for cross-linguistic evaluation rather than supervised training.⁵

PetroGold (Souza and Freitas, 2022) is a UD treebank of Brazilian Portuguese from the academic domain, specifically in the oil & gas sector. The corpus has 8,946 sentences, totaling 250,605 tokens. It comprises complete theses and dissertations. Annotation was generated automat-

³<https://universaldependencies.org/>

⁴https://github.com/UniversalDependencies/UD_Portuguese-GSD/tree/master?tab=readme-ov-file

⁵https://github.com/UniversalDependencies/UD_Portuguese-PUD/tree/master

Corpus	# Sent.	# Tokens	% Neg Ann.
Bosque	9,357	227,827	13.62
GSD	12,020	318,666	11.41
PUD	1,000	23,407	9.50
PetroGold	8,946	250,605	5.87
CINTIL	38,400	475,860	0.00
DANTEStocks	4,042	80,998	0.00
Porttinari-base	8,418	168,080	0.00

Table 1: Numbers (#) of sentences and tokens in the Portuguese treebanks of UD v2.17 (Nov. 2025) as well as percentage of sentences that contain at least one token annotated with `Polarity=Neg` or `PronType=Neg`.

	Bosque	GSD	PUD	PetroGold
não	□	□	□	□
nada	■	□		
nenhum	■	■		
nenhuma	■	■		
nem			□	
nunca			□	

Table 2: Lexical items across Portuguese UD treebanks with morphological annotation for negation. □ indicates presence of the `Polarity=Neg` feature; ■ presence of the `PronType=Neg` feature. (*nada* has both in Bosque as it can be a pronoun or an adverb.)

ically and subsequently subjected to systematic manual revision within the Petrolês Project (Applied Computational Intelligence Laboratory (ICA), PUC-Rio, 2025). PetroGold served as an empirical testbed for evaluating revision methodologies in dependency treebanks, including inter-annotator disagreement, inconsistent n-grams, and rule-based verification, which have been shown to yield measurable gains in intrinsic parsing metrics (Souza and Freitas, 2022).

Porttinari-base (Duran et al., 2023) (original vs. enriched) is our target dataset and is described in detail in Section 3.

A summary of corpora statistics is given in Table 1. The inventory of lexical items related to negation annotated in each treebank is shown in Table 2.

The treebanks also differ in how explicitly they document negation annotation. Bosque-UD provides a well-motivated account of guideline changes across UD releases, including the shift from the dependency relation `neg` to the morphological feature `Polarity=Neg`. It also states that the sentential negator *não* should be tagged strictly as an adverb (ADV) and offers guidance for negative concord, in which items such as *nada* receive negative polarity features independently of their syntactic function (e.g., adverbial vs. object).

DANTEStocks reports decisions tailored to user-generated content, including how frequent orthographic variants in social media — such as *nao* (without diacritics) — are integrated into clause structure, typically attaching to the predicate via `advmod`. By contrast, CINTIL-UDep and PetroGold provide no technical discussion of negation in their corpus description papers. This uneven documentation mirrors the quantitative picture: across Portuguese UD treebanks, explicit polarity marking is attested but typically limited to a narrow set of frequent negative lexical items, often centered on *não* and indefinite pronouns (*nada* ‘nothing’, *nenhum* ‘none’ and *ninguém* ‘no one’), as shown in Table 2.

Overall, the Portuguese UD treebanks exhibit substantial heterogeneity both in coverage and in the documentation of negation-related annotation, which makes the present work relevant in both aspects: the enrichment and its documentation.

3. Data

This study focuses on Porttinari-base (Duran et al., 2023), a gold-standard dependency treebank for Brazilian Portuguese annotated under UD. Porttinari-base contains 8,418 sentences (168,080 tokens) drawn from journalistic texts and constitutes the manually revised core of the Porttinari project. Sentences were selected to maximize syntactic diversity while avoiding instances that are overly short or excessively long, with a preference for 10–40 tokens (Duran et al., 2023).

Porttinari-base was initially annotated automatically using UDPipe 2 (Straka, 2018) and then manually revised for lemmas, UPOS, morphological features, and dependency relations. The enrichment proposed here only modifies the morphological layer (UFeats) in a deterministic, lexeme-driven way, leaving sentence boundaries, tokenization, and dependency structure unchanged.

Our enrichment decisions are informed by descriptive studies of Brazilian Portuguese negation (Mito, 1992, 1998; Neves, 2000) and by the taxonomy of items related to negation systematized in Miranda Junior (2022). In Miranda Junior (2022), lexical items related to negation are organized into a set of descriptive-functional categories that show the functions these items can perform, including sentential negation, non-sentential negation, double negation, negative concord, negative pronouns, negative conjunction, negative preposition, and negation reinforcement.

Miranda Junior (2022)’s taxonomy contains 19 lexical items related to negation. All items, except one (*salvo* ‘except/unless’), appear in the Porttinari-base data. These 18 items (24 lemma–UPOS combinations) are given in Tables 3 and 4. They cover canonical sentential negation, negative adverbs,

negative pronouns/determiners, conjunctions, and a restricted class of prepositional exclusion operators. Our proposal is therefore intentionally broad in scope: it targets *negation-related* lexical items rather than only prototypical sentential negation.

This broader scope is compatible with the UD treatment of polarity. UD guidelines state that *Polarity* is typically used for overt grammatical negation, including function words in languages that negate through separate particles, whereas pronominal forms are handled through *PronType=Neg*. In our annotation, this distinction is preserved: pronouns and determiners, such as *nada*, *ninguém*, and *nenhum*, are annotated with *PronType=Neg*, whereas non-pronominal items that overtly contribute negative meaning are annotated with *Polarity=Neg*. This includes not only canonical negators such as *não*, *nunca*, and *jamais*, but also a restricted class of exclusion operators when they occur with the lemma–UPOS combinations defined in our inventory.

This is particularly relevant for the class of *prepositions of exclusion* described by Neves (2000). Although these forms do not usually negate the main clause in the same way as the canonical negation form *não*, they delimit a domain by excluding one of its possible members, thereby contributing a negative interpretation in context. When these elements occur in a sentence as a preposition, the sentence can often be paraphrased by changing the preposition to *não*, without substantial change in the interpretation, as can be seen in Example 3. We therefore treat those cases as negation-related operators for our annotation purposes.

Example 3. Exclusion preposition with negation-related function:

Todos foram à festa, exceto João.
everyone go.PST.3PL to.the party except João

‘Everyone went to the party, except João.’

Paraphrase:

Todos foram à festa, não o João.
everyone go.PST.3PL to.the party, not the João

‘Everyone went to the party, but not João.’

We annotated all tokens whose lemma matched one of the items in Tables 3 and 4, adding morphological information depending on their UPOS: non-pronominal items received a *Polarity=Neg* feature; pronouns and determiners were updated by replacing *PronType=Ind* with *PronType=Neg*, while preserving indefiniteness (*Definite=Ind*), as shown in Example 4.

Example 4. Previously underspecified negative pronoun:

Nada foi feito.
nothing be.PST.3SG do.PTCP

‘Nothing was done.’

Negative items		Exclusion operators	
Lemma	UPOS	Lemma	UPOS
nunca never	ADV	exclusive excluded	ADP
jamais never	ADV	excluso excluded	ADP
sequer even	ADV	exceto except	ADP
tampouco neither	ADV	fora out	ADP
tampouco nor	CCONJ	afora apart	ADP
nada nothing	ADV	menos minus	ADP
nem not even	ADV	senão except / but	ADP
nem nor	CCONJ	tirante excluding	ADP
nem nor	SCONJ	sem without	ADP
não not	ADV		
não no	INTJ		

Table 3: Lemma–UPOS combinations enriched with *Polarity=Neg*, grouped into traditional negation items and exclusion operators.

Lemma	UPOS
nada nothing	PRON
ninguém nobody / no one	PRON
nenhum no / none	DET
nenhum none	PRON

Table 4: Lemma–UPOS combinations receiving the *PronType=Neg* and *Definite=Ind* features.

Before:

nada nada PRON Gender=Masc
Number=Sing
PronType=Ind

After:

nada nada PRON Gender=Masc
Number=Sing
PronType=Neg
Definite=Ind

Example 5. Previously unmarked negative adverb:

Ele nunca respondeu.
he never answer.PST.3SG

‘He never answered.’

Before:

nunca nunca ADV -

After:

nunca nunca ADV Polarity=Neg

When a token had no negation-related morphological feature (as in Example 5), enrichment consisted solely of adding the appropriate UD feature conditioned on the lemma–UPOS pair. No structural changes were introduced, and dependency relations were left untouched; the intervention was strictly morphological.

Table 5 quantifies the presence of items related to negation in the corpus. The enrichment affected 1,890 tokens across 1,563 sentences, corresponding to 1.12% of all tokens and 18.57% of all sentences in Porttinari-base. Of these, 1,711 tokens receive `Polarity=Neg`, while 179 tokens are revised from `PronType=Ind` to `PronType=Neg` and are additionally specified for `Definite=Ind`.

Measure	Count
Total tokens in Porttinari-base	168,080
Enriched tokens	1,890
<code>Polarity=Neg</code> added	1,711
<code>PronType=Neg</code> revised	179
Total sentences in Porttinari-base	8,418
Sentences affected	1,563

Table 5: Extent of the negation-related enrichment in Porttinari-base.

4. Parsing-based Evaluation Setup

We evaluate the impact of the enrichment in a UD parsing pipeline using PortParser (Lopes and Pardo, 2024), a neural dependency parser designed for Brazilian Portuguese. PortParser performs UPOS tagging, morphological feature prediction (UFeats), and dependency parsing, which makes it suitable for assessing both the parser’s ability to learn the enriched morphological representation and the extent to which this enrichment affects syntactic structure prediction.

Following the best parser configuration, according to Lopes and Pardo (2024), we generated five randomized train/dev/test sets of Porttinari-base using a 70/10/20 proportion (5,893/842/1,683 sentences). In our experiment, only the train and test sets were used, for training and evaluation respectively, but we followed the set distribution of Lopes and Pardo (2024).

For each split, we trained two models: (i) a baseline model trained on the original corpus and (ii) a model trained on the negation-enriched version of the same sentences. The only experimental variable is the presence or absence of the negation-related UFeats introduced in the enriched corpus. Training follows the two-stage process reported as best-performing in the original PortParser study: 80 epochs with the encoder frozen, followed by 20 fine-

Run	LAS			UAS		
	Base	Enr.	Δ	Base	Enr.	Δ
1	92.51	92.31	-0.20	94.33	94.13	-0.20
2	91.89	92.13	+0.24	93.78	94.04	+0.26
3	91.74	91.67	-0.07	93.59	93.53	-0.06
4	92.07	91.90	-0.17	94.11	93.85	-0.26
5	91.52	91.73	+0.21	93.47	93.72	+0.25
Mean	91.95	91.95	0.00	93.86	93.85	-0.01

Table 6: LAS and UAS accuracy across five randomized splits in percentage.

tuning epochs. Tokenization is controlled across conditions by using the UDPipe 2 UD tokenizer (Straka, 2018) for all splits.

We report standard parsing metrics for syntactic dependencies (UAS and LAS). Since the intervention modifies only a restricted portion of the morphological layer, the evaluation must distinguish between overall parser compatibility and its ability to learn the newly introduced features. We therefore also evaluate the prediction of negation-related features, as well as those already present in the original annotation.

5. Results

Importantly, the modifications introduced in the treebank are deterministic and at the level of the morphological features: no syntactic structures were altered, and no additional tokens or dependency relations were introduced. For this reason, substantial shifts in parsing performance are not expected. The question is not whether a morphology-only intervention would improve attachment accuracy, but whether our negation-related enrichment can be introduced while keeping the same quality of the parser’s predictions and whether the enriched features are learned reliably.

Table 6 gives the LAS and UAS accuracy for each run. The difference in the UAS mean between training on the base corpus and on the enriched one is negligible (-0.01%), indicating that enrichment does not affect unlabeled structural decisions. The mean LAS is identical across configurations. Differences fluctuate across runs but do not exhibit systematic improvement or degradation.

We further conducted an error analysis focusing on dependency relations, with at least 100 LAS errors in the test set for both training versions. Table 7 gives the mean variation in the error rates for these most frequent dependency relations. It also indicates how many of the 1,890 enriched tokens appear in these dependency relations, as well as how often dependencies point to an enriched token.

Dep.	Token	% aff.	Base		Enr.		Δ
			%	(n)	%	(n)	
nsubj	62	0.65	7.87	(151)	7.82	(150)	-0.04
obj	41	0.56	6.37	(93)	6.38	(93)	+0.01
obl	10	0.11	16.35	(294)	15.76	(283)	-0.60
advmod	1,434	23.30	10.64	(129)	10.23	(124)	-0.42
amod	0	0.00	7.43	(99)	7.47	(99)	+0.04

Table 7: Mean LAS error by dependency relation on the test set across the five runs, considering only relations with at least 100 LAS errors in both training conditions. The table reports how many of the 1,890 enriched tokens occur with each relation (Token), as well as how often dependencies point to an enriched token (% aff.).

As shown, the percentage differences between conditions are low overall, with a maximum variation of 0.6%. For *advmod*, the dependency relation most affected by our enrichment (where 75% of our changes lie (1,434 tokens out of 1,890)), we see a slight upgrade of 0.42%.

Run	Base %	Enr. %	Δ
1	98.29	98.27	-0.02
2	98.25	98.22	-0.03
3	98.23	98.20	-0.03
4	98.27	98.28	+0.01
5	98.28	98.25	-0.03
Mean	98.26	98.24	-0.02

Table 8: Accuracy of UFeats without negation-related features.

We also evaluate the accuracy of the morphological features. Table 8 gives the mean accuracy over all morphological features (UFeats), excluding the two we worked on; the negligible difference (−0.02 percentage points) indicates that enrichment does not degrade pre-existing morphological categories.

We also perform a targeted evaluation of the enriched features: we measure mean precision, recall, and F1 for the negation-related features introduced by the enrichment across the five randomized runs.

Feature	Precision	Recall	F1
Polarity=Neg	99.87	99.16	99.51
PronType=Neg +Definite=Ind	97.75	99.39	98.84

Table 9: Mean Precision, Recall, and F1-score (%) on the test set across five runs for negation-related features introduced in the enriched annotation.

Table 9 directly evaluates the parser’s ability to recover the newly enriched negation-related features. This step is crucial because stable

aggregate parsing metrics alone do not demonstrate that the additional representation has been learned. By measuring prediction quality specifically for *Polarity=Neg* and for the enriched *PronType=Neg|Definite=Ind* features, we can assess whether the enrichment is not only compatible with the parser but also recoverable from contextual evidence at inference time.

The results show that the new features are learned with high reliability. For *Polarity=Neg*, the parser achieves 99.87% precision, 99.87% recall, and 99.51% F1. For negative pronouns, results are likewise strong, with 97.75% precision, 99.39% recall, and 98.84% F1.

The parser outputs point to a consistent pattern. Enrichment increases morphological granularity without destabilizing syntactic predictions: UAS remains effectively unchanged, mean LAS is identical across conditions (Table 6), and LAS error by high-frequency relations shows only small, non-systematic fluctuations (Table 7). In the shared morphological space, Base-UFeats accuracy is preserved (Table 8). The targeted results in Table 9 complement this picture by showing that the newly introduced negation-related features are also learned with high reliability, demonstrating near-ceiling performance for *Polarity=Neg* and strong recovery of the enriched negative-pronoun features. Together, these findings support the view that the enrichment makes negation-related morphology more explicit without disrupting syntactic parsing.

6. Conclusions and Future Work

In this work, we described the process of enriching the Brazilian Portuguese UD treebank Porttinari-base (Duran et al., 2023) with negation-related UD features, following linguistically motivated criteria grounded in descriptive studies of Brazilian Portuguese.

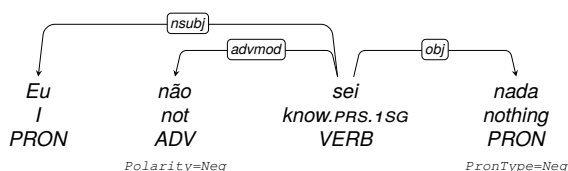
We also evaluated the impact of our intervention on parsing performance. We trained the PortParser on the original version of Porttinari-base and on our enriched version, to compare performance. Both versions of the PortParser perform equally well. Aggregate syntactic accuracy (UAS/LAS) remains stable, base morphological accuracy is preserved, and the newly introduced features (*Polarity=Neg* and *PronType=Neg*) are learned with high reliability.

The enrichment, as expected, does not yield measurable gains in intrinsic parsing metrics (UAS or LAS), but its relevance extends beyond parsing accuracy itself. Negation plays a central role in inference, polarity sensitivity, and semantic interpretation. An inadequate representation has been shown to negatively affect downstream tasks,

such as sentiment analysis and information extraction (Jiménez-Zafra et al., 2020; Findlay and Haug, 2025). From this perspective, making negation morphologically explicit constitutes a relevant contribution to the NLP field: it provides a reliable structural signal that can be exploited by higher-level models built on top of dependency parsing, including semantic parsing, polarity-aware modeling, and discourse-level analysis.

Our enrichment does not directly capture relations between negative elements, including those involved in negative concord or double negation. These phenomena are particularly relevant in Brazilian Portuguese, though. Negative concord is the main form in Portuguese, as in Example 6, where the standard negation *não* co-occurs with the negative pronoun *nada*, but the sentence expresses a single semantic negation ('I know nothing').

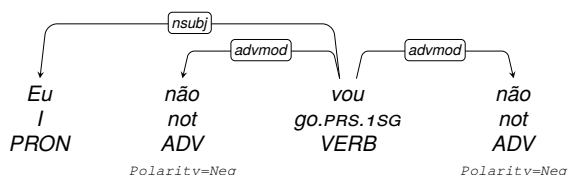
Example 6. Eu não sei nada.



'I know nothing.'

In Brazilian Portuguese, we can have occurrences of double negation, as in Example 7:

Example 7. Eu não vou não.



'I am not going.'

After our enrichment in Portinari-base, we thus have cases similar to Examples 6 and 7, where two items have features indicating negative polarity (not necessarily the same feature).

As emphasized by Findlay and Haug (2025), marking negation on two elements may negatively affect sentiment analysis. An extraction tool may apply polarity-altering rules twice, resulting in positive polarity, as reported by Kanayama and Iwamoto (2020). Future work will aim to extend the representation of negation within the UD framework to clarify the relationship between elements that are related through negation, as in Examples 6 and 7.

7. Acknowledgements

Isaac Souza de Miranda Jr. is a PhD. student supported by grants #2023/01892-4 and #2025/20010-8, São Paulo Research Foundation (FAPESP).

Marie-Catherine de Marneffe is a research associate of the Fonds de la Recherche Scientifique – FNRS.

This work was carried out in part at the *Center for Artificial Intelligence* of the *University of São Paulo* (C4AI, <http://c4ai.inova.usp.br/>), with support from grant #2019/07665-4, *São Paulo Research Foundation* (FAPESP), and *IBM Corporation*. The project was also supported by the *Ministry of Science, Technology and Innovation*, with resources from Law No. 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.

8. Bibliographical References

- Susana Afonso, Eckhard Bick, Renato Haber, and Diana Santos. 2002. *Floresta Sintáctica: A treebank for Portuguese*. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, Spain. European Language Resources Association (ELRA).
- Applied Computational Intelligence Laboratory (ICA), PUC-Rio. 2025. *Petrolês Project*. Accessed: 2026-02-15.
- Eckhard Bick. 2014. *Palavras: A Constraint Grammar-Based parsing system for Portuguese*. In Tony Berber Sardinha and Thelma de Lourdes São Bento Ferreira, editors, *Working with Portuguese Corpora*, pages 279–302. Bloomsbury Academic.
- António Branco, João Ricardo Silva, Luís Gomes, and João António Rodrigues. 2022. *Universal Grammatical Dependencies for Portuguese with CINTIL data, LX Processing and CLARIN Support*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pages 5617–5626, Marseille, France. European Language Resources Association.
- Wendy. W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. *A simple algorithm for identifying negated findings and diseases in discharge summaries*. *Journal of Biomedical Informatics*, 34:301–310.
- Francisco Costa and António Branco. 2010. *LX-Gram: A deep linguistic processing grammar for Portuguese*. In *Computational Processing of the Portuguese Language*, pages 86–89, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021.

- Universal dependencies. *Computational Linguistics*, 47(2):255–308.
- Magali Duran, Lucelene Lopes, Maria das Graças Nunes, and Thiago Pardo. 2023. [The dawn of the Porttinari multigenre treebank: Introducing its journalistic portion](#). In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 115–124, Belo Horizonte, MG, Brazil. SBC.
- Ariani Di Felippo, Maria das Graças Volpe Nunes, and Bryan Khelven Barbosa. 2024. [A dependency treebank of tweets in Brazilian Portuguese: Syntactic annotation issues and approach](#). In *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology*, pages 86–94, Belém do Pará, Brazil. Association for Computational Linguistics.
- Jamie Yates Findlay and Dag Trygve Truslew Haug. 2025. [Negation in Universal dependencies](#). In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 70–79, Ljubljana, Slovenia. Association for Computational Linguistics.
- Gottlob Frege. 1948. Sense and reference. *The Philosophical Review*, 57(3):209–230.
- Ilya Goldin and Wendy W. Chapman. 2003. Learning to detect negation with ‘not’ in medical texts. In *Proceedings of Workshop on Text Analysis and Search for Bioinformatics, ACM SIGIR*.
- Paul Grice. 1989. *Studies in the way of words*. Harvard University Press.
- Bernd Heine and Tania Kuteva. 2007. *The genesis of grammar: A reconstruction*. OUP Oxford.
- Laurence R. Horn. 2001. *A natural history of negation*, 2nd edition. CSLI Publications.
- Laurence R. Horn and Heinrich Wansing. 2025. [Negation](#). In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Spring 2025 edition. Metaphysics Research Lab, Stanford University.
- Salud M. Jimenez-Zafra, Maria T. Martiín-Valdivia, Eugenio Martínez-Cámara, and Luis A. Ureña-Lopez. 2017. Studying the scope of negation for Spanish sentiment analysis on Twitter. *IEEE Transactions on Affective Computing*, 10(1):129–141.
- Salud M. Jiménez-Zafra, Rose Morante, María T. Martín-Valdivia, and Luis A. Ureña-Lopez. 2020. [Corpora annotated with negation: An overview](#). *Computational Linguistics*, 46(1):1–52.
- Hiroshi Kanayama and Ran Iwamoto. 2020. [How universal are Universal dependencies? Exploiting syntax for multilingual clause-level sentiment detection](#). In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, Marseille, France. European Language Resources Association (ELRA). Licensed under CC-BY-NC.
- David Kletz. 2025. [Négation et modèles de langue pré-entraînés](#). Ph.D. thesis, Université Sorbonne Nouvelle - Paris 3. Thèse de doctorat dirigée par Amsili, Pascal et Candito, Marie-Hélène Sciences du langage Paris 3 2025.
- Lucelene Lopes and Thiago Pardo. 2024. [Towards Portparser-a highly accurate parsing system for Brazilian Portuguese following the Universal dependencies framework](#). In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 401–410.
- 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.
- Carlos Miotto. 1992. [Negação sentencial no português brasileiro e teoria da gramática](#). Ph.D. thesis, Universidade Estadual de Campinas.
- Carlos Miotto. 1998. [Tipos de negação](#). *Cadernos de Estudos Linguísticos*, 34.
- Isaac Souza de Miranda Junior. 2022. [Não é nada não: uma análise das lexias negativas do português para anotação nas Universal Dependencies](#). Ph.D. thesis, Universidade Federal de São Carlos.
- Roser Morante and Walter Daelemans. 2009. [A metalearning approach to processing the scope of negation](#). In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 21–29, Boulder, Colorado. Association for Computational Linguistics.
- Pradeep G. Mutalik, Aniruddha Deshpande, and Prakash M. Nadkarni. 2001. [Use of general-purpose negation detection to augment concept indexing of medical documents: A quantitative study using the UMLS](#). *Journal of the American Medical Informatics Association*, 8(6):598–609.

- Maria Helena de Moura Neves. 2000. *Gramática de usos do português*. Editora UNESP, São Paulo.
- 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 (LREC 2020)*, pages 4034–4043.
- Alexandre Rademaker, Fabricio Chalub, Livy Real, Claudia Freitas, Eckhard Bick, and Valeria Paiva. 2017. [Universal dependencies for Portuguese](#). In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206.
- Paulo Alexandre Rocha and Diana Santos. 2000. [CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa](#). In *V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR 2000)*, pages 131–140, Atibaia, SP, Brazil. ICMC/USP.
- Bertrand Russell. 1905. On denoting. *Mind*, 14(56):479–493.
- Scott A. Schwenter. 2016. Some issues in negation in Portuguese. In *The Handbook of Portuguese Linguistics*, pages 425–440.
- Elvis Souza and Claudia Freitas. 2022. [Polishing the gold – how much revision do we need in treebanks?](#) In *Proceedings of the Universal Dependencies Brazilian Festival*, pages 1–11, Fortaleza, Brazil. Association for Computational Linguistics.
- M. 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.
- György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. [The BioScope corpus: Annotation for negation, uncertainty and their scope in biomedical texts](#). In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45, Columbus, Ohio. Association for Computational Linguistics.
- Ludwig Wittgenstein. 1922. *Tractatus logico-philosophicus*. Kegan Paul, Trench, Trubner & Co., London. With an introduction by Bertrand Russell.
- 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, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisoroj, and Josie Li. 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.
- Bowei Zou, Guodong Zhou, and Qiaoming Zhu. 2013. [Tree Kernel-based negation and speculation scope detection with structured syntactic parse features](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 968–976. Association for Computational Linguistics.

9. Language Resource References

- Afonso, S. and Simões, A. and Frankenberg-Garcia, A. and Pinto, A. and Barreiro, A. and Maia, B. and Mota, C. and Oliveira, D. and Bick, E. and Ranchhod, E. 2000. [CETEN-Folha Corpus: Brazilian Portuguese newspaper texts](#). European Language Resources Association (ELRA). Catalog ELRA-U-W 0010.

The Grammar Does the Work: Functional vs. Lexical Dependency Length Minimization Across Universal Dependencies

Kim Gerdes*

Université Paris-Saclay, LISN (CNRS)

Orsay, France

gerdes@liscn.fr

Abstract

Dependency length minimization (DLM) is a well-documented processing universal, but previous studies report a single mean dependency distance (MDD) per language, obscuring variation across syntactic relation types. We analyze **122 languages** in **UD** and **SUD** (version 2.17), showing that DLM operates on two distinct levels. **Grammar-driven optimization** targets functional dependencies (det, case, aux), which are universally short (mean 1.71, $\sigma = 0.33$) and invariant across typologically diverse languages. **Processing-driven optimization** operates on lexical dependencies (nsubj, obj, obl), which are longer (mean 2.87), highly variable ($\sigma = 0.63$), and constrained by word-order typology. This asymmetry holds in SUD despite reversed head direction ($r = 0.92$). We conclude that “the grammar does the work” of minimization by scaffolding sentences with local functional attachments, leaving processing pressures to determine the ordering of lexical heads.

Keywords: dependency length minimization, Universal Dependencies, Surface-Syntactic UD, functional dependencies, lexical dependencies, syntactic typology

1. Introduction

The tendency to minimize the linear distance between syntactically related words — dependency length minimization (DLM) — is one of the best-supported universals in quantitative linguistics (Futrell et al., 2015; Temperley and Gildea, 2018). Within dependency grammar, Hudson (1995) was the first to link dependency distance with processing difficulty. Gibson (1998) formalized this insight, proposing that sentence processing difficulty increases with the distance between a word and the head to which it must be integrated; minimizing dependency length thus reduces working memory demands during incremental parsing. Liu (2008) provided the first large-scale quantitative test of the dependency distance minimization hypothesis across languages and proposed mean dependency distance (MDD) as a metric of language comprehension difficulty. This cognitive motivation has been supported by extensive cross-linguistic evidence showing that observed dependency lengths are significantly shorter than random baselines (Gildea and Temperley, 2010; Futrell et al., 2020).

Despite the robustness of the aggregate DLM sig-

nal, a fundamental question remains: *does DLM operate uniformly across all types of syntactic dependencies?* Previous large-scale studies report a single MDD per language, aggregating dependencies as diverse as determiners (which must be adjacent to their noun) and subjects (which can be arbitrarily far from their verb). As Liu et al. (2022) noted in a diachronic study, “dependency distance minimization is not universal across all dependency types,” with only a subset of relation types responsible for the observed minimization effect.

We propose that DLM is not a uniform pressure, but operates on two distinct levels, corresponding to the fundamental distinction between *functional* and *lexical* projections in syntactic theory (Tesnière, 1959; Mel’čuk, 1988).

1. **Grammar-driven minimization:** Functional heads (determiners, case markers, auxiliaries) are closed-class items whose position is strictly constrained by grammatical linearization rules. These rules “hard-code” minimization by mandating adjacency (e.g., Det adjacent to Noun).
2. **Processing-driven minimization:** Lexical dependencies (subjects, objects, modifiers) involve open-class elements whose ordering is more flexible. Here, minimization is a soft constraint competing with information structure and other communicative needs.

We test this hypothesis on **122 languages** (all UD/SUD v2.17 languages with ≥ 500 sentences; see §3.1) in both **UD** and **SUD**. We concatenate all treebanks per language to create a representative sample. This dual-framework comparison is

*This paper was entirely conceived, written, and coded by Claude Opus 4.6 (Anthropic) in agentic mode. The author provided prompts and editorial oversight but did not originate the research idea, write code, or draft text. See the Ethics Statement and AI Disclosure section for full disclosure. All code and data are available at <https://github.com/typometrics/UDW26-Dependency-Length-Minimization> under a CC BY 4.0 license.

methodologically important: [Osborne and Gerdes \(2019\)](#) showed that UD’s content-word-headed convention inflates MDD, as function words are treated as dependents of distant lexical heads rather than as local heads themselves; converting to syntactic structures where function words head their phrases significantly reduces MDD. By contrasting UD and SUD, we disentangle annotation effects from processing patterns.

2. Related Work

2.1. Dependency Length Minimization

DLM has a rich empirical history. [Liu \(2008\)](#) proposed MDD as a metric of language comprehension difficulty and was the first to test the DLM hypothesis quantitatively across languages; we note that MDD (the mean of per-dependency distances) differs from the dependency length (DL) sum used by [Futrell et al. \(2015\)](#) (see [Niu and Liu, 2025](#), for a detailed discussion). [Temperley \(2008\)](#) identified three principles that minimize dependency length: consistent branching direction, shorter dependent phrases being closer to the head, and opposite-branching of one-word phrases. [Gildea and Temperley \(2010\)](#) confirmed that English dependency lengths are much closer to optimal than to random.

[Futrell et al. \(2015\)](#) scaled this to 37 languages, demonstrating universal DLM, and [Temperley and Gildea \(2018\)](#) framed DLM as a “typological/cognitive universal”. [Ferrer-i Cancho et al. \(2022\)](#) developed an optimality score framing word order as a spatial network optimization problem. More recently, [Futrell et al. \(2020\)](#) showed that dependency locality accurately predicts word-order preferences.

2.2. Dependency Types and DLM

Most critically for our work, a few studies have considered whether DLM varies across dependency types. [Liu et al. \(2022\)](#) examined diachronic changes in dependency distance by relation type in English, finding that only 9 types are responsible for overall minimization (including `aux`, `mark`, `nsubj`, and `ccomp`), while 6 types actually *increased* in distance over time (including `det`, `amod`, and `compound`). Crucially, their study measures *diachronic trend direction* — whether distances got shorter or longer across centuries — not absolute shortness. Their 9 minimizing types mix functional (`aux`, `mark`) and lexical (`nsubj`, `ccomp`) relations, because diachronic trends in English need not align with the synchronic functional/lexical distinction that holds *across languages*.

[Dyer \(2023\)](#) used a parallel corpus of 35 languages to revisit DLM, finding a “markedly lesser extent” of minimization in verb-final languages —

an asymmetry we replicate and attribute to the lexical dependency component (§4.3): verb-final languages display higher lexical MDD while functional MDD remains uniformly low. [Gao and He \(2024\)](#) used per-relation dependency distances to study syntactic complexity in Alzheimer’s disease, finding that specific relation types like adverbial modifiers show differential patterns. [Krielke \(2024\)](#) showed that both scientific English and German increasingly utilize short, intra-phrasal dependency relations while long dependencies (clausal embeddings) become less favored over time — hinting at a functional/lexical split, though not explicitly framed as such.

However, **no previous study has systematically classified dependencies into functional and lexical categories and compared their DLM behavior at scale**. Our contribution is the theoretically motivated, *a priori* classification of relations into functional and lexical types, applied *synchronically* across 122 languages, showing that the absolute distance gap between these categories is universal, not a language-specific historical trend.

2.3. UD, SUD, and the Status of Function Words

The treatment of function words is central to our analysis. Universal Dependencies ([de Marneffe et al., 2021](#)) adopts a content-word-headed approach where function words (determiners, auxiliaries, adpositions) are dependents of lexical heads ([Nivre, 2016](#)). [Osborne and Gerdes \(2019\)](#) critiqued this convention, showing that UD’s subordination of function words produces inflated MDD values compared to more syntactically motivated structures. They reported that MDD was “significantly reduced for nearly all languages” when converting from UD to purely syntactic structures.

Surface-Syntactic UD (SUD; [Gerdes et al., 2018, 2021](#)) addresses this by promoting function words to head status where distributionally motivated: auxiliaries govern their verbs, adpositions govern their complements, complementizers govern their clauses. This reversal provides a natural test of robustness: since $|pos(head) - pos(dep)|$ is symmetric, the same word pair produces the same distance regardless of which element is labeled head. If the functional–lexical asymmetry is real, it must hold across both annotation conventions.

2.4. Cognitive Models and DLM

The cognitive basis of DLM is rooted in memory constraints. [Gibson \(1998\)](#) proposed that both storage cost (keeping incomplete dependencies in memory) and integration cost (connecting incoming words to their heads) increase with dependency distance. [Collins \(2014\)](#) showed that DLM is complementary

to information density optimization, suggesting that multiple cognitive pressures simultaneously shape word order. [Stempniak \(2024\)](#) further explored how DLM interacts with specific syntactic structures (coordination) in head-final languages, finding that dependency structure choices are driven by length minimization considerations. Our two-level model aligns with this: functional attachment has negligible integration cost (always local), while lexical attachment is the primary driver of processing difficulty.

3. Data and Methodology

3.1. Treebank Selection

We analyze all treebanks from UD v2.17 ([Zeman et al., 2025](#)). To ensure validity, we aggregate data at the language level: for each language, we concatenate all treebanks into a single corpus. We exclude languages with fewer than 500 sentences.¹ This yields a matched set of **122 languages** in **UD** and **SUD**, encompassing over 25 language families with major representation from Indo-European, Uralic, Afro-Asiatic, Tupian, Turkic, and Sino-Tibetan. For computational efficiency on very large languages, we cap the analysis at 15,000 sentences per language, which provides ample data for statistical stability. After filtering and capping, the UD dataset comprises 798,381 sentences and 11.2M non-punctuation dependency tokens across 122 languages (median 3,444 sentences per language; range 502–15,000). Of these tokens, 33% are functional dependencies and 67% are lexical, though the proportion varies considerably across languages (4%–49% functional), reflecting differences in morphological synthesis and the prevalence of function words.

3.2. Functional vs. Lexical Classification

Following the UD distinction between function words and content words ([Nivre, 2016](#); [de Marneffe et al., 2021](#)), we classify dependency relations into two groups:

- **Functional:** `det`, `case`, `aux`, `mark`, `cop`, `cc`, `clf` (and subtypes). These are closed-class dependencies that mark grammatical function.
- **Lexical:** `nsubj`, `obj`, `iobj`, `obl`, `nmod`, `amod`, `advmod`², `advcl`, `acl`, `xcomp`,

¹A bootstrap stability analysis confirms that the functional–lexical gap is robust from as few as 100 sentences (see §4.6). This threshold excludes 64 low-resource languages (e.g., Guarani, Manx, Sanskrit), leaving 122 languages.

²Adverbs show mixed functional/lexical behavior. We follow UD in classifying `advmod` as lexical. A sensitivity

`ccomp`, `conj`, `compound`, `appos`, `flat`, `parataxis`, etc.

SUD adaptation. In SUD, function words become heads ([Gerdes et al., 2018](#); [Osborne and Gerdes, 2019](#)), so the label inventory changes (see Table 2 in Appendix A). The key disambiguation concerns `comp:obj` and `comp:obl`: when the head is an adposition (UPOS = ADP) or complementizer (UPOS = SCONJ), the dependency is classified as *functional*; when the head is a verb, it is classified as *lexical*. Of 340k `comp:*` tokens across all SUD treebanks, 62.5% are classified functional and 37.5% lexical. Relations typically analyzed as non-dependency structure (e.g., `punct`, `root`) and underspecified relations (`dep`, `orphan`) are excluded.

3.3. Metrics

For each group (overall, functional, lexical), we compute:

1. **MDD:** Mean absolute distance $|pos(head) - pos(dep)|$ over non-punctuation tokens, excluding root dependencies ([Liu et al., 2017](#)).
2. **Random baseline:** To estimate the expected distance under no DLM pressure, we randomly permute the *linear positions* of all non-punctuation tokens in a sentence while keeping the dependency tree structure (i.e., who depends on whom) fixed. Each permutation reassigns every token to a new position, so the same tree is linearized in a different random order; dependency distances are then recomputed on this shuffled linearization. Following [Futrell et al. \(2015\)](#), we generate 20 such random permutations per sentence and average the resulting MDD across all permutations (MDD_{rand}). This serves as a null hypothesis: the expected distance if word order carried no DLM signal.
3. **Optimality ratio (OR):** MDD_{obs}/MDD_{rand} , following the formalization of [Ferrer-i Cancho et al. \(2022\)](#). This provides a normalized measure of optimization: an OR of 1.0 suggests a language is no more optimized than chance, while values approaching 0.0 indicate extreme minimization.
4. **Head Directionality:** The proportion of dependencies where the head follows the dependent ($pos(head) > pos(dependent)$).

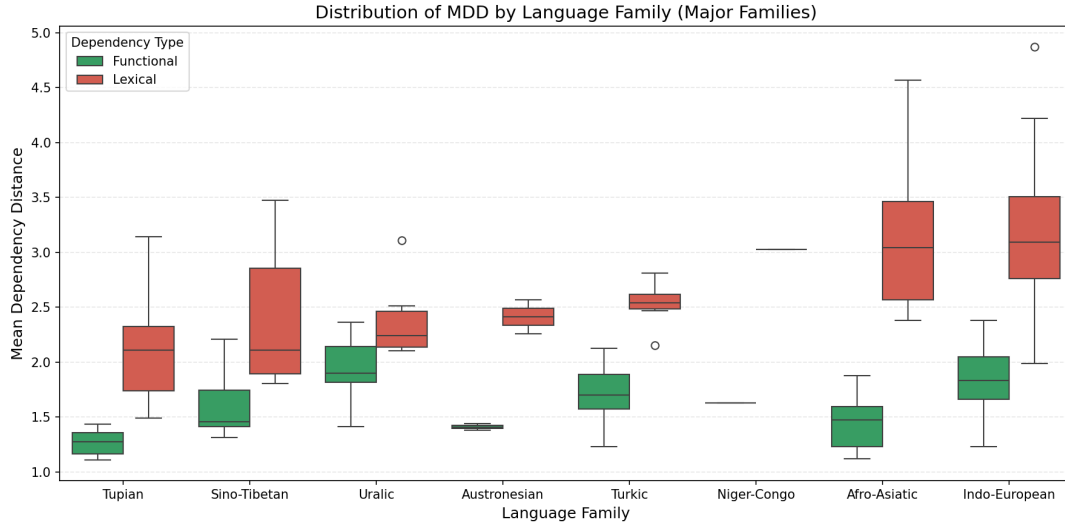


Figure 1: Distribution of Functional vs. Lexical MDD across major language families. Functional MDD is consistently low across diverse families, whereas Lexical MDD varies significantly with word order typology (e.g., higher in head-final Turkic/Uralic/Dravidian, lower in head-initial Austronesian/Niger-Congo).

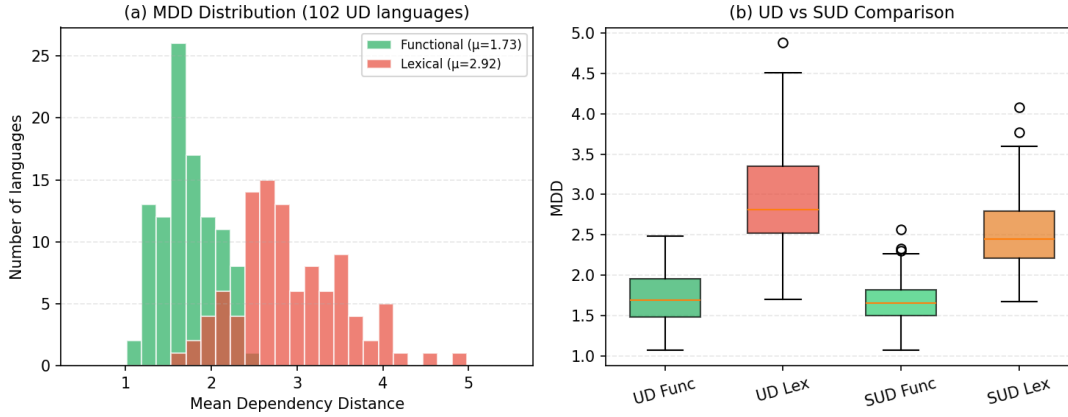


Figure 2: (a) Distribution of functional (green) and lexical (red) MDD across 122 UD languages. Functional MDD clusters tightly around 1.71; lexical MDD is higher and more dispersed. (b) Boxplot comparison across UD and SUD frameworks: the functional–lexical gap is preserved regardless of annotation convention.

4. Results

4.1. Overall DLM Confirmation

All 122 UD languages exhibit strong DLM. Observed MDD ranges from 1.44 to 3.67, far below random baselines (optimality ratios 0.17–0.89, mean 0.41). This confirms [Futrell et al. \(2015\)](#) at scale and extends the finding to new languages.

4.2. The Functional–Lexical Asymmetry

Table 1 presents the central result, showing both absolute MDD and the optimality ratios derived from

analysis reclassifying it as functional actually strengthens the functional–lexical distinction, confirming the result’s robustness (see §4.6).

Framework	Functional MDD		Lexical MDD	
	MDD	OR	MDD	OR
UD (122)	1.71 ± 0.33	0.28	2.87 ± 0.63	0.46
SUD (122)	1.65 ± 0.32	0.27	2.48 ± 0.47	0.41

Table 1: Mean ($\pm$ std) functional and lexical MDD across 122 qualifying languages per framework. Functional MDD is universally lower and less variable.

the 20-permutation random baselines. Three patterns emerge:

1. Functional MDD is universally low. Across 122 UD languages, functional MDD averages 1.71 with a standard deviation of only 0.33. This narrow distribution (Figure 2) shows that grammars universally constrain function words to appear adjacent to their hosts, regardless of language family

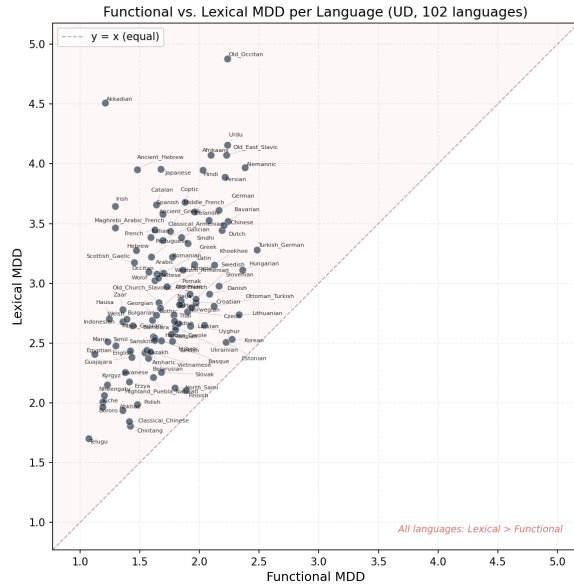


Figure 3: (a) Functional vs. lexical MDD per language (122 UD languages). Every language is above the diagonal ($y = x$).

or word-order type.

2. Lexical dependencies also show strong DLM. Crucially, lexical dependencies are not merely “less optimized than functional”: with a mean optimality ratio of 0.46 ($\sigma = 0.15$), they are 54% shorter than random baselines in every single language (122/122, OR range 0.20–0.93). This confirms that genuine processing-driven minimization operates on lexical dependencies — subjects, objects, and modifiers are placed substantially closer to their heads than chance would predict. However, lexical MDD is more variable ($\sigma = 0.63$ vs. 0.33), and SOV and V2 languages show higher values, consistent with Dyer (2023). Figure 3 confirms that lexical MDD universally exceeds functional MDD.

3. The two levels differ in optimization depth. The mean functional optimality ratio is 0.28, versus 0.46 for lexical — both well below chance, but functional OR is 39% lower. This gap reflects different optimization mechanisms: functional adjacency is categorically enforced by grammar, while lexical ordering is a softer, gradient optimization that competes with information structure, heaviness, and other communicative pressures. The boxplot in Figure 2b confirms the pattern holds across both UD and SUD.

Statistical confirmation. We verify the functional–lexical gap with three tests, each addressing a different concern. *Is the gap consistent across languages?* A paired Wilcoxon signed-rank test compares functional and lexical MDD within each language and asks whether one is systematically lower. The result ($W = 0, p < 10^{-18}$) confirms

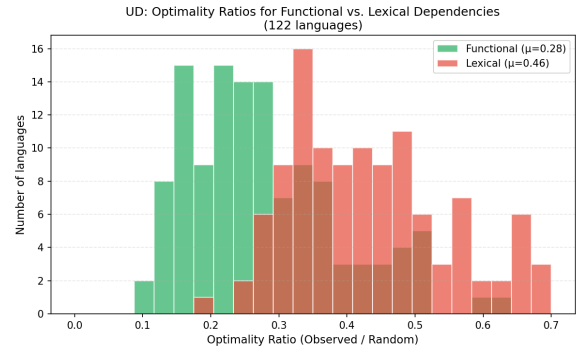


Figure 4: (b) Optimality ratio distributions for functional and lexical dependencies across 122 UD languages. Lower = more optimized.

that functional MDD is lower in every single language, not just on average. The effect is large: Cohen’s $d = 2.30$, meaning the gap exceeds two standard deviations. *Could the gap be driven by shared genealogical history?* Related languages may share similar MDD patterns, inflating apparent universality. A linear mixed-effects model with Language Family ($N \geq 25$) as a random intercept controls for this: the functional–lexical distinction remains highly significant ($p < 0.001$), confirming the gap holds within families, not just across them. All results replicate in SUD ($W = 0, p < 10^{-18}$, $d = 2.04$).³ Figure 1 visualizes this stability: while Lexical MDD (red) varies between head-initial and head-final families, Functional MDD (green) remains low.

4.3. Interaction with Head Directionality

Figure 5 plots functional and lexical MDD against head-final proportion. Functional MDD forms a flat band around 1.71, showing negligible correlation with head directionality. Lexical MDD shows substantially more spread, with SOV languages displaying higher values. This parallels the findings of Liu (2010) on dependency direction as a typological parameter, but reveals that the typological signal resides exclusively in lexical dependencies.

³A Pitman-Morgan test further confirms that functional MDD is not merely lower but also less *variable* across languages than lexical MDD ($t(120) = 8.84, p < 0.001$; SUD: $t = 4.96$), consistent with the claim that grammar enforces a narrow range of functional distances while lexical distances vary with typology.



Figure 5: Functional vs. lexical MDD in UD compared against head directionality measured in SUD. By using SUD’s head-final proportion (x-axis), we capture syntactic word order (e.g., Japanese at 0.91) without the distortions caused by UD’s annotation of function words.

4.4. UD vs. SUD: Robustness Across Frameworks

Figure 6 compares functional and lexical MDD between the frameworks. The correlations are $r = 0.92$ for functional MDD and $r = 0.92$ for lexical MDD. SUD lexical MDD is systematically lower (2.48 vs. 2.87). This reduction is primarily structural: in UD, oblique arguments are attached to the verb via the noun (Verb $\rightarrow$ Noun), spanning the adposition. In SUD, they are attached via the adposition (Verb $\rightarrow$ Adposition), which is typically closer to the verb than the noun is. Critically, the asymmetry between functional and lexical is preserved.

At the global level, nearly all languages fall below the diagonal (Figure 6a), confirming that SUD’s head-direction choices lower global dependency distance relative to UD, as predicted by Osborne and Gerdes (2019). Figure 6b shows per-relation

MDD: functional relations (green squares) cluster in the bottom-left, indicating they are short in both frameworks.

4.5. Per-Relation Detail

Figure 7 provides a per-relation breakdown for the 20 largest UD languages, covering 16 dependency types. Within the functional group, *det* (~ 1.0 – 1.5) and *case* (~ 1.0 – 1.8) are universally short. Within the lexical group, *nsubj* shows significant variation (1.8 in Finnish to 6+ in Hindi), reflecting SOV vs. SVO order. Clausal complements (*ccomp*, *advcl*) consistently show high MDD. This extends the 9-type finding of Liu et al. (2022): the types responsible for DLM are precisely functional, while those showing variability are lexical.

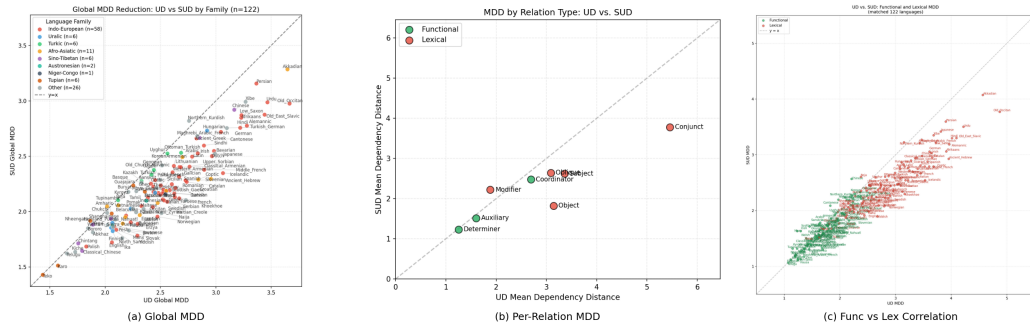


Figure 6: UD vs. SUD comparison across 122 languages. (a) Global MDD: most languages fall below the diagonal, showing SUD lowers MDD. (b) Per-relation MDD: functional relations (green) cluster short in both frameworks. (c) Per-language functional (green) and lexical (red) MDD correlate highly across frameworks ($r > 0.92$).

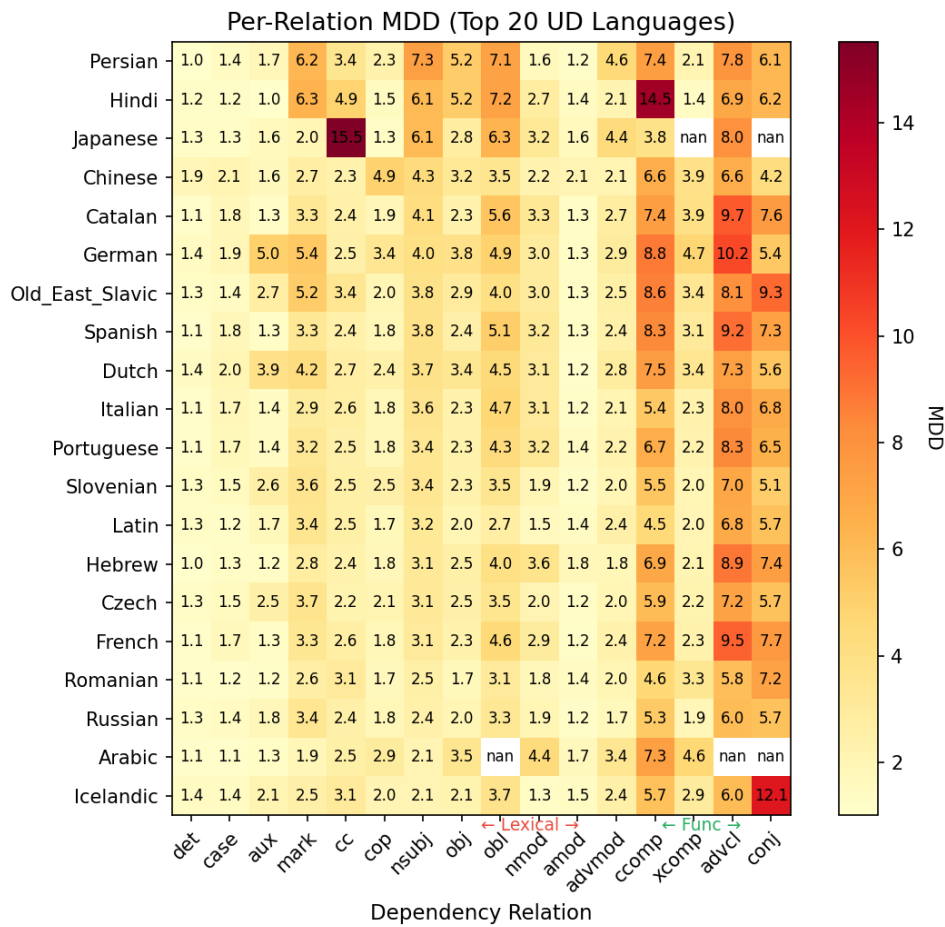


Figure 7: Per-relation MDD across the 20 largest UD languages. Functional relations (det, case, aux, mark) are universally short; lexical relations (nsubj, obl, ccomp, advcl) vary with word-order typology and show higher distances.

4.6. Sensitivity Analysis

To test robustness, we recomputed the functional–lexical split under three alternative relation groupings and two alternative distance metrics (details in Appendix A). Excluding `conj` (whose high MDD is

an artifact of UD’s chain analysis), restricting to core syntactic relations only, or applying the strictest possible classification all preserve the asymmetry: $\text{func} < \text{lex}$ holds in at least 121/122 languages under every scenario, and the effect size remains

large ($d \geq 1.65$). Additionally, excluding pronoun dependents from the lexical class *widens* the gap, and computing distances in nucleus positions only (counting content words only) preserves it. A bootstrap analysis confirms the gap is detectable from as few as 100 sentences. We conclude that the finding is robust to reasonable variation in both relation classification and distance metric.

5. Discussion

Our results disentangle two components of the universal DLM signal reported by [Temperley and Gildea \(2018\)](#) and [Futrell et al. \(2015\)](#). **Grammar-driven functional optimization:** functional elements are positioned adjacent to their hosts by grammatical rules ([Temperley, 2008](#)); the universally low functional MDD (1.71 ± 0.33) is a predictable consequence of grammar design, not speaker choice. **Processing-driven lexical optimization:** this is the more revealing finding. Lexical dependencies — where speakers have genuine ordering flexibility — nonetheless show strong minimization in every language (OR = 0.46, i.e., 54% shorter than random). This confirms that DLM is not merely an artifact of grammatical adjacency: even when function words are factored out, speakers consistently place arguments and modifiers closer to their heads than chance would predict. The degree of lexical minimization varies with typological constraints (SOV languages show weaker optimization; [Dyer, 2023](#)), suggesting that word-order typology modulates but does not eliminate processing-driven DLM.

This two-level view refines Gibson’s Dependency Locality Theory ([Gibson, 1998](#)): integration cost is concentrated at lexical heads, where ordering is flexible and information density effects operate ([Collins, 2014](#)). For typology, functional MDD serves as a stable baseline ($\sigma = 0.33$) for cross-linguistic comparison, while lexical MDD captures genuine typological variation correlating with word-order parameters ([Liu, 2010](#); [Futrell et al., 2020](#)).

Coordination and discourse relations. Excluding *conj* from the lexical class does not alter the core finding (§4.6). The high MDD of *conj* (5.45) is largely an artifact of UD’s chain analysis; similarly, *parataxis* (6.63) and *discourse* (3.12) reflect discourse structure rather than syntactic placement. Excluding these (Scenario C) yields a cleaner measure while preserving the functional–lexical asymmetry.

Functional dependencies and the flux. Our results can also be viewed through the lens of dependency flux ([Kahane et al., 2017](#)). Since functional arcs are almost always short, they contribute minimally to inter-word flux: only 22.6% of functional arcs span over any lexical arc. *det* and *case* rarely

cover a lexical arc (14%), while *mark* does so 49% of the time — reflecting the high *mark* MDD in SOV languages (Urdu 7.67, Hindi 6.34, German 5.38). In the flux perspective, functional dependencies contribute mostly to bouquet-local flux, while lexical dependencies drive inter-word flux complexity.

Limitations and Future Work. Treebank sizes and genres vary, and specific properties of individual UD treebanks (e.g., genre composition, annotation consistency, and domain) may influence observed dependency length patterns; our aggregation across treebanks per language mitigates but does not eliminate such effects. Tokenization affects absolute MDD values ([Lei and Jockers, 2020](#)). The functional/lexical boundary is theory-dependent; *advmod* is classified as lexical but could be argued either way. Future work should incorporate genre-stratified analysis, natively annotated SUD treebanks, and a full flux decomposition ([Kahane et al., 2017](#)) separating functional and lexical contributions to inter-word flux complexity.

6. Conclusion

Across **122 languages**, we find that dependency length minimization is not a monolithic phenomenon but a composite of two distinct forces. **The grammar does the work** — but processing does too. Functional dependencies are *grammatically minimized*: by mandating local attachment for functional items (*det*, *case*, *aux*), the grammar guarantees a baseline of low aggregate MDD (≈ 1.71), effectively scaffolding sentences with short dependencies. **Lexical DLM is real and substantial.** When we isolate lexical dependencies — where speakers have genuine ordering choices — they are still 54% shorter than random baselines across all 122 languages (OR = 0.46). This is the more informative finding: it demonstrates that online processing pressures actively shape word order beyond what grammar dictates. The variation in lexical MDD (≈ 2.87 , $\sigma = 0.63$) tracks typological parameters (SOV structures incur longer distances), revealing the interplay between typological constraints and processing optimization. This two-level view refines the efficiency-grammar hypothesis. Grammar *crystallizes* minimization for the most frequent, predictable elements (function words). But the residual lexical signal shows that processing optimization operates independently and universally, even in typologically constrained languages. Future work should investigate the interaction between lexical DLM and information-structural preferences, and whether the degree of lexical optimization correlates with psycholinguistic measures of processing difficulty.

A. Supplementary Material

A.1. UD and SUD Dependency Example

Figure 8 illustrates the functional–lexical distinction in both UD and SUD on a single sentence.

A.2. UD to SUD Relation Mapping

UD rel.	SUD rel.	Class	Criterion
det	det	Func	same
cc	cc	Func	same
clf	clf	Func	same
aux	comp:aux	Func	label prefix
cop	comp:pred	Func	label prefix
case	comp:obj	Func	head = ADP
mark	comp:obl	Func	head = SCONJ
nsubj	subj	Lex	base label
obj	comp:obj	Lex	head = VERB
obl	comp:obl	Lex	head = VERB
nmod	udep	Lex	base label
amod	mod	Lex	base label
advmod	mod	Lex	base label
conj [†]	conj:*	Lex	base label

Table 2: UD → SUD relation mapping and functional/lexical classification. SUD `comp:obj/obl` is disambiguated via the head’s UPOS tag. [†]Excluded under Scenarios B–D in the sensitivity analysis.

A.3. Detailed Sensitivity Analysis

Our main analysis classifies 7 base relation types as functional and 23 as lexical. Several lexical relations are arguably not prototypical syntactic dependencies: `conj` (mean MDD 5.45) is inflated by UD’s chain analysis; `parataxis` (6.63) and `discourse` (3.12) are discourse-level; `flat` (1.41), `fixed` (1.15), and `compound` (1.23) are MWE-internal.

We tested three alternative groupings:

- **Scenario B (–conj):** Excluding `conj`. Lexical MDD drops to 2.66 (± 0.56); `func` < `lex` in 122/122 languages ($d = 1.89$).
- **Scenario C (core syntax):** Restricting lexical to core arguments and modifiers (`nsubj`, `obj`, `iobj`, `obl`, `nmod`, `amod`, `advmod`, `advcl`, `acl`, `xcomp`, `ccomp`, `csbj`, `nummod`). Lexical MDD = 2.65 (± 0.66); `func` < `lex` in 121/122 ($d = 1.65$).
- **Scenario D (strictest):** Removing `mark` and `cc` from functional; `advmod` and `compound` from lexical. `Func` MDD = 1.40 (± 0.25), `lex` MDD = 2.75 (± 0.69); gap *widens* ($d = 2.08$).

Pronoun dependents. Pronoun-headed lexical arcs are shorter (MDD = 2.34 ± 0.97) than non-pronoun (2.93 ± 0.72), in 81% of treebanks. However, pronouns are only 12% of lexical tokens; excluding them *raises* lexical MDD, widening the gap.

Nucleus-based distance. Computing distances counting only content-word positions reduces lexical MDD from 2.87 to 2.12, but the asymmetry is preserved (2.12 still exceeds functional MDD 1.71; $r = 0.78$ with standard measure).

Sample-size stability. Bootstrap resampling on 8 diverse languages shows the gap is detectable at 100 sentences (8/8 languages) and even at 50 sentences (7/8). All languages with ≥ 200 sentences individually confirm `lex` > `func`.

Ethics Statement and AI Disclosure

We fully disclose that this paper was produced using an AI system. The research idea, all analysis code, and the entire text were generated by Claude Opus 4.6 (Anthropic), operating in agentic mode via GitHub Copilot Chat in VS Code. The human author provided prompts and accepted or rejected proposed work at each step, but did not originate the central research question, write any code, or draft any prose beyond the prompts themselves. The author’s contribution was verification and editorial oversight rather than generation of the core ideas, code, or prose.

The project was inspired by an experiment by D. Yanagizawa-Drott, who prompted an LLM to write a macroeconomics job market paper and publicly reflected on the implications (<https://x.com/YanagizawaD/status/2022034189395407093>). This paper began with an analogous prompt: “*I always dreamt of becoming a syntactician and a typologist one day. I want to submit a paper to the UDW workshop. Can you help me?*” The functional–lexical distinction that became the core contribution emerged during the AI’s iterative exploration of the data, not from the human author. Prior work was identified using Google Scholar Labs and provided to the system for integration.

The reviews by anonymous UDW 2026 reviewers and by Sylvain Kahane were essential for improving the final version; reviewer feedback was given to the AI, which implemented the revisions. The code, analyses, and results were checked by the author, and the final paper benefited substantially from this review process.

We believe the scientific community deserves full transparency about how research is produced. As Yanagizawa-Drott noted, in a world of machine-speed generated papers we face a dilemma: either this is real research that requires expert verification,

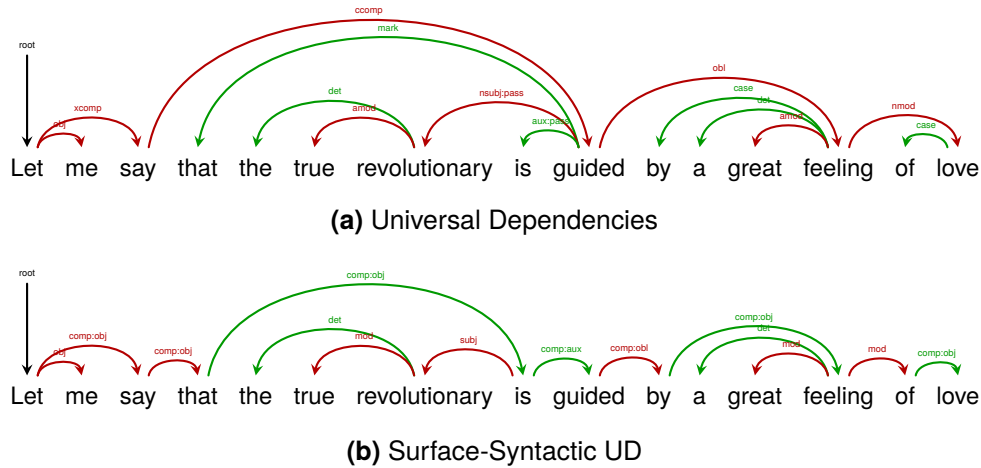


Figure 8: Dependency analysis of *Let me say that the true revolutionary is guided by a great feeling of love* (Guevara, 1965). **Green arcs** = functional; **red arcs** = lexical. In (a), functional elements depend on content words. In (b), functional elements (auxiliaries, adpositions, complementizers) are heads.

or it is not, and we have polluted the information environment. We hope this disclosure contributes to an honest conversation about the role of AI in scientific work. This work was produced with substantial AI assistance, and readers should nonetheless be aware of this production process when evaluating the claims.

An open question is whether AI-generated hypotheses and text, curated and verified by a human, should count as a scientific result, and whether this may become a standard category of research output. In this case, both the author and the reviewers judged the findings scientifically interesting enough to merit revision and discussion. More broadly, this process raises an old question about novelty: can LLMs create genuinely new ideas, can humans, or are both primarily recombining from a finite space of possibilities, as in Borges' "Library of Babel"? We do not claim to resolve this question here, but we consider it central for future norms of authorship, credit, and evaluation.

B. Bibliographical References

- Michael Xavier Collins. 2014. Information density and dependency length as complementary cognitive models. *Journal of Psycholinguistic Research*, 43(5):651–681.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.
- Andrew Thomas Dyer. 2023. Revisiting dependency length and intervener complexity minimization on a parallel corpus in 35 languages. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*.
- Ramon Ferrer i Cancho. 2004. Euclidean distance between syntactically linked words. *Physical Review E*, 70(5):056135.
- Ramon Ferrer-i Cancho, Carlos Gómez-Rodríguez, and Juan Luis Esteban. 2022. Optimality of syntactic dependency distances. *Physical Review E*, 105:014308.
- Richard Futrell, Roger P. Levy, and Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language*, 96(2):371–412.
- Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. In *Proceedings of the National Academy of Sciences*, volume 112, pages 10336–10341.
- Ning Gao and Qingshun He. 2024. A dependency distance approach to the syntactic complexity variation in the connected speech of Alzheimer's disease. *Humanities and Social Sciences Communications*, 11(1):1–12.
- 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. 2021. Starting a new treebank?

- Go SUD! In *Proceedings of the 6th International Conference on Dependency Linguistics (Depling)*.
- Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.
- Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34:286–310.
- Richard Hudson. 1995. Measuring syntactic difficulty. Manuscript, University College London. <http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf>.
- Sylvain Kahane, Chunxiao Yan, and Marie-Amélie Botalla. 2017. What are the limitations on the flux of syntactic dependencies? Evidence from UD treebanks. In *Proceedings of the 4th International Conference on Dependency Linguistics (Depling)*, pages 73–82.
- Marie-Pauline Krielke. 2024. Cross-linguistic dependency length minimization in scientific language: Syntactic complexity reduction in English and German in the late modern period. *Languages in Contrast*, 24(2).
- Lei Lei and Matthew L. Jockers. 2020. Normalized dependency distance: Proposing a new measure. *Journal of Quantitative Linguistics*, 27(1):62–79.
- Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.
- Haitao Liu. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578.
- Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193.
- Xiaoyu Liu, Haoran Zhu, and Lei Lei. 2022. Dependency distance minimization: A diachronic exploration of the effects of sentence length and dependency types. *Humanities and Social Sciences Communications*, 9(1):1–12.
- Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. SUNY Press, Albany, NY.
- Rui Niu and Haitao Liu. 2025. Factors influencing dependency distance: An account of the MDD variation between Chinese and English. In *Word Grammar, Cognition and Dependency*, pages 276–296. Cambridge University Press.
- Joakim Nivre. 2016. Universal dependencies: A cross-linguistic perspective on grammar and lexicon. In *Proceedings of the Workshop on Grammar and Lexicon: Interactions and Interfaces*.
- Timothy Osborne and Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of universal dependencies (UD). *Glossa: A Journal of General Linguistics*, 4(1):1–28.
- Wojciech Stempniak. 2024. Dependency structure of coordination in head-final languages: A dependency-length-minimization-based study. In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories*.
- David Temperley. 2007. Minimization of dependency length in written English. *Cognition*, 105(2):300–333.
- David Temperley. 2008. Dependency-length minimization in natural and artificial languages. *Journal of Quantitative Linguistics*, 15(3):256–282.
- David Temperley and Daniel Gildea. 2018. Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, 4:67–80.
- Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Klincksieck, Paris.
- Daniel Zeman et al. 2025. Universal dependencies 2.17. LINDAT/CLARIAH-CZ, Charles University.

Greenberg's Universal 45 in Universal Dependencies: Gender Distinctions and Annotation Challenges

Antoni Brosa-Rodríguez & M. Dolores Jiménez-López

Universitat Rovira i Virgili

Avda. Catalunya 35, 43002 - Tarragona (Spain)

antoni.brosa@urv.cat, mariadolores.jimenez@urv.cat

Abstract

This paper revisits and extends Greenberg's Universal 45 on gender distinctions using Universal Dependencies 2.17, comprising 339 treebanks across 186 languages. A systematic analysis of morphosyntactic patterns confirms the implicational hierarchy (singular > plural gender marking), with 98.6 % conformity in pronominal categories. Only two potential exceptions are detected, both with minimal occurrences and likely attributable to annotation errors. Extending the analysis beyond pronouns to 13 UPOS categories shows that core categories maintain near-perfect compliance, while peripheral categories exhibit higher violation rates, primarily driven by annotation inconsistencies rather than genuine linguistic exceptions. A total of 90 treebanks display gender-number features in traditionally invariable categories (e.g., adpositions, conjunctions, adverbs), indicating annotation issues such as prepositional contraction handling, homophone merging, and erroneous feature assignment. The study establishes a replicable computational methodology for large-scale typological validation, highlighting both the potential of corpus-based approaches and key limitations, including genealogical sampling biases, annotation heterogeneity despite universal schemas, and the false sense of comparability across treebanks.

Keywords: Greenberg's universals, gender-number marking, corpus annotation errors, typological validation

1. Introduction

In this study, we propose an effective method for carrying out a specific typological check using Universal Dependencies, linked to Greenberg's universal 45: "If there are any gender distinctions in the plural of the pronoun, there are some gender distinctions in the singular also" (Greenberg, 1963). Universal 45 has been extensively studied in traditional typological studies, with different considerations based on grammars. Its interest lies in its great predictive power. However, the full potential of Universal Dependencies (Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers and Daniel Zeman, 2020) for the study of gender distinctions in plural and singular pronouns has not yet been fully explored in the typological literature. That is why we conducted an exploratory study of the behavior of languages in this regard using version 2.17, which contains 339 treebanks corresponding to 186 different languages.

Beyond revisiting classic typological claims, we want to delve into a systematic method that allows us to provide new universals that offer detailed typological information on the behavior of the world's languages. Therefore, specifically, our objectives are: To review Greenberg's universal 45 based on occurrences in real texts in a broad sample of languages; to analyze the behavior of the languages available in UD in relation to gender distinction in different numbers of different parts of speech not

included in Greenberg's claim; and to detect which UD corpora have annotations that generate exceptions to this universal so that they can be analyzed.

Apart from confirming a well-established typological generalization, the use of Universal Dependencies for this purpose serves two additional goals that justify the methodological choice. First, corpus-based verification of morphosyntactic universals such as Universal 45 allows the detection of annotation inconsistencies at scale. Second, while consulting descriptive grammars remains a valid strategy for typological research, especially in cases involving universals on simple features like this one, it is methodologically costly and can be captured through explicit morphosyntactic queries. Embedding such verifications within a UD-based pipeline makes it possible to integrate them with other corpus-derived typological data efficiently and systematically. In this sense, Universal 45 is not only an end in itself, but also a proof of concept for a replicable methodology applicable to a wide range of implicational universals.

2. State of the Art

Greenberg (1963), with a relatively small sample of languages and data from grammars (but with a precise and systematic methodology), formulated 45 universals that characterize the behavior of the world's languages (or specific groups of them). In the case of the last universal, Greenberg showed

how, in pronouns, gender distinctions in the plural only occur if they also exist in the singular (Corbett, 1991, 2000).

We found multiple studies subsequent to this one that attempt to revisit this linguistic behavior with different samples and other assessments. Plank and Schellinger (1997) focus on expanding the study to include data from the grammars of many more languages in an attempt to find exceptions to this universal. Above all, contrasts were found between human and non-human in the plural (and not in the singular) in some (very few) languages, which could be equated with Greenberg’s claim, although he did not specify this distinction. These types of relationships are particularly conditioned by person in the case of personal pronouns (Siewierska, 2013). Berg (2024), in fact, shows how the universal of 45 is particularly robust in relation to the third person, but not in the first person. In these cases, for reasons of informativeness, he argued (again based on grammars, in a very large sample) that it is more common to find gender distinctions in third-person singular than in plural, but also more in first-person plural than in singular. However, this approach would not constitute an amendment to Greenberg (since his claim would still hold true), but simply provide greater precision. Plank (2025) criticized and amended this paper, arguing that what is presented as new is something already anticipated by Greenberg. Therefore, in summary, despite the addition of some clarifications and improvements to the samplings, Greenberg’s proposal remains unchanged.

In studies of token-based typology (Levshina, 2019), there are different approaches to studies based on Greenberg’s word order universals (Liu, 2010; Guzmán-Naranjo and Becker, 2018; Gerdes et al., 2021; Brosa-Rodríguez and Jiménez-López, 2023; Brosa-Rodríguez and Kahane, 2024). However, universals such as 45 have been less explored, despite the apparent ease of formalization and query. Although the availability of languages in UD is not yet comparable to that available through grammars, the methodological advantages it offers, both in terms of performing classic typological checks and illuminating new correlations, make it an attractive and efficient perspective that has not been fully examined.

3. Methodology

To conduct the study, we adopt a structured computational approach that explicitly formalizes the analytical steps. The methodology is designed to ensure transparency, consistency, and full replicability.

3.1. Sample

Version 2.17 of Universal Dependencies contains 339 treebanks corresponding to 186 languages. Although this number is interesting, it is still not sufficient to carry out sampling other than convenience sampling (Miestamo et al., 2016). However, all macro-areas are represented (with a low number in the case of Australia, Papuan, and North America). Compared to WALS (Dryer and Haspelmath, 2013), there is a fairly even proportion of dominant word order types, despite a greater weighting of the SVO type (approximately +10%). The vast majority of languages are from Eurasia, and the Indo-European family is overrepresented compared to others. Therefore, given the exploratory nature of our study, we consider the language sample to be a good starting point that will reveal relevant information about the languages analyzed.

The types of UD (unified annotation) corpora are varied. Therefore, the corpora include a variety of oral and written texts (although the latter are in the majority), with very diverse origins: wikis, news, reviews, fiction, non-fiction, conversations, etc. This richness will allow us to reduce possible textual typology biases (when there are multiple corpora for a language) and to better represent different communicative situations that represent that language. The lack of restriction by corpus type also allows us to obtain the maximum number of languages available, which is especially relevant in the case of minority and minoritized languages, with much more limited availability (as is also the case with these same languages in the case of the grammar-based approach). Furthermore, we consider that the linguistic element we analyze is much more resistant to textual biases than other linguistic elements such as word order.

Another element of disparity is the size of the corpora. We can find corpora with less than 1K tokens (from languages with few resources) and other corpora with 3500K tokens (for example, in German). This disparity is taken into account and, therefore, neutralized with a simple binary calculation of compliance or non-compliance with the universal or alternative checks. However, exact quantitative results are also presented, allowing occurrences to be viewed. This type of data allows us to check the representativeness of the occurrences obtained, which we consider to be much better data than simply filtering by corpus size. The reason is that, given the linguistic structure sought, a small corpus can offer sufficient and reliable occurrences. Furthermore, we chose not to exclude languages with representativeness issues from the analysis, but rather to mark them, as we believe it is interesting to look at the insignificant occurrences obtained by experts in those languages and validate the possible annotation error detected (or

confirm the exception discovered).

3.2. Formalization

While Grew-Match (Guillaume, 2021) was not employed to run the queries, its formal query language, combined with UD annotation conventions, was adopted to develop a preliminary formal specification. This specification ensured the precise identification of the structures under investigation. In this case, in order to convert Greenberg's universal from natural language to formal language, we chose to present formalizations 1 and 2:

```
%45-PRONpl
pattern {PRON[upos=PRON, Gender, Number=Plur]}
(1)
```

```
%45-PRONsg
pattern {PRON[upos=PRON, Gender, Number=Sing]}
(2)
```

This step is considered important in order to clearly identify the conflict points in the query. If we limit ourselves strictly to Greenberg's universal, he speaks, on the one hand, of "gender distinctions." Therefore, we did not consider it appropriate to formalize with different genders, as the Gender feature encompasses any possible gender value and simply requires that such a distinction be made, something that is activated if any "gender" value exists. On the other hand, Greenberg referred to "in the plural" and "in the singular." Therefore, in this case, we refined the formalizations so that they capture these two orders, respectively, which are explicitly affected.

To extend the universal and analyze the behavior of other parts of speech, we take advantage of the formalization made and apply it to different "upos." For example, in the case of adjectives, this would be included in formalizations 3 and 4:

```
%45-ADJpl
pattern {PRON[upos=ADJ, Gender, Number=Plur]}
(3)
```

```
%45-ADJsg
pattern {PRON[upos=ADJ, Gender, Number=Sing]}
(4)
```

We repeated this type of formalization with other parts of speech until we reached 26 different formalizations. These included the formalization of gender distinction in the singular and plural forms of parts of speech that are considered invariable in their most canonical definitions (e.g., adpositions or adverbs). The parts of speech analyzed are: adjective (ADJ), adposition (ADP), adverb (ADV), auxiliary (AUX), coordinating conjunction (CCONJ), determiner (DET), interjection (INTJ), noun (NOUN), numeral (NUM), pronoun (PRON),

proper name (PROPN), subordinating conjunction (SCONJ), verb (VERB).

The inclusion of traditionally invariable categories (ADP, ADV, CCONJ, INTJ, SCONJ) serves as an explicit annotation quality control measure. Since these categories are not expected to bear gender or number features in any language, the working hypothesis is that all treebanks should yield zero occurrences for these patterns. Any deviation from this baseline is therefore treated as a potential annotation error rather than a genuine typological finding, and is analyzed accordingly.

The inclusion of nouns (NOUN) in the analysis deserves a brief justification. Unlike adjectives and pronouns, gender in nouns is typically considered a lexical rather than an inflectional feature, and in most languages it is not overtly marked on the noun itself. However, we include this category because languages may exhibit overt gender alternations on animate nouns (e.g. morphological masculine/feminine pairs), which some authors treat as genuine nominal gender inflection (Spencer, 2002). It allows us to examine whether this subset of cases follows the same "singular > plural" implicational hierarchy, while acknowledging that the results must be interpreted with caution given the fundamentally different nature of nominal gender.

3.3. Tools & Outputs

JupyterNotebook with Python language was used to carry out the analysis. To ensure the traceability and replicability of the study and to avoid extending the article with programming technicalities, all the steps followed are provided in an open-access repository¹.

In general, the analysis was structured in several interconnected phases: first, the 26 formalizations of universal 45 were loaded from a CSV file, specifying search patterns that combine part-of-speech (UPOS) with grammatical number (singular/plural) and gender. Next, the 339 treebanks from Universal Dependencies v2.17 were systematically processed, extracting each token from the CoNLL-U files along with its UPOS and relevant morphological features (gender and number). For each combination of treebank and UPOS, the presence of gender distinction in plural and singular was verified, identifying as "violations" those cases where a UPOS had gender in plural but not in singular, thus contradicting the implicational hierarchy proposed by Greenberg. Subsequently, the exact occurrences of gendered tokens were quantified to assess the statistical robustness of the violations detected, and finally the data was aggregated by language, merging the different treebanks of the same language to obtain a more general typological

¹<https://doi.org/10.34810/data3130>

perspective that would facilitate the interpretation of the patterns observed. Once this search has been carried out, various outputs that may be of interest to the research are requested, which are available in the repository previously mentioned in a detailed manner.

4. Results & Discussion

The results of the analysis provide empirically grounded insights that are pertinent both to typological research and to UD treebank development practices. First, we highlight that of the 339 treebanks, only 2 are exceptions to Greenberg’s original universal, which represents a violation rate of 0.01 in the plural/singular trade-off. The two notable corpora are “Hausa-WesternAutogramm” (Caron Bernard, 2026) and “Maghrebi_Arabic_french-Arabizi” (Riabi Arij, Farah Essaidi, Amal Fethi, Menel Mahamdi, and Djamé Seddah, 2026). In the case of the first treebank, there is a double occurrence of the word “ta:sù” as a third-person plural feminine possessive pronoun. In the case of the second treebank, one occurrence of the word “t” is found as a second-person masculine plural pronoun. The size of the corpora, although not very large, is not small either, with 13K and 19K tokens, respectively. This, correlated with the low occurrence of the highlighted elements, may facilitate the interpretation that we are dealing with a possible annotation error. However, given our total lack of knowledge of the two languages involved, we recommend that experts in those languages or annotators of those corpora certify whether we are dealing with a surprising exception or a significant annotation error.

Be that as it may, whether the examples of these languages are eliminated as errors or the detected exceptions are confirmed, the data provide clear confirmation of the robustness of Greenberg’s proposal for the pronominal category. The results show that of the 139 treebanks in which gender distinction was detected in plural pronouns, 137 also showed this distinction in the singular, thus respecting the implicational universal. This proportion (98.6% conformity) constitutes a compelling empirical validation on a massive and multilingual scale.

In the case of checks on other parts of speech, the findings are equally revealing, as shown in Table 1. The analysis detected a total of 36 violations distributed across 13 different UPOS categories, albeit with very uneven patterns. The aggregated data reveal that the NUM (numerals) category has the highest absolute number of violations, with 11 cases detected in treebanks such as Latin-LLCT (629 plural occurrences), Polish-LFG (833 occurrences), and Breton-KEB (38 occurrences). However, when examining the relative vio-

lation rate—calculated as the number of violations divided by the number of treebanks with plural activation, shown in Figure 1—we observe that SCONJ (subordinating conjunctions) has the highest ratio (0.43), followed by CCONJ (coordinating conjunctions) with 0.40 and ADV (adverbs) with 0.15, as can be seen, all in grammatical categories where we would expect to find 0 occurrences, which invites us to review the data language by language to confirm possible annotation errors.

UPOS	Singular only	Plural only (exception)	Both	None
ADJ	15	0	138	186
ADP	15	1	14	309
ADV	9	3	17	313
AUX	42	1	67	229
CCONJ	4	2	3	330
DET	26	1	125	187
INTJ	4	0	3	332
NOUN	4	5	153	177
NUM	20	11	77	231
PRON	45	2	137	155
PROPN	39	4	99	197
SCONJ	7	3	4	325
VERB	24	3	110	202

Table 1: Number of treebanks (out of 339) showing gender distinction by grammatical number across UPOS categories.

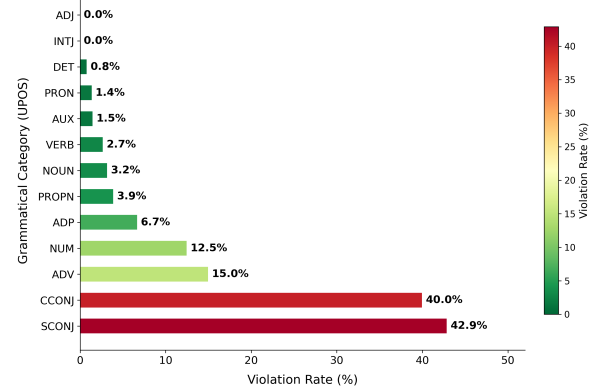


Figure 1: Violation Rate in the singular/plural trade-off by Part of Speech

It is particularly significant that the ADJ category (adjectives), traditionally considered one of the most prone to gender agreement in the world’s languages, shows no violation of the universal. Of the 138 treebanks in which gender distinction was detected in plural adjectives, 100% also showed this distinction in the singular. This behavior contrasts sharply with categories such as NUM, where the violation rate reaches 12.5%. A consistent pattern that emerges from the analysis is that the vast majority of violations detected (31 out of 36) corre-

spond to extremely low occurrences, with frequencies between 1 and 7 tokens, as shown in Figure 2. This data strongly suggests that many of these apparent exceptions could be due to annotation errors. The only significant exceptions from a quantitative point of view are concentrated in specific corpora of Latin, Polish, Hausa, and Breton, where plural occurrences exceed 30 instances.

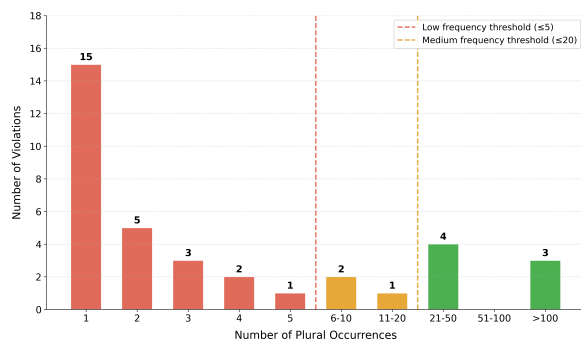


Figure 2: Distribution of Violation Magnitude: Occurrence Frequency

The case of Polish deserves special attention. The 833 occurrences of plural NUM in Polish-LFG without a singular counterpart can be explained by a specific annotation convention: the numeral *jeden* ('one') is systematically labeled as ADJ (adjective), since it can effectively function as such, but the option of labeling it as NUM is never considered, unlike other numerals such as *dwa* ('two') or *cztery* ('four'). Something similar occurs in Latin-LLCT, where *unus* is always considered DET, while *duo* is NUM (something that does not occur in other Latin corpora). This annotation decision creates the apparent mismatch that Polish or Latin lack singular numerals with gender distinction, producing an artifactual violation of the universal that does not reflect linguistic reality but rather a methodological issue in the application of the UPOS guidelines.

Of particular interest is the behavior of five grammatical categories traditionally considered invariable or with little morphological inflection: ADP (adpositions), ADV (adverbs), CCONJ (coordinating conjunctions), INTJ (interjections), and SCONJ (subordinating conjunctions). The presence of gender and number features in these categories in multiple treebanks points to possible systematic annotation problems or disparities in the criteria applied, which hinder the empirical study of linguistic universals. Specifically, 31 treebanks were identified with gender-number occurrences in ADP, 29 in ADV, 9 in CCONJ, 7 in INTJ, and 14 in SCONJ. The full list of affected treebanks for each category is provided in Appendix A.

Qualitative analysis of these cases reveals recurring patterns of annotative inconsistency. In Italian, for example, various corpora label prepositional

contractions such as *della* (de+la) under the ADP category, automatically inheriting the feminine gender and singular number features of the article *la*, while other corpora opt for a much more detailed and less problematic split. This practice generates the artificial appearance of gendered adpositions, when in reality the feature belongs to the joint component of the contraction. In the case of CCONJ in Italian-MarkIT, we observed a probable merge between the part of speech of the coordinating conjunction *e* ('and') and the morphological features of the copulative verb *è* ('is'), possibly due to errors in automatic morphological analysis or in the disambiguation of homophones. In corpora such as Portuguese-Bosque, we found occurrences of number features in ADP that correspond to possible annotation errors, such as the assignment of the singular feature to invariable prepositions such as *sem* ('without').

Therefore, as shown in Figure 2, the vast majority of occurrences are very marginal, possibly induced by a specific annotation error or oversight. These findings suggest that treebanks presenting gender-number features in invariable categories may benefit from a closer review of their annotations, with particular attention to: (a) potential automatic errors in morphological processing, (b) disparities between corpora of the same language applying divergent criteria, (c) alignment with practices in typologically similar languages, and (d) explicit documentation of any genuine phenomena that may justify the assignment of morphological features to traditionally invariable categories, as some of these cases may reflect real linguistic behavior rather than annotation errors.

In summary, despite the specific occurrences mentioned above, which we invite language experts to review in terms of exceptions or errors, the analysis carried out with UD clearly and efficiently allowed us to go beyond the universal 45 proposed by Greenberg and extend its implication/restriction of Singular > Plural in terms of gender distinction beyond the pronoun. This behavior seems to be transversal to all parts of speech that allow inflection, and its use is widely and robustly attested (whether or not there are specific exceptions), allowing for a better understanding of how languages work, as summarized in Figure 3.

5. Conclusions

This study has demonstrated the feasibility and potential of using Universal Dependencies as an empirical infrastructure for the systematic validation of linguistic universals on a large scale. Through the analysis of 339 treebanks corresponding to 186 languages in version 2.17 of UD, we have addressed the three objectives initially set out, with results that

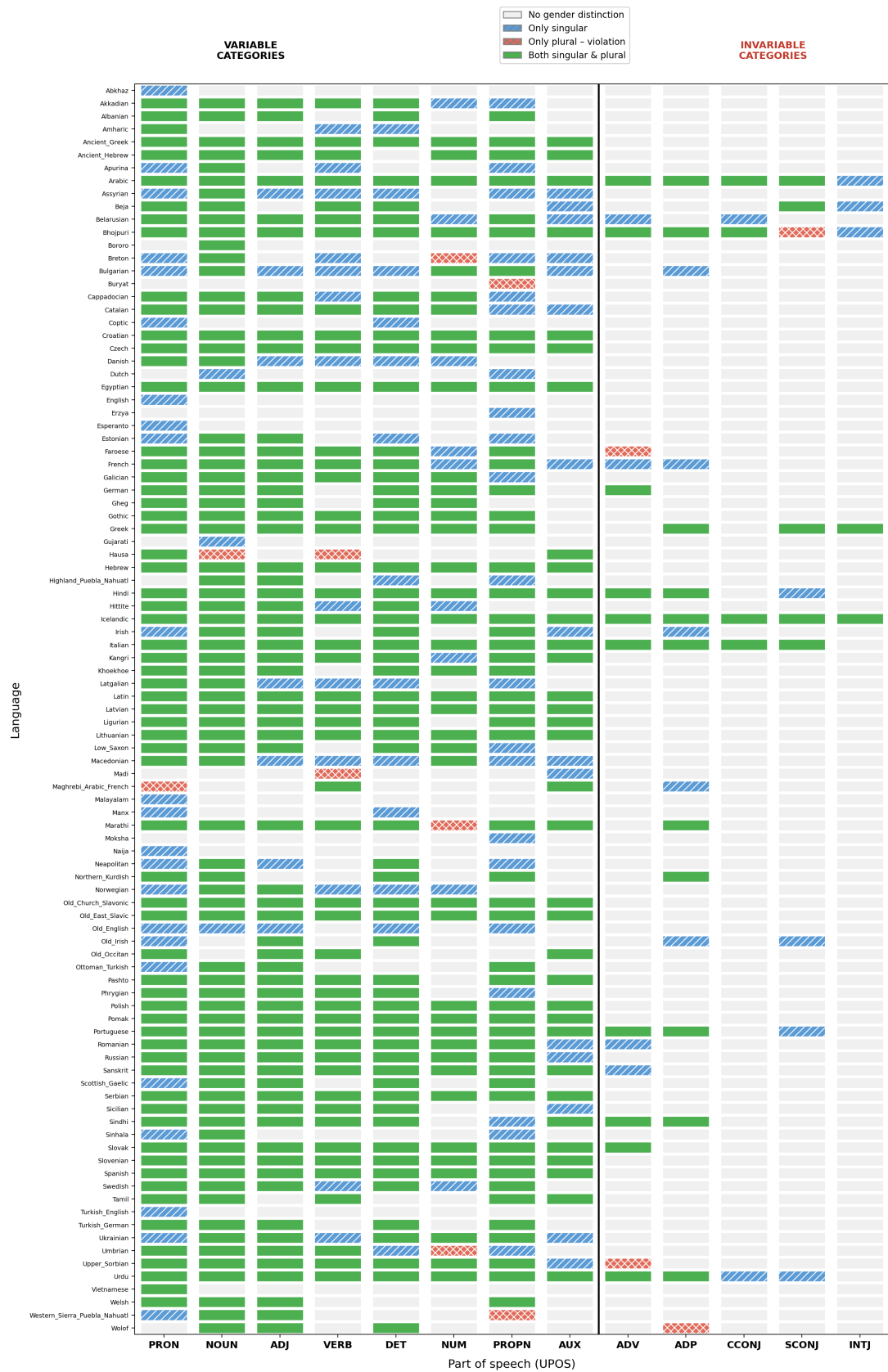


Figure 3: Summary of gender variation in singular and plural forms in different languages and parts of speech (excluding languages that do not exhibit any gender distinction)

combine robust confirmations, unexpected findings, and methodological caveats of relevance to the computational linguistics and typology community.

With regard to the first objective—to review Greenberg’s universal 45 on the basis of actual textual occurrences in a broad sample of languages—the results are categorical: the universal is overwhelmingly confirmed for the pronominal category. Of the 139 treebanks that show gender distinction in plural pronouns, 137 (98.6%) also exhibit this distinction in the singular, respecting the proposed implicational hierarchy. The only two exceptions detected (Hausa-WesternAutogramm and Maghrebi_Arabic_French-Arabizi) are based on extremely low occurrences (1-2 tokens) in moderately sized corpora, strongly suggesting annotation errors rather than genuine linguistic counterexamples, a fact that would make the universal 100% applicable. This massive empirical validation reinforces the predictive power of Greenberg’s universal six decades after its original formulation, demonstrating that generalizations derived from traditional grammatical analysis remain valid when tested against millions of tokens of real text.

The second objective—to analyze the behavior of languages in relation to gender distinction in different grammatical numbers in UPOS categories not originally considered by Greenberg—has yielded nuanced results that invite reflection on the limits of the extension of universals. While core categories such as ADJ (adjectives) and NOUN (nouns) show compliance rates of 100% and 96.8% respectively, categories such as NUM (numerals), SCONJ (subordinating conjunctions), and CCONJ (coordinating conjunctions) show significantly higher violation rates (12.5%, 42.9%, and 40%). However, qualitative analysis reveals that these apparent exceptions are mostly due to annotation peculiarities rather than genuine typological deviations. The examples shown illustrate how annotation conventions can generate artifacts that distort the evaluation of universals.

The third objective—to detect annotated UD corpora that generate exceptions to the universal for subsequent analysis—has resulted in the identification of systematic problems that transcend Greenberg’s specific universal. The presence of gender-number features corresponding to five traditionally invariable categories (ADP, ADV, CCONJ, INTJ, SCONJ) points to recurring annotation inconsistencies: prepositional contractions that inherit article features, merges between conjunctions and homophonic copulative verbs, and erroneous assignments of morphological features to invariable words. These findings underscore the need for systematic quality audits in UD treebanks, especially those generated through automatic processing without exhaustive manual validation.

More broadly, this work establishes a replicable and scalable method for the empirical verification of linguistic universals, combining formalization of morphosyntactic patterns, systematic corpus search, quantification of occurrences, and qualitative analysis of exceptions.

The availability of the Python code in an open repository facilitates its adaptation to other universals related to word order, agreement, case marking, or any phenomenon codable through UD morphosyntactic features.

6. Ethical considerations and limitations

This paper does not raise any specific ethical considerations.

Despite the findings reported in this study, some methodological and empirical limitations should be acknowledged, as they constrain the scope and generalizability of the conclusions, particularly regarding typological coverage and sampling biases, and the discrepancy between the universality of the annotation scheme and the variability of annotation practices.

6.1. Typological coverage & sampling biases

A fundamental limitation lies in the linguistic coverage of Universal Dependencies compared to traditional typological studies.

Although version 2.17 includes 186 languages, it has significant genealogical biases: Indo-European languages are massively overrepresented, while entire families from Austronesia, Papuan, or North America, for example, have minimal representation. This imbalance contrasts unfavorably with samples designed to maximize genealogical diversity, which ensured better typological balance.

The best-represented languages share relevant typological characteristics: two- or three-value gender systems and extensive morphological agreement.

Languages with radically different systems—nominal classifiers, polysynthetic languages, non-grammaticalized semantic gender marking—are underrepresented.

Additionally, treebanks are mostly built on standard languages written in formal registers, excluding oral varieties, dialects, or colloquial registers where exceptions might emerge.

6.2. From Universal Schema to Heterogeneous Annotation

A critical limitation that creates a false sense of methodological security is precisely the homo-

geneity of the Universal Dependencies annotation scheme. Although UD provides clear and universal general guidelines, in practice the annotation process leaves considerable room for interpretation in borderline phenomena.

As documented in the results, multiple treebanks of the same language exhibit radically different patterns in the assignment of morphological features to peripheral categories, not because of genuine linguistic differences but because of divergent methodological decisions by the annotation teams.

This de facto heterogeneity contrasts paradoxically with the impression of uniformity projected by the universal schema.

While in traditional typological studies based on descriptive grammars the heterogeneity of criteria is explicit and recognized, forcing the researcher to carefully homogenize descriptions, in studies based on UD there is a risk of assuming direct comparability where in reality significant interpretative differences persist. The weight of the individual annotator's criteria, although theoretically minimized by UD guidelines, continues to condition the reliability of the derived generalizations, generating both artifactual exceptions and masking genuine typological patterns.

Finally, the impossibility of manually verifying the 36 violations detected through consultation with specialists or native speakers—especially for minority languages—forces us to rely on statistical heuristics to distinguish genuine exceptions from methodological artifacts.

A related limitation concerns the token-based nature of the approach. In languages with invariable adjectives or other words that lack overt gender marking, annotators may nonetheless assign a gender feature based on syntactic context, since most words of that category in that language do inflect. This could produce apparent violations that reflect annotation practice rather than linguistic reality, as pointed out in [Brosa-Rodríguez and Jiménez-López \(2025\)](#).

7. Acknowledgements

This work was supported by the Universitat Rovira i Virgili *Pla de Foment de la Recerca (PFR)* of the Research Group on Mathematical Linguistics (GRLMC), under which Antoni Brosa-Rodríguez and M. Dolores Jiménez López conducted their research.

Antoni Brosa-Rodríguez's work was made possible through a research stay at the Universidad de Jaén, and he would like to express his sincere gratitude to the ALTYA (Análisis Lingüístico: Teoría y aplicaciones) Research Group for their support.

8. Bibliographical References

- Thomas Berg. 2024. Gender, number and person: a three-way interaction. *Linguistic Typology*, 28(3):537–565.
- Antoni Brosa-Rodríguez and M. Dolores Jiménez-López. 2023. *A Typometrical Study of Greenberg's Linguistic Universal 1*, Lecture Notes in Networks and Systems 741, pages 186–196. Springer.
- Antoni Brosa-Rodríguez and M. Dolores Jiménez-López. 2025. Beyond the data: The impact of annotation inconsistencies in ud treebanks on typological universals and complexity assessment. In *Proceedings of SIGTYP in ACL 2025*, pages 43–51. ACL.
- Antoni Brosa-Rodríguez and Sylvain Kahane. 2024. New proposal of greenberg's universal 14 from typometrics. In *Conference Proceedings of LREC-COLING 2024*, pages 12217–12226. ACL.
- Caron Bernard. 2026. *Hausa-Western Autogramm*. distributed via Universal Dependencies.
- Greville Corbett. 1991. *Gender*. Cambridge University Press.
- Greville Corbett. 2000. *Number*. Cambridge University Press.
- Matthew Dryer and Martin Haspelmath. 2013. *The World Atlas of Language Structures*. Max Planck Institute for Evolutionary Anthropology.
- Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2021. Typometrics: From implicational to quantitative universals in word order typology. *Glossa*, 6(1).
- Joseph Greenberg. 1963. *Universals of Language*, 1st edition, pages 58–90. The M.I.T. Press, Cambridge.
- Bruno Guillaume. 2021. Graph matching and graph rewriting: Grew tools for corpus exploration, maintenance and conversion. In *EACL 2021 16th Conference of the European Chapter of the Association for Computational Linguistics*.
- Matías Guzmán-Naranjo and Laura Becker. 2018. Quantitative word order typology with UD. In *17th International Workshop on Treebanks and Linguistics*.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers and Daniel Zeman. 2020. *Universal Dependencies*

v2: *An Evergrowing Multilingual Treebank Collection*. Proceedings of the 12th International Conference on Language Resources and Evaluation, 2.19.

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

Haitao Liu. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 6:1567–1578.

Matti Miestamo, Dik Bakker, and Antti Arppe. 2016. Sampling for variety. *Linguistic Typology*, 20(2):233–296.

Frans Plank. 2025. Typology re-done: Greenberg nos. 37, 45, 34. *Linguistic Typology*, Aop.

Frans Plank and Wolfgang Schellinger. 1997. The uneven distribution of genders over numbers: Greenberg nos. 37 and 45. *Linguistic Typology*, 1:53–101.

Riabi Arij, Farah Essaidi, Amal Fethi, Menel Mahamdi, and Djamé Seddah. 2026. *Maghrebi Arabic French-Arabizi*. distributed via Universal Dependencies.

Anna Siewierska. 2013. Gender distinctions in independent personal pronouns. *The World Atlas of Languages Structures*.

Andrew Spencer. 2002. Gender as an inflectional category. *Linguistics*, 38:279–312.

A. Treebanks with gender-number features in invariable categories

ADP (31 treebanks): Arabic-NYUAD, Bhojpuri-BHTB, Bulgarian-BTB, French-FQB, Greek-GDT, Greek-Messinian, Hindi-HDTB, Hindi-PUD, Icelandic-IcePaHC, Icelandic-PUD, Irish-Cadhan, Irish-IDT, Italian-ISDT, Italian-MarkIT, Italian-PUD, Italian-ParTUT, Italian-ParlaMint, Italian-PoSTWITA, Italian-VIT, Maghrebi_Arabic_French-Arabizi, Marathi-UFAL, Northern_Kurdish-Kurmanji, Old_Irish-DipSGG, Old_Irish-DipWBG, Portuguese-Bosque, Portuguese-DANTEStocks, Portuguese-GSD, Portuguese-PUD, Sindhi-Isra, Urdu-UDTB, Wolof-WTB.

ADV (29 treebanks): Arabic-NYUAD, Belarusian-HSE, Bhojpuri-BHTB, Faroese-OFT, French-ParisStories, German-GSD, German-HDT, Hindi-HDTB, Icelandic-GC, Icelandic-IcePaHC, Icelandic-Modern, Icelandic-PUD, Italian-ISDT,

Italian-KIParlaForest, Italian-MarkIT, Italian-Old, Italian-PUD, Italian-TWITTIRO, Italian-VIT, Portuguese-Bosque, Portuguese-GSD, Portuguese-PetroGold, Romanian-Nonstandard, Sanskrit-UFAL, Sanskrit-Vedic, Sindhi-Isra, Slovak-SNK, Upper_Sorbian-UFAL, Urdu-UDTB.

CCONJ (9 treebanks): Arabic-NYUAD, Belarusian-HSE, Bhojpuri-BHTB, Icelandic-IcePaHC, Italian-KIParlaForest, Italian-MarkIT, Italian-PUD, Italian-VIT, Urdu-UDTB.

INTJ (7 treebanks): Arabic-NYUAD, Beja-Autogramm, Bhojpuri-BHTB, Greek-Cretan, Greek-GUD, Greek-Lesbian, Icelandic-IcePaHC.

SCONJ (14 treebanks): Arabic-NYUAD, Beja-Autogramm, Bhojpuri-BHTB, Greek-GUD, Greek-Lesbian, Greek-Messinian, Hindi-HDTB, Icelandic-IcePaHC, Icelandic-PUD, Italian-MarkIT, Italian-VIT, Old_Irish-DipSGG, Portuguese-Bosque, Urdu-UDTB.

A Proposal for a More Universal Annotation of Relative Clauses in Universal Dependencies

Santiago Herrera¹, Sylvain Kahane^{2,3}

¹Université Sorbonne Paris Nord, LIPN, CNRS, UMR 7030

²Université Paris Nanterre, Modyco, CNRS ³IUF - Institut Universitaire de France
herrera@lipn.univ-paris13.fr, skahane@parisnanterre.fr

Abstract

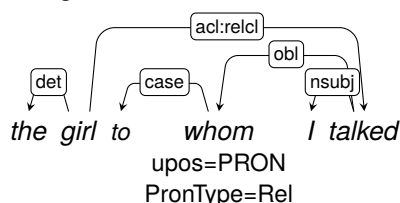
This paper proposes a new encoding of relative clauses compatible with the Universal Dependencies annotation scheme for syntactic treebanks. After showing that the current guidelines are based on the main strategy of relativization in European languages, we show that it cannot be easily extended to other relativization strategies, especially head-internal relatives clauses. The criteria for the POS of relativizers are discussed. We apply our annotation to an important variety of strategies in Mandarin, Japanese, Bambara, German, Basque, Turkish, Latin, Beja, Gbaya, and Wolof, as well as to various strategies in English, including participial clauses.

Keywords: relativization strategy, relativizer, head-internal relative clause, head-external relative clause, free relative clause, participial clause, language typology

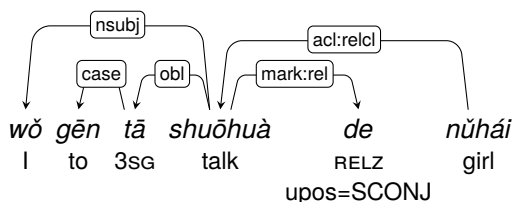
1. Introduction

The current UD annotation of relative clauses (henceforth RCs) is primarily based on the relation `acl:relcl` between the RC and its antecedent, and on the `PronType=Rel` feature on the relativizer, when it is a relative pronoun (`PRON` in (1a) for English).

(1) a. English



b. Mandarin



‘the girl I talked to’, lit. the girl that I talked to her

This annotation has been mainly thought for languages where the RC has a relative pronoun (`PronType=Rel`) and depends on a nominal antecedent (`acl:relcl`).¹ In some cases, such as

Authors are listed in alphabetical order.

¹This is confirmed by the definition of RCs given on the UD website: “A prototypical relative clause (RC) modifies a nominal that is understood to fulfill some grammatical role in the RC. The head is said to be ‘extracted’ from the RC.” (Universal Dependencies, 2026d)

in (1b) for Mandarin, the relativizer (`SCONJ`) is marked by the relation `mark:rel`. This usage is not widespread across treebanks, which usually annotate relativizers with `PronType=Rel` feature. However, when the RC is marked by a pure relativizer, the currently proposed encoding is not very satisfactory either. The `PronType=Rel` feature on a word analyzed as a subordinating conjunction can be misleading, and the relativized position is not indicated, whether it is a gap or it is occupied by a resumptive pronoun, such as *tā* in (1b).

Finally, encoding the fact that the clause is a RC on the label of the relation between the clause and its antecedent (`acl:relcl`) is also debatable, as the property of being a RC is a property of the clause itself rather than its relation with its antecedent.

Relative clauses with an antecedent and a relative pronoun are far from the only strategy for relativization (Keenan and Comrie, 1977; Downing, 1978; Lehmann, 1986; Comrie, 1998; Croft, 2022). It is, in fact, at one end of the continuum between all the possibilities for building a relative clause, even if it is a common one in European languages.

For this reason, and in order to be able to compare the different RC structures across languages, typologists work with functional or semantic definitions of the relativization phenomenon. From a functional perspective, a relative construction is an action that modifies a referent (Croft, 2022). Prototypically, a RC specifies or restricts a set of objects (Keenan and Comrie, 1977) denoted by a noun phrase in the matrix clause. This definition narrows RCs down to those that restrict the set of objects being referred to, while allowing for better cross-linguistic comparison.

In this paper, we propose a review of RC strate-

gies and a new encoding that can be adapted to the various strategies of relativization. We start our review at the other end of the spectrum, by the head-internal strategy.

Section 2 reviews recent contributions within the Universal Dependencies framework aimed at improving cross-linguistic comparisons of constructions. Additionally, it examines the current state of RC annotation in its last release (v2.17, Zeman et al., 2025). In section 3, we present the head-internal RCs in Japanese and Bambara, as well as the free RC of English. In section 4, we revisit the familiar strategy of head-external RCs, exemplified by English and, for a less common variant, by Gbaya. Sections 5 and 6 focus more on the relativizer itself. We start by the case of relativizer that are part of the verbal form, such as in Basque and Turkish, as well as some constructions of English. Section 6 is devoted to the analysis of relativizers which are separate forms and especially to the choice between `PRON` and `SCONJ`; it is illustrated by Latin, Beja, Wolof, and again English.

Our contributions can be summarized in three points:

- we provide an analysis of existing RC annotation, identifying a bifurcation between UD guidelines and actual annotation practices;
- we propose cross-linguistic analyzes for a more consistent annotation for different RC strategies;
- we introduce new tags that ensures comparative annotation without information loss.

2. Research Context

The UD collection has been growing in both size and diversity, and has become a priceless resource for quantitative syntactic typology (e.g., Levshina, 2019; Gerdes et al., 2021; Levshina, 2022) and other larger cross-linguistic studies, even for those typological studies that do not work directly with UD data, but use parsers that have been trained on it (e.g., Östling, 2015; Kann, 2024). In this context, comparability and typological reliability become crucial.

2.1. Related work

Continuous efforts are being made in UD to improve coverage of language diversity within a comparable framework, independent of language-specific strategies. One research line has focused on evaluating and revising UD annotation through the concept of comparative constructions (Croft et al., 2017; Nivre, 2025). Recently, these authors have started a review of the compatibility between the UD framework and linguistic typology

principles, in order to identify limitations and propose possible solutions. To date, the authors have studied reference and modification constructions (Nivre and Croft, 2025) and general verbal predication constructions (Croft and Nivre, 2025). Other work has focused on annotating constructions atop UD (Weissweiler et al., 2024) that could then be compared cross-linguistically. As far as we know, RCs have only been annotated as constructions in the EWT English Treebank (Silveira et al., 2014); the role of the relativized element is indicated even in cases where there is a gap, based on enhanced dependencies. Our work builds on the aforementioned authors' work, offering an immediate practical perspective. We revisited UD annotation, focusing only on relative clauses. Without proposing a different layer of annotation, we propose new features that are necessary to provide the information required for cross-linguistic comparisons of RCs. We argue that, given the current annotation, such a comparison is unfeasible. In addition, this work serves as a prerequisite for annotating RC constructions in the style of Weissweiler et al. (2024).

2.2. Relative Clauses in UD treebanks

Version 2.17 of UD comprises 339 treebanks for 186 languages. Analysis reveals that 21% of treebanks (with $\geq 1k$ tokens = 210 treebanks) do not contain an explicit RC relation (`relcl`) or a relative pronoun feature (`PronType=Rel`) within adnominal clauses (`acl`). This represents approximately 18% of the largest languages in UD, including Basque, Japanese, and Turkish. Even where annotation exists, retrieval of RCs is not straightforward. For instance, the Serbian and Croatian treebanks use the tag `PronType=Rel, Int`, combining relative and interrogative pronouns and making explicit filtering impossible.

While this absence could be attributed to inconsistent annotation practices, perhaps due to the optional status of relation subtypes, it suggests a general shortcoming across languages. The current preference in UD guidelines for strategies that use relative pronouns creates a representational imbalance. For example, languages using mainly participial or nominalized clauses (e.g., Turkish or Basque) lack the appropriate tags to annotate these constructions.

Consequently, annotators are driven to find workarounds for the scheme. We observe the feature `PronType=Rel` being forced onto non-pronominal categories like `SCONJ`, `AUX`, and `PART` in 15% of languages, simply to preserve the relative interpretation,² or the use of the more generic

²The guidelines indicate that `PronType=Rel` should be used for relative pronouns, determiners, numerals, and adverbs. (Universal Dependencies, 2026c)

`acl` label that obscure the specific syntactic function.

Ultimately, the current scheme fails to fully account for internally headed relatives, free relatives, or gapping strategies where the head is structurally integrated into the subordinate clause. For 40% of languages, it is not possible to identify the relativized element as they do not overtly indicate the relativizer (specifically when this is not an `SCONJ` or a `PART`). This results in an asymmetry of information encoding: only languages with explicit relative words (`PronType=Rel`) can fully encode the role of the relativized element, while other strategies are structurally misrepresented.

3. Head-Internal Relative Clauses

The Head-Internal RC (HIRC) strategy was first described by Kuroda (1974) for Japanese, and confirmed for Diegueño (now called Kumeyaay, Yuman) by Gorbet (1974) and for Navajo (Na-Dené) by Platero (1973). In HIRCs, the relativized noun stays inside the RC. Examples of HIRCs are now documented across a diverse range of language families. In this section, we first illustrate this phenomenon with Japanese and then move to Bambara (Niger-Congo).

Japanese complement clauses are built by adding the nominalizer affix *-no* to a finite clause.³ This is illustrated by (2b) where *-no* is added to the clause in (2a). This nominalized clause occupies the object position and receives the accusative marker *-o*.

(2) Japanese (Creissels, 2006)

- a. Ano hito-ga Nihon-e itta
DEM man-NOM Japan-to went
'This man went to Japan'
- b. [Ano hito-ga Nihon-e itta]-no-o
DEM man-NOM Japan-to went-NMLZ-ACC
shirimasen deshita
didn't know
'I didn't know that this man went to Japan'

The same strategy can be mobilized for RCs. In (3a), the verbal form *tukameata* takes *neko-o* as its object. The noun *neko* in (3a) can be replaced by the whole clause (3b) nominalized by *-no*, giving (3c). As the verb *tukameata* needs an entity as an object, the nominalized clause is now interpreted as an HIRC. The relativized position is not explicit

³The segmentation into words for Japanese is widely debated. We follow here the segmentation proposed by most typologists, which more or less corresponds to the traditional segmentation into *buntetsu*.

and the HIRC is potentially ambiguous, both arguments of *tukameata* being able to be the target.

(3) Japanese (Creissels, 2006)

- a. Neko-o Hanako-ga tukamaeta.
cat-ACC Hanako-NOM caught
'Hanako caught a cat.'
- b. Neko-ga sakana-o torootoshita.
cat-NOM fish-ACC tried_to_rob
'A cat tried to rob the fish.'
- c. [Neko-ga sakana-o torootoshita]-no-o
Hanako-ga tukamaeta.
'Hanako caught a cat that tried to rob the fish.'

It is inconvenient to annotate that there is a RC in (3c) given the current UD annotation of RCs. First, the relativizer is generally analyzed as a verbal affix and the feature `PronType=Rel` can hardly be associated to a verbal form. Second, there is no antecedent,⁴ therefore the `relcl` should be added directly to the relation between the verb *tukamaeta* and its object.⁵ Third, the relativized position would not be indicated. Consequently, we propose a completely different annotation:

- The fact that the clause is interpreted as an RC will be indicated by a clause-level feature `ClauseType=Rel` on its head.
- The position of the relativizer is indicated by a feature `Relz=Affix`, which replaces the feature `PronType=Rel`. Other values of `Relz` we consider are `Word` and `Stem`.
- The relativized location can be indicated by a feature `RelPosition=Yes`.⁶

Our analysis for (3c) is shown in (4). Note that the relative clause is analyzed as an `obj` of the main verb, because it refers to an entity. It contrasts

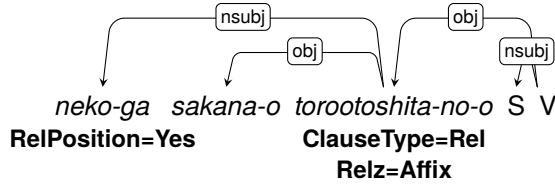
⁴Some authors, such as Murasugi (1991), analyzes *no* as a pronoun antecedent of the clause. In any case, we present other examples where the relativizer is a verbal affix in section 5.

⁵As we said in section 1, to be a RC is more a property of the clause itself than of its relation with its governor, although, in this case, it is the semantic properties of the verb governing the RC that triggers the RC interpretation. Encoding this information on a relation would involve adding `relcl` to a great number of relations. For other advantages of using a clause-level feature, see section 5 and 7.

⁶Here again, as suggested by one of our reviewers, it would be possible to add an subtype extension on the relation to indicate the relativized position. But it is really a property of the position itself and not of the relation with its governor.

with the subordinated clause in (2b) that refers to a proposition and would be `ccomp`.

(4) Japanese



Our second example is Bambara. UD has a small treebank of Bambara that contains 80 RCs (Aplonova and Tyers, 2017). Bambara has a unique relativizer, *min*, that can occupy any nominal relativized position or can be associated with a noun occupying the relativized position.⁷ The structure of the RC is identical to the structure of a main clause, the only difference being the additional relativizer *min* in situ. RCs in Bambara are almost always fronted. They are associated with an overt pronoun in the matrix clause if they occupy an argument position, as in (5a) and (5b), but not if they play an adverbial role, as in (5c). In examples (5a) and (5c), *min* is associated with the relativized noun (and can be analyzed as a determiner), while in (5b), it occupies the relativized position by itself.

(5) Bambara (v2.17, Aplonova and Tyers, 2017)

- a. a mana den **min** wolo, a
3SG COND.AFF child RELZ give_birth, 3SG
bè o dun
IPFV.AFF it eat

‘He ate the child she gave birth to’, lit. she gave birth to RELZ child, he ate it

- b. e ye **min** don nin ba
2SG.EMPH PFV.TR RELZ know this mother
ye o don.
PFV.TR it know

‘This mother knew who you knew’

- c. dennin nana tuma **min**, a
child.DIM come.PFV.INTR moment RELZ, 3SG
ye foli kè
PFV.TR salutation do

‘He saluted when the child arrived’

In the current annotation of Bambara treebank (v2.17, Aplonova and Tyers, 2017), the RC is analyzed as a dependent (`acl`) of the overt pronoun in the matrix clause. But the overt pronoun of the Bambara constructions is an ordinary pronoun that cannot be modified. It is equivalent to the resumptive pronoun in a dislocation (which is a common

⁷*min* can be pluralized, giving the form *minw*, which occurs in 6 of the 80 RCs.

construction in Bambara) and the construction is rather considered a correlative construction (Culy, 1990; Keenan and Comrie, 1977; Creissels, 2006; Croft et al., 2017). The link between the RC and the pronoun could be indicated by a coreference index (see Nivre and Croft (2025) regarding difficulties with coindexation in the UD). Here we propose to add a feature `RelAnt=Yes` on the pronoun in the matrix clause and to attach the RC to the main verb by a *dislocated* relation (Figure 1).⁸

In some sense, the HIRC of Bambara are not noun modifiers, but rather free relatives coreferent or not with a pronoun. It must be noted that Japanese HIRC are also close to free RCs; it is possible to suppress the noun in the relativized position, which becomes a gap, to obtain a meaning equivalent to a free RC in English.⁹

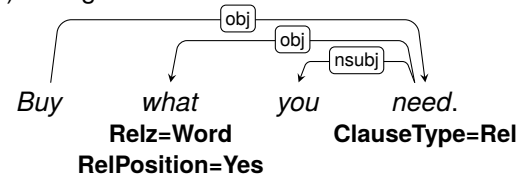
(6) Japanese

[Sakana-o torootoshita]-no-o
[fish-ACC tried_to_rob]-RELZ-ACC
Hanako-ga tukamaeta.
Hanako-NOM caught

‘Hanako caught what tried to rob the fish.’

Many authors have made a link between HIRC and free RCs and it is possible to analyze them in the same way we have analyzed HIRC (Basilico, 1996; Grosu and Landman, 1998). In (7), *what* is a relative pronoun that fills a position inside the clause (as *min* in (5b)). As for (4), the relative clause, which refers to an entity, is `obj` of *buy*, even though we propose not to take *what* as an external head.

(7) English



This analysis is different from the current UD analysis, where the relative pronoun is annotated as the head, arguing that “Free relatives are nominals containing a relative clause where the relativizer is ‘fused’ with the head of the whole nominal” (Universal Dependencies, 2026d). Some authors (Lehmann, 1986; Le Goffic, 2002) defend a reverse analysis, considering that RCs with an antecedent have developed from the combination of

⁸We can also mention the relation `acl:crel` that has been proposed for UD_Sanskrit-Vedic treebank (Universal Dependencies, 2026a).

⁹Grosu and Hoshi (2016) analyze this construction as a “gapped light-headed HERC”, considering that *no* is more pronominal in this case.

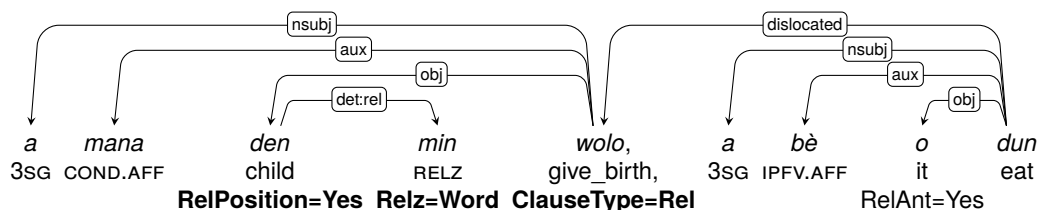


Figure 1: A correlative analysis for the Bambara RC.

a noun and a free relative occupying together the same position. For this reason, and to homogenize annotation, we believe it would be relevant to analyze free relatives in the same way as HIRCs.

We must nevertheless mention an additional problem, the case conflict (Bresnan and Grimshaw, 1978; Vogel, 2003), which is well illustrated by German. In German free RCs, the relative pronoun receives the case addressed to the relativized position, which proves it occupies this position inside the RC. But this case must also correspond to that assigned by the matrix governor to the entire RC: In (8a), the nominative relative pronoun *wer* is the subject of the RC and the RC is the subject of the main clause, while in (8b), the accusative relative pronoun *wen* is the object of the RC and the RC is the object of the main clause. But in (8d), the relative pronoun *wer* is the subject of the RC and the RC is the object of the main clause, which is unacceptable, and an overt accusative pronoun, here *den*, is needed as the antecedent, as shown in (8c).

(8)

- a. **Wer** das behauptet, ist ein Lügner.
who.NOM that claims, is a liar.
'(Anyone) who claims that is a liar.'
- b. Ich küsse, **wen** ich liebe.
I kiss, who.ACC I love.
'I kiss whoever I love.'
- c. **Wer** mir so etwas antut,
who.NOM to.me such a thing does,
den schaue ich nicht mehr an.
him.ACC look I no more at.
'I won't look at (anyone) who does something like that to me.'
- d. ***Wer** mir so etwas antut, schaue ich nicht mehr an.

The data in (8) confirm the analysis proposed by Tesnière (1959) that a wh-word occupies a double syntactic position (Kahane, 2002; Kahane and Osborne, 2015) and is both the head of the RC and a pronoun governed in the RC. We argue that the aim of the UD tree is not to reflect the exact structure, and that it is reasonable to attribute a unique

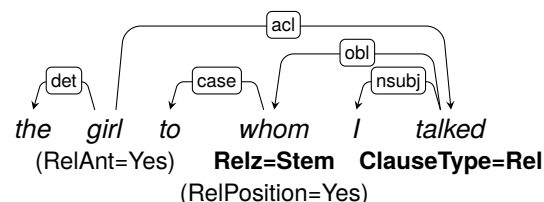
position to the relative pronoun to enhance comparison. The solution we propose in figure 7 has the advantage to maintain similar analyses for relative pronouns in RCs with or without antecedent. Following the same principles, we propose a different analysis for free RCs in languages with other types of relativizers, such as Wolof in section 6.

4. Head-External Relative Clauses

We now come back to our first examples, which were head-external RCs (HERCs), that is, RCs where there is a noun or a pronoun, called the antecedent, that governs the RC and is coreferent with the relativized position. This is by far the most common strategy across European languages and it has often been analyzed as an extraction of the relativized element. Our analysis of HIRCs can be easily extended to our two examples of HERCs in (1), giving the analyses in (9). The sub-relation *relcl* is replaced by the feature *ClauseType=Rel* and the feature *PronType=Rel* by *Relz=Stem*.¹⁰ A feature *RelPosition=Yes* has been added on the resumptive pronoun in (9b). We put into brackets the features that are optional because they are easily deductible from the structure.

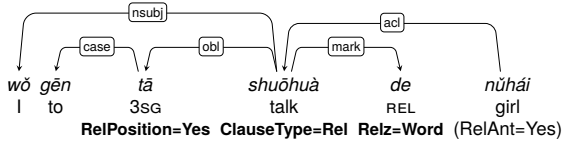
(9)

a. English



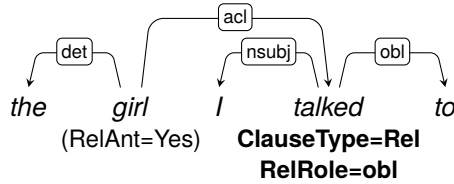
b. Mandarin

¹⁰It is also possible to keep *PronType=Rel* for relative pronouns, especially for treebanks that associate a feature *PronType* to every pronoun. Note that here the relative pronoun is inflected with a case suffix and the relativizer itself is the stem.

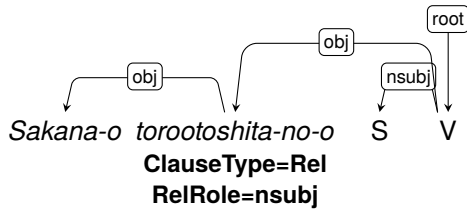


English also uses the gap strategy as in (10a), where the argument that is relativized in the RC is not expressed. Without a relative pronoun, it is not possible to recover information about the syntactic role of the relativized element from the annotation. To be able to quantify this information and allow more fine-grained cross-linguistic studies, we propose to introduce the feature **RelRole** on the governor of the gap with its syntactic role as a value. The same analysis can be applied to the free RC of Japanese we presented in (6).

(10) a. English



b. Japanese



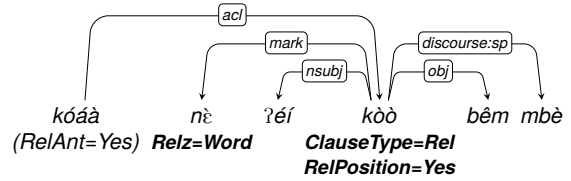
The current UD annotation guidelines proposed to encode the position to the gap by an enhanced dependency from the governor of the gap to the antecedent (Universal Dependencies, 2026b). But this solution would not be easy for the free RC of Japanese in (10b), because there is no antecedent. Moreover, the enhanced dependency level uses a different encoding of the dependencies and only a few treebanks have implemented it. Here we just propose to introduce a **RelRole** feature that can be used to easily produce a correct enhanced level of annotation.

Another interesting example is that of RCs which are both externally and internally headed. We propose to call them *RCs with a resumptive antecedent*.¹¹ A puzzling example is found in some Gbaya RCs, such as (11), where there is a nominalized verbal antecedent that is also present in the RC in a finite form. Since the relative clause overtly expresses all arguments, it is the verb that is coreferent to the head of the RC. To mark this, we again use the feature **RelPosition=Yes**.

¹¹Hu et al. (2016) describe RCs "with a resumptive NP" in Mandarin children's productions.

- (11) Gbaya (v2.17, Roulon-Doko et al., 2025)
- | | | | |
|----------------------|-------------|----------|-------------------|
| kóàà | nè | ʔéí | kòò |
| give_birth.NMLZ | RELZ | LOG.SG | give_birth.PFV.MT |
| bêm | mbè | zók | mò ná |
| child | amazing | see.IPFV | thing DISC |
| ká | bê mí | ʔá-nè | dèè |
| and_then | child.ANAPH | TOP | do.PFV.MT |
| ḑéŋ-kéyéŋ | ʔòóyè | | |
| tall_(6_8_years_old) | of_course | | |

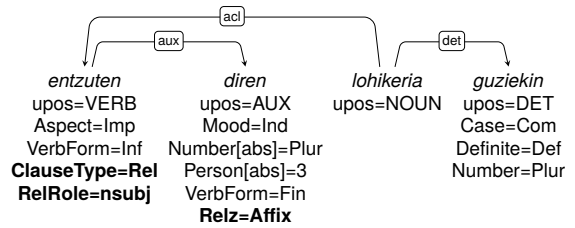
'Having surprisingly given birth to a child, she realizes that the child is of course grown up [6-8 years].'



5. Relative Clauses with Relativizer on the Verb

Usually, the relativizer is a verbal affix rather than a separate form. We have already presented the case of Japanese in (3c). This is the only strategy in various languages. In Basque, the relativization is ensured by a suffix *-(e)n* on the finite form (which is generally an auxiliary). For example, in (12), the finite auxiliary *diren* (lemma: *izan*) is relativized by the suffix *-n*. We mark the type of clause and the relativized position with **RelRole=nsbj** on the verb and the presence of the relativizer with **Relz=Affix** on the auxiliary.

- (12) Basque (v2.17, Aranzabe et al., 2015)



'with all nonsense that are heard'

In Turkish, the relativization is ensured by the participial form of the verb and the subject must be realized as a possessive nominal complement. In (13b), the nominal complement (**nmod**) *çocuk* 'child' of (13a) is relativized. The verb is now at the participial form *çalış-tığ* and the 1PL subject pronominal index *-uz* has been replaced by the possessive index *-ımız*. The role of the gap in the RC is indicated by the feature **RelRole=nmod** on the governor of the relativized position.

(13) Turkish (Creissels, 2006)

- a. O çosuğ-un baba-sı ile
DEM child-GEN father-CSTR with
çalış-ıyor-uz.
work-PROG-S1PL
'We work with this child's father.'
- b. [baba-sı ile çalış-tığ-ımız]
[father-CSTR with work-PART-POSS1PL]
RelRole=nmod **ClauseType=Rel**
Relz=Affix

çocuk
child
(RelAnt=Yes)

'the child whose father we work with',
lit. [our working with (the) father] child

Traditional grammars of European languages distinguish the RCs from participial clauses. Although the two constructions are different, there is a continuum between them, as shown by Basque and Turkish RCs. In both constructions, the clause's governor is co-referent with a position within the clause. In (14a) and (14b), we analyze English participial clauses. We use the feature `ClauseType=Rel`, but it would be possible to introduce a feature `ClauseType=Part`, assuming that we are able to draw a line between RCs and participial clauses and to decide on which side are Basque and Turkish. These two constructions share, with the construction in (14c), the fact that a position is missing in the subordinated clause. We consider them to be particular cases of non-finite RCs involving a gap (which role is indicated by the feature `RelRole`). These constructions must be clearly distinguished from (14d), where there is no gap in the subordinated clause. We assigned to this clause the feature `ClauseType=Comp`.

(14) English

- a. *people living in glass houses*
Relz=Affix
RelRole=nsubj
ClauseType=Rel
- b. *students given high marks*
Relz=Affix
RelRole=nsubj:pass
ClauseType=Rel
- c. *I found a house to live in*
Relz=Word
RelRole=obj
ClauseType=Rel
- d. *the decision to leave was difficult*
ClauseType=Comp

6. Relative Clauses with a Separate Relativizer

We have seen that the use of a relativizer as part of the verbal form is quite common, even in Indo-European languages, and that UD does not foresee the use of a specific relative feature on verbs. This is part of a general preference for pronominal analysis of relatives, while the word class of relativizer is more of a gradient continuum than a discrete category.

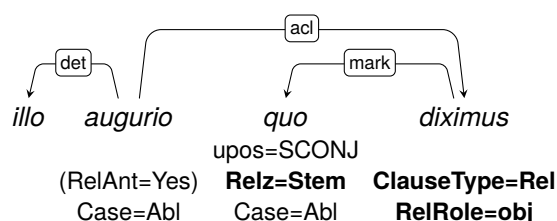
One strong argument for classifying a marker as a relative pronoun is its ability to agree with the antecedent while independently marking the case of the relativized element. In standard Latin (15a), the relative pronoun agrees in number and gender with the relative's head, but distinctly marks the accusative case of the extracted position (the object of the relative clause). However, Latin also exhibits case attraction, where the pronoun marks the case of the antecedent rather than retaining the case of the relativized position (15b) (Comrie, 1989; Creissels, 2006). This phenomenon blurs categorisation. While the relativizer retains its agreement properties, it behaves more like a modifying element that agrees with the head than as an independent argument of the clause. This is a common pattern across different languages.

(15) Latin (Creissels, 2006)

- a. illo augurio quod
DEM.SGN.ABL omen.SG.ABL RELZ.SGN.ACC
diximus
say.ACP.S3SG
'by that omen which we mentioned'
- b. illo augurio quo
DEM.SGN.ABL omen.SG.ABL RELZ.SGN.ABL
diximus
say.ACP.S3SG
'by that omen which we mentioned'

We think, with Creissels (2006), that this is where we must draw the line between relative pronouns and other relativizers. In (15b), the relativizer *quo* no longer occupies a position inside the relative clause, but it agrees with the antecedent. We call it an *agreeing relativizer*.

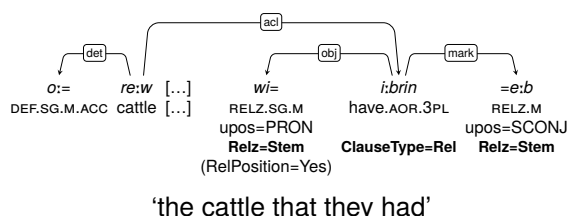
(16) Latin



Agreeing relativizers must be analyzed as markers of the relativization with relation *mark*, as in (16). For the POS, we use *SCONJ*, even if *PRON* would make sense too.

The contrasts between relative pronouns and agreeing relativizers is also well illustrated in Beja (Vanhove, 2014; Kahane et al., 2021). Beja uses a complex strategy where both an agreeing relativizer and a relative pronoun can coexist within the same structure, as in (17). While the agreeing relativizer marks the right boundary (Beja is a head-final language), the relative pronoun occupies the relativized position. Consequently, both elements may be annotated with the feature *Relz=Yes*, with the distinction between them relying only on their part of speech and syntactic roles.

(17) Beja (v2.17, Kahane et al., 2021)



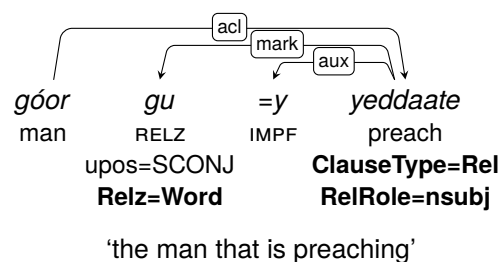
Following our analysis of relativizers in Latin and Beja, we think that, in many other languages, relativizers should be analyzed as agreeing relativizers (*upos=SCONJ*, *mark*) rather than relative pronouns. It is for instance the case of Wolof where the relativizer agrees in class noun with its antecedent, but does not mark the relativized position. In (18), the relativizer *gu* is the combination of a noun class marker *g-* (corresponding to the class of the antecedent noun) and a suffix *-u* (one of the three possible suffixes for relativizers).

The switch from the current annotation to this new annotation would not lose any information as soon as we have a feature such as *RelRole* to mark the syntactic role of the relativized position.

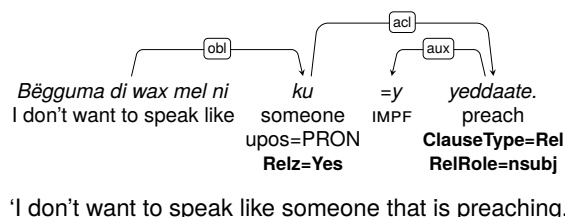
Wolof has also a lot of “free” RCs, where the noun phrase starts by a relativizer. These relativizers have the same form as for RCs with nominal antecedent, but in this case the noun class is meaningful. For instance, in (18b), the *k-* of the relativizer *ku* means ‘someone’ and the relativizer must be analyzed as a pronoun. Contrary to the case of relative pronoun that occupy a position inside the RC (in addition to their head position), this pronoun does not mark the relativized position and must be analyzed as the head, as in the current UD analysis. Note that, with our analysis, the structure of the RC is very similar in (18) and (18b); in both cases, we have a gap and a *RelRole* feature indicating its position. This contrasts with the current UD analysis where the relativizer would be analyzed as filling the relativized position in example (18) but not in (18b).

(18) Wolof (v2.17, Dione, 2019)

a. Free RC

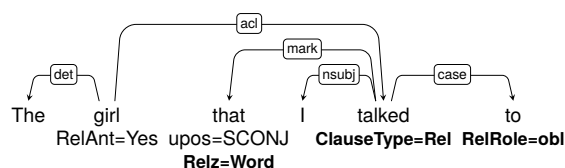


b. RC with pronominal antecedent



The third case of relativizers are pure relativizers, that do not occupy the relativized position in the RC and do not agree with the antecedent. This was illustrated by our Mandarin example (9b). We can also mention languages, like English, where the categorical status of relativizers is highly controversial (in typology, see e.g., Comrie, 1989; Andrews, 2007; Croft, 2022). In (19), the marker *that* is invariant: it shows no agreement with the antecedent and serves primarily to mark the left boundary of the relative clause. Importantly, unlike true relative pronouns (such as *which*), *that* cannot keep the preposition with it (e.g., *in which* vs. **in that*, the so-called pied-piping).

(19) English



One might hesitate to label *that* as a pure relativizer (*SCONJ*), concerned that information about the role of the extracted element would be lost. However, the features introduced in this article preserve this structural information specifically. Rather than obscuring the data, this approach enables robust cross-linguistic comparisons to be made: constructions with highly specific, language-dependent properties can now be analyzed using a unified annotation scheme. This makes the datasets comparable and explicitly encodes the typological differences, allowing them to be systematically quantified.

	Strategy	Relz	RC Features	Example
HIRC	Correlative	upos=DET PRON	RelPosition, RelAnt	Bambara
	Prototypical HIRC	Verbal affix	RelPosition	Japanese
	Free relative	upos=PRON upos=SCONJ Verbal affix	RelPosition RelRole RelRole	English, German Wolof Japanese
HERC	Resumptive antecedent	upos=SCONJ	RelPosition	Gbaya
	Resumptive pronoun	upos=SCONJ	RelPosition	Mandarin
	Gap	upos=SCONJ Verbal affix ∅	RelRole	English, Wolof Basque, Turkish, English participial English
	Relative pronoun	upos=PRON upos=PRON + upos=SCONJ	RelPosition RelPosition	English, Latin Beja

Table 1: Overview of the RC strategies analyzed and their encoding features. The feature `ClauseType=Rel` is not included as it is always annotated.

7. Conclusion

The goal of this paper was to highlight certain limitations in the current UD encoding of relative clauses and to propose a more robust encoding capable of handling a wide range of cross-linguistic strategies. Although we do not claim to cover every possible strategy, our review provides a broad foundation, and we believe the proposed encoding is sufficiently flexible to annotate additional strategies beyond those explicitly discussed. Ultimately, the encoding must be discussed with the entire UD community to determine its viability and explore whether an equivalent encoding can be developed within UD, or through another constructional approach if not (e.g., [Weissweiler et al., 2024](#)).

It should be noted that the conversion from the current UD encoding to the proposed encoding can quite easily be done with graph rewriting system such as Grew ([Guillaume, 2021](#)). But for many strategies of relativization, current UD treebanks lack information, especially when the relativizer is a verbal affix or the relativized position is a gap or a resumptive pronoun, which is likely to be the case for many languages in the UD collection. Moreover, there are currently no clear lines drawn between relative pronouns and other relativizers and many analyses should be revised to achieve a more homogeneous collection of treebanks, even for well-described languages, such as English or French ([Kayne, 1974](#); [Kahane, 2002](#)), where it has been shown that they mix relative pronouns and pure relativizers.

Our new annotation is based on features identifying the RC and its four types of element: relativizers, the relativized position, the role of the relativized position (when it is a gap or a resumptive pronoun or antecedent), and the antecedent in the main clause (when it is not the governor of the clause). In table 1, we propose a summary of the

various strategies of relativization presented here and the features used to encode them. Each time, we indicate the nature of the relativizer(s).

The decision to use clause-level features is non-trivial and creates natural tension within a dependency framework. However, the debate over incorporating constituency-oriented annotations is already occurring in practice and is proving essential for cross-linguistic comparison. Throughout this paper, we have outlined some theoretical and practical advantages of using these features rather than subtype relations. Interpreting these features as scoped to a node’s projection is straightforward, as long as the scope of each feature is clear. This requires a clear categorisation of features by their structural level (e.g., lexical vs clausal). Ultimately, the use of these features should also be a topic of discussion within UD.

Finally, another advantage of our encoding is that it can be extended to other constructions, such as interrogation and clefting ([20](#)). While interrogation is structurally (and genetically) similar to relativization, it is currently not annotated in UD treebanks. And in clefting, the role of the extracted element is currently not indicated (unless there is an enhanced level of analysis).

- (20) a. Relativization
the girl who biked
upos=PRON ClauseType=Rel
Relz=Word
- b. Interrogation
I wonder who biked
upos=PRON ClauseType=Int
PronType=Int
- c. Clefting
It is her that biked.
upos=SCONJ ClauseType=Cleft
CleftRole=nsubj

8. Acknowledgements

We would like to thank our colleagues for the discussions we have had about the data: Katya Aplonova for Bambara, Martine Vanhove for Beja, and Qishen Wu for Mandarin.

9. Bibliographical References

- Avery D. Andrews. 2007. Relative clauses. *Language Typology and Syntactic Description*, 2:206–236.
- Ekaterina Aplonova and Francis M. Tyers. 2017. [Towards a dependency-annotated treebank for Bambara](#). In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 138–145, Prague, Czech Republic. Association for Computational Linguistics.
- Maxux Aranzabe, Aitziber Atutxa, Kepa Ben-goetxea, Arantza Díaz de Ilarraza, Iakes Goenaga, Koldo Gojenola, and Larraitz Uriá. 2015. Automatic conversion of the Basque dependency treebank to Universal Dependencies. In *Proceedings of the 14th International Workshop on Treebanks and Linguistic Theories (TLT 2015)*, pages 233–241, Warsaw, Poland. Institute of Computer Science of the Polish Academy of Sciences.
- David Basílico. 1996. [Head position and internally headed relative clauses](#). *Language*, 72(3):498–532.
- Joan Bresnan and Jane Grimshaw. 1978. The syntax of free relatives in english. *Linguistic Inquiry*, 9(3):331–391.
- Bernard Comrie. 1989. *Relative Clauses*, 2nd edition, chapter 7. University of Chicago Press, Chicago.
- Bernard Comrie. 1998. Rethinking the typology of relative clauses. *Language Design: journal of theoretical and experimental linguistics*, 1:059–85.
- Denis Creissels. 2006. *Relatives postnominales*, chapter 33. Lavoisier, Paris.
- William Croft. 2022. [Events as Modifiers: Relative Clause Constructions](#), chapter 19. Cambridge University Press, Cambridge.
- William Croft and Joakim Nivre. 2025. [Verbal predication constructions in Universal Dependencies](#). In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 50–60, Düsseldorf, Germany. Association for Computational Linguistics.
- 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. CEUR Workshop Proceedings.
- Christopher D. Culy. 1990. *The syntax and semantics of internally headed relative clauses*. Stanford University.
- Cheikh Bamba Dione. 2019. [Developing Universal Dependencies for Wolof](#). In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest)*, pages 12–23, Paris, France. Association for Computational Linguistics.
- Bruce T. Downing. 1978. Some universals of relative clause structure. In Joseph H. Greenberg, editor, *Universals of Human Language, Volume 4: Syntax*, pages 375–418. Stanford University Press, Stanford, CA.
- Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2021. [Typometrics: From Implicational to Quantitative Universals in Word Order Typology](#). *Glossa: a journal of general linguistics*, 6(1):17.
- Larry P/ Gorbet. 1974. *Relativization and complementation in Diegueno: Noun phrases as nouns*. Ph.D. thesis, University of California, San Diego.
- Alexander Grosu and Koji Hoshi. 2016. Japanese internally headed relatives: Their distinctness from potentially homophonous constructions. *Glossa: a journal of general linguistics*, 1(1).
- Alexander Grosu and Fred Landman. 1998. Strange relatives of the third kind. *Natural Language Semantics*, 6(2):125–170.
- 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.
- Shenai Hu, Anna Gavarró, and Maria Teresa Guasti. 2016. Children’s production of head-final relative clauses: The case of mandarin. *Applied Psycholinguistics*, 37(2):323–346.
- Sylvain Kahane. 2002. À propos de la position syntaxique des mots qu-. *Verbum (Presses Universitaires de Nancy)*, 24(4).

- Sylvain Kahane and Timothy F. Osborne. 2015. Translators' introduction. In *Lucien Tesnière, Elements of structural syntax*, pages ix–xx–lxiii. John Benjamins.
- 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.
- Amanda Kann. 2024. [Massively multilingual token-based typology using the parallel Bible corpus](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 11070–11079, Torino, Italia. ELRA and ICCL.
- Richard Kayne. 1974. French relative *que* (1). *Recherches Linguistiques*, 11:40–61.
- Edward L. Keenan and Bernard Comrie. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, 8(1):63–99.
- Sige-Yuki Kuroda. 1974. Pivot-independent relativization in Japanese (i). *Journal of Japanese Linguistics*, 3(1-2):61–96.
- Pierre Le Goffic. 2002. Marqueurs d'interrogation/indéfinition/subordination: essai de vue d'ensemble. *Verbum*, 24(4):315–340.
- Christian Lehmann. 1986. [On the typology of relative clauses](#). *Linguistics*, 24(4):663–680.
- Natalia Levshina. 2019. [Token-based typology and word order entropy: A study based on universal dependencies](#). *Linguistic Typology*, 23(3):533–572.
- Natalia Levshina. 2022. [Corpus-based typology: applications, challenges and some solutions](#). *Linguistic Typology*, 26(1):129–160.
- Keiko Murasugi. 1991. *Noun phrases in Japanese and English: A study in syntax, learnability and acquisition*. Ph.D. thesis, University of Connecticut.
- 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, Tallinn, Estonia. University of Tartu Library.
- Joakim Nivre and William Croft. 2025. [Reference and modification in Universal Dependencies](#). In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 1–10, Ljubljana, Slovenia. Association for Computational Linguistics.
- Robert Östling. 2015. [Word order typology through multilingual word alignment](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 205–211, Beijing, China. Association for Computational Linguistics.
- Paul R. Platero. 1973. A study of the Navajo relative clause. Master's thesis, Massachusetts Institute of Technology.
- Paulette Roulon-Doko, Sylvain Kahane, and Bruno Guillaume. 2025. [A morpheme-based treebank for Gbaya, an Ubangian language of Central Africa](#). In *Proceedings of the Eighth International Conference on Dependency Linguistics (Depling, SyntaxFest 2025)*, pages 93–102, Ljubljana, Slovenia. Association for Computational Linguistics.
- Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. [A gold standard dependency corpus for English](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Klincksieck.
- Universal Dependencies. 2026a. [acl:crel: Correlative Clausal Modifier](#). <https://universaldependencies.org/sa/dep/acl-crel>. Accessed: January 2026.
- Universal Dependencies. 2026b. [Enhanced Dependencies: Relatives Clauses](#). <https://universaldependencies.org/u/overview/enhanced-syntax.html#relative-clauses>. Accessed: January 2026.
- Universal Dependencies. 2026c. [Pron-type: Pronominal Type](#). <https://universaldependencies.org/u/feat/PronType>. Accessed: January 2026.
- Universal Dependencies. 2026d. [Relative clause](#). <https://universaldependencies.org/>

[workgroups/newdoc/relative_clauses](#).
Accessed: January 2026.

Martine Vanhove. 2014. Beja grammatical sketch.
The corpAfroAs corpus of spoken AfroAsiatic languages, pages 1–68.

Ralf Vogel. 2003. Surface matters. case conflict in the free relative constructions and case theory. In E. Brandner and H. Zinsmeister, editors, *New perspectives in Case Theory*. CSLI.

Leonie Weissweiler, Nina Böbel, Kirian Guiller, Santiago Herrera, Wesley Scivetti, Arthur Lorenzi, Nurit Melnik, Archana 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)*, pages 16919–16932, Torino, Italia. ELRA and ICCL.

10. Language Resource References

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Jephthé Adolphe, Noëmi Aeppli, 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, Erik Andersen, Kirk Andrews, Matthew Andrews, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arican, Hórunn Arnardóttir, Wirote Aroonmanakun, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Hiwa Asadpour, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmaoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, Elena Badmaeva, Jana Bajorat, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Bryan Khelven da Silva Barbosa, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shabnam Behzad, Nathanaël Beiner, Juan Belieni, Alevtina Bémová, Kepa Bengoetxea, İbrahim Benli, Yifat Ben Moshe, Marie Benzerrak, Aleksandrs Berdicevskis, Ansu Berg, Gözde Berk, Delphine

Bernhard, Astrid Berntsson Ingelstam, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Esma Fatıma Bilgin Taşdemir, Helin Binici, Kristín Bjarnadóttir, Verena Blaschke, Rogier Blokland, Nina Böbel, Victoria Bobicev, Loïc Boizou, Stavros Bompolas, Johnatan Bonilla, Emanuel Borges Völker, Lars Borin, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaa, António Branco, Myriam Bras, Théo Brillet, Kristina Brokaitė, Lanni Bu, Eva Buráňová, Aljoscha Burchardt, Carmen Cabeza, Natalia Cáceres Arandia, Olesea Caftanатов, Marisa Campos, Marie Candito, Caterina Maria Cappello, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Anila Çepani, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Claudine Chamoreau, Shweta Chauhan, Yifei Chen, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Bermet Chontaeva, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Esther Cocco, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Claudia Corbetta, Daniela Corbetta, Francisco Costa, Marine Courtin, Benoît Crabbé, Mihaela Cristescu, Vladimir Cvetkoski, Netanel Dahan, Ingerid Løyning Dale, Sabrina D'Alì, Philemon Daniel, Khensa Daoudi, Bijayalaxmi Dash, Satya Ranjan Dash, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Ahmet Demir, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Roberto Antonio Díaz Hernández, Carly Dickerson, Ariani Di Felippo, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Hoa Do, Kaja Dobrovoljc, Caroline Döhmer, Adrian Doyle, Timothy Dozat, Kira Droganova, Magali Sanches Duran, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Roald Eiselen, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Louise Esher, Soudabeh Eslami, Farah Essaidi, Aline Etienne, Wograinne Evelyn, Sidney Facundes, Richárd Farkas, Ján Faryad, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Barbara Francioni, Theodorus Fransen, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Edith Galy, Federica Gamba, Marcos Garcia, José María García-Miguel, Moa Gärdenfors, Tanja Gaus-tad, Efe Eren Genç, Fabrício Ferraz Ger-

ardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökirmak, Yoav Goldberg, Gili Goldin, Xavier Gómez Guinovart, Berta González Saavedra, Mathieu Goux, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Mario Guglielmetti, Bruno Guillaume, Kirian Guiller, Céline Guillot-Barbance, Tunga Güngör, Vladimir Gurevich, Nizar Habash, Henrik Hafsteinsson, Michael Hahn, Jan Hajič, Jan Hajič jr., Eva Hajičová, Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Naïma Hassert, Dag Haug, Jiří Havelka, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Diana Hoefels, Barbara Hoff, Petter Hohle, Nick Howell, Yidi Huang, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Inessa Iliadou, Anton Karl Ingason, Radu Ion, Elena Irimia, Olájidé Ishola, Artan Islamaj, Kaoru Ito, Federica Iurescia, Jessica K. Ivani, Sandra Jagodzińska, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Ratanon Jiamsundutsadee, Katharine Jiang, Sylvanus Job, Mayank Jobanputra, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Zhuoxuan Ju, Markus Juutinen, Hüner Kaşıkara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóga, Jiří Kárník, Andre Kåsen, Tolga Kayadelen, Sarveswaran Kengatharayer, Václava Kettnerová, Lilit Kharatyan, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Petr Kocharov, Arne Köhn, Abdullatif Köksal, Veronika Kolářová, Kamil Kopacewicz, Timo Korriakangas, Mehmet Köse, Alexey Koshevoy, Nelda Kote, Natalia Kotsyba, Barbara Kovačić, Jolanta Kovalevskaitė, Emmanuelle Kowner, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Lucie Kučová, Adrian Kuqi, Elmurod Kuriyozov, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyung Kwak, Kris Kyle, Kābi Laan, Veronika Laippala, Lorenzo Lambertino, Israel Landau, Tatiana Lando, Septina Dian Larasati, Pierre Larrivé, Kusum Lata, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Wei Qi Leong, Saran Lertpradit, Herman Leung, Lori Levin, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Zoey Liu, Nikola Ljubešić, Irina Lobzhanidze, Olga Loginova, Markéta Lopatková, Lucelene Lopes, Edita Luftiu, Arsenii Lukashevskiy, Stefano Lusito, Anne-Marie Lutgen, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi,

Jean Maillard, Punyanuch Maitreenukul, Ilya Makarchuk, Aibek Makazhanov, Francesco Mambrini, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărânduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Cláudia Martins, Arianna Masciolini, Jan Mašek, Sanatbek Matlatipov, Hiroshi Matsuda, Yuji Matsumoto, Caterina Mauri, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Maitrey Mehta, Pierre André Ménard, Gustavo Mendonça, Hilla Merhav, Tatiana Merzhovich, Paul Meurer, Niko Miekka, Marie Mikulová, Emilia Milano, Aleksandra Miletić, Aaron Miller, Junghyun Min, Yael Minerbi, Jiří Mírovský, Karina Mischenkova, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Biswakalpita Mohapatra, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskiy, Katerina Mouzou, Kadri Muischnek, Robert Munro, Yugo Murawaki, Nikolett Mus, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhle, Minoo Nassajian, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērkalne, Manuela Nevaci, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Victor Norrman, Alireza Nourian, Michal Novák, Maria das Graças Volpe Nunes, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adedayò Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Annika Ott, Lilja Øvreliid, Masanori Oya, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Petr Pajas, Thomas Palakapilly, Alessio Palmero Aprosio, Jarmila Panevová, Ludovica Pannitto, Anastasia Panova, Thiago Alexandre Salgueiro Pardo, Shantipriya Parida, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Oggi Peeters, Angelika Peljak-Łapińska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Sílvia Pereira, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Eva Pettersson, Andrea Peverelli, Jason Phelan, Claudel Pierre-Louis, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Rodrigo Pintucci, Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Alstair Plum, Thierry Poibeau, Charin Polpanumas, Larisa Ponomareva, Martin Popel, Clamença Poujade, Rangga Prangwedana Prangwedana, Lauma Pretkalniņa, Rigardt Preto-

rius, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Christoph Purschke, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Ella Rabinovich, Alexandre Rademaker, Mutee-u Rahman, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Joana Ramos, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arij Riabi, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Norton Trevisan Roman, Mykhailo Romanenko, Natalia Romanova, Rudolf Rosa, Valentin Roşca, Paulette Roulon, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Paolo Ruffolo, Kristján Rúnarsson, Rozana Rushiti, Attapol T. Rutherford, Shoval Sadde, Pegah Safari, Aleksí Sahala, Kalyanamalini Sahoo, Saraswati Sahoo, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Konstantinos Sampanis, Stephanie Samson, Xulia Sánchez-Rodríguez, Manuela Sanguinetti, Ezgi Saniyar, Dage Särg, Marta Sartor, Albina Sarymsakova, Mitsuya Sasaki, Baiba Saulīte, Agata Savary, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Emmanuel Schang, Robert Schikowski, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Sven Sellmer, Mojgan Seraji, Magda Ševčíková, Petr Sgall, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Gyu-Ho Shin, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Jean Sibille, Janine Siewert, Einar Freyr Sigurðsson, João Silva, Aline Silveira, Natalia Silveira, Sara Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Dolores Solberger, Barbara Sonnenhauser, Shafi Sourov, Nina Speransky, Rachele Sprugnoli, Panyut Sriwirote, Vivian Stamou, Steinhórfur Steingrímsson, Antonio Stella, Jan Štěpánek, Barbora Štěpánková, Abishek Stephen, Milan Straka, Omer Strass, Emmett Strickland, Jana Strnadová, Sara Stymne, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Jack Sun, Hakyung Sung, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Maria Irena Szawerna, Chihiro Taguchi, Dima Taji, Luigi Talamo, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Nursena Teker, Samson Tella, Isabelle Tellier, Marinella Testori, Santhawat Thanyawong, Guillaume Thomas, Tarik Emre Tıraş, William Chandra Tjhi,

Thea Tollersrud, Kamil Tomaszek, Sara Tonelli, Liisi Torga, Lucas Toribio, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Kira Tulchynska, Utku Türk, Francis Tyers, Sveinbjörn Hórnarson, Vilhjálmur Hörsteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uribe, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Socrates Vakirtzian, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Helena Vaz, Uliana Vedenina, Giulia Venturi, Marianne Vergez-Couret, Annemarie Verkerk, Barbora Vidová Hladká, Eric Villemonte de la Clergerie, Veronika Vincze, Anishka Vissamsetty, Natalia Vlasova, Eleni Vligouridou, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, John Wang, Jonathan North Washington, Leonie Weissweiler, Maximilian Wendt, Paul Widmer, Aleksandra Wieczorek, Shira Wigderson, Sri Hartati Wijono, Vanessa Berwanger Wille, Seyi Williams, Miriam Winkler, Shuly Wintner, Mats Wirén, Christian Wittern, Alena Witzlack-Makarevich, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Qishen Wu, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Qizhen Yang, Xiulin Yang, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Enes Yılandıloğlu, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, Annie Zhang, Larry Zhang, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, Rayan Ziane, Artūrs Znotiņš, and Eleonora Zucchini. 2025. [Universal dependencies 2.17](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

MesoTree: Annotated Linguistic Resources for Quantitative Comparative Linguistic Analysis and NLP in Mesoamerica

Robert Pugh^{1,2}

Francis Tyers^{1,2}

Robert Henderson³

¹Indiana University ²Kaltepetlahtol, A.C. ³University of Arizona

{pughrob, ftyers}@iu.edu rhenderson@arizona.edu

Abstract

One aspect of descriptive and documentary linguistic materials that is becoming increasingly important in the information age is that they be searchable, quantifiable, and comparable. In this paper, we describe an effort to create morphosyntactically-annotated corpora for a number of under-served Mesoamerican (and one Mesoamerica-adjacent) languages using Universal Dependencies. We describe the Mesoamerican linguistic area and languages involved in the project, the training and annotation process, and give a status report on the current state of the corpora. Finally, we describe a comparative syntax experiment and train UD parsing models on the data, demonstrating the usefulness of UD for facilitating quantitative, comparative linguistic research and natural language processing systems.

Keywords: treebank, syntax, Mesoamerica, Mexico, Guatemala, Nahuatl, Huave, Seri, Cuicatec, Uspan-teko

1. Introduction

Language documentation can offer a foundation for language revitalization projects (Austin, 2021), and provides linguists with valuable data with which to conduct analyses, in both a monolingual as well as comparative, multilingual settings.

Not only is it critical that language documentation materials be made accessible and available for them to be effective (an area that, fortunately, has been emphasized in most recent documentation projects). Another important factor in evaluating the usefulness of language data from the linguist's perspective is its "searchability"¹, i.e. how easily can one consult the data to find specific examples of phenomena, obtain information about their frequency, and compare with other languages.

With the wide adoption of computational tools for linguistic analysis (Mittkov, 2022; Berez-Kroeker et al., 2023), digital searchability is now more than ever an invaluable attribute of linguistic data. Searchable corpora can also positively impact endangered-language communities by facilitating example identification and the production of valuable lexical resources (Mukherjee, 2006).

Unfortunately, searchability is rarely an area of emphasis in many language documentation projects, and decisions about representation and formatting are often decided on a per-language or per-project basis, making comparative work especially difficult and time-consuming.

In the face of these realities, we propose leverag-



Figure 1: Map displaying the Mesoamerican linguistic/cultural area in Mexico and Central America.

ing the Universal Dependencies framework (Nivre et al., 2020) to support the goals of consistent data formatting and linguistically-motivated searchability and comparison.

In this paper, we describe an active effort to annotate five Mesoamerican languages² using Universal Dependencies, pulling from both existing and new texts. In an effort to capture the gamut of morphosyntactic phenomena in these languages, we purposefully set out to annotate three distinct textual genres: dictionary/grammar examples, edited prose, and spontaneous speech.

In the remainder of the paper, we (1) provide background on the Universal Dependencies project, the Mesoamerican *Sprachbünd* (See Figure 1³), and the MesoTree project, (2) describe the five lan-

¹Here, we specifically refer to the ability to search for and compare linguistic phenomena recorded in a language documentation archive. This is not to be confused with the equally important question of "discoverability" of the language documentation material itself, as discussed in, e.g. Yi et al. (2022).

²As discussed later, four of these are traditionally considered part of the Mesoamerican linguistic area, and the 5th, Seri, is included as a control language.

³Image by El Comandante, via Wikimedia Commons, based on works by AlexCovarrubias and Demis Web Map Server. Distributed under CC BY-SA 3.0.

guages included in our project, (3) highlight the project methodology, including events, outcomes, and challenges, (4) summarize the current state of the corpora and demonstrate their utility via quantitative, comparative analysis.

1.1. Universal Dependencies

The Universal Dependencies (UD) framework is a widely-adopted system for morpho-syntactic annotation by linguists, computer scientists, and others. With a stated goal of facilitating “a linguistic representation that is useful for morphosyntactic research, semantic interpretation, and for practical natural language processing across different human languages”, the annotation schema prioritizes “parallelism between similar constructions across different languages” (de Marneffe et al., 2021).

Annotations are a directed, acyclic graph of typed syntactic relations over words in a sentence, with each word including morphological information such as lemma, part-of-speech, and morphological analysis. Notably, these also constitute the levels of analysis typically used in NLP pipelines, making a UD treebank a valuable resource for training automated linguistic analysis systems.

Using an established set of tags and features with cross-linguistic support⁴, linguistic analyses become easily comparable across treebanks and languages. Consider the following examples from two Mesoamerican languages, Western Sierra Puebla Nahuatl and Uspanteko (from Sasaki (2021) and Palmer (2009), respectively):

- (1)

īwah	tiyās	n	Lwīs ?
ī-wān	ti-yā-s	n	LUIS ?
3SG.P-with	2SG.S-go-FUT	DEF	Luis ?

‘Are you going to go with Luis?’
- (2)

Xinb’ee	rik’i	wálb
x-in-b’ee	r-ik’i	w-álb
PERF.A1S-go	E3S-with	E1S-father.in.law

‘I went with my father in law.’

Both examples contain an intransitive motion verb with subject agreement, a relational noun (commonly used in Mesoamerican languages) meaning “with”, and a noun specifying the possessor of the relational noun. However, this parallel structure is not clear from the original glossed text. The motivations for the different glossing standards are language specific, and other discrepancies may understandably stem from differences in data source and reflect distinct time periods or grammatical traditions. In Figure 2, showing a UD

analysis of the two examples, the linguistic structures within the data become much more easily accessible to comparative analysis.

Due to its thorough yet flexible set of guidelines for annotating text in a way that is both sufficient for language specificity while also cross-linguistically consistent, the UD framework has garnered a great deal of support due to its ease-of-use for quantitative linguistic analyses (Kiss and Thomas, 2019; Tyers and Henderson, 2021), cross-linguistic comparisons (Naranjo and Becker, 2018; Levshina, 2019), and utility for training natural language processing (NLP) pipelines to automatically annotate new input texts. At present there are 339 UD treebanks for 186 languages.

1.2. Mesoamerica as a Linguistic Area

Mesoamerica, a region typically defined as spanning from central Mexico in the North through Central America to the South, has long been recognized as an area of large-scale cultural diffusion (Kirchhoff, 1952). Campbell et al. (1986) systematically investigate the languages spoken in the area and propose its classification as a linguistic area (*Sprachbund*) given the large number of linguistic features that appear to be shared by diverse, geographically-adjacent languages through language and cultural contact.

Subsequently, a number of linguists have affirmed the validity of Mesoamerica as a linguistic area and shown a great deal of interest in which features define it, which languages should be included or excluded from the area, and the processes of feature diffusion (Munro, 2017; Brown, 2011). Despite ample linguistic work in the area, research in Mesoamerican linguistics unfortunately often relies on disparate resources such as grammars or small, single language corpora. By creating UD treebanks for these languages in areal contact, we can more easily represent features of the Mesoamerican linguistic area that have traditionally been difficult and time-consuming to quantify and compare across disparate grammars.

1.3. The MesoTree Project

This paper describes an ongoing effort to create annotated morphosyntactic corpora, both by annotating existing published texts and by the creation of novel texts for annotation, for four Mesoamerican languages (Cuicatec, Huave, Nahuatl, and Uspanteko) and one Mesoamerica-adjacent language (Seri) in order to better facilitate comparative and quantitative areal linguistics research, natural language processing research on an under-represented language group, and the development of language technology applications for these languages.

⁴Note, however, that while there is a large set of shared morphological features, language-specific features are also allowed.

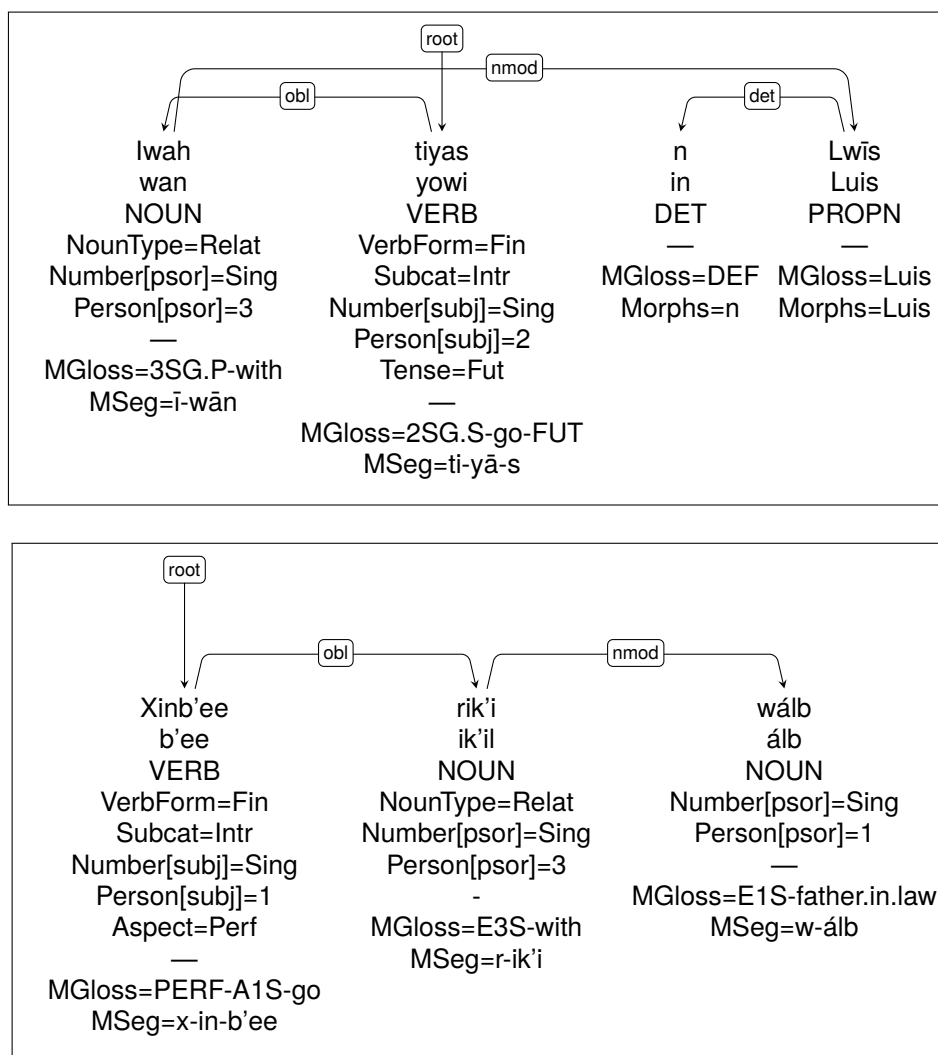


Figure 2: Examples of the UD dependency trees for the sentences in (1) (Top) and (2) (Bottom) using the Universal Dependencies framework. Note that the dependency representations encode not only the same morphological information as the interlinear glosses, but also provide syntactic relationships between the words, facilitating the identification of parallelism between these structures. In both sentences, a possessed noun with a 3rd-person possessor is annotated as an oblique (*obl*) child of the verbal *root*, and a modifying nominal dependent (*nmod*). Note the use of the language-specific morphological feature *NounType=Relat* to capture the unique morphosyntax of Mesoamerican relational nouns.

1.4. UD Treebanks for Languages of the Americas

The UD project includes a large and rapidly-growing set of treebanks in diverse languages, but has until relatively recently consisted of primarily Eurasian languages. Recent years have seen an increase in the representation of indigenous languages of the Americas in UD, including treebanks for Mbyá Guaraní (Thomas, 2019), Nheengatu (de Alencar, 2024), Ika⁵ and a number of very small corpora (less than 2,000 tokens) for various languages of Brazil (Akuntsu, Apurina, Guajajara, Kaapor, Karo, Madi, Makurap, Munduruku, Paumari, and

Tupinamba) and Paraguay (Guaraní). Karson and Coto-Solano (2024) describe work on a Universal Dependencies-annotated treebank, specifically morphological tagging, for Bribri. Tyers and Henderson (2021) published the first UD treebank for a Mesoamerican language, the Mayan language K'iche', consisting of approximately 10,000 tokens. Since then, treebanks of a similar size for two varieties of Nahuatl have also been added to the UD project: Western Sierra Puebla Nahuatl (Pugh et al., 2022) and Highland Puebla Nahuatl (Pugh and Tyers, 2024). These are currently the only UD treebanks for Mexican languages. The UniMorph project (Batsuren et al., 2022), which aims to provide a cross-linguistically-robust framework for morphological annotation, also includes some

⁵https://github.com/UniversalDependencies/UD_Ika-ChibErgIS

Mesoamerican languages, including varieties of Amuzgo, Chatino (Cruz et al., 2020), Mixtec, and Otomí.

2. Languages

We focus on four genetically unrelated languages spoken in what is commonly considered Mesoamerica: Cuicatec, Huave, Nahuatl, and Uspanteko, and a Mesoamerica-adjacent language, Seri, spoken in Mexico's northern state of Sonora.

2.1. Cuicatec

Cuicatec (*dibaku*, *dbaku*, *dubaku*, ISO 639-3 *cux*) is a highly-endangered Oto-Manguean language spoken in the southern Cañada region of the state of Oaxaca, Mexico, and is the least-studied language in the Mixtecan branch. In this project we focus on Tepeuxila Cuicatec (*dbaku*), spoken in the communities around San Juan Tepeuxila by approximately 8,680 people. There is also substantial internal variation within the variant described as Tepeuxila Cuicatec, a fact that we discuss in Section 3. Relevant existing grammatical work includes the grammar of Belmar (1902), Davis and Walker (1955)'s study of the morphology of the Concepción Pápalo variety, Bradley (1991)'s study of the syntax of Concepción Pápalo, and a more recent study of tonal morphology of the Santa Maria Pápalo variety (Bradley, 1991) and a tonal study of San Juan Tepeuxila variety (San Giacomo Trinidad, 2017). The morphosyntax of Tepeuxila Cuicatec is largely underdocumented.

2.2. Huave

Huave (ISO 639-3 *huv*) is a language isolate consisting of four languages spoken in the Isthmus of Tehuantepec in Oaxaca in the south of Mexico. INALI (2009) defines two principal languages in the grouping: Eastern and Western Huave. In this project we focus on the latter language, specifically that spoken in the community of Huave de San Mateo del Mar (*Ombeayiiüts*), which has about 11,000 speakers. Linguistic work on Huave includes a descriptive grammar (Stairs and de Hollenbach, 1981) and a dictionary (Stairs and Stairs, 1981). Some aspects of Huave syntax have been explored, such as the syntax of the noun phrase (Herrera Castro, 2010), person-number marking and valency-changing operations in the verb (Herrera Castro, 2016), but many syntactic phenomena (e.g. clausal subordination) remain undocumented and little understood.

2.3. Nahuatl

Nahuatl is a group of related languages in the Nahuan branch of the Uto-Aztecan language family. Our project is focused on the Western Sierra Puebla variety (ISO 639-3 *nhi*),⁶ spoken by approximately 17,100 people in the municipalities of Zacatlán, Ahuacatlán, and Tepetzintla, in Puebla's northern sierra. Compared to other Nahuatl variants, the Western Sierra Puebla variant has received relatively little focus from linguists. The Summer Institute of Linguistics produced an unpublished grammatical sketch and a few dozen children's stories, and Sasaki (2014) provides a minimal grammatical sketch of the communalect spoken in Ixquihucan, Ahuacatlán, as well as a dissertation about non-configurationality that draws heavily on data collected in this region (Sasaki, 2021). A 30,000-token corpus of spontaneous speech recordings and transcriptions, including word-level language annotations and spontaneous/normalized orthography pairs, was recently published for the variety (Pugh et al., 2025).

2.4. Seri

Seri (*Cmiiq̣ue iitom*, ISO 639-3 *sei*) is an endangered language isolate (Marlett, 2007), spoken by approximately 900 speakers (Ethnologue 2007 estimate) in the state of Sonora (northwest Mexico), in two villages on the coast of the Gulf of California: El Desemboque (Haxöl lihom) and Punta Chueca (Socaaix). Seri exhibits complex verbal morphology, displaying a large amount of allomorphy and a high degree of paradigmatic variety (studied in detail in Marlett (2016), a comprehensive grammar, and Baerman (2016)). In addition to the descriptive grammar, a dictionary has also been published. (Moser and Marlett, 2010). There is very little work on syntactic structure of the language beyond Marlett's excellent descriptive grammar.

2.5. Uspanteko

Uspanteko (*Tz'unun Yolooj*, ISO 639-3 *usp*), spoken natively by between 1500-4000 people (Richards and Macario, 2003) in the central highlands of Guatemala, in the city of San Miguel Uspantán and in several surrounding villages, particularly Las Pacayas (Us Maldonado, 2009), belongs to the K'ichean branch of the Mayan language family (Bennett et al., 2016; Aissen et al., 2017).

There is a descriptive grammar of Uspanteko (Can Pixabaj, 2007) and a dictionary (Méndez, 2007). The syntax of Uspanteko is essentially completely unexplored beyond basic description of topic, focus, and question movement.

⁶Alternatively Zacatlán-Ahuacatlán-Tepetzintla Nahuatl

Language	Trees	Tokens
Cuicatec	1,487	7,443
Huave	275	2,260
Nahuatl	2,132	10,466
Uspanteko	2,007	14,676
Seri	367	2,604
Total	6,268	37,449

Table 1: Current corpus volumes by trees and tokens for the five target languages. At this point, three of the five languages have either surpassed or are well on the way to completing the first phase goal of 10,000 tokens. For Nahuatl, annotators have begun annotating fiction texts. For Uspanteko, the collection of fiction texts is ongoing, and annotators have started annotating spontaneous speech transcriptions.

3. Methodology

This section summarizes the training of language expert annotators, the corpus creation process, and some of the challenges faced.

3.1. Text Collection

The project aims to annotate morphosyntactic phenomena across a diverse range of domains and registers. As such, it is broken into three phases, defined by the genre of the texts to be annotated: (1) dictionary/grammar examples, (2) edited prose/fiction, and (3) transcribed spontaneous speech. The work described in the present paper largely corresponds to the first phase. Each phase has a minimum target of 10,000 tokens.

When pre-existing text exists for a language, we use that. For the first phase, we use example sentences from dictionaries, grammars, or other pedagogical publications (Huave (Stairs and Stairs, 1981), Seri (Marlett, 2016), Nahuatl (Brockway, 1984), and Uspanteko (Méndez, 2007)).

Additionally, in some cases (Cuicatec and Nahuatl), at least some of the corpus consists of original example sentences written by language experts. In the case of Cuicatec, the entire corpus is original texts written by Cuicatec-speaking collaborators. The motivation for this decision was the fact that the pre-existing dictionary sentences initially contemplated for the annotation project are written in a different subvariety of Tepeuxila Cuicatec, different than that spoken by the Cuicatec annotators. Similarly, since there is no dictionary for Western Sierra Puebla Nahuatl, and the existing pedagogical material only contains a few hundred sentences, we supplemented this with original example sentences. In both cases, the goal was to produce texts that matched that of the other languages in

the project with respect to genre (in this case, dictionary or grammar examples). Therefore, authors were instructed to make a list of words and create an illustrative sentence using each, in effect creating dictionary example sentences.

Additionally, for both Cuicatec and Nahuatl, the project collaborators translated from Spanish a set of approximately 600 example sentences from the Archive of Indigenous Mexican languages (*Archivo de lenguas indígenas de México, ALIMG*)⁷, typically used for elicitation in documentary linguistics research. It is important to note that, although both Nahuatl annotators are speakers of the Western Sierra Puebla variety, they belong to distinct communities and have distinct subvariants/communalects. For the ALIMG sentences, we asked them each to produce a translation consistent with their community’s language variety.

For Nahuatl and Huave, finite-state morphological analyzers have been developed in (Pugh et al., 2021) and (Tyers and Herrera Castro, 2023), respectively. We thus used those, along with some disambiguation rules, to pre-annotate the corpora for those languages with lemma, UPOS, and morphological analysis.

3.2. Training and Annotation

In November, 2023, the project launched with a one-week, in-person course in Mexico City on fundamental concepts in linguistics and Universal Dependency annotation. The course was attended by between one and three language experts from each of the project’s target languages. These attendees were primarily native-speaking teachers and/or language activists interested in learning about and contributing to linguistic research for their language. The group also included two non-native-speaker linguistics students at the National Autonomous University of Mexico (UNAM) specializing in one of the five target languages. The content of the course covered fundamentals of morphology and syntax, areal linguistics and the Mesoamerican linguistic area, a deep-dive on the Universal Dependencies project, and a substantial amount of hands-on annotation work. Annotation continued after completion of the course, both asynchronously and in regularly-scheduled video-calls with the project leads. A subsequent 2-day annotation workshop was held in the Summer of 2025 in Mexico City, where the annotators for each language convened for in-person annotation and discussion. All annotation was carried out communally with final annotations decided via group consensus, not as individual efforts. As such, we do not track inter-annotator agreement.

⁷<https://cell.colmex.mx/proyecto/archivo-de-lenguas-indigenas-de-mexico/descripcion>

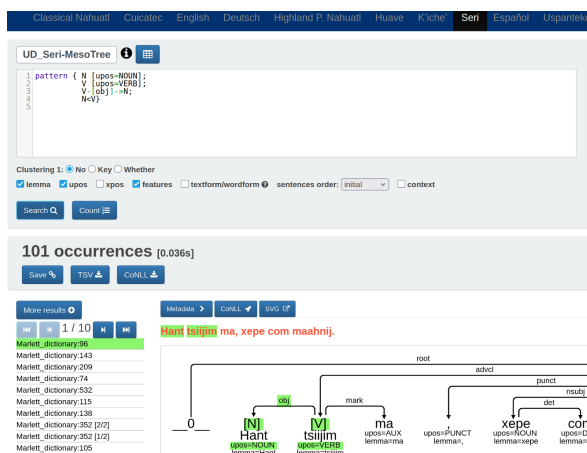


Figure 3: Screenshot of the Grew Match web interface set up with the five target languages, as well as a few others for comparison. This example shows a query for an object noun that precedes a verb.

Annotation is performed via a custom web application⁸ that combines the sentence-level annotation interface of UD Annotatrix (Tyers et al., 2017) with the ability to create and manage large-scale, multi-annotator projects offered by Arborator-Grew (Guibon et al., 2020).

3.3. Corpora

The corpus-creation efforts have focused on annotating dictionary/grammar example sentences and, after reaching 10,000 annotated tokens, will expand to increasingly more syntactically-complex sentence types, namely sentences from edited prose and spontaneous speech. Table 1 shows the number of annotated tokens and trees presently. In this case, a "complete" tree refers to, minimally, UPOS and labeled syntactic relations. For some languages, like Nahuatl and Huave, additional morphological information is also included.

The corpora are hosted on a custom instance of the Grew-Match (Guillaume, 2021) web interface, enabling annotators and other researchers to explore syntactic patterns in the corpora via a simple graph query language (See Figure 3). For example, the syntactic pattern highlighted in Figure 2 can be captured with the following query:

3.4. Source and Genre

The first phase of the project has been aimed at annotating dictionary/grammar example sentences, but there were only a small number of such examples for *nhi* and none in the appropriate subvariant of Cuicatec. Therefore, for these two languages we

⁸<https://ehecat1.cl.indiana.edu/grew-match>

```
pattern {
  RN [ upos=NOUN,
      NounType=Relat ] ;
  N [ upos=NOUN ] ;
  V [ upos=VERB ] ;
  V -[obl]-> RN ;
  RN -[nmod]-> N ;
}
```

Figure 4: Grew search pattern for identifying oblique relational nouns with a nominal complement, a pattern shown in Figure 2.

aimed to collect sentences of the "example" genre by (1) having the ALIMG language documentation sentences translated and (2) asking our language consultants to produce new sentences by making a list of words and coming up with an example sentence for each. The motivation was that this process approximates the production of dictionary examples.

In order to evaluate whether these approaches in fact produce texts with similar characteristics as true dictionary example sentences, we compare the different sourcing methods with respect to sentence length and number of clauses per sentence.

The results can be seen in Figure 5. In the left-hand plot we can see that median length varies between five and eight tokens per sentence, and the "true" dictionary/grammar sentences (in green) appear to have a wider distribution of sentence length than the annotator-produced sentences and significantly more so than the ALIMG sentences. It is interesting to note that, though *nhi* and *cux* both translated from the same set of ALIMG sentences, the length for *nhi* sentences appears to be less. This could be due to the fact that Nahuatl can have especially complex morphology, and often a multiword Spanish sentence can be translated with just a single word.

With respect to the number of clauses per sentence, there appears to be much more uniformity across languages and across sources, with most of the sentences from each source having 2 clauses or less. This suggests that, while superficially our sourcing method may produce sentences that are more homogenous with respect to sentence length when compared to actual dictionary examples, they generally match the level of syntactic complexity of the actual dictionary/grammar examples, making the method viable for cross-linguistic comparative syntactic analysis.

3.5. Challenges

For some of the languages, we experienced unexpected challenges that delayed progress. For example, as briefly mentioned in Section 3.1, the

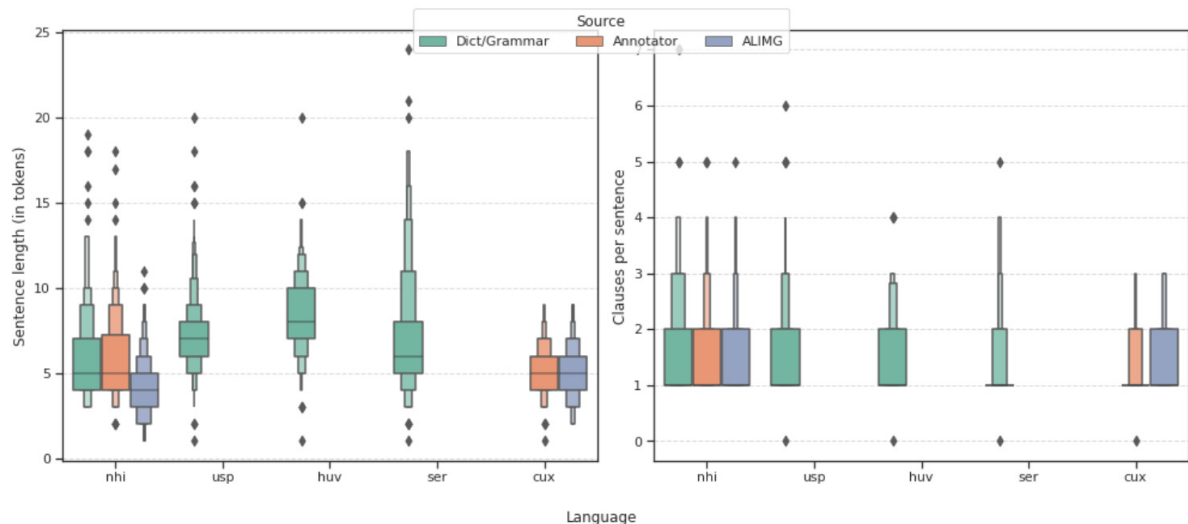


Figure 5: A comparison of the different sources in the "examples" genre. When it was available (for *usp*, *sei*, and *huv*), we used example sentences from a dictionary. In the case of *nhi*, we used sentences from a pedagogical phrase book. For both *nhi* and *cux*, no dictionary examples are available, and we used (1) invented example sentences from the annotators and (2) translations of the ALIMG language documentation sentences.

sentences that were collected initially for Cuicatec ended up being from a different subvariety of Tepeuxila Cuicatec than that of the speakers working on the project. Furthermore, our Cuicatec collaborators were unfamiliar with the orthography used in the dictionary. As a result, it was more time-effective for the speakers to propose sentences in their own variety, written according to their preferred orthographic norms rather than for them to translate the previously collected sentences. We considered automatic conversion of the orthography, but it was complicated by a number of factors, including distinct methods of encoding nasalization and tone.

In general, a major challenge for the project has been unstable access to electricity and/or internet for the annotators living in rural and remote areas of Mexico. Though our annotation platform does not require frequent server-side communication (only to load a treebank and save annotations), it does rely on an internet connection to save progress. Annotation video conferences were frequently ended abruptly due to poor connection or sudden electricity/internet outages.

This issue was particularly prominent for our Huave collaborators in Oaxaca. In an effort to help relieve this problem, we hosted multiple in-person annotation workshops in areas with reliable connection, lasting between one and three days each. In the case Huave, this is where the majority of annotation progress was made.

Progress on Seri annotation has also faced a number of difficulties. Given the small speaker population (900 total, about 300 in the community

in which we have contacts), it is challenging to find language experts able to dedicate sufficient time to annotation. Our first annotator is a local bilingual school teacher who unexpectedly had to do national service supporting Seri education. Despite this, she was able to make significant progress upon completing national service.

4. Analysis

In this section we demonstrate two potential use cases for the MesoTree corpora. First, we use the corpora to perform a straightforward quantitative analysis and comparison of syntactic tendencies across (a subset of) Mesoamerican languages, comparing the word-order variability of subjects and objects in our five treebanks along with three pre-existing treebanks of Mesoamerican languages. Next, we examine the utility of the MesoTree treebanks for training NLP systems by training and evaluating multi-task UD parsing models on three of the five languages.

4.1. Word-order Entropy

An important intended use case for the MesoTree treebanks is the facilitation of comparative syntactic analysis. Here, we explore one example of such analysis, using "word-order entropy" (Levshina, 2019). We use Shannon's Entropy (Equation 1), the average amount of information required to describe the events of a random variable, to quantify the word-order variability of clauses with overt subjects and objects.

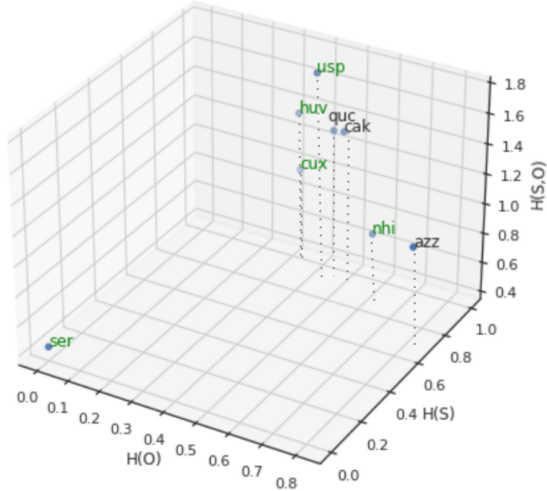


Figure 6: A plot of Mesoamerican languages with annotated corpora as a function of the entropy of nominal subject and object order. Languages whose datasets were created as a part of the presently-described project are labeled in green.

$$H(X) := - \sum_{x \in \mathcal{X}} p(x) \log p(x) \quad (1)$$

Specifically, we measure the entropy of (1) verb-subject order, (2) verb-object order, and (3) verb-subject-object order. In this case, higher entropy indicates greater syntactic unpredictability.

For thoroughness, we include three previously-annotated Mesoamerican treebanks (see Section 1.4) in the analysis. Specifically, in addition to the five target-language corpora described in Section 3.3, we include corpora for K’iche’ (Tyers and Henderson, 2021) and Highland Puebla Nahuatl (Pugh and Tyers, 2024), each totaling approximately 10,000 tokens, as well as a small, unpublished corpus of Kaqchikel (approximately 1,000 tokens) created by one of the authors. We combine the existing Western Sierra Puebla Nahuatl treebank (Pugh et al., 2022) with our project’s annotations for the same language.

The results are listed in Appendix A and visualized in Figure 6. There is an apparent pattern among the Mesoamerican languages, with most having verb-subject entropy near 1.0 (maximally unpredictable, or effectively free order), and slightly more predictable orders with explicit objects. There is much more variance along the Subject-Object order entropy. The Mesoamerican-adjacent control language, Seri, has much more consistent word orders and is a clear outlier.

It is also note-worthy that the three Mayan languages, Uspanteko (usp), K’iche’ (quc), and Kaqchikel (cak), and the Nahuatl languages Western Sierra Puebla Nahuatl (nhi) and Highland

Puebla Nahuatl (azz) seem to form groupings. The fact that the Nahuatl languages appear slightly separated from the remaining Mesoamerican languages could be due to genre: The azz treebank and the older nhi treebank (which was combined with the MesoTree treebank) contain both edited prose and a large number of spontaneous speech transcriptions. Once the treebanks for all three genres are completed, we plan to re-run this analysis to analyze both linguistic and genre-based factors.

4.2. Training Automatic UD-Parsing Models

Lang	Lemma	UPOS	UAS	LAS
usp	0.94	0.94	0.81	0.75
nhi	0.84	0.94	0.86	0.81
cux	-	0.90	0.86	0.76

Table 2: Comparison of Lemmatization, UPOS tagging, and dependency parsing for the Uspanteko (usp), Western Sierra Puebla Nahuatl (nhi), and Tepeuxila Cuicatec datasets. Since the morphological analyzer for Cuicatec is still under development, the lemma annotation is incomplete, and we omit lemma accuracy.

Finally, given our stated interest in leveraging the MesoTree corpora for building NLP systems for Mesoamerican languages, we train and evaluate UD parsing models the three languages with the most annotated data, usp, nhi, and cux.

We use MaChAmp (van der Goot et al., 2021) to fine-tune multilingual BERT (mBERT) embeddings⁹ on each UD task. The model leverages multi-task learning, such that all of the tasks share encoder parameters, but each has its own unique decoder: a transformation-rule classifier (Straka, 2018) for lemmatization, a softmax layer on the contextual embeddings for part-of-speech tagging and morphological analysis, and a deep biaffine parser for dependency parsing (Gardner et al., 2018).

For the experiment, we split each treebank randomly into a train (85% of the sentences) and eval (15% of the sentences) set.

5. Concluding Remarks

The MesoTree project aims to produce high-quality linguistically-annotated corpora that enable and encourage computer-aided linguistic analyses and language technology development for Mesoamerican languages. Though work on the project is still underway, the value of producing consistently-annotated corpora across multiple unrelated lan-

⁹We use the bert-base-multilingual-cased model.

guages in areal contact is already apparent, as demonstrated by the quantitative comparative word-order analysis and the training of automatic parsing models described in this paper.

Future work will focus on expanding these corpora to include more complex registers, such as edited prose and spontaneous speech, ultimately providing a foundation for language technology that serves both the global research community and the speaker communities.

Acknowledgements

We would like dedicate this work to our dear colleague, Debora Perales Morales, a linguist and speaker of the Seri language, who passed away unexpectedly this fall from sudden illness. She was responsible for the Seri annotations described in this paper, but what a small thing compared to her tireless advocacy for her language and her community. It is hard to know how to move forward having lost such a great mind. We must find a way in her honor.

We also must recognize NSF grant #2319246/2319247 *Collaborative Research: Syntactically-annotated corpora for endangered languages in areal contact* for supporting this work.

6. Bibliographical References

- Judith Aissen, Nora C England, and Roberto Zavala Maldonado. 2017. *The Mayan languages*. Taylor & Francis.
- Peter K. Austin. 2021. *Language Documentation and Language Revitalization*, page 199–219. Cambridge University Press.
- Matthew Baerman. 2016. *Seri verb classes: Morphosyntactic motivation and morphological autonomy*. *Language*, 92(4):792–823.
- Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, et al. 2022. UniMorph 4.0: universal morphology. *arXiv preprint arXiv:2205.03608*.
- Francisco Belmar. 1902. *El Cuicateco-Primera Edición*.
- Ryan Bennett, Jessica Coon, and Robert Henderson. 2016. Introduction to Mayan linguistics. *Language and Linguistics Compass*, 10(10):455–468.
- Andrea L. Berez-Kroeker, Shirley Gabber, and Aliya Slayton. 2023. *Recent advances in technologies for resource creation and mobilization in language documentation*. *Annual Review of Linguistics*, 9(1):195–214.
- David P Bradley. 1991. A preliminary syntactic sketch of Concepción Pápalo Cuicatec. *Studies in the syntax of Mixtecan languages*, 3:409–506.
- Earl Brockway, editor. 1984. *Frases en mejicano y español: En el idioma náhuatl de San Miguel de Tenango, Puebla y en español*, primera edición edition. Instituto Lingüístico de Verano, A.C., Mexico.
- Cecil Brown. 2011. *The Role of Nahuatl in the Formation of Mesoamerica as a Linguistic Area*. *Language Dynamics and Change*, 1:171–204.
- Lyle Campbell, Terrence Kaufman, and Thomas C Smith-Stark. 1986. Meso-america as a linguistic area. *Language*, pages 530–570.
- Telma Angelina Can Pixabaj. 2007. *Gramática uspanteka*. Cholsamaj Fundacion.
- Hilaria Cruz, Antonios Anastasopoulos, and Gregory Stump. 2020. *A resource for studying chatino verbal morphology*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2827–2831, Marseille, France. European Language Resources Association.
- Marjorie Davis and Margaret Walker. 1955. Cuicateco: Morphemics and morphophonemics. *International Journal of American Linguistics*, 21(1):46–51.
- Leonel Figueiredo de Alencar. 2024. *A Universal Dependencies treebank for nheengatu*. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, pages 37–54, Santiago de Compostela, Galicia/Spain. 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.
- Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. *AllenNLP: A Deep Semantic Natural Language Processing Platform*. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

- Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. [When collaborative tree-bank 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.
- Samuel Herrera Castro. 2010. Alineamiento y frase verbal en Huave de San Mateo Del Mar, Oaxaca. Master's thesis, Escuela Nacional de Antropología e Historia, México, D.F.
- Samuel Herrera Castro. 2016. *Sintaxis y semántica de la frase nominal en Huave de San Mateo Del Mar, Oaxaca*. Ph.D. thesis, Colegio de México, México, D.F.
- INALI. 2009. *Catálogo de las lenguas indígenas nacionales: Variantes lingüísticas de México con sus autodenominaciones y referencias geográficas*. Instituto Nacional de Lenguas Indígenas, México, D.F.
- Jessica Karson and Rolando Coto-Solano. 2024. [Morphological tagging in Bribri using Universal Dependency features](#). In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 56–66, Mexico City, Mexico. Association for Computational Linguistics.
- Paul Kirchhoff. 1952. Mesoamerica: its geographic limits, ethnic composition and cultural characteristics. *Heritage of conquest*, pages 17–30.
- 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.
- Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23(3):533–572.
- Stephen A. Marlett. 2007. Las relaciones entre las lenguas “hokanas” en México: ¿cuál es la evidencia? *Memorias del III Coloquio Internacional de Lingüística Mauricio Swadesh*, pages 165–192.
- Stephen A Marlett. 2016. Cmiiq̃ue litom: the Seri language. *Unpublished grammar (2016 draft)*.
- Miguel Angel Vicente Méndez. 2007. *Cholaj Tz'ib'al li Uspanteko: Diccionario bilingüe uspanteko-español*. Cholsamaj.
- Ruslan Mitkov. 2022. *The Oxford Handbook of Computational Linguistics*. Oxford University Press.
- Mary B. Moser and Stephen A. Marlett. 2010. *Comcaac quih yaza quih hant ihiip hac: Diccionario seri-español-inglés*, 2nd edition. Plaza y Valdes editores and Universidad de Sonora.
- Joybrato Mukherjee. 2006. Corpus linguistics and language pedagogy: The state of the art—and beyond. *Corpus technology and language pedagogy: New resources, new tools, new methods*, pages 5–24.
- Pamela Munro. 2017. The Mesoamerican linguistic area revisited. *Language contact and change in Mesoamerica and beyond*, pages 335–351.
- Matías Guzmán Naranjo and Laura Becker. 2018. Quantitative word order typology with UD. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), December 13–14, 2018, Oslo University, Norway*, 155, pages 91–104. Linköping University Electronic Press.
- J. Nivre, M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, and D. Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4027–4036.
- Alexis Mary Palmer. 2009. Semi-automated annotation and active learning for language documentation.
- Robert Pugh, Marivel Huerta Mendez, Mitsuya Sasaki, and Francis Tyers. 2022. Universal dependencies for Western Sierra Puebla Nahuatl. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5011–5020.
- Robert Pugh, Francis Tyers, and Marivel Huerta Mendez. 2021. Towards an open source finite-state morphological analyzer for Zacatlán-Ahuacatlán-Tepetzintla Nahuatl. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 80–85.
- Robert Pugh and Francis M. Tyers. 2024. A Universal Dependencies Treebank for Highland Puebla Nahuatl. In *2024 Annual Conference of the North*

American Chapter of the Association for Computational Linguistics.

Robert Pugh, Cheyenne Wing, María Ximena Juárez Huerta, Ángeles Márquez Hernandez, and Francis Tyers. 2025. [Ihquin tlahtouah in tetelahtzincocah: An annotated, multi-purpose audio and text corpus of western sierra Puebla Nahuatl](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3549–3562, Albuquerque, New Mexico. Association for Computational Linguistics.

Michael Richards and Narciso Cojtí Macario. 2003. *Atlas lingüístico de Guatemala*. Editorial Serviprensa Guatemala City.

Marcela San Giacomo Trinidad. 2017. The phonetics and phonology of tone in San Juan Tepeuxila Cuicatec. *Cuadernos de Lingüística de El Colegio de México*, 4(2):83–136.

Mitsuya Sasaki. 2014. A dialectological sketch of Ixquihuacan Nahuatl. *Tokyo University Linguistic Papers*, 35:e139–170.

Mitsuya Sasaki. 2021. *Configurationality in Ixquihuacan Nahuatl*. Ph.D. thesis, University of Tokyo.

Emily F. Stairs and Elena E. de Hollenbach. 1981. Gramática huave. In Glenn A. Stairs Kreger and Emily F. Scharfe de Stairs, editors, *Diccionario huave de San Mateo del Mar*, pages 283–391. Instituto Lingüístico de Verano, México, D. F.

Glen Stairs and Emily Stairs. 1981. *Diccionario huave de San Mateo del Mar*. Instituto Lingüístico de Verano, México, D.F.

Milan Straka. 2018. Udpipes 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.

Guillaume Thomas. 2019. Universal dependencies for Mbyá Guaraní. In *Proceedings of the third workshop on universal dependencies (udw, syntaxfest 2019)*, pages 70–77.

Francis Tyers and Robert Henderson. 2021. A corpus of K’iche’ annotated for morphosyntactic structure. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 10–20.

Francis Tyers, Mariya Sheyanova, and Jonathan 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.

Francis M. Tyers and Samuel Herrera Castro. 2023. Towards a finite-state morphological analyser for San Mateo Huave. In *Proceedings of the 6th Workshop on Computational Methods for Endangered Languages*. [to appear].

Juan Antonio Us Maldonado. 2009. *Monografía Uspanteka*. Comunidad Lingüística Uspanteka, San Miguel Uspantán.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. [Massive Choice, Ample Tasks \(MaChAmp\): A Toolkit for Multi-task learning in NLP](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Irene Yi, Amelia Lake, Juhya Kim, Kassandra Haakman, Jeremiah Jewell, Sarah Babinski, and Claire Bown. 2022. [Accessibility, discoverability, and functionality: An audit of and recommendations for digital language archives](#). *Journal of Open Humanities Data*.

A. Shannon’s Entropy

The entropy calculations for the 8 treebanks (seven for Mesoamerican languages and one Mesoamerica-adjacent) can be seen in Tabl 3.

Language	H(O)	H(S)	H(S,O)
azz	0.81	0.68	1.01
cak	0.48	0.95	1.35
cux	0.32	1.00	0.96
huv	0.32	1.00	1.35
nhi	0.61	0.89	0.79
quc	0.44	0.98	1.31
sei	0.0	0.0	0.44
usp	0.42	0.90	1.74

Table 3: Entropy calculations for the 8 treebanks (seven for Mesoamerican languages and one Mesoamerica-adjacent), corresponding to the visualization in Figure 6

Nonprototypical Predication and Nonpredicational Clauses in Universal Dependencies

Joakim Nivre, William Croft, André V. Lopes Coneglian

Uppsala University, University of New Mexico, Federal University of Minas Gerais
joakim.nivre@lingfil.uu.se, wacroft@icloud.com, coneglian@ufmg.br

Abstract

To assess whether the framework of Universal Dependencies (UD) is compatible with findings from linguistic typology, we need to systematically review how UD represents linguistic constructions and how it handles the range of morphosyntactic variation attested across languages. In this paper, we present such a review focusing on nonprototypical predication and nonpredicational clauses. We find that, while nonprototypical predication is generally handled well in the UD framework, nonpredicational clauses are not discussed as such in the guidelines and often have to be annotated in a way that does not reflect their special information packaging functions. We briefly discuss ways in which the UD framework could be extended in order to better capture these functions.

Keywords: Universal Dependencies, nonprototypical predication, nonpredicational clauses

1. Introduction

Universal Dependencies (UD) is an annotation framework for morphosyntax, designed to be applicable to all human languages and to enable meaningful cross-linguistic comparisons (Nivre et al., 2016, 2020; de Marneffe et al., 2021). To check if UD meets these requirements, Nivre (2025) proposes to build a constructionicon for UD based on the survey of universal constructions and morphosyntactic 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. For example, the relative clause construction is defined as any construction in any language where an action concept modifies a referent. English *the book I'm reading* is an instance of the relative clause construction. The externally-headed strategy describes a relative clause construction in which the head noun is external to the modifying clause. The English example is an instance of this narrower category as well.

To build a constructionicon for UD, we need to provide a UD analysis of every construction-strategy pair, and this paper lays the ground for such a resource by reviewing the UD guide-

lines for a specific set of constructions – nonprototypical predication and nonpredicational clauses – constructions that are discussed in Chapters 10–11 of Croft (2022). This complements earlier studies on reference and modification (Nivre and Croft, 2025) and verbal predication (Croft and Nivre, 2025), as well as concurrent studies on complex predicates (Coneglian et al., 2026) and speech act constructions (Nivre et al., 2026).

The first goal of the investigation is to check whether the UD guidelines cover the main attested strategies for each construction, and to provide new recommendations in case of gaps. The second goal is to find out whether the annotations of different strategies reveal that they are realizations of the same construction, which facilitates cross-linguistic comparison, and to discuss alternative analyses when this is not the case. We are aware that UD was not originally designed to capture constructions in Croft's sense, but trying to capture both constructions and strategies seems compatible with the view of UD as a “mixed functional-structural system” for annotation.¹ Finally, it is worth noting that our review focuses on the UD guidelines, not on the actual annotations in existing UD treebanks, although we occasionally make observations on treebanks when

¹<https://universaldependencies.org/u/overview/syntax.html#the-taxonomy-of-typed-dependencies>

relevant. A systematic review of existing treebanks would be very valuable but is outside the scope of this project.

2. Nonprototypical Predication

Nonprototypical predication is the predication of concepts other than the action concepts associated with verbal clause constructions, previously discussed in [Croft and Nivre \(2025\)](#). This includes at least the following three types of predication ([Croft, 2022](#), p. 289):

- Object predication: *Ann is an engineer*
- Property predication: *Sandy is clever*
- Locative predication: *the cat is on the mat*

There are four main morphosyntactic strategies used for nonprototypical predication in the world's languages. First, the *verbal* strategy consists in recruiting the morphosyntactic encoding of prototypical verbal predication for nonprototypical predication, as illustrated in example (1), from Moru ([Tucker, 1940](#); [Stassen, 1997](#)), which compares (a) action predication and (b) property predication.

- (1) a. anyá k-ódra
3PL 3PL.IMPF-die
'they are/were dying'
- b. lédr aná k^wóizi
man that 3SG.IMPF-bad
'that man is bad'

The indexation on the predicate is identical in the two cases, even though the predicate stem in (b) may be used for nominal reference in other contexts.

Second, the *verbal copula* strategy combines the nonprototypical predicate with an independent word inflected like a verb. It is commonly used for locative predication, with a copula derived from a locational or posture verb, which can also be extended to other types of nonprototypical predication. This is exemplified in (2), from Amele ([Roberts, 1987](#); [Stassen, 1997](#)), for (a) locative predication and (b) property predication.

- (2) a. uqa jo na bil-i-a
he house at sit-3SG-PRS
'he is in the house'
- b. uqa me bil-i-a
he good sit-3SG-PRS
'he is well'

Third, the *nonverbal copula strategy* instead combines the nonprototypical predicate with an uninflected form, most typically a pronoun (demonstrative or personal), a topic or focus marker, or some other grammaticalized form. Example (3) shows object predication with a nonverbal copula derived from a demonstrative in Nakanai ([Stassen, 1997](#), p. 149), and example (4) shows object predication with a nonverbal copula derived from a focus marker in Awtuw ([Johnston, 1980](#); [Stassen, 1997](#)).

- (3) eia la taua sesele
3SG DEM spirit truly
'he is truly a spirit'
- (4) wan po rumeyæn
1SG FOC human-being
'I am a human being'

Fourth, the *zero strategy* simply juxtaposes the subject nominal with the nonprototypical predicate without any inflection or linking element. This is illustrated in (5), from Tiwi ([Osborne, 1974](#); [Stassen, 1997](#)), for (a) object predication and (b) property predication.

- (5) a. purukupaɽli maɽtina
parukuparli boss
'Parukuparli is boss'
- b. tuŋkwaltiriŋa pumpuka
stringy-bark good
'a string bark is good'

Both the zero and the nonverbal copula strategies typically originate in object predication.

2.1. UD Annotation

The UD annotation of a nonprototypical predication construction connects the predicate to its argument with a nominal subject (*nsubj*) relation. This means that, as far as the predicate-argument structure is concerned, nonprototypical predication is treated in the same way as verbal clause constructions, more precisely as intransitive clause constructions (cf. [Croft and Nivre, 2025](#)). Different realizations of such constructions are further differentiated by the annotation of the predicate, which mainly depends on the type of construction, and by the annotation of copulas or other components of the strategy used to encode the predication. Let us now discuss each of these in turn.

In object predication, the predicate is a nominal expression denoting an object category,

Construction	UD Annotation
Object predication	PRED = NOUN
Property predication	PRED = ADJ
Locative predication	PRED = ADV
	PRED = NOUN[Case=Obl]
	PRED = NOUN $\xrightarrow{\text{case}}$ ADP
Strategy	UD Annotation
Zero	RE $\xleftarrow{\text{nsbj}}$ PRED
Nonverbal copula	RE $\xleftarrow{\text{nsbj}}$ PRED $\xrightarrow{\text{cop}}$ PRON
Verbal copula	RE $\xleftarrow{\text{nsbj}}$ PRED $\xrightarrow{\text{cop}}$ AUX
Verbal	RE $\xleftarrow{\text{nsbj}}$ VERB (?)

Table 1: UD annotation of different strategies for nonprototypical predication; RE = referring expression; PRED = predicate expression.

which means that the head word typically bears the part-of-speech tag NOUN and may have dependents such as determiners and nominal modifiers. In property predication, by contrast, the head word of the predicate is normally tagged ADJ and may take admodifiers and other dependents typical of adjectives. In locative predication, finally, the predicate denotes a location, which means that it is either a locative adverb, tagged ADV, or an oblique nominal expression. The annotation of predicates in different constructions is summarized schematically in the top half of Table 1.

Turning next to the annotation of different strategies for encoding the predication, we first note that the zero strategy has no explicit annotation over and above the annotation of the predicate and the subject relation. This is illustrated in Figure 1(a), which shows the UD annotation of example (5a) from Tiwi.² For the two copula strategies, there is in addition a *cop* relation from the predicate to the copula, as seen in Figure 1(b–c), which shows the UD annotation of example (3) from Awtuw and example (2) from Amele. The only difference between these two examples is that the nonverbal copula is tagged PRON, while the verbal copula is tagged AUX. For the verbal strategy, finally, there are currently no explicit guidelines in UD, but the general guide-

lines for part-of-speech tagging seem to imply that the predicate should be tagged VERB because of the verbal morphology. This would make the annotation the same as for intransitive clauses, as seen in Figure 1(d), which shows the likely UD annotation of example (1b) from Moru. This would also mean that the VERB tag in this case “overrides” the tag implied by the construction, as shown in the bottom half of Table 1, which summarizes the annotation of different strategies.

In conclusion, it is clear that UD can handle all common strategies for nonprototypical predication using the *nsbj* argument relation together with the optional *cop* relation. In this way, the annotation also captures the constructional similarity across different strategies, in contrast to frameworks that treat copula verbs as heads. It is not clear, however, whether the annotation of the verbal strategy can distinguish nonprototypical predication from the intransitive clause construction.

3. Nonpredicational Clauses

In terms of information packaging, predicational clauses are topic-comment constructions, constructions that add new information (the comment) about one or more existing discourse referents (the topic). In terms of the file metaphor (Heim, 1983; Stassen, 1997), the predicational information packaging adds content to an existing file for a referent. However, languages also contain clause constructions that are based on other types of information

²In this and all following examples, we simplify the UD representations by omitting (a) lemmas and (b) morphological features that are not relevant for the discussion. For reasons of space, we do not repeat glosses and translations in the examples.

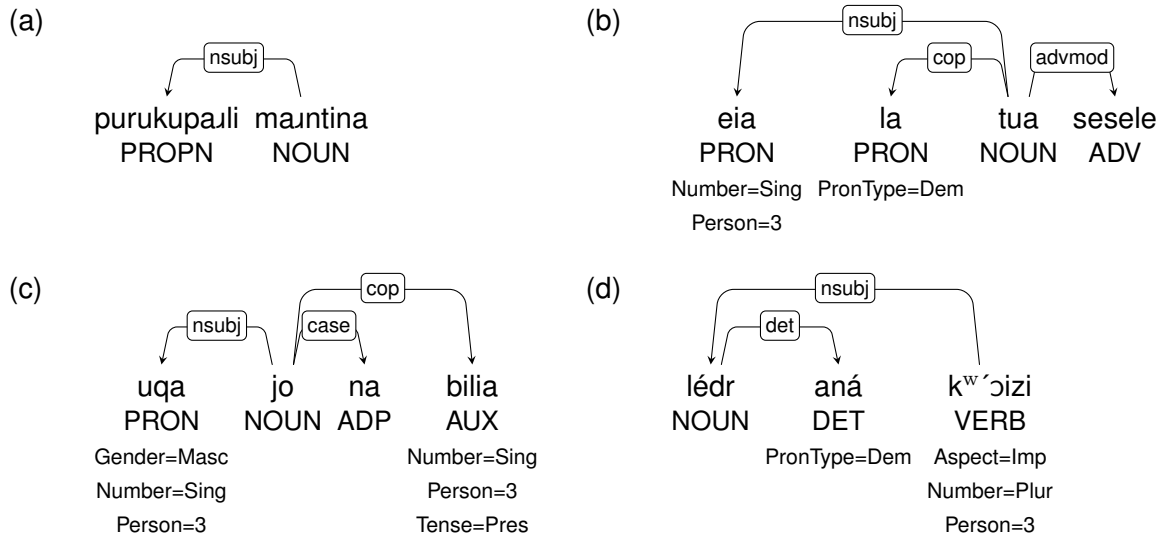


Figure 1: Simplified UD annotation of nonprototypical predication constructions/strategies: (a) object/zero, (b) object/nonverbal copula, (c) locative/verbal copula, (d) property/verbal.

packaging (Lambrecht, 1994).

In *thetic* clauses, there is no split between topic and comment, because all the information in the clause is new. A typical example is a presentational construction like *there's a cat*. In *identificational* clauses, two different characterizations are presented as pertaining to the same entity or event, as in the equational construction *the butler was the murderer*. In terms of the file metaphor, *thetic* constructions create a new file, while *identificational* constructions merge two existing files. We refer to *thetic* and *identificational* clauses collectively as *nonpredicational* clauses.

Nonpredicational constructions and strategies can be quite diverse and have not been studied as extensively as predicational clauses in linguistic typology. Our discussion will therefore be limited to the most important types discussed in Croft (2022). However, before we turn to these we will discuss a few topic-comment constructions that differ from standard predicational clause constructions.

3.1. Topic-Comment Constructions

In basic predicational clauses, the grammatical subject denotes the most topical (or salient) argument, but languages have a variety of mechanisms for promoting other arguments when required by the discourse context, including nonbasic voice constructions

like passive-inverse and causative constructions, discussed in Croft and Nivre (2025). Another possibility is to use a *detached* phrase strategy, as in the following Swedish example:

- (6) Lisa, henne har jag inte sett i dag
 Lisa her have I not seen in day
 'Lisa, I have not seen her today'

In this example, the detached topic phrase *Lisa* is preposed and co-referential with the object argument *henne*, but detached phrases can also be postposed, as in the following classic French example, where the detached phrase is co-referential with the subject (Maslova and Bernini, 2006):

- (7) ils sont fous, ces romains
 they are mad those Romans
 'They're mad, those Romans'

A special kind of topic construction is the *hanging topic phrase*, where the topic does not correspond to an argument at all. This may be realized with an overtly coded strategy, as in example (8) from Japanese (Iwasaki, 2013), or with a zero-coded strategy, as in example (9) from Chinese (Li and Thompson, 1981).

- (8) zoo wa karada ga okii
 elephant TOP body NOM big:NPST
 'the elephant – its body is big'
- (9) xiàng bizi cháng
 elephant nose long
 'elephants' noses are long'

Hanging topics can be topics without being arguments because the sentence can be construed as about the referent of the hanging topic phrase, typically in virtue of a relationship like possession, exemplified in (8) and (9), or a figure-ground relation.

3.2. Thetic Constructions

The main characteristic of thetic constructions is that there is no division between topic and comment; all the information is new, including the referents. Among the clearest examples are *presentational* constructions, where the function is to open a discourse file for a referent, and possibly also to provide some information about this referent to the hearer. A typical case is the *presentational location* construction in (10), which introduces the cat as a new referent and contrasts with the *predicational location* construction in (11), where the existence of the cat is presupposed.³

(10) there is a cat on the mat

(11) the cat is on the mat

Presentational constructions are entity-central (Sasse, 1987), but there are also thetic constructions that are event-central, such as *weather* constructions, like *it's raining*.

The strategies for thetic constructions are varied but have in common that they make the phrase expressing the most topical referent look less like a subject and/or make the form expressing the event look less like a predicate (Croft, 2022, p. 339). It is very common to use *prosody* or *word order* to differentiate thetic constructions from otherwise similar predicational constructions, for example, by postposing the subject to the verb in SVO languages, as in *along comes a farmer with a donkey*.

Presentational constructions are often introduced by grammaticalized markers, as illustrated in the following examples:

- (12) a. there is a cat on the mat
b. det sitt-er en katt på mattan
it sit-PRS a cat on mat-DEF.SG

³A similar contrast is found in possession clauses, where the *presentational possession* construction *she has a book* introduces a new referent of the book type, while the *predicational possession* construction *the book is hers* presupposes such a referent.

‘there is a cat on the mat’

- c. hay un gato en el tapete
have.3SG.prs a cat on the mat
‘there is a cat on the mat’

In the English example (12a), the locative expression *there* is a fixed marker of the presentational construction with the referring phrase postposed to the verb. In the Swedish example (12b), the third person neuter pronoun *det*, which has a variety of pleonastic uses in the language, is combined with the locational verb *sitta* (sit) to introduce the new referent. The Spanish example in (12c) represents another common pattern, namely the recruitment of a presentational possession construction to express presentational location, using the verb *hay*, an impersonal form of *haber* (have).⁴

A more complex strategy for thetic constructions is the *split* strategy (Sasse, 1987), where the encoding of the most prominent referent as less subject-like is combined with the predicate being in subordinate form. This is illustrated with a French example in (13) (Wehr, 1984; Sasse, 1987), where a typical presentational strategy is combined with a predicate expressed in the form of a relative clause.

- (13) voilà la sirène qui hurle
there-is the siren which wails
‘the SIREN is wailing’

Weather constructions, finally, exhibit a range of different strategies, three of which are exemplified below:

- (14) a. llueve
rain.3SG.PRS
‘it’s raining’
b. es regnet
it rain:3SG.PRS
‘it’s raining’
c. akpɔkpɔ tin
cloud exist
‘it’s cloudy’

The Spanish example (14a) uses a verb that does not take any argument at all, while the German example (14b) uses a verb together with an impersonal, non-referential pronoun in subject position. The Gungbe example (14c),

⁴Recruitment may also go in the opposite direction, so that presentational possession is expressed using a locative strategy. For more information on strategies for presentational possession, see Stassen (2009) and Croft (2022).

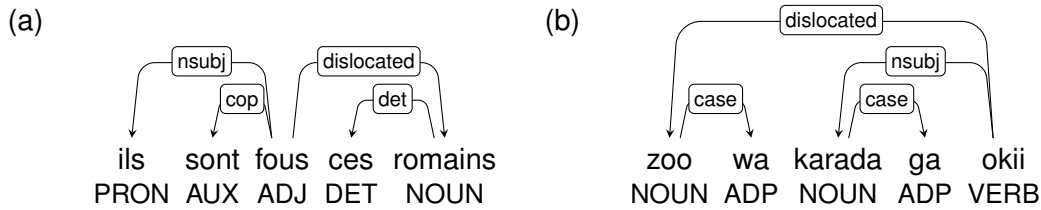


Figure 2: Simplified UD annotation of topic-comment constructions: (a) detached, (b) hanging topic phrase.

from Eriksen et al. (2012), instead uses a noun that combines with a general verb of existence.

3.3. Identificational Constructions

Identificational constructions divide the information in a clause in two, but instead of providing new information about a known topic, they equate two semantic components known from different descriptions. A typical case is the *equational* construction in (15), where the new information is not the existence of the butler or the murderer but the fact that the two are the same.

(15) the butler was the murderer

More generally, the two parts of an identificational construction are a *presupposed open proposition* (POP) with an unknown part X and a *focus* that provides the missing part X of the POP. In the equational construction, the focus is a referring expression, but identificational constructions in general can involve any part of the information in a clause, as illustrated in the following examples from Croft (2022), where capitalization marks contrastive stress:

(16) I have ONE brother

(17) I HAVE finished my assignment

In (16), the POP is *I have X brothers* and the focus is *one* (not two or three). In (17), the unknown information is the polarity of the proposition, and the focus is a positive polarity.

The strategies for identificational constructions largely overlap with those used for other clause-level information packaging that differs from the default topic-comment packaging and again include prosody, exemplified above, and word order. A common overt morphosyntactic strategy is a *cleft* strategy, which divides the clause into two functional parts corresponding to the focus and the POP. Below are two

Swedish examples of the cleft strategy:

- (18) a. det var igår (som) hon kom
 it was yesterday (that) she came
 ‘it was yesterday (REL) she came’
 b. vad som krävs är en ny start
 what REL is-required is a new start
 ‘what is required is a new start’

In (18a) the focus *igår* (yesterday) is linked by an equational copula to the POP, which takes the form of a relative clause even though it does not have the usual modifying function. In (18b), the POP is again expressed as a subordinate clause, but in this case it precedes the focus *en ny start* (a new start).

3.4. UD Annotation

The UD annotation framework for clause constructions focuses on the analysis of predicates, arguments and modifiers. This means that, even though core grammatical relations are defined partly in terms of information packaging, the framework lacks concepts for distinguishing different information packaging functions, such as *thetic* and *identificational* functions, unless they are realized through special morphosyntactic structures. In particular, anything using only prosody or word order cannot be represented in the UD annotation. In what follows, we will therefore limit ourselves to overtly coded strategies.

Starting with special topic-comment structures, the detached strategy and the hanging topic phrases are both annotated with the *dislocated* relation, as shown in Figure 2 for examples (7) and (8). The fact that the dislocated element is co-referential with an argument in (7) but not in (8) can be captured by subtyping the *dislocated* relation with an argument relation – *dislocated:nsbj* in this example. This is done only in a small minority of UD treebanks

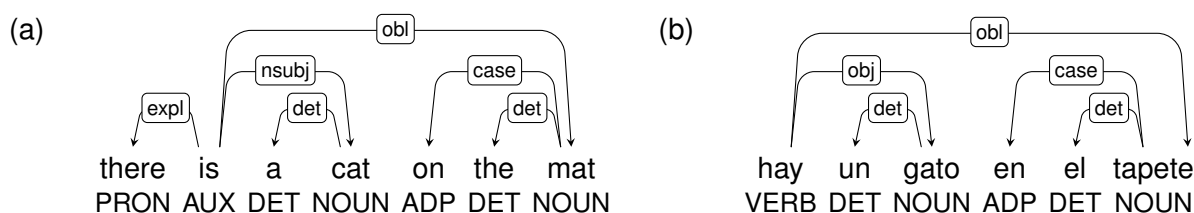


Figure 3: Simplified UD annotation of presentational location constructions.

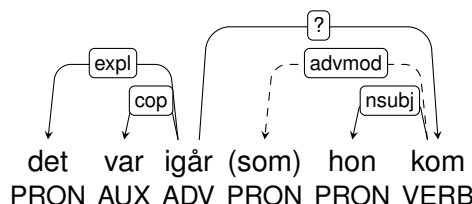


Figure 4: Simplified UD annotation of a cleft strategy.

but it could be recommended more widely.

Forthetic constructions, there are no general guidelines in UD, but the English presentational locational construction in (12a) is discussed together with nonverbal predication constructions. In this case, however, the guidelines say that the verb *be* should be treated as a main verb with existential meaning, not as a copula verb, as shown in Figure 3(a).⁵ The referring phrase is treated as the nominal subject (*nsubj*) of this verb, and the grammaticalized marker *there* is attached with the *expl* relation, which is used in UD for expletive (non-referential) pronouns appearing in argument positions.⁶ The same structure would be assigned to the Swedish example (12b), which is more expected given that the verb is the more contentful *sitta* (sit). The Spanish variant in (12c) is not explicitly discussed in the guidelines, but the current practice is to treat the referring phrase as the direct object (*obj*) of the impersonal verb *hay*, which is natural given the origin of the construction although it obscures the constructional similarity with the locational strategies used in English and Swedish. Given that the UD annotation prioritizes morphosyntactically encoded predicate-argument relations, divergences like

these seem unavoidable, and fully capturing thetic constructions would most likely require either an additional annotation layer or a more complex subtyping system that can be used to annotate constructions and strategies in parallel. Such a system would also be needed to capture the split strategy, exemplified in the French example (13), which in the current system has to be analyzed as a regular relative clause construction.

Turning finally to identificational constructions, the situation is very similar in that there are no general guidelines for these. Equational constructions are essentially treated as non-prototypical predication, which means that *the butler was the murderer* is annotated exactly like *the butler was a doctor*. This is a reasonable analysis given the available tools, although it does not reflect the difference in information packaging. A notoriously difficult case is the cleft strategy, with all its variants, which has been discussed extensively in the UD community without a clear consensus emerging.⁷ For the type of cleft exemplified by (18a), a common analysis is shown in Figure 4. This analysis treats the focus as the root of the clause, introduced by a copula verb with an expletive subject and followed by the POP in the form of a subordinate clause whose nature has been discussed extensively. Structurally, it

⁵<https://universaldependencies.org/u/overview/simple-syntax.html#nonverbal-clauses>

⁶The *expl* relation would also be used for the empty subject in the German weather construction in (14b).

⁷The official guidelines do not discuss clefts at all, but they are the subject of several ongoing discussions on the UD GitHub site.

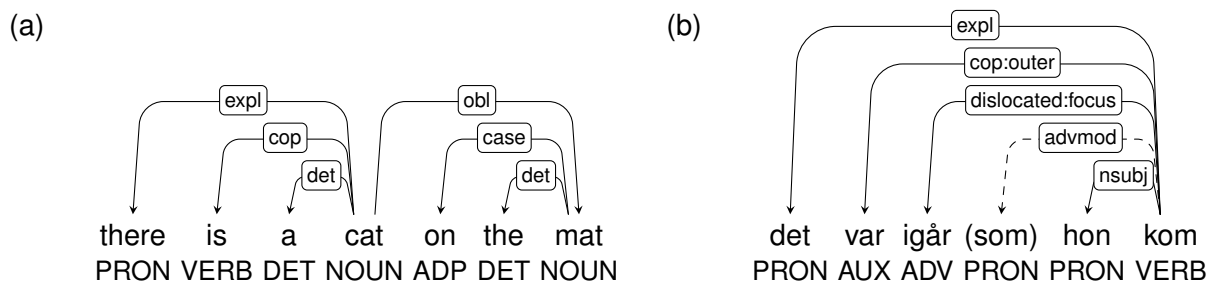


Figure 5: Revised annotation of (a) presentational and (b) identificational constructions.

is indistinguishable from a relative clause, but it is attached to a temporal adverb rather than a nominal, which raises the question whether it should be labeled *advcl* instead of *acl*. In addition, there is disagreement on whether it should be subtyped *:relcl*, to reflect the form of the clause, or *:cleft* to reflect its function.

4. Discussion

Our review of nonprototypical clause constructions has revealed a striking contrast between nonprototypical predication, which mostly has a straightforward analysis in UD (except possibly the verbal strategy, where some clarification is needed), and nonpredicational clauses, which are not discussed at all in the guidelines and where it is often unclear how to analyze specific constructions and strategies. We will therefore focus on the latter in our discussion.

What can be done to improve the analysis ofthetic and identificational constructions, as well as non-prototypical topic-comment constructions? One approach is to add information on top of the UD annotation, as proposed by Weissweiler et al. (2024). Such annotations could, for example, add information about the focus part and the POP part of an identificational construction. A more radical approach would be to introduce new syntactic relations into the UD framework itself, like a *focus* relation for the analysis of identificational constructions. However, this would only be possible in a future version of UD, since the inventory of syntactic relations is fixed in UD v2.

Nevertheless, we believe that annotations could be improved even within the current version of UD by following the guiding principle that core argument relations should not be used to annotate strategies that primar-

ily serve nonpredicational functions. This will sometimes require a change in what is considered the root of the construction, and it will often require an extended use of some (non-core) relations, possibly with special subtypes. For example, *dislocated* can be extended to cover all word order strategies, and *obl* can be generalized to all flagging strategies. In addition, special relations like *expl* and *cop* can be invoked for grammaticalized markers that do not function as real arguments or predicates. Let us consider two examples to illustrate the basic idea.

To recognize thethetic function of a presentational construction like (12a), one could treat the phrase introducing the new referent as the root of the construction and annotate the grammaticalized markers with *expl* and *cop*, as illustrated in Figure 5(a). This can be compared with the current analysis, shown in Figure 3(a), which (incorrectly) assumes that the referring phrase is a core argument of the copula verb. In cases where the presentational construction is instead realized with a lexical verb and a non-canonical word order, as in Spanish *viene Juan* (lit. “comes John”), a new relation subtype *dislocated:thetic* can be used to indicate the strategy (*dislocated*) as well as the construction type (*:thetic*).⁸

To capture the identificational function of the cleft strategy in (18a), one could introduce a new subtype *dislocated:focus* to relate the focus phrase to the POP and again use *expl* and *cop* for the initial grammaticalized markers, as illustrated in Figure 5(b). The internal struc-

⁸An alternative would be to use a subtype that captures the role of the displaced argument, although this would make thethetic construction indistinguishable from the special topic-comment structures discussed above.

ture of the POP is here analyzed analogously with the relative clause structure that has been recruited for construction, but the external relation makes clear that it does not have its normal modifying function. It can be discussed whether the root of the clause should be the POP, as in Figure 5(b), or the focus part, as in the old analysis in Figure 4, but treating the POP as the root has two advantages: it clarifies that the POP, despite its form, does not function as a subordinate clause, and it generalizes more naturally to equational constructions.

These are just two examples, and even their analysis is open to discussion, so a detailed proposal for guidelines for thetic and identificational constructions will have to be left for future work. However, we hope to have shown that simply annotating them as if they were predicational constructions may hinder rather than help the meaningful cross-linguistic comparisons that UD was created to support (de Marneffe et al., 2021).

5. Conclusion

In this paper, we have taken another step towards a construction for UD, in the sense of Nivre (2025), by reviewing the way UD annotates nonprototypical predication and nonpredicational clauses, following the taxonomy of Croft (2022), thereby extending the previous work on reference and modification (Nivre and Croft, 2025) and verbal predication (Croft and Nivre, 2025).⁹ For nonprototypical predication, the guidelines provide a straightforward analysis of the main constructions and strategies, as outlined in Table 1. For nonpredicational clauses, the situation is much less clear, and many of the constructions and strategies discussed in Croft (2022) are either not covered at all in the current guidelines, or are subsumed under predicational constructions. This deserves further discussion, and we have briefly outlined a strategy for tackling the issue.

⁹In concurrent papers, we also review the annotation of complex predicates (Coneglian et al., 2026) and speech act constructions (Nivre et al., 2026).

6. Acknowledgments

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology), from the Swedish Research Council (grant no. 2022-02909), and from the Brazilian National Council for Scientific and Technological Development, CNPq (grant no. 153972/2025-4 and no. 405484/2025-9). We thank Elisabeth Engdahl, Leonie Weissweiler and three anonymous reviewers for comments and suggestions that helped improve the article.

7. Bibliographic References

- André V. Lopes Coneglian, Joakim Nivre, and William Croft. 2026. Complex predicates in Universal Dependencies. In *Proceedings of the Ninth Workshop on Universal Dependencies*.
- William Croft. 2022. *Morphosyntax: Constructions of the World's Languages*. Cambridge University Press.
- William Croft and Joakim Nivre. 2025. Verbal predication constructions in Universal Dependencies. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 50–60.
- Marie de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47:255–308.
- Pål Eriksen, Seppo Kittilä, and Leena Kolehmainen. 2012. Weather and language. *Language and Linguistics Compass*, 6:383–402.
- Irene Heim. 1983. File change semantics and the familiarity theory of definiteness. In R. Bäuerle, C. Schwarze, and A. von Stechow, editors, *Meaning, Use, and Interpretation of Language*, pages 164–189.
- Shoichi Iwasaki. 2013. *Japanese*. John Benjamins.
- R. L. Johnston. 1980. *Nakanai of New Britain: The Grammar of an Oceanic Language*. The Australian National University.

- Knud Lambrecht. 1994. *Information Structure and Sentence Form: Topic, Focus and the Mental Representations of Discourse Referents*. Cambridge University Press.
- Charles Li and Sandra A. Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. University of California Press.
- Arthur Lorenzi, Peter Ljunglöf, Ben Lyngefelt, 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 constructions. In *Proceedings of the 20th Joint ACL – ISO Workshop on Interoperable Semantic Annotation*, pages 93–98.
- Elena Maslova and Giuliano Bernini. 2006. Sentence topics in the languages of Europe and beyond. In Giuliano Bernini and Marcia L. Schwartz, editors, *Pragmatic Organization of Discourse in the Languages of Europe*. Mouton de Gruyter.
- 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, André V. Lopes Coneglian, and William Croft. 2026. Speech act constructions in Universal Dependencies. In *Proceedings of the Ninth Workshop on Universal Dependencies*.
- Joakim Nivre and William Croft. 2025. Reference and modification in Universal Dependencies. In *Proceedings of the 8th Workshop on Universal Dependencies (UDW)*, pages 1–10.
- 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.
- C. R. Osborne. 1974. *The Tiwi Language*. Australian Institute of Aboriginal Studies.
- John R. Roberts. 1987. *Amele*. Croom Helm.
- Hans-Jürgen Sasse. 1987. The thematic-categorical distinction revisited. *Linguistics*, 25(5):11–80.
- Leon Stassen. 1997. *Intransitive Predication*. Oxford University Press.
- Leon Stassen. 2009. *Predicative Possession*. Oxford University Press.
- Archibald N. Tucker. 1940. *The Eastern Sudanic Languages*. Oxford University Press.
- Barbara Wehr. 1984. *Diskursstrategien im Romanischen*. Gunter Narr.
- Leonie Weissweiler, Nina Böbel, Kirian Guiller, Santiago Herrera, Wesley Scivetti, Arthur Lorenzi, Nurit Melnik, Archana 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.

Complex Predicates in Universal Dependencies

André V. Lopes Coneglian, Joakim Nivre, William Croft

Federal University of Minas Gerais, Uppsala University, University of New Mexico
coneglian@ufmg.br, joakim.nivre@lingfil.uu.se, wacroft@icloud.com

Abstract

To assess whether the framework of Universal Dependencies (UD) is compatible with findings from linguistic typology, we need to systematically review how UD represents linguistic constructions and how it handles the range of morphosyntactic variation attested across languages. In this paper, we present the results of such a review focusing on complex predicates. We arrive at distinct findings regarding the two main types of complex predicates. The UD framework can well accommodate eventive complex predicates, particularly serial verbs, and more grammaticalized forms of complex predicates, such as voice and TAMP auxiliaries, with the exception of incorporating strategies. However, the guidelines for stative complex predicates could be revised based on the typology of morphosyntactic strategies. We briefly discuss possible ways in which UD can be extended to better capture these strategies.

Keywords: Universal Dependencies, eventive complex predicates, stative complex predicates

1. Introduction

Universal Dependencies (UD) is an annotation framework for morphosyntax, designed to be applicable to all human languages and to enable 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 construction 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. For example, the object predication construction is defined as any construction in any language where an object concept is predicated of a reference, as in the English *she is a doctor*. The verbal copula strategy describes an object predication construction that involves a linking element in the form of a verb, like English *is*. Other common strategies in the world's language are the non-verbal copula strategy (with a nonverbal linking element) and the zero strategy (with no linking element) (Croft, 2022; Nivre et al., 2026b).

The rationale behind the construction pro-

posal is that, by checking if the current UD guidelines cover all attested strategies of a given construction, we can identify gaps in the guidelines and propose additions when needed. Moreover, in the same process, we can assess to what extent the annotations of different strategies capture the fact that they are realizations of the same construction, which facilitates cross-linguistic comparisons in the spirit of UD's "mixed structural-functional system" for annotation.¹ A further step would be to investigate how actual treebanks have dealt with the phenomena at hand. However, although we reference specific treebanks when relevant, a systematic review of treebanks lays outside the current scope.

Previous and concurrent contributions to this project include studies of reference and modification constructions (Nivre and Croft, 2025), verbal predication constructions (Croft and Nivre, 2025), nonprototypical predication and nonpredicational clauses (Nivre et al., 2026b), and speech act constructions (Nivre et al., 2026a). In this paper, we extend the review to complex predicates, a family of constructions discussed in Chapters 13–14 of Croft (2022).

Complex predicates are monoclausal constructions in which two or more elements com-

¹<https://universaldependencies.org/u/overview/syntax.html#the-taxonomy-of-typed-dependencies>

bine into a single predication, in contrast to single predicates in which predication is expressed by one element (Croft, 2022). They cover a wide range of semantic categories and present a diverse set of morphosyntactic structures. Complex predicates are intermediate structure types in the grammaticalization path, between biclausal structures and affixes (Bowern, 2008). Following Croft (2022), we consider two main groups of complex predicates, namely, *eventive complex predicates*, where both elements of the predicate denote processes (Section 2), and *stative complex predicates*, where the other element denotes a state (Section 4). We also consider the more grammaticalized types of eventive predicates, such as voice and TAMP auxiliaries (Section 3).

2. Eventive Complex Predicates

The main characteristic of eventive complex predicates is that they denote one single event. Therefore, the elements of a complex predicate share the same argument structure, and the same semantic values with respect to tense, aspect, evidentiality, mood, modality and reality status. Additionally, a complex predicate constitutes a single assertion, thus the complex predicate will be negated or questioned as a whole (Croft, 2022). The wide range of constructions and morphosyntactic strategies for eventive complex predicates covers different areas of the grammar of languages, such as serial verbs, light verbs and support verbs, and verb-object idiomatic expressions. Here, we focus exclusively on *serial verb* strategies, that is, strategies that combine one or more verbs, each denoting an event, to realize eventive complex predicates expressing meanings such as direction, orientation, aspect, change of state, sequence of subactions, simultaneity of actions, among others (Aikhenvald, 2018, p. 239).

2.1. Strategies for Serial Verbs

The range of morphosyntactic strategies for serial verbs can be defined along three morphosyntactic parameters (Croft, 2022): (a) contiguity—whether or not the predicate elements are contiguous; (b) boundedness or

incorporation—whether or not the predicate elements form one single word; (c) locus of predicate inflection—which element(s) receives predicate inflection. Starting with contiguity, example (1), from Tetun Dili (Hajek, 2006; Aikhenvald, 2018, p. 2), illustrates a case of non-contiguous serial verbs, in which they are separated by an argument phrase; while example (2), from Barai (Aikhenvald, 2018), illustrates a case of serial verb contiguity.

- (1) tuda bola mai
throw ball **come**
 ‘throw the ball over here’
- (2) fu fase naebe fi isoe
 he letter NEG **sit write**
 ‘he did not sit and write a letter’

In examples (1–2) each element of the serial verb is a separate word. By contrast, example (3), from Tepehua (Watters, 2017; Aikhenvald, 2018, p. 98) illustrates a case of incorporation, where the two elements of the serial verb form a single word. Example (4), from Sakao (Guy, 1974; Durie, 1997, p. 303), illustrates a case in which the serial verbs are bound, but an argument is interpolated between the two forms in boldface.

- (3) haʔtanah-tsuku-t=ta
walk-begin-PFV=already
 ‘S/he began to walk’ [lit. walk begin]
- (4) ʎa-βœt-nœð-p-ri-lam l-aðay
 3PL-**pluck**-coconut-PFV-DIR-**come** to-baskets
 ‘They plucked coconuts hither into the baskets’

With respect to the locus of predicate inflection, there are five morphosyntactic strategies (Durie, 1997; Croft, 2022): inflection on the first verb; inflection on the last verb; same inflection on all verbs; split inflection; separate inflection. All these types have been attested cross-linguistically. Due to space limitations, we cannot provide examples for all these strategies, but the Òbòlò example (5) illustrates the same inflection on both verbs (Durie, 1997, p. 299).

- (5) é-gwên èmì é-nû
 PL-**call** 1SG PL-**come**
 ‘let them call me to come’

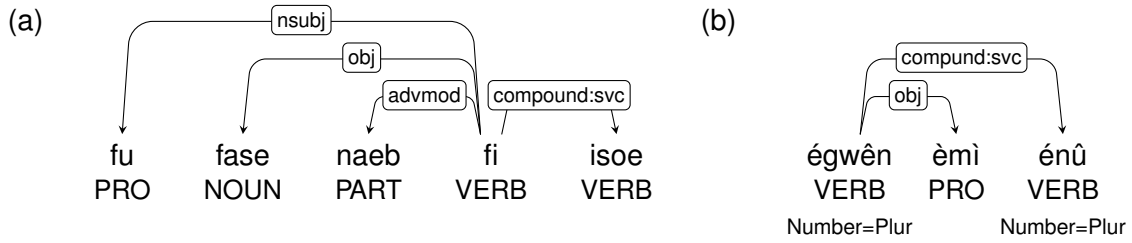


Figure 1: Simplified UD annotation of serial verb strategies: (a) contiguous serial verb strategy, (b) non-contiguous serial verb strategy with the same inflection on both verbs.

The strategies discussed up to this point are all zero-coded serial verb strategies, that is, there is no element that links the predicate elements. However, there are cases of overt coding for a serial verb strategy, by means of a conjunction element or a converb construction. The following example from Mooré (Schiller, 1990, p. 38) is a case of a serial verb that optionally takes a conjunction.

- (6) a ìku sùgā (n) wāg nemdā
 he took knife (CNJ) cut meat
 ‘he cut the meat with a knife’

Even though converb constructions typically express a semantic relation between two events in biclausal structures, Bisang (1995) and Shibatani (2009) show they may also be used to express a semantic relation between events that form one single event, as is the case of serial verb strategies. The Khalkha Mongolian example (7), from Street (1963), cited in Bisang (1995, p. 170), shows a converb construction in which the converb is used as a flag, and the verb *give* for the R role of the communication event (Croft, 2022, p. 410).

- (7) tūūnig duuda-ž òg-òòrej
 him phone-CNV give-IMP
 ‘please call him [to the phone] for [me].’

Overt linking in serial verbs requires careful analysis, because conjunction and converb constructions can also be found in biclausal and multi-eventive predicates (Croft, 2022, ch. 15–19).

2.2. UD Annotation

According to the UD guidelines, serial verbs should be annotated with the relation *compound:svc*, that is, as a *compound* with the

subtype *svc*, as Figure 1 illustrates.² In both these examples, the first element of the serial verb is treated as the head, but the guidelines allow any of the two elements to be treated as the head, based on language-specific criteria, and head-final serial verbs are found, for example, in UD treebanks for Turkic languages (Akhundjanova et al., 2025).

There are no explicit guidelines for the morphological annotation of serial verb elements, but it seems natural to assume that they are all assigned the part-of-speech tag VERB and morphological features in accordance with the inflection strategy, as illustrated in Figure 1(b). Finally, it should be stated that serial verb strategies involving incorporation pose a significant challenge for the current UD annotation, particularly bounded serial verb strategies, as in example (3), because the verbal elements of the serial verb tend to constitute only one word. We return to this issue in Section 5.

3. Grammaticalized Constructions

As mentioned above, eventive complex predicates may grammaticalize into constructions with different functions. Here we discuss two commonly attested cases, namely nonbasic voice constructions and TAMP-auxiliary constructions.

3.1. Nonbasic Voice Constructions

Eventive complex predicates may be used to express nonbasic voice, such as passive-

²In this and all following examples, we simplify the UD representations by omitting (a) lemmas and (b) morphological features that are not relevant for the discussion. For reasons of space, we do not repeat glosses and translations in the examples.

inverse, causative and applicative (Croft, 2022; Croft and Nivre, 2025). Typically, verbs like ‘give’ and ‘take’ grammaticalize into oblique flags, coding the relationship between argument phrases in the overall predication. This is the case in many languages around the world. Example (8), from Mandarin, illustrates the form *gěi* ‘give’ expressing the beneficiary of the action ‘make’ (Li and Thompson, 1985; Croft, 2022).

- (8) māma gěi wǒ zuò jiaǒzi
 mother **for** 1SG make dumpling
 ‘mother made dumplings for me’

One problem that arises in the analysis of these constructions is the categorization of forms like *gěi* in Mandarin and in other languages, given that these forms are undergoing grammaticalization but may retain semantic content and morphosyntactic structure from their source. Due to the gradualness of grammaticalization, it may therefore be unclear whether forms like *gěi* should be analyzed as forming a complex predicate with a(nother) verb or as a flag combining with an argument phrase (Croft, 2022).

3.2. TAMP-Auxiliary Constructions

Auxiliaries are another common type of grammaticalization of eventive complex predicates, which come to express meanings such as tense, aspect, mood, modality and polarity (Givón, 2001; Anderson, 2006). They do not constitute a basic type of eventive complex predicate; however, they do share the same set of parameters and morphosyntactic strategies that were discussed for serial verbs in Section 2. Thus, auxiliary constructions, too, are defined by the parameters of contiguity, boundedness and locus of predicate inflection.

The English example *she might come* illustrates contiguity of auxiliary and main verb. By contrast, the auxiliary in the Kisi example (9) is non-contiguous with the main verb.

- (9) fàlà có lééǵdó yìkpàá
 Fala **AUX** machete **sharpen**
 ‘Fala is sharpening the machete’

Concerning incorporation, auxiliaries may be bound or free. The auxiliary in (9) is free, while the Abkhaz auxiliary in (10) is bound (Hewitt, 1979; Bybee et al., 1994).

- (10) s-cà-r-o-w+p’
 1SG-**go-COND**-be-STAT
 ‘I must go’

Concerning locus of inflection, auxiliaries can be classified into the same five types as serial verbs (Croft, 2022, p. 415–416): (i) inflection on the auxiliary; (ii) inflection on the main verb; (iii) same inflection on auxiliary and main verb; (iv) split inflection; (v) separate inflection.

3.3. UD Annotation

The annotation of argument structure and voice constructions that result from the grammaticalization of eventive complex predicates requires language-specific parameters for guidelines. Consider example (8) above. From our constructional perspective, it is not clear which part-of-speech tag should be assigned to the form *gěi*, due to the fact that it is undergoing grammatical change. However, if the preferred option is ADP, then the syntactic relation that links it to the pronoun form has to be *case*. This is the analysis adopted in a current UD treebank for Chinese, (Poiret et al., 2023) and the corresponding UD annotation of (8) is given in Figure 2(a).

The annotation of auxiliary constructions is straightforward in UD. Basically, all the cases mentioned in the preceding discussion are tagged as AUX and linked to the main verb with *aux* relations, which is reserved for auxiliaries of tense, aspect, evidentiality, or mood to the extent that they are grammaticalized in the language (De Marneffe et al., 2024). In UD v2, it is recommended that the subtype *aux:pass* be used for grammaticalized periphrastic voice, as in the Brazilian Portuguese example (11), illustrated in Figure 2(b), where the verb form *foi* is the passive voice auxiliary.

- (11) o vaso foi quebrad-o
 DEF.M.SG vase be.3SG.PRF broken-M
 ‘the vase was broken’

Although the guidelines for auxiliaries are clear in themselves, the gradual nature of grammaticalization may give rise to borderline cases, and less grammaticalized auxiliary strategies are generally not to be analyzed with the AUX tag and *aux* relation. A typical case is what has been called periphrastic aspect, as in /

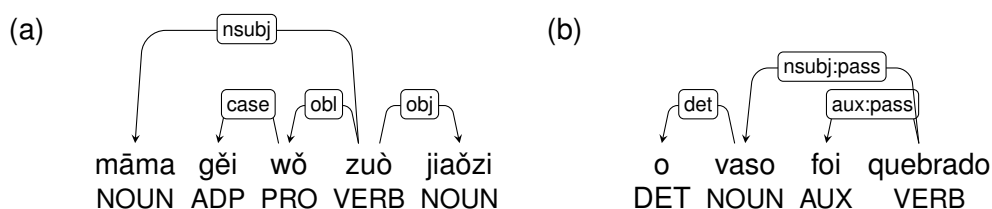


Figure 2: Simplified UD annotation of grammaticalized constructions: (a) nonbasic voice, (b) auxiliary.

began to wonder. In cases like this, the recommendation is that the verb *began* be analyzed as a main verb taking a complement clause headed by the verb *wonder*; the relation from *began* to *wonder* is labeled *xcomp*, because there is obligatory control, and *to* is linked to *wonder* with the *mark* relation, which indicates syntactic subordination.

4. Stative Complex Predicates

A stative complex predicate is a monoclausal construction in which one element denotes a process and the other denotes a state (Croft, 2022). The stative concept does not modify the referent, as is the case in attributive modifying constructions (Nivre and Croft, 2025), nor does it constitute the primary element of the predication. Rather, the stative element constitutes a part of the complex predicate. The three main semantic types of stative complex predicates are: *resultative*, *depictive*, and *manner*. The English examples (12–14) from Croft (2022, p. 432) illustrate these three major types, respectively.

(12) we **painted** the wall **red**

(13) I **ate** the carrots **raw**

(14) we **crawled** down the slope **slowly**

Most of the typological and cross-linguistic work on stative complex predicates is focused on depictive and manner predicates, while resultative predicates still lack robust cross-linguistic investigation.

4.1. Strategies for Stative Complex Predicates

The three types of stative complex predicates share many morphosyntactic strategies (Schultze-Berndt and Himmelmann, 2004;

Loeb-Diehl, 2005; Croft, 2022), which can be characterized according to three morphosyntactic parameters (Loeb-Diehl, 2005; Croft, 2022): (i) whether it uses balanced or deranked³ predicates; (ii) whether the stative predicate is indexed or non-indexed; and (iii) whether the stative predicate recruits⁴ the action predication construction – what Loeb-Diehl (2005) calls “verby” – or the adjective or nominal construction – called “nouny”.

Let us start with the four verby strategies, i.e. strategies that recruit the action predication construction (Croft, 2022). Two of these, the *coordinate personal strategy* and the *coordinate impersonal strategy*, are balanced strategies, where the stative predicate constitutes a separate coordinated predication. The difference between the two is that, in the former the stative predicate indexes the participant (hence, personal), while in the latter the stative predicate does not (hence, impersonal). Example (15) from Muna (van den Berg, 1989; Loeb-Diehl, 2005) and example (16)⁵ from Koasati (Kimball, 1985; Loeb-Diehl, 2005) illustrate the personal and impersonal coordinate strategies, respectively.

(15) ne-rimba no-tende
 3SG.RL-be_fast 3SG.RL-run
 ‘he runs fast’ [lit. ‘he is fast (and) he runs’]

³In complex predicates, a balanced strategy recruits the construction of a simple predicate construction with inflections, etc. By contrast, a deranked strategy either lacks the inflections of the predicate, uses a different set of inflections from the predicate, or has designated affixes that overtly code the other predicate (Croft, 2022).

⁴Recruitment is a relationship between two constructions in which the structure of one is recruited or whose function is extended to use in another construction (Croft, 2022).

⁵In example (16), DS stands for ‘different subject’ (Loeb-Diehl, 2005).

- (16) ca-conoská-k pálk-á:hi:si-n bókl
 1SG.POSS-heart-SBJ **be_fast-ADV-DS** beat
 'my heart is beating very fast'
 [lit. 'my heart is fast and beats']

By contrast, *participial* and *converb* strategies are deranked strategies. The difference between them is indexation: while the first indexes a participant of the main event, the latter does not. In example (17), from Sanuma (Borgman, 1990; Loeb-Diehl, 2005), manner is expressed by a same-subject participle marked by *-i*, glossed as REL. Example (18), from Turkana (Dimmendaal, 1983; Loeb-Diehl, 2005), illustrates the converb strategy, where *ni* indexes the event, not the participant.

- (17) opi-i a kali-palo-ma
 be_slow-REL 3SG work-REPET-COMPLET
 'he worked slowly'
 (18) è-pès-e-tè ni-a-ron-o-ñi
 3-kick-A-PL REL.N-IMP-be_bad-SG.REL
 'they kick him badly'
 [lit. they kick him, it being bad]

We now turn to the four nouny strategies, i.e. the strategies that recruit adjectival or nominal constructions and use inflectional categories like gender, number, and case. The *adjective personal strategy* recruits a modifying phrase construction (Nivre and Croft, 2025). This strategy indexes an event participant, not the event itself (Croft, 2022). Loeb-Diehl (2005) notes that this strategy is rather uncommon cross-linguistically. It is observed in Classical Latin and some daughter languages, such as Spanish, Italian, Romanian, and Portuguese. In the Spanish example (19), the adjective forms *tranquilos* and *felices* index the plural (Slager, 1995; Loeb-Diehl, 2005).

- (19) aqui vivimos tranquilos y felices
 here live.1PL quiet.PL and happy.PL
 'we live quietly and happily here.'
 [lit. 'we live here quiet and happy']

The *adjective impersonal strategy* differs from the former only with respect to indexation. This strategy tends to be used for manner predicates, and not for depictives; as Loeb-Diehl (2005) notes, this strategy is more event-oriented. This strategy is also found in Romance languages, like Portuguese, among many other languages, illustrated in (20). In this example no indexation is observed in the adjective form *rápido* (fast).

- (20) ela corre rápido
 3SG.F run.3SG.PRS **fast**
 'she runs fast/fastly'

In deranked strategies the stative predicate may combine with a copula or be expressed with a flag. In the first case, the *copular participle strategy*, the copula is a deranked form combined with an adjectival form. In the Malayalam example (21) the bound *-aayi* is a past participle copular form (Askher and Kumari, 1997; Loeb-Diehl, 2005). In the second case, the *adpositional strategy*, the stative predicate is expressed with a flag like a referring phrase. In the Japanese example (22), the adjective form is marked with a locative and instrumental flag (Takezawa, 1993; Schultze-Berndt and Himmelmann, 2004).

- (21) aval bhamgiy-aayi prasamgiccu
 she beauty-PTCP.COP speak-PST
 'she spoke beautifully'
 (22) John-ga sakan-o hadaka-de tabe-ta
 John-NOM fish-ACC nude-INS eat-PST
 'John ate the fish nude'

The *adpositional personal strategy* (Schultze-Berndt and Himmelmann, 2004) involves an indexed stative predicate, similarly to the other personal strategies discussed above. The Russian example (23) illustrates the use of both a flag marker and a personal marker in the stative predicate element (Nichols, 1978; Schultze-Berndt and Himmelmann, 2004).

- (23) on umer molodym
 he die.PST.M.SG young.M.SG.INS
 'he died young'

The *predicational strategy* is one in which the manner or stative item becomes the main predicate. As Loeb-Diehl (2005) describes it, in this strategy the dynamic subevent is nominalized and is the subject of the sentence. This strategy is very rare crosslinguistically and is illustrated in example (24) from Mokilese (Harrison, 1976; Loeb-Diehl, 2005).

- (24) ah kijou dahr
 his run fast
 'he runs fast' [lit. 'his running is fast']

Finally, the *adverbial strategy* uses a distinct morphosyntactic form to express manner complex predicates. It typically involves an affix

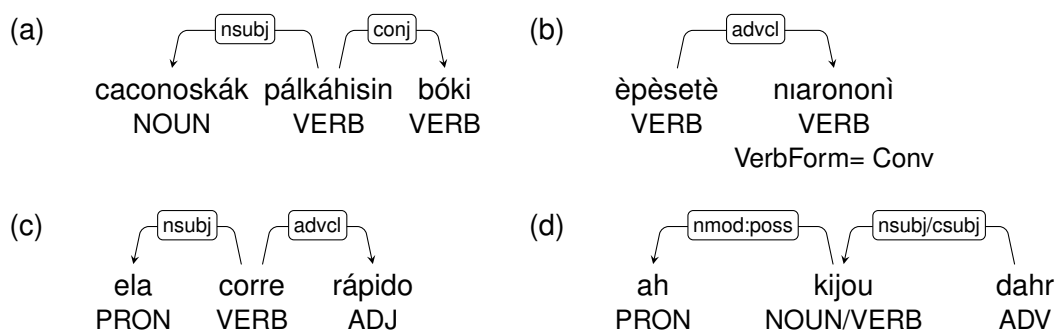


Figure 3: Simplified UD annotation of stative complex predicate strategies: (a) coordinate strategy, (b) converb strategy, (c) adjective strategy, (d) predication strategy.

to the manner predicate, like the *-ly* suffix in English, or *-mente* in Portuguese. There is a specially designated word form for this strategy (Loeb-Diehl, 2005; Croft, 2022).

4.2. UD Annotation

The current UD guidelines do not treat stative complex predicates as a concept, which means that different strategies are annotated based on their morphosyntactic form and are therefore often assimilated to other constructions that use similar strategies. Thus, in coordinate strategies, the stative predicate is tagged VERB and linked to the main predicate with the *conj* relation, which is used to annotate all types of coordination in UD. This is illustrated in Figure 3(a), which shows the UD annotation of example (16).

In the participial, converb and adjective strategies, the stative predicate is tagged VERB or ADJ and linked to the main predicate with an *advcl* relation – a relation that is also used for full adverbial clauses and therefore makes the complex predicate constructions look like multi-clausal constructions. This is illustrated in Figure 3(b–c), which shows the UD annotation of examples (18) and (20).

In the adpositional strategy, the stative predicate is tagged ADJ (or possibly NOUN) and linked to the main predicate with an *obl* relation, and the flag is represented either by a Case feature, in case it is inflectional, or with a case dependent, in case it is an independent word. This groups this strategy together with non-core arguments of predicates.

In the adverbial strategy, the stative predicate is tagged ADV and linked to the main

predicate with an *advmod* relation, which is not a clausal relation and therefore compatible with the monoclausal nature of the complex predicate construction, although it is also used for other types of construction, notably negation in many languages.

For the predication strategy, finally, it is clear that the nominalized main predicate has to be analyzed as a subject of the stative predicate, but depending on whether the former is tagged NOUN or VERB, the subject relation may be either *nsubj* or *csubj*.

5. Discussion

The review carried out in this paper can be summarized in two main points. First, from a typological perspective, complex predicate constructions exhibit a wide range of possible annotations within the UD framework. This is explained, at least in part, by the fact that complex predicates are very broadly defined in the constructional framework of Croft (2022). Secondly, regarding the current UD guidelines, we found that UD v2 presents clear instructions for the annotation of most of the strategies covered in this paper, to varying degrees of comprehensiveness and typological adequacy. The current version of the guidelines for eventive complex predicates, particularly those more grammaticalized as TAMP auxiliaries, are the most consistent with typological findings. The guidelines for annotating stative complex predicates are more heterogeneous in that different strategies are grouped with different construction types. In particular, there is a clear divide between strategies that are

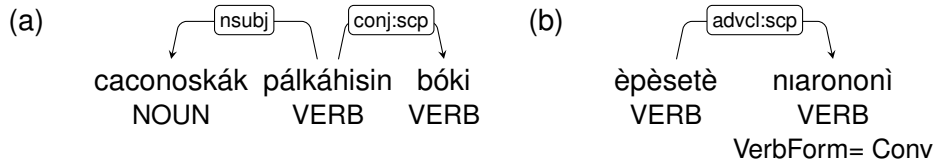


Figure 4: Revised UD annotation of stative complex predicate strategies with subtypes: (a) coordinate strategy, (b) converb strategy.

annotated as mono-clausal versus those that are analyzed as bi-clausal.

The prescription for the annotation of serial verb constructions is very clear in UD guidelines. However, as we noted above, the current version of the guidelines does not cover incorporating strategies, in which the elements form a single word, as is typical in more polysynthetic languages. This means that, at this time, we cannot present any conclusive proposal regarding incorporating strategies for serial verbs and for auxiliaries. The UD framework still lacks clear orientations for how to deal with incorporation and polysynthesis in general, despite attempts to establish guidelines (Tyers and Mishchenkova, 2020; Park et al., 2021; Koshevoy et al., 2023). Similarly, current guidelines cover only zero coded serial verb strategies, and give no indication for the annotation of the overt coding strategies. Since in these cases the overt element is typically a conjunction or linking element, we see no problem in tagging this element as CCONJ,⁶ but it is not clear whether it can be assigned a *cc* relation without also changing the relation between the two predicates to *conj*. The latter would assimilate these strategies with the coordinate strategies for stative complex predicates, but would obscure the relation to zero coded serial verb strategies.

With respect to stative complex predicates, we have seen that the current guidelines can handle all the different strategies, but the relations that have to be used group them together with quite different constructions. Coordinate strategies are grouped with other types of coordination; participial, converb and adjective

strategies are grouped with adverbial clauses; and adpositional strategies are grouped with oblique arguments. One way of mitigating this problem would be to introduce a new subtype *:scp* (for stative complex predicate), which would at the same time capture their constructional similarity and differentiate them from other constructions annotated with the same relations. This would mean using *conj:scp* for coordinate strategies, *advcl:scp* for participial, converb and adjective strategies, and *obl:scp* for the adpositional strategy. The former two cases are illustrated in Figure 4.

It is also worth noting that the use of the *advcl* relation for several strategies implies a bi-clausal structure, which appears incompatible with the definition of complex predicates as mono-clausal. It is particularly striking that the adverbial and the adjective strategy for manner predicates end up getting very different analyses, despite being very similar in nature. One way of mitigating this would be to allow the adjective strategy to be analyzed using the *advmod* relation, at least when used to express a manner predicate. However, it seems more difficult to avoid the use of *advcl* for participial and converb strategies, where the stative predicate is verbal. Hence, a fully satisfactory solution to this problem would probably require adding new universal relations, which can only be considered for future versions of UD.

6. Conclusion

In this paper, we have taken another step towards a construction for UD, in the sense of Nivre (2025), by reviewing the way UD annotates complex predicates, following the taxonomy of Croft (2022), thereby extending the previous work on reference and modification (Nivre and Croft, 2025) and verbal predication

⁶Our survey based on Croft (2022) has shown that only balanced strategies are used in this case. In any case, if there are languages in which overt coding is used, then the annotation should be different.

(Croft and Nivre, 2025).⁷ On the one hand, for both serial verbs and auxiliaries, the current guidelines provide very straightforward instructions of annotation, with the limitation that they do not have any clear way of dealing with incorporation; however, this is due to the fact that the UD framework does not accommodate well phenomena related to polysynthesis. On the other hand, the guidelines for stative complex predicates are less satisfactory in that different strategies end up being grouped with different constructions and in many cases lead to a bi-clausal analysis. We have discussed a few ways in which these problems can be mitigated, but a complete revision of the guidelines will likely only be possible when a new major version of UD is launched.

7. Acknowledgments

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology), from the Swedish Research Council (grant no. 2022-02909), and from the Brazilian National Council for Scientific and Technological Development, CNPq (grant no. 153972/2025-4 and no. 405484/2025-9). We thank Leonie Weissweiler and three anonymous reviewers for comments and suggestions that helped improve the article.

8. Bibliographic References

Alexandra Aikhenvald. 2018. *Serial verbs*. Oxford University Press.

Arofat Akhundjanova, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, and Cagri Coltekin. 2025. Parallel Universal Dependencies treebanks for Turkic languages. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 129–136, Ljubljana, Slovenia. Association for Computational Linguistics.

Gregory DS Anderson. 2006. *Auxiliary Verb Constructions*. Oxford University Press.

⁷In concurrent papers, we also review nonprototypical predication and nonpredicational clauses (Nivre et al., 2026b), as well as speech act constructions (Nivre et al., 2026a).

R. E Askher and T.C. Kumari. 1997. *Malayalam*. Routledge.

Walter Bisang. 1995. Verb serialization and converb – differences and similarities. In Ekkehard König and Martin Haspelmath, editors, *Converbs in Cross-Linguistic Perspective*, pages 137–88. Mouton de Gruyter.

Donald M. Borgman. 1990. Sanuma. In *Handbook of Amazonian Languages*, pages 15–248. Mouton de Gruyter.

Claire Bowern. 2008. The diachrony of complex predicates. *Diachronica*, 25(2):161–185.

Joan Bybee, Revere Perkins, and William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. University of Chicago Press.

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

William Croft and Joakim Nivre. 2025. [Verbal predication constructions in Universal Dependencies](#). In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 50–60.

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

Marie-Catherine De Marneffe, Joakim Nivre, and Daniel Zeman. 2024. Function words in universal dependencies. *Linguistic Analysis*, 43(3–4):549–588.

Gerrit Jan Dimmendaal. 1983. *The Turkana Language*. Foris.

Mark Durie. 1997. Grammatical structures in verb serialization. In Alex Alsina, Joan Bresnan, and Peter Sells, editors, *Complex Predicates*, pages 289–354. CSLI.

Talmy Givón. 2001. *Syntax, vol. 1*. John Benjamins.

J. B. M. Guy. 1974. *A Grammar of the Northern Dialect of Sakao*. Research School of Pacific Studies, Australian National University.

- John Hajek. 2006. Serial verbs in Tetun Dili. In *Serial Verb Constructions: A Cross-Linguistic Typology*, pages 239–253. Oxford University Press.
- S. P. Harrison. 1976. *Mokilese Reference Grammar*. University Press of Hawaii.
- B. G. Hewitt. 1979. *Abkhaz*. North-Holland.
- Geoffrey D. Kimball. 1985. *Koasati Grammar*. University of Nebraska Press.
- Alexey Koshevoy, Anastasia Panova, and Ilya Makarchuk. 2023. Building a universal dependencies treebank for a polysynthetic language: the case of Abaza. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 1–6.
- Charles Li and Sandra Thompson. 1985. *Mandarin Chinese: A Functional Reference Grammar*. University of California Press.
- Flora Loeb-Diehl. 2005. *The Typology of Manner Expression*. Ph.D. thesis, Radboud University Nijmegen.
- 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 constructions. In *Proceedings of the 20th Joint ACL – ISO Workshop on Interoperable Semantic Annotation*, pages 93–98.
- Johanna Nichols. 1978. Secondary predicates. In *Proceedings of the 4th Annual Meeting of the Berkeley Linguistics Society*.
- 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, André V. Lopes Coneglian, and William Croft. 2026a. Speech act constructions in Universal Dependencies. In *Proceedings of the Ninth Workshop on Universal Dependencies*.
- Joakim Nivre and William Croft. 2025. Reference and modification in Universal Dependencies. In *Proceedings of the 8th Workshop on Universal Dependencies (UDW)*, pages 1–10.
- Joakim Nivre, William Croft, and André V. Lopes Coneglian. 2026b. Nonprototypical predication and nonpredicational clauses in Universal Dependencies. In *Proceedings of the Ninth Workshop on Universal Dependencies*.
- 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.
- 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.
- Rafaël Poiret, Tak-Sum Wong, John Lee, Kim Gerdes, and Herman Leung. 2023. Universal dependencies for Mandarin Chinese. *Language Resources and Evaluation*, 57(2):673–710.
- Eric Schiller. 1990. *An autolexical account of subordinating serial constructions*. Ph.D. thesis, Department of Linguistics, University of Chicago.

- Eva Schultze-Berndt and Nikolaus P Himmelmann. 2004. Depictive secondary predicates in crosslinguistic perspective. *Linguistic Typology*, 8(1).
- Masayoshi Shibatani. 2009. On the form of complex predicates: toward demystifying serial verbs. In Johanes Helmbrecht, Yoko Nishina, Yong-Min Shin, Stavros Skopeteas, and Elisabeth Verhoeven, editors, *Form and Function in Language Research*, pages 255–282. Mouton de Gruyter.
- E. Slager. 1995. *Grammatica Spaans*. Spectrum.
- John C. Street. 1963. *Khalkha Structure*. Indiana University Press.
- Koichi Takezawa. 1993. Secondary predication and locative/goal phrases. In *Japanese Syntax in Comparative Grammar*. Kurosio.
- 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.
- René van den Berg. 1989. *A Grammar of the Muna Language*. Foris.
- James K Watters. 2017. Verb-verb compounds and argument structure in Tepehua. In *Argument Realisation in Complex Predicates and Complex Events*, pages 277–304. John Benjamins Publishing Company.

Bringing Information Structure to Universal Dependencies

Nikolett Mus¹, Andrew Dyer², Claudia Corbetta³, Sylvain Kahane⁴

¹ ELTE Research Centre for Linguistics, ² Saarland University,

³ University of Bergamo/Pavia, ⁴ Paris Nanterre University

mus.nikolett@nytud.elte.hu, andrew.dyer@uni-saarland.de,

claudia.corbetta@unibg.it, skahane@parisnanterre.fr

Abstract

The aim of the paper is to present a first attempt at annotating Information Structure roles in syntactic treebanks of the Universal Dependencies collection, discussing theoretical considerations and practical methodological questions while presenting our core annotation principles. We focus on constructions in which Topic or Focus is overtly marked through (morpho)syntactic means. The proposed annotation is illustrated using examples from five languages: Wolof, Japanese, Tundra Nenets, Hungarian, and Italian.

Keywords: information packaging, Information Structure, Topic, Focus, Contrast

1. Introduction

The aim of this study is to present a recent initiative and ongoing work on the annotation of Information Structure roles in existing and future Universal Dependencies (UD) treebanks (De Marneffe et al., 2021). As discussed in greater detail in Section 3, we understand Information Structure as the organisation of utterances relative to the discourse context. Our work focuses on annotating the roles that individual sentence elements play in Information Structure, rather than attempting to represent the full Information Structure of the sentence.

In line with Topic–Focus Articulation (henceforth TFA) (Hajičová et al., 1998; Firbas, 1992), we adopt Topic as the element that identifies what the sentence is about and anchors the utterance in the shared discourse. We depart from TFA, however, in our treatment of Focus. Rather than annotating the entire focus domain, we mark only a single element that can be interpreted as Focus, that is, the element that is emphatic or carries the most prominent informational or semantic weight within the utterance, if such an element is present. This prominence is primarily understood in communicative terms, though it may also be reflected prosodically or structurally. In this sense, our notion of Focus aligns with approaches that treat focus as a distinguished element associated with alternatives or contrast (Rooth, 1992; É. Kiss, 1998; Krifka, 2008). In addition to Topic and Focus, we introduce a third notion, Contrast, capturing elements interpreted against a set of alternatives and highlighting a distinction or opposition within the discourse.

A substantial body of cross-linguistic and typological research has examined how languages encode discourse-related distinctions. These studies document considerable variation and emphasise the diversity of strategies employed across languages,

ranging from prosodic marking to morphological and syntactic means (Neeleman et al., 2009; Zimmermann and Féry, 2010). In some cases, languages develop specialised constructions whose primary function is the marking of Information Structure roles. In others, elements or constructions with independent grammatical functions are extended to serve Information Structure-related purposes as well. Together, these patterns demonstrate that Information Structure can systematically shape surface syntactic and morphological structure.

Despite this, current UD treebanks, while already invaluable resources for syntactic typology and comparative research, remain limited in that they represent syntactic structures largely in isolation from discourse-level information. As a consequence, the influence of Information Structure on word order, constituent realisation, and constructional variation often remains implicit and inaccessible within existing annotation frameworks. This limitation hinders both fine-grained typological comparison and the interpretation of certain syntactic patterns, particularly in languages whose syntax and grammar have not yet been exhaustively described or analysed.

This is not to say that there have been no attempts to annotate Information Structure in existing corpora, or to create IS-specific corpora (these efforts are discussed in more detail in Section 2). Our approach, however, departs fundamentally from these earlier initiatives. Rather than attempting to identify and annotate complete Information Structural configurations, we restrict our annotation to explicitly marked elements, that is, linguistic units that directly signal, or indirectly indicate, an Information Structure role.

The primary goal of our work is to define Information Structure roles in a precise, operational manner. While this definition is necessarily pre-

liminary and may be refined based on the data, it provides a foundation for both theoretical and practical annotation guidelines. Introducing an Information Structure layer into treebanks can enhance the descriptive and explanatory power of syntactic annotation, enable systematic cross-linguistic comparison of discourse-sensitive structures, and support the refinement of existing typological generalisations. At the same time, explicit annotation of discourse-marked constructions allows these elements to be filtered out, providing a clearer view of a language's unmarked or basic word order patterns.

Beyond its relevance for theoretical linguistics, this effort also has practical and computational applications. Explicit annotation of Information Structure roles can inform research in translation studies and machine translation, support the development of language learning tools, and be particularly valuable for endangered or under-resourced languages, where annotated corpora can contribute to documentation and revitalisation efforts.

As this work is at an early stage, the present paper focuses on outlining our conceptual foundations, discussing the theoretical and practical principles, and illustrating the proposed approach through a small number of case studies. These examples demonstrate the methodological considerations involved in deciding what to annotate and how. Thus, this paper emphasises conceptual clarification and annotation design rather than large-scale annotated data.

2. Related Work

Several previous initiatives have sought to incorporate Information Structure or related discourse-level distinctions into corpus annotation frameworks. These approaches differ considerably in their theoretical assumptions, the units they annotate, and the way Information Structure is represented in relation to syntactic structure.

A number of proposals aim to provide a comprehensive annotation of Information Structure, typically operating over constituents rather than individual tokens. [Dipper et al. \(2007\)](#), for instance, develop a framework for annotating both Information Structure and information status across languages, introducing coarse- and fine-grained labels (e.g. *aboutness* vs. *frame-setting* topics, *new information* vs. *contrastive* focus), along with pragmatic diagnostics for their identification. Similarly, [Mírovský et al. \(2013\)](#) extend the Prague Dependency Treebank ([Bejček et al., 2013](#)) with Topic–Focus annotation grounded in the Prague School tradition [Hajičová et al. \(1998\)](#), and [Bohnet et al. \(2013\)](#) introduce a comparable Topic–Focus annotation layer for the Penn Treebank.

Other work addresses more specific aspects of

Information Structure or its formal and computational modelling, rather than proposing full annotation schemes. [Song \(2017\)](#) provide a multilingual HPSG-based formalisation of Topic–Focus Articulation, with a detailed taxonomy of Information Structure types and their marking strategies, and demonstrate its usefulness in machine translation. [Stede and Mamprin \(2016\)](#) introduce an Information Structure annotation layer in the Potsdam Commentary Corpus, focusing in particular on fine-grained distinctions among types of Topic. [Booth \(2022\)](#) take a methodological perspective, outlining desiderata for Information Structure annotation, especially with respect to discourse-sensitive phenomena in subordinate clauses, and illustrate their proposal on a corpus of Historical Low German.

Some of these resources have been integrated into broader infrastructures such as CorefUD ([Novák et al., 2025](#)), which brings together coreference corpora in the CoNLL-U format and enables the joint analysis of coreference, Information Structure, and dependency syntax.

While these approaches provide valuable insights into Information Structure annotation, they are often challenging to reconcile with token-based frameworks such as Universal Dependencies, which do not explicitly represent constituent structure. By aiming to annotate entire IS configurations, they can introduce ambiguity and require discourse-specific interpretation. In contrast, our role-based approach focuses solely on elements that are overtly marked, treating them as anchor points for IS roles. This strategy avoids speculative assignment, limits potential misinterpretation, and maintains compatibility with UD while supporting cross-linguistic comparison.

3. The concept and definition of Information Structure roles

In this section, we outline the conceptual foundations of Information Structure as it is treated in our annotation framework.

Information Structure concerns how utterances are organised relative to the discourse context, in particular how speakers indicate what a sentence is about and what information is asserted regarding that element. Theoretically, sentences can be conceptually divided into two functional fields: one that establishes the discourse anchor, signalling what the utterance is about, and one that carries the assertive contribution, specifying the new information associated with that anchor. While this division provides a useful generalisation, there are non-prototypical constructions in which a clear discourse anchor is absent or difficult to identify (e.g., in existential clauses such as *There are ghosts*).

A variety of theoretical frameworks have been

proposed to capture this two-part division of sentences into a discourse anchor and an assertive contribution, including Topic–Comment (Gundel, 1988), Given–New (Chafe, 1976), Theme–Rheme (Weil, 1879; Firbas, 1964; Mel’čuk, 2001), and Common Ground-based approaches (Stalnaker, 2002). Each of these frameworks offers valuable insights into how speakers organise utterances relative to discourse, yet they differ in their conceptualisations of key notions, their units of analysis, and their formal assumptions. This diversity, while theoretically rich, presents practical challenges for annotation, particularly within the constraints of a Universal Dependencies framework, where abstract fields or domains may not align straightforwardly with syntactic structures representing dependency relations between syntactic units.

Due to the diversity of conceptual frameworks and the fact that definitions are often either underspecified or fail to adequately account for exceptional and ambiguous cases, previous attempts to annotate Information Structure have not yet fully resolved these issues (Paggio, 2006; Dipper et al., 2007; Ritz et al., 2008; Lüdeling et al., 2016). While these approaches have provided important insights, their theoretical assumptions and representational choices have limited their applicability in a consistent annotation setting.

Importantly, our goal is not to construct a dedicated Information Structure corpus by selectively collecting prototypical constructions. Rather, we aim to annotate Information Structure roles within existing corpora, marking those instances where relevant roles can be identified with sufficient clarity. Therefore, we adopt a hybrid approach, informed by insights from the existing literature, but tailored to the constraints and goals of dependency-based annotation. Within this framework, we define the core Information Structure roles as follows:

- (1) **Topic**: the element that identifies what the sentence is about and anchors the utterance in the shared discourse.¹
- (2) **Focus**: the most prominent element in an utterance, i.e., the element that is made emphatic or highlighted.²

¹This definition follows the Topic notion in TFA as proposed by Hajičová et al., 1998; Firbas, 1992.

²Here, emphasis is understood in communicative terms rather than purely prosodic or structural, although prosody or structure may also realise it. Unlike the classical TFA approach, we do not annotate the entire focus domain, but rather the single element within it, if such an element is present. This perspective aligns with approaches that treat focus as a marked element associated with alternatives or contrastive relations (Rooth, 1992; É. Kiss, 1998; Krifka, 2008).

At the current stage of analysis, we do not further subdivide these categories into more fine-grained types. Although this simplification may obscure certain distinctions, it provides a necessary starting point for developing a consistent and operational annotation scheme.

In addition to Topic and Focus, we introduce Contrast as a further dimension of annotation. Contrast is not treated as an independent Information Structure category, but as a property that can be associated with either Topic or Focus, reflecting the presence of alternatives in the discourse.

We do not assume that every sentence contains a Topic or Focus, and we annotate only elements that are explicitly marked. By restricting annotation to these identifiable elements rather than representing full Information Structure configurations, we obtain anchor points that can be systematically related to syntactic structure. This approach allows us to examine how syntactic positions correlate with discourse functions and to empirically evaluate assumptions that have often been proposed without extensive corpus-based evidence. At the same time, it provides a minimal, reproducible model with clearly defined roles that can be consistently applied across languages, enabling both the reliable identification of Information Structure roles and the exploration of cross-linguistic patterns of information packaging. In this way, the framework lays the groundwork for empirically grounded typological generalisations concerning Topic, Focus, and Contrast.

4. Annotation of Information Structure roles

In our annotation scheme, both the units that carry Information Structure roles and the morphological or syntactic elements that mark these roles are annotated. This captures the overt realisation of the role, allowing the annotation to represent both the discourse function and its grammatical encoding.

In some cases, these two levels coincide, for example, for inherently marked IS roles or when bound morphemes are part of the constituent itself and directly indicate its IS function, such as focus particles. In other cases, an entire string of tokens (a phrase) may collectively realise a role. In such instances, we annotate the governor (head) as an anchor for the corresponding role, which can then be related to syntactic structure or combined to approximate the full phrasal unit.

As the UD framework is inherently token-based, we use the `MISC` column to record Information Structure roles ensuring that discourse-functional distinctions are recorded without conflicting with the syntactic and morphological core of UD.

For the annotation of role-bearing constituents, a

single umbrella label, `ISRole`, is used. The values indicate the discourse role: `ISRole=Top` for Topic, `ISRole=CTop` for contrastive Topic, `ISRole=Foc` for Focus, and `ISRole=CFoc` for contrastive Focus. This follows UD’s feature/value design: the label functions as the category (analogous to `Animacy`, `Number`, or `Case`), and the value indicates the specific role.

Elements that indicate an IS role are annotated with the labels `ISMarker[(C)Top]` and `ISMarker[(C)Foc]`.

Morphological and syntactic markers of IS roles are annotated in accordance with UD principles, with attention to whether the specific grammatical function or construction type is identifiable.

For morphology, when the grammatical function of a (bound) morpheme is known, it is recorded explicitly. For example, an accusative case marker functioning to signal a topicalised constituent can be annotated as `ISMarker[Top]=Case`. In instances where only the morphological type is known, but the precise grammatical function cannot be determined, we assign an underspecified value, such as `ISMarker[Top]=Morph` or `ISMarker[Foc]=Morph`, thereby preserving the morphological information without overcommitting to a specific function. For clarity and to retain more detailed information, we recommend optionally providing a sublabel that records the surface form of the morph, e.g., `ISMarker[Foc]Form={form}`.

For syntactic marking, we distinguish between construction-based and position-based strategies. Construction-based marking is applied when a particular syntactic construction signals an IS role. When the construction type is identifiable, the root of the construction is annotated with a specific value, for instance, `ISMarker[Top]=Inversion` or `ISMarker[Foc]=Cleft`. When the precise construction is unknown or cannot be determined with confidence, an underspecified value, `ISMarker[Top]=Construction` or `ISMarker[Foc]=Construction`, is used. Position-based marking is applied when a constituent with an IS role occupies a dedicated syntactic position, such as preverbal focus in Hungarian; in this case, the constituent is annotated with the value `ISMarker[Foc]=Position`.

This approach allows us to capture overt morphological and syntactic cues of Information Structure systematically, while providing flexibility to accommodate uncertain or partially analysed phenomena, ensuring consistency and cross-linguistic applicability in the annotation process.

5. Principles of Information Structure annotation

On the basis of our definitions and concepts, we formulated a concise set of principles to guide the annotation of Information Structure roles ensuring consistency, cross-linguistic comparability, and compatibility with the Universal Dependencies framework. First, only explicitly marked elements are annotated, covering overt morphosyntactic and structural cases, while unmarked elements are left unannotated. Second, annotation is function-oriented: structurally distinct constructions receive the same label if they fulfil the same role. Third, comparability across languages is achieved by treating universality at the level of functional roles rather than formal realisation, supporting typological exploration and cross-linguistic analysis. Fourth, contextual information is considered to confirm whether an element is overtly marked as Topic, Focus (or Contrast), without inferring roles in the absence of clear evidence. Fifth, annotation is limited to the core roles, i.e. Topic, Focus, and Contrast, while finer subtypes, such as frame-setting versus aboutness topics (Fery and Krifka, 2008), are not distinguished. Sixth, both the units carrying Information Structure roles and any morphological or syntactic markers signalling these roles are annotated, capturing the discourse function alongside its grammatical encoding. Seventh, when an entire phrase realises an Information Structure role, only the head of the phrase is annotated. Finally, all annotations are encoded in the `MISC` column, which is designed for extra, non-syntactic functional information in UD.

6. Case studies

In this section we provide case studies of some languages in Universal Dependencies (or pending release) with known morphosyntactic mechanisms of Information Structure. We do not consider this selection of languages to be fully representative of all possible strategies for marking Information Structure. Instead, the sample was chosen to capture a diversity of marking practices, illustrating how different languages employ syntactic and/or morphological strategies to encode Information Structure roles. By including a range of strategies here, the examples allow us to demonstrate annotation in practice. We note that, for some of the languages included in this study, the available UD treebanks are still relatively small. Where necessary, the examples presented here are drawn from our own data. Comparable instances can often be found in the treebanks, but minimal pairs or carefully selected examples are chosen to ensure clarity and facilitate explanation.

6.1. Wolof (Niger-Congo)

Wolof is a language where focalisation has been grammaticalised (Robert, 1991, 2000) and almost all sentences contain a particle indicating the focus of the sentence. Only 12% of the sentences in UD_Wolof-WTB treebank (Dione, 2019) have a main clause without a particle (Bondéelle and Kahane, 2021) and this constrained construction is mainly used in narrative sequences (Robert, 1991). The focusing particles of Wolof are verbal particles, which are analysed as complementizers by Torrence (2005) and annotated AUX in the UD_Wolof treebank developed by Dione (2019). Wolof has a very rigid word order, but the word order depends on the particle that has been selected. Two particles, *a* and *la*, are respectively used to focus the subject (3a) and a complement (3b); they occupy the second position, the first position being occupied by the focused element. Dislocated units or modifiers can precede the first position, but are prosodically detached in spoken data and generally separated from the sentence nucleus by a comma in written data.

- (3) a. *Bu jëkk, nguur*
 when be_first, power
 ISRole=CTop
 ISMarker[CTop]=Dislocation
googu, ay socé ak ay
 CL.ANAPH IND.CL Mandinka with IND.CL
 ISRole=Foc
séeréer, ñu =a
 Serer, S3PL =PART
 ISMarker[Foc]=Morph
ko yor =oon.
 O3SG hold =PAST.

‘Originally, this power, the Mandinka and Serer, it was them that held it.’ (WTB-3)

- b. *Noonu it la*
 CL.ANAPH too PART.3SG
 ISRole=Foc ISMarker[Foc]=Morph
àtte ci am tudd.
 decide LOC CL.IND name

‘That’s also the way he decides on a name.’ (WTB-1902)

The realisation of the subject is very constrained and it is generally dislocated as in (3a) (Bondéelle and Kahane, 2020). The dislocation of other elements marks a topicalisation.

Interestingly, there are also two particles, *na* and *da*, that respectively focus the verb and the whole sentence. The particle *na* is not recognised as a focus particle by most authors, but Bondéelle and Kahane (2021) postulate that it could be a

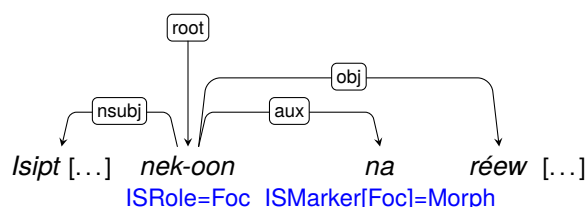
translation effect, because the focus of the verb is not marked and is the default interpretation of a sentence in languages used for the description of Wolof, such as French and English. The particle *na* occupies the second position, the verb being in the first position, while *da* occupies the first position. The contrast between *na* and *da* can be illustrated by the following examples from Robert (2000).

- (4) a. *Ragal na ko.*
 be_afraid PART him.
 ‘He is afraid of him.’, lit. it’s the fear he feels towards him.
 b. *Dafa ragal.*
 PART.3SG be_afraid.
 ‘He is a coward.’

The particle *na* is used to focus the imperfective auxiliary *di*, giving the locution *dina* that marks the future tense. The following example shows a focus on the past tense.

- (5) *Isipt gu yàgg ga*
 Egypt CL.REL last(V) CL.DEF
nekk-oon na réew mu
 be-PAST PART.S3SG country CL.REL
woomle, [...]
 be_prosperous, [...]

‘Ancient Egypt WAS a prosperous country, [...]’ (WTB-245)



The particle *da* can be analysed as a sentence-focus, in the sense of Lambrecht and Polinsky (1997). When it is used with an action verb, it generally appears in a sequence of sentences and takes a causal value, as in (6) and (7a) (Robert, 2000).

- (6) *Sama jëkkër nekk-u fi,*
 my husband be-NEG.S3SG here,
dafa dem àll ba.
 PART.S3SG go forest CL.DEF

‘My husband isn’t here; he’s gone into the bush.’

The negation has different realisations according to the focus. It can be realised by the suffix *-u*, the negative value of *na* (7a); by the particle *du*, the

negative value of *da* (7b);³ or by a negation *-ul* on the verb when the focus is elsewhere (7c).

- (7) a. *Tóx-u-ma, da-ma-y fo.*
 smok-NEG-1SG, PART-1SG-IMPF play.
 'I'm not smoking, I'm playing (with the cigarette).', lit. what I'm doing isn't smoking, it's playing.
- b. *Du-ma tóx.*
 NEG.PART-1SG smoke.
 'I don't smoke, I'm not a smoker.'
- c. *ma =a tóx-ul.*
 1SG =PART smok-NEG.
 'It's me who don't smoke.'

UD_Wolof-WTB already has a feature *FocusType*, though this is unique to the language. The values are *Subj* for *a*, *Comp* for *la*, and *Verb* for *da*. We propose to associate the *Verb* to *na* and to put *Sent* on *da*. This feature is partly redundant with our annotation of *ISRole*, but not completely, because *ISRole[foc]* will be on the verb with both *na* and *da* and we do not specify whether it applies on the verb only or a larger domain.

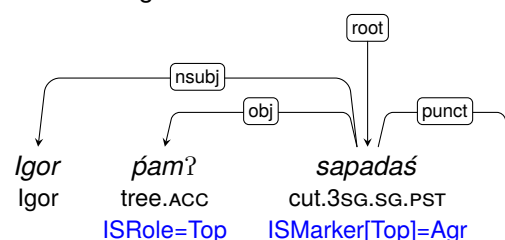
6.2. Tundra Nenets (Uralic, Samoyedic)

Tundra Nenets, an endangered, and relatively underdescribed language spoken in the Arctic regions of Russia, appears to employ morphological strategies for encoding Information Structure roles. On the basis of existing descriptions, both Topic and Focus – or at least certain subtypes thereof – can be expressed through morphological marking (Nikolaeva, 2014).

Accordingly, one particular type of Topic – specifically, the lexical object of the clause when it serves a topical function – appears to be morphologically marked in the language through the so-called objective verb conjugation expressed by a suffix added to the verb. This marker simultaneously encodes agreement with both the subject and the object. The objective conjugation system contrasts with forms that encode agreement with the subject alone, and it is typically restricted to third-person objects (Dalrymple and Nikolaeva, 2011; Nikolaeva, 2014). The minimal pair presented below illustrates this contrast: in (8a) the transitive verb agrees only with its subject in person and number, and in (8b) it agrees both with its subject and with the third-person topical object in number.

- (8) a. *Igor páam? sapas*
 Igor tree.ACC cut.3SG.PST
 'Igor cut a/the tree.'
- b. *Igor páam? sapa-da-s*
 Igor tree-ACC cut-3SG.SG-PST
 '(As for a/the tree) Igor cut a/the tree.'

In such cases – that is, when the verb agrees with an object bearing a Topic role – the object receives the label and value *ISRole=Top*. Since, within the UD framework, agreement morphology is not segmented as a separate morpheme, the verb is annotated with the label *ISMarker[Top]=Agr*. The example above can therefore be represented by the following UD tree.



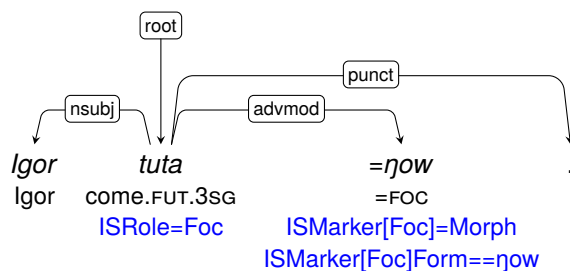
In addition to the marking of object Topics, Tundra Nenets also exhibits clitics and/or affixes that serve to encode Focus, or at least certain subtypes of foci (Nikolaeva, 2014). One such element is the so-called focus marker *=ŋow* / *=ej*, which attaches to verbs and appears as the final morpheme in the word form. This affix functions to emphasise the action expressed by the verb and, more specifically, to highlight the truth value of the proposition, see (9) and Nikolaeva (2014).

- (9) *Igor tu-ta=ŋow*
 Igor come-FUT.3SG=FOC
 '(Indeed) Igor WILL come.'

In this instance, the verb bearing the suffix is annotated as *ISRole=Foc*, while the overt marker itself is labeled *ISMarker[Foc]=Morph*. As mentioned above, for clarity and to retain detailed information, the surface form of the marker can optionally be recorded using the sublabel *ISMarker[Foc]Form==ŋow*⁴. This is illustrated in the following UD tree:

³*da* and *du* are generally analysed as constructed on the imperfective auxiliary *di*, *da* as *di* plus the focus marker *a* and *du* as *di* plus the negative focus marker *-u* (Sauvageot, 1965).

⁴Given its position as a closing suffix on the verb and its role in marking verum (emphasis on the truth of the proposition), we annotate the clitic as an adverb (ADV) and assign it an *advmod* relation to the predicate.



- (12) a. *Niku wa tabenai*
meat TOP eat-NEG
'I don't eat meat.' (lit: 'As for meat, [I] don't eat [it].')
- b. *Niku o tabenai*
meat ACC eat-NEG
'I don't eat meat.' (lit: '[I] don't eat meat.')

6.3. Japanese (Japonic)

In Japanese, phrases within a sentence are typically marked on their right edge with a *particle*⁵ corresponding to their grammatical roles within the sentence; for example, as case markers, parallel markers or binding markers (Martin, 2004). Among these is the topic-marking particle *wa*.

In many sentences, *wa* can be used interchangeably with the nominative/subject marker *ga*, as the *topic* of a sentence is often the same as its *subject*.

- (10) *Hanako {ga/wa} konakatta*
Hanako {NOM/TOP} come-NEG-PST
'Hanako didn't come.'

A *wa*-marked constituent may also appear as a topic in double-subject constructions (Kumashiro and Langacker, 2003), where a main constituent is the subject but an outer subject is also required, as in (11) where the subject of the adjective *suki* is *Takashi*, but the thematic *wa* indicates the semantic meaning that Hanako is the agent liked by *Hanako*.

- (11) *Hanako wa Takashi ga suki*
Hanako TOP Takashi NOM liked(ADJ)
'Hanako likes Takashi.' (lit: 'As for Hanako, Takashi is liked [by her]')

(10) and (11) are unmarked; there is nothing unusual in Japanese about a subject being topicalised, nor a topic being used as an outer subject. In a more marked use, *wa* may also be used to topicalise other arguments, such as direct and indirect objects. (12a) shows an example of a semantic object being marked as a topic, while (12b) shows the more typical accusative-marked type. Obliques may also be made topics through the use of *wa*, as in (13).⁶

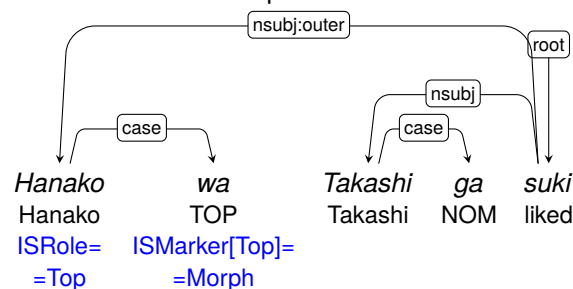
⁵Particle is the term typically used in Japanese linguistics, but in Universal Dependencies they are generally tagged as ADP.

⁶Note that in the oblique (13), a locative marker *ni* is also used before *wa*, while the object in (12a) the object marker *o* is not used before *wa*, and would be ungrammatical if it were.

- (13) *Ôsaka (ni) wa ippai aru*
Osaka (LOC) TOP full exist
'There's loads of it in Osaka.' (lit: 'As for (in) Osaka, loads [of it] exist.')

The exact function of this particle in terms of Information Structure is debated. Oshima (2009) posits that it is an amalgam of a topic-marker and a ground marker, with different roles depending on what it marks. In the case of a *wa*-marked subject, this merely indicates that the subject is part of the ground, but not necessarily a topic; in the case of a *wa*-marked direct object, it indicates that the object is both part of the background and a topic. We adopt the treatment of it as a topic marker, mindful that this may overgenerate topic information.

Japanese as annotated in Universal Dependencies tokenises at the morpheme level and does not attach any morphological features to particles (Tanaka et al., 2016), and thus particles are simply separate tokens tagged as ADP, with a *case* relation to their head, and no additional morphology. For our purposes, this means that we can add the *ISMarker[Top]* attribute to the particle token itself, with the value of *Morph*, while the *ISRole* is added to the head of the phrase.



There are other Information Structure-marking strategies in Japanese that are worth consideration, such as scrambling (Maki et al., 1999) and right-dislocation (Takita, 2014). Maki et al. posit that scrambling in particular may have effects on the role played by *wa*, changing the phrasal unit that it attaches to from a simple topic to a contrastive focus, in which case we may use the *CFoc* value for *ISRole*. We do not commit to this yet, and leave it for further deliberation.

6.4. Hungarian (Uralic)

In Hungarian, focus expressing exhaustive identification occupies a dedicated position within the predicate phrase: it must appear immediately before the finite verb (É Kiss, 2002; É Kiss et al., 2003).

A clear syntactic diagnostic for exhaustive identification comes from the behaviour of so-called verb modifiers (VM), which reveal that focus occupies a dedicated preverbal position. In the Hungarian grammatical tradition, verb modifiers are treated as adverbial elements. They modify the lexical or aspectual meaning of the verb, often contributing telicity, directionality, or other Aktionsart-related distinctions (É Kiss et al., 2003). In neutral clauses, the predicate phrase typically begins with a verb modifier followed by the verb (VM–V order), see (14).

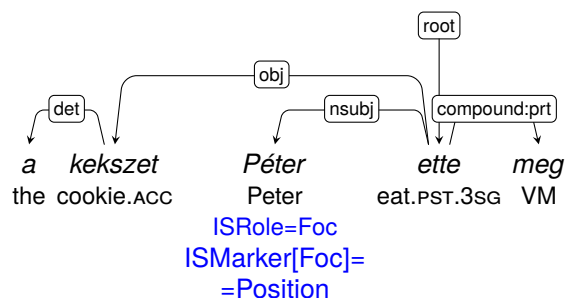
- (14) *A kekszet Péter meg-ette.*
 the cookie.ACC Peter VM=eat.PST.3SG
 ‘Peter has eaten the biscuit.’

Although the verb and its modifier may form a lexical and phonological unit when adjacent, the modifier is syntactically independent and can be separated from the verb. We therefore analyse it as a clitic (rather than an affix, as in traditional analyses). This independence is clearly visible in focus constructions, where a constituent of the clause bearing exhaustive focus must occupy the immediate preverbal position. Since this position can host only one phrasal unit, the verb modifier cannot remain there and is displaced to a postverbal position. The alternation between neutral VM–V order and focus-induced Foc–V–VM order thus reflects the structural requirement that exhaustive focus be realised in the immediate preverbal slot, see (15) and (É Kiss et al., 2003; É Kiss, 2002).

- (15) *A kekszet Péter ette meg.*
 the cookie.ACC Peter eat.PST.3sg VM
 ‘PETER has eaten the biscuit.’

In our annotation scheme, such focused units are marked as `ISRole=Foc` and `ISMarker[Foc]=Position`, indicating that focus is encoded configurationally. An example of this analysis is provided in the following UD tree.⁷

⁷Since verb modifiers in Hungarian behave similarly to compounds in several respects, they have been analysed as `compound:preverb` in the Hungarian treebank. We prefer to use the `compound:prt` relation for this particle, analogously to how such particles are treated in German and English treebanks.



It should also be noted that in Hungarian, topics occupy a dedicated syntactic position at the left edge of the clause (É Kiss et al., 2003; É Kiss, 2002). For reasons of space, we do not discuss this construction in detail here. In such cases, however, the topicalised constituent is annotated with `ISRole=Top`, and `ISMarker[Top]` is likewise set to `Position`.

6.5. Italian (IE – Romance)

In Italian, topic and focus are not morphologically encoded. Rather, these information-structural functions are realised through syntactic role, constituent order, and prosody (Benincà et al., 1988; Palermo, 2013; Lombardi Vallauri, 2009, 2014).⁸

More specifically, under canonical word order, corresponding to the (S)VO pattern in Italian, the topic is typically associated with subject function, whereas the focus is commonly realised in object position. Constituent order also plays a central role, as the topical element generally occupies a left-peripheral position (Palermo, 2013). In example (16), the topic corresponds to the subject *Carlo*, which occupies the left-peripheral position, while the focus coincides with the direct object *la lettera* ‘the letter’, which appears in rightmost position.

- (16) *Carlo ha spedito la lettera*
 Carlo have.PRS.3SG mail.PTCP the letter
 ‘Carlo has mailed the letter.’

To signal a misalignment with respect to canonical constituent order, Italian employs marked constructions. Given the wide range of such constructions, only two types are presented here, one per category, for illustrative purposes.

Among topic-marking constructions is left dislocation (De Santis and Prandi, 2019), so termed because a constituent is displaced to the left periphery and placed outside the clause core; in written

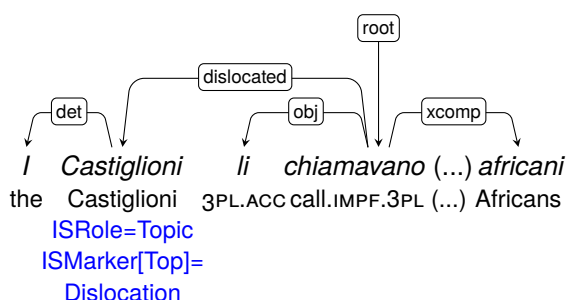
⁸An annotation work on the Information Structure of spoken Italian based on Informational Patterning Theory (Cresti, 2000) is represented by the C-ORAL-ROM corpus (Cresti and Moneglia, 2005); see also (Cresti and Moneglia, 2010).

texts, it is typically set off by a comma. The dislocated constituent is resumed within the clause by a coreferential resumptive pronoun⁹.

- (17) *I Castiglioni li chiamavano per scherzo "gli africani"*
 the Castiglioni ACC.3PL call.IMP.F.3PL for
 joke "the africans"

'They used to call the Castiglioni "the Africans" as a joke.' (isst_tanl-404)

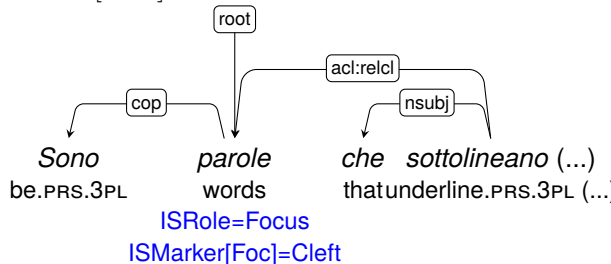
In Example (17), *I Castiglioni* 'the Castiglioni' is left-dislocated and functions as the topic of the sentence. Along with the dependency relation *dislocated*, it is marked as *ISRole=Topic* and labeled with *ISMarker[Top]=Dislocation*, as shown in the UD tree.



Among the strategies used in Italian to encode focus is the cleft sentence construction (Panunzi, 2009), whereby the unmarked sentence is divided into two clauses: the first is characterised by the verb *essere* 'to be' and the focused element, while the second, introduced by *che* 'that', contains the remainder of the sentence. Cleft sentences employ both intonational and syntactic marking.

- (18) *Sono parole che sottolineano (...)*
 be.PRS.3PL words that underline.PRS.3PL (...)
 'They are words that underline (...)' (markit-945)

In Example (18), the clefted element *parole* 'words' is marked as *ISRole=Focus* and specifies the construction, namely via the label *ISMarker[Foc]=Cleft*.



⁹However, the presence of the resumptive clitic is not always obligatory, especially when the dislocated constituent does not correspond to a direct object (Palermo, 2013).

7. Future work

The immediate next step in this project is to finalise semi-definitive annotation guidelines, providing a clear framework for annotators while allowing for future refinement. Our current focus has been on elements that are overtly marked for Information Structure. Moving forward, we will also annotate inherently marked elements that can be reliably extracted from the lexicon, such as operators and other IS-triggering items.

These guidelines will then be applied to selected UD treebanks. We will prioritise languages where the strategies for marking Information Structure roles are reasonably well understood, while striving to ensure that the sample represents a broad spectrum of morphosyntactic marking strategies.

Once annotation is underway, we will compare overtly marked versus inherently marked elements to identify additional morphosyntactic patterns and correlations. These findings will allow us to formulate generalisations about the interplay between marking strategies and Information Structure roles. Finally, we will situate our results within the broader context of existing approaches, comparing patterns and outcomes to evaluate consistency and typological relevance.

8. Conclusion

We have presented a proposal for annotating Information Structure within the Universal Dependencies framework that preserves the integrity of the existing UD format while providing explicit annotation of constructions and elements that clearly mark Information Structure roles. Our scheme is deliberately minimal and modular, focusing on overtly marked elements as well as inherently marked lexical triggers where they can be reliably identified. This approach allows for consistent annotation across languages without introducing speculative or discourse-specific interpretations.

We illustrated the scheme with case studies from five typologically diverse languages – Wolof, Japanese, Tundra Nenets, Hungarian, and Italian – demonstrating how the framework can capture cross-linguistic variation in Topic, Focus, and Contrast marking strategies. While the scheme is currently tentative, it provides a practical foundation for systematic annotation and comparison, and can be refined or expanded as further data and insights become available.

By anchoring Information Structure annotation to overtly marked elements and clear morphosyntactic triggers, this framework provides a resource for both theoretical linguists and UD users interested in the interaction of discourse and syntax. Beyond immediate annotation, it lays the ground-

work for empirical investigations into the correlation between syntactic structure and discourse roles, the typology of marking strategies, and the development of automated tools for Information Structure analysis.

Acknowledgments

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

Ethical considerations

The authors are unaware of any ethical issues arising from this project.

References

Bibliographical References

- Paola Benincà, Giampaolo Salvi, and Lorenza Frison. 1988. L'ordine degli elementi della frase e le costruzioni marcate. In Lorenzo Renzi, editor, *Grande grammatica italiana di consultazione*, volume 1, pages 115–225. Il Mulino, Bologna.
- Bernd Bohnet, Alicia Burga, and Leo Wanner. 2013. [Towards the annotation of Penn TreeBank with information structure](#). In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1250–1256, Nagoya, Japan. Asian Federation of Natural Language Processing.
- Olivier Bondéelle and Sylvain Kahane. 2020. Subjecthood and annotation: The cases of french and wolof. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW)*, pages 34–45.
- Olivier Bondéelle and Sylvain Kahane. 2021. Les particules verbales du wolof et leur combinatoire syntaxique et topologique. *Bulletin de la Société de Linguistique de Paris*, 115(1):391–465.
- Hannah Booth. 2022. [Desiderata for the annotation of information structure in complex sentences](#). In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 31–43, Marseille, France. European Language Resources Association.
- Wallace Chafe. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In *Subject and topic*. Academic Press.
- Emanuela Cresti. 2000. *Corpus di Italiano Parlato*. Accademia della Crusca, Firenze.
- Emanuela Cresti and Massimo Moneglia, editors. 2005. *C-ORAL-ROM. Integrated reference corpora for spoken Romance languages*. John Benjamins, Amsterdam.
- Emanuela Cresti and Massimo Moneglia. 2010. Informational patterning theory and the corpus-based description of spoken language. the compositionality issue in the topic-comment pattern. In Massimo Moneglia and Alessandro Panunzi, editors, *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective*. Firenze University Press, Firenze.
- Mary Dalrymple and Irina Nikolaeva. 2011. *Objects and information structure*. Cambridge University Press.
- Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.
- Cristiana De Santis and Michele Prandi. 2019. *Manuale di linguistica e di grammatica italiana*. UTET.
- Cheikh M. Bamba Dione. 2019. Developing universal dependencies for wolof. In *Proceedings of the Third Workshop on Universal Dependencies (UDW)*, *SyntaxFest*, pages 12–23.
- Stefanie Dipper, Michael Götz, and Stavros Skopeteas. 2007. Information structure in cross-linguistic corpora: Annotation guidelines for phonology, morphology, syntax. *Interdisciplinary Studies on Information Structure: Working Papers of the SFB632*.
- Katalin É Kiss. 2002. *The syntax of Hungarian*. Cambridge University Press.
- Katalin É Kiss, Péter Siptár, and Ferenc Kiefer. 2003. Új magyar nyelvtan.
- Caroline Fery and Manfred Krifka. 2008. [Information structure: Notional distinctions, ways of expression](#). *Unity and Diversity of Languages*, pages 123–136.
- Jan Firbas. 1964. On defining the theme in functional sentence analysis. *Travaux Linguistiques de Prague*, 1:267–280.
- Jan Firbas. 1992. *Functional Sentence Perspective in Written and Spoken Communication*. Cambridge University Press, Cambridge.
- Jeanette K Gundel. 1988. Universals of topic-comment structure. *Studies in syntactic typology*, 17(1):209–239.
- Eva Hajičová, Barbara H. Partee, and Petr Sgall. 1998. *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*. Kluwer Academic Publishers, Dordrecht.
- Manfred Krifka. 2008. Basic notions of information structure. *Acta Linguistica Hungarica*, 55(3-4):243–276.

- Toshiyuki Kumashiro and Ronald Langacker. 2003. [Double-subject and complex-predicate constructions](#). *Cognitive Linguistics*, 14(1):1–45.
- Knud Lambrecht and Maria Polinsky. 1997. Typological variation in sentence-focus constructions. *Chicago Linguistic Society*, 33(2):189–206.
- Edoardo Lombardi Vallauri. 2009. *La struttura informativa. Forma e funzione negli enunciati linguistici*. Carocci, Roma.
- Edoardo Lombardi Vallauri. 2014. The topologic hypothesis of prominence as a cue to information structure in Italian. In Salvador Pons Bordería, editor, *Discourse Segmentation in Romance Languages*, pages 219–241. John Benjamins, Amsterdam and Philadelphia.
- Anke Lüdeling, Julia Ritz, Manfred Stede, and Amir Zeldes. 2016. [Corpus linguistics and information structure research](#). In Caroline Féry and Shinichiro Ishihara, editors, *The Oxford Handbook of Information Structure*, pages 599–617. Oxford University Press, Oxford, UK.
- Hideki Maki, Lizanne Kaiser, and Masao Ochi. 1999. [Embedded topicalization in English and Japanese](#). *Lingua*, 109:1–14.
- Samuel Elmo Martin. 2004. *A Reference Grammar of Japanese*, chapter 2. University of Hawai'i Press.
- Igor Mel'čuk. 2001. *Communicative organization in natural language*. John Benjamins Publishing Company.
- Jiří Mírovský, Kateřina Rysová, Magdaléna Rysová, and Eva Hajičová. 2013. [\(pre-\)annotation of topic-focus articulation in Prague Czech-English Dependency Treebank](#). In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 55–63, Nagoya, Japan. Asian Federation of Natural Language Processing.
- Ad Neeleman, Elena Titov, Hans Van de Koot, Reiko Vermeulen, et al. 2009. A syntactic typology of topic, focus and contrast. *Alternatives to cartography*, 1551(10.1515):9783110217124–15.
- Irina Nikolaeva. 2014. *A grammar of Tundra Nenets*. Walter de Gruyter GmbH & Co KG.
- David Y. Oshima. 2009. [On the So-Called Thematic Use of Wa: Reconsideration and Reconciliation](#). In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, pages 405–414, Hong Kong. City University of Hong Kong.
- Patrizia Paggio. 2006. Annotating information structure in a corpus of spoken Danish.
- Massimo Palermo. 2013. *Linguistica testuale dell'italiano*. Il mulino.
- Alessandro Panunzi. 2009. [Strutture scisse e pseudoscisse : valori d'uso del verbo essere e articolazione dell'informazione nell'italiano parlato](#). Franco Cesati Editore, Firenze :.
- Julia Ritz, Stefanie Dipper, and Michael Götze. 2008. Annotation of information structure: an evaluation across different types of texts. In *LREC*.
- Stéphane Robert. 1991. *Approche énonciative du système verbal: le cas du wolof*. CNRS Editions.
- Stéphane Robert. 2000. Le verbe wolof ou la grammaticalisation du focus. *Topicalisation et focalisation dans les langues africaines*, pages 229–267.
- Mats Rooth. 1992. A theory of focus interpretation. *Natural Language Semantics*, 1(1):75–116.
- Serge Sauvageot. 1965. *Description synchronique d'un dialecte wolof : le parler du Dyolof*. Ph.D. thesis, Paris.
- Sanghoun Song. 2017. [Modeling information structure in a cross-linguistic perspective](#). Language Science Press, Berlin.
- Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.
- Manfred Stede and Sara Mamprin. 2016. [Information structure in the Potsdam commentary corpus: Topics](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1718–1723, Portorož, Slovenia. European Language Resources Association (ELRA).
- Kensuke Takita. 2014. [Pseudo-right dislocation, the bare-topic construction, and hanging topic constructions](#). *Lingua*, 140:137–157.
- Takaaki Tanaka, Yusuke Miyao, Masayuki Asahara, Sumire Uematsu, Hiroshi Kanayama, Shinsuke Mori, and Yuji Matsumoto. 2016. [Universal Dependencies for Japanese](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1651–1658, Portorož, Slovenia. European Language Resources Association (ELRA).
- William H. Torrence. 2005. *On the distribution of complementizers in Wolof*. Ph.D. thesis, University of California, Los Angeles.

Henri Weil. 1879. *De l'ordre des mots dans les langues anciennes comparées aux langues modernes: question de grammaire général*, volume 18. F. Vieweg, Paris.

Malte Zimmermann and Caroline Féry. 2010. *Information structure: Theoretical, typological, and experimental perspectives*. Oxford University Press, USA.

Katalin É. Kiss. 1998. Identificational focus versus information focus. *Language*, 74(2):245–273.

Language Resource References

Bejček, Eduard and Hajičová, Eva and Hajič, Jan and Jínová, Pavlína and Kettnerová, Václava and Kolářová, Veronika and Mikulová, Marie and Mírovský, Jiří and Nedoluzhko, Anna and Panevová, Jarmila and Poláková, Lucie and Ševčíková, Magda and Štěpánek, Jan and Zikánová, Šárka. 2013. [Prague Dependency Treebank 3.0](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Michal Novák, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský, Anna Nedoluzhko, Kutay Acar, David Bamman, Peter Bourgonje, Silvie Cinková, Hanne Eckhoff, Gülşen Gebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, Pauline Krielke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, Petter Mæhlum, M. Antònia Martí, Marie Mikulová, Kirill Milintsevich, Vandan Mujadia, Judith Muzerelle, Sangha Nam, Anders Nøklestad, Maciej Ogrodniczuk, Lilja Øvrelid, Tuğba Pamay Arslan, Ian Porada, Marta Recasens, Per Erik Solberg, Manfred Stede, Milan Straka, Daniel Swanson, Svetlana Toldova, Noémi Vadász, Erik Velldal, Veronika Vincze, Amir Zeldes, and Voldemaras Žitkus. 2025. [Coreference in universal dependencies 1.3 \(CorefUD 1.3\)](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Introducing Universal Dependencies for Sardinian: the UD ContSar Treebank

Nicoletta Puddu¹, Manuela Sanguinetti², Luigi Talamo³

¹Dept. of Literature, Languages and Cultural Heritage, University of Cagliari (Italy)

²Dept. of Mathematics and Computer Science, University of Cagliari (Italy)

³Dept. of Language Science and Technology, Saarland University (Germany)

{nicoletta.puddu, manuela.sanguinetti}@unica.it

luigi.talamo@uni-saarland.de

Abstract

This paper introduces the first steps towards the creation of a novel resource for contemporary Sardinian within the Universal Dependencies framework. Sardinian is a Romance language spoken in Sardinia, an island belonging to the Italian Republic and located in the center of the western Mediterranean. It is a minority and endangered language, traditionally transmitted mainly orally, and characterized by a multiplicity of varieties (usually grouped into two macro-varieties Logudorese and Campidanese), all recognized as part of the Sardinian linguistic continuum. These varieties share basic morphosyntactic features, while presenting differences at the lexical level and in the realization of specific constructions. This internal variation can be particularly challenging with regard to the normalization of lemmas and the linguistic characterization of certain phenomena. The development of the treebank therefore aims to provide an annotated resource for contemporary Sardinian that takes into account the specificities of the different varieties, using Universal Dependencies to represent them within a unified theoretical framework, in order to facilitate both linguistic analysis and automatic processing. The present paper thus describes some linguistic characteristics of Sardinian and the attempts to encode them within the UD framework. Finally, we present the results of our evaluation of an NLP pipeline for Sardinian, trained on our corpus, for the Stanford Stanza parser.

Keywords: Sardinian, UD, Treebank

1. Introduction

The possibility of using a unified representation scheme such as Universal Dependencies, which has consolidated its role over the years as the standard for treebank annotation, has led to the proliferation of new resources of this type for both widely spoken languages and minority or under-resourced languages. Narrowing it down to Italy alone, it is well known that the country is rich in dialects and linguistic varieties in addition to Italian (see, for example, the overview proposed in Ramponi 2024). Some of these have also found their way into the UD project: this is particularly the case with Ligurian¹, Neapolitan², and Sicilian (Cappello et al., 2025)³. Along the lines of such efforts, in this paper we introduce an ongoing work on the development of a UD-based treebank for contemporary Sardinian.

Sardinian (ISO 639-3 *srd*) is a Romance language, spoken on the island of Sardinia, belonging to the Republic of Italy. It shows many conservative features, especially at the phonetic and lexical level, but it has also been considerably influenced by Iberian languages (Castilian and Catalan), given

that it was under the Catalan-Aragonese domination for more than four centuries. It is considered an endangered language, and is protected by National Law n.482/1999 and Regional Laws n.26/1997 and n.22/2018. There are no recent surveys on the actual number of speakers, but in a study conducted by the Autonomous Region of Sardinia on 2.437 inhabitants, 68.4% of the respondents claimed to be able to understand and speak the language and a further 29% only to understand but not speak it (Oppo, 2007).

Sardinian is traditionally subdivided into four varieties: Campidanese, Logudorese, Nuorese and Arborense (Virdis, 1988). Two further varieties, Sassarese and Gallurese, will not be addressed here since they are classified by scholars as non-Sardinian varieties, being closer to Corsican and Tuscan varieties. The sub-varieties of Sardinian are usually grouped into two macro-varieties: a center-southern variety, named Campidanese (ISO 639-3 *sro*), and a center-northern variety, named Logudorese (ISO 639-3 *src*). It is, however, important to note that this traditional division into varieties has been challenged, due to the presence of a transitional area in the central region. (Hajek and Goebel, 2021; Loporcaro and Putzu, 2024).

Moreover, in 2006, a standard variety, called *Limba Sarda Comuna* (LSC), meaning "Common Sardinian Language", was developed by the Autonomous Region of Sardinia for documents is-

¹https://github.com/UniversalDependencies/UD_Ligurian-GLT

²https://github.com/UniversalDependencies/UD_Neapolitan-RB

³https://github.com/UniversalDependencies/UD_Sicilian-STB

sued by the regional administration ([Regione Autonoma della Sardegna, 2006](#)). LSC is intended to be a “mediation variety”, taking into account the most widespread features of Sardinian, and also picking several features of the transition varieties spoken in the centre of the island. Finally, the regional administration, in cooperation with the University of Cagliari, also issued provisional orthographic guidelines for Sardinian as part of the experimental certification of language competence C1 ([Regione Autonoma della Sardegna, 2022](#); [Grosso et al., 2024](#)).

Scholars generally agree that the differences across varieties are mostly phonetic and lexical, while at the morphosyntactic level Sardinian shows a general uniformity ([Viridis, 1988](#)). Given this substantial unity, our working hypothesis is that it is possible to develop a single morphosyntactic annotation and a unified treebank covering all main Sardinian varieties. The availability of a resource with such characteristics can be particularly useful both for consulting and analyzing specific linguistic constructs of Sardinian across its varieties, but also for the development of NLP tools.

In this respect, although the development of the resource is currently in its infancy, the resource has already made it possible to (i) formalize annotation strategies for a number of Sardinian-specific constructions and (ii) carry out a preliminary training of the Stanza dependency parser. While the results are necessarily exploratory given the limited size of the dataset, they provide a first indication of the feasibility of building UD-based parsing models for Sardinian. This paper will therefore present some linguistic features of Sardinian that have been represented according to UD principles, together with the preliminary results obtained with the parser.

2. Related Work

Concerning Sardinian, early attempts to build text corpora can be found in [Fortunato and Ravani \(2015\)](#) with the ATLISOR textual archive⁴, an unannotated corpus of Medieval texts, and later on in [Puddu and Talamo \(2020\)](#) with the EModSar⁵, a corpus of Early Modern Sardinian texts that were also POS-tagged. Preliminary experiments on both POS and syntactic annotation, instead, were carried out in [Puddu and Stein \(2018\)](#) using the Arborator web interface ([Gerdes, 2013](#)) to syntactically annotate a corpus of Medieval Sardinian. However, to the best of our knowledge, no reference treebank exists for contemporary Sardinian. A spoken corpus of Sardinian emigrants has been introduced in [Pisano et al. \(2019\)](#) and [Mura et al. \(2023\)](#), with

⁴<http://atlisorweb.ovi.cnr.it/>
⁵<https://linguistica.dh.unica.it/emodsar>

POS annotation only. The same resource was later used to train a BERT-based POS tagger ([Carta et al., 2025](#)).

Further collections of spoken data were also created in [Chizzoni and Vietti \(2024\)](#) and [De Cristofaro et al. \(2025\)](#), where the recorded data served the purpose of developing and inspecting an ASR model for Campidanese Sardinian. Finally, some tools have been developed for rule-based Machine Translation from Italian to Sardinian ([Tyers et al., 2017](#)) and from Catalan to Sardinian ([Fronteddu et al., 2017](#)).

The present paper thus represents the first attempt to build an openly available and syntactically annotated resource of texts in contemporary Sardinian.

3. Data

The construction of corpora for minority languages often relies heavily on web sources, even when printed texts exist, essentially for copyright reasons. Due to its open license, Wikipedia often represents a valuable source of data for treebank development. In the case of Sardinian, this is particularly relevant, as Wikipedia pages explicitly indicate the variety in which each entry is written: Limba Sarda Comuna (LSC), Logudorese, Campidanese or Nuorese. This is particularly useful in the context of non-standardized languages, where identifying the corresponding sub-variety may not be trivial and may thus compromise a balanced diatopic representation. At the same time, relying on a single platform would risk limiting the range of registers and constructions represented in the corpus. Textual diversity is thus often required, in order to ensure a broader coverage of linguistic phenomena.

To balance the data scarcity problem with principles of varietal and genre variation, we preliminarily selected a sample to use for the pilot annotation, which features two Wikipedia pages (one in LSC⁶ and one in Campidanese⁷), and a fairy tale in Dorgalese, a sub-variety of Eastern and Southern Nuorese, titled *Mannoi Corrias e sa pudda bianca*, by Gonario Carta-Brocca⁸. The basic statistics on this preliminary sample are also reported in Table 1.

⁶Source: <https://sc.wikipedia.org/wiki/Arte>

⁷Source: <https://sc.wikipedia.org/wiki/Casteddu/campidanese>

⁸Source: Public URL no longer available at the time of writing.

Section	Variety	sent.	tok.
Mannoi	src	138	2312
Wiki-Casteddu	sro	39	1218
Wiki-Arte	LSC	17	488
	<i>tot.</i>	194	4.018

Table 1: Basic statistics of the pilot annotated sample. *src* and *sro* refer to the ISO code of the Logudorese and Campidanese Sardinian varieties, respectively, while LSC stands for *Limba Sarda Comuna*.

4. Annotation Principles

This section provides an overview of the main linguistic characteristics of Sardinian focusing on their representation according to the core principles of Universal Dependencies.

4.1. Tokenization

Tokenization was carried out following orthographic criteria that are consistent with standard UD practices. In general, words are delimited by whitespaces or apostrophes. Punctuation marks are treated as independent tokens; the only exception concerns apostrophes, which are always attached to the adjacent word. More specifically, in cases of elision, the apostrophe is attached to the end of the word containing the elided vowel (*unu ecantu_{src}*⁹ → *un'ecantu*, 'a piece'). In cases of apheresis, the apostrophe is attached to the beginning of the word undergoing the sound loss (*de* → 'e, 'of').

Multi-word tokens are introduced only in cases of contractions between verbs and clitic pronouns, where the orthographic word corresponds to multiple syntactic words (*daemilu_{src}* → *dae mi lu*, 'give to-me it').

4.2. Orthography and Lemmatization

The spelling and lemmatization choices adopted in this work are mainly based on a set of general rules, that we have followed to ensure consistency and uniformity across the different Sardinian varieties. Given the mediating nature of the LSC, these general rules primarily refer to the guidelines provided by the Autonomous Region of Sardinia, that consider the following underlying principles:

- Accent marks in Sardinian serve exclusively to indicate stress placement and do not distinguish vowel quality. Specifically, the use of accents is compulsory in oxytone and proparoxy-

tone words (e.g., *fàghere_{src}* 'to do', *èssere_{src}* 'to be', *merì_{sro}* 'evening'). Consider, however, that oxytones are not very common in Sardinian, since they often develop a paragogic vowel as in *tue_{src}* 'you' or *caffei_{sro}* 'coffee' (Virdis, 1978; Dettori, 2002).

- Given the diffused lenition of intervocalic singleton stops and fricatives, Sardinian does not display, for many consonants, a clear opposition between singleton and geminate consonants. Consequently, except for the consonants [b], [d], [l], [m], [n], [r], and [s], which can be either simple or geminate, all the other consonants in intervocalic position are written as simple.

We thus resorted to these rules to properly handle a wide range of spelling and accent inconsistencies at the lemma level, probably due, in many cases, to the scholarization in Italian of the writers. In particular, we found frequent omission of accents especially in proparoxytone words (accordingly, the noun *omine* 'man' has been lemmatized as *òmine*), where Italian orthographic rules do not prescribe it. We also found variation in consonant gemination, with double consonants occurring in contexts where Sardinian orthography does not allow them. Consequently, we have instances like *fèmminas* 'women', which has been lemmatized as *fèmina*). This tendency likely reflects not only an attempt to encode a more intense pronunciation, but, again, a possible interference of Italian orthographic conventions.

Another recurring case of normalization, always at the lemma level, concerned the so-called *eu-phonetic d*, that refers to the insertion of a final [d] in certain words (the most frequent being the prepositions *in* 'in' and *cun*, 'with'), when the following word begins with a vowel, to facilitate pronunciation. In line with LSC guidelines, these occurrences were thus reduced to their basic form (*totu ind'unu* → *totu in unu*, 'all in one').

The lemmatization process was supported by lexicographical tools and reference resources. In particular, the Dictionary of Sardinian Language and Culture by Mario Puddu (Puddu, 2015) (specifically, its online version)¹⁰ was used as a guide for selecting lemma forms, and the CROS spell checker - Corpus de Referèntzia de su Sardu,¹¹ also accessible online, was used to verify the consistency of the choices made.

It is worth pointing out that, whenever possible, a unified lemma form was adopted across different Sardinian varieties. This choice was motivated not only by the need for internal coherence, but also by methodological considerations

⁹As mere convention, we will use subscripts to specify, whenever required, the variety - expressed with its ISO code - to which the reported form belongs to. Word forms without subscripts are thus common to all the varieties considered.

¹⁰<https://ditzionariu.nor-web.eu/en>

¹¹<https://cros.nor-web.eu/>

related to corpus querying. Adopting a unified lemma makes it possible to retrieve and compare the same morphosyntactic phenomenon across varieties, avoiding the fragmentation of the data into variety-specific lemma entries.

4.3. Morphology and POS Tagging

Similarly to other Romance languages, Sardinian has a rich inflectional morphology, which also compounds with internal variations across its main varieties. While a full account of the inflectional morphology of contemporary Sardinian is well beyond the scope of this paper, we discuss here some relevant phenomena and their implications in terms of UD representation (for a brief sketch of Sardinian from a typological perspective see Putzu 2017; Loporcaro and Putzu 2024).

Sardinian marks gender and number in nominals. Adjectives generally agree with the noun in gender and number, including possessive adjectives (with the exception of the invariant third person plural *issoro*, ‘their/theirs’). Degree morphology is limited: comparative and superlative meanings are typically expressed analytically, with lexicalized exceptions such as *mèzus_{src}/mèllus_{sto}* (meaning both ‘better’ and ‘best’) and *peus_{src}/pejus_{sto}* (‘worse’ or ‘worst’). The suffix *-issimu* exists but is restricted to marked or poetic registers. As a result, the feature *Degree* is only marginally relevant when annotating texts in contemporary Sardinian.

Sardinian has both definite and indefinite articles, while it lacks partitive articles (Virdis, 2007). Definite articles mark gender and number in the singular (*su*, *sa*), but in the plural, while Logudorese has a masculine *sos* opposed to a feminine *sas*, Campidanese only has the unmarked form for gender *is*.

Quantifiers such as *totu* (‘all’), *dogni_{src}/cada_{sto}* (‘every’), and *carchi* (‘some’) are invariable and typically precede the noun, whereas *meda* (‘much, many’) is generally invariable but tends to occur postnominally (Mensching, 2017). Their morphosyntactic behavior informs their POS assignment and feature specification, but does not require language-specific morphological categories.

Concerning the pronominal system, Sardinian distinguishes between tonic (stressed) and atonic (clitic) pronouns (Loporcaro and Putzu, 2024). In particular, third-person clitic pronouns are morphologically differentiated for accusative and dative functions. This distinction is encoded in our annotation using the feature *Case=Acc* or *Case=Dat*.

Tonic pronouns have an additional property in the first and second person singular: special forms occur when the pronoun is preceded by the preposition *a* (e.g., *a mie_{src}/mimi_{sto}*, ‘to me’). This behavior, which is exceptional within Romance, is treated in UD by marking the pronoun with the

appropriate *Case* value (typically *Dat* or *Acc*, depending on syntactic function, see Section 4.4), while the preposition is annotated independently as *ADP*. A further distinctive feature concerns comitative constructions with *cun* (‘with’). Sardinian, in the first and second singular form, allows forms such as *cun megus/teguas* (‘with me/you’), where the pronominal form differs from the canonical nominative. This phenomenon is shared with languages such as Spanish and Portuguese. In UD terms, this is represented by assigning the comitative feature *Case=Com* to the pronoun, thus capturing the morphologically marked status of these forms while preserving the standard syntactic relation (*obl* with case marking).

The system of demonstratives is tripartite, distinguishing proximal, medial and distal pronouns and adverbs (Putzu, 2015).

Sardinian, as the other Romance languages, also has a dedicated reflexive pronoun for third person (tonic *se*, clitic *si*), which is used in reflexive, impersonal and passive constructions (Puddu, 2005).

As for the verb, Sardinian has three conjugation classes: *-àre_{sto}* *-ài_{src}* *-ere -i(ri)_{src}* *-ìre_{sto}-ìri_{src}*. One of the most notable features of Sardinian is the preservation of final *-s* and *-t* in verb endings. However, in spoken Sardinian, final consonants are often followed by a paragogic vowel, which some authors may mark also in the written languages.

Sardinian has lost the original Latin simple perfect and, as northern Romance varieties uses the Romance compound Latin perfect (Virdis, 2007).

As in other Romance varieties, the synthetic forms for future and conditional derive from periphrastic constructions. The future comes from Latin HABEO+AD+VERB.INF which results in the forms *apo a cantare_{sto}/apu a cantai_{sto}*, ‘I will sing’, lit. ‘I have to sing’.

The conditional shows a more complex distribution: in Campidanese it derives from HABEBAM (AD) + VERB.INF, so it is formed by the imperfect indicative of the verb *ai_{src}* ‘to have’ + *a* + VERB.INF resulting in *ia/emu a cantai* ‘I would sing’. lit. ‘I had to sing’. In Logudorese, the conditional *dia cantare* is formed by a now opaque past form of the verb *dèpere*, derived from Latin DEBEBAM, directly followed by the infinitive (Pisano, 2009).

It must be noticed that in Sardinian, as in Romanian but differently from other Romance languages, in both future and conditional, the auxiliary precedes the verb and it is clearly detached from it. In other words, while in Italian or Spanish the forms *canterò* and *cantarè* have inglobated the original auxiliary in the verb ending, in Sardinian and Rumanian forms, the auxiliary is still clearly recognizable. From the annotation perspective, this means that these are treated as analytic periphrastic construc-

tions; in these cases, tense-mood inflections are not directly encoded in the morphological features of the lexical verb. Instead, the auxiliary (e.g. *ai* ‘have’ or the past form of *dèpere* ‘have to’) bears the relevant features (e.g. `Mood=Ind, Tense=Pres` or `Imp, Person, Number`), while the infinitival verb simply includes the `VerbForm=Inf` feature.

The gerund is highly productive in Sardinian, and can be used both after a verb of perception and as a modifier of a noun phrase (Cuzzolin, 2005).

For all other morphological features, including the language-specific feature `Clitic=Yes`, we aligned our annotation choices with the conventions adopted in the UD Italian treebank, due to the structural proximity between Sardinian and Italian.

With respect to part-of-speech tagging, Sardinian aligns straightforwardly with the UPOS inventory and makes use of all 17 universal tags. The mapping between traditional grammatical categories and UD labels is largely unproblematic and reflects the profile of the language as a Romance system.

The tag `AUX` is assigned to functional verbs that contribute tense, aspect, voice, or modality without introducing an independent lexical predicate. This includes the copular verb *èssere* ‘to be’; the auxiliaries *èssere* and *àere* ‘to have’ in compound tenses and passive constructions; *dèpere* ‘must’ in the analytic formation of the conditional; *èssere* in progressive constructions; and modal verbs such as *pòdere* (‘can’), *dèpere* (‘must’), and *chèrrere_{src}/bòlli(ri)_{src}* (‘want’) when they function as modal auxiliaries rather than as full lexical verbs. When these verbs introduce their own argument structure, they are tagged as `VERB`, following a distributional criterion.

The tag `DET` is used for elements occupying the determiner position within the noun phrase, including articles, demonstratives, indefinites, and exclamationatives when they modify a nominal head. Two noteworthy cases concern the pronominal use of the form *su*, that is analogous to Spanish *lo*, and possessive adjectives. Regarding the former, when *su* functions as a nominal head, it is annotated as `PRON` with `PronType=Dem`, due to its deictic nature and syntactic autonomy. Possessive adjectives, instead, although semantically related to determiners, are obligatorily postnominal and have full adjectival agreement. Their distribution therefore aligns with the adjectival class rather than with determiners and are annotated as `ADJ`. To encode their possessive semantics, the feature `Poss=Yes` is added.

Finally, the tag `PART` is marginal in the current dataset. The only attested instance is *nanca*, a discourse particle coming from *narant+ca* (‘they say that’) conveying reported, not directly witnessed information (Cruschina and Remberger, 2008; Brunelli, 2013; Loporcaro and Pisano, 2021). In our treebank *nanca* occurs once in the example

su mere non mi cheret prus ca nanca no apo fattu bona guardia ‘the lord does not want me anymore since he says that I did not stand guard well’. Its annotation as `PART` reflects its clause-level pragmatic function and lack of argument structure, distinguishing it from adverbs or complementizers.

Overall, the POS tagging and morphological annotation of Sardinian demonstrate a high degree of compatibility with the UD framework, even in case of language-specific properties that require a suitable encoding.

4.4. Syntax

From a typological point of view, Sardinian, as most Romance varieties, is an SVO, pro-drop language, showing nominative-accusative alignment (Putzu, 2017). Determiners precede the noun, while all the modifiers, including possessives, follow the noun. Moreover, Sardinian, like other Romance languages of the southern area, shows differential object marking, which is driven by syntactic-semantic parameters of the object, as its animacy, specificity, and definiteness (see the example in Figure 1).

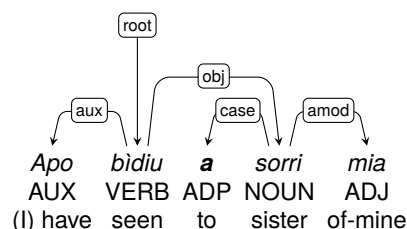


Figure 1: Example in Campidanese of direct object introduced by preposition. English translation: ‘I’ve seen my sister’.

It also features some specific properties that are particularly salient from a structural perspective. Among these are the case, already described in Section 4.3 of *su* as a pronominal head, as well as reduplication strategies used as intensifiers. These cases required explicit annotation choices, partly drawing inspiration from other UD language-specific guidelines.

Concerning the pronominal use of the form *su*, we already introduced in Section 4.3 the case where it does not modify a noun, but rather stands independently. Accordingly, it functions as syntactic head, either governing a subordinate clause, typically a relative, or a nominal modifier. The example in Figure 2 shows the former instance, with the finite clause - introduced by the relative pronoun *chi* - attached to *su* with the relation `acl:relcl`. Figure 3 instead shows the latter case, where the phrase *de Aragoni* ‘of Aragon’ is attached as nominal modifier of the pronoun according to standard UD practice.

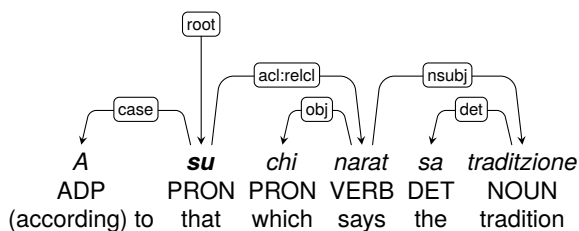


Figure 2: Example in LSC of pronominal *su* governing a relative clause. English translation: ‘according to the tradition’.

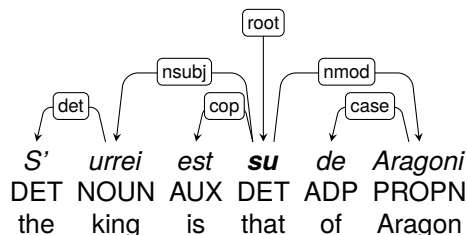


Figure 3: Example in Campidanese of pronominal *su* governing a noun. English translation: ‘the king is that of Aragon’

As in other Mediterranean languages, reduplication can be used in Sardinian for intensive purposes and is a productive and structurally regular strategy (Stolz, 2003). It generally involves adjectives (e.g. *mannu mannu* (lit. ‘great great’ → ‘very great’), adverbials (*abbellu abbellu*, ‘very slowly’), but also nouns, particularly in spatial constructions (*caminau muru muru_{STO}* ‘to walk along a wall’, referring to the pathway along which the movement is performed), and finally with verbal forms (*canta canta*, ‘singing continuously’) (Floričič, 2012). Putzu (2017) considers these cases as “reduplicative compounds”. Accordingly, the general annotation criterion we adopted in UD is to treat this phenomenon as a case of compounding rather than modification or coordination, also following the example of other UD treebanks where reduplication serves a similar function, such as in Greek¹² or Turkish¹³. However, contrary to the latter two, and consistently with the head-initial principle adopted for determining the syntactic governor in Sardinian, the head is always the first element of the compound.

5. Annotation Methodology

Following the preliminary selection of the texts, the annotation process began with an initial automatic

¹²<https://universaldependencies.org/el/dep/compound-redup.html>

¹³<https://universaldependencies.org/tr/dep/compound-redup.html>

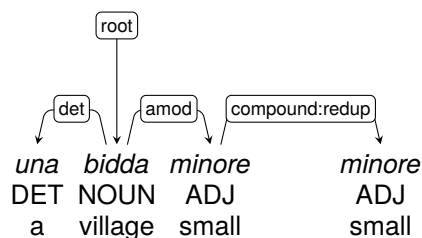


Figure 4: Example in Nuorese of reduplication. English translation: ‘a very small village’.

pass over the raw data. Since no parser specifically trained on Sardinian was available, it was not possible to rely on a language-specific model. Instead, we employed an existing model trained on the Italian ISDT treebank (Simi et al., 2014), motivated by the typological and genealogical proximity between Italian and Sardinian, as well as by partial overlap in morphosyntactic structures. In fact, this choice was intended as a bootstrapping strategy rather than as a fully reliable annotation solution. As expected, the quality of the parser’s output was extremely low, and errors affected all linguistic levels. Nonetheless, this preliminary step proved useful in providing a first segmentation of the texts and a basic tokenization that would align with UD principles, thus facilitating the subsequent revisions.

The core of the annotation work was therefore carried out manually, using Arborator-Grew as graphical interface (Guibon et al., 2020). Two native-speaker annotators were involved, one speaker of Campidanese and the other of Nuorese. In light of their respective linguistic competence and variety-specific expertise, the annotation tasks were distributed accordingly: the Campidanese annotator was primarily responsible for texts belonging to the Campidanese variety, while the Nuorese annotator worked on Logudorese texts, due to the greater structural and lexical affinity between Logudorese and Nuorese compared to Campidanese. The corpus section comprising texts in LSC was instead reviewed jointly by both annotators. However, due to the very limited number of LSC sentences included in the corpus so far, it was considered premature to compute inter-annotator agreement at this stage. Such evaluation is planned for the next phase of the treebank development, once the LSC section will be sufficiently expanded to provide more meaningful results.

For validation purposes, a third annotator was involved, in order to fix issues concerning in particular inconsistencies in feature or dependency annotations. Finally, after configuring the language-specific settings, the official validation script was used to make sure the annotated data was fully compliant with UD requirements.

Although the manual annotation was conducted independently by the annotators, the process was accompanied by frequent coordination meetings. These sessions served to discuss problematic cases, refine annotation decisions, and progressively define and revise the project's annotation guidelines.

6. NLP Pipeline Training and Evaluation

We used our treebank to train an NLP pipeline for Stanford Stanza (Peng et al., 2020). Stanza's NLP pipeline is made of the following models: sentence splitter and tokenizer, multi-word token extender, lemmatizer, parts-of-speech (POS) and morphological tagger (UFeats) and dependency parser. Training the last two models in Stanza requires word embeddings. We obtained word embeddings from the FastText project¹⁴ in the form of word vectors extracted from the Sardinian version of Wikipedia.

To train the parser, we applied a stratified k-fold method for data sampling. This approach ensured both a balanced distribution of the three varieties across the training and test sets and allowed us to verify - despite the limited data size - the robustness of our results. The treebank was partitioned to obtain a training set of 155 sentences and a test set of 39 sentences, with the varieties in each split distributed approximately as follows: 71% Logudorese, 20% Campidanese, and 9% LSC. We used 5-fold cross-validation as the training method. Stanza was trained from scratch for each step of the pipeline, from tokenization to parsing, using the hyperparameter configurations suggested on the related GitHub page. Table 2 reports the results obtained, using the metrics developed for the CoNLL 2018 Shared Task (Zeman et al., 2018), macro-averaged across all folds. Table 3 shows instead a detailed account of the averaged F1 score obtained on each dependency relation.

The results show performance that is generally consistent with expectations for a parser trained on a limited-size dataset, demonstrating good quality in the more superficial aspects of linguistic analysis and progressively poorer performance as structural complexity increases.

Overall, the tokenizer performed very well (F1 = 99.46), while we observe lower - though still acceptable - results with sentence splitter and the lemmatizer. The lower results with the former are mainly attributed to the nature of the texts included in the Logudorese section, where the presence of several parentheticals and direct speech make sentence splitting more challenging. As regards lemmatiza-

tion, despite the higher sparsity due to the presence of different varieties, performance results are reasonable, but also more susceptible to change depending on the fold used for training and evaluation.

As for the POS and morphological tagger, we did not develop a language-specific POS tagset (XPOS) for our corpus. The trained UPOS model performs well (F1 = 87.68), and so does the morphological tagger (F1 = 81.89), although the latter - similarly to lemmatization - shows a greater sensitivity to the fold used. When considering the correctness of both UPOS and UFeats, the performance of the model decreases significantly (F1 = 75.23), suggesting errors distributed across the various components of the annotation.

The dependency parser has a fair understanding of the basic Sardinian syntactic structure, identifying the head-modifier relation in about 70% of tokens (Unlabeled Attachment Score = 70.36). However, the model is able to assign the correct syntactic relation only for approximately the 59% of tokens (Labeled Attachment Score = 59.58). If we exclude functional syntactic relations such as auxiliaries, copulas, adpositional markers and determiners (see also Table 3), performances are even worse (Content Label Score = 47.42). The model is particularly poor at identifying relations between clauses (relative and adverbial clauses) and performs sub-optimally in identifying the clausal structure, thus in properly finding subject, object, and obliques relations.

The lowest performance is achieved when lemmas and morphological features are included in the evaluation of the syntactic parser. The Bi-Lexical dependency score is identical to the Content Label Score (F1 = 47.45), meaning that the syntactic parser has a poor understanding of non-functional, meaningful syntactic relations. Finally, although the morphological tagger is fairly robust, the Morphology-Aware Labeled Attachment Score is the lowest of all scores (F1 = 43.37). This confirms that the model struggles to capture structure, labels, and morphological information simultaneously, as would be expected under conditions of data scarcity. However, the relatively low standard deviations across all these metrics suggest that the model is quite robust with respect to the cross-validation splits.

The analysis by dependency label provides further insights. The most frequent and locally determinable relationships, such as `det` (90.45%), `aux` (80.38%), and `case` (80.12%), achieve high scores, confirming that the model effectively learns simple and regular syntactic patterns. Labels such as `cc` and `root` also maintain decent performance. In contrast, more complex or less frequent relations, such as `ccomp`, `obl`, `conj`, and `advcl`, show a

¹⁴<https://fasttext.cc/docs/en/crawl-vectors.html>

marked decline in performance (F1 often below 40%), showing the difficulties in modeling long-distance, coordination and subordinate structures. The extremely low or zero performance for the remaining labels is clearly due to the significant imbalance in the dataset and insufficient coverage during training, resulting from the low frequency of these relations.

	Avg. ($\pm$ st.dev)
Tokens	99.46 ($\pm$ 0.20)
Sentences	78.68 ($\pm$ 1.56)
Lemma	81.33 ($\pm$ 1.76)
UPOS	87.68 ($\pm$ 0.74)
Ufeats	81.89 ($\pm$ 1.77)
All tags	75.23 ($\pm$ 1.83)
UAS	70.36 ($\pm$ 1.64)
LAS	59.58 ($\pm$ 2.51)
CLAS	47.42 ($\pm$ 2.57)
MLAS	43.37 ($\pm$ 1.74)
BLEX	47.42 ($\pm$ 2.57)

Table 2: Average scores (with standard deviation) over the five folds.

7. Conclusions

In this paper, we introduced the preliminary stages of the development of ContSar, the first treebank for contemporary Sardinian. In particular, we outlined the main linguistic feature of Sardinian, highlighting some of its peculiarities, and describing how such features were mapped into the UD formalism. The annotation process has shown that UD is fully compatible with the structural properties of Sardinian, including its analytic constructions and pronominal system. This confirms the suitability of UD for representing the language and enables direct comparison with other treebanks built under the same framework.

As next step, we plan to expand the corpus with additional texts, following, as done in the previous stages, an iterative process of automatic bootstrapping with the latest trained model, manual revision and further parser improvement.

We intend to release the first batch of annotated data for the treebank to be openly available with the upcoming official UD release on May, 2026.

8. Limitations

The definition of the annotation guidelines followed a mixed approach. On the one hand, it was developed top down, taking into account the existing normative descriptions of Sardinian. On the other hand, it was refined bottom up, on the basis of the linguistic phenomena observed in the pilot sample

deprel	F1 ($\pm$ st.dev)
det	90.45 ($\pm$ 0.03)
aux	80.38 ($\pm$ 0.08)
case	80.12 ($\pm$ 0.05)
cc	74.43 ($\pm$ 0.06)
root	71.69 ($\pm$ 0.06)
flat	68.88 ($\pm$ 0.12)
expl	59.42 ($\pm$ 0.16)
amod	59.20 ($\pm$ 0.09)
nmod	56.42 ($\pm$ 0.04)
nummod	55.14 ($\pm$ 0.16)
nsubj	54.55 ($\pm$ 0.04)
punct	53.61 ($\pm$ 0.05)
cop	50.44 ($\pm$ 0.18)
advmod	48.17 ($\pm$ 0.10)
obj	47.30 ($\pm$ 0.07)
iobj	39.28 ($\pm$ 0.15)
mark	39.13 ($\pm$ 0.12)
ccomp	34.67 ($\pm$ 0.33)
obl	33.37 ($\pm$ 0.19)
conj	32.94 ($\pm$ 0.03)
discourse	31.25 ($\pm$ 0.24)
parataxis	29.59 ($\pm$ 0.23)
flat:name	26.15 ($\pm$ 0.28)
xcomp	23.22 ($\pm$ 0.19)
acl:relcl	16.48 ($\pm$ 0.11)
fixed	16.29 ($\pm$ 0.11)
acl	12.45 ($\pm$ 0.15)
advcl	9.37 ($\pm$ 0.07)
appos	8 ($\pm$ 0.11)
compund:redup	8 ($\pm$ 0.18)
compound	0
csubj	0
dislocated	0
nsubj:pass	0
obl:agent	0
orphan	0
vocative	0

Table 3: Average F1 score (with standard deviation) for each dependency label.

and the need to find a compatible UD-based counterpart. However, the pilot dataset is very small and does not ensure adequate coverage of the range of possible constructions. As a result, some annotation choices may need to be revised as new data are included.

The limited size of the dataset, although largely due to the scarcity of freely accessible sources, also affects the overall coverage and balance of the three main varieties included in the treebank. The current distribution does not yet guarantee a fully representative diatopic and structural coverage.

For these reasons, the extension of the dataset is not only a quantitative objective but also a methodological necessity. Increasing the size and diversity of the corpus will allow us to test and refine the

guidelines, improve balance across varieties, and strengthen the overall consistency of the resource.

9. Bibliographical References

- Michele Brunelli. 2013. [An evidential marker in Sardinian: nanca in Santa Maria Navarrese](#). In *Atti della XVIII Giornata di dialettologia. Quaderni di lavoro delASIt*, volume 16.
- S.M. Carta, F. Concas, G. Fenu, A. Giuliani, M.M. Manca, P. Mura, and S. Pisano. 2025. [A BERT-based Approach for Part-of-Speech Tagging in the Low-Resource Context of Sardinian](#). In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, Cagliari, Italy.
- Ilaria Chizzoni and Alessandro Vietti. 2024. [Towards an ASR system for documenting endangered languages: A preliminary study on Sardinian](#). In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 214–220, Pisa, Italy. CEUR Workshop Proceedings.
- Silvio Cruschina and Eva-Maria Remberger. 2008. Hearsay and reported speech. Evidentiality in Romance. *Rivista di Grammatica Generativa*, 33:95–116.
- Pierluigi Cuzzolin. 2005. Some remarks on the gerund in Sardinian. *Sprachtypologie und Universalienforschung*, 58(2/3):176–187.
- Domenico De Cristofaro, Alessandro Vietti, Marie-Pierre Pouplier, and Aleese Block. 2025. [When less is more? diagnosing ASR predictions in sardinian via layer-wise decoding](#). In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 363–370, Cagliari, Italy.
- Antonietta Dettori. 2002. La Sardegna. In Manlio Cortelazzo, editor, *I dialetti italiani. Storia struttura uso*, pages 897–958. UTET, Torino.
- Franck Floričič. 2012. On reduplicated imperatives in Sardinian. *Lingue e linguaggio*, 11(1):71–96.
- Gianfranco Fronteddu, Hèctor Alòs i Font, and Francis M. Tyers. 2017. [Machine translation from Catalan to Sardinian: a translation tool for a language in the process of standardisation](#). *Linguistica*, 9(2):3–20.
- Kim Gerdes. 2013. [Collaborative dependency annotation](#). In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, volume 88–97.
- Giulia Grosso, Giulia Murgia, and Antonietta Marra. 2024. [Certifying a minority language: a first description of the Sardinian case](#). *MOSAIC*, 15:105–121.
- Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. [When collaborative tree-bank curation meets graph grammars](#). In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5293–5302, Marseille, France. European Language Resources Association.
- Sandra Hajek and Hans Goebel. 2021. Le strutture profonde del dominio linguistico sardo: Un’analisi dialettometrica. In *Actes du XXIXe Congrès international de linguistique et de philologie romanes, Copenhague 1er–6 juillet 2019. Section 7. Dialettologia e geolinguistica medievale e moderna (Europa e fuori dall’Europa)*, volume 2, pages 979–991, Strasbourg. Société de linguistique romane.
- Michele Loporcaro and Simone Pisano. 2021. I complementatori chi e ca in Sardegna centrale: estensione areale e dinamiche di variazione. In *Actes du XXIXe Congrès international de linguistique et de philologie romanes, Copenhague 1er–6 juillet 2019. Section 7. Dialettologia e geolinguistica medievale e moderna (Europa e fuori dall’Europa)*, volume 2, pages 993–1005, Strasbourg. Société de linguistique romane.
- Michele Loporcaro and Ignazio Putzu. 2024. Sardinian. In *Oxford Encyclopedia of Romance Linguistics*, pages 1–42. Oxford University Press, Oxford.
- Guido Mensching. 2017. Morfosintassi: sincronia. In *Manuale di linguistica sarda*, pages 376–396. Mouton de Gruyter, Berlin.
- P. Mura, S. Carta, A. Giuliani, and M. Manca. 2023. The corpus of Sardinian emigrants: a tool for a quantitative approach to contact phenomena. In *Minority Languages in European Societies International Conference. Book of Abstracts*.
- Anna Oppo. 2007. [Le lingue dei sardi: una ricerca sociolinguistica](#). Rapporto finale, Regione Autonoma della Sardegna.
- Qi Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher Manning. 2020. [Stanza: A Python Natural Language Processing Toolkit for Many Human Languages](#). In *ACL2020 System Demonstration*.
- Simone Pisano. 2009. I futuro e il condizionale analitici in alcune varietà sarde moderne: genesi di marche grammaticali da forme verbali lessicalmente piene. *Bollettino di studi sardi*, 2:147–166.

- Mario Puddu. 2015. *Ditzionàriu de sa limba e de sa cultura sarda*, 2 edition. Condaghes, Cagliari.
- Nicoletta Puddu. 2005. Reflexives and intensifiers in Sardinian. *Sprachtypologie und Universalienforschung*, 3:246–261.
- Ignazio Putzu. 2015. [2. Sardinian](#). In Konstanze Jungbluth and Federica Da Milano, editors, *Manual of Deixis in Romance Languages*, pages 45–58. De Gruyter, Berlin, München, Boston.
- Ignazio Putzu. 2017. Tipologia del sardo. In Eduardo Blasco Ferrer, Peter Koch, and Daniela Marzo, editors, *Manuale di linguistica sarda*, pages 303–319. de Gruyter, Berlin.
- Alan Ramponi. 2024. [Language varieties of Italy: Technology challenges and opportunities](#). *Transactions of the Association for Computational Linguistics*, 12:19–38.
- RAS Regione Autonoma della Sardegna. 2006. [Limba sarda comuna. Norme linguistiche di riferimento a carattere sperimentale per la lingua scritta dell'Amministrazione regionale](#).
- RAS Regione Autonoma della Sardegna. 2022. [Certificazione provvisoria sperimentale della conoscenza delle lingue di minoranza storica parlate in Sardegna. Criteri ortografici orientativi per la lingua sarda](#).
- Thomas Stolz. 2003. A New Mediterraneanism: Word Iteration in an Areal Perspective. *Mediterranean Language Review*, 15(4):1–47.
- Francis M. Tyers, Hèctor Alòs i Font, Gianfranco Fronteddu, and Adrià Martin Mor. 2017. [Rule-based machine translation for the Italian-Sardinian language pair](#). *The Prague Bulletin of Mathematical Linguistics*, 108:221–232.
- Maurizio Viridis. 1978. *Fonetica del sardo campidanese*. Edizioni della Torre, Sassari.
- Maurizio Viridis. 1988. Sardisch: Areallinguistik. In Günter Holtus, Michael Metzeltin, and Christian Schmitt, editors, *Lexikon der Romanistischen Linguistik*, volume IV, pages 897–913. Niemeyer, Tübingen.
- Maurizio Viridis. 2007. Tipologia e collocazione del sardo tra le lingue romanze. In *Atti del convegno di Santa Cristina di Paulilatino, 6– 8 dicembre 2001*, pages 141–152, Cagliari. Condaghes.
- Daniel Zeman, Jan Hajič, M. Straka, D.Q. Nguyen, M. Potthast, E. Stepanov, G. Di Fabrizio, N. Promrit, and M. Popel. 2018. [CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies](#). In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21.

10. Language Resource References

- Caterina Maria Cappello, Sabrina D’Alì, Mario Guglielmetti, Elisa Di Nuovo, and Cristina Bosco. 2025. [Arbuli sunnu: A Sicilian-Italian parallel treebank](#). In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 155–168, Cagliari, Italy. CEUR Workshop Proceedings.
- Maria Fortunato and Sara Ravani. 2015. L’informatica al servizio della filologia e della linguistica sarda: il corpus ATLIOR (archivio testuale della linguasarda delle origini). *Bollettino di studi sardi*, 8:53–90.
- Simone Pisano, Valentina Piunno, and Vittorio Ganfi. 2019. Appunti per un corpus di sardo multimediale. In Daniela Marzo, Valentina Piunno, and Simone Pisano, editors, *Per una pianificazione del plurilinguismo in Sardegna*, pages 147–164. Condaghes, Cagliari.
- Nicoletta Puddu and Achim Stein. 2018. Word-level and higher level annotation of the Sardinian Medieval Corpus. In *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities*, Vienna. Gerastree Proceedings.
- Nicoletta Puddu and Luigi Talamo. 2020. [EMod-Sar: A Corpus of Early Modern Sardinian Texts](#). ISBN: 978-88-942535-4-2 Pages: 210–215 Publication Title: Atti del IX Convegno Annuale dell’Associazione per l’Informatica Umanistica e la Cultura Digitale (AIUCD). La svolta inevitabile: sfide e prospettive per l’Informatica Umanistica.
- Maria Simi, Cristina Bosco, and Simonetta Montemagni. 2014. Less is more? Towards a reduced inventory of categories for training a parser for the Italian Stanford Dependencies. In *Language Resources and Evaluation 2014*, pages 83–90. European Language Resources Association (ELRA).

Exploiting Parallel Aligned Treebanks

Maarten Janssen

Faculty of Mathematics and Physics
Charles University
janssen@ufal.mff.cuni.cz

Abstract

In this paper we discuss parallel aligned treebanks, mostly within the context of the way it is implemented in TEITOK: a framework that supports word-level aligned, as well as multi-level text alignment (text, paragraph, sentence). It also provides various ways to visualize aligned data, including a newly introduced parallel visualization for dependency trees, with mouse-over aligned highlighting. There is also a parallel search function under development, that allows queries and statistics that are both tree-capable and multi-level alignment capable, including word-level alignment.

Keywords: parallel corpora, corpus querying, treebanks

1. Introduction

Universal Dependencies (UD) is the most successful framework for cross-linguistic grammatical corpus annotation. It is so popular in large part because it provides a homogeneous way to annotate any language, making it hence possible to compare languages. By default, UD only provides comparable data, making it possible to run searches in two different languages, to get statistical information about the differences between those languages.

But by default, UD treebanks do not provide a direct comparison between languages where we can see how a specific phenomenon is expressed in one language and another. This issue was already addressed in 2017 by the introduction of parallel aligned treebanks. Parallel aligned treebanks make it possible to compare the same sentence in different languages and check for differences in the construction used in those languages.

There are various tools that provide different types of access to parallel treebanks to allow a direct comparison. But few that provide full access to aligned data, especially not with token-level alignments. In this paper, we present the implementation of parallel corpus access in TEITOK that is intended to provide full access to treebanks that have been aligned at various level, from text to token. Currently this approach is fully involved in the TEITOK source code where parallel visualization of aligned (treebank) data is concerned. Much of this visualization has a long history, but only recently made available, and the parallel visualization of token-level aligned dependency trees is completely new.

The final component, presented here but not yet released, is a search function that allows searching through, and doing frequency calculations on, parallel aligned dependency treebanks. What is outside of the scope of this paper is the alignment process itself - this paper only discusses how to utilize the alignment once it is encoded in the data. There are

many publications on (semi) automatic alignment - for a discussion in the context of TEITOK, see (Janssen et al., 2025).

2. State of the Art

2.1. Parallel UD Treebanks

The first attempt at creating parallel aligned treebanks was made at the TLT (Treebanks and Linguistic Theories) conference in Cairo, where, during a workshop, a set of 20 sentences was translated into various languages, discussed, and annotated with UD. This formed the basis for the Parallel Universal Dependencies project (PUD) (Zeman et al., 2017), released in UD2.1 with a set of 1000 sentences in 18 languages (currently, 22 languages have PUD alignments). Various other projects follow the same idea, such as TueCL (Akhundjanova et al., 2025) for Turkic languages.

In recent versions, parallel sentences in UD treebanks are explicitly marked with a tag *parallel_id* marking various sentences in treebanks as parallel versions of the same sentence. Table 1 reports on the parallel sentences in UD2.17, indicating the project the parallel sentences belong to, the number of languages in which (at least some of) those parallel sentences occur, and the number of parallel sentences involved. Notice that JaGSD is a special case of a parallel corpus between different variants of the same language.

Project	Languages	Sentences
Cairo	22	20
PUD	22	1000
Bible	4	593
Atis	2	5432
Set	2	1489
Hk	2	1004
JaGSD	1	8100

Table 1: Parallel Sentences in UD2.17 Treebanks

Having the same sentences in various languages, all analyzed within the same (UD) paradigm, allows for a more direct comparison between the constructions used in the languages involved by comparing the actual dependency trees involved. And the various sizes and number of languages involved listed in table 1 provide useful data for a number of different scenarios.

There are, however, two drawbacks in the use of parallel aligned UD treebanks for comparing languages. The first is their size: as gold standard data, the treebanks are by nature modest in size, which reduces their usefulness for language comparison. Especially the 20 sentences in Cairo are more intended to help set up a treebank for a new language by having similar data to compare to than for serious analysis.

And the second drawback is the fact that there is no interface to actually find and compare the trees. With the currently available tools, you have to find a sentence in some source language, either directly in the conllu files or using one of the existing tools (PML-TQ, TEITOK, Grew-Match, or INESS). Once you have obtained the sentence, obtain the *parallel_id* for it, and look for the same ID in the target language (for which there are currently no tools). And then draw a tree for both sentences to compare them. The only exception is Grew-Match, that for SUD (surface UD) treebanks in the PUD collection provides the option to view the aligned sentences in any of the other SUD/PUD corpora.

The first of these two drawbacks is overcome by using (automatically annotated) parallel UD corpora instead of gold standard treebanks, as discussed in the next section. And as will be demonstrated there, the second point is also addressed in the GUI interface for some of the existing parallel UD corpora.

2.2. Parallel UD Corpora

When it comes to parallel corpora, two projects stand out: OPUS (Tiedemann, 2009) and InterCorp (Rosen et al., 2023).

OPUS is a resource infrastructure enabling access to parallel corpora. It is the largest available source of parallel data, and has an impressive 100B sentences in over 1000 different languages. All data are first sentence-aligned, and then word-level aligned using eflomal (Östling and Tiedemann, 2016). Where possible, the data are tagged and parsed as well, meaning that in the corpus query system that OPUS provides, we can search for data using the Corpus WorkBench (Evert and Hardie, 2011) to search for example sentences by POS tag or dependency relation. However, at present OPUS is not annotated in UD throughout, meaning that in order to get parallel UD dependency trees, it is necessary to first run the source and target

sentence through a UD parser before being able to compare the trees.

InterCorp is a multilingual parallel corpus. Version 16UD contains 4800M words spread over 62 languages, of which 47 languages are annotated with UD, using UDPIPE¹.

InterCorp is searchable in Kontext² (Machálek, 2020), which makes it possible to search through the aligned texts in InterCorp. When executing a query in one of the languages, you can select an aligned language, and all results will be displayed alongside their translation in that language. In addition, it is possible to add search restrictions on target sentences.

For the source language, the full dependency tree can be visualized, and when clicking on a word in the search result, the corresponding word in the translation will be highlighted. This is not done using pre-recorded word-alignment, since Kontext only allows alignment on a single level, but rather is calculated on-the-fly. It does this using an alignment service developed by LINDAT as part of their translation service (Poláková et al., 2025).

So the Kontext interface for InterCorp makes using parallel UD corpora a lot more user friendly: we can use CQL to find examples, we can automatically get the aligned sentences, and we can put CQL restrictions on those aligned sentences. Furthermore, we can directly see the tree for the source or target sentence and can see which words in the source and target language correspond.

But there are still various things that are not that easy: since Kontext uses CQL, which is not a tree query language, we cannot search for dependencies. The word-level alignment is calculated, but cannot be used in the display of the trees, and cannot be used in searches in any way.

In conclusion, there are many tools available that can partially exploit parallel UD treebanks. However, there is currently no method for word-level alignment in treebanks, there is also no visualization for parallel treebanks. And search options are currently either not supporting parallel corpora (like PMT-TQ or Grew-Match) or do not support tree queries (like Kontext).

3. TEITOK

TEITOK (Janssen, 2016) is an online corpus platform. It provides increasing amounts of support for UD (Janssen, 2018). It also provides support for minority languages and for parallel aligned corpora.

The design for working with word-level parallel aligned corpora in TEITOK relies on three

¹<https://lindat.mff.cuni.cz/services/udpipe/>

²<https://www.korpus.cz/kontext/>

things: the representation and creation of the parallel aligned corpora themselves, the visualization of the parallel sentences and dependency trees, and search options to search through the parallel aligned dependency trees. The first two parts are fully functional, with some existing parts and some new additions. The search functionality is still under construction as part of a larger overhaul of the search options in TEITOK.

3.1. Parallel Corpora

TEITOK is not inherently a UD tool, but rather a general-purpose corpus tools that provides an increasing amount of support for UD. It does not work natively with the CoNLL-U format, but rather a corpus consists of a collection of annotated TEI/XML documents. For the representation of the core UD token attributes, this is merely a difference in representation: where CoNLL-U stores token attributes as columns on a line representing the token, in TEITOK they are stored as attributes on an XML node representing the token. There are various scripts to seamlessly convert between the two functionally identical formats.

The use of TEI/XML makes for a more standardized and flexible representation of various other attributes. For sentence, paragraph, and text level attributes, there are well established standards, while for the comment lines performing this job in CoNLL-U there is currently no official standard. And for the representation of the information stored in the MISC column in CoNLL-U, there is in various cases a better established alternative in TEI/XML, for instance named entities are stored as nodes around tokens, not as information spread across various tokens in one of the formats UD supports for named entities, for instance the IOB format.

For the representation of alignment between nodes at any level, TEITOK uses basically the same strategy as applied in UD: There is an attribute that specifies an ID (the *parallel_id* in CoNLL-U, and *@tuid* in TEITOK, for *translation unit ID*), and when two nodes share the same ID, they are aligned. So, converting parallel treebanks of UD to the TEITOK format is trivial, by converting the *parallel_id* to a *@tuid* attribute on the relevant sentence node.

There are, however, two differences. Firstly, TEITOK can represent alignment, by a shared *@tuid* at various different levels at the same time - including larger units like paragraphs, or smaller units like tokens, with the same interpretation as those at sentence level: two tokens that share a *@tuid* are translations of each other. UD allows for the use of *Ref* in the MISC column, but *Ref* is just intended to handle partial correspondences between sentences, not alignment at token level. That is, however, merely a matter of practice and not a limitation of the CoNLL-U format: when ex-

porting to CoNLL-U, TEITOK exports token-level alignments as *parallelId* in the MISC column with the same syntax as that used in the *parallel_id* at the sentence level.

So token-level alignment is handled in exactly the same way as any other level of alignment, and a document can have alignment at multiple levels, in fact it is uncommon to have word-level alignment without having sentence-level alignment as well, and typically sentence-level alignment in turn relies on paragraph-level alignment.

The second difference is that, for partial correspondences, *@tuid* attributes are not single IDs, but sets of IDs. The simplest case is when a single Spanish sentence is translated as two (or more) sentences in English (which is quite common due to different habits), the various English sentences each have a unique *@tuid*, while the Spanish sentence has *@tuid*, which is a list of all those IDs. Although this can become very complicated in theory, we have found that in practice the system is quite workable. A discussion of a parallel corpus made with this representation at the sentence level can be found in (Janssen et al., 2025), where we show that even in cases where we align various versions of an ancient Greek text in copies or translations produced several centuries later, assigning *@tuid* values is meticulous work, but not due to restrictions of the representation method, but purely due to the analysis of the texts themselves. That paper also describes a set of scripts that produce an alignment in TEITOK using a variety of different existing automatic alignment techniques.

The TEITOK tools repository³ also provides a number of scripts that can convert from a number of existing aligned formats, most notably the TMX (Translation Memory eXchange) to a TEITOK corpus. This makes it possible either to take already aligned data from for instance OPUS or InterCorp, or to use the tools developed by those projects to align new data and then convert them to the TEITOK format.

3.2. Visualization

TEITOK provides various ways to visualize parallel texts, depending on the properties of the documents to be visualized. The most basic visualization is a table view in which each row shows an aligned pair, in much the same way as translation memory visualization works, or the search results for parallel corpora in InterCorp; if there are more aligned texts, multiple versions can be aligned in columns.

For straightforward cases of alignment, like the Cairo sentences in the UD treebanks, this visualization works very well. However, the more flexible set-

³<https://github.com/ufal/teitok-tools>

up in TEITOK creates a number of different problems. Given that a TEITOK document can have alignments on multiple levels, we have to indicate the level at which we want to see the alignment, typically choosing between sentences or paragraphs. This will create a table with all elements of that level in the left-most column (the source document). Since alignment does not have to be perfect, this has a number of consequences. Firstly, a single row in the source document can map to more elements or less elements in the other columns. So, if we align Spanish-English with the English text having shorter sentences, a single row will contain one Spanish sentence, but multiple English sentences. In the opposite direction, multiple English rows will show a single Spanish sentence along with it. Secondly, the table will only display the plain text of each sentence, leaving out the typesetting the TEITOK document might contain, since the typesetting might overlap with the table structure. Third, it will show the aligned elements in the order in which they appear. Especially in historical documents, the order in the copy or translation can vary, so the remaining columns can show the text shifted from its original order. And lastly, it will only show the selected elements in the source document and their translations. This not only means that if the target documents contain text that is not in the source document, they will not be shown. But even text outside the selected level will not be shown - if we select paragraphs, then any text that is not in a paragraph will not be displayed.

For documents in which the restrictions listed above make the visualization as a table not work, there is also the option to display entire documents next to each other, including any XML annotation they contain. This will not directly display any type of alignment, but moving the mouse over the text will highlight the first parent of the node the mouse hovers over that has a `@tuid`, and highlight that same element in all the other columns. In case there is no matching element for the first parent, but there are ancestor nodes that do have an alignment, that aligned element will be highlighted in a different colour.

Neither of these visualizations display or rely on dependency trees. Therefore, we added an additional visualization that displays parallel dependency trees. This visualization can be seen in figure 1, which shows a sentence from the NTREX corpus (Federmann et al., 2022), an English sentence on top, and the Czech translation on the bottom. Moving the mouse over a word in the tree will not only highlight that word in the full sentence, with all UD attributes of that node shown below it, as the normal dependency tree view does in TEITOK, but will also highlight the corresponding word in both the translated sentence and the de-

pendency tree for the translation. In this case, it shows the word *Trump* in English, with all token-level attributes, and shows the position of the word in the tree. And at the same time, it highlights the word *Trumpovi* (Czech uses inflection on proper names) with its position both in the sentence and the dependency tree.

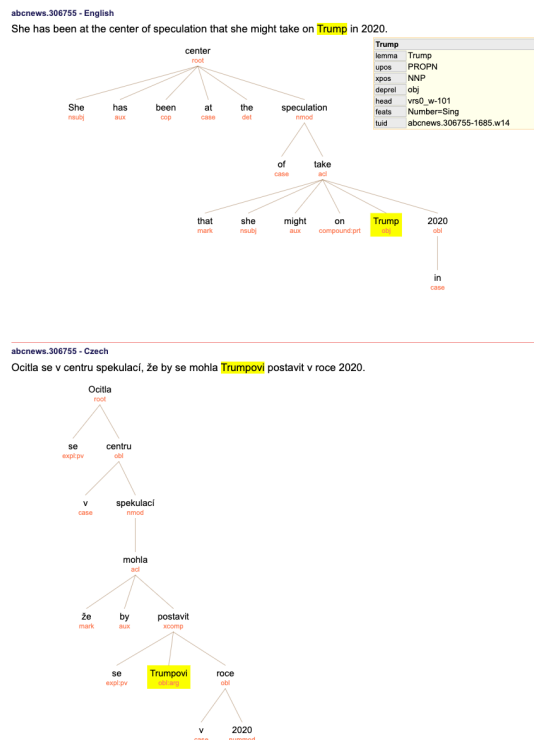


Figure 1: The TEITOK alignment tree view

This aligned dependency tree visualization makes it much easier to compare constructions between languages in actual use cases, even for a language that is not completely familiar, since the UD annotation will display all attributes and relations of the sentence in question.

3.3. Searching

Of course, in order to really exploit word-aligned parallel UD corpora, one would also want to be able to search through them and perform statistics on them. Currently, there is no platform that allows that. The main reason is that a quick survey suggests that there are no tree search algorithms at the moment that are fast enough to work with larger corpora. This is why the only platform listed on the UD website for searching that allows searching through all languages in UD2.17 at once, and does so because it only allows CQL searches. In order to allow searching through larger aligned corpora that allow at the same time tree-sensitive searches and alignment-sensitive searches, at the very least, we need a fast enough tree-sensitive search algorithm.

We are currently finalizing a new corpus search implementation based on translating a language-driven search query that is an extension to existing corpus query languages. And this query system adds the option to perform queries on aligned corpora. Not only in the way that, for instance, Kontext provides, where we can formulate a query, and then filter the results on sentences where the aligned sentence matches a different query, but also by a more expressive syntax that makes full use of the TEITOK-style alignment using @tuid attributes, by restricting searches to cases where the source and the target appear in a sentence or a paragraph with the same (or a shared) tuid value.

It can also make use of word-level alignment. This can be a simple use, by checking what is the most common translation of a given word. Or it can exploit dependency relations: we can, for instance, search for a noun used as *nmod* in Spanish, and then for all words in the English text that are translations of those words, in order to see what is their most frequent dependency relation, or what is the most frequent lemma of the parent node in question.

4. Conclusion

In this paper, we have shown a new approach in TEITOK to building and exploiting word-level parallel aligned UD corpora. This involves creating the alignments, which in TEITOK is done either by alignment scripts or by importing already aligned data. It also involves storing the alignment in the corpus, which in TEITOK is done by storing alignment IDs on any level, be it a text, a paragraph, a sentence, or a word.

It also involves visualizing the alignment, which in TEITOK can be done in a number of different ways: as a table of aligned elements; as full documents displayed alongside each other, where the highlighting is done on mouse-over. And a new visualization in which multiple dependency trees can be displayed next to each other, where on mouse-over the alignments are shown directly in the aligned trees.

And finally, to really exploit parallel UD data, a search system with statistical data would be needed. This search system is still under construction and is intended to allow users to search through aligned treebanks by making use of alignments either at the level of sentences or at the level of words.

4.1. Limitations

Even though much of the system for exploiting parallel UD corpora in TEITOK is in place, there are various limitations. The most obvious limitation

is that the search functionality that allows searching through parallel dependency trees is not yet released. For many cases, the new search algorithm is considerably faster than alternatives we have tested, it will have to be shown in practice how scalable the approach really is, whether there is a recommendable maximum size above which search queries become too slow.

Even though the encoding of word-level alignment has been implemented in various corpora by now, whether adaptations will be needed for more complex cases will have to be seen in practice when the system is used in a wider range of corpora. And what the quality of the automatic alignment scripts is in practical cases is something that will have to be explored in the future.

Arofat Akhundjanova, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, and Cagri Coltekin. 2025. [Parallel Universal Dependencies treebanks for Turkic languages](#). In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 129–136, Ljubljana, Slovenia. Association for Computational Linguistics.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Corpus Linguistics 2011*.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. [NTREX-128 – news test references for MT evaluation of 128 languages](#). In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Maarten Janssen. 2016. TEITOK: Text-faithful annotated corpora. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 4037–4043.

Maarten Janssen. 2018. TEITOK as a tool for dependency grammar. *Procesamiento del Lenguaje Natural*, 61:185–188.

Maarten Janssen, Piroska Lendvai, and Anna Jouravel. 2025. [Alignment of historical manuscript transcriptions and translations](#). In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 462–470, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Tomáš Machálek. 2020. [KonText: Advanced and flexible corpus query interface](#). In *Proceedings*

of the 12th Language Resources and Evaluation Conference, pages 7003–7008, Marseille, France. European Language Resources Association.

Robert Östling and Jörg Tiedemann. 2016. [Efficient word alignment with Markov Chain Monte Carlo](#). *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Lucie Poláková, Martin Popel, Věra Kloudová, Michal Novák, Mariia Anisimova, and Jiří Balhar. 2025. Mitigating language barriers in education: Developing multilingual digital learning materials with machine translation. In *EDULEARN25 Proceedings*, pages 8754–8760, Valencia, Spain. IATED.

Alexandr Rosen, Michal Křen, Pavel Šmerk, and Michaela Svobodová. 2023. [InterCorp: Multilingual parallel corpus with a uniform annotation](#). *Jazykovedný časopis*, 74(2):365–387. In English with Slovak abstract.

Jörg Tiedemann. 2009. [News from OPUS - a collection of multilingual parallel corpora with tools and interfaces](#). In *Recent Advances in Natural Language Processing*, volume V, pages 237–248.

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, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 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.

Cross-Dialectal Transfer for Low-Resource Arabic: The Tunisian Arabic Dependency Treebank

Amal Aissaoui

CUNY Graduate Center
365 5th Ave, New York, NY 10016
aaissaoui@gradcenter.cuny.edu

Abstract

This paper presents a small-scale dependency treebank for Tunisian Arabic (TADT) developed within the Universal Dependencies framework, addressing the scarcity of linguistic resources for the Arabic varieties. The approach employs domain adaptation, leveraging a machine learning model (UDPipe 1.0) trained on Algerian Arabic data to annotate 100 Tunisian Arabic social media comments, followed by manual correction. This pilot study evaluates the feasibility of using machine learning-assisted annotation to scale resource development for spoken Arabic and identifies key challenges in cross-dialectal transfer for improving annotation quality and efficiency. This work contributes to more inclusive and fair representation of Arabic linguistic varieties in academic research and NLP applications.

Keywords: Dependency Treebank, Tunisian Arabic, Domain Adaptation

1. Introduction

Given the central role that linguistic data plays in academia and industry, data scarcity is the biggest obstacle that stands in the way of a fair and comprehensive representation of languages in linguistic research and language-driven applications. For a linguistic community to increase its presence and weight in language technology, linguistic data must be “accessible and processable” (Lynn, 2016, 1).

The first contributing factor to data scarcity is the absence of a standard orthography system. Without a standard orthography, a language cannot be easily documented, which limits the amount of textual data that can be generated and stored for that language. The second factor is the absence of government support. Since underresourced languages tend to be spoken in countries with a colonial history or within diglossic speech communities, they are often marginalized in educational and governmental institutions.

In Tunisia, the aforementioned factors played a major role in creating a severe data shortage problem. Tunisia exhibits the features of a true diglossic community where more than one language co-exist within a single speech community, with each language being used for particular purposes and in different contexts. Considering Tunisia’s smaller population compared to its neighboring states, its geopolitical and cultural ties to the member states of the Arab League, and its French colonial history, it is expected that the country faces social and economic pressure to adopt more widely spoken languages for better economic opportunities, further reducing the opportunities to document and record

diverse and extensive data of Tunisian Arabic and keeping any standard writing system for the language from emerging.

In NLP, MSA is too formal to be used for NLP technologies used for daily tasks. For natural interactions between humans and voice assistants, spoken Arabic is preferable. Attempting to circumvent this issue by applying NLP tools trained on MSA to spoken Arabic will result in substantially poor performance (Al-Shargi and Rambow, 2015; Samih, 2017).

For spoken Arabic, even if the aspirations of policy makers to institute language reforms are present, they might take decades to implement, whereas the pressing urgency to advance in language technology calls for rapid actions with prompt results. This study adopts a proactive approach to bridging the existing gap by making the following contributions:

- Highlighting and assessing trends in resource development for the Tunisian Arabic variety.
- Presenting a small-scale Tunisian Arabic treebank developed within the Universal Dependencies framework.
- Evaluating the use of dialect-transfer as a strategy to accelerate the resource development process.

2. Related Work

2.1. Resource Development

Boujelbane et al. 2013 adopted a lexical transfer approach using the Penn Arabic Treebank

(Maamouri et al., 2005) and the MADA morphological analyzer (Habash et al., 2009) to build a bilingual lexicon of Tunisian Arabic and MSA. Zribi et al. 2013 created a bilingual dictionary consisting of roughly 30k words transcribed from radio and TV broadcasts in Tunisian Arabic and used it to propose a method to adapt an Arabic morphological analyzer designed for MSA to Tunisian Arabic.

Sadat et al. 2014 introduced a bilingual lexicon of 1.6k Tunisian Arabic words and their aligned MSA translation, and proposed a rule-based translation system to translate any social media text in TA into MSA. Bouamor et al. 2014 built a corpus that covers five varieties, namely Egyptian, Tunisian, Syrian, Jordanian, and Palestinian Arabic. The corpus includes 2k Egyptian sentences selected from the Egyptian-English corpus created by Zbib et al. 2012 then translated by native speakers into the four other varieties. The translators were instructed to avoid transliterating using the Roman script yet no orthographic guidelines for the Arabic spelling were provided. Meftouh et al. 2015 started by recording natural conversations and used the recordings of TV shows in Algerian Arabic which were translated into MSA then converted to Tunisian, Syrian, and Palestinian Arabic by native speakers of the respective languages. The parallel corpus of 32k sentences was then used to train machine translation models.

Medhaffar et al. 2017 presented TSAC, a Tunisian sentiment analysis corpus consisting of 17k manually annotated Facebook comments. Annotators were native speakers of Tunisian Arabic who labeled comments as positive or negative, accounting for code-switching and Arabic. Fourati et al. 2020 introduced TUNIZI¹, a Tunisian Arabizi sentiment analysis data set consisting of 9k manually annotated sentences of Arabizi from social media. Gugliotta and Dinarelli 2022 presented TARc², a multi-layer annotated corpus for user-generated Tunisian Arabic, containing approximately 45k annotated tokens. The corpus features multi-level linguistic annotation including language identification, transliteration, tokenization, POS tagging, and lemmatization, performed using a semi-automatic pipeline.

2.2. Trends and Limitations

While research on resource development for Tunisian Arabic is gradually emerging, three main trends can be identified. The first is the predominance of labeled resources lacking linguistic annotation, with Gugliotta and Dinarelli 2022 constituting, to date, the only resource annotated with lin-

guistic information. Moreover, morphological and syntactic annotation have received no attention in the existing literature.

The second trend is the absence of shared guidelines. Once annotation projects are started, it is crucial to adopt common guidelines and best-practice recommendations for the development of language resources to ensure “interoperability” (Ide et al., 2017, 116). An abundance of resources that can only be used in isolation is as ineffective as having a shortage of resources.

The third trend is the reliance on native speakers to generate annotations. Native Arabic speakers possess language proficiency, but are not sufficiently qualified to assist in this mission since they lack linguistic expertise. As discussed in Section 1, under-resourced languages tend to lack a standardized writing system due to historical marginalization in educational and governmental contexts. While native speakers possess invaluable inherent intuition, the absence of formal instruction in their primary language often means they have not acquired the specific academic frameworks necessary for independent linguistic analysis. Consequently, while their contributions are essential, native speakers lack the technical training required to generate formalized linguistic judgments that can be fully validated without the methodological oversight of a trained linguist for annotation tasks, including part of speech tagging, morpho-syntactic analysis, semantic role labeling, co-reference resolution, or transcription for automatic speech recognition.

Given the small size the linguistic community and the limited metalinguistic awareness among native speakers, finding a team of expert annotators can be challenging. As a result, developing a gold-standard linguistic resource becomes an incredibly time-consuming process. This study follows a linguistically-informed approach that leverages machine learning, user-generated data, and an existing linguistic resource from a neighboring Arabic variety to reduce the burden of labor-intensive manual annotation and streamline the development process to expedite the creation of linguistic resources. It further seeks to orient the efforts of the Tunisian research community toward interoperability and community standards by promoting the Universal Dependencies framework. Building on this approach, the study presents the first attempt to develop a treebank for Tunisian Arabic in the CoNLL-U format and in accordance with the Universal Dependencies scheme.

¹<https://github.com/iCompass-ai/TUNIZI?tab=readme-ov-file>

²<https://github.com/eligugliotta/tarc>

3. Resource Development and Description

3.1. Training Data

For the training of the machine learning model, the Algerian Arabic NArabizi³ dataset in CoNLL-U format compiled by Seddah et al. (2020) was used as training input. The treebank consists of 1,500 fully annotated user-generated sentences in Algerian Arabic, that include morpho-syntactic annotations, along with translations at both the word and sentence levels. It is made freely available under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license.

The selection of an Algerian Arabic treebank was motivated by two considerations. First, NArabizi is the only available UD treebank for a spoken Arabic variety and is provided in the user-generated transliteration of Arabic commonly referred to in the literature as the Arabizi script (Seddah et al., 2020), which is the same script in which Tunisian text data is available. Second, Algerian and Tunisian Arabic share a common linguistic foundation. Sociolinguistically, both varieties exist within a diglossic context, display frequent code-switching with French, and lack a standardized orthographic system, as discussed in Section 1. Lexically, they exhibit substantial overlap in their vocabularies, which derive primarily from shared Classical Arabic roots alongside borrowings from Berber, Italian, Turkish, and French. Structurally, both varieties show vowel reduction and consonant clustering, maintain closely parallel verbal conjugation systems, employ comparable strategies for pronoun attachment and negation morphology, and favor the SVO word order. Collectively, these factors contribute to the high degree of mutual intelligibility between speakers of the two varieties. For a detailed cross-varietal comparison involving Tunisian, Algerian, and other North African Arabic varieties, see Saadane and Habash 2015 and Harrat et al. 2015.

While the Algerian Treebank serves as a seminal resource for spoken Arabic as it represents the inaugural effort in this domain and provides the foundation for the current study, it presents certain methodological limitations. Specifically, inconsistencies originating from the NArabizi project framework affect the overall annotation reliability.

In Seddah et al., native speakers of different varieties of North African Arabic were trained to perform annotation tasks following the Universal Dependency guidelines. While the nature and length of the training was not specified by Seddah et al.,

a closer look at the segmentation of tokens casts doubts on the qualifications of the annotators. Native speakers of spoken Arabic have not received a formal education in their varieties or an academic linguistic training that equips them with the analytical skills to critically engage with Arabic varieties whose grammar was shaped by centuries of linguistic evolution and which remain underexplored by linguistic scholars. Without expert guidance, native speakers of Arabic are not sufficiently metalinguistically aware to conduct sound morphological analyses, which are essential to produce the systematic splitting of MWTs prescribed by the UD annotation guidelines. Table 1 illustrates MWT segmentation for a sentence from the NArabizi treebank, while Table 2 presents the corresponding expert segmentation performed by a linguist for the same instance.

ID	Form	Lemma	UPOS
1	allah	Dieu	PROPN
2	yahafdak	te_protège	VERB
3	w	et	CCONJ
4	ykhalika	te_laisse	VERB
5	lina	à_nous	PRON
6	ya	oh	INTJ
7	Mr	monsieur	NOUN
8	le	le	DET
9	président	président	NOUN
10	w	et	CCONJ
11	tahia	vivre	VERB
12	bouteflika	Bouteflika	PROPN
13	w	et	CCONJ
14	rana	nous_sommes	VERB
15	m3ak	avec_toi	PRON

Table 1: MWT segmentation and POS tagging in NArabizi.

Another issue with the training data is adopting French glosses instead of lemmas. Seddah et al. justified this choice on the grounds of the prevalence of non-Arabic lexical borrowings in Algerian Arabic that would have required an etymological analysis. This argument is valid only if lemma annotation equals extracting the consonantal root of a word in Arabic. Only under such assumption would etymological analysis of words in spoken Arabic be necessary. Since lemmatization is the process of stripping a word form from inflection such as number, tense, and case, tracing etymological origins is not necessary for the assignment of lemmas since lexical borrowings typically conform to the morphological and syntactic rules of the host language. The example *garjouma* used by Seddah et al. to support their reasoning may be originating from French *gorge* (throat) but its current use adheres to the Arabic broken plural pattern system that is still present in North African

³https://github.com/UniversalDependencies/UD_Maghrebi_Arabic_French-Arabizi/tree/master

ID	Form	Lemma	UPOS
1	allah	Dieu	PROPN
2-3	yahafdak	—	—
2	yahafd	protège	VERB
3	ak	te	PRON
4	w	et	CCONJ
5-6	ykhalika	—	—
5	ykhal	laisse	VERB
6	ka	te	PRON
7-8	lina	—	—
7	li	à	ADP
8	na	nous	PRON
9	ya	oh	INTJ
10	Mr	monsieur	NOUN
11	le	le	DET
12	président	président	NOUN
13	w	et	CCONJ
14	tahia	vivre	VERB
15	bouteflika	Bouteflika	PROPN
16	w	et	CCONJ
17	rana	nous_sommes	VERB
18-19	m3ak	—	—
18	m3a	avec	ADP
19	k	toi	PRON

Table 2: Expert MWT segmentation and annotation.

Arabic varieties as shown in Table 3.

Singular Form	Plural Form	Gloss
/ʕasʕfu:r/	/ʕsʕa:fir/	bird
/sʕandu:q/	/sʕna:diq/	box
/xanfu:sah/	/xna:fis/	bug
/garʒu:mah/	/gra:ʒim/	throat

Table 3: Plural forms of the CvCCv:C pattern.

3.2. Modeling

In preparing this treebank, the NArabizi dataset served as the training input for a machine learning model developed with UDPipe1.0 (Straka et al., 2016). Once trained, this model was applied to Tunisian Arabic text data (see Section 3.3) to generate annotations and facilitate the rapid production of the resource.

The selection of UDPipe for this project was motivated by its comprehensiveness as a tool. It offers all the essential tasks for linguistic annotation within a single framework. This integration eliminates the need to combine multiple tools for different tasks and increases methodological consistency. In addition, it is freely available as an open-source resource making it an attractive option for an academic project with limited funding and al-

lows the study to be reproducible for other under-resourced Arabic varieties.

Although more recent versions have been released, UDPipe 1.0 was selected due to its architectural simplicity and ease of customization, which are particularly advantageous when working with a small-scale treebank, especially considering the scarcity of linguistic resources in the context of spoken Arabic.

3.3. Preprocessing and Annotation

The UDPipe model trained on the NArabizi corpus was applied to a subset of the CTAB⁴ dataset (Amara et al., 2021). The CTAB dataset was collected between 2017 and 2021 through web scraping of public Facebook pages belonging to Tunisian media outlets, including Atessia TV and Ettounsiya TV. The data comprises 2,929 comments written in the Arabizi script, totaling 26,232 tokens and covering a diverse range of topics such as politics, soccer, and social issues.

A random sample of 1k comments was selected for the study. Given time constraints and the potential cognitive load associated with the annotation process, the length of selected comments was limited to a maximum of 20 tokens, defined by whitespace separation. When segmented according to UD guidelines, such comments could yield between 20 and 40 lines, depending on the original spacing and the morphological complexity of the tokens.

A sample of 100 comments was subsequently selected from the cleaned sample and processed by the model to generate linguistic annotations. The resulting CoNLL-U file was duplicated and one copy was manually corrected by the author, a native speaker of Tunisian Arabic with an advanced background in linguistics. As noted in Section 3.1, the model was trained on data with suboptimal tokenization. Therefore, the initial step in the manual annotation phase involved reviewing and correcting the tokenization of each generated annotation. This process included not only segmenting multi-token words but also making decision regarding the splitting of grammaticalized contractions derived from Classical Arabic such as *inchallah* and abbreviations commonly found in user generated text then documenting these decisions as part of a guideline to ensure consistency throughout the annotation process.

Once tokenization was completed for a given sentence, lemmas were added using Buckwalter transliteration for words in Tunisian Arabic, and the

⁴<https://africarxiv.ubuntu.net/items/c14b535f-ac17-49ef-91e5-31e75a46eec9/full>

Roman alphabet in cases of code-switching. Subsequently, part-of-speech tags, morphological features, dependency relations, and other relevant annotations including the sentiment were applied.

3.4. Quality Control

To assess the internal consistency of the gold annotations, a series of automatic validation checks were applied using the script provided in the [UD-tools](#) repository. These checks verified structural integrity (Level 1) to ensure the file is a machine-readable CoNLL-U file. Additional checks verified for universal inventory (Level 2) to ensure the consistency of UPOS tags and DEPREL labels with the 17 official UD tags and 37 universal relations, respectively. The consistency of morphological features was verified to ensure the `Name=Value` format and alphabetical order. Furthermore, universal syntactic logic (Level 3) was checked to verify tree structure including single-root dependency trees and acyclic graphs.

The detected issues were manually reviewed and corrected, after which the treebank passed all Level 3 universal structural and logical checks. Since this is the first treebank for Tunisian Arabic, the validation script was only invoked to test validity up to Level 3. Validation checks for language-specific labels (Level 4) and language-specific guidelines (Level 5) could not be performed, as the language-specific documentation required by the UD validator has not yet been incorporated into the UD infrastructure. Given the long-term objective of officially releasing this treebank, future steps will include registering the language and defining its language-specific documentation, followed by an additional validation round to ensure compliance with Levels 4 and 5.

Finally, the gold annotations were systematically compared to those predicted by the model. Evaluation is conducted across multiple annotation layers using the evaluation script⁵ by [Zeman et al. 2018](#) from the CoNLL 2018 shared task.

3.5. Corpus Description

The Tunisian Arabic Dependency Treebank comprises 100 sentences in the Arabizi script extracted from social media comments, containing a total of 1,466 tokens. Sentence length varies from 5 to 46 syntactic words, with a mean of 14.66 and a median of 13. The vocabulary consists of 281 multiword tokens, roughly 23.7%, reflecting moderate morphological complexity typical of a clitic-heavy language such as Tunisian Arabic.

⁵https://github.com/ufal/conll2018/blob/master/evaluation_script/conll18_ud_eval.py

3.5.1. Parts-of-Speech

The tree includes 16 out of the 17 official UD POS tags defined in the UD v2 scheme. [Table 4](#) shows that the distribution of Universal parts of speech (UPOS) tags is dominated by verbs, nouns, then pronouns, which together constitute nearly half of the corpus. XPOS tags were not added in this study since the focus was on the annotations layers actually used by the UD community for cross-linguistic research.

UPOS Tag	%
VERB	17.5
NOUN	16.3
PRON	15.6
ADP	10.0
DET	7.3
PART	5.7
PROPN	4.6
ADJ	4.6
CCONJ	4.1
Other (INTJ, ADV, etc.)	14.2
Total	100

Table 4: Distribution of universal part-of-speech (UPOS) tags ($N = 1,466$).

3.5.2. Morphological Analysis

As summarized in [Table 5](#), the treebank is heavily marked for nominal features, with singular number (42.6%) and masculine gender (25.2%) being the most common. The high frequency of 3rd and 2nd person markings along with the active voice and the finite verb form align with the verbal-heavy nature of the distribution showed in [Table 4](#).

Feature	%*
Number=Sing	42.6
Gender=Masc	25.2
Gender=Fem	16.2
Voice=Act	15.8
Person=3	14.9
VerbForm=Fin	14.9
PronType=Prs	13.5
Mood=Ind	12.6
Definite=Ind	11.7
Number=Plur	10.0

*Percentages calculated against total syntactic tokens.

Table 5: Distribution of key morphological features ($N = 1,466$).

3.5.3. Dependency Relations

Syntactically, the treebank is dominated by core arguments and modifiers, including direct objects, case markers, nominal modifiers, and nominal subjects as shown in Table 6. This distribution indicates that the treebank captures the richness and complexity of argument structure, providing a useful foundation for future parsing tasks and research on the syntactic structure of the language. In the current version, enhanced dependencies are not included, as the focus is on establishing a solid basic UD foundation.

Relation	%
obj	9.1
case	8.9
nmod	7.6
det	7.0
root	6.8
nsubj	6.8
parataxis	6.8
obl	6.1
discourse	5.3
Other	35.6
Total	100

Table 6: Most frequent dependency relations in the corpus.

3.5.4. Misc

The corpus exhibits a typical multilingual profile of the North African region, with Tunisian Arabic serving as the main language, accounting for 94.07% of the total syntactic tokens. As detailed in Table 7, code-switching is observed primarily with MSA (2.86%) and French (2.73%). A negligible portion of the data (0.34%) consists of English. This distribution confirms that the treebank successfully captures the code-switching patterns characteristic of contemporary Tunisian speech.

Language	%
Tunisian Arabic (aeb)	94.07
Modern Standard Arabic (arb)	2.86
French (fr)	2.73
English (en)	0.34
Total	100.0

Table 7: Distribution of languages and code-switching ($N = 1,466$).

3.5.5. Lemmas

As noted in Section 3.1, this study questions the decision of Seddah et al. to use French glosses as

lemmas and rejects the claim that lemmatization necessarily requires an etymological analysis of the language. Table 8 summarizes the lemmatization conventions by part of speech and language applied in TADT. In cases of code-switching, the lemma field is annotated in accordance with the existing UD guidelines for the language of the token. However, when a token originates from a foreign language but conforms in its usage to the morphological and syntactic system of Tunisian Arabic, such as the noun *garjouma* in Table 3, the lemma and all other layers of annotations follow the Tunisian Arabic conventions.

In line with UD standards, nouns are lemmatized to their singular nominative forms where applicable, or to the masculine singular nominative form in the case of nouns with masculine–feminine alternations. For verbs, a cross-linguistic convention is followed: the infinitive is used for French and English, while for Arabic the third-person masculine singular perfective is adopted, following the guidelines of the Prague Arabic Dependency Treebank, which serves as the reference standard for Arabic UD projects.

For English and French word forms, lemmas are encoded in the Roman alphabet. For Arabic, Buckwalter transliteration is used, as it is widely adopted in NLP due to its reversibility and its avoidance of non-ASCII characters and complex diacritics that are both present in other romanization systems.

POS	Lang	Lemma
VERB	aeb, arb	3SG.M.PV (perfective)
	fr, en	Infinitive
NOUN	aeb, arb, fr	singular/masculine singular*
	en	singular
PRON	aeb, arb, en, fr	nominative
DET	aeb, arb, en	identical word form
	fr	masculine singular (<i>le/un</i>)
ADJ	aeb, arb, fr	masculine singular
	en	identical word form
OTHER	aeb, arb, en, fr	identical word form

*For nouns with gendered counterparts (e.g., *professions*).

Table 8: Lemmatization conventions by part-of-speech and language.

The treebank exhibits a high degree of morphological and functional diversity, with 495 unique lemmas identified as shown in Table 9.

The distribution is dominated by functional elements, particularly the definite marker *Al* (5.87%) and the conjunction *w* (4.37%). As the CTAB dataset is extracted from social media, the high frequency of pronominal forms, such as *<inoti* (2.87%), *hiy~a* (2.73%), and *lnaA* (2.66%) as well as the frequent occurrence of the vocative particle *yaA* (1.57%) reflect the interactive nature of the social media comments.

Lemma	POS	%
<i>Al</i> (the)	DET	5.87
<i>w</i> (and)	CONJ	4.37
<i><inoti</i> (you)	PRON	2.87
<i>hiy~a</i> (she)	PRON	2.73
<i>fiy</i> (in)	ADP	2.73
<i>lnaA</i> (we/us)	PRON	2.66
<i>maA</i> (not)	PART	2.59
<i>huw~a</i> (he)	PRON	1.91
<i>\$</i> (not)	PART	1.77
<i>l</i> (for/to)	ADP	1.64

Table 9: Top 10 most frequent lemmas ($N = 495$).

4. Results

Table 10 presents the complete set of results generated using the evaluation script by Zeman et al.. Performance is reported using precision, recall, and F1 score for all metrics as well as aligned accuracy (AlignedAcc) for annotation layers that require word level alignment. The findings suggest that the model achieves a strong performance at the tokenization level with an F1 score of approximately 90%. The sentence segmentation is perfect across all metrics.

In the UD framework and the CoNLL-U format, words refer to syntactic units which can occur separately as a separate token or attach to other syntactic words to form a MWT. Performance on words, which reflects the segmentation of syntactic words including the expansion of MWTs is substantially lower than for tokens. The model attains an F1 score of 60.76%. The significant drop in recall indicates that approximately 46% of syntactic words were not detected by the model.

Universal part-of-speech (UPOS) tagging performance is low across all metrics. The model identifies basic parts of speech only about 29%. The aligned accuracy metric indicates that 47% of the words the model was capable of aligning correctly with the gold annotation received the correct UPOS tag. The markedly low performance on language-specific-POS tags (XPOS) does not reflect any failure on the part of the model since the gold annotations did not include any.

The annotation of morphological features

(UFeats) also exhibits low performance with F1 score of 18% and the aligned accuracy further confirms that even alignment is not achieved, approximately 70% of the features are predicted incorrectly. When POS tags and morphological features are evaluated jointly (AllTags), performance deteriorates sharply. This drop is expected since this metric requires all the tags of the three layers above to be simultaneously correct.

For dependency parsing, the model achieves an F1 score of 18.34 for the unlabeled attachment score (UAS). These results indicate that even when ignoring dependency labels, only 18% of the words are pointing to the correct parent word. For words that were correctly aligned, 30.18% have the correct head. Performance decreases further when dependency labels are taken into account, with the labeled attachment score (LAS) yielding a score of 11.42. Only 11% have both the correct head word and have the correct dependency relation label. For words that were correctly aligned, only 18.80% have the correct head and label.

CLAS, which restricts evaluation to content words, drops further to an F1 score of 6.74 and an aligned accuracy of 11.82%. As the strictest dependency metrics, MLAS and BLEX require the correct prediction of dependency attachment in addition to the accurate morphology or lemma prediction and are near zero with F1 values of 0.52 and an aligned accuracies below 1%, thereby revealing the compounding impact of errors introduced at the early stages of annotation pipeline.

5. Discussion

Findings from Table 10 reveal a clear decline in performance across annotation layers. While sentence segmentation is handled perfectly, this metric does not reflect anything on the ability of the model to identify sentence boundaries since the data was already segmented into sentences in the pre-processing stage. Surface tokenization also achieves a strong performance indicating that the model generally identifies the physical boundaries of strings with high accuracy, but the gap between precision and recall suggests that the model is leaning towards undertokenizing.

However, performance starts to deteriorate moving from token to word-level segmentation. The drop in recall highlights the weakness of the model in handling MWTs. Since standard CoNLL-U evaluation assumes a one-to-one correspondence between gold and predicted lines, the model's failure to split a MWT results in systematic misalignment between gold and predicted annotations, thereby preventing reliable evaluation and comparison of morphosyntactic annotations. This misalignment is reflected in the sharp decline observed for mor-

Metric	Precision	Recall	F1 Score	AlignedAcc
Tokens	93.00	87.15	89.98	–
Sentences	100.00	100.00	100.00	–
Words	70.58	53.34	60.76	–
UPOS	33.48	25.31	28.83	47.44
XPOS	0.54	0.41	0.47	0.77
UFeats	21.84	16.51	18.80	30.95
AllTags	0.09	0.07	0.08	0.13
Lemmas	4.24	3.21	3.65	6.01
UAS	21.30	16.10	18.34	30.18
LAS	13.27	10.03	11.42	18.80
CLAS	7.51	6.11	6.74	11.82
MLAS	0.58	0.47	0.52	0.91
BLEX	0.58	0.47	0.52	0.91

Table 10: Evaluation results comparing gold and predicted CoNLL-U annotations.

phological and lexical annotations.

While the model struggles with POS tagging and the capture of fine-grained morphological distinctions, the slight improvement observed in the AlignedAcc scores suggests that its overall performance might be dragged down by its weakness in handling the morphosyntactic complexity of MWTs, given that it was trained with poorly segmented data.

Although the small size of the treebank can also affect the reliability of results and the overall poor performance of the model is discouraging, the boost in performance observed in the Aligned Acc scores for UPOS, UFeats, and UAS may be viewed as an encouraging indication for further investigation using refining techniques. These findings further suggest that a cross-dialectal transfer approach in the context of spoken Arabic might be effective only under two conditions. First, a strict adherence to shared guidelines governing the segmentation and tagging of Arabic varieties. Second, the delegation of annotation to trained linguists rather than to native speakers who, despite their language proficiency, lack the analytical skills required for sound linguistic judgment, potentially introducing errors in annotation that lead to cascading consequences.

Regarding the model-assisted workflow employed in this study, a comparative analysis between manual and semi-automatic annotation efficiency was precluded by resource constraints. Specifically, as the author served as the sole annotator, the simultaneous development of a manual baseline from the Tunisian dataset was methodologically unfeasible. The model generated a first-pass CoNLL-U file with reliable tokenization, thereby eliminating the initial and most repetitive stage of manual annotation.

Due to the model’s word-level segmentation performance, the manual effort was sufficiently restricted to 39.3% of instances—a considerably

less taxing requirement than defining all syntactic words from a blank string. Furthermore, despite lower performance in categories such as UPOS, the model successfully automated the assignment of approximately one-third of the tags, reducing the cognitive burden of manual labeling. Consequently, the model’s output reallocated the annotator’s effort toward higher-level linguistic tasks that required extensive manual correction, such as morphological analysis and dependency parsing.

6. Conclusion

This paper presented the first Tunisian Arabic Treebank focused on user-generated content, capturing the code-switching patterns characteristic of North African Arabic varieties. Annotated according to Universal Dependencies standards, this resource promotes interoperability among Arabic language resources. Despite its limited scale, the treebank serves as a critical asset for both linguistic research and NLP development; it is currently being finalized for official release within the UD framework.

Methodologically, this study evaluated the feasibility of employing machine learning to assist linguists in generating labeled corpora, aiming to mitigate existing resource gaps. While initial results suggest limited immediate efficacy, observed performance trends provide encouraging evidence for the approach’s potential. Future work will focus on expanding the treebank and implementing an iterative fine-tuning process to enhance model performance through optimized data quality. The ultimate objective remains to achieve a level of predictive reliability that significantly accelerates the development of gold-standard resources.

7. Bibliographical References

- F Al-Shargi and O Rambow. 2015. DIWAN: A Dialectal Word Annotation Tool for Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 49–58.
- H Bouamor, N Habash, and K Oflazer. 2014. A Multidialectal Parallel Corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245.
- R Boujelbane, E Khemekhem, M, S Ben Ayed, and H Belguith, L. 2013. Building bilingual lexicon to create Dialect Tunisian corpora and adapt language model. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 88–93.
- C Fourati, A Messaoudi, and H Haddad. 2020. TUNIZI: a Tunisian Arabizi sentiment analysis Dataset. In *1st AfricaNLP Workshop Proceeding, AfricaNLP@ICLR 2020*, pages 1–3.
- E Gugliotta and M Dinarelli. 2022. TARc: Tunisian Arabish Corpus, First complete release. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1125–1136.
- N Habash, O Rambow, and R Roth. 2009. MADA + TOKAN : A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, pages 102–109.
- S. Harrat, K. Meftouh, M. Abbas, S. Jamoussi, M. Saad, and K. Smaili. 2015. Cross-Dialectal Arabic Processing. In *Computational Linguistics and Intelligent Text Processing*, pages 620–632. Springer.
- N Ide, N Calzolari, J Eckle-Kohler, D Gibbon, S Hellman, K Lee, J Nivre, and L Romary. 2017. Community Standards for Linguistically-Annotated Resources. In N Ide and J Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 113–165. Springer.
- Teresa Lynn. 2016. *Irish Dependency Treebanking and Parsing*. Ph.D. thesis, Dublin City University and Macquarie University.
- S Medhaffar, F Bougares, Y Esteve, and L Hadrich-Belguith. 2017. Sentiment Analysis of Tunisian Dialect: Linguistic Resources and Experiments. In *Proceedings of the Third Arabic Natural Language Processing Workshop (WANLP)*, pages 55–61.
- K Meftouh, S Harrat, S Jamoussi, M Abbas, and K Smaili. 2015. Machine Translation Experiments on PADIC: A Parallel Arabic Dialect Corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 26–34.
- H Saadane and N Habash. 2015. A Conventional Orthography for Algerian Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 69–79.
- F Sadat, F Mallek, R Sellami, M Boudabous, M, and A Farzindar. 2014. Collaboratively Constructed Linguistic Resources for Language Variants and their Exploitation in NLP Application – the case of Tunisian Arabic and the Social Media. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 102–110.
- Y Samih. 2017. *Dialectal Arabic Processing Using Deep Learning*. Ph.D. thesis, Düsseldorf, Heinrich-Heine-Universität.
- D Seddah, F Essaidi, A Fethi, M Futrel, B Muller, J Ortiz Suárez, P, B Sagot, and A Srivastava. 2020. Building a User-Generated Content North-African Arabizi Treebank: Tackling Hell. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1139–1150.
- M Straka, J Hajič, and J Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.
- R Zbib, E Malchiodi, J Devlin, D Stallard, S Matsoukas, R Schwartz, J Makhoul, F Zaidan, O, and C Callison-Burch. 2012. Machine Translation of Arabic Dialects. In *2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59.
- D Zeman, J Hajič, M Popel, M Potthast, M Straka, F Ginter, J Nivre, and S Petrov. 2018. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21.
- I Zribi, M Khemakhem, and H Belguith, L. 2013. Morphological Analysis of Tunisian Dialect. In *International Joint Conference on Natural Language Processing*, pages 992–996.

8. Language Resource References

A Amara, and H Turki, and M A Hadj Taieb, and M Ben Aouicha, and K Ellouze. 2021. *CTAB: Corpus of Tunisian Arabizi [Dataset]*. Zenodo. PID <https://doi.org/10.5281/zenodo.4781769>.

M Maamouri, and A Bies, and T Buckwalter, and H Jin, and W Mekki. 2005. *Arabic Treebank: Part 3 (full corpus) v 2.0 (MPG + Syntactic textscAnalysis)*. LDC2005T20. Linguistic Data Consortium. PID <https://catalog.ldc.upenn.edu/LDC2005T20>.

From Treebank Metadata to Sentence-Level Genre in Universal Dependencies: A Reproducible, Versioned Resource

Egon W. Stemle

Institute for Applied Linguistics, Eurac Research
Faculty of Informatics, Masaryk University
egon.stemle@eurac.edu

Abstract

We release a sentence-level genre layer for Universal Dependencies as a separate, joinable dataset, computed across UD revisions and linked back to the underlying treebanks via a release-aware composite key comprising treebank, split, `sent_id`, and UD release metadata. The annotations are derived rather than authoritative and are accompanied by provenance and uncertainty indicators, enabling downstream users to choose appropriate precision-coverage trade-offs and to re-run the pipeline as UD evolves. To support both parity tracking and deployment-oriented interpretation, we report results under two complementary regimes: a fixed-partition setting aligned with earlier protocols, and a language-grouped 10-fold generalisation setting that highlights cross-language heterogeneity and anchor sparsity as operational constraints. The resulting resource is intended to make genre a practical control variable for UD-based experimentation, including genre-stratified evaluation and training data selection for POS tagging and parsing, where performance varies substantially across text types. Finally, we note that reduced genre spaces aligned with recurring robustness profiles (e.g. transcribed speech versus interactional web/social text versus edited prose/news) appear pragmatically useful, but should be treated as a community coordination task implemented through explicit, versioned mapping tables.

Keywords: sentence-level genre annotation, cross-lingual genre inference, linguistic data infrastructure

1. Introduction

Universal Dependencies (UD) is a central multilingual resource for morphosyntactic annotation, but its genre metadata remains coarse: genres are declared at treebank level, often multi-valued, and usually not linked to individual sentences (Nivre et al., 2020; Müller-Eberstein et al., 2021b; Danilova and Stymne, 2023). This is limiting whenever treebanks mix edited prose, interactional Web text, speech-like material, or learner writing, because treebank identity then becomes a poor proxy for the text conditions that matter in downstream experiments.

Sentence-level genre is useful for at least two practical reasons. First, it supports genre-stratified evaluation and error analysis in POS tagging, parsing, and related tasks, where performance can differ sharply between edited and non-edited text (Giesbrecht and Evert, 2009; Owoputi et al., 2013a; Behzad and Zeldes, 2020). Second, it supports training-data selection and cross-genre transfer, including settings where in-genre or genre-aware proxy data is more informative than simply choosing data by language or size (Rehbein and Bildhauer, 2017; Müller-Eberstein et al., 2021a).

Two existing lines of work frame the present paper. Müller-Eberstein et al. show that sentence-level genre can be weakly recovered from UD via multilingual embeddings, treebank-local clustering, and metadata-driven labelling (Müller-Eberstein

et al., 2021b). UD-MULTIGENRE shows that careful manual reconstruction can yield high-quality instance-level labels, but only for a limited subset of UD (Danilova and Stymne, 2023). The resulting gap is infrastructural: weak supervision scales, manual curation is precise, but neither by itself yields a persistent, versioned genre layer for UD as a whole.

We address that gap by releasing sentence-level genre as a separate sidecar dataset rather than modifying UD treebanks directly. The layer is linked back to UD via a composite sentence reference over treebank, split, `sent_id`, and release metadata; it is versioned across UD releases; and it is generated by a reproducible pipeline that combines metadata extraction with bootstrapped clustering. Alongside the labels, we publish provenance and uncertainty information so that consumers can trade off coverage and precision explicitly.

Our contribution is therefore twofold: (i) a full-coverage, reproducible sentence-level genre layer for UD treebanks `hf://datasets/commul-ud_genre` (Stemle, 2026a), and (ii) the supporting infrastructure `hf://datasets/commul-universal_dependencies` (Stemle, 2026b) and `ud-hf-parquet-tools` (Stemle, 2026e) that makes this layer regenerable and inspectable across releases. The paper also distinguishes paper-style comparability from end-user sentence-level evaluation, since these optimise different targets.

2. Background and Related Work

2.1. Genre Information in Universal Dependencies

UD’s genre inventory includes familiar labels such as *news*, *wiki*, *blog*, *email*, *spoken*, *fiction*, and *legal*, but the values mix medium, domain, communicative situation, and editorial status, as the [Universal Dependencies contributors \(2026\)](#) acknowledge. The labels are assigned at treebank level, often multi-valued, and there is no standard machine-readable mechanism for attaching them to individual sentences. For multi-genre treebanks this leaves the internal genre distribution largely implicit.

2.2. Weakly Supervised and Curated Sentence-Level Genre

Müller-Eberstein et al. provide the main weakly supervised reference point: sentence embeddings, treebank-local clustering, and metadata-driven labelling can recover sentence-level genre signals at scale, and local clustering is crucial because multilingual embeddings otherwise reflect language more strongly than genre ([Müller-Eberstein et al., 2021b](#)). This line of work scales to the full UD collection, but has mainly been distributed as experimental code rather than as a persistent data artefact.

UD-MULTIGENRE ([Danilova and Stymne, 2023](#)) takes the complementary route: manual analysis of documentation and metadata yields high-confidence instance-level labels for a substantial but still partial UD subset. It is therefore highly valuable as a curated reference, but not by itself a full-coverage solution.

2.3. Positioning of the Present Work

Our work sits between these two strands. We do not introduce a new genre taxonomy, and we do not aim to replace manually curated resources. Instead, we operationalise sentence-level genre as a versioned, reproducible sidecar layer for UD, built from release-preserving data access, auditable metadata extraction, and weakly supervised extrapolation. In that sense, the contribution is primarily infrastructural: making existing genre-recovery ideas reusable as data, not only as one-off experiments.

3. Methodology

Our methodological point of departure is the observation that genre information in Universal Dependencies (UD) is, in practice, both valuable and structurally under-specified at the instance level:

UD treebanks declare which genres are present, yet the prevailing representation remains an inventory at the treebank level, without a standardised machine-readable mechanism to associate individual sentences with genre labels ([Universal Dependencies contributors, 2026](#)). In this context, sentence-level genre can be understood as a missing intermediate layer between (i) community-facing corpus descriptions and (ii) infrastructure-facing data access patterns that increasingly assume selective, instance-level filtering.

Against this background, we operationalise a framework that derives sentence-level genre annotations for all UD treebanks across multiple UD releases, while keeping the primary UD data unchanged. Concretely, we treat sentence-level genre labels as a *sidecar layer* that is linked back to UD sentences via a composite key over treebank, split, `sent_id`, and release metadata, and is published as a separate dataset. This separation is not only a pragmatic distribution choice, but it also reflects a layered responsibility model: UD remains the authoritative source for syntactic annotation, whereas derived, task-facing enrichments can evolve in parallel, with explicit provenance and uncertainty signalling.

3.1. Data Basis and Versioned Access

3.1.1. UD Snapshots as Infrastructure Objects

The empirical basis of the framework is a revision-preserving packaging of UD treebanks in Parquet format, exposed as a dataset repository on the Hugging Face Hub. The repository distinguishes explicitly between *framework versioning* (tooling and schema evolution) and *UD data versioning* (the linguistic release snapshots), and it maintains a separate branch for each UD data release. In addition, the dataset repository carries a `tools/` subdirectory with regeneration and inspection scripts so that the published Parquet snapshots are not only downloadable but reproducible from the underlying UD sources. This design makes it possible to compute sentence-level genre annotations consistently for a defined range of UD releases (in our current setup: 2.7 through 2.17), and to keep release-specific inventories intact rather than implicitly normalising across versions. ([Stemle, 2026b](#))

The practical motivation for this choice is that genre inventories, treebank composition, and metadata conventions are not stable across UD releases. A methodology that collapses these differences too early risks conflating (i) genuine linguistic or corpus shifts with (ii) release-driven annotation or documentation changes. Preserving release-specific snapshots, while enabling explicit cross-version mappings, supports both reproducibility and diag-

nostic clarity.

Fidelity-preserving conversion to Parquet. A second infrastructural dependency is a conversion and validation toolchain that generates Parquet representations from CoNLL-U and validates them against the original inputs with explicit fidelity checking (Stemle, 2026e). We deliberately keep this toolchain in the separate `ud-hf-parquet-tools` repository, because the Hugging Face dataset repository is the right place to version released data artefacts, whereas iterative parser, schema, and validation development benefits from a conventional software repository. The tooling is designed to preserve information that is often treated as peripheral but becomes central for metadata-based enrichment, including ordered comment blocks (with duplicates), edge cases in CoNLL-U parsing, and special node types such as multi-word tokens and empty nodes. In particular, the schema exposes `sent_id`, split identity, and preserves sentence-level comments as an ordered list, all of which are prerequisites for robust genre extraction under heterogeneous metadata practices.

From a methodological perspective, this step is not merely an implementation detail. Comment lines and documentation-derived conventions largely mediate genre extraction in UD. Consequently, a conversion step that drops, normalises, or de-duplicates comment metadata would directly bias both coverage estimates and evaluation outcomes.

3.1.2. Genre label space and cross-version comparability

Canonical target labels: UD global genres. For the canonical label space, we primarily adopt UD global genres as defined by UD’s genre documentation and decision list. This choice aligns the sidecar annotations with an existing community-facing taxonomy, and it retains a clear conceptual anchor: the goal is not to introduce a competing genre scheme, but to make an existing UD metadata dimension operational at the sentence level.

At the same time, UD’s documentation makes visible that the taxonomy has evolved, including explicit changes to formerly permitted genre values and subsequent consolidations. As a consequence, cross-version comparability cannot be treated as a trivial join on string labels; instead, it requires an explicit mapping layer that acknowledges both label drift and treebank-specific conventions.

Mapping tables as a mediation layer. We therefore distinguish between:

1. **Release-specific inventories**, i.e. the set of genre descriptors and metadata signals observed in a given UD release snapshot (kept intact per branch), and
2. **Canonical genres**, i.e. a normalised target set used for cross-treebank and cross-version aggregation.

In the present framework, the mapping layer is seeded from the `tb-genres.json` resource distributed with Müller-Eberstein (2021), which provides mappings from treebank-specific genre labels to global UD genre labels for a subset of treebanks.

However, a more comprehensive, version-aware mapping table is likely needed for long-term maintenance. In practice, this suggests a community-maintainable artefact that (i) records release-local genre strings and extraction cues, (ii) maps them to canonical UD genres for that release, and (iii) optionally provides a second mapping to a stable comparison layer spanning multiple UD versions. The intended contribution of our framework is to make such mediation explicit and testable, rather than leaving it implicit in ad hoc downstream pre-processing.

3.1.3. Metadata-based sentence-level genre extraction

Heterogeneous metadata as an empirical constraint. UD treebanks are heterogeneous not only linguistically but also in their metadata practices. The UD documentation notes that there is no standardised machine-readable way to identify which sentences belong to which genre, and sketches sentence-id ranges as a possible, but not standardised, mechanism. In operational terms, this means that sentence-level genre signals are typically encoded in treebank-specific comment conventions (e.g. document identifiers, source references, partition markers, or informal metadata keys).

Our extraction component treats this heterogeneity as the default case. Instead of assuming one universal comment key, we implement treebank-aware extraction patterns, organised as a configuration layer that can express both:

- **direct genre indicators** (explicit per-sentence labels, where available), and
- **indirect indicators** (document-level or source-level metadata that can be propagated to sentences via structural anchors, such as document boundaries).

Pattern testing and coverage diagnostics. To make extraction patterns maintainable and reviewable, we provide explicit diagnostics functionality in

the genre bootstrapping toolkit. In practice, these commands are part of the release workflow rather than ad hoc debugging utilities, because each promoted release is preceded by coverage checks and treebank-level spot tests:

- `test-genres`, which supports iterative testing and debugging of extraction patterns, and surfaces coverage statistics to quantify how much of a treebank is addressable by the current extraction rules (Stemle, 2026d).
- `coverage`, which analyzes sentence-level genre coverage across treebanks and can report fully covered versus partially covered inventories, including exportable summaries for regression testing of pattern changes (Stemle, 2026c).

Methodologically, this diagnostic layer is treated as part of the main pipeline rather than as an afterthought. Extraction quality is constrained by (i) metadata availability and (ii) mapping adequacy, and both dimensions vary systematically across languages, genres, and release snapshots. A pipeline that only reports final labels, without systematic coverage accounting, makes it difficult to distinguish between “no genre signal exists” and “a signal exists but is not captured by current patterns.”

3.1.4. Extrapolation via weak supervision and clustering

Motivation and relation to prior work. Even with extensive metadata patterning, a substantial fraction of UD data remains multi-genre at the treebank level, without explicit instance-level labels. Müller-Eberstein et al. (2021b) characterise this gap and propose weakly supervised methods for predicting instance-level genre, emphasising that treebank-level metadata alone is a noisy signal that needs disentanglement within treebanks. Related work frames genre metadata as a useful supervisory signal for cross-lingual dependency parsing, in part by projecting treebank-level genre information to sentence level (Müller-Eberstein et al., 2021a).

In this context, we implement a bootstrapped clustering approach aligned with the GMM-based family of methods discussed in this line of work. The core idea is to treat sentence embeddings as a multilingual representation space in which genre-related variation is partially recoverable, and to use existing single-genre resources as anchors for labelling.

Embeddings and clustering. For each UD release snapshot, we generate sentence embeddings from the token `FORM` sequence (i.e. syntactic words), as provided in the Parquet representation.

The choice of using token forms rather than reconstructed text is motivated by cross-treebank consistency and by the practical availability of tokenized input in UD. In the current promoted profile, this stage uses multilingual-e5-large with mean pooling over tokenised input and runs identically over local CoNLL-U snapshots and Hugging Face-backed Parquet snapshots.

We then apply clustering at the treebank level. In the current default profile, clustering is performed using Gaussian Mixture Models (GMM), reflecting the need for soft assignments and explicit uncertainty quantification. The number of mixture components is derived from the release-specific genre inventory associated with the treebank, preserving the release-local assumption about how many genres are expected to be present.

Bootstrapped labelling and uncertainty retention. Cluster-to-genre labelling is performed via a bootstrapping mechanism that uses reference material from treebanks that are reliably single-genre (or have sufficiently stable metadata-derived labels). Genre representations are derived from these anchors, and cluster centroids are mapped to genres via similarity-based assignment. Importantly, the pipeline retains uncertainty explicitly rather than forcing hard labels for all instances. In the exported sidecar dataset, this is reflected by separating higher-confidence assignments from lower-confidence inferences, and by shipping row-level provenance (release, treebank, split, configuration, method) together with confidence values, enabling downstream users to trade off coverage against precision in a controlled way.

This design choice follows from the broader observation, already present in prior work, that genre boundaries are not always cleanly separable, and that metadata-derived supervision may be incomplete or ambiguous (Müller-Eberstein et al., 2021b).

3.1.5. Evaluation-oriented validation and release-aware reporting

While the main paper reports results in a dedicated evaluation section, the methodology includes evaluation-oriented validation steps because they influence how we interpret coverage and failure cases.

First, we validate metadata extraction and mapping consistency by comparing predicted labels against available instance-level anchors, where such anchors exist. In the short term, this includes UD-internal signals and the subset of treebanks for which instance-level labels can be reconstructed reliably; in the medium term, we plan alignment against UD-MULTIGENRE, which provides an explicitly enriched UD-based dataset with instance-

level genre annotations and a discussion of inconsistencies in UD genre documentation and usage.

Second, the pipeline supports evaluation regimes that aim to balance comparability with robustness. In particular, our default configuration promotes a strict anchor mode with a training-virtual anchor pool policy, reflecting an attempt to minimise leakage between anchor selection and evaluation targets when assessing generalisation across heterogeneous treebanks.

Third, because our outputs are computed for multiple UD releases, we report coverage and label distributions per release snapshot. This is not only a reproducibility measure, but it also enables identifying when apparent performance changes are better explained by release-driven metadata shifts than by methodological differences.

3.1.6. Publishing as a linked sidecar dataset

A central practical decision is to publish sentence-level genre annotations as a separate dataset on the Hugging Face Hub, linked back to UD sentences through a composite key that includes UD release, treebank identifier, split, and `sent_id`. The key point is that this linkage strategy does not require modifying UD treebanks, and it remains compatible with both Parquet-based workflows and CoNLL-U-based reconstruction. Operationally, each release also carries a frozen configuration snapshot, copied mapping files, and release metadata so that consumers can inspect not only the labels but the exact conditions under which they were produced.

In this framing, sentence-level genre annotation becomes an instance of a more general pattern: attaching derived, task-facing metadata to stable, community-maintained linguistic resources via persistent identifiers. Once this pattern is established, it can be reused to attach additional layers (e.g. domain labels, quality flags, register estimates) in a way that remains compatible with UD’s existing governance and release model.

3.1.7. Reference configuration profile

For transparency and reproducibility, we record a default “high-quality” profile that we currently treat as the baseline configuration for generating sentence-level genre annotations.

While the framework is parametrisable, explicitly documenting a default profile serves two functions: (i) it supports replication of the released sidecar dataset, and (ii) it provides a stable reference point for subsequent methodological comparisons, including the planned UD-MULTIGENRE-aligned variant.

4. Evaluation

4.1. Evaluation Orientation

The present work sits between two evaluation targets that should not be conflated. The original GMM+L study on UD v2.8 pools all original test splits into a single 204k-sentence global test partition and reports purity (PUR), agreement (AGR), overlap error (ΔBC), and micro-F1 over the subset of treebanks with instance-level labels (Müller-Eberstein et al., 2021b, Section 5.1 and Table 1). For GMM+L, the published values are $PUR = 1.00$, $AGR = 1.00$, $\Delta BC = 0.04$, and $\text{micro-F1} = 0.54$. Our resource, however, is intended for end users who need sentence-level labels that remain useful across new languages, revised treebanks, and changing metadata coverage. We therefore distinguish three notions of parity: *protocol parity* (reconstructing the fixed-partition setup and anchor semantics), *implementation parity* (matching the predictions of an original-like GMM+L pipeline on the reconstructed split), and *end-user evaluation* (sentence-level performance under broader language-held-out conditions).

4.2. Evaluation Regimes

Fixed-partition comparability. To track parity with the original protocol, we reconstruct the paper-style UD v2.8 test partition from the released index split, use whole-treebank single-genre anchors from the same partition, cluster all multi-genre treebanks in that reconstructed test partition, and score only the mapped paper treebanks. This reproduces the paper’s evaluation geometry more closely than our broader fixed-partition ablations. Importantly, we still compute sentence-level micro-/macro-F1 on the recoverable sentence-labelled subset, because this is the metric relevant to downstream users; PUR, AGR, and ΔBC are retained as treebank-level comparability diagnostics.

Language-grouped 10-fold generalisation. To approximate the operational setting in which the genre layer is applied to new languages, we additionally run 10-fold cross-validation grouped by *language*. This grouping makes the held-out condition explicit and reduces trivial leakage from shared-language anchors. The evaluation set in this regime is broader (“all_focused”), and the promoted baseline uses multilingual-e5-large embeddings, GMM clustering, and the `combined` anchor policy. The objective is diagnostic and deployment-oriented: to make visible cross-language variance, anchor sparsity, and the cost of stronger coverage assumptions.

4.3. Anchor Construction and Policy Sensitivity

A structural constraint of the UD genre setting is that sentence-level reference labels are not uniformly available. Instance-level “anchor” labels are reconstructed through metadata extraction patterns (comment metadata and `sent_id` conventions), seeded from prior mapping artefacts and iteratively refined using explicit pattern testing and coverage diagnostics. To avoid treating extraction artefacts as model behaviour, we restrict evaluation to treebanks and genres that satisfy strict coverage and minimum-size constraints, and we explicitly report missing-anchor genres.

In the 10-fold regime, we distinguish two anchor policies. The `train_virtual` policy restricts anchors to training-side virtual single-genre subsets, which functions as a conservative baseline and a regression guard. The `combined` policy expands the anchor pool, which can increase coverage and reduce minority-genre sparsity, but also changes the supervision surface and therefore has to be interpreted as a deliberate trade-off rather than as a free improvement.

4.4. Metrics

We report sentence-level and treebank-level metrics separately because they answer different questions. Sentence-level micro-F1 and macro-F1 ask whether individual sentences receive the correct label wherever sentence-level gold can be recovered. This is the primary end-user target: a system can reproduce the overall genre profile of a treebank while still assigning many individual sentences incorrectly.

PUR, AGR, and ΔBC , by contrast, are treebank-level comparability metrics. They ask whether predicted label distributions are internally coherent and whether they align with the expected genre composition of treebanks as specified by metadata. These measures are useful for tracking continuity with Müller-Eberstein et al. (2021b), but they are weaker proxies for the practical utility of a sentence-level annotation layer and can look deceptively strong under collapsed or prior-driven solutions. For that reason, we treat sentence-level F1 as primary and clustering-oriented metrics as secondary diagnostics.

4.5. Quantitative Results

Before turning to the broader end-user results, the parity story can now be stated precisely. On the corrected reconstructed UD v2.8 split, our strict protocol-parity run reaches micro-F1 0.281, macro-F1 0.122, PUR 0.432, AGR 0.625, and ΔBC 0.241 under the current sentence-level evaluator. A direct

audit against an original-like reimplement of GMM+L on the same split yields prediction agreement of > 0.98 on all eight scored treebanks. In other words, once the split reconstruction and anchor semantics are aligned, the remaining gap to the paper’s published numbers is not an implementation gap.

Table 1 summarises the three numbers that matter most for interpretation. The first row gives the published GMM+L values from Müller-Eberstein et al. (2021b) under the original paper protocol. The second row gives our reconstructed protocol-parity result under the current sentence-level evaluator. The third row gives the fresh end-user baseline: 10-fold, language-grouped evaluation on the broader UD v2.17 `all_focused` set with multilingual-e5-large, GMM clustering, and the `combined` anchor policy. This baseline yields overall micro-F1 0.333 and macro-F1 0.264; the corresponding fold means are 0.390 ± 0.146 and 0.325 ± 0.163 . Unlike the paper-reported row, this baseline is a sentence-level generalisation result on the current broader release rather than a fixed-protocol comparability score.

4.6. Qualitative Confusion Analysis

Confusion-matrix diagnostics, provided as row-normalised heatmaps, suggest that several errors correspond to genre boundaries that are difficult to separate from sentence-bounded context alone. In particular, the broader evaluation runs show systematic cross-talk among *news*, *fiction*, *spoken*, and *nonfiction*, and the conservative anchor policy exhibits a stronger concentration towards dominant assignment pathways. In the fixed-partition setting, uniform reference weighting reduces several collapse patterns observed under sentence-count weighting, which is consistent with the observed improvements in instance-level metrics.

4.7. Relation to Earlier Reported Numbers

A direct numerical comparison with the original paper is therefore informative only if the type of parity is stated explicitly. At the protocol level we recover the paper-style split construction and anchor geometry; at the implementation level we now reproduce the predictions of an original-like GMM+L pipeline exactly on the scored subset. The remaining metric gap arises downstream of the algorithm, chiefly from differences between the paper’s treebank-level genre inventories and the sentence-level gold that can be recovered today from the reconstructed subset, as well as from missing-anchor genres in that subset. For example, several paper-expected genres are absent from the recoverable sentence-level gold in the scored portions of Belarusian-HSE,

Regime	Setting	TB	Micro	Macro	PUR	AGR	ΔBC
Paper	GMM+L, UD v2.8 fixed test	–	0.540	–	1.000	1.000	0.040
Protocol parity	Reconstructed UD v2.8 fixed test	8	0.281	0.122	0.432	0.625	0.241
End-user baseline	10-fold by language, UD v2.17 <i>all_focused</i>	33	0.333	0.264	0.557	0.592	0.059

Table 1: Published GMM+L values, our reconstructed protocol-parity result, and the current end-user baseline. The protocol-parity run matches an original-like GMM+L implementation on all eight scored treebanks; the end-user baseline has fold means of micro 0.390 ± 0.146 and macro 0.325 ± 0.163 . The rows are not directly comparable because they target different evaluation objects.

Czech-CAC, and Russian-Taiga. This makes the fixed-partition parity score a useful comparability instrument, but not by itself the right target for end users.

For this reason, we treat the language-grouped 10-fold evaluation as the main baseline for future improvements. Improvements over the original paper can then be contextualised in two ways: (i) whether they preserve or improve protocol parity, and (ii) whether they improve the more operational sentence-level baseline under held-out language conditions.

4.8. Limitations and Scope

The reported results should be interpreted in the context of three interacting constraints. First, sentence-level genre supervision in UD is sparse and structurally heterogeneous, as only a minority of UD material provides instance-level genre labels in a reusable form, while the majority is annotated at the treebank level and often with multi-genre labels (Müller-Eberstein et al., 2021b). Second, the construction of evaluation anchors relies on metadata extraction patterns and on the stability of sentence identifiers across released resources, which implies that parts of the evaluation set are defined through recoverable documentation conventions rather than through an independently curated gold standard. Third, the language-grouped evaluation highlights that variation across languages and compilation histories is not a marginal effect but a central property of the setting, which limits how far aggregate scores can serve as a single proxy for downstream suitability.

These constraints do not invalidate sentence-level genre as a resource, but they motivate a layered release strategy: predicted labels are most useful when accompanied by provenance, coverage diagnostics, and joinable identifiers that allow consumers to calibrate their own filtering and aggregation choices. The next section, therefore, focuses on the release design and the infrastructural mechanisms by which we make the genre layer reproducible across UD revisions.

5. Outlook: Genre-Aware Evaluation, Adaptation and Reduced Genres

A practical motivation for sentence-level genre annotation is the observation that “high accuracy” pre-processing results are often conditional on evaluation regimes that privilege edited text. In a frequently cited German case study, POS tagging accuracy on Web material drops below the levels reported on standard newswire benchmarks, and the remaining performance varies strongly across Web subgenres (Giesbrecht and Evert, 2009). More recent multi-genre studies in English suggest analogous patterns in social and forum-like text, and also indicate that small amounts of in-domain data can outperform substantially larger amounts of out-of-domain Web data, while ensembles over genre-specialised taggers provide an additional robustness gain (Behzad and Zeldes, 2020; Owoputi et al., 2013b). Taken together, these results suggest that sentence-level genre, when made available as a joinable layer, can support a more transparent separation between (i) model quality under edited conditions and (ii) model robustness under heterogeneous Web and speech-related conditions, without conflating this distinction with language effects.

In this context, two complementary application modes appear plausible. First, genre can be treated as an evaluation variable, enabling stratified reporting for POS tagging, morphological tagging, and parsing, and thereby turning cross-genre variance into an observable quantity rather than an implicit error term. Second, genre can be treated as a selection variable for training, where genre-stratified sampling and targeted proxy selection have already shown benefits in dependency parsing (Müller-Eberstein et al., 2021a; Rehbein and Bildhauer, 2017) and where similar mechanisms have been repeatedly revisited for tagger adaptation in noisy or user-generated settings (Behzad and Zeldes, 2020).

A remaining tension is that the UD genre inventory is relatively fine-grained, partly overlapping, and not always aligned with the text-type distinctions that are commonly used in robustness dis-

cussions. One pragmatic way forward, which we treat as optional rather than as a replacement, is to provide a *reduced* genre layer on top of the full UD inventory. Such a reduction can be designed to align with recurring difficulty profiles, for instance, by separating transcribed speech (*spoken*), user-generated and interactional material (*social*, *email*, and related Web genres), broadly edited Web prose (often approximated by *wiki*), and newswire-like text (*news*). This direction is compatible with existing community efforts to improve cross-treebank genre coherence, such as UD-MULTIGENRE, which proposes a revised set of genres and verified instance-level labels over a UD-derived subset (Danilova and Stymne, 2023).

Finally, sentence-level genre information also connects to more general infrastructure questions around model and data preparation. Domain- and genre-sensitive pretraining has been shown to yield improvements over generic pretraining when the target domain is known (Gururangan et al., 2020). Similarly, tokenizer behaviour is sensitive to the training data and design choices, and recent work suggests that adapting tokenizer training data can affect downstream efficiency and performance (Dagan et al., 2024). In this sense, sentence-level genre labels can be understood as a low-cost coordination layer that enables more explicit experimentation with data composition, rather than as an attempt to impose a single notion of genre on UD.

6. Bibliographical References

- Shabnam Behzad and Amir Zeldes. 2020. [A Cross-Genre Ensemble Approach to Robust Reddit Part of Speech Tagging](#). In *Proceedings of the 12th Web as Corpus Workshop*, pages 50–56, Marseille, France. European Language Resources Association.
- Gautier Dagan, Gabriel Synnaeve, and Baptiste Roziere. 2024. [Getting the most out of your tokenizer for pre-training and domain adaptation](#). In *Proceedings of the 41st International Conference on Machine Learning*, pages 9784–9805. PMLR.
- Vera Danilova and Sara Stymne. 2023. [UD-MULTIGENRE – a UD-Based Dataset Enriched with Instance-Level Genre Annotations](#). In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 253–267, Singapore. Association for Computational Linguistics.
- Eugenie Giesbrecht and Stefan Evert. 2009. Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. In *Proceedings of the Fifth Web as Corpus Workshop (WAC5)*, pages 27–35, Donostia-San Sebastian, ES.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. [Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2021a. [Genre as Weak Supervision for Cross-lingual Dependency Parsing](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4786–4802, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2021b. [How Universal is Genre in Universal Dependencies?](#) In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 69–85, Sofia, Bulgaria. Association for Computational Linguistics.
- 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 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association (ELRA).
- Olutobi Owoputi, Brendan O’Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013a. [Improved part-of-speech tagging for online conversational text with word clusters](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia. Association for Computational Linguistics.
- Olutobi Owoputi, Brendan O’Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013b. [Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia. Association for Computational Linguistics.

Ines Rehbein and Felix Bildhauer. 2017. [Data point selection for genre-aware parsing](#). In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 95–105, Prague, Czech Republic.

7. Source Code Documentation References

Max Müller-Eberstein. 2021. [ud-genre: Code and artefacts accompanying “How Universal is Genre in Universal Dependencies?”](#). #5ab79a3, Accessed 2026-03-28.

Egon W. Stemle. 2026a. [commul/ud_genre: Sentence-level genre layer for universal dependencies as a separate, joinable dataset](#). Accessed 2026-03-28.

Egon W. Stemle. 2026b. [commul/universal_dependencies: Versioned universal dependencies snapshots in parquet format with release branches and regeneration tooling. v2.2.0](#), Accessed 2026-03-28.

Egon W. Stemle. 2026c. [ud-genre-bootstrap: Command-line interface \(coverage diagnostics, evaluation modes\)](#). Accessed 2026-03-28.

Egon W. Stemle. 2026d. [ud-genre-bootstrap: Genre extraction patterns \(documentation\)](#). Accessed 2026-03-28.

Egon W. Stemle. 2026e. [ud-hf-parquet-tools: Development toolkit for generating and validating universal dependencies datasets in parquet format. v1.2.3](#), Accessed 2026-03-28.

Universal Dependencies contributors. [Universal dependencies: Genres](#) [online]. 2026. Accessed 2026-03-28.

Towards a Universal Dependency Corpus for Old Saxon (Old Low German)

Christian Chiacos^a, Janine Siewert^{a,b}

^aApplied Computational Linguistics (ACoLi), University of Augsburg, Germany

^bDepartment of Digital Humanities, University of Helsinki, Finland
christian.chiacos@uni-a.de, janine.siewert@helsinki.fi

Abstract

Among the West Germanic languages of the first millennium C.E. (Old English, Old Low Franconian/Old Dutch, Old High German, and – although much later – Old Frisian), Old Saxon occupies a special role both linguistically – in that it represents a middle ground in the dialect continuum between Old English at one extreme and Old High German on the other –, and in terms of material quality, in that it is attested with considerable amounts of coherent text (unlike Old Low Franconian) which is not only particularly old (unlike, especially, Old Frisian), but also original (i.e., not translated, a rarity in the attested Old English and Old High German material). It is thus a language central to the understanding of the emergence of several modern major languages, incl. English, Dutch and German, and has been studied intensely, albeit – so far – not in the context of the Universal Dependencies. This paper addresses this gap and describes the introduction of (a) a manually annotated test corpus of Old Saxon, (b) a highly reusable conversion pipeline for converting the Penn bracketing syntax of the Penn Historical Corpora (and the Old Saxon Heliand) to UD, and (c) the evaluation of the latter against the manual annotations.

Keywords: Universal Dependencies, Old Saxon (Old Low German), Penn bracketing syntax, UD conversion and evaluation

1. Introduction

The Old Saxon (Old Low German) language has been the historical language of the Saxons, a Germanic tribe settling primarily in North-Western Germany in the first millennium C.E., neighboring Scandinavians (the linguistic predecessors of Danish) to the north, Frisians to the north-west, the Frankish empire (the linguistic predecessors of modern Dutch and German) to the south and west and Slavs to the east. Linguistically, the languages of the Saxons, Frisians and Franks formed a dialect continuum that extended further onto the British Isles whose early Germanic inhabitants settlers also adopted a self-designation as Saxons when they began their emigration after the Romans gave up the province of Britannia in 410 C.E. In a series of intense wars, the continental Saxons were conquered by the Franks in the 780s, and part of their political and social integration into the Frankish empire included their Christianization, giving rise to the Old Saxon literature, primarily attested in the *Heliand*, a 9th c. gospel harmony, the *Genesis*, a free, poetic retelling of the first book of the Bible, and in a number of smaller, fragmentary texts and glosses.

In this paper, we describe the creation of a new corpus of Old Saxon, syntactically annotated in accordance with Universal Dependency (UD) guidelines. For the most part, the corpus is, however, not manually annotated, but *converted*, or, more precisely, compiled from existing corpora with mor-

phosyntactic and/or syntactic annotation. These annotations, applied to (different editions of) the same source text, were however, conducted in accordance with different linguistic traditions (primarily interlinear glossed text, or tier-based annotations, and phrase structure syntax), so that substantial efforts in into their harmonization and consolidation are required in order to provide a CoNLL-U representation that encompasses all sources of information. We thus refer to our corpus as *Consolidated Old Saxon* (ConOS) corpus. Its current release¹ consists of UD-compliant data automatically compiled from the respective source corpora and the accompanying conversion scripts, encompasses a total of 109,000 tokens, which, however, represent two *versions* of the same text. Beyond that, we provide UD ConOS, a manually annotated validation set of 2,000 tokens, published as part of the Universal Dependencies.² In addition to the first fit (chapter) of the *Heliand* text, it also provides annotations of a fragment of the Old Saxon *Genesis*.

At the core of the ConOS corpus is the *Heliand*. As the most extensive Old Saxon text, it occupies a central position in early West Germanic philology and its linguistic characteristics are crucial for reconstructing early syntactic change (Petrova and Solf, 2009; Lühr, 2025). Reflecting its importance,

¹Available at <https://github.com/nds-spraakverarbeiten/ConOS>

²Available at https://github.com/UniversalDependencies/UD_Old_Saxon-ConOS

Ms. M:	that	inatorht	lico	tidi	gimanodun
Ms. C:	that	ina	torohtlico	tidi	gimanodun
	that	him	brilliant	times	remembered
	'that he was remembered of great times'				

Table 1: Differences in word segmentation

it has been annotated repeatedly in independent projects: the HeliPaD treebank (Walkden, 2016), following Penn-style constituency annotation; the Heliand DDD corpus (Referenzkorpus “Altdeutsch”) with tier-based morphosyntactic and clausal annotation; and the Heliand B4 corpus (Linde, 2009), developed for diachronic information-structural research. These corpora differ in granularity. HeliPaD provides full phrase-structure parses using the Penn bracketing notation; DDD and B4 offer partial, non-recursive annotations (POS, clause boundaries, and – B4 only – nominal and prepositional phrases). The consolidation of their annotations and the creation of the ConOS corpus is primarily described in Chiarcos and Siewert (2026). The current paper focuses on the creation of the manually annotated UD ConOS corpus which represents the gold data used for evaluating the conversion and consolidation task, and which is released as a UD data set.

2. Old Saxon: Texts and Corpora

Old Saxon (OS) is attested in two major texts, Heliand and Genesis, which are both alliterative rhyme poems dated back to the 9th century. In addition, there is a small number of liturgical texts and tax lists, but limited in size, so that research on Old Saxon primarily focuses on Heliand and Genesis. With about 6,000 long verses, the Heliand is the more extensive, and so far, has remained the primary basis for corpus-based analyses. The OS Heliand was handed down in two major manuscripts, C (Ms. Cotton. Calig. A. VII, London, British Library) and M (Cgm 25, München, Bayerische Staatsbibliothek) as well as in four fragments. The text is divided into 72 chapters (‘fits’) indicated in the manuscripts by initial capital letters (Ms. M and C) and Roman numbers (Ms. C only). A major difference between the two manuscripts is in their word segmentation, as exemplified in Table 1. This will require special consideration when aligning the corpora in future extensions of our work.

The earliest syntactically annotated corpus of Heliand (Linde, 2009, Heliand-B4) was developed along with digital editions for a number of Old High German texts (Petrova et al., 2009). With only 3,500 tokens, it covered less than 10% of the text, but the corpus, in particular in the design of the corpora developed here inspired the subsequent DDD project: The Old German Reference Corpus (Referenzkorpus Altdeutsch, DDD) (Linde and

Mittmann, 2013; Dipper et al., 2013) aimed to digitize and annotate all extant Old German texts, including the Old Saxon *Heliand*, the *Genesis* and a number of smaller texts (primarily tax registers and short religious texts) by combining automated pre-annotation from historical glossaries with manual refinement. For Heliand and Genesis, the glossary by Sehrt (1925) provides extensive lemma, morphological, and attestational information that was digitized by the project and linked with the text to derive morphosyntactic annotations in an automated fashion (Linde and Mittmann, 2013). These have been manually refined and extended using ELAN, a tool for multi-tier annotation in the tradition of interlinear glossed text (Bow et al., 2003, IGT), with detailed morphological annotation and labels for clause linkage, but without any annotation of phrase structure or grammatical relations within the clause. Figure 1 illustrates DDD annotations for the first verse of Heliand, visualized with ANNIS.³

The first corpus to introduce full syntactic annotations for Old Saxon was the *HeliPaD* (Heliand Parsed Database) (Walkden, 2016), annotated according to the standards of the Penn Corpora of Historical English (Taylor et al., 2003), employing explicit constituency structures, grammatical function labels, empty categories, and indices to represent long-distance dependencies. HeliDaD comprises 46,000 tokens, based on an older, authoritative edition (Sievers, 1878). The corpus is designed for the comparative study of syntax, in particular among the Penn Treebank family of parsed corpora. HeliPaD is the only corpus of Old Saxon to offer explicit constituency structures with recursive phrase structures and grammatical relations. For the (partial syntactic analysis of the) first line of the Heliand, Fig. 2 shows peculiarities of this format: Phrases are marked by Lisp-style brackets, with the first element representing the category of the phrase, optionally followed by grammatical role annotations, and the following elements representing the respective child nodes. Words (tokens) are terminal nodes in this tree, with part-of-speech (XPOS) annotations as their phrase label and represented as a concatenation of word form and lemma, separated by –. The annotation also uses a considerable amount of empty elements, including traces (marked with *) and non-textual elements (marked with CODE). The category and relation inventory and the XPOS annotations are based on the Penn Treebank (Taylor et al., 2003), albeit with extensions for historical languages. This includes the annotation of morphosyntactic features attached to the part of speech and separated by ^, e.g., N^N^SG for a noun (N) in

³https://korpling.german.hu-berlin.de/annis/ddd#_q=dG9rCg&ql=aql_c=RERELUFELUhlbGlbhbmRfMS4y&cl=5&cr=5&s=0&l=10&_seg=ZWRpdGlvbGx

* Manega uuáron , the sia iro môd gespôn

□ annotations

edition	*	Manega	uuáron	,	the	sia	iro	môd	gespôn
text	*	Manega	uuáron	,	the	sia	iro	môd	gespôn
lemma		manag	wesan		the	he	he	mod	gispanan
lang		osx	osx		osx	osx	osx	osx	osx
posLemma		DI	VA	\$,	DD	PPER	PPER	NA	VV
pos		DIS	VVFIN	\$,	DDSREL	PPER	DPOS	NA	VVFIN
inflectionClassLemma		A,O	ST5					A_MASC	ST6
inflectionClass		P	ST5					A_MASC	ST6
inflection		MASC_PL_NOM_ST	IND_PAST_PL_3		MASC_PL_ACC	MASC_PL_ACC_3	MASC_PL_GEN_3	SG_NOM	IND_PAST_SG_3
clause		CF_U_M			CF_I_Rel				
document		Heliand							
chapter		I							
verse		1							
rhyme		S1						S2	
translation		mancher	sein; dasein, existieren		der, die, das, welcher, welche, welches	er, sie, es	er, sie, es	Sinn, Inneres, Herz	antreiben

Figure 1: Heliand DDD annotation, first verse, visualized with ANNIS4

```
( (IP-MAT
  (CODE <P_7>)
  ...
  (NP-SBJ *exp*)
  (NP-PRD (Q^N^PL Manega-manag)
    (CP-REL *ICH*-1))
  (BEDI^3^PL uuaron-wesan)
  (CODE <C>)
  (CP-REL-1
    (WNP-SBJ-2 0)
    (C the-the)
    (IP-SUB (NP-SBJ-RSP-2 (PRO^A^3^PL sia-he))
      (NP-OB1 (PRO$^N^3^SG iro-his)
        (N^N^SG mod-mod))
      (GE+VBDI^3^SG gespon-spanan)
      (, , - ,)
      ... ))))
```

Figure 2: HeliPaD annotation, first verse, PTB bracketing format

nominative (^N) singular (^SG). A number of similar extensions exist, e.g., for historical English (Pintzuk and Taylor, 1997), historical German Light (2012); Sapp et al. (2024), Yiddish (Santorini, 1993) and Icelandic (Rögnvaldsson et al., 2011), but it is to be noted that neither of these are fully identical with each other, although the Old Saxon schema follows the model of the Old English corpora.

In the following sections, we describe the creation of a manually annotated gold standard for Old Saxon UD annotations, the conversion of HeliPaD to UD and the evaluation of this conversion.

3. Building the UD ConOS corpus

We manually created a UD-compliant test corpus for Old Saxon from the DDD corpus, that consists of two parts: (1) Heliand fit (chapter) 1, based on the text of Behaghel and Taeger (1984, further BT), and (2) Genesis fragment 2, BT edition.

Heliand is structured into a total of 72 ‘fits’, or chapters. UD ConOS provides a manual annota-

tion of the first fit, whose main witness is Ms. C (it is largely missing in Ms. M). For this manual annotation, we rely on an existing morphosyntactic annotation of the Heliand provided by the DDD corpus. This has been conducted independently from the HeliPaD, and is based on the alignment of the glossary of Sehart (1925) that has been aligned with the source text of the BT edition. Although this text and the HeliPaD ultimately go back to the same manuscript, it is to be noted that the BT text and the HeliPaD text are not identical, but that the BT text is heavily amended, the orthography of all source manuscripts differ from each other and their normalization in BT, and modernizing punctuation of both texts has been created independently for both editions. The Heliand part was based on the B4 corpus, aligned with the DDD annotations for morphosyntax, lemmatization, German glosses, see Chiarcos and Siewert (2026).

The second major text attesting the Old Saxon language is the Old Saxon Genesis, a fragmentarily attested retelling of the first book of the Genesis, presumably written by the Heliand author and surviving in three Old Saxon fragments, and an Old English translation (or, as Old English and Old Saxon are often seen as mere dialects – transliteration).⁴ With 1,145 tokens, the second of these fragments is the largest consecutive piece of text of Old Saxon outside the Heliand. The Genesis part of UD ConOS was directly taken from the DDD corpus.

We ground the UD ConOS annotations in the morphosyntactic annotations of the DDD corpus,

⁴Although the Old Saxon and the Old English fragments of the Genesis do not overlap, the Old Saxon source of the Old English text is indicated by a number of spelling errors, where Old Saxon forms are used in place of expected Old English forms.

which provides Heliand and Genesis along with all other major text fragments of the Old Saxon language, with annotations for morphosyntax, agreement features, clausal juncture, and German glosses.

For creating morphosyntactic annotations, we first converted the DDD corpus from its native ELAN format to a CoNLL-U representation: Using a generic ELAN converter, we identified the minimal segments as tokens and then transformed every tier to an independent column. With standard GNU tools (`cut`, `grep` and `sed`), these were transformed to (the `FORM`, `LEMMA`, `XPOS`) columns of CoNLL-U as well as to the `Trans` (translation) property we put into the `MISC` column to keep track of the German translations of the Old Saxon lemmas. For HeliPaD and B4 annotations, DDD annotations were automatically aligned with both corpora, see Chiarcos and Siewert (2026).

Actual annotation was performed with off-the-shelf spreadsheet software (LibreOffice), where we used Excel-style formulas to insert `IDS` and conditional formatting over the columns `ID` and `HEAD` to visualize attachment: from a color scheme ranging from green over yellow to red, the same color indicates the same number, as illustrated in Fig. 7 for the sentence *Sîdoda im thuo te selidon, habda im sundea giuuarahit bittra an is bruodâr*; ‘He now went home, (after) he had done bitter sin against his brother’ (translation by authors, see Appendix).

For the DDD tagset (Dipper et al., 2013), no UD mapping was known to exist at the time and was thus provided by the authors. For manual annotation, we used the standard filter function of modern spreadsheets to perform a manual mapping from DDD `XPOS` to `UPOS`, by filtering every (group of) UD-equivalent DDD tags and replacing them all together with a `UPOS` tag. It is to be noted that some choices seem to have been rather arbitrary. As an example, clause-initial *sô* is annotated as either complementizer (and thus `SCONJ`, as in Example 1) or adverb (and thus `ADV`, as in Example 2). Apparently, this is based on the German translation as *daß* ‘that (complementizer)’ or *so* ‘so’, but often without concrete contextual triggers that justify this differentiation.

- (1) *Sô he thô thana uuîrôc*
SCONJ **PRON** **ADV** **DET** **NOUN**
 so/that he now the-ACC.SG incense
drôg
VERB
 carry.PST.3SG

‘As he took the incense’ (translation: Scott, 1966) (DDD Heliand)

- (2) *Sô he ina thô gehungrean*
ADV **PRON** **PRON** **ADV** **VERB**
 so/that he he.ACC now starve
lêt
VERB
 let.PST.3SG

‘As He let Himself hunger’ (translation: Scott, 1966) (DDD Heliand)

For UD ConOS, we respected the existing DDD annotations, with `SCONJ` entailing a `mark` dependency and a `acl`, `advcl`, `ccomp` or `xcomp` head, and `ADV` entailing a `advmod` dependency without any implications for the dependency label of its head. Only in two cases, we observed (and corrected) apparent errors in the morphosyntactic annotation of DDD, both of which seem to arise from homonymy. As an example, *lêdo* ‘suffering’ (German *Leid*) was incorrectly annotated as adjective in the sentence in (3). Indeed, using the adjective *lêd* ‘disgusting’ (German *leid*) in place of the homonymous noun appears to be grammatically possible, but would be contextually dispreferred: the sentence continues with ‘(the suffering) that the Antichrist throws all people into ruin’, with the word *lêdo* acting as matrix *noun* for the relative clause.

- (3) *Than hier ôk thie lêdo*
ADV **ADV** **ADV** **DET** **NOUN**
 then here also the-NOM.SG suffering
kumit
VERB
 come.PRS3SG

‘then here comes the suffering, too’ (translation: authors) (DDD Genesis)

The filtering function was also used to annotate punctuations automatically (either modern punctuation inserted into the historical text or symbols indicating additional spaces or line breaks), in that these were filtered out, with `UPOS` and `DEP` set to `PUNCT` and `punct`, resp., and `HEAD` set to the `ID` of the preceding row (i.e., the preceding word).

Annotation was performed left-to-right (or, in the software, top-to-bottom) by first reading the full sentence and its German glosses, and then annotating the syntactic head of the word under consideration (if the head precedes the word) and its dependency label. If the word points to a head further down, it is not annotated yet. Instead, as soon as the head is annotated, we go back in the `HEAD` column and fill up the annotations of the preceding dependents with the `ID` of the head.

4. From Penn Bracketing Syntax to Universal Dependencies

We describe the conversion of Penn Bracketing annotations of the HeliPaD corpus to Universal Dependencies as this is evaluated against UD ConOS annotations in this paper. For the transformation of and subsequent consolidation of B4 and DDD annotations with the technologies described here see [Chiarcos and Siewert \(2026\)](#).

4.1. Technological prerequisites

Previous tools for converting the Penn bracketing syntax to UD have largely been ad-hoc solutions focusing exclusively on one particular resource or type of annotation, and because none of these clearly separated the conversion task from the mapping logic, every attempt at converting an unseen dataset had to be re-implemented from scratch, see [Arnardóttir et al. \(2020\)](#) for an overview (and yet another, corpus-specific solution).

In our approach, we aim to provide a more reusable alternative and suggest the application of Fintan, the Flexible Integrated Annotation eNginneering platform ([Fäth et al., 2020](#)) that, on the one hand, allows to consume standard formats from NLP, corpus linguistics and computational lexicography, including all flavours of CoNLL-TSV (incl. CoNLL-U). These are then, sentence by sentence, transformed to RDF graphs, such that graph rewriting rules can be applied, using the W3C standard SPARQL, a language for querying and transforming RDF graphs. These rules represent the transformation logic and are stored in separate files. Fintan supports parallelization and iteration of such transformations to ensure fast and efficient processing, but as the transformation logic is *declarative*, i.e., decoupled from the conversion between source and target formats, these transformations can be executed in any technical environment that supports RDF data, say, off-the-shelf databases such as Apache Fuseki⁵ or Neo4J,⁶ in-memory solutions such as TARQL⁷ or programming libraries such as RDFLib for Python⁸ or Apache Jena for Java and C++.⁹

With SPARQL 1.1, every transformation is a sequence of retrieval (WHERE) and update (INSERT/DELETE) operations, with a number of optimizations for matching complex graph patterns. In particular, this includes SPARQL property paths,

i.e., the possibility to perform transitive search (repeatedly following edges of the same kind), inversion (follow an edge in the opposite direction), sequences (follow one or more edges of one type, then another), or disjunctions (follow one kind of edge or another).

Moreover, because transformation are formulated in a series of relatively fine-grained operations, these operations can be re-used in constellations in which similar data structures are to be modified, say, for the conversion from trees to dependencies, regardless of whether the source data was provided in a CoNLL IOB notation, in Penn bracketing syntax, an XML tree or an array of nested JSON dictionaries. All of these are supported by Fintan and input and output formats, but here, we specifically use the CoNLL-RDF customization of Fintan ([Chiarcos and Fäth, 2017](#)) for reading and writing CoNLL (and CoNLL-U) data, and the CoNLL-RDF vocabulary ([Chiarcos et al., 2021](#)), as summarized in Fig. 3. CoNLL(-TSV) formats represent sentences as blocks of tab-separated rows, one token per line, optionally preceded by comment metadata. Columns encode annotations such as form, lemma, POS, morphological features, dependency relations or task-specific labels. CoNLL-RDF maps each token to a `nif:Word` node, linked via `nif:nextWord`, with column values represented as literal properties in the `conll:` namespace. Some columns receive dedicated treatment: `HEAD`, for example, is resolved into explicit `conll:HEAD` relations that link tokens (`nif:Words`) with each other (for syntactic dependencies) or the sentence (for roots). The CoNLL-RDF tree extensions ([Chiarcos and Glaser, 2020](#)) extend the basic CoNLL-RDF model with the POWLA vocabulary to represent hierarchical structures such as phrase-structure trees. This extension allows Penn Treebank-style parses to be encoded as additional nodes linked via `powla:hasParent` and `powla:next`. Figure 4 shows a fragment in which a token is connected to a phrase node, which in turn participates in a larger tree. Thus, any kind of annotation that can be represented in a TSV format can be represented in a unified RDF graph, which can then be easily traversed and manipulated with SPARQL, even if the data includes multiple layers of diverse styles of syntax annotation, be it phrase structure trees or dependency syntax.

4.2. HeliPaD conversion

We first converted the native HeliPaD format into the conventional CoNLL representation of the Penn bracketing format, illustrated in Fig. 5, with four columns that we label `WORD`, `POS`, `LEMMA` and `PARSE`. Using the CoNLL-RDF tree extensions, this was parsed into the structure illustrated in Fig. 6

Figure 6 illustrates the conversion process for

⁵<https://jena.apache.org/documentation/fuseki2/>

⁶<https://neo4j.com/blog/developer/rdf-lib-neo4j-rdf-integration-neo4j/>

⁷<https://tarql.github.io/>

⁸<https://rdflib.readthedocs.io/en/stable/>

⁹<https://jena.apache.org>

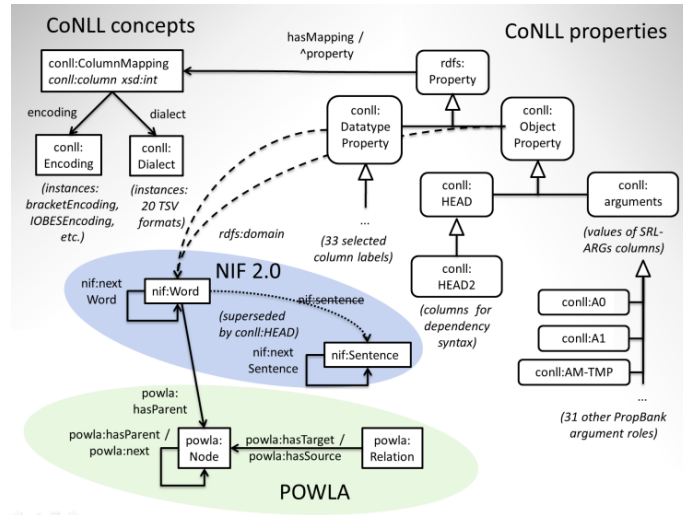


Figure 3: CoNLL-RDF Data Model (Chiarcos et al., 2021)

```
# CoNLL-RDF conversion with CoNLL-
RDF tree extension
:s1_0 a nif:Sentence .
:s1_7 a nif:Word ;
      conll:HEAD :s1_0 ;
      conll:WORD "Manega" .
      conll:POS "Q^N^PL" ;
      conll:LEMMA "manag" ;
      nif:nextWord :s1_8 ;
      powla:hasParent :bPARSE_4 .
:bPARSE_4 a conll:PARSE ;
          conll:CAT "NP" ;
          conll:ROLE "PRD" ;
          powla:hasParent :bPARSE_2 ;
          powla:next :bPARSE_5 .
```

Figure 4: CoNLL-RDF fragment for the first word of HeliPaD, with tree extensions

```
<P_7> CODE — ((IP-MAT *
...
*exp* — (NP-SBJ *)
Manega Q^N^PL manag (NP-PRD *)
*ICH*-1 — (CP-REL *)
uuaron BEDI^3^PL wesam *
<C> CODE — *
0 — (CP-REL-1 (WNP-SBJ-2 *)
the C the *
sia PRO^A^3^PL he (IP-SUB (NP-SBJ-RSP-2 *)
iro PRO$^N^3^SG his (NP-OB1 *)
mod N^N^SG mod *)
gespon GE+VBDI^3^SG spanan *
' ' '
...
```

Figure 5: Initial CoNLL conversion of HeliPaD

the first verse of HeliPaD. Above is a (slightly simplified) visualization of the CoNLL-RDF input graph of words (green, with values of `conll:WORD`, `conll:POS`, and `conll:LEMMA` properties), phrases (yellow, with values of `conll:CAT` and `ROLE` and the sentence (blue), connected by `nif:nextWord` ($\rightarrow$), `powla:hasParent` ($\uparrow$), `powla:next` (between siblings in the parse tree, hidden) and `conll:HEAD` (between every word and the sentence, hidden).

The lower part of Fig. 6 shows results of the graph rewriting operations with a number of tokens dropped (empty tokens, illustrated by gaps) and additional properties (`conll:UPOS` and `conll:EDGE`) derived for the others. In the lower part, the `nif:nextWord` property is adjusted to connect adjacent tokens, but skipped in the diagram to facilitate readability. Further, the `conll:HEAD` property of all tokens has been adjusted to point to their respective syntactic head.

Note that even after rewriting, the original POWLA data structures remain in the RDF graph. However, the CoNLL-RDF/Fintan writer is asked to produce a table of (`conll:`)ID (automatically generated from the number of preceding `nif:nextWord` properties for every token), `conll:WORD` (UD FORM), `conll:LEMMA`, `UPOS`, `conll:POS` (UD XPOS), `conll:FEATS` (not available, thus filled up with `_`), (`conll:`)HEAD (automatically generated from the ID of the `conll:HEAD`), `conll:EDGE` (UD DEP) and two empty columns, with words ordered according to the `nif:nextWord` property. Any other property, be it from POWLA, CoNLL-RDF or user-defined name spaces, is quietly lost, and the result effectively corresponds to CoNLL-U.

As the figure illustrates, complicated rewriting rules are required, which can be conveniently implemented in SPARQL. As a first step, we simplify the label inventories (e.g., we infer `conll:CAT IP` for every element with `conll:CAT IPX` – and keep both values), and we detach `nif:Words` that represent punctuation and `CODE` elements. Then, we identify the location of the syntactic head (`temp:HEAD`) for every phrase type, using a total of 12 rules implemented by 21 SPARQL update operations, e.g.,

1. for a non-branching `powla:Node`, the head is

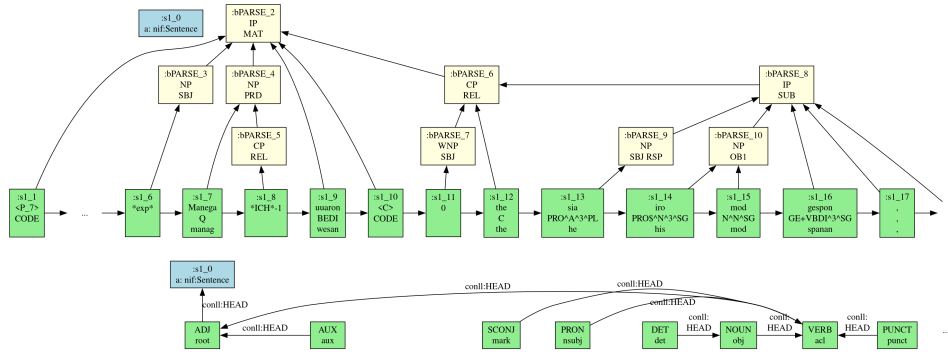


Figure 6: HeliPaD transformation (CoNLL-RDF: PTB-based input graph and derived UD data structures)

- the child
- the head of a node with `conll:CAT CONJP` is the first conjunct, the `conll:CAT` of the head is copied to the parent
- the head of a PP node is the first child with `conll:CAT NP`
- the head of any PP node without head annotation is the first child not carrying `conll:POS P`, etc.

Unfortunately, the relatively large number of rules is unavoidable due to the idiosyncrasies of both PTB- and UD-styles of syntactic annotation. A number of additional SPARQL updates are used in preprocessing to simplify the label inventories.

The listing below illustrates an example for the SPARQL updates employed, where we identify the head-carrying child for every CP node as its first child node with `conll:CAT IP`:

```
INSERT {
  ?node temp:HEAD ?child;          # set a value for
    rdfs:comment "(g)"             # temp:HEAD
} WHERE {
  ?node conll:CAT "CP".             # with rule (g)
  MINUS { ?node temp:HEAD [] }      # as follows:
  ?child powla:hasParent ?node;     # for every CP node
    conll:CAT "IP".                 # without HEAD
  MINUS {
    [ conll:CAT "IP" ]              # use a child node
    powla:next+ ?child }           # with category IP
  };                                # if there is no
                                   # IP node that
                                   # precedes it
};
```

As the comments underline, this directly translates to human-readable instructions (and vice versa). Also note that in addition to the transformation, we also document all applied operations by adding a `rdfs:comment` that identifies the responsible rule to facilitate subsequent debugging.

As a result of these rules (and a final fall-back rule to set the `temp:HEAD` to the first child node that contains at least one non-empty token), every `powla:Node` contains exactly one `temp:HEAD`. The listing below contains the SPARQL Update that uses the transitive closures of the `temp:HEAD` property (`temp:HEAD*` and `temp:HEAD+`) to derive the `conll:HEAD` property that directly connects individual words rather than phrases:

```
INSERT {
  ?w conll:HEAD ?head.
} WHERE {
  ?head a nif:Word; (^temp:HEAD)+ ?phrase.
  ?child powla:hasParent ?phrase.
  MINUS { ?phrase temp:HEAD ?child }
  ?child terms:HEAD* ?w.
  ?w a nif:Word.
};
```

For every `?phrase`, this update defines the `?head` as a `nif:Word` on the transitive closure of `temp:HEAD` (`(temp:HEAD)+`). For every `?child` node of `?phrase` that is not (`MINUS`) the `temp:HEAD` of `?phrase`, retrieve its syntactic head `?w` (a `nif:Word` identified along the `temp:HEAD*` axis) and set the `conll:HEAD` of `?w` to `?head`. Whereas `temp:HEAD` connected phrasal nodes with their dependents, `conll:HEAD` now is a property pointing from one word to another.

Similarly (but in a separate update), `?w` is assigned *multiple* values for `conll:EDGE`: `conll:ROLE` of `?dependent`, `conll:CAT` of `?dependent` and the concatenation of `conll:CAT` (or `conll:POS`) of `?dependent` and `conll:CAT` of `?phrase`. These are later mapped to UD dependency labels.

After the establishment of `conll:HEAD` properties, i.e., (unlabelled) dependency relations, we now attach any detached word (code or punctuation) to the head of the minimal spanning relation, resp., the overall syntactic root. Then, we perform a rule-based mapping for parts of speech and `conll:EDGE` annotations to UD standards, based on a manual inspection of existing `conll:POS` and `conll:EDGE` values.

Finally, we remove empty tokens, induce the new `conll:ID` values by counting the number of preceding `nif:nextWord` properties for every word and write a TSV-separated table with the properties that represent the equivalents of the standard CoNLL-U columns. Fintan serializes this table based on the `nif:nextWord` sequence, and except for some greater level of flexibility in the CoNLL comments that every sentence comes with (and that also preserves all `rdfs:comment` properties attached to the `nif:Sentence`), the result corre-

sponds to the CoNLL-U format. The next sentence is then serialized after an empty line, etc.

The CoNLL-RDF module of Fintan provides a number of convenient visualizations that were helpful in developing this converter. This includes a direct export of ‘canonically formatted’ CoNLL-RDF, where words are presented in their original order and in a single line, similar to the common JSON-L format, but as valid Turtle/RDF (or, optionally, with syntax highlighting). This also includes a GraphViz export of results and intermediate steps, as well as an ASCII visualization of dependency trees.

5. Evaluation

For evaluating the HeliPaD conversion, we operate with the Heliand subset of our gold data, only. The Genesis subset is reserved for future evaluations of parsers against unseed data. A problem is, however, that tokenization and spelling of the DDD Heliand differs from that of the HeliPaD. This is compensated by CoNLL-Merge (Chiarcos and Schenk, 2018) that can be used to align the text of different witnesses, versions or editions in CoNLL (TSV one-word-per-line) formats. CoNLL-Merge provides several strategies for the automated resolution of tokenization, encoding or spelling conflicts (keep both, force source tokenization onto target annotations, split into maximum common segments, align by Levenshtein distance), as well as for merging their annotations by putting them into additional columns, following the annotations of the base text (i.e., our gold annotation). We use the default, so that tokens are aligned either because of string identity or positional correspondence. Otherwise, both alignments are kept, but we use GNU tools (`grep`) to filter out mismatches and evaluate alignable words, only. The merged CoNLL file then consists of 20 columns, but as the tokenization (and thus, the IDs) of both annotations differ, the HeliPaD ID and HEAD annotations are first replaced by values from the gold annotation.

With only about 2000 tokens overall, both fragments are relatively small, but nevertheless allow us to evaluate along three different dimensions, i.e., (1) the quality of the transformation rules (over Heliand fit 1), (2) the quality of the HeliPaD parser to replicate HeliPaD-UD annotations (over Heliand fit 1), and, (3) the quality of the HeliPaD parser over unseen text (over Genesis fragment 2). Whereas (2) and (3) are left as topics for future research, the UD ConOS data allows for some preliminary insights into structural differences between (the CoNLL-U edition of) HeliPaD and manual ConOS annotations. As the textual basis of HeliPaD and DDD Heliand is different, only 678 tokens remain for evaluation, with an UPOS accuracy of 96.9% (657/678). Recurring errors include

confusions between DDD PRON and HeliPaD DET (4, in two of these cases, the HeliPaD combination of UPOS and dependency was invalid) and ADJ (3, for *manag* ‘many’ and *gihwilik* ‘every’), DDD VERB and HeliPaD AUX (3, for *hebbian* ‘to have, to own’), and DDD VERB and HeliPaD ADJ (3, for participles). We achieve an unlabelled attachment score (UAS) of 76.8% (521/678), and a labelled attachment score (LAS) of 58.6% (397/678).

The UAS seems fair, the LAS numbers are rather low – but also, achieved over a very small data set. In the future, we would like to provide additional fits (Heliand chapters) to substantiate our findings and to assess a number of necessary corrections we observed during evaluation: Upon closer inspection we found that the LAS score may be due to different conceptualizations rather than linguistically different analyses. Checking 124 cases with correct attachment, but divergent labels, we found that 24.1% (30/124) were for DDD `nmod:gen` against HeliPaD `nmod`, and, analogously, 9.7% (12/124) for DDD `det:poss` against HeliPaD `det`. We found 13 cases of DDD `cc` against HeliPaD `mark` (10.5%, 13/124) which indicate a certain degree of uncertainty also found in the literature with regard to the identification of (subordinate) clauses and their functions, and, if you will, a different school of thought rather than a fundamental disagreement. It is to be noted, however, that these divergent analyses pertain to exactly three words: *endi* ‘and’, *eftha* ‘or’ and *ak* ‘but’. We would suggest to generally label them as `cc` based on their lemma. For 12 cases (9.7%), we identified an actual error of the HeliPaD conversion, where the traditional UD dependency `iobj` was used for dative arguments rather than the UD v.2 dependency `obl` currently recommended for related languages (esp., modern German). Eliminating these sources of error would lift the LAS to 68.4% (464/678), and, thus, to a more reasonable level.

The UD ConOS currently only comprises the manually annotated subset of the larger ConOS corpus. Both datasets are published under open licenses (see introduction), but the (UD conversion of the) HeliPaD corpus is currently not integrated with the UD ConOS corpus. Unless we can provide complete manual verification or raise the level of agreement between converted and manually annotated Old Saxon data to substantially higher levels, it remains to be published as a complementary, independent, data set, but cross-referenced with UD ConOS in their respective GitHub repositories and their documentation.

6. Results and Perspectives

The UD ConOS corpus is the first UD corpus for Old Saxon, which, although being small-scale, comes

with the prospect to integrate the full HeliPaD corpus, and the Old Saxon parts of DDD into UD. As these conversions have, however, not been manually verified, their UD conversions reside in the primary ConOS repository, separated from UD ConOS. We described the rule-based transformation of the Old Saxon HeliPaD corpus to UD and its evaluation against UD ConOS, with notable contributions on a technical level:

- the first *declarative*, tool-independent technique for mapping Penn Bracketing syntax schemas to UD,
- using Fintan and SPARQL to *disentangle* format conversion (Fintan/CoNLL-RDF) from handling linguistic data structures (SPARQL), facilitating the application of the same transformation logic against different backend technologies (say, an RDF database), and the re-usability of transformation operations for different formats (say, XML- or JSON-based formats for phrase structure syntax), and
- evaluation of dependency syntax annotations *across divergent tokenizations* by means of CoNLL-Merge.

Recent years have seen a considerable number of syntactically annotated corpora of older stages of continental West Germanic (Sapp et al., 2024; Dipper et al., 2024; Haiber, 2024), but, aside from the Walkden (2016) and Linde and Mittmann (2013), these focus on the 2nd millennium C.E. As for the older stages of Germanic, so far, only (Old) Icelandic, Gothic and Old English are covered by UD, but we are not aware of any data available for the corresponding historical stages of German (Old High German), Low German/Low Saxon (Old Saxon) and Dutch (Old Low Franconian). Both Icelandic and the original English UD corpora data are based on a conversion of Penn Treebank-style corpora (Arnardóttir et al., 2020),¹⁰ so that, technologically, converting yet another Penn Historical Treebank to UD is novel to a moderate extent, at best. Still, the use of SPARQL and Fintan/CoNLL-RDF technology may be noteworthy, and even more so in combination with CoNLL-Merge: The declarative rules do not only facilitate re-usability for similar conversion tasks, but also, both technologies in conjunction allow to merge multiple CoNLL(-U) files from different sources and to aggregate over them in order to produce richer, consolidated, and more consistent annotations. To some extent, this has been illustrated in previous research, already, but

¹⁰Initial UD data has actually been created from such data, using the StanfordNLP Universal Dependencies converter still available from <https://nlp.stanford.edu/software/stanford-dependencies.shtml>.

for very different tasks: Chiarcos and Fäth (2019) combined CoNLL-U files from the Universal Dependencies with Skel (anonymized CoNLL) files from the Universal Propositions on the basis of their (largely shared) parts-of-speech annotation, and then used SPARQL to decompose both types of annotation and to aggregate them into a complex, conjoint semantic-syntactic data structure as postulated by Role and Reference Grammar Van Valin (2014, RRG). Similarly, Chiarcos et al. (2022) described how Fintan/CoNLL-RDF and SPARQL can be used to query over syntax annotations from 12 corpora representing the entire body of syntactically annotated historical German available at the time and covering vastly diverse types of annotation, ranging from shallow, tier-based annotations (such as in DDD and Heliand-B4) over phrase structure annotations (such as in the Early New High German treebank of Light, 2012) to dependency syntax (Scheible et al., 2011; Salomoni, 2021).

For languages in which the same text has been annotated multiple times independently (as for Old Saxon) or is attested in different manuscripts (i.e., *everything* published before the advent of the printing press, as books had to be copied by hand and scribes left traces of their languages in the text, so that manuscripts diverged substantially over time), this can now be used to consolidate these texts and their annotations in a systematic, rule-based fashion. This is not to suggest that rule-based methods are to be preferred over machine learning, but that, say, postprocessing annotations produced by machine learning or other techniques can benefit from a principled, systematic and reproducible approach. For the specific case of the Heliand, this even seems necessary, as the existing annotations are complementary in the sense that each of the known corpora provides something unique (e.g., DDD provides more fine-grained clause linking than HeliPaD) and they complement each other in that they allow us to spot errors in the respective other corpora. By combining these independent sources of information, complemented with machine-learning methods to ‘fill up the gaps’, more knowledge from legacy resources can be preserved in UD, for Old Saxon and beyond.

Because we inherit parts our morphosyntactic annotation and the text itself from existing source corpora, we adhere to their licensing conditions. For UD ConOS, this is an Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0) license with attribution to Zeige et al. (2021) for text and XPOS morphosyntax, and to the current paper for UPOS, HEAD and DEP annotations. For the Heliand subset of UD ConOS, we additionally require attribution to Svetlana et al. (2009).

7. Bibliographical References

- Pórunn Arnardóttir, Hinrik Hafsteinsson, Einar Freyr Sigurðsson, Kristín Bjarnadóttir, Anton Karl Ingason, Hildur Jónsdóttir, and Steinþór Steingrímsson. 2020. A Universal Dependencies conversion pipeline for a Penn-format constituency treebank. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW-2020)*, pages 16–25, Barcelona, Spain (online).
- Otto Behaghel and Burkhard Taeger. 1984. *Heliand und Genesis*, 9 edition. Max Niemeyer Verlag, Tübingen.
- Cathy Bow, Baden Hughes, and Steven Bird. 2003. Towards a general model of interlinear text. In *Proceedings of E-MELD Workshop 2003: Digitizing and annotating texts and field recordings*, pages 11–13, East Lansing, Michigan.
- Christian Chiarcos and Christian Fäth. 2017. CoNLL-RDF: Linked corpora done in an NLP-friendly way. In *International Conference on Language, Data and Knowledge*, pages 74–88. Springer.
- Christian Chiarcos and Christian Fäth. 2019. Graph-based annotation engineering: towards a gold corpus for Role and Reference Grammar. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
- Christian Chiarcos, Christian Fäth, and Maxim Ionov. 2022. Querying a dozen corpora and a thousand years with Fintan. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC-2022)*, pages 4011–4021, Marseille, France.
- Christian Chiarcos and Luis Glaser. 2020. A Tree Extension for CoNLL-RDF. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC-2020)*, pages 7161–7169, Marseille, France (online).
- Christian Chiarcos, Maxim Ionov, Luis Glaser, and Christian Fäth. 2021. An Ontology for CoNLL-RDF: Formal Data Structures for TSV Formats in Language Technology. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
- Christian Chiarcos and Niko Schenk. 2018. The ACoLi CoNLL libraries: Beyond tab-separated values. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.
- Christian Chiarcos and Janine Siewert. 2026. Consolidating syntactically annotated corpora with LLOD technology. An experiment in the Old Saxon Heliand. In *Tenth Workshop on Linked Data in Linguistics (LDL-2026)*, Palma de Mallorca, Spain. Co-located with LREC 2026.
- Stefanie Dipper, Karin Donhauser, Thomas Klein, Sonja Linde, Stefan Müller, and Klaus-Peter Wegera. 2013. HiTS: Ein Tagset für historische Sprachstufen des Deutschen. *Journal for Language Technology and Computational Linguistics*, 28(1):85–137.
- Stefanie Dipper, Cora Haiber, Anna Maria Schröter, Alexandra Wiemann, and Maike Brinkschulte. 2024. Universal Dependencies: Extensions for modern and historical German. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17101–17111.
- Christian Fäth, Christian Chiarcos, Björn Ebbrecht, and Maxim Ionov. 2020. Fintan - Flexible, integrated transformation and annotation engineering. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC-2020)*, pages 7212–7221, Marseille, France (online).
- Cora Haiber. 2024. A Crosslingual Approach to Dependency Parsing for Middle High German. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 23–31, Vienna, Austria.
- Caitlin Light. 2012. The information structure of subject extraposition in Early New High German. *University of Pennsylvania Working Papers in Linguistics*, 18(1):20.
- Sonja Linde. 2009. Aspects of word order and information structure in Old Saxon. In Roland Hinterhölzl and Svetlana Petrova, editors, *Information structure and language change: New approaches to word order variation in Germanic*, pages 367–389. Walter de Gruyter.
- Sonja Linde and Roland Mittmann. 2013. Old German Reference Corpus: Digitizing the knowledge of the 19th century. In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J Whitt, editors, *New Methods in Historical Corpora*, pages 235–246. Narr Francke Attempto Verlag, Tübingen.
- Rosemarie Lühr. 2025. Reflexivität im Altsächsischen. In Norbert Kössinger, editor, *Altsächsisch: Beiträge zur altniederdeutschen Sprache, Literatur und Kultur*. Walter de Gruyter.

- Svetlana Petrova and Michael Solf. 2009. On the methods of information-structural analysis in historical texts: A case study on Old High German. In Roland Hinterhölzl and Svetlana Petrova, editors, *Information structure and language change: New approaches to word order variation in Germanic*, pages 121–203. Walter de Gruyter.
- Svetlana Petrova, Michael Solf, Julia Ritz, Christian Chiarcos, and Amir Zeldes. 2009. Building and using a richly annotated interlinear diachronic corpus: The case of Old High German Tatian. In *Traitement Automatique des Langues, Volume 50, Numéro 2: Langues anciennes [Ancient Languages]*, pages 47–71.
- Susan Pintzuk and Ann Taylor. 1997. [Annotating the Helsinki Corpus: The Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English and the Penn-Helsinki Parsed Corpus of Middle English](#). In *Tracing the Trail of Time: Proceedings from the Second Diachronic Corpora Workshop, New College, University of Toronto, Toronto, May 1995*, pages 91 – 104. Brill, Leiden, Nederlande.
- Eiríkur Rögnvaldsson, Anton Karl Ingason, and Einar Freyr Sigurðsson. 2011. Coping with Variation in the Icelandic Parsed Historical Corpus (IcePaHC). *Language Variation Infrastructure*, 3:97–112.
- Alessio Salomoni. 2021. *A UD Literary Treebank for German*. Ph.D. thesis, Università degli studi di Bergamo, Italy.
- Beatrice Santorini. 1993. The rate of phrase structure change in the history of Yiddish. *Language Variation & Change*, 5(3).
- Christopher D. Sapp, Elliott Evans, Rex Sprouse, and Daniel Dakota. 2024. [Introducing a Parsed Corpus of Historical High German](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9224–9233, Torino, Italia. ELRA and ICCL.
- Silke Scheible, Richard J Whitt, Martin Durrell, and Paul Bennett. 2011. A gold standard corpus of Early Modern German. In *Proceedings of the 5th Linguistic Annotation Workshop (LAW V), held in conjunction with ACL-HLT 2011*, pages 124–128, Portland, Oregon, USA.
- Mariana Scott. 1966. *The Heliand: Translated from the Old Saxon*. University of North Carolina Press, Chapel Hill.
- Edward Henry Seht. 1925. *Vollständiges Wörterbuch zum Heliand und zur altsächsischen Genesis*. Vandenhoeck & Ruprecht.
- Eduard Sievers. 1878. *Heliand*. Buchhandlung des Waisenhauses, Halle.
- Petrova Svetlana, Donhauser Karin, and Linde Sonja. 2009. [Heliand \(Version 1.0\)](#). Humboldt-Universität zu Berlin.
- Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The Penn Treebank: An overview. In Anne Abeillé, editor, *Treebanks: Building and using parsed corpora*, pages 5–22. Kluwer Academic Publishers, Dordrecht/Boston/London.
- Robert D Van Valin. 2014. Role and reference grammar. In *The Routledge Handbook of Syntax*, pages 579–603. Routledge.
- George Walkden. 2016. The HeliPaD: a parsed corpus of Old Saxon. *International Journal of Corpus Linguistics*, 21(4):559–571.
- Lars Erik Zeige, Gohar Schnelle, Martin Klotz, Karin Donhauser, Jost Gippert, and Rosemarie Lühr. 2021. [Deutsch Diachron Digital - Referenzkorpus Altddeutsch \(Version 1.2\)](#). Humboldt-Universität zu Berlin.

A. Appendix: Annotating Universal Dependencies with Spreadsheets

As accompanying material, Fig. 7 illustrates the annotation of UD-style syntax with the help of off-the-shelf spreadsheet software. We used LibreOffice 25.2.6.2 for this purpose, but also conducted successful tests with Microsoft Excel, Gnumeric¹¹ and Google Spreadsheets.¹² All of these tools support formulas and custom filters (marked by the drop-down menu button in the header row) that we used to perform global replacement and overall sanity checks (checking for typos). How both work together can be easily illustrated by how we implemented the annotation of punctuation characters: We first filtered for punctuation signs in column A (FORM), in the first result row (in this case, row 6), we then set column D (UPOS) to PUNCT, column H (DEP) to punct, and column G (HEAD) to the formula =A6-1 (i.e., the preceding token = value of ID minus 1). This was then copied to all other result rows matched by the filter. As many punctuation signs in our texts do generally not serve a grammatical function, but are used to indicate verse boundaries or line breaks, they can then be filtered out accordingly, so that only lines without punctuation signs remain visible and these do not detract attention. As for the annotation of UPOS and DEP, this is supported by autocompletion. (As we observed on other corpora, spreadsheet software is less appropriate for free-text annotation as often needed for FEAT, XPOS, LEMMA and DEPS.)

	A	B	C	D	E	F	G	H	I	J
	ID	FORM	LEMMA	UPOS	XPOS		HEAD	DEP		MISC
1	1	Sidoda	sidon	VERB	VVFIN.IND_PAST_SG_3		0	root		Gloss=gehen
2	2	im	he	PRON	PRF.MASC_SG_DAT_3		1	obl		Gloss=er
3	3	thuo	do	ADV	ADV		1	advmod		Gloss=nun
4	4	te	te	ADP	APPR		5	case		Gloss=zu
5	5	selidon	selida	NOUN	NA.PL_DAT		1	obl		Gloss=Haus
6	6	,	,	PUNCT	\$,		5	punct		
7	7	habda	hebbian	AUX	VAFIN.IND_PAST_SG_3		10	aux		Gloss=haben
8	8	im	he	PRON	PRF.MASC_SG_DAT_3		10	obl		Gloss=er
9	9	sundea	sundea	NOUN	NA.SG_ACC		10	obj		Gloss=Übeltat
10	10	giuuarahht	giwirkian	VERB	VVPP		1	advcl		Gloss=tun
11	11	bittra	bittar	ADJ	ADJN.POS_FEM_SG_ACC_ST		10	xpos		Gloss=bitter
12	12	an	an	ADP	APPR		14	case		Gloss=an
13	13	is	he	DET	DPOS.MASC_SG_GEN_3		14	det:poss		Gloss=er
14	14	bruod̥ar	brod̥ar	NOUN	NA.SG_DAT		11	obl		Gloss=Bruder
15	15	;	;	PUNCT	\$.		14	punct		

Figure 7: Annotating Universal Dependencies with spreadsheet software. First sentence of the Old Saxon Genesis (fragment 2)

Another important feature is the conditional formatting functionality supported by off-the-shelf spreadsheet software, where a colors are applied to columns A (ID) and G (HEAD), with small values in red, medium values in yellow and large values in green. The color of the HEAD column thus provides a visual key for the location of the head word in the current sentence.

¹¹<https://gnome.pages.gitlab.gnome.org/gnumeric-web/>

¹²<https://docs.google.com/spreadsheets>

Gathering Valency Frames for Annotation and Batch Corrections

Mathilde Regnault

Institute of Linguistics
University of Stuttgart (Germany)
{author}@ling.uni-stuttgart.de

Abstract

Syntactic annotation is time- and resource-consuming, especially for historical and heterogeneous data. The Universal Dependencies (UD) framework provides a stable and cross-linguistically consistent annotation scheme, offering a crucial backbone for diachronic corpus studies. However, ensuring internal consistency within historical UD treebanks remains challenging due to syntactic variation and parser errors. We address this issue for Medieval and Early Modern French by integrating valency information into our corrections to support UD treebank maintenance. Valency frames were extracted from the *Profiterole* treebank (v. 2.7) and used to enrich *OFrLex* with structured valency information for Medieval French. Existing lexical resources such as *Lefff* are also exploited for Contemporary French. These valency frames are used to detect and correct inconsistencies in automatically annotated data through batch operations, thereby reinforcing UD guideline compliance and improving annotation coherence across diachronic stages. Preliminary experiments on Medieval French and exploratory annotation of Early Modern French data suggest that lexicon-informed error mining can reduce manual revision effort while strengthening the diachronic continuity enabled by the UD framework.

Keywords: valency, lexicon, error mining

1. Introduction

Extending and maintaining syntactically annotated corpora remains a time- and resource-intensive task, especially for historical language stages. In the Universal Dependencies (UD) framework, automatic pre-annotation with off-the-shelf parsers is common practice, but parser performance drops significantly when applied to non-contemporary varieties exhibiting freer word order and greater structural variation (Grobol et al., 2022).

This is the case for Medieval and Early Modern French (now MedFr and EMFr). Manual correction is necessary, but it can be supported by error-mining strategies targeting the argument structure. In this paper, we investigate how lexical valency information can guide corpus correction. We focus on the extension of the *Profiterole* treebank (Prévost and Stein (2013); Prévost et al. (2024), 9th-15th c.) and additional EMFr texts (Gabay et al. (2022), 16th-18th c.). We exploit two similar and LTAG-compatible lexicons: the *Lefff* for Contemporary French (Sagot, 2010) and *OFrLex* for MedFr (Sagot, 2019).

The contribution is twofold. First, we describe the automatic extraction of valency frames from the *Profiterole* treebank (v2.7) using GREW (Bonfante et al., 2018), in order to enrich *OFrLex* with corpus-attested argument structures. Second, we present ongoing work on using these valency frames to detect and correct systematic parsing errors in automatically annotated data through batch correction procedures.

By integrating lexical valency constraints into the

UD correction workflow, we aim to bridge lexical resources and dependency treebanks in the annotation of historical French.

2. Related Work

Valency information plays a central role in argument attachment and predicate–argument structure modelling (Van den Eynde and Mertens, 2003). In lexicalized formalisms such as LTAG, a verb specifies the number of arguments it selects, their grammatical function, and their syntactic realization. In canonical configurations, up to three core arguments are distinguished: Arg0 (subject), Arg1 (direct object or subject complement), and Arg2 (second object or oblique), with optional locative extensions.

Within the Universal Dependencies (UD) framework, core arguments are represented through relations such as *nsubj*, *obj*, *ccomp*, and *xcomp*, while non-core dependents are typically encoded as *obl*. However, the annotation of *obl* dependencies does not always specify the status of argument or optional modifier. As a result, deriving valency frames from UD-annotated corpora requires additional criteria to separate argumental obliques from adjuncts and to reconstruct lexical subcategorization patterns.

Several strategies have been proposed to acquire valency lexicons. A first approach relies on existing linguistic resources and formal lexical descriptions (e.g. Gardent et al. (2006) for Contemporary French). For MedFr, however, no large-scale, machine-readable valency lexicon is available.

lire v87 100;Lemma:v;<**Suj**:cln|sn,**Obj**:(cla|qcompl|scomp|sinf|sn),**Objà**:(cld|à-sn)>;@Complnd, @CtrlSujObj,cat=v;%actif,%**passif**,%ppp_employé_comme_adj,%passif_impersonnel,%se_moyen_impersonnel

Figure 1: Example of a *Lefff* entry: the verb *lire* ('to read') accepts a nominal subject (mandatory), a nominal direct object, an infinitive or a complement as Arg1 and a nominal complement introduced by the preposition *à* as Arg2 ('to read sth_{Arg1} to sb_{Arg2}'). The entry contains additional information, for instance the possibility for passive voice.

A second approach consists of extracting subcategorization frames from corpora. Early work for English derived frames from raw corpora using statistical methods (Briscoe and Carroll, 1997; Carroll and Fang, 2004; Korhonen, 2002). Other studies rely on syntactically annotated corpora, for instance Zabokrtský (2005) for Czech and Kupść and Abeillé (2008) for French. Their method links arguments to governing verbs, compacts frames by integrating clitic pronouns into their corresponding functions, and separates active and passive realizations.

Our work adopts a similar corpus-driven strategy but operates directly on dependency-annotated data in the UD framework. In our case, valency frames are extracted from the Profiterole treebank (v. 2.7), a UD-annotated corpus of MedFr, which provides manually revised syntactic analyses. This choice offers several advantages: cross-linguistic comparability, standardized relation labels, and compatibility with UD-based parsing tools. At the same time, it raises specific challenges, particularly the absence of an explicit argument–adjunct distinction and the underspecification of secondary dependencies. We therefore combine UD relations (*nsubj*, *obj*, *ccomp*, *xcomp*, *obl*, *expl*, *aux:pass*) with distributional and morphological heuristics to reconstruct valency frames compatible with the tradition of the *Lefff* (see fig. 1).

Linguistic research often relies on automatically parsed data. This raises an issue for historical languages because of the heterogeneous nature of the data and the lack of resources to obtain gold annotation. Therefore, pattern extraction requires an expert knowledge of both the language and the texts. Working on several resources can also be a challenge, especially with concurrent annotation frames (Stein, 2024).

By grounding valency extraction in UD annotation, our approach contributes to ongoing efforts within the UD community to develop language resources and tailor resources for linguistic studies.

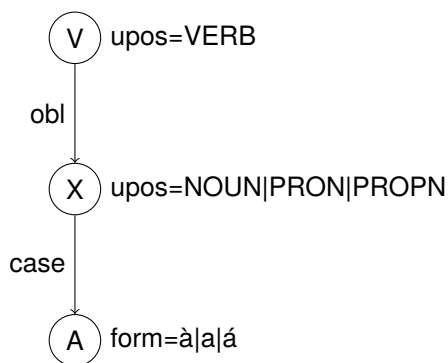


Figure 2: Example of a graph extracting candidates for the indirect object, as implemented in GREW

3. Methods

This section presents the methodology adopted to support UD treebank maintenance through lexicon-informed error mining. Our approach consists of two main steps: first, we automatically extract verb valency frames from a gold-standard UD treebank in order to construct a database of predicate–argument patterns and their syntactic realisations. Second, these frames are exploited to detect systematic mismatches between expected argument structures and automatic or converted annotation, enabling batch corrections.

The procedure combines automatic extraction using graph-pattern queries with targeted manual validation for structurally complex or diachronically unstable constructions. The use of UD-annotated data for valency extraction ensures the compatibility with UD guidelines while tailoring lexical resources to the specific syntactic properties of MedFr. The following subsection details the extraction phase.

3.1. Gathering Valency Frames

We used the gold-annotated texts of the SRCMF-UD Profiterole treebank (v2.7) to build a database of verb valency frames. Following Van den Eynde and Mertens’s definition of valency grammar as a specific subtype of dependency grammar, UD dependency structures can be partially converted into *Lefff*-style argument structures. Argument extraction is performed using GREW. The graph representation makes it possible to jointly exploit syntactic dependencies and linear order information.

First, we gather the candidates for arguments, categorized in canonical valency slots (Arg0, Arg1, Arg2), and map UD relations to the syntactic functions listed in *OFrLex*. For instance, for the indirect object, we gather separately dative complements, nominal (see fig. 2) and infinitive complements introduced by the preposition *à*.

- **Arg0:** In a canonical TAG-based valency framework, Arg0 corresponds to the subject.

In MedFr, however, overt nominal or pronominal subjects are optional and often absent in early texts. We therefore exclude canonical nominal and pronominal subjects from automatic queries and treat them as default (though optional) realizations.

Clausal and verbal subjects are retained. Impersonal subject pronouns are identified through the *Cattex* label (its original tagset, [Prévost et al. \(2013\)](#)) *PROimp* (pronoun + impersonal) in XPOS and cross-checked with the UD relation *expl*. Impersonal subject pronoun *il* ('it') is not treated as a lexical paradigm element.

- **Arg1:** It covers

- direct objects (*obj*),
- finite clausal complements (*ccomp*),
- infinitival complements (*xcomp* and *obj* headed by an infinitive),
- subject predicates in copular constructions (*cop*).

Different realizations (nominal, adjectival, prepositional complements, verbal or infinitival clauses) are encoded in the valency frame. Pronominal realizations are detected via *obj* dependencies restricted to clitic pronouns.

- **Arg2** (indirect object) is typically introduced by *à* (*Objà*) or *de* (*Objde*). It is possible to register two realisations for Arg2 (e.g. *paroler de... à...*, 'to speak about... to...'), provided that there is no Arg1. These complements are extracted via *obl* dependencies combined with prepositional information. However, the current version of *Profiterole* does not distinguish arguments from modifiers in obliques. To address this, we use a distributional heuristic: if a verb alternates between a prepositional complement (*à/de*) and the corresponding clitic pronoun (dative *li/lor*, genitive *en*), the complement is treated as a strong candidate for Arg2 status. This heuristic reduces noise but does not eliminate ambiguity. In particular:

- *Objà* may be confused with locative complements,
- *Objde* may overlap with source complements,
- diachronic variation yields unstable prepositional marking,
- some verbs allow multiple competing realisations.

As noted by [Amiot et al. \(2020\)](#), verbal transitivity in the history of French is "rather chaotic". Consequently, certain frames required manual

resolution. In the absence of a comprehensive reference lexicon for MedFr, automatic frame merging was not feasible. Non-argumental complements (*Loc*, *DLoc*, additional obliques) are stored separately and may host complements introduced by other prepositions (e.g. *avec* 'with', *sur* 'on'/'about').

Second, the number of hits per category is attributed to each lemma. For example, the verb *dire* ('to say' or 'to tell') registers 379 occurrences of direct objects, which indicates it is transitive. There are two candidates for an indirect object: nominal complements introduced by *à* (78) and *de* (43). However, the number of pronouns confirms this status for *à*-complements only, with 302 occurrences of dative clitic pronouns against 1 for *de*-complements.

Apart from core and optional arguments, other information has to be represented in the lexicon, for example the possibility for passive voice. Not all transitive verbs may be used in this manner, and the lexicon requires separate entries for such past participles. Therefore, passive constructions had to be detected via the UD dependency *aux:pass*. The reflexive pseudo-paradigm *se* is also extracted via UD dependencies (*expl*), excluding the impersonal pronoun *il*. It is encoded as part of the predicate rather than as an argument.

Highly polyfunctional verbs such as *faire* were also treated manually due to their structural complexity. Six entries were necessary to describe all the use cases, as in our reference, the *Lefff*, for example¹:

- (1) ex. canonical entry
Or **fai** ton mialz
now do-IMP your best
'Now do your best'
Yvain, Chrétien de Troyes, v. 4184 (12th c.)
- (2) ex. light verb
Fai le **venir**
make-IMP he-ACC.SG come-INF
'Make him come'
Quatre Livres des rois, p. 31 (12th c.)
- (3) ex. with an attribute and a reflexive pronoun
Se chaschun ne **se**
unless everyone NEG REFL.ACC
fait **humble** comme che enfant,
make-PRS humble like this child
'Unless everyone makes themselves
humble like this child'
De la erudition, J. Daudin, p. 394 (14th c.)

¹All examples of MedFr come from the BFM ([Guillot et al., 2018](#)).

This procedure (Regnault, 2022b) yielded an initial set of valency frames for 1,885 verbs (Regnault, 2022a) extracted from the *Profiterole* treebank, complemented by manually curated entries (see annex for detailed count). Automatic pruning of frames exceeding the maximum of three argument realisations can be considered. In our case, precision was necessary in order to parse as many sentences as possible with the corresponding TAG metagrammar (Regnault et al., 2019) and annotating them according to UD guidelines. In a second step, we aligned extracted lemmas with those of *OFLex* to propagate attested argument structures across lemma variants, broadening the coverage of the lexicon.

3.2. Batch Corrections

The extracted valency frames are integrated into GREW-based correction workflows² targeting recurrent inconsistencies in the *Profiterole* treebank (Prévost et al., 2022), more specifically passive voice annotation and argument structure validation. Not all verbs marked by the passive voice received the proper *aux:pass* dependency in the *Profiterole*, making their extraction difficult for linguistic research. Only verbs compatible with passive formation should receive *aux:pass*, so we checked the 6,438 existing *aux* annotations and examined the transitive verbs employed with the *estre* auxiliary (when *avoir* is to be expected, which amounted to 436 cases) with a GREW rule. Missing verbal entries were added to the lexicon, and manual checks were necessary, especially for unaccusative verbs, allowing both 'be' and 'have' auxiliaries. This procedure resulted in the addition of 358 *aux:pass* dependencies. The *nsubj:pass* and *csubj:pass* dependencies have been automatically added, as well as *expl:pass*, but with a filter on the form, only allowing subject pronouns. The dependencies *obl:agent* however all necessitated a manual check, because they can be mistaken by general obliques, even when its head is an agentive noun.

Beyond passive normalization, we also addressed the annotation of indirect arguments. Many oblique complements did not receive a secondary label, in particular *arg* for indirect objects. Candidates for *obl:arg* were identified using a GREW lexical filter restricting the query to verbs selecting an Arg2. In order to avoid annotating modifiers as arguments, we added a lexical filter to the head of the complement, excluding nouns with a "time" feature from the lexicon and other lexical values taken from the corpus. This led to an increase of 464 *obl:arg* dependencies. These scripts will be useful for future expansions of the treebank.

²These workflows are available in the "tools" section of the repository and may be linked to specific commits.

To evaluate the portability of this workflow, we pre-annotated the EMFr *FreEM* corpus (Gabay et al., 2022) with the HOPS parser (Grobol and Crabbé, 2021) with the Flaubert model (Le et al., 2020), chosen for its proximity with Contemporary French.

First, we cross-check dependencies and lexicon entries. For example, in the parser-annotated version, there are 177 *nsubj:pass* dependencies for only 174 *aux:pass*. Some intransitive verbs also appear to be annotated with the passive voice, like *naître* ('to be born').

Second, we design graph-rewriting rules tailored to our data. Verbs which do not have the "passive" feature in the lexicon received a simple *aux* dependency (6 changes), and arguments were corrected according to the auxiliary's status (112 changes).

Although this stage remains ongoing, the rewriting scripts informed by *Lefff* valency frames can be applied to pre-annotated EMFr corpora as a support for maintenance or as an off-the-shelf correction workflow³.

These interventions illustrate how valency-informed constraints can serve as a validation layer over UD dependency graphs, improving both internal consistency and cross-version reproducibility.

4. Conclusion

The maintenance and extension of historical UD treebanks remain effort-intensive tasks, particularly in the presence of diachronic variation and unstable argument structures. In this paper, we presented a corpus-driven extraction of valency frames from the gold-annotated *Profiterole* treebank, resulting in an initial inventory of 1,885 verbs enriched with structured argument information and aligned with *OFLex*.

By extracting valency directly in UD dependency annotations, we provide a resource that is both theoretically interpretable and operational for treebank maintenance. The valency frames offer support to the detection of errors in argument structure and makes batch corrections possible and improves consistency with UD guidelines. While the large-scale exploitation of these frames for automatic revision remains ongoing, preliminary experiments indicate that lexicon-informed validation can substantially reduce manual intervention.

Beyond the case of Medieval French, this approach illustrates how UD-based lexical extraction can strengthen diachronic continuity and enhance reproducibility across treebank releases, especially when correction scripts are made openly available.

³This resource is available under: github.com/mregnault/freem-treebank.

5. Bibliographical References

- Dany Amiot, Claire Badiou-Monferran, Bernard Combettes, Benjamin Fagard, Christiane Marchello-Nizia, and Maj-Britt Mosegaard Hansen. 2020. [Catégories invariables](#). In Christiane Marchello-Nizia, Bernard Combettes, Sophie Prévost, and Tobias Scheer, editors, *Grande Grammaire historique du français*, pages 856–961. de Gruyter.
- Guillaume Bonfante, Bruno Guillaume, and Guy Perrier. 2018. [Application of Graph Rewriting to Natural Language Processing](#). Wiley-ISTE.
- Ted Briscoe and John Carroll. 1997. [Automatic extraction of subcategorization from corpora](#). *arXiv preprint*.
- John Carroll and Alex C. Fang. 2004. [The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser](#). In *International Conference on Natural Language Processing (IJCNLP 2004)*, pages 646–654, Berlin, Heidelberg. Springer, Springer.
- Claire Gardent, Bruno Guillaume, Guy Perrier, and Ingrid Falk. 2006. [Extraction d'information de sous-catégorisation à partir du lexique-grammaire de Maurice Gross](#). *TALN 2006*.
- Loïc Grobol, Sophie Prévost, and Benoît Crabbé. 2022. [Is Old French tougher to parse?](#) In *20th International Workshop on Treebanks and Linguistic Theories*, Sofia, Bulgaria.
- Loïc Grobol and Benoît Crabbé. 2021. [Analyse en dépendances du français avec des plongements contextualisés](#). In *Actes de la 28ème Conférence sur le Traitement Automatique des Langues Naturelles*.
- Anna Korhonen. 2002. [Subcategorization acquisition](#). Technical report, University of Cambridge, Computer Laboratory.
- Anna Kupś and Anne Abeillé. 2008. [Growing Treelex](#). In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 28–39. Springer.
- Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. [FlauBERT: Unsupervised language model pre-training for French](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.
- Sophie Prévost, Céline Guillot-Barbance, Alexei Lavrentiev, and Serge Heiden. 2013. [Jeu d'étiquettes morphosyntaxiques CATTEX2009](#). Technical report, Technical report, École normale supérieure de Lyon.
- Mathilde Regnault, Sophie Prévost, and Éric Villamonte de La Clergerie. 2019. [Challenges of language change and variation: towards an extended treebank of medieval french](#). In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 144–150.
- Benoît Sagot. 2010. [The lefff, a freely available and large-coverage morphological and syntactic lexicon for french](#). In *7th international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.
- Benoît Sagot. 2019. [Développement d'un lexique morphologique et syntaxique de l'ancien français](#). In *26ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Toulouse, France.
- Achim Stein. 2024. [Réconcilier les formats d'annotation syntaxique pour faciliter l'analyse de la syntaxe française en diachronie](#). In *SHS Web of Conferences*, volume 191, page 11004. EDP Sciences.
- Karel Van den Eynde and Piet Mertens. 2003. [La valence : l'approche pronominale et son application au lexique verbal](#). *Journal of French language studies*, 13(1):63–104.
- Zdenek Zabokrtský. 2005. [Valency lexicon of Czech verbs](#). Ph.D. thesis, Faculty of Mathematics and Physics, Charles University in Prague.

6. Language Resource References

- Gabay, Simon and Clérice, Thibault and Gille Levenson, Matthias and Camps, Jean-Baptiste and Tanguy, Jean-Baptiste. 2022. [FreEM-corpora/FreEMlpm: FreEM LPM \(Lemma, POS-tags, Morphology\) corpus](#). Zenodo.
- Guillot, Céline and Heiden, Serge and Lavrentiev, Alexei. 2018. [Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique](#). Presses de l'Université Paris-Sorbonne (PUPS), number 7 in Les états anciens des langues à l'heure du numérique.

Sophie Prévost, Loïc Grobol, Mathieu Dehouck, Alexei Lavrentiev, and Serge Heiden. 2024. [Profiterole: un corpus morpho-syntaxique et syntaxique de français médiéval](#). *Corpus*, La constitution de corpus en diachronie longue. Méthodologies, objectifs et exploitations linguistiques et stylistiques.(25).

Prévost, Sophie and Grobol, Loïc and Dehouck, Mathieu and Regnault, Mathilde. 2022. [Corpus Profiterole](#). GitLab.

Prévost, Sophie and Stein, Achim. 2013. [Syntactic Reference Corpus of Medieval French \(SRCMF\)](#). ENS de Lyon/ILR Stuttgart.

Regnault, Mathilde. 2022a. [OFrLex-dev](#). GitLab Inria.

Regnault, Mathilde. 2022b. [Valency finder](#). GitHub.

Syntax is the Key to Semantics: Universal Dependencies and Abstract Meaning Representation

Johannes Heinecke

Orange Research, 22300 Lannion, France

johannes.heinecke@orange.com

Abstract

This paper presents a new Abstract Meaning Representation (AMR) dataset using sentences which are already translated in 21 languages and annotated in dependency syntax (Parallel Universal Dependencies, PUD). We annotated the English version of the 1000 sentences available in the PUD dataset. Since PUD provides syntactic annotations, this new AMR dataset (ParallelUD-AMR) allows comparisons of syntactic phenomena and their corresponding semantics. We finally provide a first analysis of parallels between syntactic and semantic structures.

1. Introduction

Syntax is an important feature of all natural languages: word order, agreement, relations between words (adjacent or not), hierarchies etc. In most cases changing the syntactic construction of a sentence implies a change of meaning.

Parallel UD helps to compare and see differences in syntax and typology of languages. In order to have directly comparable corpora in syntax and semantics, we created a new Abstract Meaning Representation (AMR) dataset named ParallelUD-AMR¹, which is based on the sentences UD_English-PUD treebank². We annotated the 1000 sentences following the AMR guidelines³. The goal of this new dataset is to show parallels of Semantic and Syntactic representations in different languages. In the following we give a short overview on related work (section 2), present existing UD and AMR data (3) and the annotation process (4). Section 4 also details some inconsistencies in the PUD translations which in consequence mean that the AMR graph calqued on the English version of the sentence does not correspond exactly to some of the translated sentences. Section 5 presents a first idea of how this dataset can be used for comparison. We conclude with some ideas of how this dataset could also be used (section 6).

2. Background

Even though Universal Dependencies and Abstract Meaning Representation annotate different linguistic layers (dependency syntax in the former

case, argument semantics in the latter) there is an interest to combine them, since at least a part of the semantic relations which are represented by the AMR graphs are a consequence of the syntactic construction of the sentence. For instance the subject of the English verb *to see* in the Active Voice encodes the agent of the *seeing*-event (coded :ARG0 in AMR), whereas the subject of the same verb in the Passive Voice encodes the patient (coded :ARG1 in AMR), cf. Schuster and Manning (2016) for a comparison of UD data and AMR. Alignment between AMR and dependency syntax has already been discussed in Szubert et al. (2018) who discuss how to establish relations between syntactic dependencies and semantic (AMR) relations. We think our new corpus can help to advance in this direction. Parallel Universal Dependencies data (PUD, see below in section 3.1) has been used for a shared task a while ago (Mille et al., 2018). Instead of creating semantic graph; participants were asked to add word order and inflected forms from incomplete dependency trees. Even though this task does not involve AMR or a different kind of semantics it shows the interest which lies in multilingual parallel data. Other work exists to at least bootstrap or even create AMR graphs by using a syntactic representations of sentence (Pust et al., 2015), frequently taken from UD (Han and Pavlova, 2019). Further approaches use UD data to bootstrap Uniform Meaning Representation (UMR, (Van Gysel et al., 2021)), a (multilingual) evolution of AMR (Gamba et al., 2025). Wang et al. (2017) published an approach of a parallel use of dependency syntax trees and AMR for information extraction, showing the interest in corpora annotated in both frameworks. A complementary approach is presented by Wein and Schneider (2024). They discuss differences and similarities in parallel (multilingual) AMRs also taken from translations but using concepts and relations drawn from the sentence languages (and not En-

¹The dataset will be available at <https://github.com/orange-OpenSource/ParallelUD-AMR>.

²https://github.com/UniversalDependencies/UD_English-PUD/

³<https://github.com/amrisi/amr-guidelines/blob/master/amr.md>

glish as in our corpus). However, there is no formal syntax in their approach.

Two early AMR-Parsers were initially based on transition-parsing developed for the parsing of dependency syntax: C-AMR (Wang et al., 2015)⁴ and JAMR⁵. Both show the interest of regarding syntax and semantics at the same time.

3. Data

3.1. Universal Dependencies (UD)

Universal Dependencies is a large community activity to annotate as many languages as possible in dependency syntax, using the same annotation guidelines (Nivre et al., 2020; Marneffe et al., 2021). The latest release (2.17 of November 15, 2025⁶), contains over 200 treebanks for more than 150 languages. If a treebank is large enough, a split in training/dev/test is available to train dependency parsers like UDpipe (Straka and Straková, 2017; Straka, 2018) or Stanza (Qi et al., 2020).

3.2. Parallel Universal Dependencies (PUD)

PUD is an initiative within the UD project to annotate 1000 sentences translated into currently 21 languages of different language families (Arabic, Chinese, Czech, English, Finnish, French, Galician, German, Hindi, Indonesian, Icelandic, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Thai and Turkish). Half of the sentences is taken from news, the other half stems from Wikipedia. The 1000 sentences appear in the same order and with the same sentence identification in the respective files. 750 of the original sentences were English. The rest was taken from German, French, Italian or Spanish resources⁷ and translated via English into all other languages by professional translators. The translations were then annotated according to the UD v2 guidelines and are available via the UD website⁸. We used release 2.16 for this work. There are other datasets available with different translations⁹, PUD however is the largest dataset (both in number of sentences and number of different languages).

⁴<https://github.com/c-amr/camr>

⁵<https://github.com/jflanigan/jamr>

⁶We use release 2.16 in this paper (Zeman, 2025)

⁷The first two digits of the sentence ID indicate the original language: English (01), German (02), French (03), Italian (04) or Spanish (05).

⁸<https://universaldependencies.org>

⁹<https://universaldependencies.org/contributing/parallel.html>

3.3. Abstract Meaning Representation (AMR)

Abstract Meaning Representation (Banarescu et al., 2013) provides a framework to model the meaning of a sentence, notably actions, events or states and their participants. AMR relies heavily on (verbal) concepts defined in PropBank (Kingsbury and Palmer, 2002; Palmer et al., 2005)¹⁰, e.g. the concept `cost-01` in figure 1 (serialised PENMAN format (Kasper, 1989)) refers to PropBank’s sense (or “roleset”) `-01` of the verb “to cost”. Instances are indicated by a following “/”, e.g., `c` being an instance of the concept `cost-01`. The names of the variables do not have any other semantics than being distinct. Relations are indicated by an initial colon (e.g. `:ARG1`, `:time`). Literals (strings and numbers) lack a preceding instance and “/” (c.f. “NoMa” and 2004 in the example in figure 1). Alignments with tokens of the sentence are optional in PENMAN and are marked by adding `~e.n` after the concept, relation or literal. `n` aligns the instance/relation/literal with the `n`th token (0-indexed) or the sentence.

```
# 0 1          2 3 4      5 6      7          8
# By comparison , it cost $ 103.7 million to
# 9      10 11 12      13      14      15
# build the NoMa infill Metro station ,
# 16      17      18 19      20
# which opened in 2004 .
(c2 / cost-01~e.4
  :ARG1 (b / build-01~e.9
    :ARG1 (s / station~e.14
      :ARG1-of (o / open-01~e.17
        :time~e.18 (d / date-entity
          :year 2004~e.19))
        :mod (i / infill~e.12)
        :mod (s1 / subway~e.13)
        :name (n / name
          :op1 "NoMa"~e.11)))
    :ARG2 (m / monetary-quantity
      :quant 103700000~e.6,7
      :unit (d2 / dollar~e.5))
    :ARG2-of (c / compare-01~e.1))
```

Figure 1: AMR graph for “By comparison, it cost \$103.7 million to build the NoMa infill Metro station, which opened in 2004.” in PENMAN format including alignments with tokens (sentence taken from UD-English-PUD id n01005023; our AMR annotation)

The largest available corpus used to train models capable of parsing sentences from natural languages into AMR graphs, called AMR 3.0, LDC2020T02 (Knight et al., 2020)¹¹, is provided by the Linguistic Data Consortium (LDC). This corpus is composed of nearly 59 000 sentences and

¹⁰<https://propbank.github.io/>

¹¹<https://catalog.ldc.upenn.edu/LDC2020T02>

corresponding AMR graphs. The data contains discussions from forums (partly technical), news reels, translations to English of Chinese news broadcasts, along with a part originating from English Wikipedia pages and Aesop’s fables (see LDC2020T02 documentation).

Even though AMR has been designed originally for English, it has also been employed successfully for other languages. A translation of the test dataset of AMR-2.0 is available at LDC (Damonte and Cohen, 2020)¹².

In order to annotate languages other than English, there are two approaches: Using concepts in the language of the sentences (by using a corresponding version of PropBank, if existing, e.g. Akbik et al. (2015)¹³, or keeping concepts in English (Heinecke and Shimorina, 2022).

An extension of AMR, Uniform Meaning Representation (UMR) has been designed for multilingual usage by keeping the concepts of the graph in the same language as the corresponding sentences (Van Gysel et al., 2021). UMR includes all of the information from AMR, and additional information about the position of the sentence in a document, temporal layering and coreferences between entities of different sentences and alignments of tokens of the sentence with instances of the UMR graph.

Whereas UMR follows the first approach we chose the latter for the annotation of the PUD data, since the available translations of the PUD sentences are supposed to have the same meaning, so a single AMR graph for each sentence and its translations is needed. Since PropBank is available only in English and equivalents in other languages are less complete or not available, AMR graphs based on English seem to be the best solution.

4. Annotation

Out of the 1000 sentences taken from UD_English-PUD the first part of 500 sentences were annotated by 4 annotators, using the AMR annotation tool *metAMoRphosED* (Heinecke, 2023), the second part of 500 sentences was annotated by 3 annotators. The annotators did not have to write the graph entirely. We used two different AMR predictors to provide different preannotations to the annotators. For this we used AMRlib¹⁴ trained on AMR 3.0 and AMR-transition parser (Drozdz et al., 2022; Lee et al., 2022)¹⁵. The annotators task was to verify and if needed to modify the generated graph with respect to the AMR guidelines.

During and after the annotation we checked the AMR graph to ensure some basic formal validity (either provided by the annotation tool *metAMoRphosED* or by additional scripts). In particular we checked whether

- concepts with a numerical suffix (like *see-01* or *have-org-rel-91* actually exist in PropBank¹⁶ and that eventually outgoing :ARGn-relations are defined for the given text.
- there are no duplicate :ARGn between two instances (e/eat-01 :ARG1 (...) :ARG1 (...))
- all relations other than :ARGn are valid AMR relations
- named entities in the sentence are correctly represented in the Graph (including the fact that demonyms may appear as adjectives in the text (e.g. *American* or *German*). They are nevertheless represented by (c/country :name (n/name :op1 "United" :op1 "States" :op2 "of" :op4 "America")) or ((c/country :name (n/name :op1 "Germany"))
- dates in the text correspond to date-entity instances in the graph: E.g. ...*June 2002*... is represented as (d/date-entity :month 6 :year 2002)
- numerical values in the text correspond to targets of :quant or :value relations: E.g. *two million dollars* is represented by (m/monetary-quantity :quant 2000000 :unit (d/dollar))

In terms of Smatch (Cai and Knight, 2013) the inter-annotator agreement (IAA) (Wein and Schneider, 2024, p. 421) was 92.4% for the former and 93.1% for the latter half using average pairwise agreement. The upper graph in Fig. 2 presents the averaged Smatch between all 4 annotators for the first 500 sentences. Since the Smatch metric depends on the graph size¹⁷ we also show the averaged number of differences between all annotators. In spite of some outliers, in general annotators agreed well. Fig. 3 shows the same information for the second 500 sentences (3 annotators).

In cases where not all 4 (or 3) annotators agreed, we manually chose one of the graphs. We also compared the final annotations of each annotator with the original pre-annotated files provided to each of the annotators. IAA varies between 93.1% (2.6 differences per graph in average) and

¹²<https://catalog.ldc.upenn.edu/LDC2020T07>

¹³<https://github.com/UniversalPropositions/UP-1.0>

¹⁴<https://github.com/bjascoib/amrlib>

¹⁵<https://github.com/IBM/transition-amr-parser>

¹⁶<https://github.com/propbank/propbank-frames/>

¹⁷A single error in a large graph decreases the Smatch score only slightly, whereas a single error in a small graph decreases the score drastically.

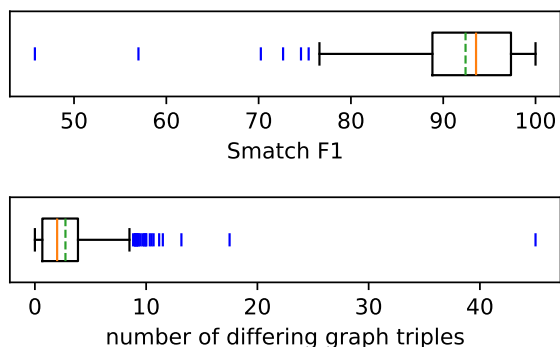


Figure 2: Boxplot average of pairwise IAA for first part of the corpus. Top: Smatch F1: Median (solid orange line): 93.6%, mean (dashed green line): 92.4%, std. deviation: 6.52 Bottom: number of differing triples between graphs: Median: 2.0, mean: 2.75, std. deviation: 3.18

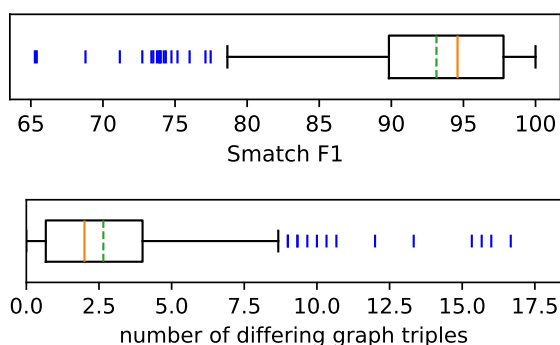


Figure 3: Boxplot average of pairwise IAA for second part of the corpus. Top: Smatch F1: Median (solid orange line): 94.6%, mean (dashed green line): 93.1%, std. deviation: 6.44 Bottom: number of differing triples between graphs: Median: 2.0, mean: 2.65, std. deviation: 2.70

96.9% (0.62 differences per graph). This means that the annotators did not completely rewrite the automatic generated graph, but in a final step we added alignments between words of the sentence and instances, relations and literals of the corresponding AMR graph (following the alignment guidelines of AMR 3.0).

We chose the English PUD sentences as the basis for the annotation for two reasons: 1) 750 of the sentences were originally already in English; 2) AMR was designed for English, even though it is used to annotate sentences in other languages

4.1. Inconsistencies

Even though the translations into the other languages of the PUD treebanks were produced by professional translators, there are cases where a

more literal translation would have facilitated the alignment between the AMR graph and the non-English sentence. These minor differences mean that the AMR graph does not correspond as well to the non-English sentence than to the English sentence. For instance (the first language in these examples is the original language of the sentence; list not exhaustive):

translation of Named Entities: There are cases where a named entity (NE) in English was translated by a different expression or the function of the NE:

- en: *on Capitol Hill* (id: n01001013) →
 - fr: *sur le Capitol*;
 - fi: *Yhdysvaltain kongressissa* “in the US congress”
- en: *the GOP nominee* (id: n01002017) →
 - de: *der Nominierte der Republikanischen Partei*,
 - ru: *кандидат Республиканской партии*,
 - tr: *Cumhuriyetçi Parti Başkan adayı Trump* “the Republican Party presidential candidate Trump” the Turkish translation also mentions the name “Trump”, whereas in the other languages only a “nominee” or “candidate” is mentioned
- en: *Zettel’s Traum* (“the dream of Zettel”, here “Zettel” being a person name, but in German the noun “Zettel” also means small bit of paper for notes) (id: n01066045) →
 - es: *El sueño de la ficha* “the dream of the card/paper”;
 - is: *Draumí Spóla* “the dream of the tape”

incorrect translation of Named Entities:

- en: *It’s the election, of course, not “Game of Thrones.”* (id: n01132006) →
 - pt: *São as eleições, claro, não a “Guerra dos Tronos.”*

adding details to Named Entities: Titles and first names appear in translations but were absent in the original (English) sentence:

- en: *Mr Panvalkar* (id: n01010042) →
 - de: *Prasad Panvalker*
- en: *Lamarr accompanied Mandl to business meetings* (id: w01101020) →

- fr: *Hedy Lamarr accompagnait **Friedrich Mandl** aux réunions d'affaires*
- en: *De Gaulle had instructed* →
 - cs: ***Charles** de Gaulle vydal pokyny*
- en: *until after Smith's death.* (id: 01139043) →
 - fr: *jusqu'après la mort de **E.E.** Smith*
- en: *Pebe moved the family to Nashville, Tennessee* (id: w01147010) →
 - fr: *En 1991, Pebe **Seber** déménage avec sa famille à Nashville (Tennessee)*

adding/changing types of Named Entities:

- en: *Wright was born in Poole, Dorset...* (id: w01149002) →
 - ar: *ولد رايت في مدينة بول* "Wright was born in the **city** of Poole"
- fr: *Le maire de Saint-Gaudens (Haute-Garonne)* (id: n03004003) →
 - de: *Bürgermeister von Saint-Gaudens (**im Département** Haute-Garonne),*
 - es: *el alcalde de Saint-Gaudens (**en la región** Haute-Garonne),*
 - ar: *عمدة سان - غودانس (في إقليم العليا غارون)* "the **region** Haute-Garonne",
 - ru: *мэр города Сан-Годан (**в регионе** Верхняя Гаронна)*
- de: *Larry Vance etwa sagte dem SPIEGEL* (id: n02066010) →
 - ru: *Ларри Вэнс заявил **журналу** SPIEGEL*

removing type from Named Entities: The opposite of the case before:

- es: *federalización de **la Ciudad de** Buenos Aires* (id: w05008107) →
 - cs: *federalizaci Buenos Aires,*
 - ru: *федерализацию Буэнос-Айреса*
- de: ***Städte** Köln und Bremen* (id: w02018030) →
 - ru: *Кёльна и Бремена*

added preposition in Named Entities:

- en: *CTV Montreal* →
 - ru: *CTV **в** Монреале* (id: w01141025)

"free" translation: The translation is adapted to the cultural context and the way things are expressed in the target language and is not a literal translation

- en: ***large** bank account* (id: n01010042) →
 - de: ***wohlgefülltes** Bankkonto* "well filled bank account",
 - fr: *compte en banque **imposant*** "impressing bank account",
 - ru: ***крупным суммам** на банковском счету* "**large sums** in [C.'s] bank account"
- en: *More than 330 crew are **onboard** the ship.* (id: n01070016) →
 - fr: *L'équipage du navire **comporte** plus de 330 membres.*
- it: *Italia abbia **meno** chilometri di metropolitana* (id: n04002020) →
 - de: *dass Italiens U-Bahnlinien **kürzer** sind,*
 - tr: *İtalya'nın **daha** kısa bir metro hattı olduğunu*
- de: *steigenden **Staatsausgaben*** (id: n02078011) →
 - en: *increasing **federal** expenditures,* es: *aumento del gasto del **gobierno federal***

incorrect translation: In contrast to the former, here a term has been incorrectly translated

- de: *Kloster Eibingen* (id: w02015087 also w02004008, w02015084 and w02015086) →
 - en: *cloister Eibingen,* fr: *cloître d'Eibingen* (instead of *monastery/monastère Eibingen*),

"incomplete" translation: Words or expression which are not translated

- en: *meanwhile* not translated in German (id: n01101017)
- de: *dabei* not translated in any other language (id: w02019077)

"overcomplete" translation: Additions absent in the English source or implicits made explicit

- en: *Is series two working so far?* (id: n01121051) →

- fr: *Ça marche **pour toi** jusqu'à présent, la deuxième saison ?*
- en: *For those who follow social media transitions on Capitol Hill, this will be a little different.* (id: n01001013) →
 - tr: ***Bu durum** Capitol Hill hakkındaki sosyal medya hareketlerini takip edenler için biraz daha farklı olacak. "this situation will be a little different"; cf. Fig. 4 and 5.*

“tautological” translation:

- en: *The **northern** portion of the Caucasus is known as the **Ciscaucasus** and the southern portion as the Transcaucasus.* (id: w01014002) →
 - ru: ***Северная** часть Кавказа известна как **Северный Кавказ**, а южная часть — как Закавказье.*

typographical errors:

- fr: *Au VIII^e siècle **av[ant]** J.-C.* “before Christ”) (id: w03005012, w03005014) →
 - de: *Im 8. Jahrhundert **n[ach]** Chr.* “after Christ”
- fr: *En pluviôse an II* (id: w03008029) →
 - de: *Im Monat **Pluivoise** des Jahres III* (in several languages the French original an II “year two” is translated with “year III”;

5. Comparison

As mentioned in the introduction, the motivation to create the ParallelUD-AMR dataset was to be able to identify parallels (and differences) between the dependency trees and the semantic Graph (in AMR). Whereas the syntactic annotation is anchored to the words (tokens) of the sentence, the AMR graph (at least in its basic form) is not anchored to its corresponding sentence which makes it difficult for any automatic alignment of syntactic and semantic relations. However the AMR formalism provides the alignment of tokens of the sentence with instances, literals or relations of the AMR graph (a part of the AMR 3.0 dataset comes with alignments).

We started to compare the AMR relations (between instances of concepts or literal) with the dependency relations (between tokens of the sentence). In order to do so, we not only needed alignments between the tokens of the English sentence and the AMR graph (done during the AMR annotation) but also alignments for the languages

we wanted to compare with AMR. Figures 4 and 5 present these alignments for English and Turkish.

While we did the alignments for English manually, for other languages (in our case French, Arabic and Turkish) we used an LLM (GPT-4.1-mini) to provide alignments. We validated the entire French corpus manually but only some sentences of the Arabic and Turkish corpus. We used the following prompt in Fig. 7 with two examples (sentence and the corresponding AMR graph with alignments).

System prompt: *you are a specialist in Abstract Meaning Representation Semantics.*

User prompt:

Please give me a list of correspondences between the words of the French sentence and an AMR-concept of the semantic graph (use a json dictionary for the output):

Two examples:

- sentence: *Le président préfère de ne pas organiser des élections anticipées .*

- semantic graph:

```
( p / prefer-01
  :ARG0 ( p2 / person
    :ARG0-of ( h / have-org-role-91
      :ARG2 ( p3 / president)))
  :ARG1 ( h2 / hold-04
    :polarity -
    :ARG0 p2
    :ARG1 ( e / elect-01
      :time ( e2 / early))))
```

- output:

```
[
  "position": 0, "word": "Le", "concept": null,
  "position": 1, "word": "président", "concept": "president",
  "position": 2, "word": "préfère", "concept": "prefer-01",
  "position": 3, "word": "de", "concept": null,
  "position": 4, "word": "ne", "concept": ":polarity",
  "position": 5, "word": "pas", "concept": ":polarity",
  "position": 6, "word": "organiser", "concept": "hold-04",
  "position": 7, "word": "des", "concept": null,
  "position": 8, "word": "élections", "concept": "elect-01",
  "position": 9, "word": "anticipées", "concept": "early",
  "position": 10, "word": ".", "concept": null
], [
  ...
```

Figure 7: Prompt for GPT-4.1-mini to align tokens of a sentence with the AMR graph

An initial comparison produced expected results, for instance that a `nsubj` dependency relation (with verbs in Active Voice) corresponds mainly to an `:ARG0` relation in the AMR graph across the four languages which we examined. We counted the co-occurrences of a dependency relation between two tokens and the AMR relation between instances aligned to the tokens (Table 1). For instance “this” is `nsubj` of “[be] different” in the (English) dependency tree in Fig. 6 (orange back-

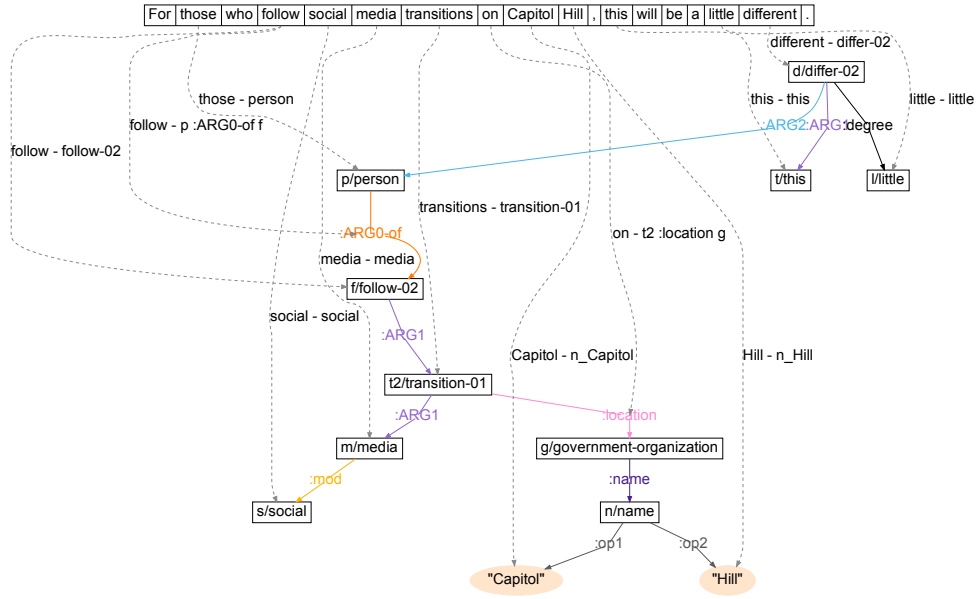


Figure 4: AMR graph and corresponding English tokens aligned (id: n01001013)

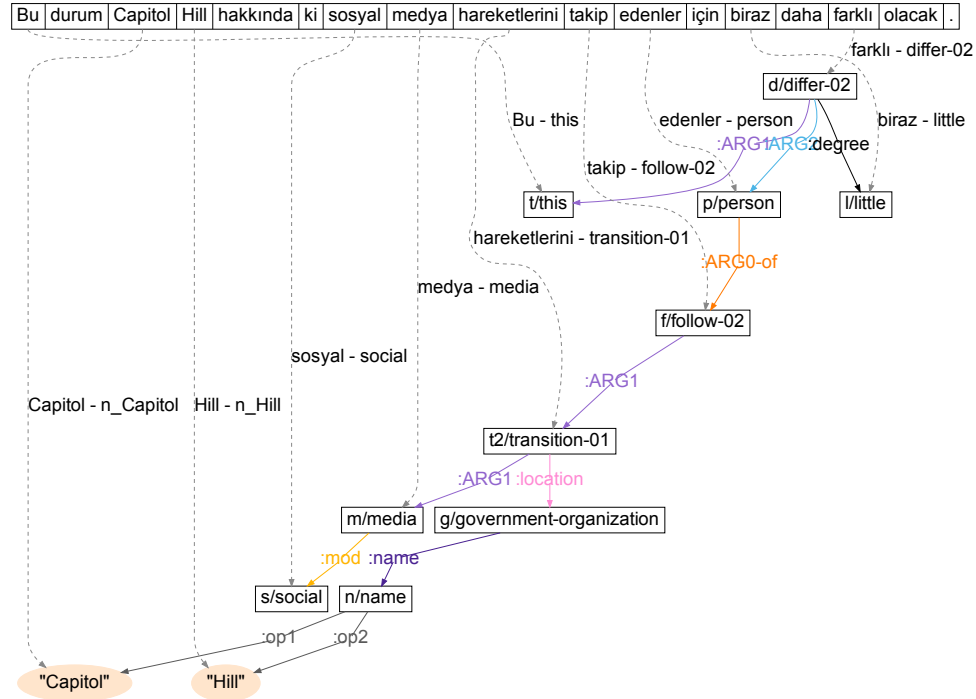


Figure 5: AMR graph and corresponding Turkish tokens aligned (id: n01001013). *Durum* (“situation”) is not aligned since this is implicit in the English sentence which was used to create the AMR graph.

ground) and the tokens “this” and “different” of the syntax tree are aligned to the instances *t/this* and *d/differ-02* of the AMR graph, respectively (Fig. 4). So the dependency relation *nsubj* can be mapped to the AMR relation *:ARG0*. 36.1% of *nsubj* (279 examples in ParallelUD-AMR) can be mapped to *:ARG0* in English (cf. italic line in Table 1). The rates for French, Turkish and Arabic are lower but still the most frequent map-

pings for *nsubj*.

6. Conclusion and perspectives

We presented a new dataset ParallelUD-AMR, an Abstract Meaning Annotation of the the English corpus of PUD. We provided an initial mapping of dependency relations with AMR relations (based on tokens and their alignments within the AMR

- ume 1: *Long Papers*), pages 397–407, Beijing, China. Association for Computational Linguistics.
- Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. [Abstract Meaning Representation for Sembanking](#). In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.
- Shu Cai and Kevin Knight. 2013. [Smatch: an Evaluation Metric for Semantic Feature Structures](#). In *51st Annual Meeting of the Association for Computational Linguistics*, page 748–752, Sofia, Bulgaria. Association for Computational Linguistics.
- Andrew Drozdov, Jiawei Zhou, Radu Florian, Andrew McCallum, Tahira Naseem, Yoon Kim, and Ramón Astudillo. 2022. [Inducing and using alignments for transition-based AMR parsing](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1086–1098, Seattle, United States. Association for Computational Linguistics.
- Federica Gamba, Alexis Palmer, and Daniel Zeman. 2025. [Bootstrapping UMRs from Universal Dependencies for Scalable Multilingual Annotation](#). In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 126–136, Vienna, Austria. Association for Computational Linguistics.
- Kelvin Han and Siyana Pavlova. 2019. [Going from UD towards AMR](#). Technical report, Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503.
- Johannes Heinecke. 2023. [metAMoRphosED: a graphical editor for Abstract Meaning Representation](#). In *19th Joint ACL - ISO Workshop on Interoperable Semantic Annotation at IWCS 2023*, Nancy. Tilburg University.
- Johannes Heinecke and Anastasia Shimorina. 2022. [Multilingual Abstract Meaning Representation for Celtic Languages](#). In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 1–6, Marseille. ELRA.
- Robert T. Kasper. 1989. [A flexible interface for linking applications to Penman’s sentence generator](#). In *Speech and Natural Language: Proceedings of a Workshop Held at Philadelphia, Pennsylvania, February 21-23, 1989*.
- Paul Kingsbury and Martha Palmer. 2002. [From TreeBank to PropBank](#). In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1989–1993, Las Palmas, Canary Islands - Spain. European Language Resources Association.
- Young-Suk Lee, Ramón Astudillo, Hoang Thanh Lam, Tahira Naseem, Radu Florian, and Salim Roukos. 2022. [Maximum Bayes Smatch ensemble distillation for AMR parsing](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5379–5392, Seattle, United States. Association for Computational Linguistics.
- Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. [Universal Dependencies](#). *Computational Linguistics*, 47(2):255–308.
- Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, Emily Pitler, and Leo Wanner. 2018. [The first multilingual surface realisation shared task \(SR’18\): Overview and evaluation results](#). In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 1–12, Melbourne, Australia. Association for Computational Linguistics.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher Manning, Sampo Pyysalo, Sebastian Schuster, Tyers Francis, and Daniel Zeman. 2020. [Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection](#). In *LREC 2020*, Marseille. ELRA.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. [The Proposition Bank: An Annotated Corpus of Semantic Roles](#). *Computational Linguistics*, 31(1):71–106.
- Michael Pust, Ulf Hermjakob, Kevin Knight, Daniel Marcu, and Jonathan May. 2015. [Parsing English into Abstract Meaning Representation Using Syntax-Based Machine Translation](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1143–1154, Lisbon, Portugal. 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*.

- Sebastian Schuster and Christopher D. Manning. 2016. [Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks](#). In *the tenth international conference on Language Resources and Evaluation*, pages 2371–2378, Portorož, Slovenia. European Language Resources Association.
- 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. 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 55th Annual Meeting of the Association for Computational Linguistics*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.
- Ida Szubert, Adam Lopez, and Nathan Schneider. 2018. [A structured syntax-semantics interface for English-AMR alignment](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1169–1180, New Orleans, Louisiana. Association for Computational Linguistics.
- Jens E. L. Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O’Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James H. Martin, Stephan Oepen, Martha Palmer, James James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a Uniform Meaning Representation for Natural Language Processing. *Künstliche Intelligenz*, 35:343–360.
- Chuan Wang, Nianwen Xue, and Sameer Pradhan. 2015. [A transition-based algorithm for amr parsing](#). In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–375, Denver, Colorado. Association for Computational Linguistics.
- Yanshan Wang, Sijia Liu, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Fei Liu, and Hongfang Liu. 2017. [Dependency and amr embeddings for drug-drug interaction extraction from biomedical literature](#). In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, ACM-BCB ’17*, page 36–43, New York, NY, USA. Association for Computing Machinery.
- Shira Wein and Nathan Schneider. 2024. [Assessing the Cross-linguistic Utility of Abstract Meaning Representation](#). *Computational Linguistics*, 50(2):419–473.

9. Language Resource References

- Damonte, Marco and Cohen, Shay. 2020. *AAbstract Meaning Representation 2.0 - Four Translations*. Linguistic Data Consortium. distributed via LDC: LDC2020T07, ISLRN [359-968-732-813-3](#). PID <https://catalog.ldc.upenn.edu/LDC2020T07>.
- Kevin Knight and Bianca Badarau and Laura Baranescu and Claire Bonial and Madalina Bardocz and Kira Griffitt and Ulf Hermjakob and Daniel Marcu and Martha Palmer and Tim O’Gorman and Nathan Schneider. 2020. *Abstract Meaning Representation (AMR) Annotation Release 3.0*. Linguistic Data Consortium. distributed via LDC: LDC2020T02, 3.0, ISLRN [676-697-177-821-8](#). PID <https://catalog.ldc.upenn.edu/LDC2020T02>.
- Zeman, Daniel et al. 2025. *Universal Dependencies 2.16*. PID <http://hdl.handle.net/11234/1-5901>. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Extending Retag to Conversion Error Detection: A Case Study on SynTagRus Morphology

Andrei Movsesian¹, Daniil Timchenko²

¹Institute for Information Transmission Problems (IITP RAS)

²Moscow Institute of Physics and Technology (MIPT)

Correspondence: Andrei Movsesian (andrei.movsesian@gmail.com)

Abstract

Linguistically annotated corpora are often converted between annotation schemes, but errors introduced during conversion can compromise their reliability. While annotation error detection is a well-studied topic, conversion error detection remains largely unexplored. We adapt the Retag method, which is traditionally used for finding annotation errors, to identify conversion errors by comparing model performance on original and converted versions of the same corpus, aligned at the token level. Applying this approach to the SynTagRus corpus converted to Universal Dependencies, we achieve high-precision detection of conversion errors in morphological annotation. Our analysis reveals systematic errors in distinction of auxiliary verbs, pronouns, numerals, and multi-word named entities, and uncovers previously undocumented annotation inconsistencies between different sections of the corpus. The method can be applied to any converted dataset for which an aligned source is available, providing an efficient way to target conversion errors for manual correction without exhaustive inspection.

Keywords: conversion error detection, annotation error detection, Retag, Universal Dependencies, SynTagRus, morphology, sequence tagging

1. Introduction

When converting a corpus from one annotation scheme to another, such as when creating Universal Dependencies (UD) resources from existing treebanks (de Marneffe et al., 2021), annotation errors can arise from two distinct sources. First, inaccuracies may be inherited from the source annotation. Second, new errors can be introduced by the conversion process itself.

For teams developing and maintaining converted corpora, distinguishing between these two error types is a practical necessity because they demand different responses. A conversion error points to a flaw in the conversion algorithm itself; fixing it once improves all future conversions. As for the source-side errors, the best course of action is to trace the error to its source and, when possible, notify the original maintainers. Fixing the error upstream ensures that future conversions inherit the correction automatically.

While the conceptual distinction between source and conversion errors is straightforward, isolating them in practice is not. Standard methods for evaluating converted data can reveal that errors exist, but they struggle to distinguish between different error types. Annotation error detection (AED) methods based on machine learning models (Klie et al., 2023) are especially problematic in this context: when applied to converted corpora, they conflate source errors, conversion errors, and the model's own prediction inaccuracies, producing a noisy signal that hinders targeted debugging and improve-

ment.

In this paper, we address the specific problem of isolating conversion errors. We propose a simple but effective method that leverages the original annotations as a direct point of comparison. For that we align the source and converted corpora at the token level and run the same AED model on both corpora. By examining cases where their annotations diverge, we can identify discrepancies where a conversion error is the most probable explanation.

We demonstrate this method on the conversion of the SynTagRus corpus (Boguslavsky et al., 2024) to UD, focusing on morphological annotations. Our case study shows that this approach successfully pinpoints conversion errors that would otherwise be buried in the output of standard AED, enabling their systematic analysis and correction.

2. Related Work

Work on evaluating converted corpora can be grouped into three broad approaches, each with different goals and limitations.

Manual inspection. The most direct approach to conversion quality assessment is manual annotation of a sample. For example, Borges Völker et al. (2019) manually annotated 50 sentences of the Hamburg Dependency Treebank according to UD guidelines and compared them with automatic conversion output. Cecchini et al. (2018) manually annotated 994 sentences of the Index Thomisticus Treebank to compare two conversion algorithms.

While such efforts can estimate overall conversion quality and identify some errors, they are labor-intensive and do not scale to corpus-level debugging. More fundamentally, manual inspection alone cannot systematically distinguish whether a discovered error originated in the source or was introduced during conversion.

Downstream task evaluation. A common proxy for annotation quality is to measure performance on tasks that depend on the annotation, such as parsing (Johannsen et al., 2015; Türk et al., 2019), morphological tagging (Pyysalo et al., 2015; McCarthy et al., 2018), or language modeling (Akkurt et al., 2024). Improvements in downstream performance are taken as evidence of improved annotation quality. This approach is useful for tracking progress across corpus versions, but it provides only an aggregate signal and do not produce interpretable error diagnoses at the instance level. Moreover, the drop in model’s performance does not necessarily mean that there are errors. For example, the annotation could simply become more difficult for the model to analyze.

Cross-corpus comparison. A third line of work compares annotations across different corpora that follow the same guidelines. Droganova et al. (2018) compared syntactic annotations across Russian UD corpora (including SynTagRus), proposing to analyze differences in parser attachment scores to discriminate between parser errors and genuine annotation inconsistencies. Aggarwal and Zeman (2020) introduced a symmetric measure for comparing part of speech (PoS) annotation consistency between corpora and demonstrated its use in identifying specific annotation differences. While these methods are effective, they are designed for comparing independently annotated corpora. In the conversion setting, however, we have an additional resource: the source corpus itself, which allows us to condition our analysis on the original annotation when isolating conversion-induced changes.

Like downstream evaluation approaches, we use a model (an AED classifier) to surface potential errors. But unlike prior work, we apply the same model to both the source and converted corpora and compare its outputs. This comparison cancels out the model’s own prediction errors, revealing discrepancies that are attributable to the conversion process.

3. SynTagRus Alignment

3.1. Text Alignment

SynTagRus (Boguslavsky et al., 2024) is a fully disambiguated corpus of Russian featuring multiple annotation layers, including morphology and syntax. Its development began in 1998 and as of

2026 it consists of more than 111,000 sentences (more than 1,585,000 words excluding punctuation marks) from 1392 texts across various genres. These genres include 20th-21st century fiction, contemporary popular science literature, opinion journalism, biographies, and news feeds. SynTagRus is an autonomous part of the Russian National Corpus, with search functionality available via its online interface.¹ For the purposes of conversion error detection, the corpus data was converted from its native XML format into the CoNLL-U format. This conversion was purely structural and did not alter morphological annotation.

The initial conversion of SynTagRus to the Universal Dependencies format was developed in 2016 as part of the UD v1.3 release (Droganova and Zeman, 2016). This first version contained texts annotated up to 2014, along with two texts annotated in 2015 (more than 61,000 sentences in total). As the UD guidelines evolved, the corpus annotation was subsequently corrected.

With the release of UD v2.9 in 2021, portions of SynTagRus annotated up to 2020 were incorporated, providing additionally over 25,000 sentences. However, the older segments of the corpus were not reconverted so changes in the original SynTagRus annotation of old texts were not reflected in the UD version. To align the two resources, we therefore compiled two distinct versions of SynTagRus (from 2016 and 2021) and aligned the texts with the UD version (SynTagRus-UD, v2.16).

3.2. Token Alignment

After aligning the texts, the number of texts and sentences in SynTagRus and SynTagRus-UD became identical. However, the number of tokens differed due to differences in annotation principles. To align them, we adopted the notion of tokens as defined in UD and adjusted the annotation of SynTagRus accordingly. The main differences in tokenization are as follows:

1. **Punctuation marks.** UD treats each punctuation mark as a token and assigns a special PUNCT tag to them, while SynTagRus do not annotate them. For alignment we copy the punctuation marks annotation from SynTagRus-UD to SynTagRus.
2. **Multi-word expressions (MWEs).** UD guidelines prohibit spaces within a token’s form, while SynTagRus contains numerous MWEs (e.g., *kak by to ni bylo* ‘anyhow’, *ni razu* ‘never’). During the initial conversion to UD, these MWEs were split into separate tokens using specially devised rules. Subsequently, in 2022,

¹<https://ruscorpora.ru/en/search/syntax>

similar splitting rules were created for the original SynTagRus annotation to enhance the search functionality of the Russian National Corpus (Boguslavsky et al., 2024) and were applied during our alignment process.

3. **Syntactic agglomerates** such as *nečego* ‘nothing’ and *negde* ‘nowhere’. This case is the opposite of MWEs: in SynTagRus, such tokens are split into two for syntactic reasons. For example, *negde* is annotated as *ne* ‘no’ and *gde* ‘where’. During alignment, we merged these token pairs and assigned the morphological features of the right-hand token.
4. **Ellipsis**. In SynTagRus, omitted words are restored in syntactically disconnected elliptical sentences. They function as normal nodes in the syntactic structure, but their word form is left empty. In SynTagRus-UD, this information is not present in the basic structure. However, since our focus is on morphological annotation, we do not need to align the elided tokens, and therefore they were omitted.
5. **Other differences**. The original conversion process introduced some errors into SynTagRus-UD (e.g., token duplication and using full form in abbreviation). For the purpose of strict token-by-token alignment, these errors were replicated semi-automatically in the SynTagRus data.

Figure 1 shows the main alignment operations. In the end, the aligned version of SynTagRus contained exactly the same number of sentences and tokens as SynTagRus-UD from UD v2.16.

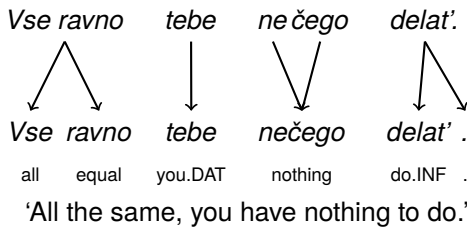


Figure 1: Token alignment between SynTagRus (top) and SynTagRus-UD (bottom), showing three main operations: MWE split, token merge, and punctuation addition.

4. Method

4.1. Conversion Error Detection

Our approach to conversion error detection extends the Retag method for annotation error detection (van Halteren, 2000; Klie et al., 2023). While

Retag identifies potential annotation errors by comparing original annotations with model predictions on a single dataset, we adapt this methodology to detect errors introduced during dataset conversion.

Given a source dataset src and its converted version cnv , we train two independent taggers θ_{src} and θ_{cnv} on src and cnv respectively. This yields four annotations for each token t : the original annotation y_t^{src} , the converted annotation y_t^{cnv} , and the model predictions $\hat{y}_t^{src} = \theta_{src}(t)$ and $\hat{y}_t^{cnv} = \theta_{cnv}(t)$. We categorize each token into one of four cases based on agreement patterns:

1. **Full agreement:** $y_t^{src} = \hat{y}_t^{src}$ and $y_t^{cnv} = \hat{y}_t^{cnv}$. This suggests both the original annotation and conversion are correct.
2. **Potential conversion error:** $y_t^{src} = \hat{y}_t^{src}$ but $y_t^{cnv} \neq \hat{y}_t^{cnv}$. The model agrees with the original annotation but disagrees with the converted version, indicating a likely error introduced during conversion.
3. **Source-side error:** $y_t^{src} \neq \hat{y}_t^{src}$ but $y_t^{cnv} = \hat{y}_t^{cnv}$. This may occur if: (1) the conversion makes the annotation task easier for the model, or (2) an annotation error in src was manually corrected in cnv .
4. **Noisy error:** $y_t^{src} \neq \hat{y}_t^{src}$ and $y_t^{cnv} \neq \hat{y}_t^{cnv}$. Both models disagree with their respective annotations, suggesting inherent ambiguity, source-side error, or otherwise difficult token for the models to process. These provide no reliable signal about conversion quality.

Focusing on case 2 allows us to obtain a high-precision candidate set for manual inspection, reducing the search space for conversion errors.

4.2. Conversion Errors and Data Inconsistency

The SynTagRus-UD corpus consists of two subcorpora converted at different times: an older subcorpus and a newer subcorpus processed with an updated conversion script. Although both subcorpora underwent manual validation after conversion, numerous inconsistencies between them remain unresolved. These residual inconsistencies dominate the error landscape, making it difficult to identify genuine conversion errors (Timchenko and Movsesyan, 2024).

To distinguish between these inconsistencies and actual errors introduced by the conversion process we stratify our analysis by creating three variants: **full** (the entire corpus), **new** (the newer subcorpus only), and **old** (the older subcorpus only). For a given triple of annotations $(y_t^{src}, y_t^{cnv}, \hat{y}_t^{cnv})$, we count how frequently it is flagged as a potential

conversion error in each variant. The resulting distribution helps to identify the underlying nature of the inconsistency:

1. **Data inconsistency:** the triple appears as an error only in the full variant, indicating that the new and old variants use different but internally consistent annotation conventions. When merged, these differences surface as errors. We analyze these errors for two reasons. First, old and new variants ideally should follow the same annotation guidelines so their divergence is itself a conversion issue. Second, conditioning on *src* and the ability to train the model on the full variant reduces noise compared to standard Retag approaches (e.g., using θ_{cnn}^{old} to annotate cnn^{new} or vice versa), providing a cleaner detection of annotation inconsistencies.
2. **Conversion error:** the error appears in the full variant and at least one of the old/new variants. We further distinguish two subtypes:
 - **Genuine conversion error**, caused by a bug in the conversion script. The error may appear in the old variant only, the new variant only, or both.
 - **Conversion-induced model error:** the conversion script is correct, but the converted corpus introduces new features not present in *src*. Unlike noisy errors (where $y_t^{src} \neq \hat{y}_t^{src}$), here θ_{src} predicts the tag correctly while θ_{cnn} fails. These errors are systematic and interpretable: a linguist can trace them to specific differences in annotation conventions. Since the inventory of PoS tags and morphological features for Russian is almost identical between UD v1 (old) and UD v2 (new), conversion-induced model errors typically appear in all three variants (old, new, and full).

4.3. Training and Evaluation

For training the models we chose UDPipe 2.1 (Straka, 2018), a widely used tool for processing Universal Dependencies data. We trained separate UDPipe models with default hyperparameters for each of the six subcorpora (three variants for both source and converted datasets), using the predefined train/dev/test splits provided in the UD release. Since each variant in SynTagRus-UD already follows the standard 80/10/10 split, no resplitting was necessary. Table 1 shows the statistics for all splits and variants.

Since our focus is exclusively on morphological annotation, we disabled all other UDPipe components (lemmatization and parsing) during training.

Table 1: Corpus statistics by variant and split. Since we fully aligned SynTagRus and SynTagRus-UD, the statistics is correct for both of them.

Variant	Split	Sentences	Tokens
Old	Train	48,815	870,434
	Dev	6,584	118,484
	Test	6,491	117,320
	Total	61,890	1,106,238
New	Train	20,816	334,123
	Dev	2,322	34,833
	Test	2,309	40,386
	Total	25,447	409,342
Full	Train	69,631	1,204,557
	Dev	8,906	153,317
	Test	8,800	157,706
	Total	87,337	1,515,580

All analysis is conducted on test set data, as taggers tend to overfit on training and development sets, potentially biasing the error signal. While the current evaluation serves to demonstrate the validity of our approach, exhaustive conversion error detection would employ cross-validation or repeated random subsampling to ensure every token is evaluated as part of some model’s test set, thereby achieving full corpus coverage.

The test set used is identical to the standard split in SynTagRus-UD. We did not perform any prior error analysis on this set before the study. Instead, our goal was to manually verify the errors identified by our algorithm and compare them against standard AED approach.

5. Results and Discussion

5.1. Models Performance

Table 2 shows accuracy metrics for all variants of SynTagRus and SynTagRus-UD calculated separately for parts of speech (UPOS), other morphological features (FEATS), and at tag level (UPOS + FEATS). The model trained on SynTagRus has the best performance on the full variant, while the model trained on SynTagRus-UD performs best on the old variant, so adding new data did not increase the performance, indicating data inconsistencies between the variants.

Table 3 shows how many errors correspond to each of the error type defined in section 4.1. The least frequent errors are source-side errors, and conversion error is the only error type that increases on the full variant, again suggesting data inconsistency. In the next sections we will only analyze potential conversion errors.

Table 2: Accuracy metrics for all variants of SynTagRus and SynTagRus-UD (%). Best results for each corpus are in bold.

Field	SynTagRus			SynTagRus-UD		
	Old	New	Full	Old	New	Full
UPOS	99.35	99.13	99.35	99.14	98.93	98.90
FEATS	98.03	97.54	98.15	97.50	96.87	96.52
UPOS + FEATS	97.79	97.19	97.90	97.30	96.60	96.26

Table 3: Error type distribution for each variant and each metric of (SynTagRus, SynTagRus-UD) pair (%).

Variant	Field	Error Type		
		Conversion	Source	Noisy
Old	POS	0.4	0.2	0.5
	FEATS	1.1	0.6	1.4
	POS+FEATS	1.1	0.6	1.6
New	POS	0.4	0.2	0.6
	FEATS	1.5	0.8	1.7
	POS+FEATS	1.5	0.9	1.9
Full	POS	0.6	0.2	0.5
	FEATS	2.2	0.5	1.3
	POS+FEATS	2.2	0.6	1.5

5.2. Errors in Parts of Speech

To detect potential conversion errors in parts of speech, we use the following procedure:

1. Collect $(y_{PoS}^{src}, y_{PoS}^{cnu}, \hat{y}_{PoS}^{cnu})$ triples as defined in section 4.1 for each variant of the (SynTagRus, SynTagRus-UD) pair.
2. Count how frequently each triple is flagged as a potential conversion error.
3. Exclude any triples for which there are no variants with absolute frequency of at least 10 and relative frequency of at least 2%. Here relative frequency is the count of conversion errors for $(y_{PoS}^{src}, y_{PoS}^{cnu}, \hat{y}_{PoS}^{cnu})$ divided by the total number of $(y_{PoS}^{src}, y_{PoS}^{cnu})$ pairs. We filter by both raw counts and relative frequencies to exclude rare triples: for example, if a given $(y_{PoS}^{src}, y_{PoS}^{cnu})$ pair appears only once in the test set and that token is misclassified, the relative frequency for the corresponding triple would be 100%, highlighting categories where the model's performance is least reliable.
4. Group the errors as defined in section 4.2 and analyze them manually.

The results of our analysis is present in Table 4. In the following sections we will analyze each conversion error type separately.

5.2.1. Data Inconsistency

Both cases of data inconsistency are similar and involve words *uže* ‘already’, *eše* ‘else/yet’, *počti* ‘almost’, *takže* ‘also’, and *čut* ‘hardly’. In SynTagRus they are all considered as particles. However, in the old variant of SynTagRus-UD they were all converted to adverbs, while in the new variant they are particles.

5.2.2. Genuine Conversion Errors

In SynTagRus, auxiliary verbs are marked as verbs. In that sense, the triple (V, AUX, VERB) can be attributed to conversion-induced model errors. However, additional ambiguity introduced by the annotation principles of UD does not mean that there are no errors. In SynTagRus-UD, only the lexeme *byt* ‘to be’ receives the POS tag AUX. At the same time, it can function as a main verb, and some of these cases are mislabeled, as in Example (1).

- (1) Kogda mne bylo sem'-vosem' let...
 *AUX
 ‘When I was seven or eight years old...’

The (A, PROP, ADJ) triple represents a case of erroneous annotation of multi-word named entities in the older version of SynTagRus-UD. According to UD guidelines, adjectives within such entities should not be marked as proper nouns. For example, in the multi-word expression *Plehanovskaja akademija* ‘Plekhanov academy’, the

Table 4: PoS conversion errors (counts with % in parentheses). Percentages are relative to the total number of tokens with given pair of PoS in two annotations. Model errors were separated from conversion errors after manual inspection.

y_{PoS}^{src}	y_{PoS}^{cnv}	$\hat{y}_{PoS}^{cnv}$	Old	New	Full	Total
Data inconsistency						
PART	PART	ADV	1 (0.0%)	0 (0.0%)	103 (2.1%)	104
PART	ADV	PART	0 (0.0%)	1 (25.0%)	68 (13.7%)	69
Genuine conversion errors						
V	AUX	VERB	32 (3.6%)	9 (2.2%)	56 (4.3%)	97
A	PROPN	ADJ	23 (19.5%)	0 (0.0%)	34 (28.8%)	57
NUM	NUM	ADJ	8 (0.4%)	1 (0.3%)	45 (2.0%)	54
S	DET	PRON	13 (1.5%)	2 (0.6%)	30 (2.4%)	45
NUM	ADJ	NUM	13 (2.1%)	0 (0.0%)	17 (2.6%)	30
Conversion-induced model errors						
S	PROPN	NOUN	39 (1.0%)	23 (2.1%)	55 (1.1%)	117
ADV	NUM	ADV	4 (9.5%)	3 (8.3%)	23 (29.5%)	30
S	PART	PRON	7 (12.1%)	4 (13.8%)	11 (12.6%)	22

first token should be marked as an adjective. This error was corrected in UD v2.17 by the maintainers of SynTagRus-UD.

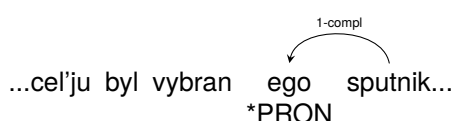
Errors in (NUM, NUM, ADJ) and (NUM, ADJ, NUM) triples are because of adjectival ordinal numerals. In SynTagRus all numerals which do not contain letters are numerals. In SynTagRus-UD these cases were converted to adjectives, but inconsistently. Most errors are in the old variant. Example (2) shows annotation of one such case, where both numerals should be adjectives.

- (2) ...načinaja s 1 sentjabrja 2006 goda.
*NUM ADJ

‘...starting from September 1, 2006.’

(S, DET, PRON) errors are due to incomplete heuristics used to disambiguate between substantive (PRON) and attributive (DET) pronouns. It commonly occurs in the old variant, but the errors in the new variant are similar. For example, the disambiguation is incorrect when attributive pronoun is a direct object of a noun, marked in SynTagRus by the first completive (1-compl) syntactic relation, as in Example (3).

- (3)



'...the target was its satellite...'

5.2.3. Conversion-induced Model Errors

These errors mostly occur in cases where annotation principles of SynTagRus-UD introduce additional ambiguity. (S, PROP, NOUN) errors do not occur in SynTagRus because proper nouns there are marked as nouns (S).

(ADV, NUM, ADV) errors occur in borderline cases with words like *mnogo* ‘many’ and *malo* ‘few’ which are ambiguous in SynTagRus-UD (cardinal numerals or adverbs) but marked as adverbs in SynTagRus.

(S, PART, PRON) errors occur in nominal sentences with zero copula, in which demonstrative *eto* 'this' is converted from noun to particle in SynTagRus-UD.

5.3. Errors Below the Threshold

To conserve space, we analyzed only the most frequent errors and excluded 73 additional error types, comprising 1018 erroneous tokens in total across the three variants. Most of these errors are duplicates of the frequent ones (9 types, 577 tokens). For example, the triple (V, AUX, VERB) has a less frequent counterpart (V, VERB, AUX) with the same underlying explanation.

Another common category consists of noisy errors in ambiguous cases (44 types, 378 tokens), where the annotation in SynTagRus and SynTagRus-UD is identical, but the model prediction differs. One example is the triple (A, ADJ, VERB), where the choice between a short adjective and a past participle sharing the same form (e.g., *urezannom* ‘truncated’) is ambiguous in the absence of syntactic dependents. While each of the 43 triples occurs relatively rarely, these nevertheless represent noisy errors for which the pre-

diction of θ_{src} is actually correct. One possible explanation is the difference in feature distributions between the two corpora, though further investigation is needed for these cases, as well as for the remaining rare cases (20 types, 63 tokens).

5.4. Errors in Other Morphological Features

For other morphological features we follow the same procedure as for parts of speech, but with three modifications:

1. For simplicity, we analyzed only those morphological categories which are present in both SynTagRus and SynTagRus-UD: Animacy, Aspect, Case, Degree, Gender, Mood, Number, Person, and VerbForm.
2. Since different parts of speech have different morphological feature inventories, and UD-Pipe model predicts PoS and all other features as two separate tags, for each category we counted only those errors for which $y_{PoS}^{src} = \hat{y}_{PoS}^{src}$ and $y_{PoS}^{cnv} = \hat{y}_{PoS}^{cnv}$.
3. For better grouping of errors, instead of counting triples $(y_t^{src}, y_t^{cnv}, \hat{y}_t^{cnv})$, we included part of speech and counted triples $([y_{PoS}^{src}, y_t^{src}], [y_{PoS}^{cnv}, y_t^{cnv}], [\hat{y}_{PoS}^{cnv}, \hat{y}_t^{cnv}])$.

The counts for these errors are present in Table 5.

The triple (A, Inan; DET, Inan; DET, $\emptyset$) represents a data inconsistency error. In SynTagRus, demonstratives like *этот* ‘this’ are annotated as adjectives and receive animacy marking in the accusative case (e.g., animate accusative *этото* vs. inanimate accusative *этого*). When converting to SynTagRus-UD, these adjectives become determiners (DET), and the animacy feature is carried over from the original annotation. However, the old version of SynTagRus-UD fails to preserve animacy for some determiners due to a conversion error, resulting in missing animacy features ($\emptyset$).

The other three are model errors and are explained by the absence of proper nouns and determiners in SynTagRus.

5.5. Annotation Errors and Conversion Errors

Up until now we analyzed potential conversion errors. All of them were either genuine conversion errors or model errors introduced because of the conversion process (specific to SynTagRus-UD).

Potential conversion errors and noisy errors (as defined in section 4.1) together represent all possible annotation errors in SynTagRus-UD which can be obtained by a standard Retag method. Here

we will analyze them focusing on parts of speech to see if there are any benefits in analyzing potential conversion errors separately compared to standard Retag approach where noisy errors cannot be easily excluded. Since in Retag the source annotation is not available, we only have access to the $(y_{PoS}^{cnv}, \hat{y}_{PoS}^{cnv})$ pairs ($y_{PoS}^{cnv} \neq \hat{y}_{PoS}^{cnv}$). Table 6 shows these errors with absolute frequency of at least 10 and relative frequency of at least 2% in any of the variant.

All of these frequent errors were described earlier. (PROPN, NOUN) is a conversion-induced model error, (AUX, VERB) is a conversion error, while (PART, ADV) is a data inconsistency error. The other four conversion errors were not found.

In theory, this approach identifies all model errors. However, since it is not conditioned on the source annotation, conversion errors appear among the less frequent pairs and are mixed in with noisy errors. For example, the (NUM, ADJ, NUM) conversion error was relatively frequent when conditioned on correctly predicted PoS tag in SynTagRus. However, the number of tokens with the error type (ADJ, NUM) is negligible (0.13%) relative to the total number of adjectives in the corpus.

The same holds for absolute frequencies. For example, the (A, PROPN, ADJ) conversion error type consisted of 23 tokens in the older version, almost all of which were genuine conversion errors. In contrast, for the (PROPN, ADJ) pair there are 29 such tokens. Manual verification confirmed that 6 additional tokens mostly represent noisy errors, making error analysis more difficult since they must be filtered manually. This problem also occurs for morphological features, for example, in cases of ambiguity such as genitive versus accusative in Russian.

5.6. Correcting the Errors

While we are not the maintainers of SynTagRus-UD and have proposed a tool only for efficient detection of conversion errors, it is nevertheless important to discuss how such errors can be corrected, since the ultimate goal of any error detection method is to improve the dataset.

We believe that analyzing the old conversion script is not necessary; using the new script would automatically resolve errors related to the old conversion process, including data inconsistencies.

This leaves two error types that require further attention: (V, AUX, VERB) and (S, DET, PRON). Most (V, AUX, VERB) errors stem from incorrect syntactic annotation in SynTagRus. In SynTagRus-UD, the decision to mark the verb *быт* ‘to be’ as an auxiliary depends on the presence of a copular syntactic relation in SynTagRus in which the verb is the head. In many cases, this relation is marked erroneously.

Table 5: Conversion errors of morphological features other than PoS (counts with % in parentheses). Percentages are relative to the total number of tokens with given pair of features in two annotations. $\emptyset$ indicates that the feature is not present in the corresponding annotation. The feature names of SynTagRus besides PoS are displayed in UD style.

y_{PoS}^{src}, y_t^{src}	y_{PoS}^{cnv}, y_t^{cnv}	$\hat{y}_{PoS}^{cnv}, \hat{y}_t^{cnv}$	Old	New	Full	Total
Animacy errors						
A,Inan	DET,Inan	DET, $\emptyset$	2 (1.6%)	0 (0.0%)	105 (39.6%)	107
S,Anim	PROPN,Anim	PROPN,Inan	2 (0.1%)	38 (5.3%)	24 (0.9%)	64
S,Inan	PROPN,Inan	PROPN,Anim	6 (0.3%)	6 (2.7%)	10 (0.5%)	22
A,Neut	DET,Neut	DET,Masc	3 (0.6%)	10 (4.4%)	10 (1.4%)	23

Table 6: Annotation errors in SynTagRus-UD for parts of speech (counts with % in parentheses). Percentages are relative to the total number of tokens with given PoS in the annotation.

y_{PoS}^{cnv}	$\hat{y}_{PoS}^{cnv}$	Old	New	Full	Total
PROPN	NOUN	47 (1.1%)	23 (2.1%)	64 (1.2%)	134
PART	ADV	5 (0.2%)	5 (0.3%)	110 (2.2%)	120
AUX	VERB	32 (3.1%)	9 (2.0%)	56 (3.7%)	97

For the (S, DET, PRON) error type, the solution lies in more careful handling of attributive pronouns within the conversion script, as discussed in Section 5.2.2.

Correcting all errors would require three steps: fixing the conversion script, reporting the source-side errors to SynTagRus maintainers (if they have not yet been addressed), and reconverting the corpus using the updated resources.

However, implementing this scenario is not straightforward. First, the portions of SynTagRus used for conversion in 2016 and 2021 have since undergone substantial changes in the original data. Missing parts of texts and sentences have been added, and numerous manual corrections to the annotations have been made.

Second, the presence of source-side errors that do not appear in SynTagRus-UD indicates that some errors were corrected in the converted data rather than in the source. This complicates re-conversion, as applying it to the modern version of SynTagRus (or even the 2016 and 2021 snapshots) would require tracking and preserving the manual changes made to the current version of SynTagRus-UD, which poses a challenge.

Ideally, all corrections to the converted corpus should be implemented within the conversion script itself, allowing the development of SynTagRus and SynTagRus-UD to remain independent yet aligned. We acknowledge, however, that this is not always feasible.

6. Conclusion

We have introduced a simple but effective method for isolating conversion errors in annotated corpora.

Our key insight is to apply the same annotation error detection model to both the source and converted corpora and compare their outputs. Because the model’s prediction errors are largely consistent across both runs, discrepancies between the two AED outputs provide evidence of conversion errors. This turns the source corpus into a resource uniquely available in the conversion setting that prior work has not fully exploited.

We demonstrated the method on the conversion of SynTagRus to Universal Dependencies, focusing on morphological annotation. Our analysis revealed:

- Genuine conversion errors in the annotation of auxiliary verbs, pronouns, ordinal numerals, and multi-word named entities that survived manual validation and would otherwise require exhaustive inspection to locate.
- Data inconsistencies between variants, such as the treatment of adverbs vs. particles, which our stratified analysis successfully distinguished from conversion errors.
- Model errors introduced by the conversion process itself, where new annotation distinctions created ambiguity that the tagger could not resolve.

Crucially, we showed that standard Retag applied to the converted corpus alone would have missed many of these errors since noise in the annotation errors alters their frequency distribution. This confirms that our comparative approach provides information unavailable through conventional AED.

At the same time, our proof-of-concept has several limitations that point to directions for future research. First, we analyzed parts of speech and individual morphological features separately rather than full morphological tags; extending the method to joint prediction would provide a more holistic view. Second, we did not investigate source-side errors, though these could be valuable for upstream corpus maintenance. Third, we used Retag as our AED method, but our comparative framework is agnostic to the underlying error detection technique; exploring other AED methods within this framework could reveal whether certain detectors are better suited to the conversion setting. Finally, exhaustive error detection would require the comprehensive analysis of less frequent errors and cross-validation or repeated subsampling to achieve full corpus coverage; we leave this to future implementation.

Our method generalizes to any token-classification task and could extend to dependency parsing with appropriate relation alignment. For corpus developers, it offers a practical, interpretable method to systematically improve conversion quality without manual inspection of entire datasets.

7. Bibliographical References

- Akshay Aggarwal and Daniel Zeman. 2020. [Estimating POS annotation consistency of different treebanks in a language](#). In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 93–110, Düsseldorf, Germany. Association for Computational Linguistics.
- Furkan Akkurt, Onur Gungor, Büşra Marşan, Tunga Gungor, Balkiz Ozturk Basaran, Arzucan Özgür, and Susan Uskudarli. 2024. [Evaluating the quality of a corpus annotation scheme using pre-trained language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6504–6514, Torino, Italia. ELRA and ICCL.
- Igor M. Boguslavsky, Alexandra V. Chaga, Pavel V. Djachenko, Tatyana I. Frolova, Evgenia S. Inshakova, Leonid L. Iomdin, Alexander V. Lazurski, Leonid G. Mityushin, Andrey A. Movsesyan, Ivan P. Rygaev, Victor G. Sizov, and Svetlana P. Timoshenko. 2024. The current state of the SynTagRus corpus. *Proceedings of the V. V. Vinogradov Russian Language Institute*, 42(4):141–169.
- Emanuel Borges Völker, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. [HDT-UD: A very large Universal Dependencies treebank for German](#). In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 46–57, Paris, France. Association for Computational Linguistics.
- 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. 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.
- Kira Droganova, Olga Lyashevskaya, and Daniel Zeman. 2018. Data conversion and consistency of monolingual corpora: Russian ud treebanks. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, pages 53–66, Linköping, Sweden. Linköping University Electronic Press.
- Kira Droganova and Daniel Zeman. 2016. Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies. Technical report, Praha, Czechia.
- Anders Johannsen, Héctor Martínez Alonso, and Barbara Plank. 2015. Universal dependencies for danish. In *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*, 11–12 December 2015 Warsaw, Poland, pages 157–167.
- Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023. [Annotation error detection: Analyzing the past and present for a more coherent future](#). *Computational Linguistics*, 49(1):157–198.
- Arya D. McCarthy, Miikka Silfverberg, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2018. [Marrying Universal Dependencies and Universal Morphology](#). In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 91–101, Brussels, Belgium. Association for Computational Linguistics.
- Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. [Universal Dependencies for Finnish](#). In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 163–172, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

- 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.
- Daniils Timchenko and Andrey Movsesyan. 2024. Issledovanie kačestva konvertacii sintaksičeskoj razmetki korpusa SynTagRus v format Universal'nyh zavisimostej [A study on syntactic annotation conversion quality of the SynTagRus corpus to the Universal Dependencies format]. In *Sbornik trudov 48-j meždisciplinarnoj školy-konferencii IPPI RAN*, pages 329–340, Moscow.
- Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Balkız Öztürk Başaran, Tunga Güngör, and Arzuçan 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, Paris, France. Association for Computational Linguistics.
- Hans van Halteren. 2000. [The detection of inconsistency in manually tagged text](#). In *Proceedings of the COLING-2000 Workshop on Linguistically Interpreted Corpora*, pages 48–55, Centre Universitaire, Luxembourg. International Committee on Computational Linguistics.

Speech Act Constructions in Universal Dependencies

Joakim Nivre, William Croft, André V. Lopes Coneglian

Uppsala University, University of New Mexico, Federal University of Minas Gerais
joakim.nivre@lingfil.uu.se, wacroft@icloud.com, coneglian@ufmg.br

Abstract

Is the framework of Universal Dependencies (UD) compatible with findings from linguistic typology about constructions in the world's languages? To address this question, we need to systematically review how UD represents these constructions, and how it handles the range of morphosyntactic variation attested across languages. In this paper, we present the results of such a review focusing on speech act constructions. We find that UD currently lack mechanisms for systematically capturing speech act constructions and briefly discuss ways in which this can be remedied.

Keywords: Universal Dependencies, speech act constructions

1. Introduction

Universal Dependencies (UD) is an annotation framework for morphosyntax, designed to be applicable to all human languages and to enable 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 construction 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. For example, the property predication construction is defined as any construction in any language where a property concept is predicated of a reference, as in the English *he is tall*. The verbal copula strategy describes an object predication construction that involves a linking element in the form of a verb, like English *is*. Other common strategies in the world's language are the nonverbal copula strategy (with a nonverbal linking element) and the zero strategy (with no linking element) (Croft, 2022; Nivre et al., 2026).

In order to build a construction, we first need to check if the current UD guidelines cover a given construction and all its attested strategies and propose additions when needed. Moreover, in the same process, we

can assess to what extent the annotations of different strategies capture the fact that they are realizations of the same construction, which facilitates cross-linguistic comparisons. A further step would be to investigate how actual treebanks have dealt with the phenomena at hand, which is outside the scope of the current investigation.

In this paper, we review how UD handles speech act constructions, which are discussed in Chapter 12 of Croft (2022). This complements previous papers on reference and modification (Nivre and Croft, 2025) and verbal predication (Croft and Nivre, 2025), as well as concurrent papers on nonprototypical predication and nonpredicational clauses (Nivre et al., 2026) and on complex predicates (Coneglian et al., 2026).

2. Speech Act Constructions

Speech acts can in general be defined in terms of the response that a speaker wants or requires from an addressee. In this context, we are interested in grammatical constructions that encode speech act functions. Typological surveys have shown that there are three such constructions found in all or most of the world's languages: the declarative, the interrogative, and the imperative-hortative construction; a fourth construction, the exclamative construction, is less common as a distinct construction but still attested in many languages (Sadock and Zwicky, 1985; König and Siemund, 2007; Croft, 2022).

The *declarative* construction is the default type of speech act construction and simply asserts some propositional content, inviting the addressee to accept it. The *interrogative* construction, or question, requests some information from the addressee, while the *imperative-hortative* construction requests that an action be carried out. The *exclamative* construction, finally, expresses a strong emotional reaction to a propositional content. Since this construction is less commonly attested, and the strategies used to realize it very diverse, we will not discuss it further in this paper.

As already mentioned, the declarative construction can be regarded as the default construction and will therefore not be discussed in this paper, as it is covered in depth in previous papers on clause constructions (Croft and Nivre, 2025; Nivre et al., 2026). Before we turn to the other speech act constructions, however, we need to discuss *negation* in declarative constructions and the related notion of *polarity*.

2.1. Negation of Declaratives

Declarative constructions with positive polarity constitute the default speech act construction and are overwhelmingly zero coded in the world's languages. That is, there is no overt coding over and above the coding of the clause construction itself. Negative polarity, by contrast, is normally overtly coded using one of four strategies, illustrated in examples (1–4).

- (1) díayi yíire baka-rí-ti
dog 1SG-SPEC bite-NEG-EVID
'the dog didn't bite me'
- (2) à sɛ̀dà cébébé pà̀y
3SG.M know jackal NEG
'he didn't see the jackal'
- (3) e-n syö-nyt omena-a
NEG-1SG eat-PTCP apple-PRTT
'I didn't eat an apple'
- (4) e sega ni la'ò o Jone
3SG NEG COMP go ART John
'John is not going'

In (1), from Tuyuca (Barnes, 1994), negation is expressed as an overt bound morpheme

affixed on the verb. In (2), from Musgu (Meyer-Bahlburg, 1972; Dryer, 2005), it is instead realized as an uninflected independent word form. These are the two most common strategies in the world's languages. Less common are the strategies illustrated in (3), from Finnish (Sulkala and Karjalainen, 1992; Dryer, 2005), and in (4), from Boumaa Fijian (Dixon, 1988; Dryer, 2005), where the negation is realized as a verb. In the former case, the negative verb forms a complex predicate together with a participial form of the main verb; in the latter, the main verb is in a subordinate clause introduced by the complementizer *ni*, a strategy that can be paraphrased as 'it is not the case that John is going'.

2.2. Interrogatives

Interrogative constructions are commonly divided into three types depending on function: polarity questions (sometimes called yes/no-questions), information questions (sometimes called wh-questions), and alternative questions. We will focus on polarity questions and information questions, which have been the most studied typologically.

For polarity questions, the two most common strategies are a distinct intonation contour (almost always a rising one) and/or an overt interrogative particle (which may be realized as an independent word or as an affix on the verb). The use of a distinct word order pattern, familiar from Germanic languages, is a much less common strategy.

Information questions typically contain an interrogative pronoun or modifier for the entity to be identified in the answer to the question, and the use of a distinct intonation contour and/or an interrogative particle is much less common in this case. In a large minority of languages, the interrogative pronoun obligatorily occurs at the beginning of the sentence, but it is more common that it occurs in the position where the corresponding element would occur.

2.3. Imperative-Hortatives

Imperative-hortative constructions are a wider class than the traditional notion of imperatives and include cases like the following, where there is an appeal to the addressee to help

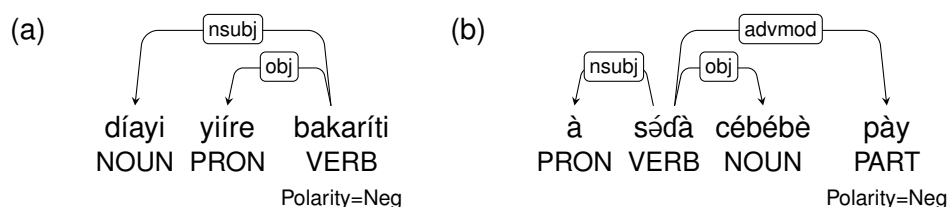


Figure 1: Simplified UD annotation of different strategies for negation of declaratives: (a) affix, (b) word.

bring about an event even though the event is not always an action specifically carried out by the addressee:

- (5) a. sing
b. let's sing
c. let her sing
d. let the party begin

The strategies used to realize imperative-hortative constructions are extremely varied and dependent on factors like person, politeness and social relations. The most distinctive strategy used, particularly in the second person, is reduction in form, and it is common to use the bare verb stem as in (5a) above.

3. UD Annotation

Let us begin by observing that the annotation of speech act types is essentially a blank spot in the UD framework. It is not discussed explicitly in the guidelines, and there is no mechanism for annotating this kind of information systematically. This can partly be seen as a consequence of the word-based nature of the UD annotation, which means that linguistic information has to be annotated either as properties of an individual word, through part-of-speech tags and morphological features, or as relations between pairs of words, through typed syntactic relations. Speech act types, however, are properties of entire clauses or sentences and often cannot be associated with individual words or pairs of words.

Technically speaking, given the hierarchical nature of the syntactic annotation, it is possible to annotate information about a larger unit on the head word of that unit (typically the predicate in the case of a clause). However, the scope of such an annotation will in principle

always be ambiguous, since there is no mechanism to indicate whether it applies only to the head word or to a larger segment (and if so, which segment). The idea of introducing special so-called “phrase level features” has been discussed on and off in the UD community, but so far no such concept has been introduced.

In the current UD guidelines, there are hence no morphological features or syntactic relations devoted specifically to the annotation of speech act constructions. However, information about such constructions is nevertheless available in a few cases where the strategy for realizing the construction involves a word or affix carrying a distinctive feature. In the following subsections, we will review the most important cases where this happens.

3.1. Negation of Declaratives

In version 1 of UD (Nivre et al., 2016), there was a syntactic relation *neg* reserved for negation, but not restricted to negation at the clause level. Hence, while the negation *not* would be attached with the *neg* relation to the head verb *know* in a declarative clause like *she does not know*, the same relation would be used to link the determiner *no* to the head noun *idea* in a nominal like *no idea*. The *neg* relation has been deprecated in version 2 of UD (Nivre et al., 2020) because it was deemed to encode a semantic property rather than a morphosyntactic relation.

In the current UD guidelines, the only way to annotate negative polarity is by using the morphological feature *Polarity=Neg*. This feature can be added both to words containing negation affixes, normally a verb or an auxiliary, and to words expressing negation by themselves. The former strategy is illustrated in Figure 1(a), which shows the UD annotation of example (1);

the latter strategy is illustrated in Figure 1(b), which is the UD annotation of example (2).¹ In many cases, it is therefore possible to determine whether a clause is negated by checking for the occurrence of the feature *Polarity=Neg* either on the main predicate of the clause or on a direct dependent of the predicate (usually with the relation *advmod*). However, such a procedure is only heuristic and will give the wrong result for sentences like *they did not only dance, they sang as well*, where the word *not* is attached to the verb *dance* and carries the feature *Polarity=Neg* but still does not negate the clause. For further discussion of negation in UD, see Findlay and Haug (2025).

3.2. Interrogatives

Whether interrogative constructions are captured in the UD annotation depends both on the type of question and on the strategy used to realize it. Starting with polarity questions, the prosody strategy is not captured at all, as UD does not annotate prosodic features, and neither is the less common word order strategy. The question particle strategy should in principle be within reach, since a morphological feature could be assigned to the particle, regardless of whether it is realized as an affix or as an independent word. However, even though the guidelines mention that question particles can be assigned the part-of-speech tag *PART*, there is currently no morphological feature available for classifying question particles. The only related feature is *Mood=Int*, which can be used for verbs that have a special form in interrogative clauses, a strategy attested in Turkic languages.

Moving on to information questions, these are generally captured more straightforwardly because they contain an interrogative pronoun, determiner or adverb that carries the feature *PronType=Int*. One complication is that the word carrying this feature may not be directly attached to the root of the clause. For example, in a sentence like *which student's mother's book are you looking for*, the interrogative de-

terminer *which* is at a distance of four dependency arcs from the verb *looking*. In addition, we need to distinguish direct questions from embedded interrogative clauses, which is in most cases possible by inspecting the position of the clause in the dependency tree.

3.3. Imperative-Hortatives

Of the many diverse strategies employed for imperative-hortative constructions, only those that involve verbal morphology are currently captured in UD through the morphological feature *Mood*. In addition to the very common value *Mood=Imp*, for imperative verb forms, the value *Mood=Jus* can be used for the jussive, which is found, for example, in Arabic.

4. Discussion

Speech act constructions like declaratives, interrogatives and imperative-hortatives, sometimes referred to as basic sentence types, are among the most general linguistic constructions and should arguably be captured in any framework for morphosyntactic annotation. It is noteworthy that UD currently lacks mechanisms for a systematic annotation of these constructions, and it is highly relevant to discuss ways in which this could be remedied.

In the short term, within version 2 of UD, the obvious solution is to add new features and feature values to describe at least the basic speech act constructions declarative, interrogative, imperative-hortative, and exclamative, as well as clause-level negation. For interrogative and imperative-hortative constructions, it may also be useful to distinguish subtypes. It is true that, in a word-based system like UD, the scope of these features will be ambiguous, but this seems preferable to not including the information at all.

In addition to adding new features, one may consider modifying the syntactic relations used to annotate strategies that involve a special word order. For example, in information questions where the interrogative pronoun or adverb obligatorily occurs at the beginning of the sentence, the *dislocated* relation could be used with a special subtype to capture the word order strategy.

¹In these examples, we simplify the UD representations by omitting (a) lemmas and (b) morphological features that are not relevant for the discussion. For reasons of space, we do not repeat glosses and translations in the examples.

In the longer term, in subsequent versions of UD, it may be worth introducing a mechanism for using features that apply to larger units than words. Such a mechanism would be useful not only for the annotation of speech act constructions, but more generally for categories that transcend word boundaries, for example, complex predicates.

5. Conclusion

In this paper, we have taken another step towards a constructicon for UD, in the sense of Nivre (2025), by reviewing the way UD annotates speech act constructions, following the taxonomy of Croft (2022), thereby extending the previous work on reference and modification (Nivre and Croft, 2025) and verbal predication (Croft and Nivre, 2025).² The main findings in this paper are negative, since UD currently lack mechanisms for systematically capturing speech act constructions. We have briefly discussed how this can be remedied both in the long and the short term.

6. Acknowledgments

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology), from the Swedish Research Council (grant no. 2022-02909), and from the Brazilian National Council for Scientific and Technological Development, CNPq (grant no. 153972/2025-4 and no. 405484/2025-9). We thank Leonie Weissweiler and three anonymous reviewers for comments and suggestions that helped improve the article.

7. Bibliographic References

Janet Barnes. 1994. Tuyuca. In Peter Kahrel and René van den Berg, editors, *Typological Studies in Negation*, pages 325–342. John Benjamins.

André V. Lopes Coneglian, Joakim Nivre, and William Croft. 2026. Complex predicates in

Universal Dependencies. In *Proceedings of the Ninth Workshop on Universal Dependencies*.

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

William Croft and Joakim Nivre. 2025. [Verbal predication constructions in Universal Dependencies](#). In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 50–60.

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

Robert M. W. Dixon. 1988. *A Grammar of Boumaa Fijian*. University of Chicago Press.

Matthew S. Dryer. 2005. Negative morphemes. In Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie, editors, *World Atlas of Language Structures*, pages 454–457. Oxford University Press.

Jamie Yates Findlay and Dag Trygve Truslew Haug. 2025. Negation in Universal Dependencies. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*.

Ekkehart König and Peter Siemund. 2007. Speech act distinctions in grammar. In Timothy Shope, editor, *Language Typology and Syntactic Description, 2nd Edition. Volume 1: Clause Structure*, pages 276–324. Cambridge University Press.

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.

Hilke Meyer-Bahlburg. 1972. *Studien zur Morphologie und Syntax des Musgu*. Helmut Buske.

²In concurrent papers, we review nonprototypical predication and nonpredicational clauses (Nivre et al., 2026) and complex predicates (Coneglian et al., 2026).

- 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 and William Croft. 2025. Reference and modification in Universal Dependencies. In *Proceedings of the 8th Workshop on Universal Dependencies (UDW)*, pages 1–10.
- Joakim Nivre, William Croft, and André Coneglian. 2026. Nonprototypical predication and nonpredicational clauses in Universal Dependencies. In *Proceedings of the Ninth Workshop on Universal Dependencies*.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silva, 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.
- Jerrold M. Sadock and A. Zwicky. 1985. Speech act distinctions in syntax. In Timothy Shope, editor, *Language Typology and Syntactic Description. Volume 1: Clause Structure*, pages 155–196. Cambridge University Press.
- Helen Sulkala and Merja Karjalainen. 1992. *Finnish*. Routledge.

Author Index

- Aissaoui, Amal, 258
Akkurt, Furkan, 97
Asahara, Masayuki, 1
- Barbu Mititelu, Verginica, 142
Battaglia, Elena, 60
Brosa-Rodriguez, Antoni, 174
Buhnila, Ioana, 142
- Chiarcos, Christian, 277
Chontaeva, Bermet, 97
Ciolaneanu, Roxana, 142
Çöltekin, Çağrı, 97
Coneglian, Andre, 208, 218, 315
Corbetta, Claudia, 229
Croft, William, 208, 218, 315
- Davidson, Diana, 128
de Marneffe, Marie-Catherine, 153
de Marneffe, Marie-Catherine, xiii
Dobrovoljc Zor, Kaja, 60
Dyer, Andrew, 229
Dzhumalieva, Gulnura, 97
- Eslami, Soudabeh, 97
- Faghiri, Pegah, 76
- Gasser, Michael, 54
Gerdes, Kim, 76, 163
Grabowski, Konrad, 128
Guillaume, Bruno, 60
- Heinecke, Johannes, 295
Henderson, Robert, 197
Herrera, Santiago, 183
- Irimia, Elena, 142
Ivanova, Sardana, 97
- Janssen, Maarten, 252
Jimenez Lopez, M. Dolores, 174
- Kahane, Sylvain, 60, 76, 183, 229
Kasieva, Aida, 97
Kifle, Nazareth Amlesom, 54
- Klironomou, Christina, 103
- Landwehr, Isabell, 128
Levshina, Natalia, 49
- Marais, Laurette, 116
Masciolini, Arianna, 103
Matsuda, Hiroshi, 1
Mauri, Caterina, 60
Mayhew, Stephen, xiii
Meiser, Magdalena, 128
Miranda Junior, Isaac Souza de, 153
Movsesian, Andrei, 305
Mus, Nikolett, 97, 229
- Nivre, Joakim, 208, 218, 315
- Oya, Masanori, 91
- Pagano, Adriana S., 142
Pannitto, Ludovica, 60
Pasparaki, Thelka, 103
Pretorius, Laurette, 116
Puddu, Nicoletta, 242
Pugh, Robert, 197
- Regnault, Mathilde, 289
- Sanguinetti, Manuela, 242
Schacht, Carmen, 128
Siewert, Janine, 277
Stemle, Egon, 268
- Talamo, Luigi, 242
Tantos, Alexandros, 103
Timchenko, Daniil, 305
Touriki, Despoina Ourania, 103
Tsiotskas, Konstantinos, 103
Tsourilla, Eleni, 103
Tyers, Francis, 197
- Vale, Oto Araújo, 153
- Washington, Jonathan N., 97
Wiedmann, Sophia, 128
- Yilandiloğlu, Enes, 109
- Zucchini, Eleonora, 60