

TLT 2024

**The 22nd Workshop on Treebanks and Linguistic Theories
(TLT 2024)**

Proceedings of the Conference

December 5-6, 2024

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

University of Hamburg, Indiana University, supported by the DAAD in the program University Partnerships with Eastern Europe with funds from the Federal Foreign Office of Germany, and the SFB 1102



Gefördert durch:



©2024 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN None

Introduction

The 22nd International Workshop on Treebanks and Linguistic Theories (TLT 2024) follows an annual series that started in 2002 in Sozopol, Bulgaria. TLT addresses all aspects of treebank design, development, and use. “Treebank” is taken in a broad sense, comprising any spoken, signed, or written data augmented with computationally processable annotations of linguistic structure at various levels.

For the first time, TLT is being hosted by the Hamburg KorpusLab on December 5-6, 2024 in Hamburg, Germany. The KorpusLab is a research group led by Heike Zinsmeister at the Institute for German Language and Literature (Institut für Germanistik) at the University of Hamburg. Information about the group’s research projects and other activities are collected on the KorpusLab website.

From the papers submitted to TLT 2024, we accepted 8 archival submissions as well as 1 non-archival submission. The papers range in topics from UD treebanks for new languages to coreference and information status annotations on top of UD annotations for the literary domain, and a novel approach to parsing dependencies using shallow information. For the first time, TLT offered the option of non-archival submissions.

When the call for the workshop was published, a member of the community expressed concerns about the relevance of linguistically annotated resources in the area of large language models (LLMs) and questioned the appropriateness of continuing research on creating and analyzing such resources. Addressing this concern, we organized a panel discussion on “Treebanks and linguistic annotation in the area of LLMs”.

Our sincere thanks go to everyone who is making this event possible: everybody who submitted papers; all the reviewers; our invited speakers and panelists, and the SFB 1102 at Saarland University for funding an invited speaker.

Finally, we would also like to thank ACL SIGPARSE for its endorsement and the ACL Anthology for publishing the proceedings.

Daniel Dakota, Sarah Jablotschkin, Sandra Kübler, Heike Zinsmeister (TLT2024 Chairs)
December 2024

Organizing Committee

TLT2024 Chairs

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

Program Committee

Reviewers

Ann Bies, University of Pennsylvania
Gosse Bouma, University of Groningen
Miriam Butt, Universität Konstanz
Éric Villemonte de la Clergerie, INRIA
Eva Hajicova, Charles University Prague
Lori Levin, Carnegie Mellon University
Wolfgang Menzel, University of Hamburg
Adam Meyers, New York University
Jiří Mírovský, Charles University Prague
Kaili Müürisep, University of Tartu
Joakim Nivre, Uppsala University
Petya Osenova, Bulgarian Academy of Sciences
Daniel Zeman, Charles University Prague

Keynote Talk: Multilingual Coreference and Treebanking: Benefits of Interaction

Anna Nedoluzhko
Charles University, Prague

Abstract: Several years ago, we created CorefUD, a harmonized collection of coreference datasets for multiple languages. This collection has grown steadily, with new languages and datasets added each year. Currently, CorefUD 1.2 includes 21 datasets across 15 languages. CorefUD is compatible with morphosyntactic annotations in the Universal Dependencies (UD) framework, highlighting the close relationship between two types of linguistic annotation: coreference and syntax. But how do these annotations interact? Do UD tree structures correspond to mention spans in coreference annotations? Are syntactic heads in UD equivalent to the head mentions in coreference annotation? Can reconstructed empty nodes in enhanced UD effectively align with zero anaphora? And how do zeros in coreference relate to syntactic structures across the diverse languages in the collection? In the talk, I will address these questions with a specific focus on zero anaphora which was the special topic of the recent CRAC shared task on multilingual coreference resolution.

Keynote Talk: Increasing Language Diversity in NLP

Marcel Bollmann
Linköping University

Abstract: Linguistic diversity in NLP remains an important challenge, with many languages lagging behind in terms of available data and resources for training and evaluation of NLP models. In this talk, I will present CreoleVal, a project aimed at providing an evaluation benchmark for several Creole languages. I will discuss why we chose to work on Creoles in particular, what kinds of data and annotations we produced for CreoleVal, and what challenges we encountered in the process. Finally, I will give an outlook on challenges around data and data annotation in the TrustLLM project, an ongoing EU-funded project on creating trustworthy LLMs for the Germanic languages.

Treebanks and Linguistic Annotation in the Area of LLMs

The panel discussed the impact Large Language Models (LLMs) have had on the current state of treebank design and development, as well as their continued impact on the future of the field. Topics considered included:

- Do LLMs make treebanks redundant?
- What can we learn from treebanks that we can't learn from LLMs?
- Is it still justified to spend money on creating and maintaining treebanks?

Invited Panel Members

Marcel Bollmann, Linköping University
Daniel Dakota, Indiana University
Anna Nedoluzhko, Charles University Prague
Sandra Kübler, Indiana University
Juri Opitz, University of Zurich

Non-Archival Abstracts

UD for German Poetry

Stefanie Dipper and Ronja Laarmann-Quant
Ruhr-Universität Bochum

This article deals with the syntactic analysis of German-language poetry from different centuries. We use Universal Dependencies (UD) as our syntactic framework. We discuss particular challenges of the poems in terms of tokenization, sentence boundary recognition and special syntactic constructions. Our annotated pilot corpus currently consists of 20 poems with a total of 2,162 tokens, which originate from the PoeTree.de corpus. We present some statistics on our annotations and also evaluate the automatic UD annotation from PoeTree.de using our annotations.

Table of Contents

<i>Developing the Egyptian-Ujaen Treebank</i> Roberto Antonio Díaz Hernández and Marco Carlo Passarotti	1
<i>Symmetric Dependency Structure of Coordination: Crosslinguistic Arguments from Dependency Length Minimization</i> Adam Przepiórkowski Przepiórkowski, Magdalena Borysiak, Adam Okrański, Bartosz Pobożniak, Wojciech Stempniak, Kamil Tomaszek and Adam Głowacki	11
<i>A First Look at the Ugaritic Poetic Text Corpus</i> Tillmann Döncke, Clemens Steinberger, Max-Ferdinand Zeterberg and Noah Krill	23
<i>LuxBank: The First Universal Dependency Treebank for Luxembourgish</i> Alistair Plum, Caroline Döhmer, Emilia Milano, Anne-Marie Lutgen and Christoph Purschke	30
<i>Building a Universal Dependencies Treebank for Georgian</i> Irina Lobzhanidze, Erekle Magradze, Svetlana Berikashvili, Anzor Gozalishvili and Tamar Jalaghonia	40
<i>Introducing Shallow Syntactic Information within the Graph-based Dependency Parsing</i> Nikolay Paev, Kiril Simov and Petya Osenova	46
<i>A Multilingual Parallel Corpus for Coreference Resolution and Information Status in the Literary Domain</i> Andrew Dyer, Ruveyda Betul Bahceci, Maryam Rajestari, Andreas Rouvalis, Aarushi Singhal, Yuliya Stodolinska, Syahidah Asma Umniyati and Helena Rodrigues Menezes de Oliveira Vaz	55
<i>Dependency Structure of Coordination in Head-final Languages: a Dependency-Length-Minimization-Based Study</i> Wojciech Stempniak	65

Developing the Egyptian-UJaen Treebank

Roberto Antonio Díaz Hernández,¹ Marco Carlo Passarotti²

¹ University of Jaén (radiaz@ujaen.es)

² Università Cattolica del Sacro Cuore (marco.passarotti@unicatt.it)

Abstract

This paper presents preliminary results of the development of the Egyptian-UJaen treebank, the first dependency treebank created for pre-Coptic Egyptian in Universal Dependencies. It describes the current state of the treebank, explains the approach adopted for the morphosyntactic annotation and discusses some issues concerning the adoption of the CoNLL-U format for the annotation of Egyptian texts. This treebank will surely become a useful linguistic tool for understanding the synchronic and diachronic use of pre-Coptic Egyptian.

1 Introduction

Over the last decade, there has been a growing interest in less-resourced languages that has led to a boom in treebanks for such languages in Universal Dependencies, a useful framework that provides a systematic annotation of grammar across languages (de Marneffe *et al.*, 2023).¹ The creation of the Egyptian-UJaen treebank (henceforth EUJA treebank) aims to contribute to the development of UD by applying the universal inventory of categories developed therein to the morphosyntactic annotation of Egyptian texts. It is the first dependency treebank² created for pre-Coptic Egyptian. Texts are annotated morphosyntactically at the University of Jaén, according to the structuralist approach to Egyptian philology (see Polotsky’s key works, 1944, 1976 and Schenkel’s, 2012).

The EUJA treebank started as UD release 2.14 on 15 May 2024 with 5,515 words and 707 sentences from Old Egyptian texts. The data and results of the present paper are based on the current

state of the treebank consisting of 1,573 sentences and 14,650 words (UD release 2.15 to appear on 15 November 2024).

The aim of this paper is to describe the methodology used in the development of the EUJA treebank. It provides a brief overview of Egyptian and its scripts (2) and a description of the sources selected for the treebank (3). There follows a discussion on the annotation of Egyptian texts (4) and the evaluation of an NLP model trained on the treebank (5). Finally, the next stages of the development of the EUJA treebank are outlined in the conclusion (6).

2 Egyptian language and scripts

Egyptian is an Afroasiatic language that knew the following stages:

- 1) Old Egyptian (ca. 2700–2000 BC).
- 2) Middle Egyptian (ca. 2000–1550 BC).
- 3) Late Egyptian (ca. 1550–700 BC).
- 4) Demotic (7th century BC to 5th century AD).
- 5) Coptic (4th century to 14th century AD).

These stages can be classified into Earlier Egyptian, which includes Old Egyptian and Middle Egyptian, and Later Egyptian, which includes Late Egyptian, Demotic and Coptic. While the syntax of Earlier Egyptian is mainly synthetic, Later Egyptian is characterised by an analytic syntax. It should be noted that in the Middle Kingdom (ca. 1980–1760 BC) Old Egyptian was used as a sacred language for the transmission of the Pyramid Texts, even though Middle Egyptian was spoken, while Middle Egyptian became a standardised classical language from the 18th Dynasty (ca. 1539–

¹ <https://universaldependencies.org/>

² For the Coptic treebank in UD see Zeldes and Abrams (2018).

1292 BC) onwards, when other stages of Egyptian were spoken.

Different scripts were used for Egyptian. Hieroglyphs were usually the monumental script for Old Egyptian, Middle Egyptian and eventually Late Egyptian. Hieratic script was mainly used for documents, letters and copies of religious and literary texts in Old Egyptian, Middle Egyptian and Late Egyptian. This script was used exceptionally in monuments and steles. The hieroglyphic and hieratic scripts evolved throughout history, for example the Old Kingdom (ca. 2543–2436 BC) hieroglyphic and hieratic scripts are both different from those used in the New Kingdom (ca. 1539–1077 BC). Finally, Demotic and Coptic were written in Demotic and Coptic script respectively.

The EUJA treebank annotates Egyptian texts using the Tübingen transcription system (see 4.1, below). Hieroglyphs of Old Egyptian, Middle Egyptian and Late Egyptian texts are written in the MISC column (see 4.7, below). The same is planned for Demotic signs. Hieratic texts are transliterated into hieroglyphic script.

3 Sources

Egyptology or rather Egyptian philology is the discipline that studies Egyptian texts. Its official beginning dates back two centuries ago when Jean François Champollion deciphered hieroglyphs in 1822, not without the help of Thomas Young’s earlier attempts. Plenty of textual sources make Egyptian a well-documented ancient language, comparable to Akkadian, Ancient Greek or Latin. Considering such richness it is regrettable that only a handful of universities in the world offer the possibility of studying Egyptology as a fully official degree.

The amount of textual sources for Egyptian depends on their state of preservation. As a rule, the younger the linguistic stage, the more sources there are—Old Egyptian sources are scarcer than those of Middle Egyptian and Late Egyptian. An exception to this is the number of texts written in Classical Egyptian, for it is much larger than for Late Egyptian. Since the aim of the EUJA treebank is to provide a linguistic resource for the morphosyntac-

tic study of pre-Coptic Egyptian, it purports to contain the most representative texts of each stage, namely:

- 1) Old Egyptian: The Pyramid Texts (PT, Sethe, 1908–1922), Old Kingdom and First Intermediate biographical texts (Sethe, 1933 and Clère/Vandier, 1948).
- 2) Middle Egyptian: The Coffin Texts (CT, de Buck, 1935–1961), Middle Kingdom biographical texts (Lange/Schäfer, 1902–1925) and literary texts, such as Sinuhe (Koch, 1990) and the Eloquent Peasant (Parkinson, 1991).
- 3) Classical Egyptian: The Book of the Dead (BD, Naville, 1886), 18th Dynasty biographical texts (Sethe, 1906–1909 and Helck, 1955–1958) and literary texts, such as Neferti (Helck, 1970) and Ipuwer (Enmarch, 2008)
- 4) Late Egyptian: New Kingdom biographical texts (Kitchen, 1975–1990) and literary texts (Gardiner, 1932).
- 5) Demotic: Literary texts, such as the teaching of Onchsheshonqy (Glanville, 1955).

Several editions of these Egyptian texts were published in the first half of the twentieth century and are now available online as pdf files, such as the Coffin Texts.³ As the linguistic usage of Egyptian varies not only in sources from different stages, but also in some sources from the same period, each sentence in the EUJA treebank is assigned a bibliographic reference and an ID in order to identify and classify all sentences by source. The ID consists of the acronym EUJA, followed by a hyphen and a numeral, for example EUJA-1. Each sentence is also provided with a reference indicating the exact paragraph in the original text, its origin, date, the genre and source’s language stage, for example:

sent-id =EUJA-44

ref = PT § 1a T, Saqqara, 6th Dynasty, rel, OE⁴

EUJA-1 is a test sentence. EUJA-2 to 43 are multiword expressions taken from various Old Egyptian text corpora.

The systematic annotation of the Pyramid Texts begins with EUJA-44. 14,404 words, correspond-

³ <https://isac.uchicago.edu/research/publications/oriental-institute-publications-oip> (OIP 34, 49, 64, 67, 73, 81, 87 and 132).

⁴ The abbreviation “rel” stands for religious text, and “OE” for Old Egyptian.

ing to 1/5 of the whole corpus, have been annotated in the treebank, which means that the Pyramid Texts consist of 60,000–80,000 words.

From the beginning of the EUJA treebank the question whether to create a repository for each stage or for all stages of pre-Coptic Egyptian was discussed with Daniel Zeman.⁵ The latter option was chosen in order to have an overview of the evolution of Egyptian in a single CoNLL-U file.




If particular linguistic features of a text corpus or a stage are to be studied, the corpus name, e.g. Pyramid Texts, or the stage name, e.g. OE (Old Egyptian), should be mined to find all instances that match the search. The README file also contains a classification of the sentences according to the stage of Egyptian and the text corpus in order to facilitate searching:

sent_id = EUJA-	language and text corpus
1–1573	Old Egyptian
1, 3, 4, 11–15, 23, 25, 26, 30–34, 36–40, 43	biographical texts
2, 6, 7, 9, 10, 16–21, 24, 27–29, 41, 42, 44–1573	Pyramid Texts
5, 35	Letters to the Dead
8, 22	Captions to everyday life scenes

Table 1: Sentence classification in the treebank

4 Annotation

4.1 Transcription

Egyptian scripts consist of phonetic signs and classifiers. Phonetic signs are reproduced by means of transcription characters to make reading easier. Classifiers are signs that give information about a word (Goldwasser, 2022: 192), for example ⁶ is a classifier in the word   *ht(i)* “travel downstream”.

Some colleagues attending the 13th International Conference of Egyptologists held in Leiden from 6 to 11 August 2023 established the Leiden Unified Transliteration (LUT),⁷ and there has been constant pressure since then to adopt it for the transcription of Egyptian texts in digital resources, text editions and publications. However, the LUT is clearly a scientific regression, as it keeps traditional

signs, such as *t̄* and *d̄*, which were adopted in the 19th century only for typographical reasons.

The Tübingen transcription system (Schenkel, 2012: 19–25; Schneider, 2023: 4)⁸ has been followed for the annotation of Egyptian texts in the EUJA treebank, for its suitability for linguistic analysis; for example, as in the Slavic languages *č* stands for the sound /tʃ/, whereas the LUT’s *t̄* may confuse a linguist for it is used to transcribe /θ/ (ث) in Arabic. A table with both systems and the Unicode codes used for the transcription signs in the EUJA treebank is included in the appendix (see table 3, below).

As is usual with sources of extinct languages, Egyptian texts occasionally contain gaps and errors, which must be noted in their transcription. The Leiden system for editing ancient texts is consequently used in the EUJA treebank (Schenkel, 2012: 28–29). It includes the following critical signs:

1. Brackets () add a conventionally omitted element, for example the suffix pronoun of the first person singular *ʾt* is usually omitted in Old Egyptian as vowels or weak consonants such as *ʾt* were not written.
2. Square brackets [] enclose a restored text that was missing.
3. Curly brackets {} enclose typographical errors, for example the reduplication of a consonant (i.e. dittography). Such errors are labelled as Typo according to the CoNLL-U format in the EUJA treebank.
4. Angle brackets <> add an element that has been erroneously omitted from the text, for example a missing consonant due to haplography.

4.2 Sentence splitting

It is generally assumed that no punctuation marks are used in Old Egyptian and Middle Egyptian. However, the annotation of the Pyramid Texts in the EUJA treebank has revealed that a line is occasionally used as a punctuation mark (see fig. 1, below) to indicate the end of a spell (e.g. EUJA-1309) or to separate a recitation text from a ritual remark (e.g. EUJA-178). The line in the hieroglyphic text is transcribed by means of the vertical bar (|).

⁵ Thanks to Daniel Zeman for his support.

⁶ The hieroglyphs used in this paper are drawn from the hieroglyphic text processor JSesh.

⁷ <https://www.iae-egyptology.org/the-leiden-unified-transliteration>

⁸ See also Rössler, 1971: 263–326.

bank, the lemmata of derivatives, such as nisba adjectives, are the words from which they are derived, for example the lemma of the nisba adjective¹¹ *im.t* “one who is in” is the preposition *m* “in”. Likewise, participles, relative forms and infinitives are lemmatised after the verb stem, for example the passive participle *mr.y* “beloved” corresponds to the lemma *mri* “love”. Causative verbs are also lemmatised after the verb stem without the causative prefix, for example *š:w'b* “make pure” (i.e. “purify”, “cleanse”) corresponds to the lemma *w'b* “be pure”.

4.5 Universal Part-of-Speech tags and Morphological analysis

Fifteen Universal Part-of-Speech tags (cf. Petrov *et al.*, 2012) are documented in Old Egyptian according to the current state of the EUJA treebank.¹²

- 1) Adjective (ADJ; 528/3.60%): There are a few primary adjectives, for example *nb* “every”, “all”. Most of them are deverbal adjectives such as *nfr* “be good” and nisba adjectives such as *im(.t)* “one who is in”, derived from the preposition *m* “in”. In an attributive function, adjectives usually agree in gender and number with the noun they follow. The boundary between adjective and noun is occasionally diffuse in Old Egyptian, as it is unclear if a nisba is used as an adjective in an attributive function or as a noun in apposition.
- 2) Adverb (ADV; 46/0.31%): This part of speech is only used sporadically. Among the Old Egyptian adverbs, *im* “there” is common in the Pyramid Texts, although it is occasionally unclear whether it is the adverb *im* or the preposition *m* in *status pronominalis* with an omitted suffix pronoun. Instead of adverbs, adpositions (ADPs; 1,901/12.98%) are usually used in Old Egyptian, consisting of a preposition and a noun phrase. Nouns with an adverbial function, such as *č.t* “eternally” or *hrw* “day” are also found in Old Egyptian.
- 3) Interjection (INTJ; 66/0.45%): *hʹ* “O” and *i* “O” are interjections common in the Pyramid Texts. They precede a noun and have a vocative function, for example *hʹ Wniš* “O Unas”.
- 4) Noun (NOUN; 4,036/27.55%): There is no case distinction of nouns in Egyptian scripts.¹³ They have two genders, masculine and feminine. The ending *t* is used to mark the feminine gender and to form the neuter gender, especially in participles and relative forms, for example *nfr.t* “that which is good” i.e. “(the) good”. Nouns have three numbers, singular, plural, and dual.
- 5) Proper Noun (PROPN; 1,788/12.20%): Names of deities, kings and mythological places are common in the Pyramid Texts. All of them are annotated as PROPN.
- 6) Verb (VERB; 2,490/17.00%): The EUJA treebank follows the structuralist approach, reinforcing and developing Polotsky’s theoretical framework of the Egyptian verbal system. In Old Egyptian, there are two verb conjugations, the “suffix pronoun conjugation” (SPC) and the “Old Semitic suffix conjugation” (OSSC). The former needs a noun or a suffix pronoun as a subject in a similar way as non-pro-drop languages, such as English. Most of the exceptions to this rule are due to phonographic reasons. The OSSC consists of personal endings added to the verb stem similar to the verbs of pro-drop languages, such as Spanish. The SPC is based on a system of tenses:¹⁴ the past I *ščm ʹf* (SPC= Past-1), the past II *ščm.n ʹf* (SPC=Past-2), the present *ščm ʹf* (SPC=Pres), the future *ščm ʹf* (SPC=Fut), the bireferent future *ščm.t ʹf* (SPC=Bi-Fut)¹⁵ and the contingent tenses *ščm.in ʹf* (SPC=ContPast), *ščm.hr ʹf* (SPC=ContPres) and *ščm.kʹ ʹf* (SPC=ContFut).¹⁶ The SPC also has the subjunctive mood *ščm ʹf*

¹¹ In Semitic languages, such as Arabic, “nisba” is used to label an ending added to nouns, and rarely to prepositions and pronouns, to form (relative) adjectives and nouns (Schulz 2010, 86). The addition of the nisba ending to prepositions to form adjectives and nouns is a common feature in Egyptian.

¹² The absolute and relative frequency of each part of speech is given between brackets.

¹³ The genitive case is expressed by two consecutive nouns (“direct genitive”) or by the adjective nisba *n.i* “belonging to” (“indirect genitive”).

¹⁴ The keys in brackets are used in the XPOS column of the treebank.

¹⁵ The bireferent future has two reference points in time, one in the past and one in the future.

¹⁶ The contingent tenses are conditioned on the verbal action of the main clause.

(SPC=Sub). The identification of the present, future and subjunctive is usually circumstantial, depending on the context, since strong verbs in hieroglyphic writing have the same consonant spelling. The impersonal construction (Imprs=Man) corresponding to “one” in English is rendered by adding the noun *tt* / *tw* to the SPC verb form, for example:

n fh-tt n zk (EUJA-658)

“No one got rid of you.”

In addition, there are two passive verb forms in the SPC, the past passive *ščm.w šf* (SPC=PastPass) and the future passive *ščmm šf* (SPC=FutPass). The past II *ščm.n šf*, the present *ščm šf*, the future *ščm šf* and the passive forms can be used as abstract relative verb forms (Type=Abstrel), i.e. nominal finite verb forms used syntactically as nouns, especially in the emphatic construction, the Egyptian cleft sentence with an adverbial phrase as focus, for example:

pr.n šf hr ir.t Hr.w (EUJA-248)

“It is with the Eye of Horus that he came forth.”

The SPC may consist of adjective finite verb forms, known as “relative verb forms” (VerbForm=Relform), which match the gender and number of the antecedent, for example:

Wšr(.w) Wniš m n zk ir.t Hr.w šnm.tn šf (EUJA-222)

“Osiris Unas, take the Eye of Horus, which he rejoined.”

There are syntactic rules for the use of the OSSC in relation to SPC tenses. Thus, the tense, aspect and mood of the OSSC varies according to its syntactic function. The Early Egyptian verb system has an imperative (Imp) and infinite verb forms. The infinitive (Inf) is the nominal infinite verb form, as opposed to the nominal finite verb forms i.e. the abstract relative verb forms. In addition, there are two adverbial infinitives, the so-called negatival complement (NegCom) and the complementary infinitive (ComplInf). Participles (Part) are adjective infinitive verb forms as opposed to the adjective finite verb forms i.e. the relative forms. Both participles and relative forms are occasionally used as nouns.

- 7) Adposition (ADP; 1,901/12,98%): In Old Egyptian, adpositions are usually prepositions used before a noun. Prepositions occasionally show different spellings in status pronominalis (Status=Pron) and status constructus (Status=Cons), for example *im* (Status=Pron) and *m* (Status=Cons) “in”. Complex prepositions such as *m-ʿ* “in the hand” i.e. “from” are considered multiword expressions (MWEs). Old Egyptian also knows the use of postpositions, for example *is* “like”.
- 8) Auxiliary (AUX; 45/0.31%): The particle *tw* is considered an auxiliary as it is used to express the present perfect in combination with the past II *ščm.n šf* and the habitual aspect with the present *ščm šf*, for example: *tw rč.n (št) t' n hkr* (EUJA-1) “(I) have given bread to the hungry.” *tw phr n šf hʿ(.w)* (EUJA-1274) “Thousands (usually) serve him.”
- 9) Coordinating Conjunction (CCONJ; 8/0.05%): The use of CCONJs in Old Egyptian is exceptional. In the current state of the Egyptian-UJaen treebank, only *isč* “and” is attested as CCONJ (e.g. EUJA-548).
- 10) Determiner (DET; 369/2.52%): No articles are used in Old Egyptian. There are four types of demonstrative pro-adjectives (Dem) with three genders, masculine, feminine and neutral.
- 11) Numeral (NUM; 159/1.09%): There are ordinal and cardinal numbers in Egyptian. While “1” and “2” are adjectives, the remaining cardinals are nouns. Ordinal numbers usually follow a noun as attributives.
- 12) Particle (PART; 288/1.97%): Old Egyptian has many particles, three types of which are present so far in the EUJA treebank—negative particles (*n* and *ny*), emphatic particles (*tn*, *is*, *wnn.t*, *hm* and *mi*) and modal particles (*ʿ* and *my*).
- 13) Pronoun (PRON; 2,708/18.48%): There are three types of personal pronouns in Old Egyptian—the independent (IndPron), dependent (DepPron) and suffix (SFP). The keys to the three types are annotated in the XPOS column of the EUJA treebank.
- 14) Subordinating conjunction (SCONJ; 4/0.03%): Two SCONJs have been annotated so far in the EUJA treebank, *n-n.tt*

“because” (UDE-385) and *wn.t* “that” (UDE-1380).

- 15) Symbol (SYM): Although no symbols are found in the current state of the EUJA treebank, some signs may have been used as symbols in exceptional cases.

4.6 Universal Dependency Relations

Nominal core arguments (nominal and clause subject, object and indirect object), non-core arguments (oblique nominal, vocative, expletive and dislocated element) and nominal dependents (nominal modifiers, appositional modifier and numeric modifier) are widespread in Egyptian. It should be noted that the vocative is usually used in the Pyramid Texts (418/2.85%), as these are ritual texts addressed to the deceased king.

The dependency relation between verbal clauses is often established by “adordination” (Díaz Hernández, 2013: 5, footnote 20), i.e. the syntactic dependency relation caused by a temporal reference of the verb form in the “adordinate” clause:

mḥ-ib n(.t) nšw ḥnt ʃ (EUJA-32)

“One who earns the king’s trust (i.e. king’s confidant) (when) he sails upstream.”

Here the present tense *ḥnt ʃ* “he sails upstream” is syntactically dependent on the head of the preceding clause because of the temporal reference of the verb form.

The current state of the EUJA treebank also contains cases of modifier words and function words. The three types of universal modifier words are adverbial modifiers (172/1.17%, e.g. the negative particle *n* in EUJA-1072), discourse elements (151/1.03%, e.g. the particle *m ʃk* in EUJA-916) and adjectival modifiers (500/3.71%, e.g. the adjective *nfr.t* in EUJA-923). Among the function words, Old Egyptian has the particle *tw* used as an auxiliary (45/0.31%, e.g. EUJA-1), the demonstrative determiner *pt* or *pw* used as a copula (96/0.66%, e.g. EUJA-417), markers (54/0.37) such as the subordinating conjunction *wn.t* (EUJA-1380), determiners (246/1.68%) and prepositions usually used to mark a case of relation. In Egyptian, classifiers are not words, but signs that provide semantic information about the word they accompany.

Conjuncts (293/2.68%) are usually connected to other elements without coordinating conjunctions. The “fixed” relation is only used for com-

plex prepositions (111/0.76%), such as *m-ḥt* “behind” and the flat relation for names consisting of two or more elements, for example *Ḥr.w-nḥn(.y)* “Horus of Nekhen”. Egyptian multiword expressions are not annotated as elements in a “fixed” relation because they are expressions with an idiosyncratic meaning whose morphological and syntactic structure can change.


No list has yet been annotated in the EUJA treebank, although chains of items are found in Egyptian inventories.

Parataxis (303/2.07%) is a common relation, as it is used in reported speech (e.g. EUJA 973) and to link pairs of sentences in the so-called “balanced sentence” (e.g. EUJA-645).




The “orphan” relation used to indicate ellipsis is documented in the EUJA treebank (15/0.10%, e.g. EUJA-916) and combinations of lexemes considered morphosyntactically as single words are annotated as compounds (93/0.63%), for example *psč.t-nčš.t* “Little Ennead”. Unspecified dependency (dep; 48/0.33%) occurs when the relation between words cannot be determined due to the absence of vowels in the hieroglyphic script, for example in the offering formula (EUJA-168).

4.7 Hieroglyphs

Hieroglyphs have been annotated manually (over 15,000 signs) using Unicode characters in the MISC column. When hieroglyphs are omitted for phonographic or conventional reasons, the key “Hiero=No” is annotated.

It should be noted that the Unicode extended repertoire of Egyptian hieroglyphs and control characters are still not supported by computer systems (Suignard, 2023 and Glass et al., 2021). Thus, only hieroglyphs from Gardiner’s list are annotated (Gardiner, 1957: 438–548). The key “UC_No” means that there is no Unicode character for a given hieroglyph, whereas when a Unicode character for a hieroglyph of the extended repertoire (Suignard, 2023) cannot be used because still under development, its code is annotated with the key “UC_Code”, for example  is annotated “UC_1397B”.

As Unicode control characters cannot be used yet to arrange hieroglyphs, they are annotated using the following signs:

1. Colon (:) to indicate subordination of signs, for example  :  corresponds to  *pn* “this”.

2. Brackets () to segment groups of hieroglyphs.
3. Asterisk (*) to indicate the juxtaposition of hieroglyphs, for example ($\square^* \triangle$): corresponds to $\square \triangle p.t$ “sky”.

5 Training and Evaluating an NLP model

We used the CoNLL-U file containing the EUJA treebank to train a model of UDPipe 1 (Straka *et al.*)¹⁷ to automatically perform tokenisation, morphological analysis, part-of-speech tagging, lemmatisation and dependency parsing. The test set consisted of 160 sentences chosen randomly. These sentences were EUJA-181–200, 351–370, 491–510, 671–690, 811–830, 960–979, 1221–1240 and 1431–1450. The training set consisted of the remaining 1,413 sentences. Table 2 shows the results of the evaluation process.¹⁸

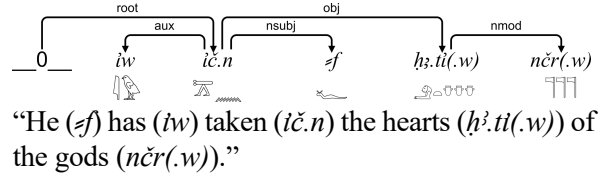
Metric	F1 Score
UPOS	90.30
XPOS	76.01
UFeats	75.87
AllTags	65.39
Lemmata	89.38
UAS	82.52
LAS	71.97
CLAS	69.13
MLAS	56.14
BLEX	63.27

Table 2: Evaluation of an NLP model trained on the treebank

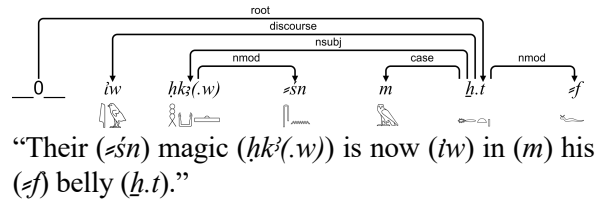
This table shows promising results as all categories get an F1 score over 50. The accuracy of lemmata (89.38) and Universal Part of Speech tags (UPOS: 90.30) is especially high. The Labeled Attachment Score (LAS), the Bilexical Dependency Score (BLEX), the Language Specific Part of Speech tags (XPOS) and the Morphological Features (UFeats) show an F1 score of between 60.00 and 80.00.¹⁹ The Morphology-Aware Labeled Attachment Score (MLAS) is the only category with a F1 score between 50.00 and 60.00.

The UDPipe 1 trained model usually provides a high accuracy rate on UPOS tags, especially nouns and nisba adjectives. As for parsing, it can automatically and accurately annotate short sentences, for example EUJA-1280 and 1287:

EUJA-1280:

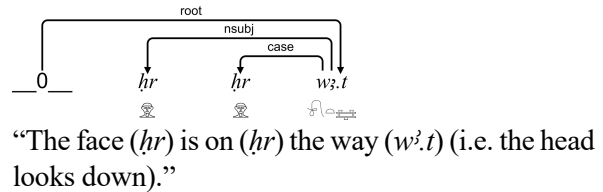


EUJA-1287:



The trained model reveals to be sufficiently good in assigning the correct morphological features and dependency relations to two words with the same spelling, for example:

EUJA-1324:



6 Conclusion

When Joris F. Borghouts published his Middle Egyptian grammar in 2010, with a large number of examples and references to Egyptian texts, Wolfgang Schenkel predicted that a digital database of syntactically analysed sentences would be the next step in Egyptian philology.²⁰ The EUJA treebank makes Prof. Schenkel’s prediction come true, as it contains morphosyntactically annotated sentences from the most representative texts of each pre-Coptic stage. It will be an auxiliary tool for the study of Egyptian grammar, facilitating the synchronic and diachronic parsing of structures and words.

The development of the EUJA treebank includes two further phases:

¹⁷ <https://ufal.mff.cuni.cz/udpipe/1>

¹⁸ The evaluation was performed using the eval.py script provided among other UD tools at <https://github.com/UniversalDependencies/tools>

¹⁹ For LAS and UAS see Buchholz and Marsi, 2006. For CLAS, MLAS, BLEX see Zeman *et al.*, 2018.

²⁰ Personal communication to R.A.D.H.

1) Annotation of the remaining part of the Pyramid Texts.

2) Annotation of the Old Kingdom and First Intermediate Period biographical texts.

Once these corpora are annotated, the treebank will certainly hold over 100,000 Old Egyptian words, and annotation of the Middle Egyptian corpus will begin.

Acknowledgments

This paper is the result of a three-month Short Term Scientific Mission (STSM) Grant awarded to Roberto Antonio Díaz Hernández by “UniDive” (COST Action 21167). The STSM was carried out under the supervision of Marco Carlo Passarotti at the *Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell’Espressione* (CIRCSE) at the *Università Cattolica del Sacro Cuore* in Milan from 1 May to 31 July 2024, with funding from the European Union. Thanks to the organisers of “UniDive” for this grant, and to Flavio Cecchini, Amir Zeldes and Daniel Zeman for discussing the use of Egyptian classifiers and Unicode characters in an issue published in June 2024 on the GitHub website of Universal Dependencies.

References

- Joris F. Borghouts. 2010. *Egyptian. An Introduction to the Writing and Language of the Middle Kingdom*. Peeters, Leuven.
- Sabine Buchholz, Erwin Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*. New York City and Association for Computational Linguistics: 149–164.
- Adriaan de Buck. 1935–1961. *The Egyptian Coffin Texts*, (7 vols.). The University of Chicago Press, Chicago.
- Jacques J. Clère and Jacques Vandier. 1948. *Textes de la Première Période Intermédiaire et de la XI^{ème} Dynastie*. Bibliotheca Aegyptiaca X, Brussels.
- Roberto A. Díaz Hernández. 2013. *Tradition und Innovation in der offiziellen Sprache des Mittleren Reiches*. Wiesbaden.
- Roberto A. Díaz Hernández. 2021. The Man-impersonal *šçm.n-ti/tw(šf)* Form in Earlier Egyptian. *Lingua Aegyptia* 29: 37–59.
- Roberto A. Díaz Hernández. 2022. The Man-impersonal Verb Forms of the Suffix Pronoun Conjugation in Earlier Egyptian. *Lingua Aegyptia* 30: 25–90.
- Roland Enmarch. 2005. *The Dialogue of Ipuwer and the Lord of All*. Griffith Institute, Oxford.
- Adolf Erman and Hermann Grapow (eds.) 1926–1961. *Das Wörterbuch der Aegyptischen Sprache*, (5 vols.) Berlin.
- Hans-W. Fischer-Elfert. 2021. *Grundzüge einer Geschichte des Hieratischen*, (2 vols). Lit, Berlin.
- Alan H. Gardiner. 1932. *Late-Egyptian Stories*. Bibliotheca Aegyptiaca 1, Brussels.
- Alan H. Gardiner. 1957. *Egyptian Grammar. Being an Introduction to the Study of Hieroglyphs*. Griffith Institute, Oxford.
- Andrew Glass (et al.) 2021. Additional control characters for Ancient Egyptian hieroglyphic texts. <https://www.unicode.org/L2/L2021/21248-egyptian-controls.pdf>
- Orly Goldwasser. 2022. L’écriture énigmatique: distancée, cryptée, sportive. In Stéphane Polis (ed.) *Guide des écritures de l’Égypte ancienne*. Cairo: 192–199.
- Stephen R. Glanville. 1955. *The Instructions of ‘Onchsheshonqy* (British Museum Papyrus 10508), (vol. 2). London.
- Wolfgang Helck. 1955–1958. *Urkunden der 18. Dynastie*, (vols. 17–22). Berlin.
- Wolfgang Helck. 1970. *Die Prophezeiung des Nfr.tj*. Otto Harrasowitz, Berlin.
- Kenneth A. Kitchen. 1975–1990. *Ramesseid Inscriptions. Historical and Biographical*, (8 vols.) Oxford.
- Roland Koch. 1990. *Die Erzählung des Sinuhe*. Bibliotheca Aegyptiaca XVII, Brussels.
- Hans O. Lange and Heinrich Schäfer. 1902–1925. *Catalogue général des antiquités égyptiennes du Musée du Caire. Grab und Denksteine des Mittleren Reichs*, (4 vols.). Berlin.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre and Daniel Zeman. 2021. Universal Dependencies. Association for Computational Linguistics, 47(2): 255–308. https://doi.org/10.1162/COLI_a_00402.
- Edouard Naville. 1886. *Das aegyptische Tottenbuch der XVIII. bis XX. Dynastie*. Berlin.
- Richard Parkinson. 1991. *The Tale of the Eloquent Peasant*. Oxford.
- Slav Petrov, Dipanjan Das and Ryan MacDonald. 2012. A Universal Part-of-Speech Tagset. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 12)*: 2089–2096.

Alexandre Piankoff. 1968. *The Pyramid of Unas*. New York.

Hans J. Polotsky. 1944. *Études de syntaxe copte*. Publications de la Société d'Archéologie Copte, Cairo.

Hans J. Polotsky. 1976. Les transpositions du verbe en égyptien classique. *Israel Oriental Studies* 6: 1–50.

Otto Rössler. 1971. Das Ägyptische als semitische Sprache. In Franz Altheim and Ruth Stiehl (eds.) *Christentum am Roten Meer*, vol. 1, de Gruyter, Berlin: 263–326-

Wolfgang Schenkel. 2012. *Tübinger Einführung in die klassisch-ägyptische Sprache und Schrift*. Pagina, Tübingen.

Thomas Schneider. 2023. *Language Contact in Ancient Egypt*, Lit, Berlin.

Eckehard Schulz. 2010. *A Student Grammar of Modern Standard Arabic*. Cambridge.

Kurt Sethe. 1908–1922. *Die altägyptischen Pyramidentexte nach den Papierabdrücken und Photographien des Berliner Museums*, (4 vols.). Leipzig.

Kurt Sethe. 1906–1909. *Urkunden der 18. Dynastie*, (vols. 1–16). Leipzig.

Kurt Sethe. 1933. *Urkunden des Alten Reichs*. Leipzig.

Milan Straka (et al.) 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)*: 4290–4297.

Michel Suignard. 2023. Encoding proposal for an extended Egyptian Hieroglyphs repertoire. <https://www.unicode.org/L2/L2023/23181-n5240-hieroglyphs.pdf>

Amir Zeldes and Mitchell Abrams. 2018. The Coptic Universal Dependency Treebank. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*. Brussels and Association for Computational Linguistics: 192–201.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, Slav 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*. Brussels and Association for Computational Linguistics: 1–21.

A Appendix

	LUT	Tübingen	Unicode
	ʒ	ʒ	A723
	i	i	A7BD
	y	y	
	ī	ī	00EF
	‘	‘	A725
	w	w	
	b	b	
	p	p	
	f	f	
	m	m	
	n	n	
	r	r	
	h	h	
	ḥ	ḥ	1E25
	ḥ	ḥ	1E2B
	ḥ	ḥ	1E96
	z	s	
	s	ś	015B
	š	š	0161
	q	ḳ	1E33
	k	k	
	g	g	
	t	t	
	ṯ	č	010D
	d	ṯ	1E6D
	ḏ	č	010D+0323

Table 3: The LUT, the Tübingen transcription system and the Unicode signs used in the EUJA treebank

Symmetric Dependency Structure of Coordination: Crosslinguistic Arguments from Dependency Length Minimization

Adam Przepiórkowski

ICS Polish Academy of Sciences
and University of Warsaw

Magdalena Borysiak

University of Warsaw

Adam Okrański

University of Warsaw

Bartosz Pobożniak

University of Warsaw

Wojciech Stempniak

University of Warsaw

Kamil Tomaszek

University of Warsaw

Adam Głowacki

University of Warsaw

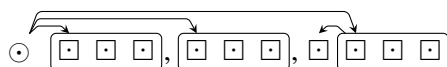
Abstract

The aim of this paper is to replicate and extend recent treebank-based considerations regarding the syntactic structure of coordination. Overall, we confirm the previous results that, given the principle of Dependency Length Minimization, corpus data suggest that the structure of coordination is symmetric. While previous work was based on 2 English datasets, we extend the investigation to 3 more English datasets, 3 Polish datasets, and UD corpora for a number of diverse languages. The results confirm the symmetric structure of coordination, but they also make it possible to question some of the previous findings regarding the exact symmetric structure of coordination.

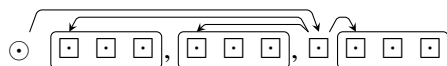
1 Introduction

There is no agreement in theoretical linguistics about the syntactic structure of coordination. Within dependency approaches alone, 4 basic structures have been proposed with a number of variants (see Popel et al. 2013 and Przepiórkowski and Woźniak 2023; the latter henceforth abbreviated to PW23), as schematically presented in (1)–(4):

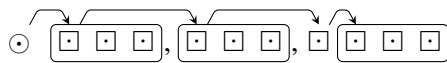
(1) Multi-headed/London:



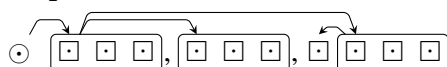
(2) Conjunction-headed/Prague:



(3) Chain/Moscow:



(4) Bouquet/Stanford:



In these schemata, \odot marks the governor (e.g., *saw* in (5)), \square marks tokens within coordination, with tokens belonging to the same conjunct grouped; the single ungrouped \square is the conjunction (e.g., *and*).

- (5) Maggie *saw* [[a brown dog], [a grey cat], and [a green tree]].

Moreover, these schemata follow syntactic theory in assuming that heads of conjuncts are typically near the beginning of these conjuncts in English, given that it is a head-initial language; e.g., the DP conjuncts in (5) are headed by the determiners.¹

Prague and London approaches (1)–(2) are symmetric in the sense that all conjuncts bear the same relation to the governor of the coordinate structure: in (1) they are direct dependents of the governor, while in (2) they are all direct dependents of the conjunction. By contrast, in the asymmetric (3)–(4), only the first conjunct is a direct dependent of the governor of the coordinate structure, with the other conjuncts being direct (in (4)) or possibly indirect (in (3)) dependents of the first conjunct.

PW23 give a novel corpus-based argument for a symmetric structure of coordination. The argument assumes the principle of Dependency Length Minimization (DLM) – a robustly demonstrated tendency for natural languages to strive for maximally local dependencies.² As argued in Hawkins 1994 and Futrell et al. 2020, this tendency operates both at the level of use and at the level of grammar.

At the level of use, when both orders of two dependents are grammatical, the longer one of these dependents gets, the stronger the pressure for the other dependent to occur closer to the governor. For example (cf. Przepiórkowski and Woźniak 2023), consider an intransitive verb, e.g., *sing*, and its two PP dependents: a durative PP_{for}, e.g., *for two hours*, and a locative PP_{in}, e.g., *the short in that club* or

¹This should be contrasted with Universal Dependencies (UD; <https://universaldependencies.org/>; Nivre et al. 2016, de Marneffe et al. 2021, Zeman et al. 2024), where the nouns – i.e., typically conjunct-final tokens – are assumed to be heads; see Osborne and Gerdes 2019 for discussion.

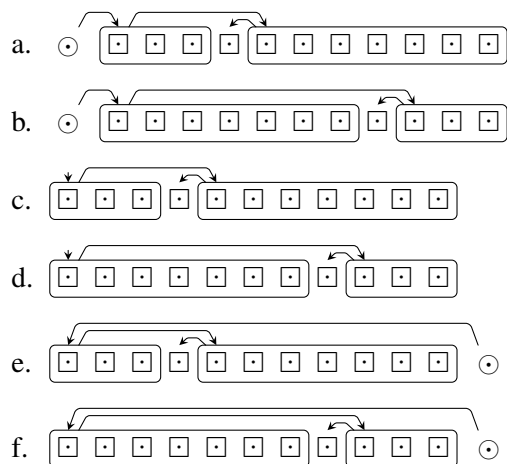
²See, e.g., Behaghel 1909, Hawkins 1994, Gibson 1998, Ferrer-i-Cancho 2004, Gildea and Temperley 2007, 2010, Liu 2008, Liu et al. 2017, Futrell and Levy 2017, Temperley and Gildea 2018, and many others.

the much longer *in the most famous American jazz club*. Then, if the likelihood of the order [V PP_{for} PP_{in}] (as opposed to [V PP_{in} PP_{for}]) is p_1 for the shorter PP_{in} (i.e., for *sing [for two hours] [in that club]*), then it will be $p_2 > p_1$ for the longer PP_{in} (i.e., for *sing [for two hours] [in the most famous American jazz club]*).

Such at-use pressures may become conventionalized, i.e., they may become at-grammar tendencies. For example (cf. PW23 again), given that NPs are on average shorter than PPs (which consist at least of a preposition and an NP), DLM will be more often satisfied by [V NP PP] than by the [V PP NP] order. Hawkins (1994: 90) argues that this tendency became conventionalized in English into a general preference for the former order, active even when the lengths of the NP and the PP are equal, i.e., when there is no at-use DLM gain. For example, despite the similar lengths of the two dependents, *I sold [my mother's ring] [for five dollars]* is preferred to *I sold [for five dollars] [my mother's ring]*. On the other hand, this at-grammar pressure may be overridden by the at-use pressure when length differences are large: the [V PP NP] order becomes more natural again in *I sold [for five dollars] [my mother's silver engagement ring that she got from my father]*, despite the violation of the at-grammar preference for [V NP PP].

Now, the general idea of PW23's argument is to compare the predictions of each of the four proposed structures of coordination to what is observed in corpora. For example, consider binary coordinations in the asymmetric Stanford approach.

(6) Bouquet/Stanford:



(6a–b) illustrates coordination with the governor on the left (as in (5)), (6c–d) – coordinations with no governor (e.g., coordination of sentences), (6e–f) – those with the governor on the right (e.g., *Bart and*

Lisa laughed). Each pair compares two orders of conjuncts: in the first the first conjunct is shorter, in the second – the second is shorter.

If DLM operates in coordinate structures only at the level of use, then the following tendencies are predicted. First, as seen in (6a–b), when the governor is on the left, there is a pressure for the first conjunct to be shorter: the total sum of dependency lengths is smaller in (6a) than in (6b).³ So, there is an at-use pressure for the shorter conjunct to occur as the first conjunct when the governor is on the left. Moreover, the difference between the aggregate dependency lengths in (6a–b) is equal to the difference of lengths of the two conjuncts. Hence, this pressure for the shorter conjunct to be first is greater when the conjunct length differences are greater. These considerations translate into a clear prediction: when the difference between the lengths of conjuncts is greater, the proportions of coordinations with the shorter first conjunct should be greater. Formally, let $p_L(n)$ be the proportion of those binary coordinations with a governor on the left with the absolute length difference between the two conjuncts being $n > 0$ in which the first conjunct is shorter. The prediction of the Stanford approach is that $p_L(n)$ should be a monotonically increasing function of n .

It is easy to see that exactly the same prediction is made when there is no governor (see (6c–d)) and when the governor is on the right (see (6e–f)): in all three cases, when the first conjunct is shorter, the aggregate dependency length is smaller. Moreover, in all three cases the difference between the sum of lengths is the same and equal to the length difference between the two conjuncts. That is, $p_L(n)$, $p_-(n)$ (the proportion function when there is no governor), and $p_R(n)$ (the proportion function when the governor is on the right) should all be equally monotonically increasing.

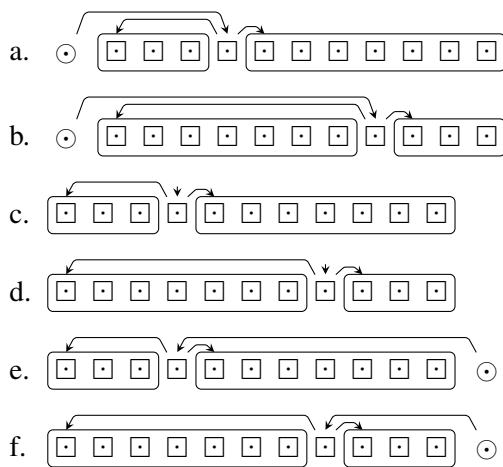
In order to verify such predictions, PW23 examined the distribution of binary coordinations in PTB_& (Ficler and Goldberg 2016), a version of Penn Treebank (PTB; Marcus et al. 1993) which improves on PTB by offering explicit and relatively consistent information about coordinations. Out of 21,825 binary coordinations they extracted from around 49.2K sentences in PTB_&, 13,106 had gov-

³Dependencies *within* conjuncts are not shown here, as they do not depend on the order of conjuncts, i.e., they do not matter for the comparison of aggregate dependency lengths. Also, unlike some of the previous work reported below, we only consider lengths measured in words here – not syllables or characters.

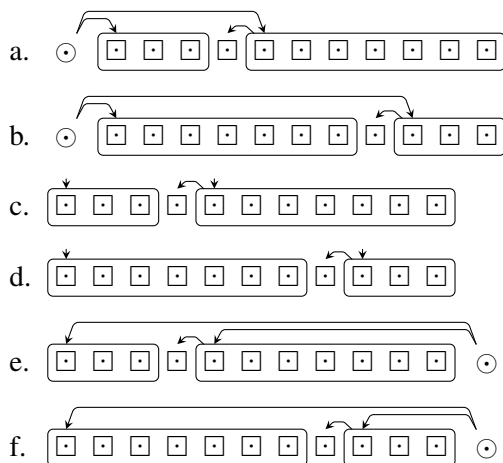
error on the left, 4,000 had no governor, and 4,719 had governor on the right. For each subpopulation, they fitted a monofactorial logistic regression model to estimate $p_L(n)$, $p_-(n)$, and $p_R(n)$. The result was that $p_L(n)$ and $p_-(n)$ were monotonically increasing, as predicted by the Stanford approach, but $p_R(n)$ was more or less constant, with confidence bands compatible with the true $p_R(n)$ being either decreasing or increasing. However, they also performed a multifactorial binary logistic regression analysis, which showed that the slope is statistically significantly flatter when the governor is on the right than when it is on the left or missing. This is not predicted by the Stanford approach, where all three slopes should be the same.

As binary coordinations have almost exactly the same dependency relations on the two asymmetric – Stanford and Moscow – approaches, the above observations and conclusions also hold for the Moscow approach, which we will not consider further. However, the predictions of the two symmetric approaches are more interesting. The relevant schemata are presented in (7)–(8):

(7) **Conjunction-headed/Prague:**



(8) **Multi-headed/London:**



PW23 note that the Prague approach is directly compatible with their corpus analyses: shorter first conjuncts minimize aggregate dependency length when the governor is on the left (see (7a–b)) or absent (see (7c–d)), but not when it is on the right (see (7e–f)). In the latter case, the aggregate dependency lengths are the same in (7e–f). This directly corresponds to the observed monotonically increasing $p_L(n)$ and $p_-(n)$, and constant $p_R(n)$.⁴

Finally, in the case of the London approach, the prediction is that $p_L(n)$ is increasing (cf. (8a–b)), $p_-(n)$ is constant (cf. (8c–d)), and $p_R(n)$ is decreasing (cf. (8e–f)). This is not directly compatible with PW23’s corpus-based models when DLM is only considered at use. However, PW23 also consider an at-grammar DLM effect, related to the well-known fact (which they confirm on the basis of PTB_&) that most of the coordinate structures have their governor on the left in English. As – on any approach to coordination – the shorter first conjunct minimizes the aggregate dependency length in such situations, this means that in most cases it pays to have the first conjunct shorter and that this tendency could have plausibly been conventionalized to the at-grammar pressure for shorter first conjuncts in general.

The existence of such a hypothetical at-grammar tendency does not change anything in the case of asymmetric approaches (they still predict that all three $p_*(n)$ functions should be equally monotonically increasing), but it makes a difference in the case of the London approach. If such an at-grammar tendency is present, then $p_L(n)$ is still predicted to be monotonically increasing, but now also $p_-(n)$ is predicted to be monotonically increasing, by virtue of the at-grammar pressure alone. Moreover, the at-use pressure for the shorter *second* conjunct observed in (8e–f) is counterbalanced by the hypothetical at-grammar pressure for the shorter *first* conjunct, resulting in the roughly constant $p_R(n)$ observed in PTB_&.

One of the limitations of PW23 is the relative scarcity of data: the number of coordinations with the governor on the right was not sufficient to train

⁴Note that the Prague approach predicts stronger pressure when the governor is on the left (see (7a–b); the difference between the two orders is *twice* that of the conjunct length difference) than when it is absent (see (7c–d); the difference is that of the conjunct length difference). PW23’s multifactorial analysis confirms the corresponding difference of slopes between $p_L(n)$ and $p_-(n)$ when length is measured in characters or syllables, but it detects no statistically significant difference when it is measured in words.

a logistic regression model that would give a statistically significant answer concerning the monotonicity of $p_R(n)$. In an attempt to remove this limitation, [Przepiórkowski et al. 2024](#) (henceforth, **PBG24**) replicate [PW23](#)’s study on the basis of the Corpus of Contemporary American English (COCA; [Davies 2008–2023](#)) automatically parsed with Stanza ([Qi et al. 2020](#)) to the UD format. Unlike [PW23](#), they considered the first and last conjunct in *all* coordinations, noting that over 86% of them were binary and that restriction to binary coordinations does not affect the results. From a subset of COCA containing almost 21.8M sentences, they extracted over 11.5M coordinations and fitted those with considerable length differences between conjuncts (at least 4 words) into a logistic regression model. As in [PW23](#), the estimated $p_L(n)$ and $p_-(n)$ were monotonically increasing – $p_L(n)$ more so than $p_-(n)$ – but this time $p_R(n)$ was monotonically *decreasing* (statistically significantly with $p \ll 0.001$). This is clearly incompatible with asymmetric theories, on which all should be similarly increasing, not fully compatible with symmetric Prague approach, on which $p_R(n)$ should be constant if DLM only operates at use or increasing if it also operates at grammar, but fully compatible with the London approach, on the assumption that the at-grammar tendency is strong enough to make $p_-(n)$ – constant at the level of use – increasing, but not strong enough to make $p_R(n)$ – decreasing at use – constant or increasing. **PBG24** conclude that their study makes it possible to sharpen the results of [PW23](#), as it not only provides evidence for symmetric approaches to coordination in general, but for a particular such approach (London).

However, **PBG24** note a major limitation of their approach that was absent in [PW23](#), namely, the low quality of their automatically parsed data. For each governor position (left, absent, right) and each conjunct length difference (from 1 to 20 words), they sampled 15 coordinations from the 11.5M coordinations automatically extracted from COCA, resulting in 900 coordinations altogether, and checked whether they were extracted correctly, i.e., had the right information about governor position and identified the two conjuncts correctly, as only this information matters for the statistical model. They found that only slightly over 50% of coordinations were extracted correctly in this sense. While there are no reasons to think that the distribution of errors significantly influenced their results, such a prob-

lem cannot be *a priori* excluded, leading them to the conclusion that “further replication studies, also based on languages other than English, are needed to make these results even more robust” ([Przepiórkowski et al. 2024](#): 1029).

2 New Studies

In order to validate the results of [PW23](#) and **PBG24**, we performed a number of similar studies on different datasets: 2 for English, 2 for Polish (another head-initial language), and further studies based on UD corpora of a number of languages, including English and Polish. Because of relatively small sizes of the datasets for languages other than English, some of the results by themselves are not statistically significant, but taken together they largely confirm the conclusions of previous studies.⁵

2.1 English

Two English studies follow **PBG24**: they are based on automatically-parsed COCA, i.e., on a low quality but large resource. The difference with respect to **PBG24**’s study is that in the current studies COCA was not parsed to the UD format.

As is well known (see, e.g., [Przepiórkowski and Patejuk 2019](#)), the representation of coordinate structures in the basic UD standard is not optimal: certain structures cannot be represented unambiguously. For example, there is just one UD representation of the sequence *lazy cats and dogs*, whether *lazy* modifies *cats* alone, or whether it modifies the whole coordinate structure (so that *dogs* are also *lazy*).⁶ This is a problem, as it is not clear whether the two conjuncts in *lazy cats and dogs* are of same lengths (this is the case if *lazy* modifies the whole coordination) or whether the first conjunct is longer (if *lazy* modifies *cats* alone), and this information is crucial for the argument at hand. While **PBG24** implemented various heuristics for disambiguating such representations, they are imperfect, so this ambiguity problem contributes to the low quality of input data in their study.

⁵All statistics and visualizations were performed using R ([R Core Team 2024](#)), with the statistical significance of slope differences estimated using the `emmeans` commands from the `emmeans` package ([Lenth 2024](#)).

⁶This difference is easy to represent in some other approaches, including the Prague approach and the enhanced version of UD ([Schuster and Manning 2016](#)). Unfortunately, the main dependency parsers currently in use only provide the basic UD structures.

2.1.1 COCA parsed with Stanza/SUD

In order to alleviate this problem, two different representations were used here. In the first study, the Surface-syntactic Universal Dependencies (SUD; Gerdes et al. 2018, 2021a) format was used, which makes it possible to represent information about shared dependents explicitly.⁷ We trained Stanza on a treebank consisting of SUD versions of three English UD corpora: EWT (Silveira et al. 2014), GUM (Zeldes 2017), and (the English part of) ParTUT (Sanguinetti and Bosco 2014), all downloaded from <https://surfacesyntacticud.github.io/data/>.

In order to assess the quality of coordinations extracted using this SUD-trained parser, we also trained Stanza on the original UD versions of the same corpora, and compared the 1526 coordinations extracted from the testing parts of these corpora by the two trained parsers. The two parsers agreed on 1075 coordinations. In the case of the remaining 451, SUD-based procedure correctly identified 260 (57.6%) coordinations, and UD-based procedure – 252 (55.9%) coordinations. That is, the coordination extraction process based on the SUD-trained parser turned out to be only slightly better than that based on the UD-trained parser. (This difference was not statistically significant, according to McNemar’s test.) Hence, while we did not evaluate the quality of extracted coordinations using the same procedure as PBG24, we do not expect coordinations based on SUD-trained Stanza to be of significantly better quality than those based on UD-pre-trained Stanza in PBG24.

Despite this only marginal improvement, the SUD-trained Stanza was used to replicate PBG24’s study. The whole COCA was parsed and, as a result, 14,341,063 coordinations were identified, including 12,476,392 binary coordinations. Three logistic regression models were trained, as before, with results presented in the left column of Figure 1. The relations between the three slopes are as expected: the slope is most positive in the case of $p_L(n)$ (top graph) and least positive in the case of $p_R(n)$ (bottom graph). The fact that all slope differences are highly statistically significant ($p \ll 0.001$) is consistent with symmetric approaches, but not with asymmetric approaches.

⁷Also other aspects of SUD representations, especially, the fact that constructions are headed by function rather than content words (e.g., PPs are headed by prepositions rather than nouns), make the resulting structures less ambiguous and easier to work with.

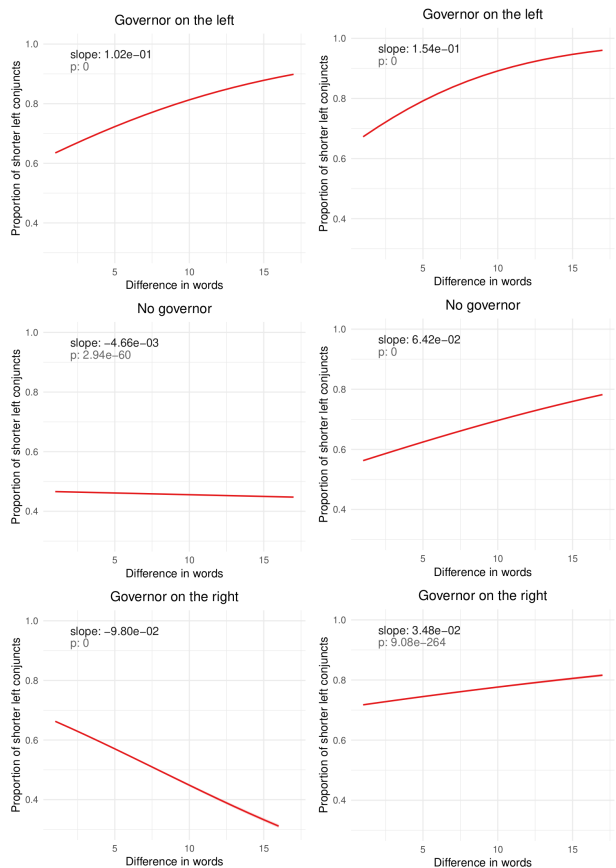


Figure 1: Logistic regression models of COCA coordinations extracted with Stanza trained on SUD (left column) and UD (right column) corpora

Moreover, the significantly negative slope of $p_R(n)$ is only consistent with the symmetric London approach. However, what is unexpected and not witnessed before is that also the slope of $p_-(n)$ was significantly negative. This is incompatible not only with asymmetric approaches and the Prague approach, which all predict that it should be positive, but also with the London approach, on which it should be constant (if there is no at-grammar pressure) or positive (if there is additional at-grammar pressure). This effect is specific to SUD-trained Stanza.⁸

Interestingly, when the same full COCA was parsed with Stanza trained by us on the UD versions of the same training corpora, the slopes of all logistic regression models were significantly pos-

⁸Moreover, it seems that, at least to some extent, this effect was caused by the inclusion of spoken parts of COCA in the current study, unlike in PBG24, where only written parts of COCA were processed. After removing two conversational genres – spoken and TV/movies – the slope of $p_-(n)$, while still significantly negative, was much flatter (-0.00104 vs. -0.00466 in Figure 1), with p not reaching the < 0.001 significance level.

itive (see the right column of Figure 1), unlike in the SUD-based study (see the left column again), but also unlike in PBG24, where a large subset of COCA was parsed with Stanza pre-trained on UD and where $p_R(n)$ was monotonically decreasing. However, while this difference awaits explanation, the positive slope of $p_R(n)$ is compatible with both symmetric approaches to coordination: on the assumption of *any* at-grammar pressure, the slope of $p_R(n)$ is expected to be positive on the Prague approach, and if this at-grammar pressure is sufficiently strong, then the positive slope of $p_R(n)$ is also expected on the London approach. Moreover, it is important to note that the relations between slopes of these UD-based models are again as expected by symmetric theories of coordination: the slope is most positive in the case of $p_L(n)$ (top graph) and least positive in the case of $p_R(n)$ (bottom graph), with all relevant differences statistically significant ($p \ll 0.001$).

2.1.2 COCA Parsed with BNP

Another way to avoid the problems of UD representation of coordination was to use a constituency parser. To this end, we utilized the Berkeley Neural Parser (BNP; Kitaev and Klein 2018, Kitaev et al. 2019) with the `benepar_en3` model. All of COCA apart from the spoken genre was parsed – around 59.5M sentences. Only simple binary coordinations were extracted – constituents consisting of three children, where the middle child is a conjunction (e.g., *Lisa and Bart*) and constituents consisting of four children, where the first and third child constitute a conjunction (e.g., *either Marge or Homer*). This way the problem of the exact extents of conjuncts was avoided. However, unlike dependency representations, the PTB format produced by BNP does not contain a clear information about governors, so heuristics similar to those used in PW23 were employed. In the process, information about 13,543,340 coordinations was extracted.

The quality of the resulting data was evaluated using exactly the same procedure as in PBG24: for each governor position (left, absent, right) and each conjunct length difference (from 1 to 20 words), 15 coordinations were sampled and checked for correctness (understood as in PBG24: the right conjuncts and the right position of the governor). The data quality was much higher than in the case of Stanza-parsed dataset used in PBG24: 78.11% of coordinations were judged as correct here, as opposed to 50.1% in PBG24.

The results are analogous to those based on UD-trained Stanza reported in the previous section: 1) all three slopes were significantly positive (with $p \ll 0.001$), 2) that of $p_L(n)$ was most positive (0.112), followed by $p_-(n)$ (0.085), and by $p_R(n)$ (0.029), with all differences highly statistically significant ($p \ll 0.001$). Again, this is compatible with both symmetric approaches to coordination (assuming different strengths of at-grammar pressure), but not with asymmetric approaches.

2.2 Polish

Two Polish studies follow PW23: they are based on manually-annotated treebanks, i.e., on small but relatively high-quality resources.

2.2.1 Składnica Constituency Parsebank

The first is based on Składnica, a manually-disambiguated constituency parsebank of Polish (Woliński et al. 2011, 2018) containing 14K sentences. As this is a much smaller corpus than PTB_& (49.2K sentences), the first and last conjuncts of all coordinations, not just binary, were taken into account, resulting in 5395 extracted coordinations (including 4800 binary; vs. 21,825 in PTB_&).

The results are similar to those obtained by PW23. First of all, both $p_L(n)$ and $p_-(n)$ are monotonically increasing with statistically highly significant ($p \ll 0.001$) positive slopes; this is compatible with all approaches, although in the case of the London approach only with the assumption of an at-grammar tendency for shorter first conjuncts. Second, the slope of $p_L(n)$ is statistically significantly ($p < 0.05$) greater than that of $p_-(n)$ (0.18 vs. 0.09); this is only explained by the symmetric approaches. Third, while the slope of $p_R(n)$ is also positive (0.025), this value is not significantly different than 0 ($p > 0.05$); this is again more in line with symmetric approaches. Finally, while the difference of slopes of $p_-(n)$ and $p_R(n)$ is not statistically significant, the difference of slopes of $p_L(n)$ and $p_R(n)$ is ($p < 0.05$), which is not compatible with asymmetric approaches (on which all slopes should be the same), but immediately explained by both symmetric approaches.

2.2.2 Polish Dependency Bank

The second study is based on Polish Dependency Bank (PDB; Wróblewska 2014), a pre-UD dependency treebank in which coordinations are annotated according to the Prague approach (so they were free from the ambiguity problem mentioned

above). The version of PDB used in this study contains over 22K sentences. Again, all coordinations were taken into account: 13,247 were extracted, including 11,635 binary coordinations.

The results of this study are similar to those of the previous one. First, all three slopes are monotonically increasing, but this time all positive slopes are statistically significant ($p \ll 0.001$ for $p_L(n)$ and $p_-(n)$, $p < 0.05$ for $p_R(n)$). Second, the relation between the three slopes is as expected by symmetric theories of coordination: greatest for $p_L(n)$ (0.093), smaller for $p_-(n)$ (0.073), and smallest for $p_R(n)$ (0.055); however, this time the differences between these slopes did not turn out to be statistically significant.⁹

In summary, the results based on PDB alone are not sufficient to distinguish between symmetric and asymmetric approaches to coordination: the relevant differences, while in line with symmetric approaches, are not statistically significant. However, these results are compatible with those based on Składnica, where most of the crucial differences are statistically significant and, hence, provide an argument from Polish for the symmetric structure of coordination.

2.3 Partial Summary and Discussion

The results of previous work and our own studies are presented in Table 1.¹⁰ In the $L/-$ col-

Table 1: Summary of studies described above: number of sentences, number of extracted coordinations, comparisons of slopes (see explanation in text)

	sents	coords	$L/-$	$-/R$	L/R	R
PW23	49.2K	21.8K	-	+***	+**	-
PBG24	21.8M	11.5M	+***	+***	+***	-***
St./SUD	69.2M	14.3M	+***	+***	+***	-***
St./UD	69.2M	10.8M	+***	+***	+***	+***
BNP	59.5M	13.5M	+***	+***	+***	+***
Składnica	14.0K	5.4K	+*	+	+*	+
PDB	22.2K	13.2K	+	+	+	+*

umn, ‘+’ means that the slope of $p_L(n)$ is greater (more positive) than that of $p_-(n)$, and ‘-’ that it is smaller (more negative), and analogously in

⁹Recall that we assume that lengths are measured in words. When they are measured in syllables, the difference between $p_L(n)$ and $p_R(n)$ turns out to be statistically significant ($p < 0.05$), while the character metric renders the difference between $p_L(n)$ and $p_-(n)$ statistically significant ($p < 0.05$).

¹⁰The small ratios of coordinations to sentences in St./(S)UD and BNP rows is probably caused by the inclusion of conversational genres (spoken in all three, TV/movies also in St./(S)UD), characterized by a very large number of very short – coordination-free – sentences.

the next two columns. Recall that the prediction of both symmetric approaches is that the slope of $p_L(n)$ is greatest and that of $p_R(n)$ smallest, so an ideal confirmation of such approaches would have a sequence of statistically significant +’s in these three columns. On the other hand, according to asymmetric approaches there should be no slope differences, so a sequence of statistically insignificant differences is expected. The final R column presents the sign of the slope of $p_R(n)$; the negative sign, ‘-’, is compatible with the London approach, but not with the Prague approach. The number of asterisks reflects levels of statistical significance: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$; additionally, when $p > 0.1$, + or - is in grey.¹¹

What all English models have in common is that the differences between the slopes are as predicted by the symmetric approaches to coordination: when the difference is significant, it is always +. However, the studies reported here also show that the effect of parser is clearly visible, with the slope of $p_R(n)$ – crucial for a potential argument for the superiority of the London symmetric approach over the Prague approach – sometimes significantly negative (PBG24, Stanza/SUD), and sometimes significantly positive (Stanza/UD, BNP). Given that – as shown by the evaluation of extracted coordinations – the quality of input to these models was highest in the case of COCA parsed with BNP, these results seem to be most reliable. Hence, the conclusion of PBG24 that – given the negative slope of $p_R(n)$ – the London approach is the only one compatible with corpus data might have been premature. That is, the current conclusion must be that asymmetric approaches are clearly incompatible with corpus data, but – contrary to the conjecture of PBG24 – the resulting models are not sufficiently reliable to distinguish between the two symmetric approaches. Note that this conclusion is compatible with PW23’s results, which were inconclusive about the slope of $p_R(n)$, as well as with the results of our Polish studies, according to which the slope of $p_R(n)$ is positive (significantly so, according to the PDB-based study).

2.4 Other Languages (UD Corpora)

We also performed similar studies on the basis of UD corpora of 10 languages (version 2.14; Zeman et al. 2024). We only considered clearly head-ini-

¹¹PBG24 do not report the levels of statistical significance for slope differences; we estimated these levels on the basis of their raw data, made available to us.

tial languages with at least 700K tokens in UD corpora, i.e., 5 Romance languages (Italian, Latin, Portuguese, Romanian, Spanish), 2 Germanic (English, Icelandic), and 3 Slavic (Czech, Polish – exceptionally, even though it had less than 700K tokens, Russian).¹² See Table 2.

Table 2: Sizes of – and results based on – UD datasets: number of tokens, number of extracted coordinations, comparisons of slopes (see explanation in text)

	tokens	coords	$L/-$	$-/R$	L/R	R
it	864K	25,426	+***	+	+	+**
la	983K	39,510	+	+***	+***	–
pt	1,361K	29,255	+***	+	+**	+**
ro	938K	37,247	+	+***	+***	+
es	1,002K	28,666	+***	+	+	+**
en	718K	21,013	–	+**	+**	–
is	1,183K	43,852	+***	–	+	+
cs	2,249K	90,566	–***	+***	+***	–*
pl	497K	16,684	–	+	+	+
ru	1,896K	61,004	+	+***	+***	–

While many differences are statistically insignificant, a fact that may be explained by the relatively small sizes of corpora used, it is clear that the results of this study are overall only compatible with the symmetric approaches.

This is most clear in the case of the Romance languages, where all differences are in the positive direction expected by symmetric approaches and the L/R difference is always statistically significant. While only in the case of Latin are all three differences statistically significant, for all Romance languages at least one of the differences $L/-$ and $-/R$ is highly statistically significant ($p < 0.001$), contra asymmetric approaches.

All statistically significant differences are in the ‘right’ positive direction also in the case of the two Germanic languages, English and Icelandic, and similarly for two of the Slavic languages, Russian and Polish, even if only one difference reaches the level of statistical significance in the case of Polish (probably because of the very small dataset).

¹²We used Typometrics (<https://typometrics.elizianet/>; Gerdes et al. 2021b) to estimate headedness: we considered a language head-initial if it scored over 50% on two measures: the percentage of adpositional constructions with the adposition preceding its proper noun object (ADP-comp:obj-PROP) and the percentage of verbal phrases with the verb preceding its proper noun object (VERB-comp:obj-PROP). While all selected languages scored close to 100% on the adpositional measure, they differed widely on the verbal measure: from 51% for Latin, 72% for Czech, 81% for Russian, and 85% for Polish, to over 99% for English, Portuguese, and Italian. By contrast, two prototypically head-final languages – Korean and Turkish – scored 0% on the adpositional measure and, respectively, 0% and 3% on the verbal measure. German is not included, as it scored below 50% on the verbal measure.

Finally, Czech is an outlier in this study, in that the $L/-$ difference is highly significantly *negative*. However, the difference of slopes of $p_L(n)$ and $p_-(n)$ is relatively small: 0.0404 vs. 0.0553. Moreover, while both these slopes are positive, the slope of $p_R(n)$ is statistically significantly ($p < 0.05$) negative, -0.0167 , which speaks not only against asymmetric approaches, on which it should be positive, but also – a little ironically – against the Prague approach, on which the slope of $p_R(n)$ should be 0 or positive. We leave the investigation of this outlier for future work.

3 Conclusion

At the most general level, the main contribution of this paper is a demonstration that extensive replication is crucial not only in psychology, medicine, and social sciences, but also in formal and computational linguistics. PBG24’s replication of PW23’s argument for the symmetry of coordination seemed to narrow down potentially valid representations of coordination from the two symmetric approaches to just the London approach, but the current more extensive replication invalidates this conjecture. While two of our studies (Stanza/SUD and UD/cs) result in models with negative slopes of $p_R(n)$, compatible only with the London approach, 7 other studies (Stanza/UD, BNP, PDB, UD/it, UD/pt, UD/es, UD/is) – including two based on English data, just as Stanza/SUD and PBG24 – result in models with significantly positive slopes of $p_R(n)$. Clearly, the choice of parser and dataset is important for the argument, and future research should determine how exactly it influences the results.

Nevertheless, the current studies add strong cross-linguistic arguments for the main claim of PW23 and PBG24, namely, that corpora provide quantitative evidence for the symmetry of coordination. Apart from UD/cs, where the unexpected statistically significant $L/-$ difference was observed, and PDB, where relevant differences were not statistically significant, in all other 13 models statistically significant slope differences were found that are only compatible with symmetric approaches.

An important limitation of this paper is that it only considers head-initial languages, as the above reasoning assumes that heads of conjuncts are conjunct-initial on average. An investigation of the structure of coordination in two head-final languages, Korean or Turkish, may be found in Stempniak 2024a.

Acknowledgements

The corpus-based studies discussed in this paper were performed within the Bachelor’s Group Research Project “Corpus-based investigation of the symmetry of coordination” carried out in 2022–2024 at the Cognitive Science programme at the University of Warsaw. The project was supervised by the first author, all authors were participants. Particular studies are described in more detail in:

- **Borysiak 2024** – COCA parsed with Stanza/SUD (§2.1.1),^{13,14}
- **Pobożniak 2024** – COCA Parsed with BNP (§2.1.2),¹⁵
- **Okraśniński 2023** – Składnica Constituency Parsebank (§2.2.1),¹⁶
- **Tomaszek 2023** – Polish Dependency Bank (§2.2.2),¹⁷
- **Stempniak 2024b** – UD Corpora (§2.4).¹⁸

We are grateful to Bartosz Maćkiewicz for his statistical assistance in some of these studies, as well as to TLT 2024 reviewers for their comments.

References

Otto Behaghel. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*, 25:110–142.

Magdalena Borysiak. 2024. Dependency structure of English coordination: A surface-syntactic approach. Bachelor’s thesis, University of Warsaw.

Mark Davies. 2008–2023. The Corpus of Contemporary American English (COCA). Available online at <https://www.english-corpora.org/coca/>.

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

Ramon Ferrer-i-Cancho. 2004. Euclidean distance between syntactically linked words. *Physical Review E*, 70:056135.

Jessica Fidler and Yoav Goldberg. 2016. Coordination annotation extension in the Penn Tree Bank. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 834–842, Berlin, Germany.

¹³<https://github.com/bmagdab/sud-coords>

¹⁴<https://github.com/glowak/dlm>

¹⁵<https://github.com/KattGaii/coca-thesis>

¹⁶<https://github.com/Adokr/korpus>

¹⁷<https://github.com/kvmilos/PracaLicencjacka>

¹⁸<https://github.com/wjstempniak/Dependency-Structure-of-Coordination>

Richard Futrell and Roger Levy. 2017. **Noisy-context surprisal as a human sentence processing cost model**. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL2009): Volume 1, Long Papers*, pages 688–698, Valencia, Spain.

Richard Futrell, Roger P. Levy, and Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language*, 96(2):371–412.

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. Association for Computational Linguistics.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2021a. **Starting a new treebank? Go SUD!** In *Proceedings of the Sixth International Conference on Dependency Linguistics (DepLing, Syntax Fest 2021)*, pages 35–46, Sofia, Bulgaria.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2021b. **Typometrics: From implicational to quantitative universals in word order typology**. *Glossa: A Journal of General Linguistics*, 6(1):1–31.

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

Daniel Gildea and David Temperley. 2007. **Optimizing grammars for minimum dependency length**. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 184–191, Prague.

Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.

John A. Hawkins. 1994. *A Performance Theory of Order and Constituency*. Cambridge University Press, Cambridge.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. **Multilingual constituency parsing with self-attention and pre-training**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. **Constituency parsing with a self-attentive encoder**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Russell V. Lenth. 2024. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.10.0.

- Haitao Liu. 2008. [Dependency distance as a metric of language](#). *Journal of Cognitive Science*, 9(2):159–191.
- Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. [Universal Dependencies v1: A multilingual treebank collection](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).
- Adam Okraśniński. 2023. Minimalizacja długości zależności w strukturach współrzędnie złożonych: badanie korpusowe na podstawie banku drzew Składnica. Bachelor’s thesis, University of Warsaw.
- 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(17).
- Bartosz Pobożniak. 2024. Analysing dependency structure of coordination using a constituency parser and Dependency Length Minimization. Bachelor’s thesis, University of Warsaw.
- Martin Popel, David Mareček, Jan Štěpánek, Daniel Zeman, and Zdeněk Žabokrtský. 2013. [Coordination structures in dependency treebanks](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 517–527, Sofia, Bulgaria.
- Adam Przepiórkowski, Magdalena Borysiak, and Adam Głowacki. 2024. [An argument for symmetric coordination from Dependency Length Minimization: A replication study](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1021–1033, Torino, Italy. ELRA and ICCL.
- Adam Przepiórkowski and Agnieszka Patejuk. 2019. [Nested coordination in Universal Dependencies](#). In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 58–69. Association for Computational Linguistics.
- Adam Przepiórkowski and Michał Woźniak. 2023. [Conjunct lengths in English, Dependency Length Minimization, and dependency structure of coordination](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15494–15512, Toronto, Canada. Association for Computational Linguistics.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A Python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.
- R Core Team. 2024. [R: A Language and Environment for Statistical Computing](#). R Foundation for Statistical Computing, Vienna.
- Manuela Sanguinetti and Cristina Bosco. 2014. Part-TUT: The Turin University Parallel Treebank. In Roberto Basili, Cristina Bosco, Rodolfo Delmonte, Alessandro Moschitti, and Maria Simi, editors, *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, pages 51–69. Springer-Verlag.
- Sebastian Schuster and Christopher D. Manning. 2016. [Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, pages 2371–2378, Portorož, Slovenia. European Language Resources Association (ELRA).
- Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. [A gold standard dependency corpus for English](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 2897–2904, Reykjavík, Iceland. European Language Resources Association (ELRA).
- Wojciech Stempniak. 2024a. 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 (TLT 2024)*, Hamburg, Germany.
- Wojciech Stempniak. 2024b. Struktura zależnościowa koordynacji – analiza korpusów Universal Dependencies. Bachelor’s thesis, University of Warsaw.
- David Temperley and Daniel Gildea. 2018. Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, 4:67–80.
- Kamil Tomaszek. 2023. Minimalizacja długości zależności w strukturach współrzędnie złożonych: badanie korpusowe na podstawie Polish Dependency Bank. Bachelor’s thesis, University of Warsaw.
- Marcin Woliński, Katarzyna Głowińska, and Marek Świdziński. 2011. A preliminary version of Składnica—a treebank of Polish. In *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 299–303, Poznań, Poland.

- Marcin Woliński, Elżbieta Hajnicz, and Tomasz Bartosiak. 2018. *A new version of the Składnica treebank of Polish harmonised with the Walenty valency dictionary*. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1839–1844, Paris, France. European Language Resources Association (ELRA).
- Alina Wróblewska. 2014. *Polish Dependency Parser Trained on an Automatically Induced Dependency Bank*. Ph.D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw.
- Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.
- Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Erik Andersen, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arican, Þórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basnov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Yifat Ben Moshe, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaa, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Lauren Cassidy, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Daniela Corbetta, Marine Courtin, Mihaela Cristescu, Philemon Daniel, Elizabeth Davidson, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Drogonova, Puneet Dwivedi, Hanne Eckhoff, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaz Erjavec, Aline Etienne, Wograiné Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabricio Ferraz Gerardi, Kim Gerdes, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Oľájdé Ishola, Kaoru Ito, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóga, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Oğuzhan Kuyrukçu, Asli Kuzgun, Sookyong Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Krister Lindén, Nikola Ljubešić, Olga Loginova, Stefano Lusito, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Tatiana Merzhevich, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHosseini Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisepp, Pinkey Nainwani, Mariam Nakhilė, Juan Ignacio Navarro Horňáček, Anna Nedoluzhko, Gunta Nešpore-Bėrzkalne, Manuela Nevaci, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayo Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik,

Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Alessio Palmero Aprosio, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelika Peljak-Łapińska, Siyao Peng, Cene-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleks Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Saniyar, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Dmitry Sichinava, Janine Siewert, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Shafi Sourov, Carolyn Spadine, Rachele Sprugnoli, Vivian Stamou, Steinþór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilian Wendt, Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrummyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir

Zeldes, He Zhou, Hanzhi Zhu, Anna Zhuravleva, and Rayan Ziane. 2024. [Universal Dependencies 2.14](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A First Look at the Ugaritic Poetic Text Corpus

Tillmann Dönicke^{1,2}, Clemens Steinberger^{1,3}, Max-Ferdinand Zeterberg^{1,2}, Noah Kröll^{1,3}

¹University of Göttingen

²Göttingen State and University Library, Platz der Göttinger Sieben 1, D-37073 Göttingen

³Faculty of Theology, Platz der Göttinger Sieben 2, D-37073 Göttingen

Contact: ugarit@uni-goettingen.de

Abstract

For the Ugaritic poetic texts there is currently no digital corpus including extensive philological and poetological annotations. Within the research project “Edition des ugaritischen poetischen Textkorpus” (EUPT), these texts are digitised and provided as an online-accessible corpus. This paper briefly introduces the project and outlines the principles of the data model. The focus is on the different annotation levels and their connection with each other.

1 Introduction

1.1 Ugarit and Ugaritic

The kingdom of Ugarit, located on the northern Syrian Mediterranean coast, had its heyday in the 14th and 13th centuries BC.¹ Its territory covered large parts of today’s Syrian province of Latakia with its most important archaeological sites being Tell Ras Shamra and Ras Ibn Hani. Ugarit was a significant trading hub. Since the middle of the 14th century BC, it was a vassal state of the Hittite Empire. Shortly after 1200 BC, Ugarit was destroyed by unknown conquerors and the kingdom fell into oblivion.

In 1929, archaeologists discovered the first clay tablets preserving texts written in cuneiform alphabetic script. This script was probably brought into use in Ugarit in the 13th century BC. It was mainly employed to record texts in Ugaritic (the local Northwest Semitic language), including a number of poetic texts (e.g., epics/myths, prayers and incantations). The Ugaritic alphabet covers 30 signs. The script is primarily consonantal; the texts’ vocalisation is to be reconstructed as part of the modern philological analysis.

¹For up-to-date summaries and exhaustive references to secondary literature on Ugarit and Ugaritic, see [Tropper and Vita \(2020, p. 15–41\)](#).

1.2 Related Work

To date, most collections of Ugaritic texts have been published in print only (e.g., [Smith, 1994](#); [Smith and Pitard, 2009](#); [Pardee, 1997](#); [Parker, 1997](#)). A notable exception is the “Ras Shamra Tablet Inventory” (RSTI)² ([Prosser, 2018](#)), which provides a digital collection of Ugaritic texts as part of the University of Chicago’s OCHRE Data Service.³ RSTI includes metadata and transliterations for each tablet. Several texts are vocalised, translated and morphologically annotated. Further, the transliterations from [Cunchillos et al. \(2003\)](#) (Ugaritic Data Bank) are offered as a module of the *Accordance Bible Software* (for a fee; selected texts are morphologically annotated and translated). [Zemánek \(2007a,b\)](#) outlines the construction of a treebank for Ugaritic, but did not make such a treebank available.

For other texts from ancient West Asia there are more electronic resources available than for the Ugaritic texts (e.g., for Sumerian, Akkadian and Hittite sources; an overview is given on the openDANES website⁴). For instance, there has been developed a Universal Dependencies⁵ treebank based on a sample of Akkadian royal inscriptions ([Luukko et al., 2020](#)).

1.3 EUPT

Although the Ugaritic poetic texts have already been treated several times, there is still no comprehensive digital corpus reflecting the latest state of research. Also, there is no existing corpus of Ugaritic texts (or other cuneiform texts from ancient West Asia) that includes annotations of their poetic structure or their stylistic and motivic features. The research project “Edition des ugaritischen poetischen Textkorpus” (EUPT) aims to

²<https://voices.uchicago.edu/rsti/>

³<https://digitalhumanities.uchicago.edu/project/ochre/>

⁴<https://opendanes.org/nav/DANES-resources.html>

⁵<https://universaldependencies.org/>

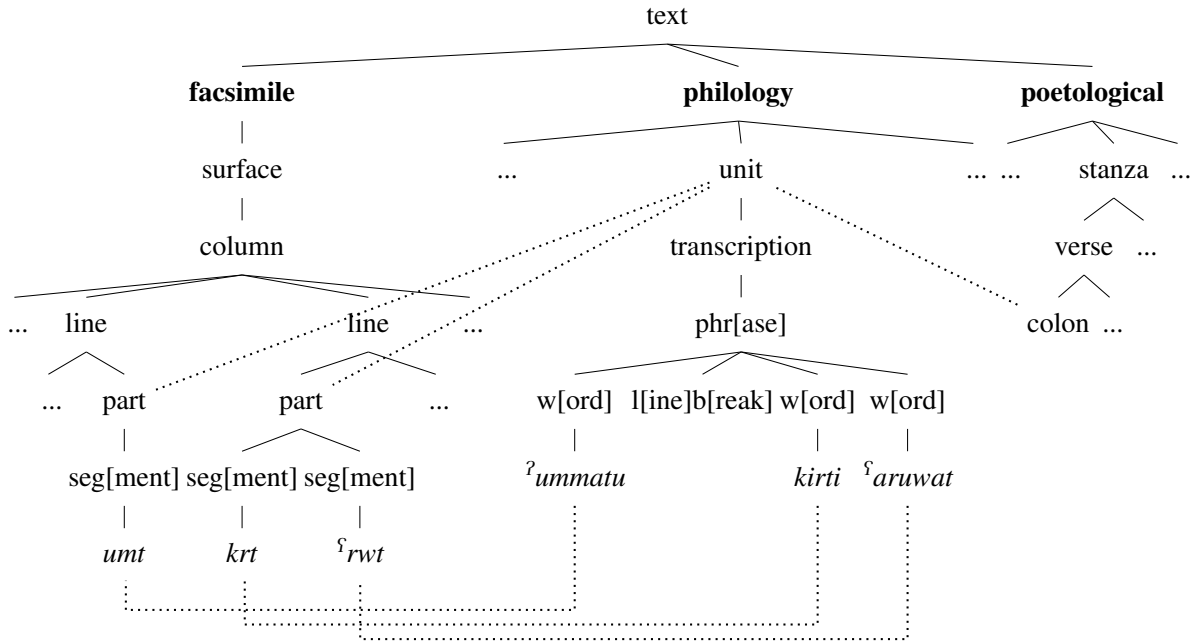


Figure 1: Pruned tree representation of a text, showing a single clause in the epigraphic (left), philological (middle) and poetological (right) (sub-)tree. Ellipses (...) indicate siblings of the same type; parentheses (e.g., w[ord]) write out abbreviations; dashed lines indicate connections via IDs.

close this gap: EUPT is preparing a digital edition of all known Ugaritic poetic texts. The texts are transliterated, vocalised and translated, as well as morphosyntactically and poetologically annotated; further included are hand copies of the tablets and commentaries on the philological reconstruction.

A key focus lies on the analysis of the texts' poetic characteristics, especially their verse structure, the forms of parallelism, various stylistic features and the motifs that the texts revolve around.⁶ Philological and poetological analysis are closely intertwined. This is evident in the linguistic structure of the texts: Lines on a tablet do often not correspond to a specific syntactic or poetic unit. In EUPT, the texts are not only prepared line-by-line according to the original tablets' layout, but also in their reconstructed linguistic/poetic structure. Reconstructing the texts' verse structure is a crucial prerequisite for an adequate edition of the texts.

2 EUPT's Three-Fold Annotation Scheme

In EUPT, the texts are annotated on three linguistic levels: the epigraphic level (named "facsimile" in our data and the corresponding figures), the philological level and the poetological level. Figure 1 shows a simplified excerpt from the corpus, com-

prising only one clause / three words. The three levels constitute separate sub-trees under the root node, but linguistic units that correspond to each other are connected via unique identifiers (IDs).

2.1 Epigraphic Tree

The epigraphic tree aims to represent the original tablet surface. It is structured into columns, lines, parts and segments, where segments correspond to words and parts are auxiliary elements that build up units (= clauses) in the philological and the poetological tree.⁷ The segment nodes contain the transliterations of the respective cuneiform signs. Since alphabetic cuneiform generally does not represent vowels, the transliteration also does not include vowels (see the leaf nodes in the left sub-tree in Figure 1). Damaged parts and segments on the tablet, or such that are completely broken-off, are annotated accordingly. Furthermore, potential misspellings by the scribe and their most probable corrections are annotated (e.g., when the scribe wrote *d* but probably meant *b*).

2.2 Philological Tree

The philological tree aims to capture the linguistic/ logical structure of the text. A text is segmented

⁶A subset of the project's poetological glossary is already published on the EUPT website: https://eupt.uni-goettingen.de/lab/Glossar_der_ugaritischen_poetischen_Formen.html

⁷In case of enjambment a clause can also be annotated as several units. These cases are specifically annotated in the poetological tree.

into units/clauses, which are connected to their corresponding parts in the epigraphic tree. Since a unit can contain a line break but a part cannot exceed a line, one unit can correspond to multiple parts (see Figure 1). Each unit has a German translation⁸ (not shown in Figure 1) and a transcription, that is further segmented into (possibly nested) phrases and words. The word nodes contain the vocalised words (see the leaf nodes in the middle sub-tree in Figure 1) and are connected to the corresponding segment nodes that contain the unvocalised words. Elements that are annotated as damaged in the epigraphic tree are also annotated as damaged in the transcription and the translation. All words are annotated with lemma and morphological analysis.

2.3 Poetological Tree

The poetological tree segments a text into stanzas (= strophes), verses and cola. The colon corresponds to the clause annotated as unit in the philological tree (see the connection in Figure 1). The colon nodes of the poetological tree do not contain any string content, since the unvocalised and the vocalised text is already represented in the epigraphic and the philological tree. Poetic devices, such as semantic/grammatical parallelism, enjambement, metaphor and others, are annotated on stanza, verse and colon level.

3 Corpus

The digital corpus construction is divided into three phases. In the first phase, the *Kirtu* epic (KTU 1.14–16⁹), the *Aqhatu* epic (KTU 1.17–19) and the *Rāpi'ūma*-fragments (KTU 1.20–22) are digitised. The second phase is devoted to the *Ba^llu* cycle (KTU 1.1–6). In the third phase, the shorter mythological texts, prayers, incantations and ritual texts that contain poetic forms will be digitised. The entire corpus contains 68 tablets with about 4,000 lines.

3.1 Annotation Process

The project's Ugaritology team consists of three Ugaritic experts (one PhD candidate, one postdoc and one professor), who are assisted by five students with adequate knowledge of Ugaritic. In a first step, a text is transliterated, vocalised and translated by one of the Ugaritic experts. After that, a student assistant segments the text into

⁸English translations of the texts are planned to be added in the future.

⁹The KTU identifiers refer to [Dietrich et al. \(2013\)](#)

words, phrases, parts etc. and annotates each word's lemma and morphology,¹⁰ while an Ugaritic expert performs the poetological annotation. All annotations are reviewed collaboratively by the three experts and corrected when necessary.

3.2 Commentary

Another special feature of EUPT is that the Ugaritic experts also add philological comments/notes to all texts. In theory, each element in a text's tree can be annotated with notes. In practice, most lines or even words have been annotated with notes about their epigraphy, reconstruction, lexis, grammar, content or poetology. An example follows in the next subsection.

3.3 Format

The Ugaritic texts and their annotations are stored as XML trees (based on TEI-XML¹¹), where each column of a tablet has its own XML file. The outer structure of an XML file looks as follows:

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <edxml
   xmlns="http://sub.uni-goettingen.de/edxml#"
   xmlns:tei="http://www.tei-c.org/ns/1.0">
3 <header>
4 ...
5 </header>
6 <text>
7 <facsimile xml:id="fac1_1">
8 ...
9 </facsimile>
10 <philology xml:id="phil_1">
11 ...
12 </philology>
13 <structure type="poetological" xml:id="poet_1">
14 ...
15 </structure>
16 </text>
17 </edxml>

```

The header element (ll. 3–5) contains metadata, such as title and license. The text element (ll. 6–16) contains the three sub-trees (epigraphic, philological, poetological).

To keep it short, we only present an excerpt from the philological tree in Figure 2, showing the same unit as in Figure 1. The unit element (ll. 4–18) contains a transcription, a translation and a notes element. The transcription element (ll. 5–12) stores the syntactic structure of the clause, which consists of one phr[ase] element (ll. 6–11) that contains three w[ord] elements (ll. 7, 9, 10) and one l[in]e[b]reak element (l. 8). The corresp attributes in the unit and the w elements store the IDs

¹⁰Our tagset is based on [Tropper \(2012\)](#) and can be found at https://eupt.uni-goettingen.de/Einfuehrung/Editorische_Prinzipien_Kommentar.html (> Editorische Prinzipien > Kommentar).

¹¹<https://tei-c.org/>

```

1 <philology xml:id="phil_1">
2 <units>
3 ...
4 <unit xml:id="unit_1.14_I_7" n="i 6b-7a" corresp="#line_1.14_I_6_2 #line_1.14_I_7_1">
5 <transcription type="vocalisation">
6 <phr xml:id="phr_vqv_34m_q1c">
7 <w xml:id="opc_rlr_pzb" corresp="#seg_jcy_2jv_4zb" lemma="lemma:umt-1" ana="Nom.f.Sg.
  St.cstr.">iummatu</w>
8 <lb n="i 7"/>
9 <w xml:id="u1m_rlr_pzb" corresp="#seg_vpx_d4v_4zb" lemma="lemma:krt-1" ana="PN
  Gen.m.Sg."><tei:damage>kirti</tei:damage></w>
10 <w xml:id="r2b_slr_pzb" corresp="#seg_k3l_jjv_4zb" lemma="lemma:rw-1" ana="G-SK 3.f.Sg. / alt. Vok.:
  arawat"><tei:damage degree="low">sa</tei:damage>ruwat</w>
11 </phr>
12 </transcription>
13 <translation xml:lang="de">Die Sippe <tei:damage>Kirtu</tei:damage> war entblöbt,</translation>
14 <notes>
15 <note type="lx" target="#r2b_slr_pzb">
16 <label>i 6b-7a: <textBlock>iRWT</textBlock></label>
17 <p><textBlock>iRWT</textBlock>, wörtl. <quote>"sie war entblöbt / nackt"</quote>, im übertragenen Sinn
  <quote>"sie war vernichtet (/ leer)"</quote> (KWU 20 s.v. <hi>srw</hi> G; vgl. auch <bibl
  zotero="eupt:SBKFHDM"de Moor, 1987: 192 Anm. 4</bibl>).</p>
18 <p>Etymologisch ist <textBlock>iRW</textBlock> &#60; <textBlock>iRWT</textBlock> mit der in
  verschiedenen semitischen Sprachen bezugten Wurzel
  <textBlock>iry</textBlock>/<textBlock>w</textBlock> <quote>"nackt sein"</quote> zu verbinden (KWU 20
  s.v. <hi>srw</hi>). Anders del Olmo Lete / Sanmartín (DUL&#179; 182 s.v. <hi>s-r-w</hi>) und
  <hi>s-r-y</hi>): Sie verknüpfen ug. <textBlock>iRW</textBlock> (<quote>"to be consumed"</quote>) mit
  ar. /<textBlock>sarā</textBlock>/ / <textBlock>srw</textBlock>. Das ar. Verb bedeutet jedoch nicht
  <quote>"aufbrauchen"</quote> oder <quote>"vernichtet sein"</quote> o. Ä. (beachte die Form
  <textBlock><tei:choice><tei:corr>ITBD</tei:corr></tei:choice></textBlock> <quote>"[das Haus] war
  völlig zerstört"</quote>, die in der <hi>Kirtu</hi>-Passage parallel zu <textBlock>iRWT</textBlock>
  steht), sondern <quote>"aufsuchen, besuchen; heimsuchen, überkommen"</quote> (<bibl
  zotero="eupt:AV8XGRME">AEL 2027-2028</bibl>; <bibl zotero="eupt:KSEIKH8R">Wehr / Kropfitsch, 2020:
  609</bibl>; ausgehend vom vermeintlichen ar. Kognat analysieren del Olmo Lete / Sanmartín
  <textBlock>iRWT</textBlock> in KTU 1.14 i 6b-7a als Gp-Form). Von <textBlock>iRW</textBlock> &#60;
  <textBlock>iRWT</textBlock> unterscheiden sie <textBlock>sRY</textBlock> <quote>"to be
  naked"</quote>. Für <textBlock>sRY</textBlock> verweisen sie auf die Form <textBlock>sRYT</textBlock>
  in KTU 2.38 24-25 (dort bezogen auf ein Schiff; nach UG&#178; 569 wahrscheinlich <quote>"es [scil.
  das Schiff] wurde entleert"</quote> oder <quote>"es wurde 'entkleidet' [i. e. die Segel des Schiffes
  wurden entfernt]"</quote>; DUL&#179; 182 s.v. <hi>s-r-y</hi>: <quote>"[it] is unrigged"</quote>; del
  Olmo Lete / Sanmartín analysieren <textBlock>sRYT</textBlock> als G-SK-Form, Tropper [UG&#178; 569]
  als Dp-SK-Form [alt. als D-SK-Form]). Vermutlich sind die Formen <textBlock>iRWT</textBlock> (in der
  <hi>Kirtu</hi>-Passage) und <textBlock>sRYT</textBlock> (in KTU 2.38 24-25) jedoch beide auf das sem.
  Verb <textBlock>sry</textBlock>/<textBlock>w</textBlock> <quote>"nackt sein"</quote> zurückzuführen
  (zu <textBlock>sRYT</textBlock> vgl. UG&#178; 195 / 569).</p>
19 </note>
20 ...
21 </notes>
22 </unit>
23 ...
24 </units>
25 </philology>

```

Figure 2: XML excerpt of a text, showing the clause from Figure 1 in the philological tree.

of the corresponding part and seg[ment] elements in the epigraphic tree (not shown in Figure 2, but in Figure 1). The lemma and ana attributes of the w elements store a word's lemma and morphological analysis, respectively. The embedded tei:damage elements (ll. 9, 10) mark up damaged signs on the tablet. The translation element stores a translation of the unit. The notes element (ll. 14–17) contains the philological notes (see previous subsection) as individual note elements. Each note element (l. 15) has a type attribute (here lx for “lexicographic”) and a target attribute with the ID of the element that the note refers to (here the ID of the third w element). Notes are in German; a translation of this note can be found in Appendix A.

3.4 Ambiguities and Alternatives

The analysis of the Ugaritic texts is beset with ambiguities and uncertainties, primarily due to the fact that most tablets are incomplete and, moreover, the grammatical analysis and vocalisation of the consonantal texts remain a matter of debate. EUPT provides various annotations to indicate that remnants of a sign are unidentifiable (<pc type="non_identifiable_sign_single"/>), that the identification or reconstruction of a grapheme, phoneme or word *x* is uncertain (<w cert="low">*x*</w>), or that the scribe mistakenly included a sign *x* (<tei:surplus>*x*</tei:surplus>), omitted *x* (<tei:supplied>*x*</tei:supplied>), erased or overwrote *x* (<tei:del>*x*</tei:del>), or wrote *x* instead

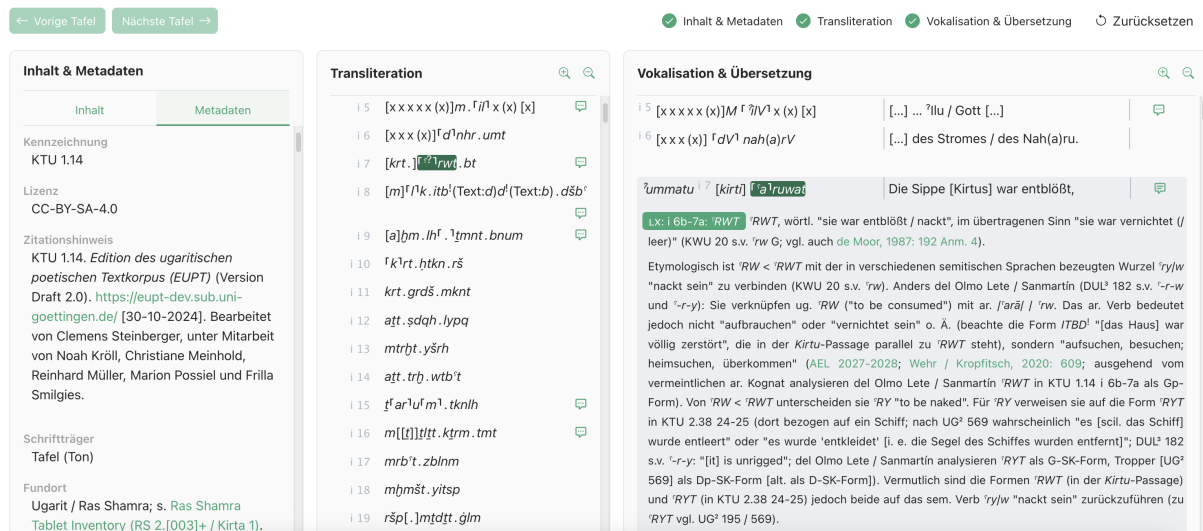


Figure 3: TIDO with three panels (left to right): “Content & Metadata”, “Transliteration” (for epigraphic information), and “Vocalisation & Translation” (for philological and poetological information). The excerpt from Figures 1 and 2 can be seen from line i 6 (last word) to line i 7 (first and second word). Note that lines in the inner panel correspond to actual lines on the tablet, while lines in the right panel correspond to clauses (hence the line number i 7 appears in the middle of the line). When a user hovers over a word (here: $\text{r}^{\text{f}}\text{a}^{\text{r}}\text{ruwat}$), it is highlighted in both panels. Philological notes can be expanded by buttons at the end of the corresponding lines (speech-bubble symbol).

of y ($\langle \text{tei:choice} \rangle \langle \text{tei:corr} \rangle y \langle / \text{tei:corr} \rangle \langle \text{tei:sic} \rangle x \langle / \text{tei:sic} \rangle \langle / \text{tei:choice} \rangle$). These annotations correspond to the notation conventionally used in Ugaritic transliterations/vocalisations, e.g., $m \langle \text{tei:del} \rangle \langle \text{t} \rangle \langle / \text{tei:del} \rangle \langle \text{t} \rangle \langle \text{t} \rangle$ is typically represented as $m[\langle \text{t} \rangle] \langle \text{t} \rangle \langle \text{t} \rangle$. Any more far-reaching uncertainties relating to the philological analysis are discussed in the commentaries.

3.5 Access

Access to the corpus data is provided on the EUPT website at <https://eupt.uni-goettingen.de/edition.html>. On the back-end side, the XML files are converted to HTML using handwritten XSLT rules. The HTML files are then embedded into the website using the interactive Text Viewer for Digital Objects (TIDO)¹² (Göbel et al., 2024). Figure 3 shows the TIDO interface on the website.

A workflow for versioned releases of the raw XML and HTML files is currently under development. Meanwhile, it is possible to view the raw files at <https://gitlab.gwdg.de/subugoe/eupt/eupt-textapi/-/tree/main/assets>. Note that only one tablet has been fully digitised and made accessible so far. Given the nascent state of the project, we advise contacting the authors directly to request access to the data.

¹²<https://www.sub.uni-goettingen.de/en/digital-library/digital-tools/text-viewer-for-digital-objects-tido-textapi/>

4 Future Work

Until 2032, the entire Ugaritic poetic text corpus shall be digitised and fully annotated, including the annotation of grammatical roles. New features will be successively implemented on the project website, including display of hand copies of the tablets, visualisation of all philological and poetological annotations, online publication of the project’s lexical glossary, search tools, and additional options for users to configure the corpus view.

In the future, the XML data will be extended by a graph database—currently, the team is developing the graph data model. The text-as-graph approach (cf. Kuczera, 2016) will open up new possibilities for the granular annotation of different elements and their relationships to each other.

Acknowledgments

The research project “Edition des ugaritischen poetischen Textkorpus” (EUPT) is funded by the Deutsche Forschungsgemeinschaft (project number: 506107876). The first version of the XML scheme was prepared by Uwe Sikora. The developer team was supervised by Michelle Weidling and Mona Orloff. The Ugaritological team was supervised by Reinhard Müller. For the current members of the project team, please visit EUPT’s website at <https://eupt.uni-goettingen.de/Einfuehrung/Team.html>.

References

- Jesús-Luis Cunchillos, Juan-Pablo Vita, José Ángel Zamora, and Raquel Cervigón. 2003. *The texts of the Ugaritic data bank / Ugaritic Data Bank*. Laboratorio de Hermeneumática, Madrid.
- Johannes de Moor. 1987. *An Anthology of Religious Texts from Ugarit*, volume 16 of *Nisaba*. Brill, Leiden / New York / Copenhagen / Köln.
- Gregorio del Olmo Lete and Joaquín Sanmartín. 2015. *A Dictionary of the Ugaritic Language in the Alphabetic Tradition*. Translated and edited by Wilfred G. E. Watson, third, revised edition, volume 112 of Wilfred H. van Soldt, editor, *Handbook of Oriental Studies. Section 1. The Near and Middle East*. Brill, Leiden / Boston.
- Manfried Dietrich, Oswald Loretz, and Joaquín Sanmartín. 2013. *Die keilalphabetischen Texte aus Ugarit, Ras Ibn Hani und anderen Orten / The Cuneiform Alphabetic Texts from Ugarit, Ras Ibn Hani and Other Places*, third, enlarged edition, volume 360/1 of Manfried Dietrich, Oswald Loretz, and Hans Neumann, editors, *Alter Orient und Altes Testament*. Ugarit-Verlag, Münster.
- Mathias Göbel, Michelle Weidling, and Paul Pestov. 2024. [TextAPI and TIDO - An example for research software product development](#). In *Book of Abstracts, deRSE24 - Conference for Research Software Engineering in Germany*, pages 76–77. Julius-Maximilians-Universität Würzburg.
- Andreas Kuczera. 2016. [Digital Editions beyond XML – Graph-based Digital Editions](#). In *Proceedings of the 3rd HistoInformatics Workshop on Computational History (HistoInformatics 2016) co-located with Digital Humanities 2016 conference (DH 2016)*, pages 37–46, Krakow, Poland.
- Edward William Lane. 1863–1893. *An Arabic-English Lexicon*. Williams and Norgate, London / Edinburgh.
- Mikko Luukko, Aleksis Sahala, Sam Hardwick, and Kristin Lindén. 2020. [Akkadian Treebank for early Neo-Assyrian Royal Inscriptions](#). In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 124–134, Düsseldorf, Germany. Association for Computational Linguistics.
- Dennis Pardee. 1997. The Ba^šlu Myth / The Kirta Epic / The [?]Aqhatu Legend. In William W. Hallo, editor, *The Context of Scripture. Volume I. Canonical Compositions from the Biblical World*, pages 241–274 / 333–343 / 343–356. Brill, Leiden / Boston.
- Simon B. Parker, editor. 1997. *Ugaritic Narrative Poetry*, volume 9 of Simon B. Parker, editor, *Writings from the Ancient World*. Society of Biblical Literature, Atlanta.
- Miller C. Prosser. 2018. [Digital Philology in the Ras Shamra Tablet Inventory Project: Text Curation through Computational Intelligence](#). In Vanessa Bigot Juloux, Amy Rebecca Gansell, and Alessandro Di Ludovico, editors, *CyberResearch on the Ancient Near East and Neighboring Regions: Case Studies on Archaeological Data, Objects, Texts, and Digital Archiving*, chapter 10, pages 314–335. Brill, Leiden, Niederlande.
- Mark S. Smith. 1994. *The Ugaritic Baal Cycle. Volume 1. Introduction with Text, Translation and Commentary of KTU 1.1–1.2*, volume 55 of John A. Emerton et al., editors, *Supplements to Vetus Testamentum*. Brill, Leiden / New York / Köln.
- Mark S. Smith and Wayne T. Pitard. 2009. *The Ugaritic Baal Cycle. Volume 2. Introduction with Text, Translation and Commentary of KTU/CAT 1.3–1.4*, volume 114 of Hans M. Barstad et al., editors, *Supplements to Vetus Testamentum*. Brill, Leiden / Boston.
- Josef Tropper. 2008. *Kleines Wörterbuch des Ugaritischen*, volume 4 of Reinhard G. Lehmann, and Josef Tropper, editors, *Elementa Linguarum Orientis*. Harrassowitz, Wiesbaden.
- Josef Tropper. 2012. *Ugaritische Grammatik. Zweite, stark überarbeitete Auflage*, volume 273 of Manfried Dietrich, Oswald Loretz, and Hans Neumann, editors, *Alter Orient und Altes Testament*. AUgarit-Verlag, Münster.
- Josef Tropper and Juan-Pablo Vita. 2020. *Lehrbuch der ugaritischen Sprache*. Zaphon, Münster.
- Hans Wehr and Lorenz Kropfisch. 2020. *Arabisches Wörterbuch für die Schriftsprache der Gegenwart. Arabisch – Deutsch. 6., von Lorenz Kropfisch völlig neu bearbeitete und erweiterte Auflage*, sixth edition. Harrassowitz, Wiesbaden.
- Petr Zemánek. 2007a. [A treebank of Ugaritic: Annotating fragmentary attested languages](#). In *Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT2007)*, pages 213–218.
- Petr Zemánek. 2007b. [Morphological Tagging of Ugaritic](#). In *Chatreššar 2007. Electronic Corpora of Ancient Languages. Proceedings of the International Conference*, pages 135–149, Prague.

A Translation

This is a translation of the note in Figure 2:

^šRWT, literally “she was naked”, figuratively “she was destroyed (/ empty)” (KWU [= Tropper, 2008] 20 s.v. ^šrw G; cf. also de Moor, 1987: 192 note 4).

Etymologically, ^šRW < ^šRWT is to be connected with the root ^šry/w “to be naked” attested in various Semitic languages (KWU 20 s.v. ^šrw). Differently del Olmo Lete / Sanmartín (DUL³ [= del Olmo Lete and Sanmartín, 2015] 182

s.v. r-w and r-y): They link ug. RW (“to be consumed”) with ar. $\text{arā} / \text{rw}$. However, the ar. verb does not mean “to consume” or “to be consumed” or the like (note the form *ITBD*¹ “[the house] was completely destroyed”, which is parallel to RWT in the *Kirtu* passage), but “to visit; come upon” (AEL [= Lane, 1863–1893] 2027-2028; Wehr and Kropfitch, 2020: 609; based on the supposed ar. cognate del Olmo Lete / Sanmartín analyze RWT in KTU 1.14 i 6b-7a as a Gp form). From $\text{RW} < \text{RWT}$ they distinguish RY “to be naked”. For RY they refer to the form RYT in KTU 2.38 24-25 (there referring to a ship; following UG² [= Tropper, 2012] 569 probably “it [scil. the ship] was emptied” or “it was ‘stripped’ [i. e. the sails of the ship were removed]”; DUL³ 182 s.v. r-y : “[it] is unrigged”; del Olmo Lete / Sanmartín analyze RYT as G-SC form, Tropper [UG² 569] as Dp-SC form [alternatively as D-SC form]). Presumably, however, the forms RWT (in the *Kirtu* passage) and RYT (in KTU 2.38 24-25) are both to be derived from the semitic root ry/w “to be naked” (on RYT cf. UG² 195 / 569).

LuxBank: The First Universal Dependency Treebank for Luxembourgish

Alistair Plum, Caroline Döhmer, Emilia Milano,

Anne-Marie Lutgen, Christoph Purschke

University of Luxembourg

Esch-sur-Alzette, Luxembourg

{alistair.plum, caroline.doehmer, emilia.milano}@uni.lu

{anne-marie.lutgen, christoph.purschke}@uni.lu

Abstract

The Universal Dependencies (UD) project has significantly expanded linguistic coverage across 161 languages, yet Luxembourgish, a West Germanic language spoken by approximately 400,000 people, has remained absent until now. In this paper, we introduce LuxBank, the first UD Treebank for Luxembourgish, addressing the gap in syntactic annotation and analysis for this ‘low-research’ language. We establish formal guidelines for Luxembourgish language annotation, providing the foundation for the first large-scale quantitative analysis of its syntax. LuxBank serves not only as a resource for linguists and language learners but also as a tool for developing spell checkers and grammar checkers, organising existing text archives and even training large language models. By incorporating Luxembourgish into the UD framework, we aim to enhance the understanding of syntactic variation within West Germanic languages and offer a model for documenting smaller, semi-standardised languages. This work positions Luxembourgish as a valuable resource in the broader linguistic and NLP communities, contributing to the study of languages with limited research and resources.

1 Introduction

The Universal Dependencies (UD) project has facilitated the production of treebanks across many languages, although some languages are still not represented almost 10 years after its original release (Nivre et al., 2016). With 161 languages represented as of the latest release, and a total of 283 treebanks across these languages, the language coverage is undeniably vast.¹ The range of languages includes many of the major world languages, as well as varieties and dialects. However, some languages are still not represented at all, and Luxembourgish was one such case until recently.

¹Latest release at the time of writing: 15.05.2024.

A West Germanic language closely related to German, Luxembourgish is spoken by roughly 400,000 people, mainly in Luxembourg (Gilles, 2019). Historically, Luxembourg has had a complex multilingual society where French and German have been predominantly used for official and formal (written) communication. In contrast, Luxembourgish was mostly a spoken language used informally between Luxembourgers until recently. With the rise of digital and social media, however, Luxembourgish has started to develop in the written domain and significant amounts of text data have started to become available, coupled with active language policies promoting Luxembourgish. Research in Natural Language Processing (NLP) for Luxembourgish has been limited until now, often in favour of French, German, and English. This has resulted in a situation where Luxembourgish is considered by some to be a ‘low-research’ language, as opposed to a low-resource language.

In addition, large-scale syntactic annotation and analysis has not been undertaken before for Luxembourgish, making Luxembourg one of the few countries whose national language is not represented in the UD treebanks. This remains true despite the fact that four treebanks are available for Standard German (Völker et al., 2019; McDonald et al., 2013; Zeman et al., 2018; Basili et al., 2017), as well as three non-standard treebanks for Swiss German (Aeppli, 2018), Low Saxon (Siewert et al., 2021) and Bavarian (Blaschke et al., 2024). None of these represent a Middle-German variety, however, indicating an opportunity to extend the coverage for varieties of (or related to) German.

Aiming to address this gap in research, we present LuxBank, the first UD treebank for Luxembourgish. This project will be the first large-scale quantitative analysis of Luxembourgish syntax, and with this paper, we introduce the first formal guidelines for Luxembourgish language annotation. To this end, we present work related to Luxembourgish

in Section 2 and describe the creation of LuxBank in Section 3, including highlighting notable syntactic phenomena. We discuss difficulties encountered in the creation process in Section 4 and conclude the paper with Section 5.

2 Related Work

Four UD treebanks exist for German, GSD (McDonald et al., 2013), PUD (Zeman et al., 2018), LIT (Basili et al., 2017) and the largest, HDT (Völker et al., 2019), at around 189k sentences. For non-standard varieties of German there are three UD treebanks: the UZH for Swiss German (Aeppli, 2018), the LSDC for Low Saxon (Siewert et al., 2021) and as of recently, MaiBaam for Bavarian (Blaschke et al., 2024).

Two sets of guidelines for the UD project have been released since its inception, the first for version 1 (Nivre et al., 2016) and the second for version 2 (Nivre et al., 2020). As the current iteration of the project is version 2, we adhered to these guidelines, although we will discuss some aspects of the version 1 guidelines that could have been useful for our project in Section 4.

2.1 Luxembourgish Syntax

Early work on the syntax of Luxembourgish can be found in Schanen (1980) and in a few chapters of grammar books (Schanen and Zimmer, 2012). Certain characteristics of Luxembourgish syntax were later on investigated by dialectologists working on syntactic phenomena in West Germanic (Glaser, 2006) or presented in overview papers on Luxembourgish (Gilles, 2023). A more in-depth analysis of syntactic features was conducted by Döhmer (2020), and there are studies on neighbouring topics, namely pronominal reference for female persons (Martin, 2019) and variation in inflectional morphology (Entringer, 2022), but linguistics research on Luxembourgish syntax and on grammar in general is still in its beginnings. As there is relatively little research literature, we will invest more time into detecting, discussing, and categorising syntactic phenomena parallel to the annotation.

2.2 Luxembourgish NLP

Luxembourgish is underrepresented in NLP compared to its linguistic neighbours, French and German. Early research includes resources for NLP tasks (Adda-Decker et al., 2008), analysis of writing patterns (Snoeren et al., 2010), and a corpus

for language identification (Lavergne et al., 2014). Recent advancements feature sentiment analysis pipelines (Sirajzade et al., 2020; Gierschek, 2022), an orthographic correction pipeline (Purschke, 2020), a zero-shot topic classification approach (Philippy et al., 2024), and automatic comment moderation (Ranasinghe et al., 2023). LUX-ASR provides Automatic Speech Recognition for Luxembourgish (Gilles et al., 2023a,b), while language models like LUXGPT leverage transfer learning from German (Bernardy, 2022). Additionally, LUXEMBERT matches multilingual BERT’s performance in Luxembourgish tasks (Lothritz et al., 2022, 2023), and ENRICH4ALL supports a multilingual chatbot in administrative contexts (Anastasiou, 2022). While some tools and models exist for basic language processing, such as a limited spaCy integration² and the python tool spellux for lemmatisation³, there is no published work on these tasks.

3 LuxBank

In this section, we set out the methodology for the first round of annotations for LuxBank and reflect on specific linguistic conditions, such as standardisation and structural properties of Luxembourgish. The initial steps include translating the Cairo CILing sentences, setting up preprocessing, as well as defining the annotation process. For the continuation of this project, we present the next steps in section 3.4, which are focused on adding further sentences from various domains of writing.

The project group working on LuxBank is made up of researchers from a range of different disciplines and specialisations: Two PhD researchers from the research project TRAVOLTA⁴ with a background in linguistics, one expert for Luxembourgish grammar and syntax, and one computational linguist specialising in NLP for Luxembourgish. This is of central importance to our approach, as we are trying to incorporate computational processing and linguistic analysis on an equal footing in the development of the project. This is also due to the fact that linguistic experts are often underrepresented in computational linguistics projects. In the following, we describe the data annotation and analysis process.

²<https://github.com/PeterGilles/Luxembourgish-language-resources/blob/master/spaCyforLuxembourgish.ipynb>

³<https://github.com/questoph/spellux>

⁴<https://purschke.info/en/travolta>

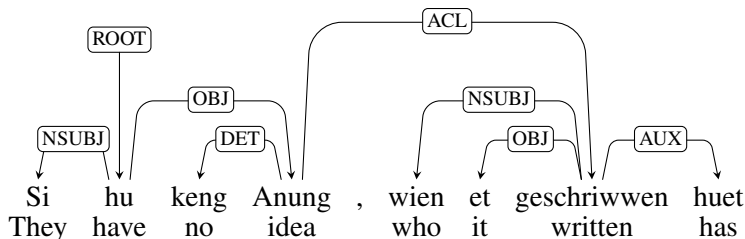


Figure 1: Auxiliary verb in sentence c12.

The Luxembourgish language is not fully standardised and presents a considerable amount of variation, be it lexical, grammatical, or phonological (Entringer et al., 2021). For this project, we decided to use written Luxembourgish according to the official spelling rules.⁵ Luxembourgish has an ‘emerging standard’ and regional variants are being levelled. It is unclear whether there is significant syntactic variation stemming from the different dialects. Given the small size of the country and the ongoing efforts at standardisation, we argue that the variant of written Luxembourgish we are using comes very close to a standard language. The syntactic variation we find in the data is limited and can in most cases be explained through structural reasons.

For our first annotation set, we translate the 20 sentences from the Cairo CICLing corpus into Luxembourgish to ensure comparability. For the second round of annotations we will focus on news texts (journalistic language), as they represent a domain of formal writing and comply with the latest version of the spelling rules published in 2022.⁶ The choice of this specific written data is mainly due to practicality reasons, as those texts are easily accessible and offer a good starting point for the project. In the future, we will be open to add texts from different genres to cover a broader range of written language use in practice.

3.1 CICLing Sentences

The first 20 sentences are translated from the Cairo CICLing⁷ sentences, as recommended in the UD guide for submitting new treebanks.⁸ We use the English sentences as source language, and ask native speakers to perform the translations. We em-

⁵D’Lëtzebuenger Orthografie, Zenter fir d’Lëtzebuenger Sprooch (ZLS) 2022.

⁶D’Lëtzebuenger Orthografie, ZLS 2022.

⁷<https://github.com/UniversalDependencies/cairo>

⁸https://universaldependencies.org/release_checklist.html

ploy the available NLP resources for Luxembourgish to perform tokenisation, that is, the available Luxembourgish model for spaCy and spellux for obtaining lemmas.

Of note for our tokenisation is that we split contracted prepositions and determiners manually, which we adopt from Standard German. For the same reason we do not split hyphenated compound words. We deviate from the German guidelines with the determiner *d’*, which does not exist in German, and for which we follow the French standard of tokenising it as *d’*, therefore keeping the punctuation intact.

3.2 Annotation

After the corpus selection, the two PhDs working on this project discuss each sentence. The discussion includes analysing the syntactic structure and dependencies by referring to the UD guidelines for German⁹ and current work on Luxembourgish syntax (Döhmer, 2020). The analysis starts by annotating the Part-of-Speech (POS) tags for every token. Then, the PhDs adhere to the classic UD process by starting with the main clause, detecting the root and its dependencies with the constituents of the clause. Afterwards, the secondary clause is the main focus of the discussion, looking at the connection with the main clause and its dependencies. Then, as a further step, the two linguists consult the syntactic expert for Luxembourgish to discuss their previous decisions, make additional changes and have a final validation of the dependency annotation.

The difficulties encountered during the annotation process mainly relate to the following reasons: First, the number of people available to work on this project is limited. Since Luxembourgish grammar is not taught in school, finding student assistants who could be trained as annotators is difficult; Second, the two PhDs working on the annotations have limited experience with UD annotation; and

⁹<https://universaldependencies.org/de/>

	hunn (have)	sinn (be)	goen (go)	ginn (give)	kréien (get)	wäert (will)
main verb	+	+	+	+	+	-
copula	-	+	-	+	-	-
past tense	+	+	-	-	-	-
passive voice	-	+	-	+	+	-
subjunctive mood	-	-	+	+	-	+/-
future tense	-	-	-	-	-	+/-

Table 1: Functional properties of Luxembourgish auxiliary verbs, adapted from Nübling (2006) by Döhmer (2020).

third, sometimes there is a missing overlap of Luxembourgish grammatical phenomena with the available UD tags.

3.3 Special Linguistic Features

In this section, we introduce the syntactic phenomena that need a more thorough explanation, as the tags offered by the UD are not sufficient to cover all the grammatical details unique to the Luxembourgish sentence structure.

3.3.1 The Verbal Domain

We first focus on the verbal domain, describing the categorisation of different functional verb classes during the initial period of the project.

Auxiliary Verbs As with most of the Germanic and Romance languages, Luxembourgish has a set of auxiliary verbs to serve different grammatical purposes, such as periphrastic constructions to express the past tense, subjunctive mood, or passive voice. In general, there are six auxiliaries in Luxembourgish, namely *hunn*, *sinn*, *goen*, *ginn*, *kréien*, and *wäert*, which can also occur as lexical verbs with the meaning of, respectively, ‘to have, to be, to go, to give, to get’, with the exception of *wäert* (‘will’) which has a defective paradigm and only works as a function verb. Each of these verbs, when used as an auxiliary, has a specific function, e.g. tense or mood. When used as main verbs, these verbs are marked as *root*, while, when used as auxiliaries, they are marked as *aux*, together with modal verbs. Table 1 summarises the functional properties of the Luxembourgish auxiliary system, and Figure 1 shows an annotated sentence from LuxBank.

Modal Verbs Like other Germanic languages, Luxembourgish has a set of modal verbs that indicate the modality of the verbal phrase, i.e., if a situation/action is likely, possible, required etc. These are: *kënnen*, *mussen*, *sollen*, *däerfen* and *wëllen*, meaning, respectively, ‘can, must, shall,

may, want’. Since there is no dedicated tag for modal verbs in the UD, this category too goes under the *aux* tag. In some grammatical descriptions, they are referred to as ‘modal auxiliaries’ (Barbiers and Van Dooren, 2017). Therefore, in LuxBank grammatical auxiliaries and modal verbs are marked with the same dependency tag. An annotated example from LuxBank is shown in Figure 2.

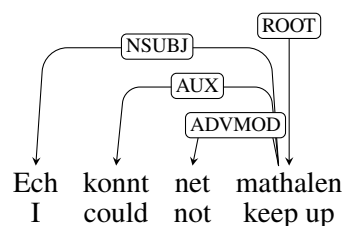


Figure 2: Modal verb in sentence c18.

Copular Verbs It is worth underlining here that Luxembourgish, like many other Germanic languages, has more than one verb which can form a copular construction, e.g. *ginn* (‘to give’) or *sinn* (‘to be’). As it is not possible to have more than one copular verb in the UD, at present, *sinn* is registered as copula, while *ginn* is only mentioned as an auxiliary. Figure 3 shows an annotated example from LuxBank.

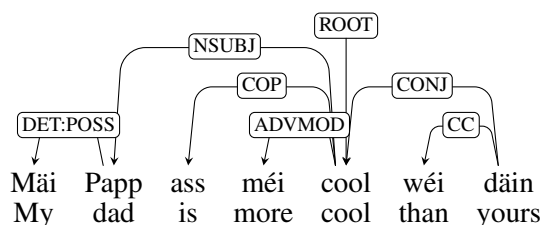


Figure 3: Copular verb in sentence c8.

Causative Verbs The verb *doen* ‘to do’ can be used to form a causative construction. Causatives indicate that a person or event is causing an action

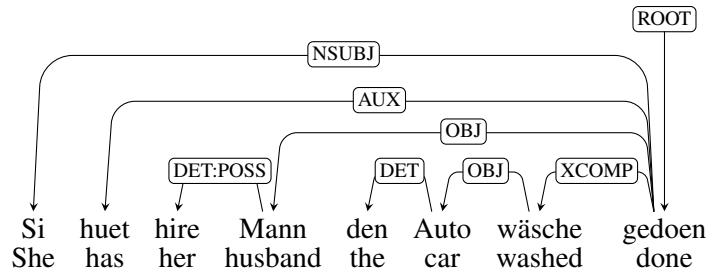


Figure 4: Causative verb in sentence c6.

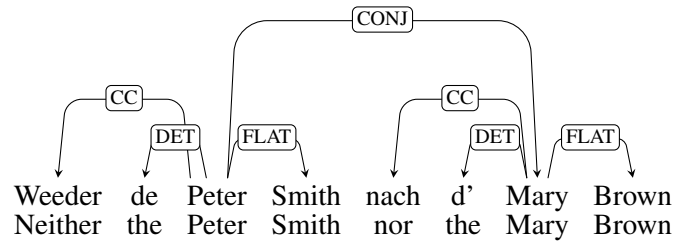


Figure 5: Determiner and proper name in sentence c11.

to happen. This auxiliary was already attested in Old and Middle High German (Hans-Bianchi and Katelhoen, 2011) and persists in Luxembourgish but not in Modern Standard German. However, the use of *doen* is very selective towards its governed verbal phrase, as it can only be combined with specific main verbs. Its status is unclear because it has the functional and structural properties of an auxiliary but the semantic properties of a lexical verb. We tag it as *root* to identify it as a lexical head rather than an auxiliary, considering its limited use and to maintain consistency within the under-specified auxiliary category. An annotated sentence featuring a causative verb is shown in Figure 4.

3.3.2 The Nominal Domain

When focusing on further syntactic elements, we find that Luxembourgish also shows a few structural peculiarities in the nominal domain which are worth mentioning.

Determiner and Proper Name A common phenomenon in Luxembourgish is the obligatory definite article before proper names. Like in any other noun phrase, the determiner is inflected based on number, gender, and case. Therefore, two or more dependencies in simple noun phrases are quite frequent, especially if the complete name of the person is mentioned. In these cases, we use the tag *det* for the determiner, and following the UD guidelines, *flat* for the second name or surname of

the person. The annotated example sentence from LuxBank is shown in Figure 5.

Possessive Constructions The genitive is not an active case in the Luxembourgish language. Possessive relations can be expressed with an adnominal dative (only for animate possessors) or with a *van*-PP (Döhmer, 2020). An annotated example sentence is shown in Figure 6.

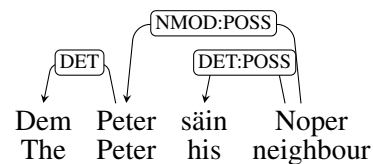


Figure 6: Possessive construction in sentence c7.

3.3.3 Other Domains

Since not every phenomenon in Luxembourgish can be analysed with the UD tagset, we decided to use the miscellaneous attributes for the annotation to explicate the phenomena. The miscellaneous attributes, labelled in the MISC column, are intended for the annotators to put in additional information about a tag.¹⁰ At the moment, there are two phenomena that are covered by this tag, the negation and the agreement marker, described in the tag set as *s* clitic.

¹⁰<https://universaldependencies.org/misc.html>

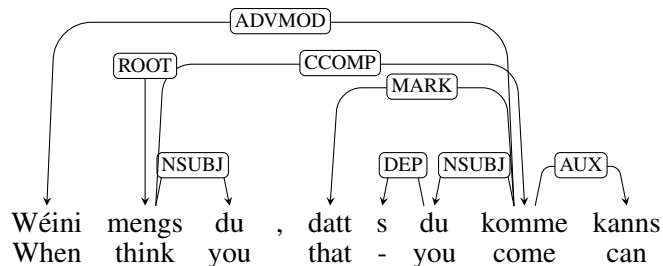


Figure 7: Agreement marker in sentence c14.

Negation The negation in Luxembourgish is typically expressed as a negation particle with *net*. In the first version of the UD tagset, the negation was a proper tag, but in the second version the tag is no longer available and is now tagged as *advmod*. We will use the feature *Polarity=NEG* for the negation particle, as is the custom in other UD treebanks.

Agreement Marker In subordinate clauses, where the subject is the second person singular (*du/de*), the complementiser is followed by the agreement marker *s*. The *s*-marker is mandatory in this sentence structure and has an orthographically isolated position between the initial element of the subordinate clause and the *du/de*-pronoun (Döhmer, 2020). It developed out of a reanalysis of the inflectional (verbal) *s*-suffix (2nd person singular) and became a clitic before the subject pronoun. Over time it grammaticalised into an obligatory *s*-marker with a fixed syntactic position. As there is no available tag to properly describe this phenomenon, we decided to use the *dep* tag and describe it in the miscellaneous column with *clitic*. In general, this is not a case of clitic doubling as in some West Germanic dialects because the subject pronoun itself is not always used as a clitic. Moreover, the *s*-clitic appears after any element in the complementiser position, not only subordinating conjunctions, but also after interrogative phrases or long prepositional phrases (Döhmer, 2020). Therefore, it should not be linked to the complementiser. Given the fact that it is syntactically bound and very predictable in terms of the sentence type in combination with a specific subject pronoun, attaching it to the verb with the *expl* relation (as per the UD guidelines) would not be justified. Although it doesn't behave like a regular clitic, the *clitic*-tag seems to be the most suitable, because of the strong dependence on the subject pronoun *du/de*. This phenomenon has different structural properties in the Continental West Ger-

manic varieties (it doesn't appear in other standard languages, though) and the terminology may vary in some descriptions (Renkowitz, to appear).

Figure 7 gives an example sentence from LuxBank where this phenomenon is annotated.

3.4 Planned Work

Extending the coverage of LuxBank is our primary objective, with the next batch of sentences currently being annotated. This batch comprises 50 randomly sampled sentences¹¹ from news articles from RTL, the main news broadcaster of Luxembourg. For further extensions, we plan to translate sentences from xSID (van der Goot et al., 2021) to support comparability across further NLP tasks in various languages. While working on this extension, we will also add the morphological features in the initial and future set of sentences.

4 Discussion

After applying the UD guidelines and analysing the Luxembourgish sentences, we now discuss practical and theoretical aspects related to the syntactic structure of the 20 CICLing sentences, including under-specified tags and potential challenges when incorporating different languages. Although the CICLing sentences are drawn from simple everyday language, the analysis of such sentences can be quite complex, e.g., when they contain elliptical constructions. Ellipses are a common phenomenon in many European languages, but it is difficult to determine syntactic dependencies, when different parts of the sentence have been elided. Among the 20 CICLing sentences, at least five contain some sort of elliptical structure. As a consequence, CICLing corpus might not be the best starting point for developing new treebanks, since some of the fundamental basic syntactical structures are not as well represented.

¹¹Sentences longer than 25 tokens were not considered.

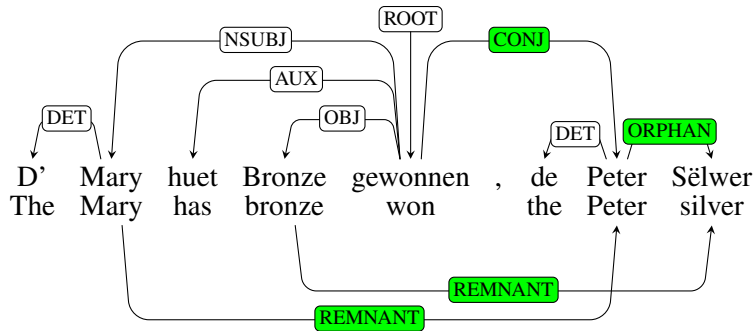


Figure 8: UD v2 versus v1 (below) annotation of ellipsis in sentence c9.

To better understand their structure, we analyse the sentences with elliptical structure following both the UD guidelines of version 1 and version 2, see the respective syntactical analysis in Figure 8. Although the version 2 UD guidelines are currently in use, where the dependency between the head of the elliptic sentence and the element depending on the omitted verb is marked as *orphan*, we find the version 1 to be more accurate from a linguistic point of view. In a verb phrase ellipsis, connecting the two *nsubj* under the tag *remnant* and leaving the other dependencies unvaried (i.e. as the verb phrase were there) would better reflect the underlying structure of these sentences.

A further discrepancy between linguistic theories and UD guidelines, as already mentioned in 3.3, concerns the *aux* tag. This tag is under-specified and used for two classes of functional verbs: auxiliaries and modal verbs. While the miscellaneous column can be helpful to deal with the limits of the UD guidelines in practice, it is still a makeshift solution that does not do full justice to phenomena not yet covered by the guidelines. As the feature column is still not enough to distinguish between different verb classes, a dedicated tag to allow better differentiation between auxiliary and modal verbs would be more precise from a linguistic point of view. Moreover, limiting the classification to a single copular verb further reduces the linguistic accuracy of the UD. The possibility to add more than one copular verb would then result in a more realistic representation of the class of copular verbs in Luxembourgish, without compromising the comparability with other languages.

Another aspect regarding the CICLing sentences concerns the modeling of gendered languages. As English usually does not mark the grammatical gender of common nouns, languages with marked gender then need to decide on the grammatical gender

of these nouns. Although this is not strictly related to the syntactic dependencies in the sentence, it could lead to a different interpretation and therefore an inaccurate translation of the original sentence. The following example from the CICLing sentences (c7) illustrates this:

- (EN) Peter’s **neighbour** painted the fence red.
- (DE) **Der Nachbar** von Peter hat den Zaun rot (an)gemalt.
- (LB) Dem Peter **säin Noper** huet den Zonk rout ugestrach.

As can be seen in the example sentences (marked in bold), even if the grammatical gender is unmarked in English, in both target languages the translators chose the male version of the word, arguably perpetuating the unaware gender bias of male and female roles in society (Bolukbasi et al., 2016). While we do not foresee cases like this in future additions to LuxBank, since we will be using original Luxembourgish material instead of translations, we feel it is important to point this out.

LuxBank is an ongoing project and the main goal is to add more annotated sentences to the treebank. Since this is the beginning of the project, we are continuously adapting the guidelines for Luxembourgish while annotating the data. More linguistic features for Luxembourgish will need to be specified in the future, as they weren’t covered in the initial 20 sentences, e.g., loanwords, verb cluster variation, and doubly filled complementisers.

Given the amount of language contact phenomena in Luxembourgish, especially loanwords from German, French, or English are a frequently occurring phenomenon that needs to be addressed. In the nominal domain, further guidelines must be created for French and English compounds, aside from using the *flat* tag, as they are sometimes written as one

word, as separate units, or hyphenated, depending on either the spelling norms of the source language or on Luxembourgish orthography.¹² French compounds often appear as multi-word units and are therefore close to syntactic expressions (Goethem and Amiot, 2019). Some of those expressions are directly borrowed into Luxembourgish, e.g. *Projet de loi* ‘bill (draft law)’ or *Carte d’identité* ‘identity card’. These expressions will need to be tagged according to French morphology and left-headedness. It should also be avoided that the French prepositions *de* and *d’* are automatically tagged as Luxembourgish definite articles.

Another common pattern in Luxembourgish syntax is verb cluster variation. The order of elements in 2-, 3-, and 4-verb clusters is variable, when modal verbs or subjunctive auxiliaries appear in subordinate clauses (Döhmer, 2020). In general, word order variation will not affect the deep structure of the sentence, i.e., the dependencies remain the same, but the surface structure will be different. Concerning the left periphery of subordinate clauses, the initial element of the subordinate clause is sometimes extended by a second complementiser, namely *dass/datt* (Döhmer, 2020). Sentences with a doubly filled complementiser, such as *obwuel dass et reent* ‘(lit.) although that it rains’, could cause difficulties in the annotation process because in most cases the complementiser position can only contain a single constituent. All of these phenomena (among others) have to be addressed in the future to develop appropriate guidelines for Luxembourgish.

5 Conclusion

In this paper, we introduce LuxBank as the first treebank for Luxembourgish. As the discussion of structural characteristics and challenges encountered when developing annotation guidelines for Luxembourgish show, building a new treebank for a small language represents a theoretical as well as practical challenge. This is particularly true in view of the structural variation in Luxembourgish and its ongoing standardisation. In this context, the decision to bring together a mixed team of linguistic and computational experts has proven crucial to the successful implementation of UD for Luxembourgish.

LuxBank will facilitate a more in-depth understanding of Luxembourgish as a ‘low-research’ lan-

guage, making it an invaluable resource not only for linguists but also for language teaching. This treebank project can serve as an aid for spell-checking tools as well as for future grammar checking applications. A tailor-made tagging system derived from earlier versions of LuxBank could ensure higher accuracy and consistency in Luxembourgish text processing and modelling, to help to better organise existing text archives, and to extend the treebank further. In the future, LuxBank will enable easier quantitative exploration of linguistic data, providing insights that were previously more difficult to obtain.

From a typological perspective, it is important to complete the data in the UD treebanks for West Germanic varieties. So far, mainly large standard languages have been incorporated, whereas regional varieties and/or smaller languages are underrepresented. LuxBank adds the first Middle German language description to the UD. This can help to explore syntactic variation and to understand the structural aspects of these languages.

LuxBank will also be beneficial for NLP research and text processing in general. Presently, the support for Luxembourgish is limited to certain tasks (lemmatisation, POS), and the available resources do not use the UD tagset for POS tagging. Building a dedicated treebank for Luxembourgish will make it possible to extend the support for the language in industry-standard tools like *spaCy* to the grammatical level and to offer a comparable tag set for the analysis of syntactic structures. In doing so, LuxBank is laying the foundation for a better representation of Luxembourgish in NLP, both for further research and for the development of customized tools and pipelines.

Luxembourgish can also serve as a model case for describing other small languages and varieties, as these often possess unique characteristics – and resulting challenges – like those discussed in this paper: a limited amount of available resources, a small number of trained linguistic experts, a high amount of linguistic variation (be it lexical, grammatical, or orthographic), a structural influence from other (standard) languages, and a complex multilingual language situation. With this contribution, we aim to position Luxembourgish as a valuable resource for comparable language situations. We also hope to highlight the importance of foundational research for small and non-standardised languages to preserve linguistic diversity in the digital age and make it more visible in NLP.

¹²D’Lëtzebuerg Orthografie, ZLS 2022.

Limitations

The work presented in this paper is still in progress, and subsequent modifications may be made as the project evolves. It is important to note that finding and recruiting domain experts for data annotation is challenging. Additionally, the amount of variation within the language sometimes makes it difficult to reach a consensus on the classification of phenomena, which has introduced additional complexity to our research.

Ethics Statement

All data used in this project is freely available and obtained from publicly accessible sources. The human annotators involved in this project were fully compensated for their contributions, as this work forms part of their regular employment responsibilities. Additionally, all data is appropriately licensed for the intended use in this research, ensuring compliance with legal and ethical standards. This adherence to ethical guidelines ensures the integrity and responsible conduct of our research.

Acknowledgements

This research was supported by the Luxembourg National Research Fund (Project code: C22/SC/117225699).

References

- Martine Adda-Decker, Thomas Pellegrini, Eric Bilinski, and Gilles Adda. 2008. [Developments of “Lëtzebuergesch” Resources for Automatic Speech Processing and Linguistic Studies](#). In *Proceedings of LREC’08*, Marrakech, Morocco. ELRA.
- Noëmi Aepli. 2018. *Parsing approaches for swiss german*. Master’s thesis, University of Zurich.
- Dimitra Anastasiou. 2022. [ENRICH4ALL: A First Luxembourgish BERT Model for a Multilingual Chatbot](#). In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 207–212, Marseille, France. ELRA.
- Sjef Barbiers and Annemarie Van Dooren. 2017. [Modal Auxiliaries](#). In Martin Everaert and Henk C. Van Riemsdijk, editors, *The Wiley Blackwell Companion to Syntax*, 2nd ed. edition. Wiley, Hoboken, NJ, USA.
- Roberto Basili, Malvina Nissim, and Giorgio Satta. 2017. *Toward a treebank collecting german aesthetic writings of the late 18th century*. In *Proceedings of CLiC-it*, volume 11, page 12.
- Laura Bernardy. 2022. *A Luxembourgish GPT-2 Approach Based on Transfer Learning*. Master’s thesis, University of Trier.
- Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank. 2024. [MaiBaam: A Multi-Dialectal Bavarian Universal Dependency Treebank](#).
- Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. [Man is to computer programmer as woman is to homemaker? debiasing word embeddings](#). In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.
- Caroline Döhmer. 2020. *Aspekte der luxemburgischen Syntax*. Current Trends in Luxembourg Studies. Melusina Press.
- Nathalie Entringer. 2022. *Vun iwwerfëlltene Bussen bis bei déi beschte Witzer. Morphologische Variation im Luxemburgischen – eine variations- und perzeptionslinguistische Studie*. Ph.D. thesis, University of Luxembourg.
- Nathalie Entringer, Peter Gilles, Sara Martin, and Christoph Purschke. 2021. [Schnëssen. surveying language dynamics in luxembourgish with a mobile research app](#). *Linguistics Vanguard*, 7:20190031.
- Daniela Gierschek. 2022. *Detection of Sentiment in Luxembourgish User Comments*. Ph.D. thesis, University of Luxembourg.
- Peter Gilles. 2019. [39. Komplexe Überdachung II: Luxemburg. Die Genese Einer Neuen Nationalsprache](#). In Joachim Herrgen and Jürgen Erich Schmidt, editors, *Sprache und Raum - Ein internationales Handbuch der Sprachvariation. Volume 4 Deutsch*, pages 1039–1060. De Gruyter Mouton, Berlin, Boston.
- Peter Gilles. 2023. [Luxembourgish](#). In Sebastian Kürschner and Antje Dammel, editors, *Oxford Encyclopedia of Germanic Linguistics*. Oxford University Press, Oxford.
- Peter Gilles, Léopold Edem Ayité Hillah, and Nina Hosseini Kivanani. 2023a. [Asrlux: Automatic speech recognition for the low-resource language luxembourgish](#). In *Proceedings of the 20th International Congress of Phonetic Sciences*. Guarant International.
- Peter Gilles, Nina Hosseini Kivanani, and Léopold Edem Ayité Hillah. 2023b. [Lux-asr: Building an asr system for the luxembourgish language](#). In *Proceedings of SLT*.
- Elvira Glaser. 2006. [Zur Syntax des Lëtzebuergesch: Skizze und Forschungsprogramm](#). In Claudine Moulin, editor, *Perspektiven einer linguistischen Luxemburgistik. Studien zu Diachronie und Synchronie*, number 25 in Germanistische Bibliothek, pages 227–246. Winter.

- Kristel Van Goethem and Dany Amiot. 2019. [Compounds and multi-word expressions in french](#). In Barbara Schlücker, editor, *Complex Lexical Units*, pages 127–152. De Gruyter, Berlin, Boston.
- Barbara Hans-Bianchi and Peggy Katelhoen. 2011. Kann man tun und lassen, was man will? Verben zwischen Lexik und Grammatik. *Estudios Filológicos Alemanes*, 201:75–88.
- Thomas Lavergne, Gilles Adda, Martine Adda-Decker, and Lori Lamel. 2014. [Automatic language identity tagging on word and sentence-level in multilingual text sources: a case-study on Luxembourgish](#). In *Proceedings of LREC*, pages 3300–3304, Reykjavik, Iceland. ELRA.
- Cedric Lothritz, Saad Ezzini, Christoph Purschke, Tegawendé François D Assise Bissyande, Jacques Klein, Isabella Olariu, Andrey Boytsov, Clement Lefebvre, and Anne Goujon. 2023. Comparing Pre-Training Schemes for Luxembourgish BERT Models. In *Proceedings of KONVENS*.
- Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2022. [LuxemBERT: Simple and Practical Data Augmentation in Language Model Pre-Training for Luxembourgish](#). In *Proceedings of LREC*, pages 5080–5089, Marseille, France. ELRA.
- Sara Martin. 2019. [Hatt or si? Neuter and feminine gender assignment in reference to female persons in Luxembourgish](#). *STUF - Language Typology and Universals*, 72(4):573–601.
- 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 ACL*, pages 92–97.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. [Universal Dependencies v1: A multilingual treebank collection](#). In *Proceedings of LREC'16*, pages 1659–1666, Portorož, Slovenia. ELRA.
- Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.
- Damaris Nübling. 2006. Auf Umwegen zum Passivauxiliar - Die Grammatikalisierungspfade von GEBEN, WERDEN, KOMMEN und BLEIBEN im Luxemburgischen, Deutschen und Schwedischen. In Claudine Moulin and Damaris Nübling, editors, *Perspektiven einer linguistischen Luxemburgistik. Studien zu Diachronie und Synchronie*, number 25 in Germanistische Bibliothek, pages 171–201. Winter, Heidelberg.
- Fred Philippy, Shohreh Haddadan, and Siwen Guo. 2024. [Forget NLI, use a dictionary: Zero-shot topic classification for low-resource languages with application to Luxembourgish](#). In *Proceedings of LREC-COLING*. ELRA and ICCL.
- Christoph Purschke. 2020. Attitudes Toward Multilingualism in Luxembourg. A Comparative Analysis of Online News Comments and Crowdsourced Questionnaire Data. *Frontiers in AI*, 3:536086.
- Tharindu Ranasinghe, Alistair Plum, Christoph Purschke, and Marcos Zampieri. 2023. Publish or hold? Automatic comment moderation in Luxembourgish news articles. In *Proceedings of RANLP*.
- Julia Renkwitz. to appear. The agreement of subclause initial elements in Continental West Germanic: Realizations and explanations. In *Syntax aus Saarbrücker Sicht 6*, ZDL Beihefte. Steiner.
- François Schanen. 1980. *Recherche sur la syntaxe du luxembourgeois de Schengen: l'énoncé verbal*. Thèse pour le Doctorat d'État, Paris IV, Paris.
- François Schanen and Jacqui Zimmer. 2012. *Lëtzebuergesch Grammaire*. Éditions Schortgen.
- Janine Siewert, Yves Scherrer, and Jörg Tiedemann. 2021. Towards a balanced annotated low saxon dataset for diachronic investigation of dialectal variation. In *Conference on Natural Language Processing*, pages 242–246. KONVENS 2021 Organizers.
- Joshgun Sirajzade, Daniela Gierschek, and Christoph Schommer. 2020. An Annotation Framework for Luxembourgish Sentiment Analysis. In *Proceedings of SLTU-CCURL 2020*, page 172—176, Marseille. Language Resources and Evaluation Conference (LREC 2020).
- Natalie D. Snoeren, Martine Adda-Decker, and Gilles Adda. 2010. [The study of writing variants in an under-resourced language: Some evidence from mobile n-deletion in Luxembourgish](#). In *Proceedings of LREC*, Valletta, Malta. ELRA.
- Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. [From masked language modeling to translation: Non-english auxiliary tasks improve zero-shot spoken language understanding](#). *Preprint*, arXiv:2105.07316.
- Emanuel Borges Völker, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. HDT-UD: A very large Universal Dependencies treebank for German. In *Proceedings of udw, syntaxfest 2019*, pages 46–57.
- Daniel Zeman, Jan Hajic, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav 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.

Building a Universal Dependencies Treebank for Georgian

Irina Lobzhanidze, Erekle Magradze, Svetlana Berikashvili, Anzor Gozalishvili,
Tamar Jalaghonia

Ilia State University

irina_lobzhanidze@iliauni.edu.ge, erekle.magradze@iliauni.edu.ge,
svetlana.berikashvili@iliauni.edu.ge,
anz2.gozalishvili@gmail.com, jalaghonia98@gmail.com

Abstract

This paper presents the design and development of the Georgian Syntactic Treebank within the Universal Dependencies (UD) framework, addressing the unique morphosyntactic challenges of Georgian, a Kartvelian language. We describe the methodology for selecting and annotating 3,013 sentences from Wiki, mapping existing tagsets to the UD scheme, and converting data into the CoNLL-U format. The paper also details the training of a UDPipe model using this preliminary treebank.

1 Introduction

The development of syntactic treebanks is essential for advancing natural language processing (NLP) across diverse languages, enabling computational models to better understand and process linguistic structures. The Universal Dependencies (UD) (Nivre et al., 2017) framework provides a standardized approach to syntactic annotation that facilitates cross-linguistic consistency and data sharing. The data freely available on GitHub is generally used for training various models like UDPipe (Straka, 2016), UDify (Kondratyuk et al., 2019), Stanza (Qi et al., 2020) and others.

However, many languages, particularly those with complex morphosyntactic characteristics, remain underrepresented in these resources. Georgian, a Kartvelian language, is one such language that presents challenges due to its split-ergative structure, free word order, and rich inflectional morphology. This paper addresses the compilation of a Georgian Syntactic Treebank consisting of 151 utterances (2123 tokens) from the Georgian Language Corpus (GLC) and 3013 utterances (54116 tokens) from Wiki; totaling 3164

utterances (56239 tokens). This work contributes to the development of computational tools for under-resourced languages.

The paper consists of five sections. The first section provides a brief review of previous work concerning the Georgian language. The second section offers a detailed description of the data selection, annotation process, tagset mapping, and conversion to the CoNLL-U format. The third section includes information on the training of the UDPipe model, and the fourth section presents the results and their analysis. The fifth section summarizes the findings.

2 Background on Georgian Language Treebank

The development of treebanks for Kartvelian languages, a family characterized by its unique morphosyntactic structure and phonological properties, can be considered as new within the field of natural language processing (NLP). From this perspective the syntactic Treebank of the Laz language, another Kartvelian language, can be considered as the first attempt to create the Universal Dependencies Treebank and to make it available online (Turk et al. 2020). Common features shared by Georgian and other Kartvelian languages include the following:

- A relatively uniform sound system;
- A well-developed system of word inflection and derivation;
- Agglutinating and inflecting systems that make use not only of a large variety of grammatical affixes, but also of ablaut and other types of processes typical of internal stem inflection;

- The split-ergativity (Boeder 1979; Harris 1981, 1985; Hewitt 1983, 1987; Tuite 2017; Baker and Bobaljik 2017; Berikashvili 2024 and others).

All of these features pose unique difficulties at all levels of language processing and present interesting challenges for the compilation of robust language processing systems.

Prior to the efforts documented in this paper, Georgian had been largely underrepresented in major syntactic annotation initiatives such as the UD framework. While, various research groups (Datukishvili, 1997; Gurevich, 2006; Kapanadze, 2009 and others) have developed some tools for the processing of Modern Georgian morphology or for the creating of corpora (Gippert et al., 2011; Doborjginidze et al., 2012), the problem of syntax remained unsolved. Early attempts to create syntactic resources for Georgian included efforts to develop the ParGram Treebank within the Lexical Functional Grammar (LFG) framework (Sulger et al., 2013) and the GRUG treebank combining constituency-based and dependency-based structures (Kapanadze, 2017). But tagsets (Erjavec, 2004; Meurer, 2007 and others) and annotation schemes were not fully compatible with the UD framework, preventing their integration and wider use. Thus, it was important to adapt the existing tagsets and the mapping of Georgian linguistic features to the UD framework, to ensure that the syntactic annotation of Georgian could align with the UD, allowing the possibility of comparative linguistic studies.

As a result, the initial test version was limited in coverage and consisted of 151 utterances, that did not fully capture the linguistic characteristics of Georgian. Additionally, the tools available for syntactic parsing, such as the UDPipe model, had not been trained on sufficient Georgian data.

3 Methodology and Annotation Process

The development of the Georgian Syntactic Treebank followed a systematic approach to address the specific challenges posed by the Georgian language’s complex morphosyntactic structure. The methodology encompassed several key strategies: determining syntactic functions and compiling annotation guidelines, improving the

initial annotation scheme developed for the initial 151 utterances from the GLC by revising and standardizing the use of dependency relations, selecting and annotating data, and contributing to the UD GitHub repository. Additionally, the training of the UDPipe model using the annotated data is described in detail.

3.1 Data Selection

The Georgian Language Corpus (GLC) (Doborjginidze et al. 2012) served as the initial source for the treebank, offering a collection of texts of different genres and periods (6th-21st centuries). From this corpus, a total of 151 sentences reflecting Modern Georgian were selected. The selection criteria focused on ensuring a representative sample of Georgian syntax, including various sentence lengths, structures, and complexity levels. But these data were not enough to train the model and to complement the data from the GLC and introduce a more diverse linguistic style were also selected from Georgian Wikipedia. As a result, 3,013 sentences were selected from Wikipedia, covering 131 different scientific domains. The selection process prioritized sentences that demonstrate a variety of syntactic constructions, including simple, coordinated and subordinated complex clauses, as well as those that feature unique or less common linguistic phenomena. All these sentences were checked to include different morphosyntactic features.

3.2 Data annotation

The annotation process was preceded by the compilation of annotation guidelines and the development of the UD annotation scheme for Georgian. These guidelines were made available in the language-specific documentation section of the UD GitHub repository¹. The development of the scheme for Georgian involved adapting the tags used in the Georgian morphological analyzer (Lobzhanidze 2022) to ensure compatibility with UD standards. After the mapping of tagsets, a special Python code was written to convert the analyzer’s output into the CoNLL-U format and to provide additional tokenization. It was especially important to provide segmentation of multi-word tokens, which were not covered by the analyzer’s

¹

<https://universaldependencies.org/ka/index.html>

output and to fill information on lemmas, part-of-speech (POS) tags, and morphosyntactic features. The main differences between the analyzer’s output and the UD scheme like tokenization as well as different linguistic phenomena connected to split-ergativity and other features of Georgian can be summarized as follows: a) the main core dependency arguments, which are used in Georgian are nominal subject, direct and indirect objects. While in Indo-European languages, the verb generally agrees with the subject of the sentence, in Georgian the verb agrees not only with the subject, but with its objects (direct and indirect) as well. However, as a result of the strong Person Case Constraint (PCC) effect, the direct object is always the third person in ditransitive constructions, and the third person agreement is always null. Therefore, there are no cases where all three arguments agree simultaneously. As a result, Georgian verbs have core and peripheral arguments. A core argument agrees morphologically with the verb by means of person and number markers, while a peripheral argument does not. In Georgian, a nominal subject is a nominal that serves as the subject of the verbal predicate in ergative or nominative or dative cases; a direct object is a nominal or noun phrase that serves as the object of the verbal predicate in nominative or dative; the indirect object of a verb is a dative-marked complement. The Georgian treebank uses all the main non-core dependent’s tags except of `expl` and `dislocated`. All nominal dependent tags are used except of classifier (`clf`). As a result each sentence was annotated to capture syntactic dependencies, including subject, object, and modifier relationships. Taking into consideration the complexity of Georgian syntax - characterized by split-ergativity and free word order - special attention was paid to accurately representing the syntactic roles of words within sentences and to the case-marking of subject, direct and indirect objects.

3.3 UDPipe Model Training

To evaluate the quality of the annotations and provide a baseline for further development, a UDPipe model was trained using the annotated data. The training set consisted `ka_glc-ud-dev.conllu` (470 utterances), `ka_glc-ud-test.conllu`

(481 utterances) and `ka_glc-ud-train.conllu` (2213 utterances). The UDPipe model was trained on the Georgian data using the default parameters. Performance metrics, including tokenization accuracy, POS tagging accuracy, and parsing accuracy (both Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS)), were calculated to assess the model’s effectiveness.

3.4 Validation and Corrections

Following the automatic annotation and model training, a manual validation process was implemented. This involved reviewing a sample of the annotated sentences to identify and correct errors in tokenization, POS tagging, and syntactic annotation. Corrections were made directly in the CoNLL-U files, and the model was retrained as necessary to incorporate these improvements.

3.5 Contribution to the UD repository

The validated treebank files, including `ka_glc-ud-test.conllu` and `ka_glc-ud-train.conllu`, were uploaded to the repository, along with related documentation files such as `README.md`. The treebank passed the UD validation process². At this moment the treebank is available in the dev branch of the repository and will be unified with the master branch after the twenty-first release of annotated treebanks in Universal Dependencies, v2.15, to be implemented in November.

4 Model training

UDPipe (Straka et al. 2016) is a trainable pipeline for tokenization, tagging, lemmatization, and dependency parsing of CoNLL-U files. UDPipe is language-agnostic and can be trained given annotated data in CoNLL-U format. For the Georgian language case, we have used Version 1.3.1-dev. Data training has been implemented on 3164 utterances (sentences) consisting of 56239 tokens. We trained UDPipe models (tokenizer, tagger, parser) using training set. The method used for training was "morphodita_parsito" which is the only supported method in `udpipe` version 1.3. We used default parameters for each model in a pipeline. The training results are as follows:

- Tokenizer: Epoch 44, logprob: -1.6215e+03, training acc: 99.87%, heldout tokens:

[ator/cgi-bin/unidep/validation-report.pl?UD_Georgian-GLC](http://quest.ms.mff.cuni.cz/udvalidator/cgi-bin/unidep/validation-report.pl?UD_Georgian-GLC)

²

<https://quest.ms.mff.cuni.cz/udvalid>

99.83%P/99.84%R/99.84%, sentences:
98.08%P/97.87%R/97.97%;

- Tagger: Iteration 20: done, accuracy 99.85%, heldout accuracy 89.49%/91.80%/85.38%;
- Parser: Iteration 8: training logprob - 2.0778e+04, heldout UAS 79.04%, LAS 74.75%

While the testing for accuracy on ka_glc-ud-test.conllu gives the following results:

- Tokenizer: Number of SpaceAfter=No features in gold data: 1523; Tokenizer tokens - system: 9288, gold: 9283, precision: 99.69%, recall: 99.74%, f1: 99.71%; Tokenizer multiword tokens - system: 742, gold: 751, precision: 97.71%, recall: 96.54%, f1: 97.12%; Tokenizer words - system: 10035, gold: 10039, precision: 99.15%, recall: 99.11%, f1: 99.13%; Tokenizer sentences - system: 497, gold: 481, precision: 92.35%, recall: 95.43%, f1: 93.87%
- Tagger: Tagging from gold tokenization - forms: 10039, upostag: 93.34%, xpostag: 93.34%, feats: 85.42%, alltags: 85.18%, lemmas: 89.89%
- Parser: Parsing from gold tokenization with gold tags - forms: 10039, UAS: 80.34%, LAS: 76.01%

Comparing the results some frequent misinterpretations were noted concerning the complex subordinate clauses. The gold standard files included more complex structures, while the parser tried to simplify them. For example, the parser sometimes had difficulties distinguishing the subject and object of sentences marked with `Case=Nom` or `Case=Dat`, which can be explained by the split-ergativity of Georgian. Additionally, it assigned the modifier relation differently depending on sentence context or positional emphasis, and showed discrepancies in the representation of clitics like postpositions and particles.

5 Results and discussion

The primary outcome of this project is the creation of an initial Georgian Syntactic Treebank, consisting of 3164 sentences (56239 tokens). This

treebank was developed by mapping existing Georgian linguistic resources to the UD framework, ensuring compatibility with cross-linguistic standards. The treebank was validated and made available for use within the UD community, representing a significant milestone for the computational processing of the Georgian language. The main components of the treebank are the following:

- Universal POS Tags (UPOS): The mapping of Georgian part-of-speech tags to the UD's UPOS tags ensured the cross-linguistic consistency of the treebank. The main difference revealed is as follows: NOUN and PROPN. as opposed to +Noun+Com and +Noun+Prop;
- Morphological Features (FEATS): The detailed morphological features in the FEATS column allowed the representation of Georgian's morphosyntactic properties. We have added `AdpType`, `AdvType`, `PartType`, `NameType`, `VerbType`, `Subcat`, `PunctType` to `Lexical Features`; `NumForm` to `Inflectional Features` and `Person[subj]`, `Person[obj]`, `Person[io]`, `Number[subj]`, `Number[obj]`, `Number[io]` to `Verbal Features`;
- Syntactic Dependencies (DEPREL): The syntactic annotation, including the identification of heads and dependency relations, provided a structured representation of Georgian syntax. We have used all tags except `expl`, `dislocated`, `clf`, and `reparandum`.

At the same time, the implementation of the project revealed some areas for further improvement:

- Mapping and Compatibility: The mapping of Georgian morphosyntactic tagset to the UD revealed that some features and categories were not directly compatible with existing UD tags. For instance, some of the tags indicating voice and connected to the category of diathesis are not compatible with the UD framework (e.g. `autoactive`, `inactive` (inverse active));

- **Annotation Accuracy:** The treebank was validated through a series of automated and manual checks by two annotators, ensuring the accuracy of the syntactic annotations. The reliance on existing tools like the morphological analyzer and the UD validator can be considered as effective, but the manual correction highlighted the importance to add some additional syntactic dependencies like `flat:foreign`, `flat:name` etc.;
- **Challenges in Complex Structures:** The analysis identified particular difficulties in accurately annotating sentences with complex syntactic structures, such as those involving multiple clauses, valency-changing operations, and free word order. The Georgian verb reflects relations between two or three arguments and provides a mapping between morphology and syntactic features such as the roles of participants. Especially, impersonal verbs do not have a subject at all, intransitive verbs take a subject only, indirect transitive verbs take two arguments: a subject and an indirect object; transitive verbs take two arguments: a subject and a direct object and, ditransitive verbs take three arguments: a subject and a direct and indirect object. As a result, the subject can be marked by the nominative, ergative or dative cases, while the objects are marked by the nominative or dative case with or without a postposition. All these affected the correct marking of arguments at the level of syntactic dependencies.

6 Conclusions

By this study we tried to represent an advancement in the development of linguistic resources for the Georgian language, particularly through the creation of a syntactic treebank within the Universal Dependencies (UD) framework. The implementation of this project has provided a resource for the computational processing of Georgian, addressing main challenges related to the complex morphosyntactic structure and contributing to the broader field of natural language processing (NLP) for under-resourced languages. Expanding the treebank with the complete GLC data, updating the UDPipe model can be considered as important future steps to improve the accuracy of Georgian NLP tools.

Acknowledgments

This study was funded by the Shota Rustaveli National Science Foundation (No FR-22-20496) and the JESH (Joint Excellence in Science and Humanities) grant, financed by the Austrian Academy of Sciences. The authors would like to thank the anonymous reviewers for their helpful comments and feedback.

References

- Mark Baker and Jonathan Bobaljik. 2017. On inherent and dependent theories of ergative case. In *J. Coon, D. Massam, & L. Travis, The Oxford Handbook of Ergativity*. Oxford: Oxford University Press, pages 111–134. <https://doi.org/10.1093/oxfordhb/9780198739371.013.5>.
- Svetlana Berikashvili. 2024. *Differential Subject Marking in Georgian*. Göttingen: Georg-August-Universität Göttingen.
- Winfried Boeder. 1979. Ergative syntax and morphology in language change: the South Caucasian languages. In *F. Plank, Ergativity: towards a theory of grammatical relations*. Orlando: Academic Press, pages 435–480.
- Ketevan Datukishvili. 1997. Some questions of computer synthesis of verb in Georgian. In *Proceedings of the Second Tbilisi Symposium on Language, Logic and Computation*. Tbilisi: t'bilisi saxelmcip'o universiteti (Tbilisi State University), pages 83-85.
- Nino Doborjginidze, Irina Lobzhanidze, and Irakli Gunia. 2012. *Georgian Language Corpus (GLC)*. <http://corpora.iliauni.edu.ge/> (Accessed: 16 August 2024).
- Tomaž Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal: European Language Resources Association (ELRA).
- Jost Gippert, Paul Meurer, Manana Tandashvili. 2011. *The Georgian National Corpus (GNC)*. <http://gnc.gov.ge/> (Accessed: 16 August 2024).
- Olga Gurevich. 2006. A Finite-State Model of Georgian Verbal Morphology. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*. New York: Association for Computational Linguistics, Pages 45-48.

- Alice Harris. 1981. *Georgian Syntax: a study in Relational Grammar*. Cambridge: Cambridge University Press.
- Alice Harris. 1985. *Diachronic syntax: the Kartvelian Case*. New York: Academic Press.
- Oleg Kapanadze. 2009. "Describing Georgian Morphology with a Finite-State System." In *Proceedings of the 8th international conference on Finite-State Methods and Natural Language Processing*. Pretoria: Springer. 114-122.
- Oleg Kapanadze. 2017. *Multilingual GRUG Parallel TreeBank — Ideas and Methods*. Saarbrücken: LAMBERT Academic Publisher.
- Dan Kondratyuk, Milan Straka. 2019. 75 Languages, 1 Model: Parsing Universal Dependencies Universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong: China, pages 2779–2795.
- Irina Lobzhanidze. 2022. *Finite-State Computational Morphology: An Analyzer and Generator for Georgian*. Cham: Springer.
- Paul Meurer. 2007. A Computational Grammar for Georgian. In *Lecture Notes in Computer Science*. Berlin: Springer, pages 1-15.
- Joakim Nivre, Daniel Zeman, Filip Ginter, Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.
- 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*. Online: Association for Computational Linguistics. 101–108.
- Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pages 4290–4297.
- Sebastian Sulger, Miriam Butt, Tracy King, Paul Meurer, Tibor Laczko, Gyorgy Rákosi, Cheikh Dione, Helge Dyvik, Victoria Rosen, Koenraad De Smedt, Agnieszka Patejuk, Ozlem Çetinoğlu, I Wayan Arka, Meladel Mistica. 2013. ParGramBank: The ParGram Parallel Treebank. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Sofia, pages 550–560.
- Kevin Tuite. 2017. Alignment and orientation in Kartvelian. In *J. Coon, D. Massam, & L. Travis, The Oxford Handbook of Ergativity*. Oxford: Oxford University Press, pages 1114–1138.
- Utku Turk, Kaan Bayar, Aysegul Dilara Ozercan, Gorkem Yigit Ozturk, and Saziye Betul Ozates. 2020. First Steps towards Universal Dependencies for Laz. *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*. Barcelona, Spain (Online): Association for Computational Linguistics, pages 189–194.

Introducing Shallow Syntactic Information within the Graph-based Dependency Parsing

Nikolay Paev, Kiril Simov, Petya Osenova
Artificial Intelligence and Language Technology
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
Bulgaria

nikolay.paev@iict.bas.bg, kivs@bultreebank.org, petya@bultreebank.org

Abstract

The paper presents a new BERT model, fine-tuned for parsing of Bulgarian texts. This model is extended with a new neural network layer in order to incorporate shallow syntactic information during the training phase. The results show statistically significant improvement over the baseline. Thus, the addition of syntactic knowledge - even partial - makes the model better. Also, some error analysis has been conducted on the results from the parsers. Although the architecture has been designed and tested for Bulgarian, it is also scalable for other languages. This scalability was shown here with some experiments and evaluation on an English treebank with a comparable size.

1 Introduction

In this paper we present a transformer-based architecture for dependency parsing which is extended to accommodate some predefined shallow dependency information. The predefined information came from two sources: lexicons and shallow grammars. The Dependency information — dependency relations (arcs and labels) — are represented within the lexicon at least in two varieties: (1) representation of valency frames, and (2) representation of multiword expressions (MWEs). For a recent overview see (Giouli and Barbu Mititelu, 2024). In our in-house lexicons we use partial dependency trees in order to represent the obligatory grammar information such as the object and clitic relations of the verbal head and the modification relations of the nominal head. For example, the MWEs “kick the bucket” is expected to have in the lexicon two dependency relations — from the article “the” to the head noun “‘bucket’ the relation is “det” and from “bucket” to the head verb “kick” the relation is “obj”.

Our goal was set to implement a parser that is able to incorporate preliminary dependency relations among words — even partial — from the

lexicon, thus before parsing of the whole sentence this step has been performed. Similarly, shallow grammars provide sets of rules over the word forms and their grammatical annotation. These grammars are known to produce partial but reliable analyses mostly achieving 100 % accuracy. In the experiments reported in this paper both sources have been explored. The experiments and evaluation were performed for Bulgarian and English.

The structure of the paper is as follows: the next section provides a focused overview of related work; section 3 describes the Dependency parsing architecture that was implemented in our work. This section also elaborates on the modification of the initial architecture towards the incorporation of some sure information from lexicons and shallow grammars. In section 4 the experimental settings are described in detail. Here also the results are presented and discussed. In section 5 the manual evaluation of the results is outlined. Section 6 concludes the paper and presents some future directions of research.

2 Related Work

Zhou et al. (2023) show that prepositional phrase attachment poses the biggest challenge to understanding syntax by LLMs. The case study on training the dynamics of the LLMs revealed that the majority of syntactic knowledge is learned during the initial stages of training. For these reasons, we started with the injection of partial but sure linguistic information into the model. Shen et al. (2021) propose a new syntax-aware language model — Syntactic Ordered Memory (SOM). The model explicitly models the structure with an incremental parser and maintains the conditional probability setting a standard language model (left-to-right). The related experiments show that SOM can achieve strong results in language modeling, incremental parsing and syntactic generalization

tests, while using fewer parameters than other models. The model uses constituency trees for English and these trees are embedded in a grid-like memory representation. The authors report improvement on phenomena like gross syntactic states and long-distance dependencies. In our case instead of incremental approach we use predefined partial syntactic information. Yoshida et al. (2024) propose a novel method called tree-planting. This means to implicitly “plant” trees into attention weights of Transformer LMs to reflect syntactic structures of natural language. Transformer LMs trained with tree-planting are called Tree-Planted Transformers (TPT). They learn syntax on small treebanks via tree-planting and then scale on large text corpora via continuous learning with syntactic scaffolding. Our approach is similar since it uses dependency subtrees but it relies only on chunks and MWEs as ‘islands of certainty’. Another difference is that we add syntactic information within the transformer network during the fine-tuning phase, but in future we plan to pre-train a model on partially annotated corpora.

The combination of information from different sources in order to improve the overall performance of the parser is not a new idea. This is especially true for the combination of various machine learning techniques with sure symbolic knowledge under the motto “*why to guess if we already know?*”. With respect to dependency parsing Özates et al. (2020) use special rules to introduce dependency relations between certain word forms in the sentences. Each rule identifies some arcs within the dependency tree. The rules are applied recursively up to the moment when no more applications are possible. The result from the application of the rules is encoded as additional token embeddings which are concatenated with embeddings used by the actual neural network parser. The parser used in their experiments is an LSTM-based dependency parser — Stanford’s Graph-based Neural Dependency Parser (Dozat et al., 2017). The baseline is the parser trained without these extended embeddings, and thus later trained with them. The paper reports on improving UAS (about 2 %) and LAS (near 3 %). Our approach differs from theirs in several ways: (1) We fine-tune a BERT¹ language model as a dependency parsing model. The fine-tuning step requires the existence of a depen-

dependency treebank². In addition to the treebanks we relied on the supplement of “suggested” arcs which would facilitate arcs prediction and labeling (see below). These arcs we considered to be the linguistic knowledge added to the respective treebank. (2) the linguistic knowledge added during the training is not necessary to be the same as during the inference time. In this way the approach could be used when there are no reliable sources of such linguistic knowledge. The existence of a treebank, of course, is obligatory. (3) The additions of dependency relations in parallel to the treebank look like redundant information, but it plays an important role during the parsing of new texts.

In the next section we present the specifics of the dependency parsing model that has been implemented for the experiments reported in this paper.

3 Graph-based Dependency Parsing

In the implementation of our dependency parsing based on LLMs we follow the approach of McDonald et al. (2006) about a graph-based dependency parsing performed in two steps: (1) determination of dependency arcs in the syntactic tree — the immediate domination relation over the tokens in the sentence — for each token to find its immediate parent token (adding special token for the root of the sentence); and (2) labeling the selected arcs with the appropriate dependency relations. This approach was adopted by many of the recent dependency parsers ((Dozat and Manning, 2017), for instance) — where a transformer-based model is used for determining the context-aware token embeddings, and an additional model for the selection of the arcs (Head selection model) as well as for the labels.

In our implementation both - the transformer model and the head selection model - are directly connected - the head selection model is integrated as an additional layer over the last layer of the transformer model. The head selection model is similar to any other token classification model, except that the number of classes is dynamic — the number of possible heads in the sentence varies. When sub-word tokenization is performed, only the first token of each word is used, while the others are ignored during training and inference phases. In the next sections the implementation of the parser is presented in more details.

²In our experiments the available Bulgarian and one of the English Universal Dependency Treebanks — <https://universaldependencies.org/>

¹BERT model is introduced by Devlin et al. (2018).

3.1 Head Classification — (UAS)

As it was mentioned above, the first step of the parser is to identify the arcs. This is done by selecting the head of each word form in the sentence. The head could be any of the other word forms in the sentence, or a specially included token for the root of the sentence.

Since the number of the possible heads in a sentence is dynamic — (it depends on the number of tokens within the sentence) — a simple linear (affine) transformation is not applicable. Instead, a self-attention mechanism is used due to its ability to aggregate information from sequences with different lengths.

Let $s = (w_0, w_1, \dots, w_S)$ denote a sentence of length S , where w_0 is the special token for the head of the root. The representation of w_0 within the transformer encoding of the sentence is associated with the [CLS] token. In order to use a technique similar to self-attention, we exploit some parts of the corresponding matrices for each word form in the sentence. Thus we define the following matrices and vectors:

- Let $h_i = Model(w_i)$ be the embedding of w_i , produced by an encoder model. ($h_i \in \mathbf{R}^{d_e}$) for $i \in [0, S]$;
- Let $q_i = QueryMatrix(h_i)$ and $k_j = KeyMatrix(h_j)$ ($q_i, k_j \in \mathbf{R}^{d_k}$) be linear (affine) transformations of h_i for $i \in [1, S]$ and of h_j for $j \in [0, S]$;
- Let $K \in \mathbf{R}^{d_k \times S+1}$ be the matrix with rows - k_j for $j \in [0, S]$.

The distribution over all possible heads of w_j is obtained with softmax across the multiplication of q_i and K ($q_i K \in \mathbf{R}^{S+1}$). The encoder model weights and the transformations are trained with a cross entropy loss between the distribution over the heads and the one-hot encoded label of the correct head:

$$Loss = - \sum_{i=1}^S \sum_{k=0}^S y_{i,k} \log((\text{softmax}(q_i K))_k)$$

where:

- S is the sequence length.
- $y_{i,k} = \begin{cases} 1, & \text{if } w_k \text{ is the head of } w_i \\ 0, & \text{otherwise} \end{cases}$

- $q_i \in \mathbf{R}^{d_k}$ is the output vector of the query matrix transformation for the word w_i .
- $K \in \mathbf{R}^{d_k \times S+1}$ is the matrix with rows - the outputs of the key matrix transformation for all words in the sentence.

This end-to-end training fine-tunes the encoder model weights. It is done simultaneously over all words in a sentence and over multiple sentences in a batch. During inference the model produces distribution over the possible heads for each word. A simple strategy to predict the head of each word is to calculate the *argmax* of the distribution.

$$Prediction(w_i) = \text{argmax}_{k=0}^S (q_i K)_k$$

We use this prediction for validation during training. However, it is well known that this greedy prediction does not guarantee a construction of a tree (although in more than 95 % of the cases a tree is produced). Thus, we adopted the Chu-Liu-Edmonds algorithm for the construction of a Maximum Spanning Tree over the full graph of all potential dependency arcs of the sentence to implement and to select the most probable tree (McDonald (2006)). The full graph in our case is presented as a transition matrix composed of the vectors for the distribution of the possible heads for each word in the sentence. Over this graph we apply the Chu-Liu-Edmonds algorithm.

Our head classification layer can be seen as a simplified version of the (Dozat and Manning, 2017) Deep Bi-affine attention. While they use an LSTM to produce embeddings, we use BERT and our scores are just the dot products of the heads and dependents transformations while they transform the outputs and then use a Bi-affine transformation to produce the scores. We argue that a simpler layer is sufficient, because of the expressive power of the pre-trained BERT.

3.2 Incorporating the Information from the Lexicons and Shallow Grammars

As it was mentioned above, our goal is to incorporate “sure” information about the syntactic structure of a given sentence in the process of parsing in such a way that it improves the performance of the parser. Such information could be used during the training time (fine-tuning) of the parser as well as during the inference time.

Let $s = (w_0, w_1, \dots, w_S)$ denote a sentence of length S , and w_0 is the special token for the root as above. T is a dependency tree for s if and only if

$$T = \{(w_i, w_j, l_i) | W_i \in s, w_j \in s, l_i \in DL\},$$

where T is a tree, the root of T is w_0 , w_j is the head node of w_j , and DL is a set of labels for the dependency relations. The “sure” syntactic information for the sentence s is a subset of arcs in the dependency tree T : $T_P \subseteq T$. We call T_P a set of prompting arcs. In the experiments reported in this paper we use only unlabeled arcs, because we would like to see their influence on the unlabeled parsing. In our opinion, the additional information — the labels and grammatical features — might be incorporated in a similar manner. Thus, T_P contains arcs from some of the words in the sentence to their heads.

Our *main intuition* is that we could use the prompting arcs to urge the model to pay more attention to the “sure” heads provided by T_P .

The incorporation of the additional information from the set T_P can be done in at least two ways. First, through a simple extension over the model, described in the previous subsection, is to modify the scores for the corresponding arcs predicted by the model to an infinitely large score. In this way we will force the Chu-Liu-Edmonds algorithm to always select these arcs. A similar approach could be used with the argmax head selection algorithm.

One disadvantage of this method is that it has an effect only during the actual head selection phase. Thus, the transformer model cannot take advantage of the predefined arcs. Motivated by the intuition that incorporating the information as early as possible would facilitate the model predictions for the other words, as a second solution we propose an extension to the layers of the encoder model, which are used to prompt the model with the predefined information. Since only a fraction of the dependency arcs are predefined, this prompting is done only on a small number of words in the sentence. Thus, to implement this intuition within the model we modify the typical architecture of the transformer model.

Let us consider the standard architecture of the transformer block for the encoding model. It consists of two major elements: the first element includes a Multi-Headed Self-Attention layer (MHA) with following residual connections and normalization. The second element is a Feed-Forward Network (FFN), also with following residual connections and layer normalization (LN). The output of each of the elements are denoted in

the following way:

$$O_{attn}(X) = LN(X + MHA(X))$$

is the result of the first element — Multi Headed Self-Attention, residual connections and layer normalization. Then

$$O_{ffn}(X) = LN(X + FFN(X))$$

is the result of the second element — Feed-Forward Network, residual connections and layer normalization.

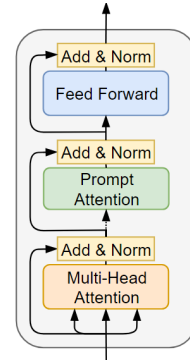


Figure 1: The modified encoder layer with prompt attention

Our modification introduces a bias to the embedding of each token towards the embedding of its predefined head. This is done by the prompting attention (PA) sublayer:

$$O_{pa}(X) = \left(LN(w_i + \sum_{j=1}^S I(i, j) * Prompt(w_j)) \right)_{w_i \in X},$$

where

- $Prompt(x)$ is a learnable linear (affine) transformation which transforms the head embedding.

$$I(i, j) = \begin{cases} 1, & \text{if } w_j \text{ is the predefined head} \\ & \text{of } w_i \\ 0, & \text{otherwise} \end{cases}$$

The function $I(i, j)$ is an input to the model and it behaves as a matrix which indicates the predefined arcs. The final modified encoder layer looks like this:

$$Layer(X) = O_{ffn}(O_{pa}(O_{attn}(X))).$$

The graphical representation of the modified Transformer block is depicted in Fig. 1.

The Multi Head Attention and FFN parameters are initialized from the weights of the pre-trained model, while the prompt attention parameters are randomly initialized and later learned by fine-tuning. The current implementation allows for a prompt attention sublayer only in some pre-selected layers in the BERT architecture. In this way we could use it only for some of BERT layers. Adding the prompt attention to the last few layers of the model produces best results. We prove this by performing experiments with different settings. In our opinion the reason for this is as follows: adding it to more layers of BERT introduces too many newly initialized parameters and they require longer training.

After the modifications, the head classification layer from 3.1 is appended to the model and it is fine-tuned by an end-to-end training.

3.3 Relation Classification — (LAS)

After receiving the structure of the syntax tree, another model is trained to predict the labels of the word - head arcs.

Let $h_i = Model(w_i)$ and $h_j = Model(w_j)$ be the embeddings of the words w_i and w_j and let w_j be the head of w_i .

The number of relation classes are of a fixed size, so a linear (affine) classifier can be used. The embeddings of the word and its head (predicted in the previous step) are concatenated, then passed through the transformation. The distribution across the possible classes for the relation between w_i and its head w_j is $c_i = Classifier(Concat(h_i, h_j))$. The model is end-to-end trained with cross-entropy loss. Currently the addition of some predefined information for the label classification is outside the scope of our work.

4 Experiments

In this section we present the experiment settings that were used to evaluate the new dependency architecture as well as the results from the different experiments. We performed experiments with two Universal Dependency Treebanks: *BTB Bulgarian Treebank* (Osenova and Simov 2015) and the *GUM English Treebank* (Zeldes, 2017). The Bulgarian treebank was selected because we are mainly interested in Dependency Parsing for Bulgarian. Also, we have access to many language resources and tools for Bulgarian like Chunk grammars for recognition of noun chunks, verbal com-

plex chunks, prepositional chunks, lexicon with MWEs (still quite modest as a coverage). By performing experiments also for English, we wanted to provide some initial evidence that our architecture is not language specific. The GUM English Treebank was selected because its size is similar to that of BTB Bulgarian Treebank.

For the experiments with the Bulgarian parser we used a pre-trained BERT model with 355M parameters as an encoder which produced the initial embeddings of the tokens. The BERT model was trained by us on 20B of Bulgarian tokens. Our pre-training dataset consists of mainly Web data, literature, administrative and scientific documents, as well as Wikipedia articles. The model was trained for 3 epochs and the pre-training took 23 hours for a single epoch on 16 Nvidia A100s. The models will be uploaded on Huggingface.

For the experiments with the English parser, BERT large uncased was used (Devlin et al., 2018) because the model architecture is similar to our pre-trained BERT. The difference in the parameter count comes from the bigger embedding layer because of the larger vocabulary size of our model.

The notion of $T_P \subseteq T$ — the set of “sure” arcs in the dependency tree T , can easily extended to a whole treebank by applying the same procedure to each sentence in a given treebank. We denote the set of arcs for the whole treebank as $T_P(TreeBankName)$ with additional superscripts if necessary. For Bulgarian we adopted the available constituent-based cascaded chunk grammars where each rule was applied over each sentence annotated with grammatical features (the XPOS column of the CoNLLU format was used as defined for all Universal Dependency treebanks). The rules are ordered and applied according to the specified order on the basis of the result from the previous rules. A very simple example is the following one: if the current sentence contains a preposition (R) and a noun chunk ($NChunk$), then the following rule can be applied:

$$R, NChunk \rightarrow PChunk$$

When the whole grammar is applied, the arcs within each of the chunks are selected for the corresponding sentence. For example, for the above PChunk we could predict an arc from the preposition to the head of the $NChunk$ with label “case”. The arcs and their labels could be defined uniquely on the basis of the chunks and grammatical features of the words in them.

The lexicon of the MWEs contains a uniform rep-

Model	Training set	$T_P(BTB)^{ChMWE}$		$T_P(BTB)^0$		$T_P(BTB)^{20}$	
		UAS	LAS	UAS	LAS	UAS	LAS
Corrected Argmax	$T_P(BTB)^0$	0.9640	0.9361	0.9614	0.9335	0.9694	0.9409
<i>Corrected MST</i>	$T_P(BTB)^0$	0.9640	0.9361	0.9615	0.9337	0.9695	0.9410
Prompted-10	$T_P(BTB)^{10}$	0.9655	0.9370	0.9626	0.9340	0.9690	0.9400
Prompted-20	$T_P(BTB)^{20}$	0.9641	0.9360	0.9606	0.9324	0.9700	0.9411
Prompted-0-40	$T_P(BTB)^{0+40}$	0.9672	0.9392	0.9640	0.9362	<u>0.9718</u>	<u>0.9433</u>
Prompted-ChMWE	$T_P(BTB)^{ChMWE}$	0.9655	0.9374	0.9510	0.9231	0.8307	0.8665

Table 1: Accuracy of UAS and LAS of the models on the UD_Bulgarian-BTB test set with different subsets of predefined arcs. The different models are trained on different training sets. The first two models were trained on the treebank without any prompting arcs. For these two models the MST ones are performing generally better. Thus, we selected *Corrected MST* as a baseline model (highlighted in bold and italics) because it achieved the best result on the treebank without any prompting arcs. As the best new model we selected *Prompted-0-40* because it achieved the best result (highlighted in bold) over $T_P(BTB)^{ChMWE}$ — ($ChMWE = \mathbf{C}$ hunk grammars and \mathbf{M} WE lexicon). This is a realistic scenario, because the prompting arcs are produced by shallow grammars and the lexicons which could be applied over new texts. This model also produced better results over the treebank without prompting arcs. *Prompted-0-40* produced even better results (underlined in the table) over the test set with 20 % random prompting arcs. But this is an unrealistic scenario because we do not have reliable sources for such prompting arcs.

resentation of each MWE which contains not only the strings, but also dependency relations for the structure of the MWE and some grammatical features of its internal elements — see (Osenova and Simov, 2024). Here only MWEs that are realized continuously in the text, and which are unambiguous are used.

The set of arcs selected in this way for the universal BTB treebank is $T_P(BTB)^{ChMWE}$. It contains around 25 % of all arcs in the treebank. When there are not available grammars, lexicons, or just for experiments appropriate sets of arcs could be selected randomly from the treebank itself. In these cases we could have sets such as: $T_P(BTB)^{10}$, $T_P(BTB)^{20}$, $T_P(BTB)^{30}$ containing 10 %, 20 %, 30 % and so on of the arcs in the treebank. Similarly, $T_P(GUM)^{10}$, $T_P(GUM)^{20}$, $T_P(GUM)^{30}$ for the English treebank. In some cases we also use $T_P(BTB)^0$ for the treebank without any prompting arcs selected. Also we use $T_P(BTB)^{0+40}$ to denote the shuffled union of two copies of the treebank: one without any prompting arcs and one with 40 % of prompting arcs.

These sets of arcs could be used during the training, validation and testing of the corresponding parsers. Obviously, the analyses of the new texts will require shallow grammars and/or lexicons of MWEs. If not available, the parser will be used without the predetermined sure arcs.

When the predefined sets of arcs are used to influence only the final decisions of the head selection algorithms (as opposed to injecting data into the

model) we call the setup *Corrected models*. Thus, we have *Corrected Argmax* and *Corrected MST models*. As it was mentioned before, for a baseline of the models we use *Corrected MST models*. The corrected Argmax will be reported in order to demonstrate the performance of the parser model strictly by itself.

Also, the fact that in most of cases *Corrected Argmax* and *Corrected MST* produced very close results is an evidence that the transformer model does some reasoning which ensures a tree-like structure of the output graph. Our intuition is that this reasoning happens in the last layers of the encoder, just before the head selection layer, since the head selection layer predicts the head of each word independently. Thus no information sharing can happen there. This motivates the decision to incorporate outside information about the tree in these layers.

4.1 Training and testing set-up

We considered 3 training and testing set-ups:

- Training and testing with predefined arcs produced by arbitrary fixed size subsets of all arcs in the treebanks - ($T_P(BTB)^{10}$, $T_P(BTB)^{20}$ and so on).
- Training and testing with predefined arcs produced by a shallow grammar parser and a lexicon ($T_P(BTB)^{ChMWE}$) — only for Bulgarian.
- Training with predefined arcs produced by ar-

bitrary subsets ($T_P(BTB)^k$) and testing with predefined arcs produced by a shallow grammar parser and a lexicon ($T_P(BTB)^{ChMWE}$) — again only for Bulgarian.

Using an arbitrary predefined subset of arcs as prompts is sufficient to train the model to recognize the prompts and produce good results when testing with both — arbitrary or custom predefined arcs. The ability to train with an arbitrary subset of prompts is important in case of insufficient linguistic resources.

The fine-tuning training is done for 10 epochs with a learning rate of 5e-5 with linear decay and batch size of 384. The fine-tuning takes around 5 minutes on 8 Nvidia A100s. The best performing model checkpoint over the 10 epochs on the validation set is selected.

4.2 Results

In this subsection we present some results from the experiments. In Table 1 the evaluation of Bulgarian parsers is given. The best performing Prompted model for Bulgarian on UAS was trained on a shuffled union of the BTB train set with no predefined arcs and the BTB train set with 40% predefined arcs ($T_P(BTB)^{0+40}$), which enables the model to 'see' sentences without any predefined arcs during training. A paired t-test over 10 training sessions with different random seeds for weight initialization was done and it shows that the improvement in accuracy of the *Prompted models* in comparison to the *Corrected MST* is statistically significant and thus not a product of lucky weights initialization.

Different experiments were made regarding the size of the set of predefined arcs during training and the number of modified encoder layers: We found that the size of the predefined subset should be neither very large nor very small: if too small (under 5% of all arcs) — the model cannot learn the meaning of the prompts and does not use the predefined information. If too large (more than 50%) — the model cannot learn to do parsing on its own. We found that 20% of predefined arcs is an optimal overall size. In addition to the size, it is important to be mentioned that when the prompts are selected randomly, they also belong to different kinds of arcs representing different phenomena within the dependency trees. Having different types of prompt arcs enables the model to better generalize over the meaning of the prompts. This is one explanation why the set of arcs, produced by the shallow gram-

mars and the lexicons ($T_P(BTB)^{ChMWE}$) is not so good for fine-tuning of the dependency graphs comparing to the set with 20 % randomly selected arcs ($T_P(BTB)^{20}$). This is a consequence of the nature of the shallow grammars and most of the MWEs in our lexicon.

The number of the modified layers also matters. We made experiments with adding the prompt attention to half or even all layers, but doing so introduces too many new parameters thus making the training require more examples. In our case the models with modified 2 to 4 of the last layers performed best. The results reported in this paper were produced by models with attention modification considered only for the last 4 out of the 24 layers of BERT (layers 21-24).

The quality of the pre-trained encoder model is also important. Our previous experiments for Bulgarian were done with a BERT model pre-trained on a significantly smaller dataset (4B tokens). While the improvement by adding the modification was still present, the general scores were lower.

In Table 2 the evaluation of the English parsers is given. The results show that not only the improvement with the proposed modification is maintained but there is even some improvement in the case when there are no predefined arcs.

5 Manual Evaluation and Discussion

We performed some manual comparison of the result from the baseline model *Corrected MST* and the best model **Prompted-0-40** for Bulgarian. The two models are correct with respect to the fixed expressions. Thus, the influence of adding some preliminary syntactic knowledge seems to affect the overall analyses and help in cases of correct head identification and direction as well as other phenomena like the PP attachment, apposition relations, etc.

Here two cases are considered: a) the baseline makes errors while the best model is correct, and b) in the opposite direction - the best model makes errors while the baseline is correct.

The baseline makes errors. When inspecting the errors of the baseline where the best model has taken correct decisions, the following main cases have been identified:

- *wrong head direction*: for example, the subject of a copula should be dependant on the

Model	Training set	$T_P(GUM)^{20}$		$T_P(GUM)^0$	
		UAS	LAS	UAS	LAS
Corrected Argmax	$T_P(GUM)^0$	0.9436	0.9246	0.9299	0.9125
<i>Corrected MST</i>	$T_P(GUM)^0$	0.9447	0.9256	0.9308	0.9133
Prompted-10	$T_P(GUM)^{10}$	0.9467	0.9273	0.9321	0.9143
Prompted-20	$T_P(GUM)^{20}$	0.9471	0.9280	0.9310	0.9138

Table 2: The accuracy of UAS and LAS of the models on the UD_English-GUM test set with different subsets of pre-defined arcs. The experiments reported here only demonstrate that the behaviour of the models follows the same pattern. The models trained on treebanks augmented with prompting arcs achieve better results even on treebank data without prompting arcs.

content word of the copula predicative but it erroneously was analyzed as depending on the copula

- *wrong head selection*: for example, in Bulgarian NN construction with the first noun indicating quantity, the head is the first noun, while in the baseline the second one was chosen.
- *wrong head assignment*: for example, the subject should be related to the main verb of a sentence but it was assigned to the modal verb instead; in an appositive structure the modifier of the head noun in the dependant structure is wrongly attached to the head of this dependant structure
- *wrong root assignment*: for example, in complex sentences, the baseline assigns the root relation to both verbs — in the main sentence as well as in the clause
- *wrong PP attachment*: for example, instead of depending on the noun, the head of the PP is made dependant on the verb
- *wrong non-PP attachment*: for example, the adverb is adjacent to the preceding noun but has to be attached to the following verb. However, it was wrongly attached to the noun; when the complementizer ‘che’ (that) is used in non-typical for it structures like after the negative particle ‘ne’ (not), the complementizer is wrongly attached to the negative particle instead of to the verb

The best model makes errors. When inspecting the errors of the best model where the baseline had taken correct decisions, the following main cases have been identified:

- *wrong head direction*: the same error as in the baseline error list
- *wrong head assignment*: for example, in more embedded clauses, the last verb is wrongly attached to the very initial one instead of the nearest governor
- *wrong PP attachment*: the same error as in the baseline error list

As it can be seen from the observations above, both models generally make identical types of errors. At the same time it seems that the best one has more attachment-related issues while the baseline — more head-related ones. From the statistics over the errors it can be seen that the baseline makes more errors per category than the best one. Both models have almost the same difficulties with the following labels: *obl* and *advmod*. Thus, all adverbials — despite being expressed by adverbs or nominals, cause problems towards the proper analyses. Also, it seems that the processing of the relation *obj* is easy for the best model while not for the baseline; the processing of the relation *discourse* is easy for the baseline while not for the best model.

The manual validation of the results from the two models shows that the extension of the transformer architecture with the new prompt attention layer improves the general performance of the head selection model. But it also shows that the improvement is not an extension of the baseline model. Instead, the extended model covers a different part of the search space. Thus we plan to address this discrepancy in several ways: (1) through the improvement of the prompt attention layer by including more linguistic information such as higher order arc information, grammatical features, shallow semantic information — ontological information for NEs, terms and key words, where this information is reliable; (2) through the extension of the treebank with

new sentences selected using some active learning procedure. (3) through the improvement of the shallow grammar and the coverage of the MWE lexicon as well as the related algorithms for their better prediction and consequent recognition in text.

6 Conclusion and Future Work

In this work we extended the standard transformer block architecture with a new *prompt attention* layer which incorporates the information from some external knowledge sources like shallow grammars and MWE lexicons. In this way the BERT-based dependency parsing model was internally modified to produce a better dependency parsing model. Here some experimental settings were described where the inclusion of shallow syntactic knowledge and knowledge from MWE lexicons improves the parsing model for Bulgarian. Our assumption is that this architecture would be applicable to any other language. To initially prove this, we also performed experiments with the English UD treebank — GUM.

In our future work we plan to use deeper syntactic knowledge as well as improved shallow syntactic knowledge and semantic information — not only during the fine-tuning stage but also during the pre-training. We plan to make experiments with some variant of the multilingual dependency parsing where the models are simultaneously trained on more than one UD treebank.

Acknowledgments

The reported work has been supported by CLaDA-BG, the *Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage*, part of the *EU infrastructures CLARIN and DARIAH*. We also acknowledge the provided access to the *e-infrastructure of the Centre for Advanced Computing and Data Processing* (the Grant No BG05M2OP001-1.001-0003).

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: pre-training of deep bidirectional transformers for language understanding](#). *CoRR*, abs/1810.04805.

Timothy Dozat and Christopher D. Manning. 2017. [Deep Biaffine Attention for Neural Dependency Parsing](#). *Preprint*, arXiv:1611.01734.

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. [Stanford’s Graph-based Neural Dependency Parser at the CoNLL 2017 shared task](#). In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30, Vancouver, Canada. Association for Computational Linguistics.

Voula Giouli and Verginica Barbu Mititelu, editors. 2024. [Multiword expressions in lexical resources](#). Number 6 in *Phraseology and Multiword Expressions*. Language Science Press, Berlin.

Ryan McDonald. 2006. [Discriminative Training and Spanning Tree Algorithms for Dependency Parsing](#). Ph.D. thesis.

Ryan McDonald, Kevin Lerman, and Fernando Pereira. 2006. [Multilingual dependency analysis with a two-stage discriminative parser](#). In *Proc. of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X ’06, page 216–220, USA. ACL.

Petya Osenova and Kiril Simov. 2015. [Universalizing BulTreeBank: a linguistic tale about glocalization](#). In *The 5th Workshop on Balto-Slavic Natural Language Processing*, pages 81–89, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Petya Osenova and Kiril Simov. 2024. [Representation of multiword expressions in the Bulgarian integrated lexicon for language technology](#). In Voula Giouli and Verginica Barbu Mititelu, editors, *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, *Phraseology and Multiword Expressions*, chapter 6, pages 117–146. Language Science Press, Berlin.

Saziye Betül Özates, Arzucan Özgür, Tunga Güngör, and Balkiz Öztürk. 2020. [A hybrid approach to dependency parsing: Combining rules and morphology with deep learning](#). *CoRR*, abs/2002.10116.

Yikang Shen, Shawn Tan, Alessandro Sordani, Siva Reddy, and Aaron Courville. 2021. [Explicitly modeling syntax in language models with incremental parsing and a dynamic oracle](#). In *Proc. of the 2021 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 1660–1672, Online. ACL.

Ryo Yoshida, Taiga Someya, and Yohei Oseki. 2024. [Tree-planted transformers: Unidirectional transformer language models with implicit syntactic supervision](#). *Preprint*, arXiv:2402.12691.

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

Houquan Zhou, Yang Hou, Zhenghua Li, Xuebin Wang, Zhefeng Wang, Xinyu Duan, and Min Zhang. 2023. [How well do large language models understand syntax? an evaluation by asking natural language questions](#). *Preprint*, arXiv:2311.08287.

A Multilingual Parallel Corpus for Coreference Resolution and Information Status in the Literary Domain

Andrew Dyer, Ruveyda Betul Bahceci, Maryam Rajestari,
Andreas Rouvalis, Aarushi Singhal, Yuliya Stodolinska,
Syahidah Asma Umniyati, Helena Rodrigues Menezes de Oliveira Vaz

Language Science and Technology
Saarland University
Saarbrücken, Germany

{firstname.lastname}@uni-saarland.de

Abstract

Information status — the newness or givenness of referents in discourse — is known to affect the production of language at many different levels. At the morphosyntactic level, information status gives rise to special word orders, elisions, and other phenomena that challenge the notion that morphosyntax can be considered independent of discourse context. Though there are many language-specific corpora annotated for information status and its related phenomena, coreference and anaphora resolution, what is not available at present is a cross-lingually consistently annotated corpus or annotation scheme that would allow for comparative study of these phenomena across many diverse languages. In this paper we present our work to build such a resource. We are annotating a parsed, parallel corpus of prose in many languages for information status and coreference resolution, so that like-for-like cross-lingual comparisons can be made at the intersection of discourse and syntax. Our corpus can and will be used both for corpus analysis and for model training.

1 Introduction

When speakers¹ produce sentences, utterances and meanings, they usually do so not in isolation, but in the context of a longer discourse and in a communicative context between speakers with shared knowledge of the world that coincides or differs in important ways. The shared world knowledge between speakers mediates what meaning can be interpreted from utterances (Beyer, 2015), while common ground in conversation mediates what information need be explicitly stated (Karttunen, 1974).

Central within this dynamic is information status: broadly, whether information communicated is

¹We use *speakers* as it is the term for those who produce language that will be most readily understood, but these arguments apply equally to signed languages, as well as the written modality.

new – that is, being encountered or asserted for the first time; or *given* – the information has been introduced before, or is otherwise already inferable by the receiver (Chafe, 1976). Broadly speaking, information, referents and arguments that are considered known in the common ground of the discourse may be reordered, reduced, receive special (intonational) markers, or may even be omitted altogether in the aid of information flow, allowing processing time and emphasis for the assertion of more novel or surprising information (Fenk-Oczlon, 2001).

How this plays out varies widely across languages. Languages such as English have definiteness as a grammatical feature within the noun phrase, thus allowing their hierarchy of givenness to be visible through word forms (Gundel et al., 1993). Other languages, such as Czech and Hungarian, convey the givenness of information through word order, and are considered *discourse configurational* languages as a result of this expectation (Kiss, 1995). Additionally, many languages allow for given information to be omitted from the sentence entirely, either relying on indexing arguments through morphological processes on root words, or by relying on speakers to infer arguments from context. Japanese is an example of such a language (Vermeulen, 2012). In these languages, information flow is handled by simple elision of overt arguments.

The role of information status on language production has been well studied in individual languages using both psycholinguistic experimentation and corpus study. Seminal studies include Arnold et al. (2000) on word order in English, Skopeteas and Fanselow (2010) on cross-linguistic differences in the expression of focus, and Wang et al. (2012) on the so-called Chomsky illusion, showing how focus is a determinant of depth of syntactic processing in Mandarin Chinese.

There has also been increasing interest in cross-lingual comparison of the way information status

is signalled and the way it affects language production in the world’s languages.

To study the influence of information status on syntax cross-lingually – for example, the shifting of given information to a sentence-initial position, or the use of pronominal forms for an entity that is currently active in the discourse – we need corpora that are multilingual and consistently annotated both for syntax and information status (Lüdeling et al., 2014).

Information status is closely related to the task of coreference and anaphora resolution: the identification of expressions in a text that refer to the same entity, and there are several corpora that combine these two tasks (Markert et al., 2012; Zeldes, 2017). In the interests of cross-lingual natural language processing, there have been efforts to bring diverse corpora for coreference and anaphora resolution together into a common format (Nedoluzhko et al., 2022), and there are beginning efforts towards consistent multilingual annotation (Poesio et al., 2024). However, as of yet there is no resource that fully meets the criteria that we need met in order to pursue multilingual comparative studies.

We introduce our work to develop such a resource. We annotate on top of a parallel corpus of modern literature in translation, predictively parsed according to Universal Dependencies annotation. We annotate spans of entity mentions, with coreference chain annotation to track mentions of the same underlying entity; and information status and mention type annotations to describe the mention. In this way, we can use the underlying syntactic annotation of sentences to follow the placement of referring expressions, to quantify how the information status of such expressions, their mention type, and the recency of mentions of the same entity in the discourse, affect the order in which they are placed.

We annotate texts in a diverse variety of languages, with common annotation guidelines applying to each language that is added. As each new language is added, we work to ensure that our annotation principles and guidelines apply consistently to each language, ensuring that like-for-like comparisons can be made between languages.

Our corpus has the following benefits:

- **Parallel:** The texts used in the corpus are direct translations of works of prose in each language. This makes it easier to make direct comparisons of phenomena between languages.

- **Minimalist:** We are conservative with regard to mention spans, including only the most relevant information and minimising overlap. This makes annotation easier and faster, and visually clearer for users and programs.
- **Feature modularity:** We make features modular, increasing the efficiency and precision of annotation. This allows flexible and granular descriptions of mentions while avoiding feature explosion, and is simpler for annotators and readers than a lengthy list of features.

In this paper, we will describe and motivate our annotation scheme in the context of existing resources; and discuss our current workflow and progress in annotation.

2 Related Work

There are many monolingual corpora for coreference resolution and/or information status that have been used for quantitative study of the effects of word order. For example, *RefLex* (Baumann and Riester, 2012) and *ISNotes* (Markert et al., 2012) are corpora in German and English respectively, with span annotation of entity mentions, coreference links, and nuanced categories of mention type. *OntoNotes* (Hovy et al., 2006) is among the most widely used corpora for coreference and anaphora resolution, and covers English, Arabic and Mandarin Chinese. The *Georgetown University Multi-layer corpus (GUM)* (Zeldes, 2017) is a multimodal corpus of English annotated with UD syntactic structure, coreference, and information status, among many other layers of annotation. The information status and mention type labels of *GUM* are inherited by our scheme.

Many coreference resolution corpora — including *GUM* – from a variety of European languages have been assembled and harmonised in *CorefUD* (Nedoluzhko et al., 2022), where coreference annotation is joined with predictive Universal Dependencies parsing. The harmonisation of many schemes into a common format has been the basis of considerably many experiments and advances in training cross-lingual and multilingual coreference resolution models (Ogrodniczuk et al., 2023).

Universal Anaphora² (Poesio et al., 2024) is a Universal Dependencies-inspired effort to create a common framework for annotation of coreference resolution so that coreference and information sta-

²<https://universalanaphora.github.io/UniversalAnaphora/>

tus can be compared across languages in a similar manner to Universal Dependencies. As of the time of writing, Universal Anaphora has contributed an enhanced file format for representation of coreference resolution (the conll-UA format), and a wide range of tools for scoring and validation of coreference resolution models, but work to create a common linguistic annotation scheme has not yet been undertaken.

To our knowledge, there are no currently existing *parallel* multilingual corpora annotated for both coreference resolution and information status, and this is where we seek to make our contribution.

3 Data and format

3.1 Data

The corpus that we use as the base for our annotation is *mini-CIEP+* (Verkerk and Talamo, 2024). *mini-CIEP+* is a multilingual parallel³ corpus of modern prose in translation. The corpus is predictively parsed according to Universal Dependencies (Nivre et al., 2020) using Stanza (Qi et al., 2020).⁴ The corpus is thus represented in conllu format⁵. The corpus covers 40 languages at the time of writing, with more to be added.

From among this data, we have annotated data from books in seven languages: English, Ukrainian, Modern Greek, Portuguese, Hindi, Turkish, and Indonesian. The choice of these languages is motivated by linguistic diversity: the languages come from a variety of families (Indo-European, Turkic, Austronesian), and exhibit varying degrees of word order freedom, morphological indexing and pro-drop. Accommodating these languages early on allows us to address the linguistic challenges that arise from them.

The data being drawn from the literary domain presents its own challenges. Compared with the more formal styles favoured in many resources such as OntoNotes and GUM, the literary genre includes complicated annotation issues such as asymmetry of knowledge between characters, changes in entities, and lexical variation in entity description (Han et al., 2021). The benefit of this challenge is that we expect to encounter more idiosyncratic and

³Parallel in the sense that the same work is represented - either in original or translation - in each language, and thus the context is the same. The texts are not strictly bitexts, or sentence- or token-aligned, but contain in theory the same content, ensuring comparability.

⁴<https://stanfordnlp.github.io/stanza/>

⁵<https://universaldependencies.org/format.html>

diverse language use, which has benefits both for diversity of sampling and for model training.

3.2 Format

Our annotation of the corpus is output in the *CorefUD* format⁶. The CorefUD format follows that of conllu, but places mention span annotation in the *misc* column, along with other data concerning coreference relations. The building blocks in this format are *spans* and *clusters*. Mentions of entities in the discourse are represented by a tuple-like span object, opening on the token where the span begins, and closing on the token where it ends. Within this span is contained an entity ID, specifying the ID of the underlying entity of the mention, as well as various other attributes. We refer to the CorefUD file format description for more details and examples, but we give an example of the output of our corpus in Fig 1.

We choose to follow this format as closely as possible so as to be able to integrate our corpus with existing resources, including CorefUD corpora and evaluation scripts, so that we can train models to parse more of the corpus and further corpora.

4 Annotation design

4.1 General principles

In the interests of speedy annotation, and to avoid overburdening annotators with too many labels, we try to keep our labels simple and modular. That is to say, that rather than giving annotators a deep hierarchy of labels to choose from, we aim to give a set of attributes with limited options, as shown in Table 1.

For example, we only use two labels for information status: *given* and *new*. There are finer grained measures of coreference, such as the near-identity relations used by Recasens et al. (2011), that we do not include. We also do not include focus, often cited as a central part of information structure, due to the difficulty of defining this in a cross-lingually satisfactory way (Matić and Wedgwood, 2013).

4.2 Markables

4.2.1 Markable spans

A markable is a span of text that may constitute an entity mention. (Dipper et al., 2007) The following structures are always annotated as markables:

⁶<https://ufal.mff.cuni.cz/~popel/corefud-1.0/corefud-1.0-format.pdf>

```

# sent_id = Alquimista_English_006_2
# text = During the two hours that they talked, she told him she was the merchant's daughter ...
1   During during ADP      IN      _      4      case      _      TokenRange=74:80
2   the the DET      DT      Definite=Def|PronType=Art 4      det      _      Entity=(e7-time-3-CorefType:coref,InfStat:new|TokenRange=81:84
3   two two NUM     CD      NumForm=Word|NumType=Card 4      nummod   _      TokenRange=85:88
4   hours hour NOUN   NNS    Number=Plur 10     obl      _      Entity=e7)|TokenRange=89:94
5   that that PRON   WDT    PronType=Rel 7      obj      _      TokenRange=95:99
6   they they PRON   PRP    Case=Nom|Number=Plur|Person=3|PronType=Prs 7      nsubj    _      Entity=(e8-person-1-CorefType:ana,InfStat:given)|
SplitAnte=e2<e8,e1<e2|TokenRange=100:104
7   talked talk VERB   VBD    Mood=Ind|Number=Plur|Person=3|Tense=Past|VerbForm=Fin 4      acl:relcl _      SpaceAfter=No|TokenRange=105:111
8   , , PUNCT   _      _      10     punct   _      TokenRange=111:112
9   she she PRON   PRP    Case=Nom|Gender=Fem|Number=Sing|Person=3|PronType=Prs 10     nsubj    _      Entity=(e2-person-1-CorefType:ana,InfStat:
given)|TokenRange=113:116
10  told tell VERB   VBD    Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin 0      root     _      TokenRange=117:121
11  him he PRON   PRP    Case=Acc|Gender=Masc|Number=Sing|Person=3|PronType=Prs 10     iobj     _      Entity=(e1-person-1-CorefType:ana,InfStat:
given)|TokenRange=122:125
12  she she PRON   PRP    Case=Nom|Gender=Fem|Number=Sing|Person=3|PronType=Prs 17     nsubj    _      Entity=(e2-person-1-CorefType:ana,InfStat:
given)|TokenRange=126:129
13  was be AUX    VBD    Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin 17     cop      _      TokenRange=130:133
14  the the DET     DT      Definite=Def|PronType=Art 15     det      _      Entity=(e2-person-4-CorefType:pred,InfStat:new(e3-person-2-
CorefType:coref,InfStat:given)|TokenRange=134:137
15-16 merchant's _      _      _      _      _      _      Entity=e3)|TokenRange=138:148
15  merchant merchant NOUN   NN      Number=Sing 17     nmod:poss _      _
16  's 's PART    POS    _      15     case     _      _
17  daughter daughter NOUN   NN      Number=Sing 10     ccomp   _      Entity=e2)|SpaceAfter=No|TokenRange=149:157

```

Figure 1: An example sentence from our corpus (from the English portion) in CorefUD format, with entity annotation in the *misc* column. Note that mention spans may open on one token and close on another, and that two mentions may start or end on the same token (but may not cross each other).

- Referring pronouns (excluding dummy pronouns and relative pronouns)
- Referring noun phrases (excluding idiomatic instances)

Additionally, we annotate as markables these structures if they are coreferred by an anaphoric mention:

- Interrogative and quantifying pronouns, e.g. *whoever, anything*
- Verbal and other non-nominal phrases that are referred to anaphorically as discourse deixis (Dipper and Zinsmeister, 2009); for example "[He said no]. [That] surprised me"
- Pro-adverbs such as *here* and *then*

Pronominal clitics may also be annotated as markables provided that they are not part of an introverted reflexive verb phrase (Haspelmath, 2008). These are common in many Indo-European languages such as Portuguese and Dutch, where they simply reinforce that the agent of a verb is the same as its patient.

The greatest divergence with most other schemes in *CorefUD* in terms of annotation philosophy is that we are more minimalist with what we include in a markable. Such schemes typically cover the full syntactic noun-phrase, including all determiners, modifiers, adjuncts and clausal expansions. By contrast, we opt for an approach where only the most relevant information used to identify the entity is included. This always includes the syntactic head, but adjuncts and modifiers are only included if they provide information that is essential to understanding and referring to the referent.

We use some linguistic tests to decide on what information should be included when annotating a markable span:

- *Question test*: If we form a question to which the entity being referred to is the answer, would the same wording typically be used in the answer?
- *Repeated mention test*: Would the same wording be used (or is it used) in a subsequent mention to refer to the entity?
- *Contrast test*: Does the wording of the mention serve to contrast this referred entity with another similar entity?

Likewise, we also do not include possessive pronouns as part of the markable span (but may include them in their individual spans, see ex (1))⁷, and we do not mark conjunctions as a single markable (see ex (2)). We are just as often interested in the order of possessor and possessum in such expressions, and if we need the full expression, it is easy to recover this from the dependency tree.

- (1) a. [Our] [house] is on fire
b. * [[Our] house] is on fire
- (2) a. [Tom], [Dick] and [Harry] were there.
b. * [[Tom], [Dick] and [Harry]] were there.

4.2.2 Zero anaphora

Many corpora in *CorefUD* use *zero tokens* to represent dropped or omitted arguments of verbs, or

⁷In this paper we use * to indicate that we do not identify mentions this way; not that the text itself is ungrammatical.

in some case of nouns. For example, the Ancora corpus for Spanish (Taulé et al., 2008) annotates the referent of indexed subjects of inflected verbs; the SzegedKoref corpus for Hungarian annotates indexed subjects and objects of verbs and possessors of nouns (Vincze et al., 2018); and the Turkish ITCC corpus also annotates indexed subjects and possessors, potentially leading to multiple mention spans to be annotated on the same token (Pamay Arslan and Eryiğit, 2025).

The languages to which this is applied are typically those with extensive pro-drop, and particularly those where arguments and possessors are indexed with morphology on the verb or noun. The motivation behind this is that in such languages, indexed arguments constitute the majority of anaphoric expressions. In terms of information status, the topicalisation of arguments is also a factor in whether an overt pronoun is used or not (Givón, 1983).

To make an annotation scheme *universal*, we believe that this needs to be accommodated for all languages. Our corpus includes, on one end of the spectrum, languages such as English, which allows only minimal and restricted pro-drop; and on the other end, languages such as Turkish and Portuguese, which employ extensive and free pro-drop. In both cases, we allow for the annotation of dropped arguments as zero tokens with the following conditions:

1. The expression’s syntactic role supports a category relevant to the argument.
2. The argument is indexed through morphology on the expression, however minimally.
3. The argument is not overtly mentioned in the same or a head clause.

Keeping to these rules allows us to apply zero tokens to any language while maintaining like-for-like comparisons, and is less burdensome for annotators.

4.3 Coreference relations

The basic coreference relation in our corpus is identity. This is a symmetric relation that implies that the entity referred to by mention A is one and the same as the one that is referred to by mention B. In the output, identity coreference is represented by two mentions sharing the same entity ID: in other words, a cluster representation. For two entities to be identity coreferential, they must share the same underlying entity. An anaphoric mention, for example, will have the same entity ID as its antecedent.

This may also apply to both mentions in a predicative statement. For example, in ex (3), all three mentions are identity coreferent. Likewise, two appositional mentions may also be identity coreferent, as in ex (4)

(3) This is John Snow_i, he_i’s King in the North_i.

(4) Narcissus_i, a youth_i who knelt daily...

Split antecedence is also represented in our annotation scheme. Unlike identity coreference, this is an asymmetric relation that signifies that the entity referred to by mention A is a superset to one or more antecedent entities. For example These entities may be in conjunction or free configuration (Yu et al., 2020). In the output, split antecedence is represented using the SplitAnte feature in the CorefUD format. An example of this can be seen in Fig 1.

4.4 Attributes

Key attributes relating to information status, mention type and other important linguistic phenomena are carried in attributes annotated onto mention spans. These are represented in key-value pairs, and these are listed in Table 1.

Our motivating principle for the attributes is modularity. While each of the attributes is quite simple, reducing the effort at annotation time, combinations of attributes may build a granular description of the mention’s characteristics, while avoiding combinatorial explosions of discrete features. Modularity also allows flexibility in annotation, giving greater freedom to annotate unusual mentions.

Attribute	Values	Required
InfStat	new, given	true
CorefType	ana, cata, pred, disc, appos, coref	true
Indexing	NullSubj, NullObj, NullPoss	false
Bridging	<i>boolean</i>	false
Deixis	<i>boolean</i>	false

Table 1: The set of attributes that we can apply to a mention.

4.4.1 Information Status

We use only two values for information status: *new* and *given*. Unlike some other schemes (e.g. GUM), we do not include *accessible* – i.e. a mention of

an entity that is considered given simply due to cultural or environmental context, such as *God* or *the sky* – as a tag. The reason for this is that it is difficult to fully define cross-lingually what can be considered accessible, due to the different cultural contexts of each book.

We apply information status to *mentions*, not to entities themselves. The *new* value applies to the first mention of an entity, but it may also apply to another mention that substantially expands the known information about that entity. In ex (5), that the referent is named Jon Snow and that he is King in the North is *new* information, even if the entity is already introduced. We consider this more reflective of human packaging and processing of information, recognising that a speaker might employ information status-related strategies to convey this new information.

(5) [This]_(new,cata) is [Jon Snow]_(new,pred).
 [He]_(given,ana)'s [King in the
 North]_(new,pred).

4.4.2 Coreference type

CorefType is the attribute that we use to classify the type of mention: for example, an anaphoric mention such as a pronoun; a predicate mention, such as in an *is* statement (e.g. ex (3)); or a general *coref* mention, for all kinds of open class referring expressions. We inherit the coreference types used in *GUM* for mentions which are coreferent with another mention. These are *ana* (anaphor); *cata* (cataphor); *pred* (predicate); *appos* (apposition); *disc* (discourse deixis); and *coref* (lexical coreference). These are applied to individual mentions, and may also be applied to singletons (sole mentions of an entity).

4.4.3 Deixis

To investigate the effect of deixis in conjunction with givenness, we introduce the attribute *Deixis*. This applies to any deictic mention of an entity; one where the reference can only be fully understood from the spatiotemporal perspective of the utterer. This attribute applies to:

- Any anaphoric first- or second-person reference. These are also considered *given* from the utterer's perspective, giving all such mentions the combination (*given, ana, deixis*).
- Spatial demonstratives, such as *here*, *over there*, and nouns with spatial determiners such as *that guy*, *this place* if relying on the loca-

tion of the utterer. Such references may be either new or given, depending on the context.

- Temporal adverbs or noun phrases, such as *now*, *yesterday*, *next year*.

4.4.4 Bridging

Bridging refers to a relation between two entities in a discourse where a target entity is not strictly the same as its antecedent, but bears a strong semantic link and is inferrable (Clark, 1977). In ex (6), *the trees* is inferrable as part of *the woods*, and therefore does not need introducing in the same way that other entities might.

(6) Lost in **the woods**, **the trees** devour me.

In *CorefUD* corpora that include it, bridging is a relation between two entities, requiring a link from one mention to another. It is represented, like split antecedence, by a pointer from the entity ID of the mention to the entity ID of the antecedent.

Though we are interested in bridging, since it affects the manner in which entity mentions are introduced, for the sake of simplicity we represent bridging as a boolean attribute which applies only to the target mention. The antecedent, from which the target is inferrable, is not annotated. This choice is motivated by the need for rapid annotation and simplicity among a mixed team of annotators. Finding the antecedent of a bridging mention is often difficult, and indeed the antecedent may not be a nominal at all, but may only appear at a phrasal or even discourse level. The representation of bridging is thus contained within the mention span, rather than in the *misc* column. Table 2 shows an example of our bridging annotation.

Lost	
in	
the	Entity=(e1-object-2-InfStat:new
woods	Entity=e1)
,	
the	Entity=(e2-object-2-InfStat:new,Bridging:True
trees	Entity=e2)
devour	
me	

Table 2: An example of bridging annotation in our current scheme. Bridging is represented as a boolean value, without pointing to the antecedent; and is annotated within the mention span, rather than in the *misc* column.

4.4.5 Indexing

As explained in Section 4.2.2, in many languages anaphoric subject, agent, patient or nominal possessor arguments are indexed through morphology

on the syntactic head phrase, with the option of omitting a (pro)nominal mention.⁸

These indexed mentions are included as zero tokens, and we use the attribute *Indexing* to identify the argument that they index. The three basic types are inherited from the CorefUD scheme:

1. NullSubj: An indexed subject
2. NullObj: An indexed object
3. NullPoss: An indexed possessor

5 Annotation Procedure

We performed our annotation using Brat (Stenetorp et al., 2012)⁹, hosted on a webserver. Brat was chosen primarily for its ease of use and customisation of the configuration.

The annotators are each native speakers of the language that they annotate. All annotators begin by annotating sentences in English to practice, with a native English-speaker reviewing, before moving on to their own languages. Practice annotation in English is done collaboratively and different annotators’ decisions are compared. Once annotators are confident of their understanding of the guidelines, they move on to annotation in their own languages. Again, we keep an open forum for discussion of linguistic issues that arise in new languages, and policies evolve based on new linguistic scenarios encountered.

Annotating the full text of each book in one document would be impossible due to the limitations of Brat (and other coreference annotation software): the sheer amount of text and arrows between elements would overwhelm the GUI. For this reason, we chunk each book in each language into chunks of 10 sentences each, and perform annotation on each chunk. Information status is carried over between chunks, so that an entity that has been seen in a previous chunk of the same book will be considered *given* in its next appearance. Attachment of coreference chains between chunks, however, is a task that will need to be completed later.

6 Progress

At the time of writing, our progress in number of sentences annotated is as shown in Table 3. We have completed scripts to serialise from Brat annotation to CorefUD format, and so can output

⁸See features GB089-GB094 in Grambank (Skirgård et al., 2023) for descriptions and a list of languages with these features.

⁹<https://brat.nlplab.org/>

this data in the appropriate format to be used in CorefUD scripts.

Language	Sentences annotated (approx.)
English	3130
Portuguese	2320
Greek	900
Ukrainian	750
Indonesian	190
Hindi	270
Turkish	130

Table 3: Approximate number of sentences annotated per language covered so far, as of November 2024.

7 Future Plans

Now that we have a large amount of annotated data in several languages, we are closer to being able to train multilingual models to predictively annotate data and speed up our annotation process (Pražák and Konopik, 2022), as well as to evaluate the consistency and intrinsic strengths of our annotation (Chai and Strube, 2023). We may also apply techniques such as annotation projection to speed up pre-annotation.

A major shortcoming of our work so far is that we have not instituted quantitative quality control measures such as inter-annotator agreement. One reason for this is that we have only one annotator for each language other than English; while for English we prioritised annotating as much data as possible. Another is that the annotation platform, Brat, does not easily facilitate such measures or annotation of the same data by multiple annotators. In the long term we would like to move to another tool such as INCEPTION¹⁰ (Klie et al., 2018), which would facilitate this when we are able to recruit more annotators.

A shortcoming of Brat is that annotators are unable to see the underlying syntax trees when annotating. Since our goal is to be able to analyse syntax and discourse annotation together, it would be beneficial to ensure that, for example, annotation spans do not cross subtree boundaries (Popel et al., 2021).

Finally, the shortness of our chunks is a problem for studying long range coreferences, and an important next step is to concatenate chunks to form full

¹⁰<https://inception-project.github.io/>

documents for single works and to link coreference chains referring to the same entities.

8 Conclusion

We have presented our annotation scheme, design, and ongoing work on a multilingual corpus that will enable large scale corpus-based analyses of the interplay of information status and word order in a cross-section of the world's languages. Our corpus is now at the stage where we can experiment with model training and evaluation, with sentences annotated in seven languages so far, and annotation guidelines continuously evolving to meet the demands of new languages. We look forward to our first release and to the first applications of the data to answer questions regarding the intersection of information status, information theory, and word order variability.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

Special thanks to Martin Popel, who provided an improved script to convert the Brat annotation to a form compliant with CorefUD.

Ethics

Data availability

Our data contains annotations of works which are protected under copyright. As a result of this we cannot make our corpus open-source and open-access. However, the copyright law of our country allows us to share portions of copyrighted works with researchers for non-commercial purposes, and we are happy to do this on request per the conditions explained in [Verkerk and Talamo \(2024\)](#).

Annotators

Annotation was performed variously by members of the research group working on this project, collaborators in other departments, and currently enrolled students at our institution who were employed under a student assistant contract.

Student assistants were recruited via a call for assistants circulated by email within our institution. Shortlisted candidates were interviewed, and from these candidates annotators were selected based on the interview, their linguistic experience, and their language skills. Student assistant annotators were paid above the minimum wage of our country,

and working time was flexible and limited to be compatible with the demands of full-time study.

All annotators, regardless of status, played an important role in the project and were treated with respect and kindness. All were offered to be named as co-authors and are so named in this paper.

References

- Jennifer E Arnold, Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1):28–55.
- Stefan Baumann and Arndt Riester. 2012. [Referential and lexical givenness: Semantic, prosodic and cognitive aspects](#).
- Christian Beyer. 2015. [Meaning, Context, and Background](#). In Thomas Metzinger and Jennifer M. Windt, editors, *Open MIND, 2-vol. set*. The MIT Press.
- Wallace L Chafe. 1976. [Givenness, contrastiveness, definiteness, subjects, topics, and point of view](#). *Subject and topic*.
- Haixia Chai and Michael Strube. 2023. Investigating multilingual coreference resolution by universal annotations. *arXiv preprint arXiv:2310.17734*.
- Herbert H Clark. 1977. [Bridging](#). In *Thinking: Readings in Cognitive Science*, pages 411–420. Cambridge University Press, Cambridge.
- Stefanie Dipper, Michael Goetze, and Stavros Skopeteas, editors. 2007. *Information structure in cross-linguistic corpora : annotation guidelines for phonology, morphology, syntax, semantics and information structure*. Universitätsverlag Potsdam.
- Stefanie Dipper and Heike Zinsmeister. 2009. [Annotating discourse anaphora](#). In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 166–169, Suntec, Singapore. Association for Computational Linguistics.
- Gertraud Fenk-Oczlon. 2001. [Familiarity, information flow, and linguistic form](#). In *Frequency and the Emergence of Linguistic Structure*. John Benjamins.
- Talmy Givón. 1983. *Topic Continuity in Discourse: A Quantitative Cross-language Study*. John Benjamins, Amsterdam; Philadelphia. 2010.
- Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. [Cognitive status and the form of referring expressions in discourse](#). *Language*, 69(2):274–307.
- Sooyoun Han, Sumin Seo, Minji Kang, Jongin Kim, Nayoung Choi, Min Song, and Jinho D. Choi. 2021. [FantasyCoref: Coreference Resolution on Fantasy Literature Through Omniscient Writer's Point of View](#). In *Proceedings of the Fourth Workshop on*

- Computational Models of Reference, Anaphora and Coreference*, pages 24–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Martin Haspelmath. 2008. [A frequentist explanation of some universals of reflexive marking](#). *Linguistic Discovery*, v.6, 40-63 (2008), 6.
- Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. [OntoNotes: The 90% solution](#). In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.
- Lauri Karttunen. 1974. [Presupposition and linguistic content](#). *Theoretical Linguistics*, 1(1-3):181–194. Publisher: De Gruyter Mouton Section: Theoretical Linguistics.
- Katalin É Kiss, editor. 1995. *Discourse configurational languages*. Oxford studies in comparative syntax. Oxford University Press, New York.
- 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. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).
- Anke Lüdeling, Julia Ritz, Manfred Stede, and Amir Zeldes. 2014. *Corpus linguistics and information structure research*. Oxford, Oxford University Press.
- Katja Markert, Yufang Hou, and Michael Strube. 2012. [Collective classification for fine-grained information status](#). In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 795–804, Jeju Island, Korea. Association for Computational Linguistics.
- Dejan Matic and Daniel Wedgwood. 2013. [The meanings of focus: The significance of an interpretation-based category in cross-linguistic analysis](#). *Journal of Linguistics*, 49(1):127–163.
- Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. [CorefUD 1.0: Coreference meets Universal Dependencies](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. [Universal Dependencies v2: An evergrowing multilingual treebank collection](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.
- Maciej Ogrodniczuk, Vincent Ng, Sameer Pradhan, and Massimo Poesio, editors. 2023. *Proceedings of The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)*. Association for Computational Linguistics, Singapore.
- Tuğba Pamay Arslan and Gülşen Eryiğit. 2025. [Enhancing Turkish Coreference Resolution: Insights from deep learning, dropped pronouns, and multilingual transfer learning](#). *Computer Speech & Language*, 89:101681.
- Massimo Poesio, Maciej Ogrodniczuk, Vincent Ng, Sameer Pradhan, Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, Amir Zeldes, Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, and Daniel Zeman. 2024. [Universal anaphora: The first three years](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17087–17100, Torino, Italia. ELRA and ICCL.
- Martin Popel, Zdeněk Žabokrtský, Anna Nedoluzhko, Michal Novák, and Daniel Zeman. 2021. [Do UD Trees Match Mention Spans in Coreference Annotations?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3570–3576, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ondřej Pražák and Miloslav Konopik. 2022. [End-to-end multilingual coreference resolution with mention head prediction](#). In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 23–27, Gyeongju, Republic of Korea. 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*.
- Marta Recasens, Eduard Hovy, and M. Antònia Martí. 2011. [Identity, non-identity, and near-identity: Addressing the complexity of coreference](#). *Lingua*, 121(6):1138–1152.
- Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latache, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowers, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Natalia Hübner, Biu Huntington-Rainey, Jessica K.

- Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L.M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tônia R.A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O.C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W.P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. [Grambank reveals global patterns in the structural diversity of the world’s languages](#). *Science Advances*, 9.
- Stavros Skopeteas and Gisbert Fanselow. 2010. Focus types and argument asymmetries: A cross-linguistic study in language production. *Contrastive information structure*, pages 169–197.
- Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. 2012. [brat: a web-based tool for NLP-assisted text annotation](#). In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.
- Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. [AnCora: Multilevel annotated corpora for Catalan and Spanish](#). In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco. European Language Resources Association (ELRA).
- Annemarie Verkerk and Luigi Talamo. 2024. [mini-CIEP+ : A shareable parallel corpus of prose](#). In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*, pages 135–143, Torino, Italia. ELRA and ICCL.
- Reiko Vermeulen. 2012. [The information structure of Japanese](#), pages 187–216. De Gruyter Mouton, Berlin, Boston.
- Veronika Vincze, Klára Hegedús, Alex Sliz-Nagy, and Richárd Farkas. 2018. [SzegedKoref: A Hungarian coreference corpus](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Lin Wang, Marcel Bastiaansen, Yufang Yang, and Peter Hagoort. 2012. Information structure influences depth of syntactic processing: Event-related potential evidence for the chomsky illusion. *PLoS One*, 7(10):e47917.
- Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, and Massimo Poesio. 2020. [Free the plural: Unrestricted split-antecedent anaphora resolution](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6113–6125, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Amir Zeldes. 2017. [The GUM corpus: Creating multilayer resources in the classroom](#). *Language Resources and Evaluation*, 51(3):581–612.

Dependency Structure of Coordination in Head-final Languages: a Dependency-Length-Minimization-Based Study

Wojciech Stempniak
 University of Warsaw
 and Saarland University
 w.stempniak@student.uw.edu.pl

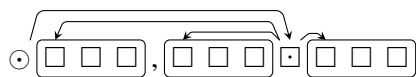
Abstract

There is no single accepted model of the dependency structure of coordination. Universal Dependencies (UD, De Marneffe et al. 2021) enforces in its corpora an asymmetrical model privileging the coordination’s first conjunct as a standard. Kanayama et al. (2018) criticize that approach stating that this model is incompatible with the grammatical structure of head-final languages. Recent research (Przepiórkowski and Woźniak 2023, Przepiórkowski et al. 2024a) provides a DLM-based argument for the symmetrical models of the dependency structure of English coordination. This paper shows the result of the analysis of coordinations found in UD corpora of two head-final languages, namely Korean and Turkish. Based on the analysis of coordinations and theoretical arguments, an alternative approach to the dependency structure of coordination in head-final languages is suggested.

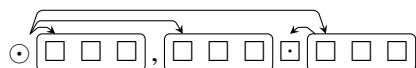
1 Introduction

There is no single universally accepted approach to the dependency structure of coordination. Przepiórkowski and Woźniak (2023) (henceforth PW23) enumerate four main models¹:

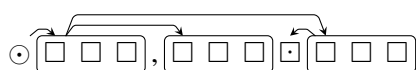
(1) a. **Conjunction-headed/Prague**



b. **Multi-headed/London**

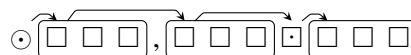


c. **Bouquet/Stanford**



¹The following diagrams are based on those in the work of PW23. The governor is marked by ⊙, the conjunction by □, and other tokens by □. Tokens belonging to the same conjunct are grouped. The names of the approaches in (1a)–(1d) are based on those in PW23. Apart from the approach shown in (1b), they were originally named by Popel et al. (2013).

d. **Chain/Moscow**

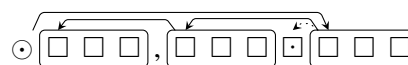


PW23 show that the asymmetrical approaches (1c)–(1d) cannot describe the English coordination structure correctly. Their argument is based on Dependency Length Minimization (DLM) – an universal and well-documented tendency to order words in sentences in a way so that long dependencies are avoided. (Temperley 2007, Futrell et al. 2015).

PW23’s findings are replicated by subsequential studies including Przepiórkowski et al. (2024a) (from now on PBG24). The latter indicate that the London approach is probably the best description of the English coordination structure.

However, these conclusions cannot be extended for head-final languages such as Korean or Turkish. Kanayama et al. (2018) suggest that the coordination in head-final languages (HFL) may be asymmetrical. They propose a different approach taken from the work of Choi and Palmer (2011)²:

(2) **Right-headed/Inverted Moscow**



This paper aims to show that the approach shown in (2) might be the only one describing dependency coordination structure in HFL correctly. Using the methodology of PW23 and assuming the Dependency Length Minimization, it is demonstrated that using this approach the change in the tendency to put shorter conjunct at the beginning of coordination can be predicted more accurately than when using the other approaches.

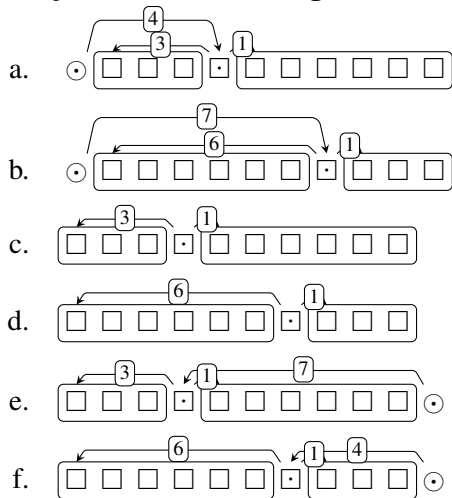
²This approach assumes the head of the right conjunct to be the technical head of coordination and that each token is a dependent of the subsequent conjunct head. Those assumptions are inverted with respect to the Moscow approach. Note that Choi and Palmer (2011) do not specify which token is the governor of the conjunction. For reasons explained in §6.1 it is assumed that in this approach the conjunction is the dependent of the head of its closest conjunct.

2 Previous Work

PW23 examine the tendency to put the shorter conjunct of the coordination as the first. They show that, assuming DLM, each approach to the dependency structure of coordination can predict the change in this tendency as the absolute difference of the conjunct length grows. Their study takes into account only binary coordinations.

To summarize their method, let me present the predictions of one dependency structure of coordination model, namely the Prague approach. They compare the total dependency length in six cases:³

(3) Conjunction-headed/Prague



PW23 compare the total dependency length in cases with the same governor position, i.e. (3a) vs (3b), (3c) vs (3d) and (3e) vs (3f). E.g. the total length of dependencies in (3a) is $4 + 3 + 1 = 8$ tokens, and in (3b) it is $7 + 6 + 1 = 14$ tokens, so the absolute length difference is equal to $|8 - 14| = 6$ tokens. In scheme (3) the difference of the conjunct length is $|3 - 6| = 3$ tokens. This means that the Prague approach assumes that when the governor is on the left, the total dependency length is smaller when the shorter conjunct is on the left (3a) than when it is on the right (3b). With the growth of the conjunct length difference, the total dependency length difference also grows.

PW23 point out that out of each pair, the arrangement with the smallest total dependency length is the more probable the greater the difference between the conjuncts' length is⁴. Therefore, because

³The governor can be in one of the three positions (left, absent, right) and the shorter conjunct can be either the first (left) or the last (right). Technically, coordinations with the governor in the middle (between conjuncts) are possible but they are too uncommon to be analyzed. Edge labels show the length of the dependency measured in tokens.

⁴Note that the DLM is not the only factor taken onto account while ordering the conjuncts. There is a general tendency

of DLM, the greater the difference is, the more coordinations are expected to have the shorter conjunct on the left. This can be demonstrated as a change in a function $p_*(n)$, where $n > 0$ is the absolute difference between the conjunct lengths and $* \in \{L, -, R\}$ is the governor position.⁵ The function value is the proportion of the coordinations with the shorter conjunct on the left to all coordinations with a given governor position.

PW23 show that each approach can predict the direction of $p_*(n)$ function slope by comparing the total dependency length in pairs. Moreover, they determine the true values of the proportions function by analyzing 21.8K English coordinations in PennTree Bank. Table 1 summarizes the predictions of the direction of the $p_*(n)$ tendencies in English. The predictions are compared with the actual tendencies found by PW23 and PBG24.

	L	$-$	R
Prague	+	+	0
London	+	0	-
Stanford	+	+	+
Moscow	+	+	+
PW23	+	+	0
PBG24	+	0	-

Table 1: Predictions of the change of the $p_*(n)$ tendencies in English and the tendencies observed in previous works.

PW23 argue that only the symmetric approaches (namely the Prague and London models) predict the changes in the proportions correctly. While the predictions of the Prague approach match the observed tendencies, there is a difference between the actual changes in the proportions and the predictions of the London approach. They state that this difference can be explained by the DLM effect at grammar. PW23 point out that coordinations with the left governor are most frequent in English and the $p_L(n)$ is positive. Therefore, a tendency to put the shorter conjunct as the first has become a gen-

to put the shorter conjunct as the first one, which has multiple explanations (Lohmann, 2014). However, PW23 explain that the DLM is the only factor that depends on the governor position and the conjunct length at the same time. They assume that in the analysis of thousands of coordinations the influence of the remaining factors even out.

⁵The possible governor positions shall be understood as follow: L means the governor is on the left, R denotes a governor on the right and $-$ stands for a coordination with no governor. The $p_*(n)$ function is not defined explicitly in PW23 and it is taken from the work of Przepiórkowski et al. (2024b).

eral, grammatical rule in English. With the growth of the difference between the conjuncts length, the tendency is stronger. Hence, the observed $p_{-}(n)$ and $p_{R}(n)$ tendencies are distorted. This means that the important thing to compare is not the actual and predicted $p_{*}(n)$ tendencies, but rather the actual and predicted differences between various $p_{*}(n)$ tendencies.

PBG24 replicate PW23’s study, analyzing the larger (11.5M coordinations) Corpus of Contemporary American English (COCA). Tendencies observed by them match the predictions of the London model without the need to refer to DLM-at-grammar. Because of that, they narrow down possible models to the London approach.

PW23 research only covers the matter of the structure of coordination in English, which is a head-initial language. Kanayama et al. (2018) claim that the dependency structure of coordination in head-final languages can be different. They point out that the development of the models in (1) was based on the research, arguments and intuitions regarding only head-final languages. They especially criticize the asymmetrical Stanford approach used in Universal Dependencies as incompatible with the head-final languages’ conditions.

Kanayama et al. (2018) claim that forcing Japanese and Korean UD annotators to use the Stanford approach resulted in lowering the quality of their corpora. They show linguistic and empirical arguments towards an alternative approach proposed by Choi and Palmer (2011) and urge UD to allow using that model in HFL corpora.

This paper argues that allowing the Right-headed approach to the dependency structure of coordination in UD corpora of head-final languages would be beneficial. The claim is based on the results of the analysis of Turkish and Korean UD corpora using PW23 and PBG24’s methodology and theoretical arguments taken from the work of Kanayama et al. (2018).

3 Data

Three Korean (Kaist, GSD and PUD) and nine Turkish (Kenet, Penn, Tourism, Atis, GB, FrameNet, BOUN, IMST and PUD) corpora have been analyzed. All corpora have been annotated in Universal Dependencies v. 2.13 and downloaded from UD’s website (<https://universaldependencies.org/> in December 2023). The data has been annotated manually. Four

Turkish corpora (Atis, GB, BOUN and PUD) have been annotated natively in UD style, others have been automatically converted from different style. In total 21.5K Korean and 19.6K Turkish coordinations have been analyzed.

Table 2 shows the number of coordinations with a specific position of the governor and the shorter conjunct.⁶

shorter conjunct	governor position		
	left	absent	right
Korean			
left	294	4054	3999
right	89	1093	964
Turkish			
left	894	3263	4257
right	111	1052	880

Table 2: The number of coordinations with a specific position of the governor and the shorter conjunct in the HFL corpora.

4 Methods

In order to replicate the methodology of PW23, only binary coordinations should be taken into account. However, ignoring every coordination with more than two conjuncts could severely impact the result of the study. On the other hand, to analyze the impact of the DLM effect on the length of every conjunct of the coordination a new methodology would be needed. Therefore, in this study every coordination is treated as binary, i.e. no matter how many conjuncts are in it, the first conjunct is considered the left one and the last conjunct is considered to be the right one. If a coordination have more than two conjuncts, the middle ones are ignored.

To determine the slope of all $p_{*}(n)$ functions as well as the differences between them the coordinations are extracted and the lengths of their conjuncts are measured.

The process of delimiting the conjunct basing on dependency trees is non-trivial and cannot be automated with high accuracy⁷ (Patejuk and

⁶Since the DLM effect can only be noticed when there is a difference between the length of the conjuncts, the coordinations in which both conjuncts have the same length are not taken into account in this analysis. Also, because of the small number of the coordinations with the governor in the middle, those are also ignored.

⁷E.g. in a phrase such as *long days and nights* it is not syntactically determined if the word *long* describes only the word *days*, or both *days* and *nights*. The conjuncts can be delimited as either *[[long days] and [nights]]* or *[long [[days*

Przepiórkowski 2018, Przepiórkowski and Woźniak 2023). Therefore a heuristic-based algorithm has been used to extract coordinations. It is an HLF-adjusted version of the algorithm used in PBG24’s analysis. It is depicted in the Appendix A.

Since the automated process is not fully reliable, it has been evaluated by a native speaker of the Turkish language. 60 coordinations have been sampled randomly and evaluated using two criteria: 1. the governor position has been determined correctly and 2. both conjuncts have been delimited exactly as they should be (putting aside the punctuation, as it does not affect the word count). 35 of Turkish coordinations have been extracted correctly, which resulted in overall 58% algorithm accuracy. The algorithm has not been evaluated in the Korean language analysis.

The dependency length can be measured in various ways (e.g. characters, syllables, words). The DLM effect is mostly connected with the number of new objects in the discourse. Since objects correspond to words, for the DLM analysis dependency length is measured in words understood as non-punctuation tokens (Futrell et al., 2020).

Once the conjunct lengths and governor positions are determined, the monofactorial logistical regression is performed⁸ to calculate the slope of each of the $p_*(n)$ functions.

The code used for extracting coordinations and statistical analysis is publicly available in the repository at <https://github.com/wjstempniak/Dependency-Structure-of-Coordination>.

5 Results

Figure 1 depicts changes in the tendency to put the shorter conjunct at the beginning of the coordination with the growth of the absolute difference between the conjuncts length.

Table 3 shows the differences between the slopes of $p_L(n)$, $p_-(n)$ and $p_R(n)$ the tendencies computed using R’s `emmeans::emtrends` function (Lenth, 2024). The $L/-$, $-/R$ and L/R columns

and `[nights]]]`.

⁸In PW23, “due to the low number of coordinations with large length differences when the governor is on the right, observations were collected into five buckets defined by the vector $\delta = \langle 0, 1, 2, 3, 6, 25 \rangle$ ”, where “bucket i contains coordinations with absolute length differences within the interval $(i, i + 1]$ ”. In the HFL analysis, due to a low number of coordinations with the governor on the left, similar method was applied. To fit the data, the values of the δ vector have been adjusted to $\langle 0, 1, 2, 3, 6, 18 \rangle$. The computations were performed using the R’s `glm` function (R Core Team, 2023).

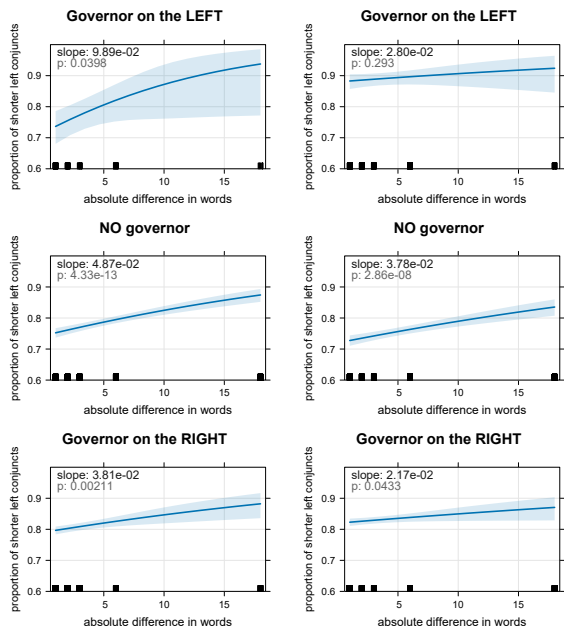


Figure 1: Observed $p_*(n)$ in Korean (left) and Turkish (right).

	$L/-$		$-/R$		L/R	
	diff	p	diff	p	diff	p
ko	0.05	0.30	0.01	0.45	0.06	0.22
tr	-0.01	0.72	0.02	0.21	0.01	0.83

Table 3: Differences between the steepness of observed $p_*(n)$ tendencies in HFL.

show the difference of the steepness of two respective $p_*(n)$ tendencies, e.g. if $L/-$ is positive, the $p_L(n)$ tendency is more increasing than the $p_*(n)$ tendency. The difference of the steepness of the $p_*(n)$ and $p_{\dagger}(n)$ tendencies is henceforth referred as the $*/\dagger$ contrast.

Although it may seem that for both languages almost all contrasts are positive,⁹ it is crucial to notice that the differences between the tendencies are highly insignificant. For all three contrasts, both in Korean and Turkish, p was greater than 0.2. In Turkish, for the L/R contrast (which had been expected to be the greatest and therefore most significant) p was equal to 0.83.

The insignificance of differences can either mean that there is no difference or may indicate the lack of sufficient quality and quantity of the data. How-

⁹Note that the only negative tendencies are the L and $L/-$ tendencies in Turkish. However, those are the tendencies concerning Turkish coordinations with the governor on the left. Because Turkish is a HFL, this type of coordination structure is rare in its corpora – in this study, there are only 1.8k Turkish coordinations with the governor on the left (opposed to 12k coordinations with the governor on the right).

ever, the total number of coordinations analyzed in Korean and Turkish was similar to the number of English coordinations analyzed by PW23 (Korean: 21.5K, Turkish: 19.6K vs. PW23: 21.8K). This suggests that if such differences exist, examining around 20k coordinations should be enough to find them. The fact that in this analysis no significant differences were found does not *prove* that there are no differences, but certainly indicates that this is probable. Given that, since most of the (insignificant) observed tendencies are positive, it is more probable that those differences are positive or neutral than negative.

The next section explains how these results can be interpreted in the context of the approaches to the dependency structure of coordination shown in (1) and (2).

6 Discussion

6.1 Dependency structure of coordination in head-final languages

For the analysis of binary coordinations in HFL several assumptions have to be made.

Firstly, in case of binary coordinations the Stanford and Moscow approaches are essentially the same.¹⁰ The predictions of these approaches are the same regardless of the position of the governor, so for the analysis' sake one of them can be omitted.

Additionally, it is known that in HFL heads tend to be at the end of phrases. For the sake of the analysis, the head is assumed to be the final token of a coordination conjunct.¹¹

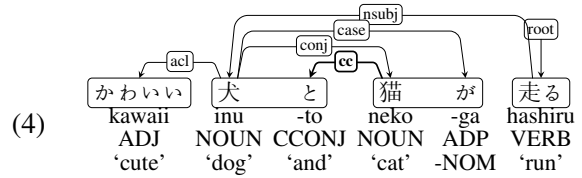
Finally, it is safe to assume the conjunction is always dependent to the conjunct head next to it. This is a conclusion from the fact that in HFL the conjunction is often an agglutinate, suffix, or part of the word or phrase unit¹² that is the head of the conjunct. This is visible in Japanese and Korean examples below.¹³

¹⁰The only difference between these models is the two relations between the head of the right conjunct and the conjunction. The sum of these relations is the same in all six cases so the influence of this relations on the total dependency length is none.

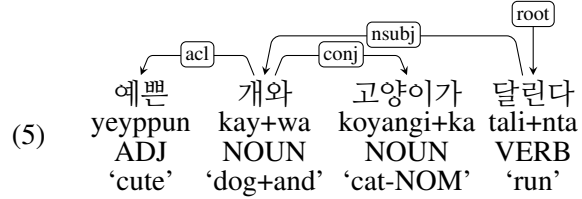
¹¹It is a simplification, because the head is not always the final token. However, the relevant factor is that there are more potential dependents on the left than on the right side of the head. Appendix B shows the information about the relative position of the head within the conjuncts in the used data.

¹²Such as Japanese *bunsetsu* or Korean *eojeol*. See Kanayama et al. (2018) for details and examples.

¹³The example sentences are adopted from the work of Kanayama et al. (2018) and are annotated according to the



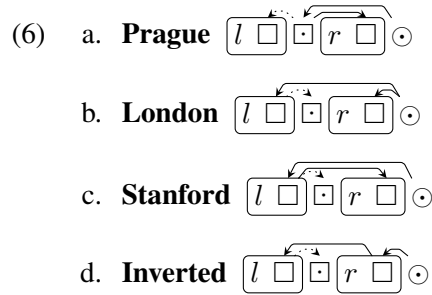
In (4) the conjunction token ‘と’ (‘and’) is a part of *bunsetsu* ‘犬と’ (‘dog and’). According to the Stanford approach ‘と’ (‘and’) should be treated as a dependent of ‘猫’ (‘cat’). Assuming common sense and basic semantic intuition, there is no reason to do that (Kanayama et al., 2018).



In (5), the conjunctive particle ‘와’ is a suffix in *eojeol* ‘개와’ (‘dog and’), therefore it does not constitute an individual token and is not a dependent to any conjunct head.

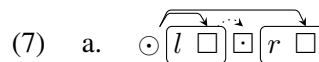
For these reasons assume that in HFL if the conjunction is a separate token, it is connected to the right head. Thus, the dependency length between the conjunction and the head is constant and its influence on the total dependency length is negligible. Therefore, that dependency is ignored.

Taking into account the assumptions stated above, the following approaches are analyzed:¹⁴



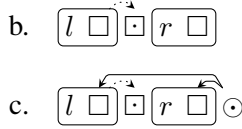
6.2 Predictions of different approaches

In order to determine the predictions, the pairs of cases with the three different governor positions are compared (now using the London approach as the example):

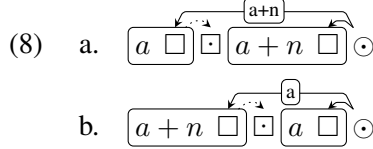


Stanford approach.

¹⁴The non-head tokens which are parts of the conjuncts (henceforth called the conjunct *body*) are replaced with *l* or *r* symbol for clarity. In these schemata, the governor is placed on the right side of the coordination to reflect the fact in HFL it strongly tends to be in that position. The analysis still covers coordinations with all three governor positions, so that does not interfere with the models' predictions.



For each possible governor position, the difference of the total dependency length is compared between the case where the first conjunct is shorter and the case where it is longer. Let a and $a + n$ be the length of the conjuncts, where $n > 0$. For the case shown in (14a), those cases are:



In (8a) the total dependency length is equal to $a+n$ and in (8b) it is equal to a . Therefore, the absolute difference is equal to n . The total dependency length in (8a) is greater than the total dependency length in (8b). Thus the prediction of the London approach for HFL is that when the governor is on the right, the proportion of coordinations with the shorter conjunct on the left is decreasing with the growth of the absolute difference of the conjunct length. In other words, the London approach predicts that $p_R(n)$ is decreasing.

To shorten the calculations let me formalize them. Let S_* be the sum of the *relevant*¹⁵ dependencies' length in the case where the first conjunct is shorter, and S'_* be the sum of the *relevant* dependencies' length in the other case. The model predicts that the function $p_*(n)$ is increasing if and only if S'_* is greater than S_* . Finally, let $e_*(n)$ be a function such that

$$(9) \quad e_*(n) = S'_* - S_*$$

The function $e_*(n)$ estimates the direction of the slope of $p_*(n)$ in the way that the sign of $e_*(n)$ is equal to the direction of the slope of $p_*(n)$ for $* \in \{L, -, R\}$.

Let \bar{l}_* (\bar{r}_*) be the number of dependencies that go over the left (right) conjunct's body. In (8a) there are 0 dependencies going over the left conjunct body and there is 1 dependency going over the right conjunct's body. Since the conjunction is ignored, to compute the total dependency length the product of the number of dependencies and their length, which is equal to the conjunct length, is simply added. The total dependency length in (8a) is $0a + 1(a+n) = a+n$.

This can be generalized as

¹⁵The dependencies within the conjunct are ignored as they are constant and independent from changes in the conjunct order and the governor position.

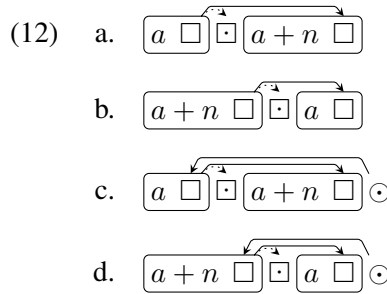
$$(10) \quad \begin{aligned} \text{a. } S_* &= \bar{l}_*a + \bar{r}_*(a+n) \text{ and} \\ \text{b. } S'_* &= \bar{l}_*(a+n) + \bar{r}_*a. \end{aligned}$$

Recall from (9) that $e_*(n) = S'_* - S_*$. Thus

$$(11) \quad \begin{aligned} \text{a. } e_*(n) &= \bar{l}_*(a+n) + \bar{r}_*a - (\bar{l}_*a + \bar{r}_*(a+n)), \text{ which can be simplified to} \\ \text{b. } e_*(n) &= (\bar{l}_* - \bar{r}_*)((a+n) - a), \text{ so} \\ \text{c. } e_*(n) &= (\bar{l}_* - \bar{r}_*) \cdot n \end{aligned}$$

Because $n > 0$, the estimating function $e_*(n)$ is increasing if and only if $\bar{l}_* - \bar{r}_* > 0$. That means the prediction of the model of the direction of $p_*(n)$ slope is equal to sign of $\bar{l}_* - \bar{r}_*$.

The DLM can be understood as a probability function from total dependency length in a given case to a probability that this case occurs in natural language. Assuming that the function is monotonous, the differences between the slopes for each governor position pair can be predicted. Let me show how to do that using the example of the $-/R$ contrast in the Stanford approach for HFL:



In (12), the $e_-(n) = a - (a+n) = -n$, so the slope of $p_-(n)$ is expected to be negative. However, the $e_R(n) = 2a - 2(a+n) = 2n$, so the slope of $p_R(n)$ is expected to be positive and smaller than the slope of $e_-(n)$.

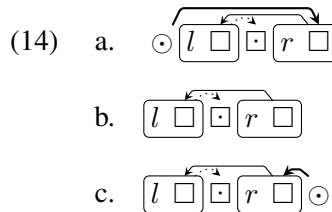
This observation can be generalized using the $e_{*/\dagger}(n)$ function such as

$$(13) \quad \begin{aligned} \text{a. } e_{*/\dagger}(n) &= e_*(n) - e_{\dagger}(n) \text{ or} \\ \text{b. } e_{*/\dagger}(n) &= (\bar{l}_* - \bar{r}_* - (\bar{l}_{\dagger} - \bar{r}_{\dagger})) \cdot n \end{aligned}$$

for $*, \dagger \in \{L, -, R\}$ and $n > 0$.

Hence, the sign of the contrast between $p_*(n)$ and $p_{\dagger}(n)$ functions is equal the sign of $e_{*/\dagger}(n)$.

Consider again the three cases described in (7).¹⁶



¹⁶In the following example, the coordinations are annotated according to the Inverted approach. However, the implications below are true for every model of the dependency structure of coordination.

All dependencies that are present in (14b) are also present in (14a) and (14c). Moreover, the dependencies going over the left conjunct body can be divided into two groups: 1. those present when there is no governor (in (14b) there are \overline{l}_- of them) and 2. those connecting the governor with its dependents (in (14a) and (14c), they are thickened). Let \overline{L}_* (and \overline{R}_*) be the number of relations connecting the governor with its dependents going over the left (right) body conjunct. From the observation above it is visible that

$$(15) \quad \begin{aligned} \text{a. } & \overline{l}_L = \overline{l}_- + \overline{L}_L, \text{ or } \overline{L}_L = \overline{l}_L - \overline{l}_- \\ & \text{and, similarly} \\ \text{b. } & \overline{r}_L = \overline{r}_- + \overline{R}_L, \text{ or } \overline{R}_L = \overline{r}_L - \overline{r}_-. \end{aligned}$$

Since the dependencies present in (14b) are also present in (14a) and (14c), when comparing the difference between different slopes they can be omitted. In other words, while computing the contrast between $p_*(n)$ functions the only *relevant* dependencies are those connecting the governor with its dependents.

Recall from (11c) and (13) that:

$$(16) \quad e_{*/\dagger}(n) = e_*(n) - e_{\dagger}(n)$$

$$(17) \quad e_*(n) = (\overline{l}_* - \overline{r}_*) \cdot n$$

From (11c):

$$(18) \quad \begin{aligned} \text{a. } & e_L(n) = (\overline{l}_L - \overline{r}_L) \cdot n, \\ \text{b. } & e_-(n) = (\overline{l}_- - \overline{r}_-) \cdot n, \\ \text{c. } & e_R(n) = (\overline{l}_R - \overline{r}_R) \cdot n. \end{aligned}$$

From (13):

$$(19) \quad \begin{aligned} \text{a. } & e_{L/-}(n) = e_L(n) - e_-(n), \\ \text{b. } & e_{L/-}(n) = (\overline{l}_L - \overline{r}_L) \cdot n - (\overline{l}_- - \overline{r}_-) \cdot n, \\ \text{c. } & e_{L/-}(n) = (\overline{l}_L - \overline{l}_- - (\overline{r}_L - \overline{r}_-)) \cdot n, \\ \text{d. } & e_{L/-}(n) = (\overline{L}_L - \overline{R}_L) \cdot n. \end{aligned}$$

Similarly, it is provable that

$$(20) \quad \begin{aligned} \text{a. } & e_{-/R}(n) = e_-(n) - e_R(n), \\ \text{b. } & e_{-/R}(n) = (\overline{l}_- - \overline{r}_-) \cdot n - (\overline{l}_R - \overline{r}_R) \cdot n, \\ \text{c. } & e_{-/R}(n) = (\overline{l}_- - \overline{l}_R - (\overline{r}_- - \overline{r}_R)) \cdot n, \\ \text{d. } & e_{-/R}(n) = (\overline{L}_R - \overline{R}_R) \cdot n. \end{aligned}$$

To sum up, the predictions of a given model are signs of functions shown in (18), (19d) and (20d) for $n > 0$. Using these formulae, predictions for HFL can be computed for every model (see Table 4).

Model	$e_L(n)$	$e_-(n)$	$e_R(n)$	$e_{-/R}(n)$	$e_{L/-}(n)$
Prague	0	-n	-2n	-n	-n
London	n	0	-n	-n	-n
Stanford	0	-n	-2n	-n	-n
Inverted	-n	-n	-n	0	0

Table 4: Values of estimating function for HFL.

Recall from Table 1 that all three observed $p_*(n)$ tendencies are positive. This means that none of the considered approaches predict the slope direction itself correctly. This may be due to a strong, universal tendency to put the shorter conjunct at the beginning of the coordination which has a different cause.¹⁷ For this reason it is important to compare the predicted and observed differences between the tendencies steepness (i.e. the contrasts) rather than the predicted and observed $p_*(n)$ tendencies.

As it is visible in Table 3, every approach predicts that the $L/-$ and $-/R$ contrasts are either negative or none. However, the results of the study suggest that the contrast is more likely to be positive. One might say that there can be other approaches to the dependency structure of coordination in HFL which predict the contrasts to be positive. The following section proves that such an approach is impossible.

6.3 All possible approaches

To cover all approaches, let me analyze possible dependents of the governor of the coordination. The governor's dependent can be either to the left of the left conjunct's body (as in 21c), to the right of the right conjunct body (as in 21b), or between the conjuncts' bodies (as in 21a).

$$(21) \quad \begin{aligned} \text{a. } & \odot \overline{l} \square \square \square \overline{r} \square \odot \\ \text{b. } & \odot \overline{l} \square \square \square \overline{r} \square \odot \\ \text{c. } & \odot \square \overline{l} \square \square \square \overline{r} \odot \end{aligned}$$

From (21a)–(21c) it is visible that irrespective of the assumed approach for every dependency connecting the governor on the left going over the left conjunct body either this dependency is also over the right conjunct body (21b) or there is another dependency over the right conjunct body when the governor is on the right (21a).

The same can be said about the dependency connecting the governor on the right going over the right conjunct body and the dependency goes over the left conjunct body when the governor is on the left. This can be written as

$$(22) \quad \begin{aligned} \text{a. } & \overline{L}_L + \overline{R}_L = \overline{L}_R + \overline{R}_R, \text{ or} \\ \text{b. } & \overline{L}_L - \overline{L}_R = \overline{R}_R - \overline{R}_L. \end{aligned}$$

¹⁷This tendency is observed in multiple previous works and explained in a numerous ways, including arguments based on pragmatics (Lohmann, 2014), psycholinguistics (McDonald et al., 1993), stress patterns (Wright et al., 2005) and DLM-at-grammar (PW23).

Given that:

- (23) a. $e_{L/-}(n) = (\overline{L_L} - \overline{L_R}) \cdot n$
 b. $e_{L/-}(n) = (\overline{R_R} - \overline{R_L}) \cdot n$
 c. $e_{L/-}(n) = e_{-/R}(n)$

Furthermore, from (21b) it is visible that if there is a dependency connecting the left governor going over the right conjunct body, this dependency goes also over the left conjunct body. Thus

- (24) a. $\overline{L_L} \geq \overline{L_R}$ or $\overline{L_L} - \overline{L_R} \leq 0$, and since
 b. $e_{L/-}(n) = \overline{L_L} - \overline{L_R}$ and
 c. $e_{L/-}(n) = e_{-/R}(n)$, therefore
 d. $e_{L/-}(n)$ and $e_{-/R}(n)$ are non-increasing functions.

Therefore, irrespective of the assumed approach to the dependency structure of coordination, a model can either predict that the slope of $p_L(n)$ can be either the same or more decreasing than the slope of $p_-(n)$. The same can be said about the $p_-(n)$ and $p_R(n)$ slopes. This is true for all possible models consistent with the assumptions made in this paper. This leads to the conclusion that a model predicting positive contrasts between the slopes is impossible.

It is well known that the DLM affect word order both in individual sentences (Futrell et al., 2015) and at the grammatical level (PW23). However, it is possible that the DLM influences the shape of the grammatical structure itself as well.¹⁸

In case of head-initial languages, a symmetric coordination allows to use DLM efficiently, as putting the short conjunct first in coordinations with the governor on the left indeed shortens the total length of the dependencies (because the $L/-$ and $L/-$ contrasts are negative). However, in case of HFL, there is no possible approach that would allow shortening the total length of the dependen-

¹⁸It is intuitive that a grammatical structure for a simple clause should have a simple dependency structure with as short dependencies as possible. For that reason the Stanford approach declares the head of the left conjunct as the “technical head of the coordination” (De Marneffe et al., 2021) – because in the head-initial languages the governor of coordination tends to be at the left, and the head of the left conjunct tends to be at the beginning of this conjunct, i.e. next to the governor. Therefore, the head of the left conjunct is a dependent to the governor. In case of HFL, it is exactly opposite – the governor tends to be at the end of a coordination, which is next to the head of the right conjunct. Because of that, the Stanford and Inverted approaches can be intuitive for respectively head-initial and head-final languages users and not intuitive for the other language group users. The need to minimize the length of dependencies is also the reason why Choi and Palmer (2011) decided to invert the Moscow approach, and not the Stanford model – because a hypothetical Inverted Stanford model would have to long dependencies between conjunct heads in case of long coordinations.

cies by putting the short conjunct last in coordinations with the right governor (because the $-/R$ and L/R contrasts are also negative or none). Using any of the four main approaches shown in (1) in HFL would cause almost every coordination to have excessively long dependencies.

This may explain why head-final languages may have formed an inverted-approach dependency structure of coordination, opposed to head-initial languages which evolved a symmetric one.

7 Limitations

The main drawback of the presented research is the quality and quantity of used data. As stated in the work of Kanayama et al. (2018), UD-imposed bounds “tied the hands” of HFL corpora annotators and forced them to work out compromises, which reduced the corpora quality. This include “dropping the conjunction category entirely in the case of Japanese” (Kanayama et al., 2018, p. 82), which made an analysis of coordination in Japanese UD corpora impossible. Given these restrictions, the relatively small¹⁹ Korean and Turkish corpora were analyzed. Once more corpora of the head-final languages with coordinations marked consistently will be created, revisiting this study with more and better-quality data will become possible.

Apart from that, the algorithm used for extracting coordinations was highly imperfect. As stated before, determining the exact conjunct length based solely on dependency trees is a well-known common issue and cannot be solved automatically (Kanayama et al. 2018, Patejuk and Przepiórkowski 2018). The evaluation process showed that 58% of all Turkish coordinations were extracted correctly.²⁰ However, the algorithm was not evaluated for the Korean language analysis.

Lastly, the author does not know Turkish nor

¹⁹There are 446K tokens in Korean and 735K in Turkish UD corpora opposed to 2.6M in Japanese UD corpora.

²⁰The recurrent issue with Turkish corpora has been that two unrelated simple sentences have been treated as one coordination without a governor. In fact, 20% of coordinations in the sample have been a part of a run-on sentence that had been incorrectly marked as a coordinations. However, this is an issue with corpora, not with the algorithm. However, this problem does not necessarily affect the results of the study. There is no pattern in the incorrectly extracted coordinations, so in a long run all influence from the invalid data points should even out. Moreover, this effect seems to apply only to the coordinations with no governor. If the issue affected strongly the $p_-(n)$ slope, the $L/-$ and $-/R$ contrasts would be significantly different. This, however, did not happened. Overall, while there is no reason to state that this might affect the study result, that cannot be ruled out.

Korean and therefore does not have head-final intuitions internalized, leaving space for an error arising from head-initial-based unconscious assumptions. A native speaker of Turkish has been consulted for the development of heuristics and has performed the evaluation of the algorithm. However, no Korean native speaker was involved in the analysis.

8 Conclusions

The main subject of the study is the analysis of the contrast between the tendencies to put the shorter coordination conjunct at the beginning as the absolute difference between the conjunct length grows with a given governor position. For head-initial languages there are approaches that predicts that the $L/-$ and $-/R$ contrasts are negative, so it may be intuitive to say that for the head-final languages there are approaches that predict that the contrast are positive. The main novel contributions of this paper is the proof that no model of the dependency structure of coordination can predict that. In other words, this paper proves that irrespective of an assumed approach either all tendencies are predicted to be either the same or more descending in the order: left-absent-right governor. This is true for both head-initial and head-final languages.

Additionally, the paper explains why the arguments for a symmetric approach to the dependency structure of coordination provided by **PW23** and **PBG24** cannot be extend to head-final languages. Moreover, the negative replication of aforementioned works in addition to experiments provided by [Kanayama et al. \(2018\)](#) and their theoretical arguments leads to a conclusion that [Choi and Palmer's \(2011\)](#) Inverted Stanford/Moscow approach probably describes best the dependency structure of coordination in head-final languages. However, the lack of significant differences in slopes does not *prove* that there is no difference between them, so the result of the experimental part of this research remains negative. A further analysis with more and better-quality data is needed to strengthen these claims.

As stated in [Kanayama et al. \(2018\)](#), better data could be obtained if Universal Dependencies allowed the HFL corpora annotators to annotate coordinations according to approaches that are more intuitive for HFL users. That would increase the number of correctly annotated coordinations in Korean and Japanese at the cost of universality. However,

since it is possible for head-initial and head-final languages to have a different coordination structure, the dependency structure of coordination may not be universal across languages. Therefore, there is a need for a possibility to annotate coordinations differently in the Universal Dependency standard.

Acknowledgements

The analysis described in this paper was part of the author's bachelor's thesis at the Cognitive Science programme at the University of Warsaw ([Stempniak, 2024](#)). The idea to conduct this research was devised by the supervisor of the thesis, Professor Adam Przepiórkowski. The paper benefited from comments by him, the three anonymous TLT 2024 reviewers, and Magdalena Borysiak. I am grateful to Berke Şenşekerci for his help with the development of the heuristics for the head-final languages analysis as well as for evaluating the algorithm for the analysis of Turkish coordinations.

References

- Jinho D Choi and Martha Palmer. 2011. Statistical dependency parsing in Korean: From corpus generation to automatic parsing. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, pages 1–11.
- Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.
- 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. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.
- Hiroshi Kanayama, Na-Rae Han, Masayuki Asahara, Jena D Hwang, Yusuke Miyao, Jinho D Choi, and Yuji Matsumoto. 2018. Coordinate structures in Universal Dependencies for head-final languages. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 75–84.
- Russell V. Lenth. 2024. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.10.0.
- Arne Lohmann. 2014. *English coordinate constructions*. Cambridge University Press.
- Janet L McDonald, Kathryn Bock, and Michael H Kelly. 1993. Word and world order: Semantic, phonological,

and metrical determinants of serial position. *Cognitive Psychology*, 25(2):188–230.

Agnieszka Patejuk and Adam Przepiórkowski. 2018. *From Lexical Functional Grammar to Enhanced Universal Dependencies: Linguistically informed treebanks of Polish*. Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warszawa.

Martin Popel, David Mareček, Jan Štěpánek, Daniel Zeman, and Zdeněk Žabokrtský. 2013. Coordination structures in dependency treebanks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 517–527.

Adam Przepiórkowski, Magdalena Borysiak, and Adam Głowacki. 2024a. An argument for symmetric coordination from dependency length minimization: A replication study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1021–1033.

Adam Przepiórkowski, Magdalena Borysiak, Adam Okrański, Bartosz Poboźniak, Wojciech Stempniak, Kamil Tomaszek, and Adam Głowacki. 2024b. Symmetric dependency structure of coordination: Crosslinguistic arguments from dependency length minimization. In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*, Hamburg, Germany.

Adam Przepiórkowski and Michał Woźniak. 2023. Con-junct lengths in English, dependency length minimization, and dependency structure of coordination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15494–15512.

R Core Team. 2023. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Wiedeń, Austria.

Wojciech Stempniak. 2024. *Struktura zależnościowa koordynacji – analiza korpusów universal dependencies*. Bachelor’s Thesis, University of Warsaw.

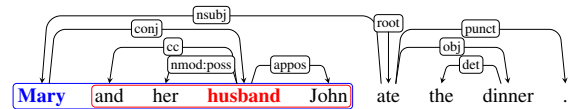
David Temperley. 2007. Minimization of dependency length in written English. *Cognition*, 105(2):300–333.

Sandra K Wright, Jennifer Hay, and Tessa Bent. 2005. Ladies first? Phonology, frequency, and the naming conspiracy. *Linguistics*, 43(3):531–561.

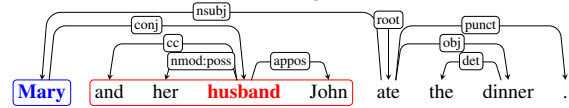
A Algorithm for determining the conjunct contents

This algorithm assumes that the coordination structure is annotated according to the current UD guidelines.

Consider all descendants of the conjuncts’ heads:

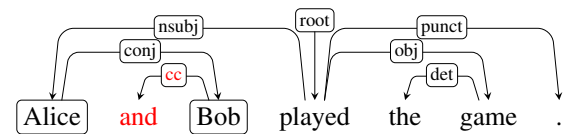


Exclude the head of the right conjunct and its descendants from the left conjunct:



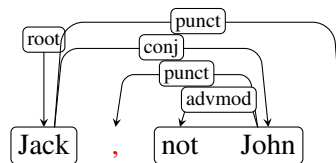
Then apply the following heuristics:

- (H1) A conjunct cannot begin with a conjunction (a word that is connected to the head of the conjunct by a cc relationship).



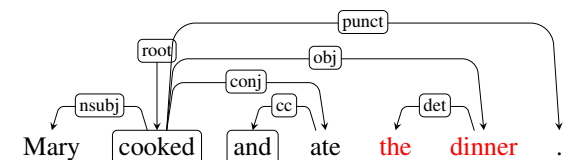
In the example above, *and* is not considered to be a part of the right conjunct because of (H1).

- (H2) A conjunct cannot begin with a punctuation mark (specifically, a comma, semicolon, colon, or dash).



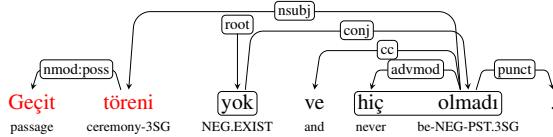
Though the comma is a descendant of the right conjunct head *John*, it is not a part of right conjunct because of (H2).

- (H3) Left head descendants on the right side of the right conjunct are not a part of the left conjunct.



The goal of this heuristic is to exclude the tokens describing the both conjuncts that are dependents of the head of the left conjunct. The intuition supporting it is that the “private” dependents of the head of the left conjunct are almost always near this head and almost never at the right side of the right conjunct.

- (H4) Right head descendants on the left side of the left conjunct are not a part of the right conjunct.

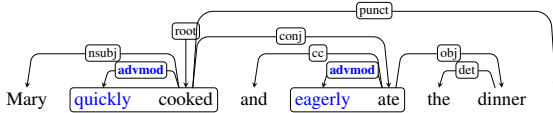


'There is no parade and there never was.'

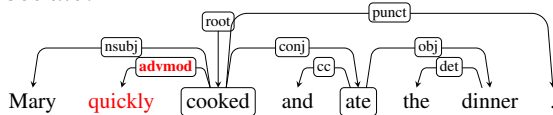
This heuristic has been developed specifically for head-final languages. It is an inverted version of (H4).

- (H5) The child of the left conjunct head on the left side of the left head is not a part of left conjunct, if its relation with left head is *unique*, i.e. there is no relation between any other head and its child identical to it.

This is by far the most unreliable heuristic. Its goal is to tell apart the dependents of the head of the left conjunct describing the left conjunct exclusively from those describing all coordination's conjuncts.

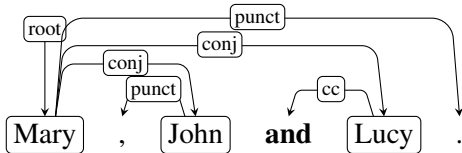


In the example above, the heads of both conjuncts have a dependent with an *advmod* relation. This means that this relation is not *unique*. Therefore, *quickly* describes *cooked*, and *eagerly* describes *ate*.

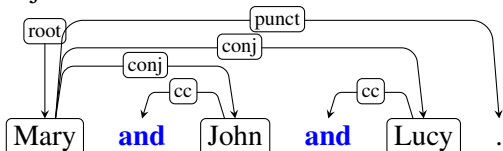


In this example, only the left conjunct head has a dependent with the *advmod* relation. Because of that, this relation is considered to be *unique*. According to (H5), *quickly* describes both *cooked* and *ate*.

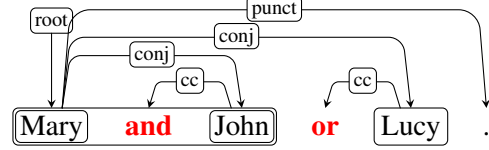
- (H6) If there are multiple different **conjunctions** in a coordination, there is an extra coordination nested in it.



There is only one conjunction in this coordination – *and*. Therefore, this is one coordination with 3 conjuncts.



Here, there are two instances of the same conjunction – *and*. Because of this, this is also one coordination with 3 conjuncts.



In the example above, there are two distinct conjunctions – *and* and *or*. Therefore, in this sentence there are two coordinations with 2 conjuncts each, one nested in another.

B The relative head position within the conjunct

To confirm the assumption that most of the descendants of the conjunct heads are at the left side of the head, the relative position of the head within the conjunct is computed. The relative head position is defined by the formula

$$P = \frac{H - 1}{N - 1} \text{ for } N \geq 2,$$

where P denotes a relative position of the head; H is equal the absolute position of the head within the conjunct (i.e. the ordinal number of the token that is the head); and N is the conjunct length measured in tokens.

Table 5 shows the mean relative position of the heads of the right and left conjuncts of the coordinations found in the Korean and Turkish corpora.

	left conjunct		right conjunct	
	N	mean	N	mean
ko	6801	0.78	12951	0.65
tr	7994	0.64	12763	0.69

Table 5: The relative position of the head within the conjunct in HFL.

In all cases, the heads tend to be in the right half of the conjunct. This is confirmed using the Student's t-test testing the difference from 0.5 (i.e. the middle of the conjunct). All differences are highly significant ($p < 0.001$). Computations were performed using the R's `t.test` function (R Core Team, 2023).

Author Index

- Antonio Díaz Hernández, Roberto, 1
Asma Umniyati, Syahidah, 55
- Berikashvili, Svetlana, 40
Betul Bahceci, Ruveyda, 55
Borysiak, Magdalena, 11
- Carlo Passarotti, Marco, 1
- Dyer, Andrew, 55
Döhmer, Caroline, 30
Dönicke, Tillmann, 23
- Gozalishvili, Anzor, 40
Głowacki, Adam, 11
- Jalaghonia, Tamar, 40
- Krill, Noah, 23
- Lobzhanidze, Irina, 40
Lutgen, Anne-Marie, 30
- Magradze, Erekle, 40
Milano, Emilia, 30
- Okraśniński, Adam, 11
Osenova, Petya, 46
- Paev, Nikolay, 46
Plum, Alistair, 30
Pobożniak, Bartosz, 11
Przepiórkowski, Adam Przepiórkowski, 11
Purschke, Christoph, 30
- Rajestari, Maryam, 55
Rodrigues Menezes de Oliveira Vaz, Helena, 55
Rouvalis, Andreas, 55
- Simov, Kiril, 46
Singhal, Aarushi, 55
Steinberger, Clemens, 23
Stempniak, Wojciech, 11, 65
Stodolinska, Yuliya, 55
- Tomaszek, Kamil, 11
- Zeterberg, Max-Ferdinand, 23