University of Warsaw
Faculty of Psychology

Kamil Tomaszek
Record book number: 432044

Middle Polish Dependency Treebank in Universal Dependencies Format: Design, Implementation, and Analysis

Master's thesis
in COGNITIVE SCIENCE

The thesis was written under the supervision of
Dr. Alina Wróblewska
Institute of Computer Science
Polish Academy of Sciences
Dr. Grzegorz Krajewski
University of Warsaw

Warsaw, September 2025

Summary

This thesis presents a rule-based approach to converting the Middle Polish Dependency Treebank (MPDT), annotated in a Polish-specific scheme, into the Universal Dependencies (UD) format. After introducing the project motivation, data sources, and target standard, the thesis outlines general design assumptions behind the conversion, the mapping strategy, and the validation workflow. It reports overall outcomes of the conversion and sketches applications and extensions, including releasing MPDT-UD and implications for research in historical language processing within cognitive science.

Keywords

Middle Polish, dependency trees, treebank conversion, Universal Dependencies

The title of the thesis in Polish

Średniopolski Bank Drzew Zależnościowych w formacie Universal Dependencies: projekt, implementacja i analiza

Streszczenie

Praca przedstawia podejście regułowe do konwersji Średniopolskiego Banku Drzew Zależnościowych (MPDT), anotowanego w polskim schemacie, do formatu Universal Dependencies (UD). Po krótkim omówieniu motywacji, danych i standardu docelowego zaprezentowano ogólne założenia projektu, strategię odwzorowań oraz schemat walidacji. Przedstawiono ogólne wyniki konwersji oraz możliwe zastosowania i kierunki rozwoju, w tym udostępnienie MPDT-UD i znaczenie dla badań nad przetwarzaniem języka historycznego w kognitywistyce.

Słowa kluczowe

język średniopolski, drzewa zależnościowe, konwersja korpusu, Universal Dependencies

The title of the thesis in English

Middle Polish Dependency Treebank in Universal Dependencies Format: Design, Implementation, and Analysis

Contents

1. Introduction . . . 5
- 1.1. Motivation . . . 5
- 1.2. Objectives . . . 6
- 1.3. Contributions . . . 6
- 1.4. Structure of the Thesis . . . 6

2. Background . . . 7
- 2.1. Dependency Grammar . . . 7
- 2.2. Universal Dependencies . . . 9
- 2.3. Middle Polish Linguistic Resources . . . 12
  - 2.3.1. KorBa . . . 12
  - 2.3.2. MPDT . . . 14

3. Linguistic Features of Middle Polish . . . 17
- 3.1. Orthography and Punctuation . . . 17
  - 3.1.1. Orthography and Transliteration . . . 17
  - 3.1.2. Punctuation . . . 18
- 3.2. Morphology . . . 18
  - 3.2.1. Additional Parts of Speech and Forms . . . 18
  - 3.2.2. Gender System and Declension . . . 20
- 3.3. Syntax . . . 21
  - 3.3.1. Word Order and Non-projectivity . . . 21
  - 3.3.2. Predicate Ellipsis . . . 21
  - 3.3.3. Clause Linking and Subordination . . . 22
- 3.4. Summary . . . 23

4. Conversion Design and Implementation . . . 24
- 4.1. Design Overview and Pipeline . . . 24
- 4.2. POS and Morphological Mapping . . . 24
- 4.3. Dependency Relation Conversion . . . 24
- 4.4. Logging, Testing, and Traceability . . . 24

4.5. Processing Workflow . . . 24

5. Validation and Outcomes . . . 25
5.1. Evaluation Data . . . 25
5.2. UD Validation Setup . . . 25
5.3. Results Overview . . . 25
5.4. Qualitative Error Analysis . . . 25
5.5. Known Limitations and Outstanding Issues . . . 25

6. Applications and Cognitive Science Perspective . . . 26
6.1. Usefulness and Audience . . . 26
6.1.1. Who benefits and how . . . 26
6.1.2. Packaging and License . . . 26
6.1.3. Repository and UD ecosystem integration . . . 26
6.2. Use Cases . . . 26
6.2.1. Historical Syntax and Diachrony . . . 26
6.2.2. Parser Training and Evaluation . . . 26
6.3. Cognitive Science Perspective . . . 26
6.3.1. Processing Constraints . . . 26
6.3.2. Category Change Over Time . . . 26
6.4. Future Work . . . 26
6.4.1. Coverage and Phenomena . . . 26
6.4.2. Generalization and Automation . . . 26

Chapter 1

Introduction

1.1. Motivation

Natural-language preprocessing tools and comparative treebank research have standardized around Universal Dependencies (UD), which enables typologically informed analyses and cross-lingual transfer (Nivre et al. 2020). For Polish texts from the 17th and 18th centuries, however, key resources remain outside UD: texts in the KorBa corpus (Gruszczyński et al. 2022) and the emerging Middle Polish Dependency Treebank (MPDT) are annotated in a Polish-specific scheme (Wieczorek 2025). KorBa is a corpus of historical Polish texts, while MPDT adds a syntactic dependency layer to selected portions of this corpus. However, these resources being annotated in a different format creates challenges for interoperability with UD-based tools and limits straightforward comparative studies with other languages.

A natural solution is to convert these resources to the UD format. From an engineering perspective, however, a faithful, auditable conversion is non-trivial: historical orthography, abbreviations, clitic mobility, numeral complexes, and multiword conjunctions/prepositions interact with head rules and label inventories. Prior conversion experience for contemporary Polish (PDB → PDB-UD; NKJP1M → NKJP1M-UD) offers valuable guidance (Wróblewska 2018; Wróblewska 2020), yet historical data introduce additional phenomena that require explicit, rule-based handling and transparent traceability.

As Wieczorek (2025) notes, MPDT's current format is well-suited to comparative studies with contemporary Polish syntax; at the same time, she highlights the advantages of moving to UD for cross-linguistic comparability, wider intelligibility, and representational options such as enhanced dependencies for shared dependents and shared governors in coordination—even if some information may be lost in conversion. This thesis operationalizes that rationale by delivering a documented, UD-oriented converter for MPDT and preparing the current version of MPDT-UD suitable for validation and downstream use.

The intended users include historical linguists needing UD-compatible data and NLP practitioners interested in diachronic Polish or cross-lingual experiments.

1.2. Objectives

The thesis pursues the following research goals:

(R1) **Design a UD-oriented conversion strategy for MPDT.** Specify mapping principles that respect Middle Polish specifics while aligning with UD guidelines.

(R2) **Implement an auditable conversion pipeline.** Provide modular components for morphosyntax mapping and dependency restructuring, with token-level logging.

(R3) **Ensure UD conformance and evaluability.** Produce output that passes the official UD validator (on all levels) and supports downstream analysis.

(R4) **Document decisions.** Record non-obvious mapping choices and edge-case policies to enable maintenance and reuse.

1.3. Contributions

This project delivers concrete, reusable artifacts:

(C1) **A rule-based MPDT → MPDT-UD converter.** A modular pipeline with fine-grained logging, selectively adapting ideas from PDB→PDB-UD while targeting Middle Polish phenomena. The code will be released in a public repository under an open-source license, together with this thesis, which documents the design and implementation.

(C2) **An initial public release of MPDT-UD.** A set of MPDT (2018 sentences at the time of writing) converted automatically and validated with the official UD validator.

1.4. Structure of the Thesis

- **Chapter 2: Background.** Presents the foundational concepts and resources, including dependency grammar, Universal Dependencies, KorBa and MPDT.
- The remaining chapters will be incorporated here as they are finalized.

Chapter 2

Background

This chapter provides the essential background for understanding the Middle Polish Dependency Treebank conversion to Universal Dependencies. It begins with the theoretical foundations of dependency grammar and its specific Polish manifestation in the Polish Dependency Bank (PDB) scheme. Then it outlines Universal Dependencies as the target framework, highlighting its advantages for cross-linguistic research. Finally, it describes the key resources: KorBa as the source corpus and MPDT as the dependency-annotated dataset that forms the input to our conversion pipeline.

2.1. Dependency Grammar

Dependency grammar is a theory of syntactic structure organized around asymmetric governor–dependent relations. A *dependency* links two lexical items: a *governor* that selects and constrains a dependent, and a *dependent* that is licensed by the governor. One item can be a governor for multiple dependents, but each dependent has a single governor. Sentence structures are modeled as directed trees whose nodes correspond to tokens and whose edges encode these governor–dependent links. The tree has a single *root* (a node with no governor), and every other node is reachable from it along directed edges. In addition to purely structural links, dependency grammar is used here in a morphosyntactic sense, focusing on grammatical relations rather than semantic or prosodic dependency representations.

The dependency scheme used in Middle Polish follows the conventions established for the Polish Dependency Bank (PDB), which is adapted specifically for Polish syntax (Wróblewska 2023). The PDB annotation scheme uses a comprehensive set of morphological categories and dependency relations designed specifically for Polish morphosyntax.

The PDB tagset includes the following morphological categories:

- **Nouns:** `subst` (noun), `depr` (depreciative noun)

- **Pronouns:** `ppron12` (non-third person pronoun), `ppron3` (third person pronoun), `siebie` (reflexive pronoun)

- **Adjectives:** `adj` (adjective), `adja` (adjectival adjective), `adjc` (predicative adjective), `adjp` (prepositional adjective)

- **Verb forms:** `fin` (finite non-past), `praet` (past tense), `imps` (impersonal), `impt` (imperative), `inf` (infinitive), `aglt` (agglutinate of 'być'), `bedzie` (future form of 'być'), `winien` (modal verbs like 'winien'), `pred` (predicative), `ger` (gerund), `pcon` (contemporary adverbial participle), `pant` (anterior adverbial participle), `pact` (active adjectival participle), `ppas` (passive adjectival participle)

- **Numerals:** `num` (cardinal numeral), `numcomp` (numeral compound)

- **Conjunctions:** `comp` (subordinating conjunction), `conj` (coordinating conjunction)

- **Other categories:** `adv` (adverb), `brev` (abbreviation), `dig` (Arabic numeral), `romandig` (Roman numeral), `emo` (emoticon), `fill` (filler), `frag` (fragment), `interj` (interjection), `interp` (punctuation), `part` (particle), `prep` (preposition), `ign` (unrecognized form)

The PDB annotation scheme distinguishes several classes of dependency relations:

- **Core arguments:** `subj` (subject), `obj` (direct object), `obj_th` (thematic object), `comp` (complement), `comp_fin` (finite complement), `comp_inf` (infinitive complement), `comp_ag` (agent complement)

- **Adjuncts and modifiers:** `adjunct` with semantic subtypes such as `adjunct_temp` (temporal), `adjunct_loc` (locative), `adjunct_dur` (duration), `adjunct_caus` (causal), `adjunct_mod` (manner), `adjunct_emph` (emphatic particle), `adjunct_compar` (comparative)

- **Predicate-related:** `pd` (predicate), `aux` (auxiliary), `neg` (negation)

- **Coordination:** `conjunct` (coordinated element), `pre_coord` (pre-coordinator)

- **Multiword expressions:** `mwe` (multiword expression), `ne` (named entity), `ne_foreign` (foreign named entity)

- **Special relations:** `punct` (punctuation), `vocative` (vocative), `refl` (reflexive), `orphan` (orphaned dependent), `discourse` (discourse marker), `parataxis` (parataxis), `root` (sentence root)

The example dependency trees below illustrate the scheme of a PDB-annotated sentence alongside its UD counterpart, showing the structural differences.

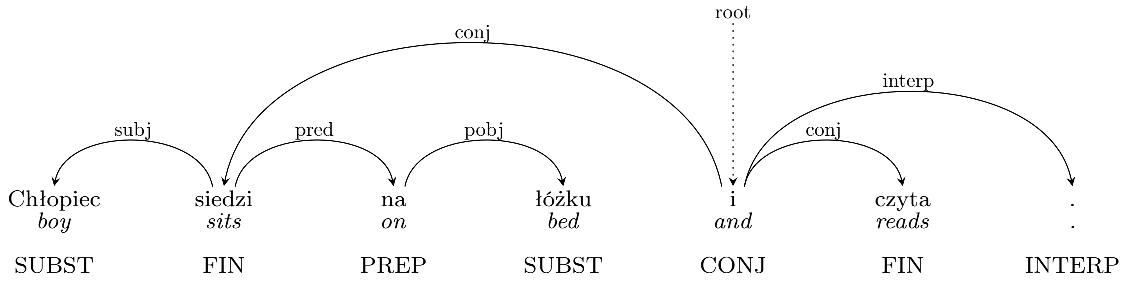


Figure 1: Example dependency tree in the PDB format

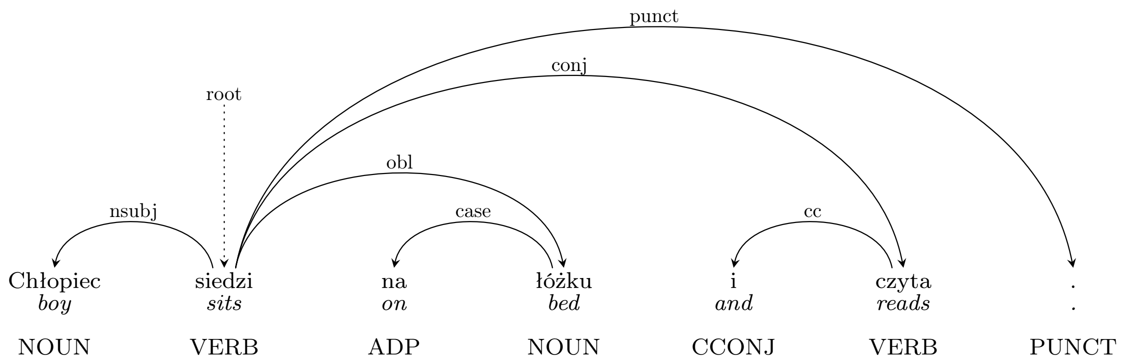


Figure 2: A dependency tree of the same sentence in UD format

Dependency formalisms differ on certain design choices (e.g., whether adpositions are heads or dependents inside adpositional phrases; how to encode coordination; whether and how to mark valency vs. modification). The PDB scheme takes specific positions on these issues, treating prepositions as heads (note the on→bed relation in Figure 1), using a coordination-centric approach where conjunctions govern coordinated elements (*i* being the `root` of both *siedzi* and *czyta* in Figure 1), and distinguishing arguments from adjuncts through detailed semantic role marking.

2.2. Universal Dependencies

Universal Dependencies (hereafter UD) is a cross-linguistic annotation framework designed to harmonize morphosyntactic and syntactic representations across languages within a dependency-based, lexicalist model (Nivre et al. 2020; de Marneffe et al. 2021). UD serves as both a theoretical framework and a practical collection of treebanks—currently the largest repository of multilingual dependency trees with over 200 treebanks for more than 150 languages.[1] It is widely adopted in NLP and linguistic typology, and is maintained by an open community with regular releases.

The scheme provides three aligned layers for sentence-level annotation:

1. **Tokenization.** UD defines dependencies between *syntactic words*. To handle orthographic contractions or clitic clusters, it uses *multiword tokens*, ensuring a faithful word-level analysis. A multiword token is a single orthographic unit that is split into multiple syntactic words, each receiving its own morphological analysis and syntactic role.

   For example, Middle Polish *kiedym* 'when I' is annotated as:

   ```
   14-15   kiedym    _       _     ...
   14      kiedy     kiedy   ADV   ...
   15      m         być     AUX   ...
   ```

   Here, the single orthographic token *kiedym* (ID 14-15) splits into two syntactic words: *kiedy* 'when' (ID 14) and *m* (clitic form of 'I am', ID 15).

   Similarly, *jeszcześ* 'still you are' becomes:

   ```
   7-8   jeszcześ   _          _      ...
   7     jeszcze    jeszcze    PART   ...
   8     ś          być        AUX    ...
   ```

2. **Morphology.** Each syntactic word is associated with a `LEMMA`, a universal part-of-speech tag (hereafter part of speech tag=`POS`; universal part of speech tag=`UPOS`) from a fixed 17-tag set, and a bundle of `FEATS` (morphological features). The UPOS tags cover open-class words (adjectives `ADJ`, adverbs `ADV`, interjections `INTJ`, nouns `NOUN`, proper nouns `PROPN`, verbs `VERB`), closed-class words (adpositions `ADP`, auxiliary verbs `AUX`, coordinating conjunctions `CCONJ`, determiners `DET`, numerals `NUM`, pronouns `PRON`, particles `PART`, subordinating conjunctions `SCONJ`), and other categories (punctuation `PUNCT`, symbols `SYM`, other `X`). UD v2 standardized features and values across languages and clarified tag boundaries, e.g. extending auxiliary verbs to copulas and tense–aspect–mood particles while narrowing particles.

3. **Syntax.** The syntactic layer is a single-rooted tree with possible 37 universal dependency relations organized according to functional and structural categories. Sentence structures are modeled as directed trees whose nodes correspond to syntactic words and whose edges encode governor–dependent links. Relations include:

[1] Universal Dependencies, `https://universaldependencies.org`, accessed 2025-10-10.

- core arguments (nominal subject `nsubj`, direct object `obj`, indirect object `iobj`, and clausal complement `ccomp`),
- non-core dependents (oblique `obl`, dislocated element `dislocated`, adverbial clause modifier `advcl`, adverbial modifier `advmod`, discourse element `discourse`, auxiliary `aux`, and copula `cop`),
- nominal dependents (nominal modifier `nmod`, numeral modifier `nummod`, adjectival modifier `amod`, determiner `det`, and case marker `case`),
- coordination (conjunct `conj`, coordinating conjunction `cc`),
- multiword expressions (fixed `fixed`, flat `flat`),
- special relations (list element `list`, parataxis `parataxis`, orphan `orphan`, punct `punct`, root `root`, other dependent `dep`).

The framework also allows language-specific subtypes (e.g., `nsubj:pass` for passive subjects, `det:poss` for possessive determiners) and defines semi-mandatory subtypes that should be used when the relevant phenomenon exists in the language. A full list of relations and subtypes, along with their descriptions, is available in the UD webpage.[2]

In addition to the *basic* representation, UD also defines an *enhanced* graph that adds extra arcs (and occasionally null nodes) to capture phenomena such as shared dependents in coordination, control and raising, relativization, and ellipsis. In Figure 2, the basic tree structure is shown; an enhanced representation would add an additional edge to represent the dependent (in this case: the subject) of *czyta* (reads) as also being the *Chłopiec* (boy), as shown in figure Figure 3.

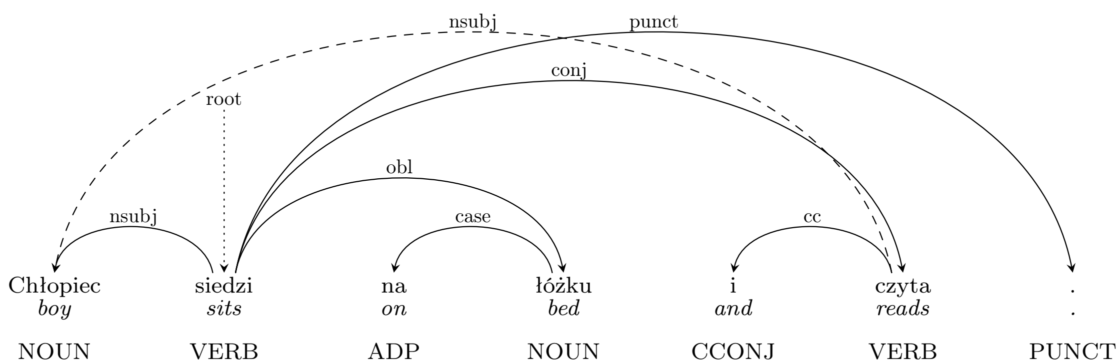


Figure 3: A dependency tree with enhanced dependencies (dashed lines)

[2]Universal Dependencies relations list: `https://universaldependencies.org/u/dep/index.html`

Format: For practical implementation and data sharing, UD annotations must be encoded in a standardized format. UD uses the CoNLL-U format, a ten-column tabular specification with the fields:

- `ID` - a syntactic word index (or range for multiword tokens);
- `FORM` - the surface form;
- `LEMMA` - the dictionary form;
- `UPOS` - the universal POS tag;
- `XPOS` - a language-specific POS tag;
- `FEATS` - a pipe (|) separated list of morphological features;
- `HEAD` - the index of the head syntactic word (or 0 for the root);
- `DEPREL` - the dependency relation to the head;
- `DEPS` - for enhanced dependencies;
- `MISC` - for miscellaneous annotations.

Here is a CoNLL-U snippet for the sentence "Chłopiec siedzi na łóżku i czyta.", with the enhanced dependencies.

```
# sent_id = test-sentence
# text = Chłopiec siedzi na łóżku i czyta.
1   Chłopiec   chłopiec   NOUN    subst    Gender=Masc|Number=Sing|Case=Nom                        2   nsubj   _          _
2   siedzi     siedzieć   VERB    fin      Aspect=Imp|Mood=Ind|Tense=Pres|Person=3|Number=Sing     0   root    _          _
3   na         na         ADP     prep     AdpType=Prep|Case=Loc                                   4   case    _          _
4   łóżku      łóżko      NOUN    subst    Gender=Neut|Number=Sing|Case=Loc                        2   obl     _          _
5   i          i          CCONJ   conj     _                                                       2   cc      _          _
6   czyta      czytać     VERB    fin      Aspect=Imp|Mood=Ind|Tense=Pres|Person=3|Number=Sing     2   conj    1:nsubj    _
7   .          .          PUNCT   interp   PunctType=Peri                                          2   punct   _          _
```

2.3. Middle Polish Linguistic Resources

2.3.1. KorBa

KorBa (Gruszczyński et al. 2022) – from Polish *Korpus Barokowy*, "Baroque Corpus" – is a 13.5-million-token corpus of Polish texts from 1601–1772, compiled from over seven hundred sources and annotated morphosyntactically (lemmas, POS, features). It is searchable via MTAS (Multi Tier Annotation Search; Brouwer et al. 2017), and provides parallel transliteration/transcription layers, structural and language markup, and rich metadata (period, region, text type, genre) that enable stratified analyses.

The corpus includes diverse text types ranging from literary works (epic poetry, drama, lyric poetry) to non-literary materials (scientific-didactic texts, persuasive writings, factual literature, official documents, press releases) and biblical texts. Geographically, texts span the Polish-Lithuanian Commonwealth, with approximately 27% of the corpus being of unknown origin. As shown in Figures 4 and 5, the corpus maintains careful balance across regions and text types to ensure representativeness of Middle Polish.

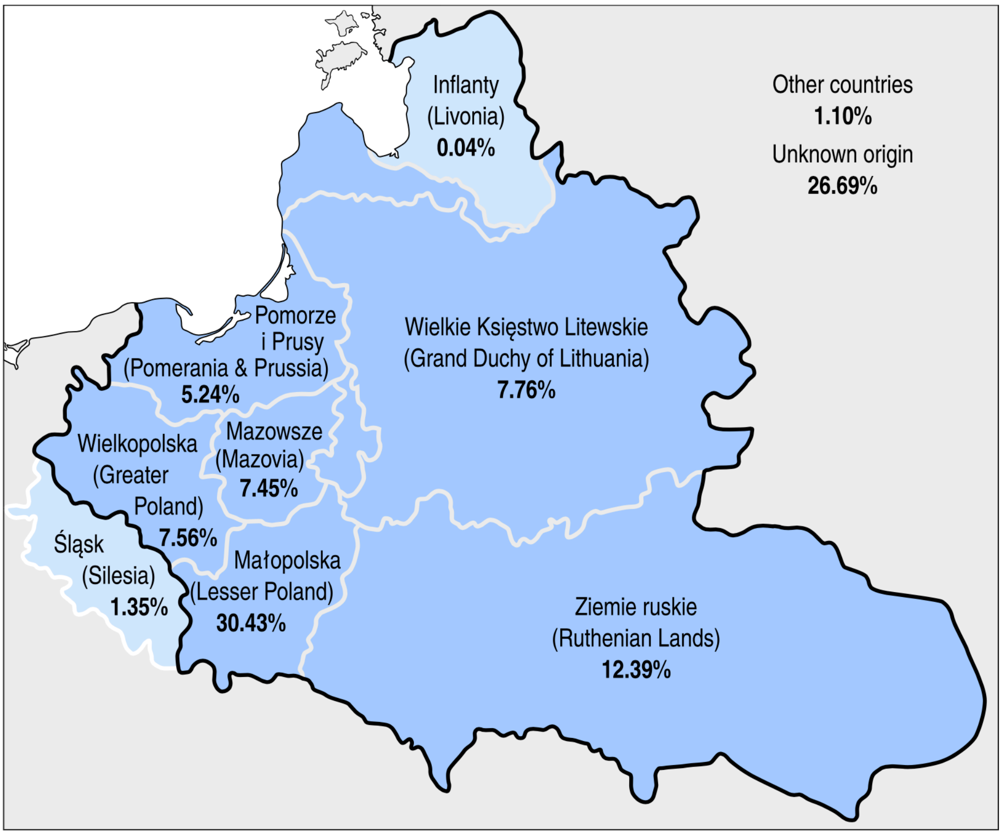


Figure 4: Geographical distribution of texts in the corpus displayed on the map of the Commonwealth after the Union of Lublin of 1569. Source: Gruszczyński et al. (2022), p. 315, CC BY 4.0.

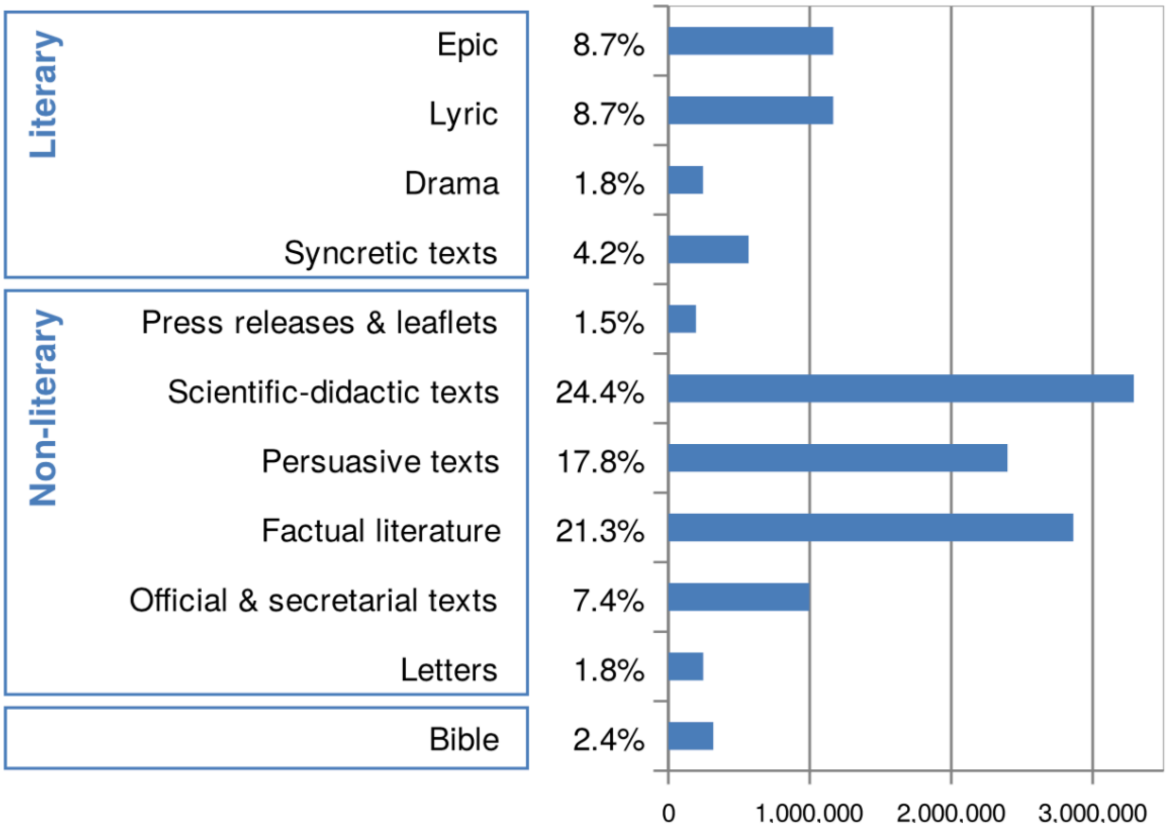


Figure 5: Types of texts in KorBa. Source: Gruszczyński et al. (2022), p. 316, CC BY 4.0.

2.3.2. MPDT

The Middle Polish Dependency Treebank (MPDT) is a manually curated, syntactically annotated subset of the KorBa corpus, capturing key syntactic phenomena of 17th–18th-century Polish texts. Sentences chosen for MPDT exclude poetry and Latin insertions. This ensures high-quality tokenization, lemmas, POS tags, and morphological features, while maintaining representativeness across period, region, and text type.

Annotation workflow:

1. **Automatic pre-annotation.** Two parsers trained on contemporary PDB data (MaltParser, COMBO) generate initial dependency analyses.

2. **Manual correction.** Two linguist annotators independently revise parser outputs, leveraging complementary error profiles.

3. **Adjudication.** Conflicting annotations are resolved by an adjudicator to produce a single gold-standard tree.

4. **Format and tooling.** Final annotations are encoded in CoNLL-U with KorBa's extended tagset (e.g., dual number `Dual`) and processed using MaltEvalAnnotator.

To accommodate Middle Polish morphology, which is described in detail in chapter 3, several POS and feature categories were added:

- `adjb`, `ppasb`, `ppraet` for short-forms and past participles
- Feature `Number=Dual` alongside `Sing` and `Plur` for the dual number.

Corpus statistics

- Total sentences: 2 018
- Total tokens: 47 273
- Distinct POS tags: 45
- Distinct dependency relations: 27
- Non-projective edges: 3 748 across 879 sentences
- Average sentence length: 23.43 tokens

Figure 6 presents the 20 most frequent MPDT POS tags, highlighting the prominence of nouns (`subst`: 11,374 occurrences), punctuation (`interp`: 7,971), adjectives (`adj`: 5,315), and prepositions (`prep`: 4,391).

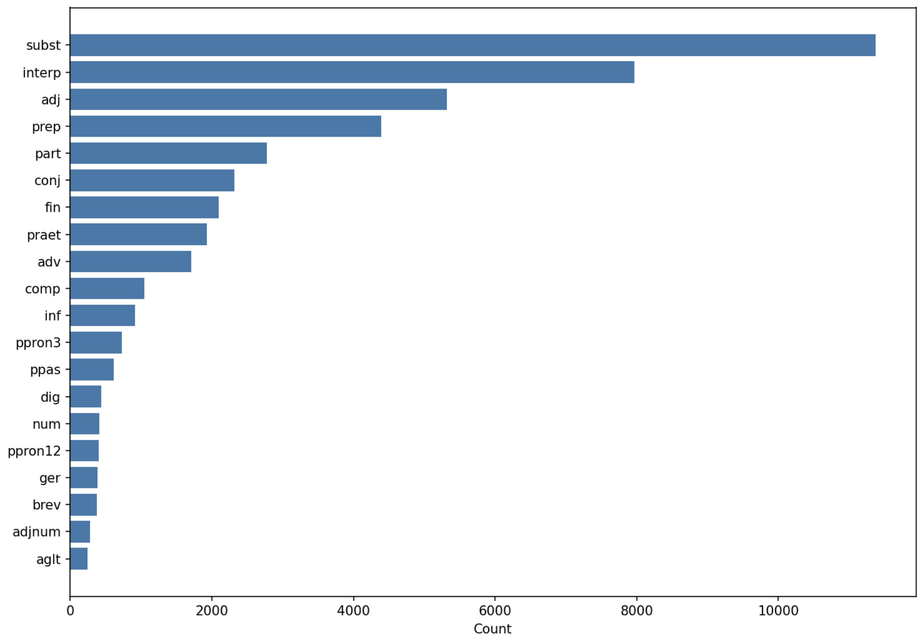


Figure 6: Top 20 MPDT POS tag frequencies

Figure 7 shows the distribution of the top 20 dependency relation bases. Adjuncts (`adjunct`: 13,276) and complements (`comp`: 8,539) are most common, followed by punctuation (`punct`: 6,896), coordination elements (`conjunct`: 6,071), and core arguments (`obj`: 3,423; `subj`: 2,286).

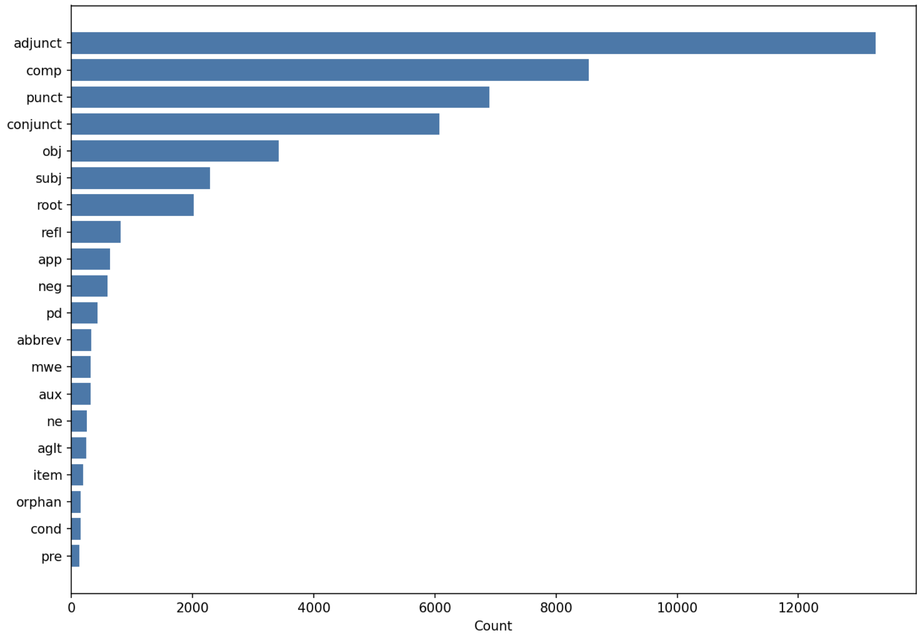


Figure 7: Top 20 MPDT dependency relation base frequencies

Chapter 3

Linguistic Features of Middle Polish

This chapter characterizes the linguistic system of Middle Polish as represented in the KorBa corpus and the Middle Polish Dependency Treebank (MPDT). It outlines key differences from modern Polish orthography, morphology, and syntax, emphasizing those that directly affect dependency annotation and conversion to Universal Dependencies (UD).

3.1. Orthography and Punctuation

3.1.1. Orthography and Transliteration

The KorBa corpus preserves two parallel orthographic layers: *transliteration* (a faithful rendering of the historical text) and *transcription* (a normalized spelling approximating contemporary Polish orthography). As noted in the KorBa manual, transliteration reflects the original graphic form of seventeenth- and eighteenth-century sources, while the transcription adapts them to modern conventions, keeping key phonetic and morphological features of Middle Polish.

Middle Polish orthography was far from standardized. The same word could appear in several spelling variants, sometimes even within a single text. Graphemes were often used interchangeably (e.g. *i/y*, *u/v*, *ć/ci*, *rz/ż*). Long vowels (*á*, *é*) and palatalization (*ć*, *ź*, *ś*, *ń*) were marked inconsistently. The KorBa transliteration layer preserves this variation, while the transcription layer normalizes it (e.g. *rodźicow* → *rodziców*).[3]

Orthographic conventions also influenced tokenization. Many expressions that are today written separately were then written together, and vice versa. Following the Wieczorek (2020), elements that would now be written jointly are linked with the `mwe` relation during annotation.

Examples:

[3] See Gruszczyński et al. (2022), p. 317.

- Historical joint writing: *zchęći* (modern *z chęci*; 'from willingness') → split into two syntactic words in UD.
- Historical separate writing: *dla tego* (modern *dlatego*; 'because', lit. 'for this') is treated as a single prepositional unit.

3.1.2. Punctuation

As described by Wieczorek (2025), punctuation in Middle Polish reflected the rhythm and pauses of speech rather than syntactic boundaries. Marks were used inconsistently and sometimes idiosyncratically: slashes (/) often functioned as commas, semicolons as commas, and colons as semicolons or dashes. Conversely, long unpunctuated stretches also occur. During syntactic annotation, punctuation is interpreted according to its syntactic role, not its original mark, e.g. a slash (/) introducing a new clause is annotated as `punct`.

> *Powstawszy raz z bárzo ćięszkiey choroby/ ták rzekł Nie nagorzey śię zemną stáło: Bo mię chorobá vpomniáłá/ ábym się w pychę nie podnośił/ ponieważem iest śmiertelny.*
> ('Having once recovered from a very severe illness/ he said thus: It did not go too badly with me: For the illness reminded me/ that I should not lift myself up in pride/ since I am mortal.')
> *Source: MPDT corpus / metadata.*

3.2. Morphology

The morphological system of Middle Polish differs significantly from the modern language, both in its inventory of forms and in category values. These distinctions were codified in the KorBa 2.0 tagset and later adopted in the MPDT.

3.2.1. Additional Parts of Speech and Forms

The Middle Polish tagset introduces several POS categories absent in contemporary Polish:

(a) Short-form adjectives (`adjb`). These forms—like *żyw*, *godzien*—are indeclinable or partially declined adjectives, often used predicatively without the copula. They correspond to UD `ADJ` with `Variant=Short`.

> *Iak długo ia* ***żyw*** *iestem, żyie Pán moy poty, Czuię bol y wesołość, czuię y kłopoty.*

('As long as I live, my Lord lives likewise, I feel pain and joy, I also feel troubles.')
Source: MPDT corpus / metadata.
In modern Polish, the short form *żyw* ('alive') would be considered archaic or poetic; the modern equivalent is *żywy*.

*Chcesz się zemną równać: nie **godzien**eś tego.*
('You want to match yourself with me: you are not worthy of this.')
Source: MPDT corpus / metadata.
In modern Polish, *godzien* still exists, along with words like *pewien* ('certain'), however their usage is now limited, and the standard forms are *godny*, and *pewny*.

(b) Short passive participles (`ppasb`). Passive participles in the uninflected short form (e.g. *zbawion*, *pisan*) co-occur with finite forms of *być*. They are annotated as `ADJ` with `VerbForm=Part, Voice=Pass, Variant=Short`.

*Kto vwierzy, á okrzći się, **zbáwion** będźie, ále kto nie vwierzy będźie **potępion**.*
('Whoever believes and is baptized will be saved, but whoever does not believe will be condemned.')
Source: MPDT corpus / metadata.
In modern Polish, these short forms are archaic; the standard forms are fully inflected *zbawiony*, *potępiony*.

Pisań na zamku pileckim, dnia 23 miesiąca lipca, roku Pańskiego 1620.
('Written at the castle of Pilec, on the 23rd day of July, in the Year of Our Lord 1620.')
Source: MPDT corpus / metadata.
In modern Polish, the short form *pisan* is archaic; the standard form is *pisany*.

(c) Past participles (`ppraet`). These forms, such as *osłabiałe*, *opuchłymi*, *zasiniałymi*, represent an older stage of adjectival participles derived from past tenses, intermediate between `ppas` and `pact`. They are mapped to `ADJ` with `VerbForm=Part` and `Tense=Past`.

*Częstokroć ábowiem były widáne z twarzámi **opuchłymi**/ **záśiniáłymi**.*
('For often they were seen with swollen/ bruised faces.')

Source: MPDT corpus / metadata.

In modern Polish, some of those forms are still in use, but some are archaic or poetic. For example, *opuchłymi* would be rather replaced by *opuchniętymi*, while *zasiniałymi* is still acceptable.

(d) Dual number (`Number=Dual`). Middle Polish still preserved dual forms for certain nouns, numerals, adjectives, and verbs. The KorBa manual documents the explicit tag `du`. These forms gradually merged with the plural after ca. 1740, though fossilized duals like *ręce*, *oczy* survive in modern Polish (sg. *oko* ['eye'] → pl. *oczy* when referring to the organ, but also pl. *oka* when used in other sense, e.g. *oka w rosole* ('eyes in the broth'); similarly sg. *ucho* ['ear'] → pl. *uszy*, when about body parts, or pl. *ucha*, when referring to cup handles).

*6. Po przepędzonych przez **dwie lecie** tych okrutnych boleściách, pokazał się iey Pan mowiąc: Iż bez lat pięc nie miałaby iadać ani mięsa. ani nabiáłu.*
('6. After two years spent in these cruel pains, the Lord appeared to her saying: That for five years she should not eat either meat or dairy.')
Source: MPDT corpus / metadata.

Here, *dwie lecie* is dual accusative of *dwa* (two) and *lato* ('summer; year'). In modern Polish, the dual form is archaic; the standard form (both the nominative and accusative) is *dwa lata*.

3.2.2. Gender System and Declension

The masculine gender system was less differentiated than today. KorBa distinguishes three values: `m` (general masculine), `manim1` (personal), and `manim2` (non-personal). In early texts, these values overlap; many forms do not yet reflect consistent distinctions in case endings. For example, *ptaki* and *ptacy* alternate for 'birds' depending on context.

*(m) 6. Vbogáćiłeś ich chybkośćią i lotem nád wszytkie loty prędszym i bystrzeyszym/ i bystrym ták/ iż i strzały/ i **ptaki**/ i pioruny poprzedzáć/ á wszytkie rzeczy/ mury/ skáły/ przenikać mogą.*
('You have enriched them with speed and with flight swifter and sharper than all flights/ so that even arrows/ and birds/ and thunder they can outpace/ and penetrate all things/ walls/ rocks.')
Source: MPDT corpus / metadata.

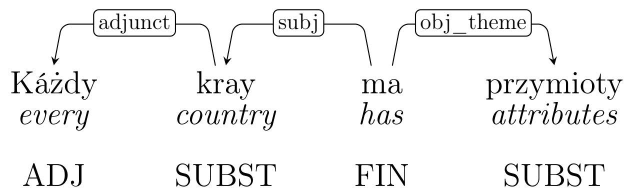
*(manim1) 122. Czemu **ptacy** ktorzy ogona nie máią dlugie nogi maią?*
('Why do birds that do not have a tail, have long legs?')
Source: MPDT corpus / metadata.

3.3. Syntax

3.3.1. Word Order and Non-projectivity

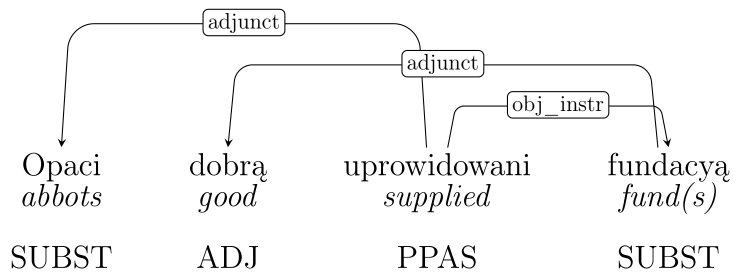
Middle Polish syntax exhibits high flexibility of word order, frequent inversion, and long-distance dependencies. As noted by Wieczorek (2025), discontinuous structures—especially in noun phrases with adjectival modifiers—often yield non-projective trees.

Example 1 (linear order): *Káżdy kray ma przymioty* → no crossing edges.



Source: adapted from Wieczorek (2025), Fig. 6, p. 12.

Example 2 (discontinuous order): *Opaci dobrą uprowidowani fundacyą* ('Abbots supplied with good funds') → crossing edges between *dobrą* and *fundacyą*.



Source: adapted from Wieczorek (2025), Fig. 7, p. 12.

These inversions complicate automatic parsing and were one motivation for explicit rule-based conversion to UD.

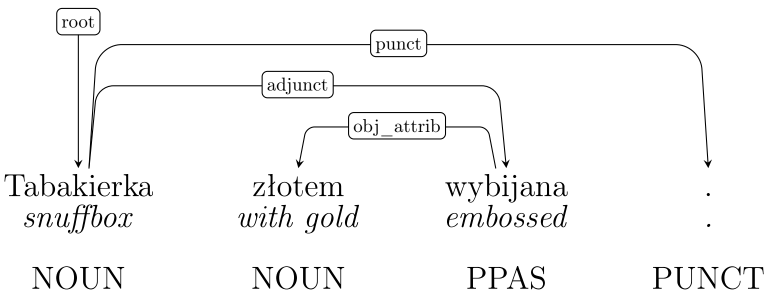
3.3.2. Predicate Ellipsis

As noted by Wieczorek (2025), it is rare in modern Polish for sentences to lack a predicate (at least in texts written in careful language), but this was quite common in 17th–18th-century Polish. In dependency analysis, the predicate is considered the centre of the sentence (`root`)—most often a finite verb. In the absence of a predicate, another sentence element serves as the centre. Most often, this centre becomes the subject, which receives the label `root` instead of `subj`.

Example (missing verbal predicate):

Tabakierka złotem wybijana .

('A snuffbox, embossed with gold.')



Source: adapted from Wieczorek (2025), Fig. 8, p. 13.

3.3.3. Clause Linking and Subordination

Middle Polish frequently employs conjunctions that have since changed meaning, an example of which could be the token *jako*. It is frequently employed in two functions: (i) as a comparative/similative marker in the sense of modern *jak* ('like/as; when/how'), and (ii) in the modern-like role/identity sense 'as'.

> *Ale iako nowi Obywatele tam przybywać poczeli z Kir, czy Syr Kraiu w Medii leżącego, Kirya, to Syria zwać się poczęła.*
> ('But as/when new inhabitants began to arrive there from the land of Kir, or Syr, lying in Media, Kirya then began to be called Syria.')
> *jako* in the sense of *jak* 'when'
> *Source: MPDT corpus / metadata.*

> *Iako Roża rozpuszcza z przyrodzenia swego zapach przyiemny, tak Serce dobroczynne wydaie bez przyniewolenia uczynki dobre.*
> ('As a rose by its nature gives off a pleasant fragrance, so a charitable heart produces good deeds without compulsion.')
> *jako* in the sense of *jak* 'how/as'
> *Source: MPDT corpus / metadata.*

> *Ja zaś w tych terminach stawam jako mediator, prowadząc do zgody obiedwie strony.*
> ('And I, for my part, in these proceedings stand as a mediator, leading both sides to agreement.')
> *jako* = 'as (in the role of)'
> *Source: MPDT corpus / metadata.*

3.4. Summary

Middle Polish exhibits substantial divergence from modern Polish in orthography, morphology, and syntax:

- Orthography: inconsistent, variable, often merging or splitting tokens differently from modern norms.
- Punctuation: prosodic rather than syntactic, with slashes and colons used irregularly.
- Morphology: additional forms (`adjb`, `ppasb`, `ppraet`), productive dual number, and fluid gender distinctions.
- Syntax: high non-projectivity, frequent inversion, ellipsis, and loose coordination.

These properties directly inform the design of the MPDT $\rightarrow$ MPDT-UD conversion pipeline, motivating special conversion rules and additional validation layers to preserve linguistic authenticity while ensuring formal compatibility with Universal Dependencies.

Chapter 4

Conversion Design and Implementation

4.1. Design Overview and Pipeline

4.2. POS and Morphological Mapping

4.3. Dependency Relation Conversion

4.4. Logging, Testing, and Traceability

4.5. Processing Workflow

Chapter 5

Validation and Outcomes

5.1. Evaluation Data

5.2. UD Validation Setup

5.3. Results Overview

5.4. Qualitative Error Analysis

5.5. Known Limitations and Outstanding Issues

Chapter 6

Applications and Cognitive Science Perspective

6.1. Usefulness and Audience

6.1.1. Who benefits and how

6.1.2. Packaging and License

6.1.3. Repository and UD ecosystem integration

6.2. Use Cases

6.2.1. Historical Syntax and Diachrony

6.2.2. Parser Training and Evaluation

6.3. Cognitive Science Perspective

6.3.1. Processing Constraints

6.3.2. Category Change Over Time

6.4. Future Work

6.4.1. Coverage and Phenomena

6.4.2. Generalization and Automation

Bibliography

Brouwer, M., Brugman, H., and Kemps-Snijders, M. (2017). *MTAS: A Solr/Lucene based multi-tier annotation search solution*. CLARIN. URL: http://www.ep.liu.se/ecp/136/002/ecp17136002.pdf.

De Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). "Universal Dependencies". In: *Computational Linguistics* 47.2, pp. 255–308. ISSN: 0891-2017. DOI: 10.1162/coli_a_00402. URL: https://doi.org/10.1162/coli_a_00402.

Gruszczyński, W., Adamiec, D., Bronikowska, R., Kieraś, W., Modrzejewski, E., Wieczorek, A., and Woliński, M. (2022). "The Electronic Corpus of 17th- and 18th-century Polish Texts". In: *Language Resources and Evaluation* 56.1, pp. 309–332. ISSN: 1574-0218. DOI: 10.1007/s10579-021-09549-1. URL: https://doi.org/10.1007/s10579-021-09549-1.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). "Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S. Marseille, France: European Language Resources Association, pp. 4034–4043. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.497/.

Wieczorek, A. (2020). *Instrukcja anotowania drzew zależnościowych: Doyeark dla tekstów XVII- i XVIII-wiecznych*. Institute of Polish Language, Polish Academy of Sciences.

Wieczorek, A. (2025). "Towards the Middle Polish Dependency Treebank". In: *Native Language in the 21st Century: System, Communication Practices and Education*. V & R Unipress.

Wróblewska, A. (2018). "Extended and Enhanced Polish Dependency Bank in Universal Dependencies Format". In: *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*. Ed. by de Marneffe, M.-C., Lynn, T., and Schuster, S. Brussels, Belgium: Association for Computational Linguistics, pp. 173–182. DOI: 10.18653/v1/W18-6020. URL: https://aclanthology.org/W18-6020/.

Wróblewska, A. (2020). “Towards the Conversion of National Corpus of Polish to Universal Dependencies”. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S. Marseille, France: European Language Resources Association, pp. 5308–5315. ISBN: 979-10-95546-34-4. URL: `https://aclanthology.org/2020.lrec-1.653/`.

Wróblewska, A. (2023). *Instrukcja anotowania drzew w Polskim Banku Drzew Zależnościowych.*