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

Chapter 1

Introduction

1.1. Motivation

Natural-language tools and comparative treebank research have standardized around Universal Dependencies (UD), which enables typologically informed analyses and cross-lingual transfer (Nivre, de Marneffe, Ginter, Hajič, Manning, Pyysalo, Schuster, Tyers, and Zeman 2020). For 17th–18th-century Polish, however, key resources remain outside UD: Middle Polish texts in KorBa (Gruszczyński, Adamiec, Bronikowska, Kieraś, Modrzejewski, Wieczorek, and Woliński 2022) and the emerging Middle Polish Dependency Treebank (MPDT) are annotated in a Polish-specific scheme (Wieczorek 2025). This limits their interoperability with UD-based tools and does not allow for straightforward comparative studies with other languages.

From an engineering perspective, a faithful, auditable conversion is non-trivial: historical orthography, abbreviations (`brev`), clitic mobility (*by*, *że*), 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 PDB-consistent 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 translation. This thesis operationalizes that rationale by delivering a documented, UD-oriented conversion for MPDT and preparing an initial MPDT-UD subset suitable for validation and downstream use.

1.2. Objectives

The thesis pursues the following goals:

(O1) Design a UD-oriented conversion strategy for MPDT. Specify mapping principles that respect Middle Polish specifics while aligning with UD guidelines.

(O2) Implement an auditable conversion pipeline. Provide modular components for morphosyntax mapping and dependency restructuring, with token-level logging.

(O3) Ensure UD conformance and evaluability. Produce output that passes the official UD validator (on all levels) and supports downstream analysis.

(O4) 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 → UD converter. A modular pipeline with fine-grained logging, selectively adapting ideas from PDB→UD while targeting Middle Polish phenomena. The code will be released in a public repository under an open-source license, together with this paper, which documents the design and implementation.

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

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

1.4. Structure of the Document

- **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

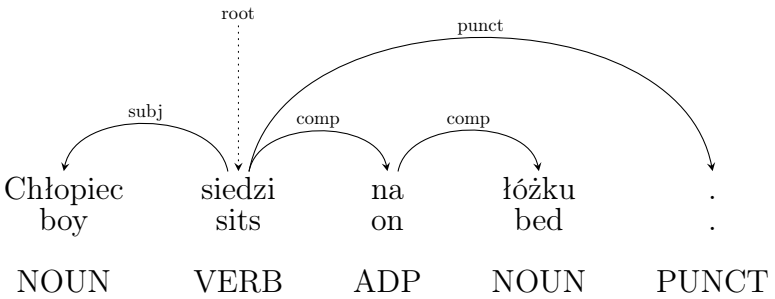
Background

2.1. Dependency Grammar

Dependency grammar is a theory of syntactic structure organized around asymmetric head–dependent relations. A *dependency* links two lexical items: a *head* that selects and constrains, and a *dependent* that is licensed by the head. Sentences are modeled as directed trees whose nodes correspond to tokens and whose edges encode these head–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, we use dependency grammar here in a strictly morphosyntactic sense, leaving semantic or prosodic dependency representations aside.

The example dependency trees below illustrates the notation used in this thesis. Edge labels are function-like and purely illustrative.

Example tree



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). One universal standard is Universal Dependencies.

2.2. Universal Dependencies

Universal Dependencies (further references as UD) is a cross-linguistic annotation framework designed to harmonize morphosyntactic and syntactic representations across

languages within a dependency-based, lexicalist model (Nivre, de Marneffe, Ginter, Hajič, Manning, Pyysalo, Schuster, Tyers, and Zeman 2020). It is widely adopted in NLP and linguistic typology, and serves as the target formalism for the conversion presented in this thesis. The scheme provides three aligned layers:

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. Conversely, several orthographic tokens may be combined into one syntactic word in well-motivated, language-specific cases.

2. **Morphology.** Each token is associated with a `LEMMA`, a universal POS tag (`UPOS` from a fixed 17-tag set), and a bundle of `FEATS`. UD v2 standardized features and values across languages and clarified tag boundaries, e.g. extending `AUX` to copulas and tense–aspect–mood particles while narrowing `PART`.

3. **Syntax.** The syntactic layer is a single-rooted tree with universal dependency relations such as `nsubj`, `obj`, `obl`, `amod`, `case`, `cc`, and `conj`. UD v2 unified earlier divergences by, for example, replacing `dobj` with `obj`, reserving `obl` for predicate-level obliques (distinct from nominal `nmod`), and constraining the use of `cop` to pure grammatical linkers.

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.

Format: UD uses the CoNLL-U format, a ten-column tabular specification with the fields:

- `ID` - a token 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 token (or 0 for the root);
- `DEPREL` - the dependency relation to the head;
- `DEPS` - for enhanced dependencies;
- `MISC` - for miscellaneous annotations.

Here is an example CoNLL-U snippet:

```
# sent_id = test-sentence
# text = Chłopiec siedzi na łóżku.
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   .           .           PUNCT   interp   PunctType=Peri                                        2   punct    _   _
```

2.3. Resources

2.3.1. KorBa

KorBa 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), and provides parallel transliteration/transcription layers, structural and language markup, and rich metadata (period, region, text type, genre) that enable stratified analyses (Gruszczyński, Adamiec, Bronikowska, Kieraś, Modrzejewski, Wieczorek, and Woliński 2022).

In this thesis, KorBa supplies the textual and morphosyntactic substrate.

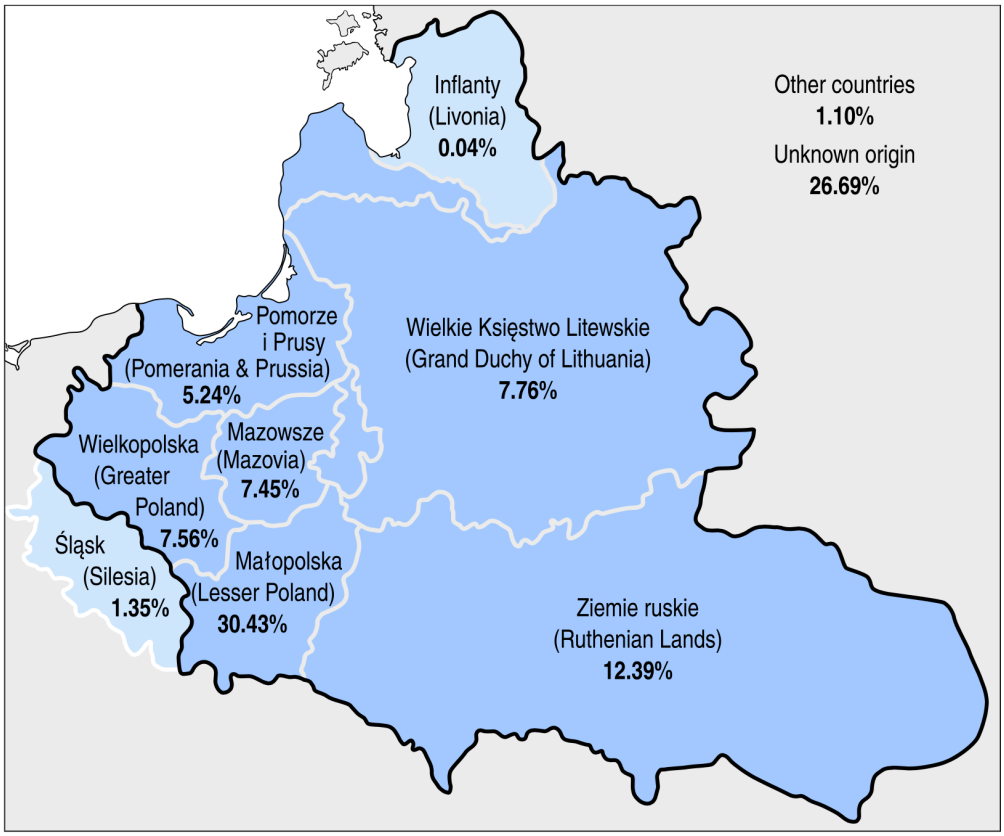


Figure 1: Geographical distribution of texts in the corpus displayed on the map of the Commonwealth after the Union of Lublin of 1569

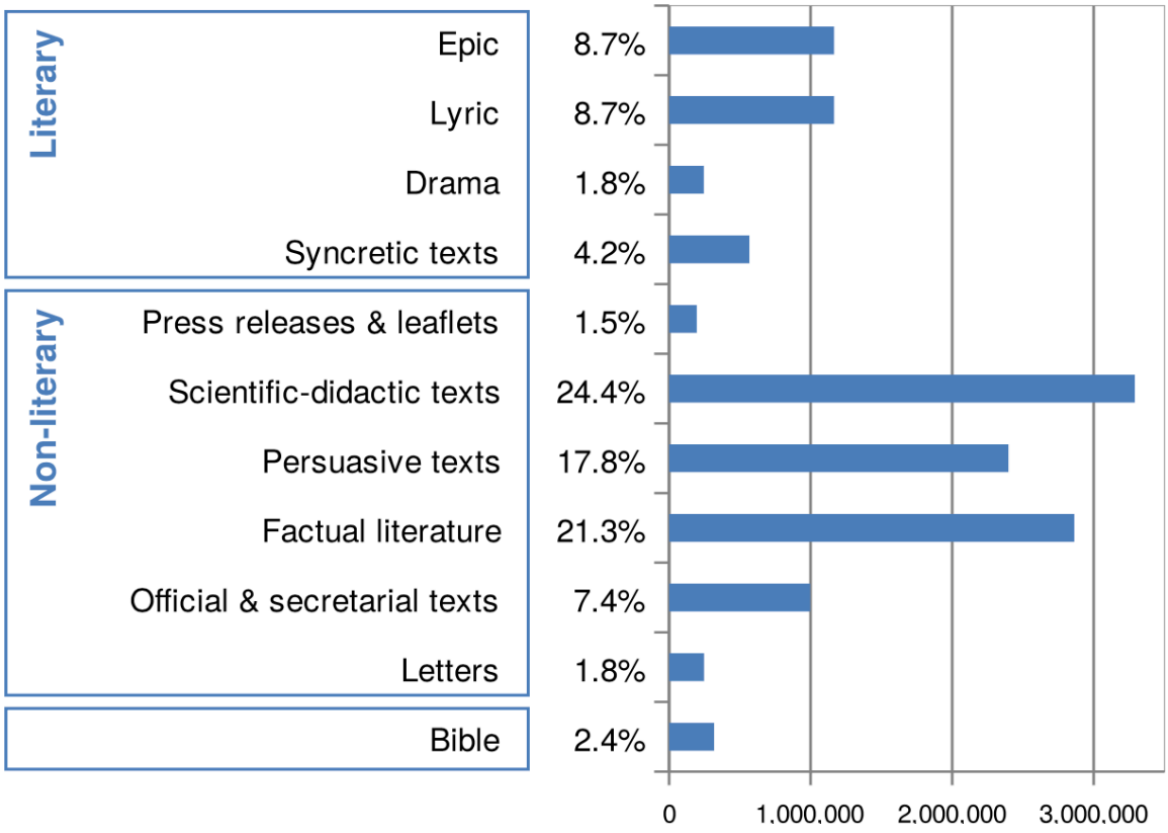


Figure 2: Types of texts in KorBa

Source for Figures 1–2: Gruszczyński, Adamiec, Bronikowska, Kieraś, Modrzejewski, Wieczorek, and Woliński (2022), CC BY 4.0.

2.3.2. MPDT

Middle Polish Dependency Treebank is a syntactically annotated subset of KorBa: it adds a dependency layer on top of KorBa's tokenization, lemmas, POS, and FEATS for selected Middle Polish texts (Wieczorek 2025). The annotation scheme is compatible with PDB conventions, and the resource is under active development (not publicly released at the time of writing).

In this thesis, the converter consumes MPDT data—i.e., KorBa's tokenization, lemmata, POS and features *together with* the MPDT dependency layer—and transforms them to UD (CoNLL-U). Only those KorBa segments that belong to MPDT are converted, since a dependency layer is a prerequisite for UD conversion.

Chapter 3

Linguistic Features of Middle Polish

3.1. Middle Polish: Key Linguistic Features Relevant to Conversion

3.1.1. Orthography and Punctuation

3.1.2. Morphology

3.1.3. Syntax

3.2. Annotation Principles for This Work

3.2.1. Scope and Exclusions

3.2.2. Tokenization and Normalization

3.2.3. Extended POS (ExtPos)

3.3. Summary

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

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. et al. 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. (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. et al. Marseille, France: European Language Resources Association, pp. 5308–5315. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.653/.