

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 wali-dacji. 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	5
1.3. Contributions	6
1.4. Structure of the Document	6
2. Background	8
2.1. Dependency Grammar	8
2.2. Universal Dependencies	8
2.3. Resources	8
2.3.1. KorBa	8
2.3.2. MPDT	10
2.3.3. PDB / PDB-UD (analogy and rule reuse)	10
3. Linguistic Features of Middle Polish	12
3.1. Middle Polish: Key Linguistic Features Relevant to Conversion	12
3.1.1. Orthography and Punctuation	12
3.1.2. Morphology	12
3.1.3. Syntax	12
3.2. Annotation Principles for This Work	12
3.2.1. Scope and Exclusions	12
3.2.2. Tokenization and Normalization	12
3.2.3. Extended POS (ExtPos)	12
3.3. Summary	12
4. Conversion Design and Implementation	13
4.1. Design Overview and Pipeline	13
4.2. POS and Morphological Mapping	13
4.3. Dependency Relation Conversion	13
4.4. Logging, Testing, and Traceability	13
4.5. Processing Workflow	13

5. Validation and Outcomes	14
5.1. Evaluation Data	14
5.2. UD Validation Setup	14
5.3. Results Overview	14
5.4. Qualitative Error Analysis	14
5.5. Known Limitations and Outstanding Issues	14
 6. Applications and Cognitive Science Perspective	15
6.1. Usefulness and Audience	15
6.1.1. Who benefits and how	15
6.1.2. Packaging and License	15
6.1.3. Repository and UD ecosystem integration	15
6.2. Use Cases	15
6.2.1. Historical Syntax and Diachrony	15
6.2.2. Parser Training and Evaluation	15
6.3. Cognitive Science Perspective	15
6.3.1. Processing Constraints	15
6.3.2. Category Change Over Time	15
6.4. Future Work	15
6.4.1. Coverage and Phenomena	15
6.4.2. Generalization and Automation	15

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 et al. 2020). For 17th–18th-century Polish, however, key resources remain outside UD: Middle Polish texts in KorBa (Gruszczyński et al. 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*, *ze*), numeral complexes, and multiword conjunctions/prepositions interact with head rules and label inventories. Prior conversion experience for contemporary Polish (PDB → UD) offers valuable guidance (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 (strict settings) 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 on all of the levels.

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.** Dependency grammar from first principles; UD design; CoNLL-U and validation; a brief note on KorBa, MPDT, and PDB.
- **Chapter 3: Linguistic Features of Middle Polish.** Data and annotation context; orthography, morphology, and syntax relevant to conversion; annotation principles used here.
- **Chapter 4: Conversion Design and Implementation.** Pipeline modules; POS/morph mapping; relation conversion (function words, coordination, copulas, numerals, MWEs); testing, logging, and traceability.

- **Chapter 5: Validation and Outcomes.** Evaluation setup; validator configuration; aggregate results and warning profiles; qualitative error analysis; limitations.
- **Chapter 6: Applications and Cognitive Science Perspective.** Usefulness and audience; packaging and repository integration with UD; exemplar use cases (historical syntax, parser baselines); future work.

Chapter 2

Background

2.1. Dependency Grammar

2.2. Universal Dependencies

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 et al. 2022).

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

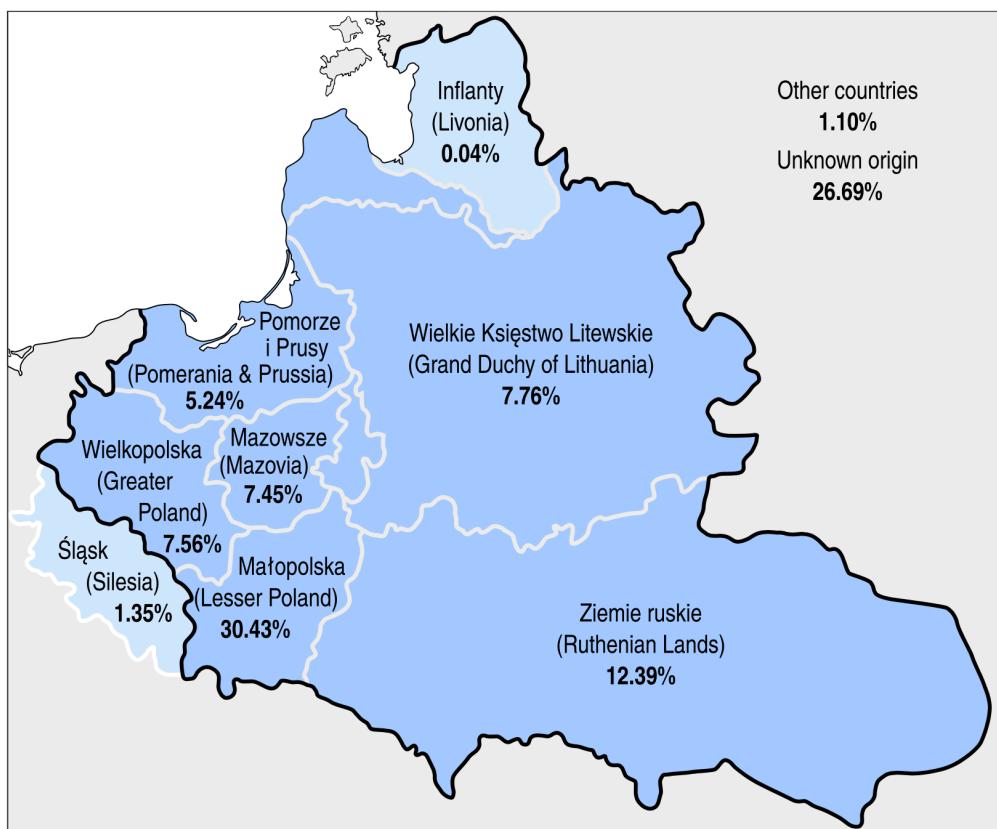


Figure 1: Geographical distribution of texts in the corpus displayed on the map of the Commonwealth after the Union of Lublin of 1569

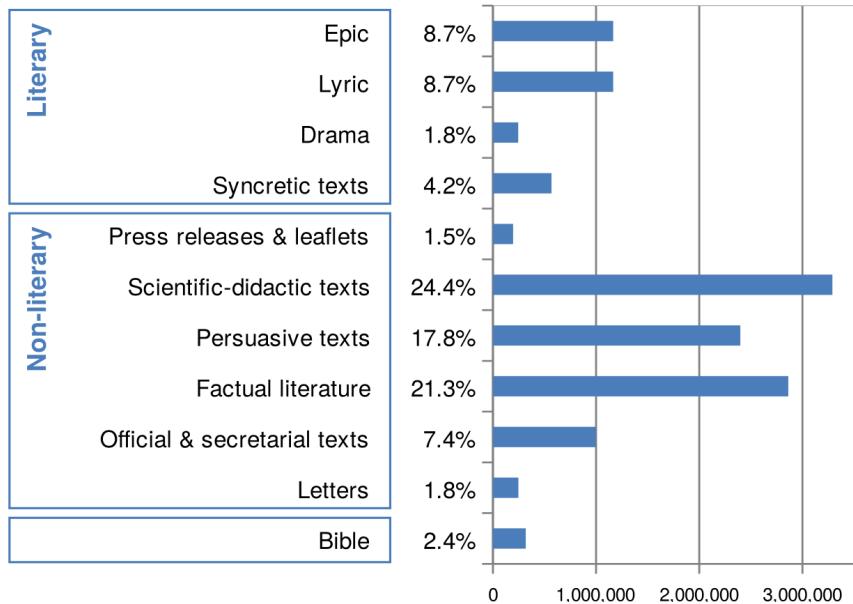


Figure 2: Types of texts

Source for Figures 1–2: Gruszczyński et al. (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, lemmas, 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.

2.3.3. PDB / PDB-UD (analogy and rule reuse)

Polish Dependency Treebank and its UD counterpart (PDB-UD) are not data sources for this work, but they serve as useful analogies. Where appropriate, we reuse or adapt established Polish-specific conversion patterns (e.g., treatment of function words,

coordination, and numerals) described for PDB → UD, tailoring them to Middle Polish phenomena (Wróblewska 2020).

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. et al. (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. et al. (2020). “Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection”. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by N. Calzolari 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. (2020). “Towards the Conversion of National Corpus of Polish to Universal Dependencies”. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by N. Calzolari 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/>.