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

Motivation and Background

1.1. Motivation

1.2. Project Scope

1.3. Objectives and Contributions

1.4. Key Concepts and Resources

1.4.1. Dependency Trees

1.4.2. Universal Dependencies (UD)

1.4.3. Resources Overview (KorBa, MPDT, PDB)

1.5. Structure of the Document

Chapter 2

Linguistic Features of Middle Polish

2.1. Corpora and Annotation Context

2.1.1. KorBa

2.1.2. MPDT (Middle Polish Dependency Treebank)

2.1.3. PDB and Relation to MPDT

2.2. Middle Polish: Key Linguistic Features Relevant to Conversion

2.2.1. Orthography and Punctuation

2.2.2. Morphology

2.2.3. Syntax

2.3. Annotation Principles for This Work

2.3.1. Scope and Exclusions

2.3.2. Tokenization and Normalization

2.3.3. Extended POS (ExtPos)

2.4. Summary

Chapter 3

Conversion Design and Implementation

3.1. Design Overview and Pipeline

3.2. POS and Morphological Mapping

3.3. Dependency Relation Conversion

3.4. Logging, Testing, and Traceability

3.5. Processing Workflow

Chapter 4

Validation and Outcomes

4.1. Evaluation Data

4.2. UD Validation Setup

4.3. Results Overview

4.4. Qualitative Error Analysis

4.5. Known Limitations and Outstanding Issues

Chapter 5

Applications and Cognitive Science Perspective

5.1. Release

5.1.1. Packaging and License

5.1.2. Repository and UD Integration

5.2. Use Cases

5.2.1. Historical Syntax and Diachrony

5.2.2. Parser Training and Evaluation

5.3. Cognitive Science Perspective

5.3.1. Processing Constraints

5.3.2. Category Change Over Time

5.4. Future Work

5.4.1. Coverage and Phenomena

5.4.2. Generalization and Automation

Bibliography

Adamiec, D., W. Gruszczyński, M. Majdak, and M. Żółtak (May 2023). "Barokowa polszczyzna w internecie, czyli Elektroniczny słownik języka polskiego XVII i XVIII wieku". In: *LingVaria* 18.1(35), pp. 113–124. DOI: 10.12797/LV.18.2023.35.08. URL: https://journals.akademicka.pl/lv/article/view/5082.

Bronikowska, R., W. Gruszczyński, M. Ogrodniczuk, and M. Woliński (2016). "The Use of Electronic Historical Dictionary Data in Corpus Design". In: *Studies in Polish Linguistics* 11.2, pp. 47–56. ISSN: 1732-8160. DOI: 10.4467/23005920SPL.16.003.4818. URL: https://ejournals.eu/en/journal/studies-in-polish-linguistics/article/the-use-of-electronic-historical-dictionary-data-in-corpus-design.

Bronikowska, R. and K. Kryńska (Dec. 2020). "Łacina w KorBie. Użyteczność Elektronicznego Korpusu Tekstów Polskich XVII i XVIII wieku dla filologa neolatynisty". In: *Polonica* 40.1. DOI: 10.17651/POLON.40.8. URL: https://polonica.ijppan.pl/index.php/polonica/article/view/157.

Gruszczyński, W., D. Adamiec, R. Bronikowska, W. Kieraś, et al. (Mar. 1, 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.

Gruszczyński, W., D. Adamiec, R. Bronikowska, and A. Wieczorek (2020). "Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. – problemy teoretyczne i warsztatowe". In: *Poradnik Językowy* 8, pp. 32–51.

Instrukcja korzystania z wyszukiwarki do Elektronicznego Korpusu Tekstów Polskich z XVII i XVIII w. (do 1772 r.) (Jan. 2024). Zespół KorBa.

Nivre, J. et al. (May 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. (2020). *Instrukcja anotowania drzew zależnościowych: Dodatek 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.

Wieczorek, A. and A. Wróblewska (n.d.). "Middle Polish Dependency Treebank and its conversion to the UD format". Unpublished manuscript.

Wróblewska, A. (May 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/.

— (June 23, 2023). *Instrukcja anotowania drzew w Polskim Banku Drzew Zależnościowych*. Tech. rep. Institute of Computer Science, Polish Academy of Sciences.

Zespół KorBa (2025). *KorBa: Elektroniczny korpus tekstów polskich XVII i XVIII wieku — o korpusie/overview*. O korpusie. Polish. URL: https://korba.edu.pl/overview (visited on 08/27/2025).