Journal of Open Humanities Data

PrevHist: A Dataset of Old and Middle Hungarian Preverb Constructions

DATA PAPER

ÁGNES KALIVODA

ubiquity press

ABSTRACT

This data paper presents a dataset of Old and Middle Hungarian preverb-verb constructions, covering a time span from the late twelfth century to 1772. The resource consists of 68,458 records. It was extracted automatically from two corpora: the Old Hungarian Corpus and the Old and Middle Hungarian corpus of informal language. Each record is annotated for a wide range of morphosyntactic features and metadata, notably the year of writing and the register. The dataset can be used to explore the diachronic trajectory of preverb-verb constructions and, eventually, to gain a better understanding of preverbs' grammaticalization process in Hungarian.

CORRESPONDING AUTHOR:

Ágnes Kalivoda

Institute for Lexicology, ELTE Research Centre for Linguistics, Budapest, Hungary

kalivoda.agnes@nytud.elte.hu

KEYWORDS:

Old Hungarian; Middle Hungarian; preverb-verb constructions; preverb; verbal particle

TO CITE THIS ARTICLE:

Kalivoda, Á. (2026). PrevHist: A Dataset of Old and Middle Hungarian Preverb Constructions. *Journal of Open Humanities Data*, 12: 95, pp. 1–5. DOI: https://doi.org/10.5334/johd.582

(1) OVERVIEW

REPOSITORY LOCATION

DOI: https://doi.org/10.5281/zenodo.19927110

CONTEXT

Present-day Hungarian has a prolific system of complex predicate formation combining a separable preverb and a verb. However, this has not always been the case. In the earliest surviving texts of Old Hungarian, only a few lexical items function as preverbs – e.g. *meg* 'perfective', *le* 'down', *fel* 'up' – and their frequency is vanishingly low compared to the present state. The preverb system's expansion was probably accelerated toward the end of the Old Hungarian period (É. Kiss, 2014, p. 70) and it is still an ongoing process.

The dataset presented here is a collection of Old and Middle Hungarian preverb-verb constructions, suitable for both large-scale quantitative and fine-grained qualitative studies of their diachronic trajectory. It was produced within the research project *The emergence of preverb constructions in Hungarian: a corpus-driven approach*.

(2) METHOD

STEPS

The source files were obtained from the Old Hungarian Corpus (OHC) (Simon, 2014)[1] and the Old and Middle Hungarian corpus of informal language (MHC) (Novák, Gugán, Varga, & Dömötör, 2018).[2] The first corpus (OHC) is a comprehensive sample of the Old Hungarian period (896–1526), later extended with Middle Hungarian Bible translations; its texts are almost all religious literature. Of its three layers (raw text in original orthography, 3.2 million tokens; manually normalized text, 1.3 million; and text that is both normalized and morphosyntactically annotated, ca. 300 thousand), the present dataset was built from the last, since morphosyntactic annotation was crucial for finding the relevant constructions reliably. The second corpus (MHC) aims at approximating the vernacular; in practice it is Middle Hungarian (1526–1772), focused on informal text types, primarily court records of witch trials and private letters. It is normalized and annotated throughout, and could therefore be used in this project in its entirety. It has a size of ca. 1.5 million tokens.

The dataset design considerations and the corpus processing workflow were inherited from the PrevDistro (Preverb Distributions) project (Kalivoda, 2022).[3] Considerable fine-tuning was nonetheless required for the two historical corpora and for phenomena absent from present-day Hungarian, most notably complex past tenses.

The dataset was produced following these steps (each time using custom Python 3.8 scripts)[4]:

1. Harmonizing the two corpora structurally. OHC and MHC use almost identical normalization principles and morphosyntactic annotation, but OHC consists of TSV files while MHC is stored as a MySQL database, with annotation layers and metadata held in separate tables. Both were converted to a common four-column TSV structure (original form, normalized form, lemma, morphological analysis). For OHC, only the morphologically annotated files were kept; rows tagged as errors, fragments, or foreign-language material, as well as Latin passages in MHC, were discarded. Both corpora were re-tokenized so that punctuation and the clitic question particle *-e* form separate tokens, and a unified metadata file was compiled.
2. Obtaining all clauses likely to contain a preverb-verb combination, whether as a finite verb, an infinitive, or a verbal derivative. Preverbs inside derivations are not marked as such in the source annotation, so they were located by a string-based search: lemmas were matched against a hand-curated list of preverbs and their variants (compiled for

1 https://omagyarkorpusz.nytud.hu/en-search.html (last accessed: 2026-04-30).

2 https://tmk.nytud.hu/3/ (last accessed: 2026-04-30).

3 https://zenodo.org/records/6349410 (last accessed: 2026-04-30).

4 The scripts are available at https://github.com/kagnes/prevhist_scripts (last accessed: 2026-06-06).

DOI: 10.5334/johd.582

this project), and lemmas containing a derivational boundary were split into a candidate preverb and a stem. The morphological analyses already present in the corpora – which follow the emMorph annotation scheme (Novák, Siklósi, & Oravecz, 2016), whose coding distinguishes the relevant derivational categories – were then used to assign each candidate a construction type, with the word ending serving as a fallback where the analysis was insufficient. A candidate was retained only if it received a recognized construction type and its preverb matched the curated list; this step was fully automated, with no manual annotation.

3. Finding the associated verb for detached preverbs and extracting distributional information; the overall strategy follows Kalivoda (2022). For each detached preverb, the clause was scanned outward in both directions for a verb to which it could belong: a verb in the right context yields a discontinuous order, one in the left context an inverted order. Candidates were validated automatically against their morphological analysis and the intervening material, allowing only short adverbial or pronominal sequences between preverb and verb and rejecting candidates separated by punctuation or a clause boundary. Complex past tenses and copular constructions, which are absent from present-day Hungarian, were handled explicitly so that the preverb is linked to the full verbal complex. For each accepted construction, the verb lemma, construction type and subtype, preverb position, intervening words, and the keyword-in-context concordance (original and normalized) were recorded.
4. Extracting all available metadata for each valid hit.
5. Compiling the dataset into an easy-to-use TSV format. A detailed description of the extracted linguistic features and metadata can be found in the repository of the dataset.

QUALITY CONTROL

Inconsistencies in the lemmatized forms of verbs and preverbs were identified by inspecting frequency lists of lemma variants. For example, slightly different forms of the same preverb appeared as different lemmas (*bele, belé, bel*', each of these meaning 'into'). These were re-normalized to the most frequent variant (in the latter example, it was *bele* 'into'). The same applies to the most common verbs where the frequency of each lemma variant was above 50 occurrences.

In the case of detached preverbs, a random sample of 50 hits for each preverb position was selected and checked manually for potential systematic errors (or the whole set in case of preverb positions with fewer than 50 hits). All inspection was carried out by the author, the dataset's sole creator, who checked in each sampled case whether the detached preverb had been linked to its correct associated verb; the recurring error types found this way were then filtered automatically with regular expressions.

It must be noted, however, that accounting for annotation errors in the original corpora was not feasible, and the resulting dataset might have inherited these errors. That is, the dataset has undergone substantial quality control, but it is not a gold standard quality.

(3) DATASET DESCRIPTION

REPOSITORY NAME

Zenodo

OBJECT NAME

PrevHist

FORMAT NAMES AND VERSIONS

TSV file with Unicode (UTF-8) character encoding and Unix/Linux (LF) line endings.

CREATION DATES

2025-10-01 – 2026-04-30.

DATASET CREATORS

The dataset was designed and created exclusively by the author.

LANGUAGE

Data: Old and Middle Hungarian. Metadata: English and Hungarian.

LICENSE

PUBLICATION DATE

2026-04-30.

(4) REUSE POTENTIAL

The dataset is primarily intended for researchers working on the historical morphosyntax of Hungarian. Historical and diachronic linguists can trace how preverb-verb constructions evolved over approximately six centuries, examining how variables such as word order, preverb-verb stem distance, construction type, and intervening elements change over time. The annotation of intervening words is especially promising, allowing researchers to ask whether lightweight adverbs or heavier nominal constituents tend to intervene between preverb and verb. Uralists and typologists may also find the data relevant for comparative work on the grammaticalization of preverbs beyond the Uralic family.

The dataset shares PrevDistro's design principles and variable names wherever possible, so the two can be aggregated by researchers extending their analysis into present-day Hungarian. Early modern Hungarian remains unrepresented, however, so a substantial period would appear as a gap in any such aggregated analysis.

The main limitations concern the comparability of data across the two source corpora. The OHC is relatively small and composed almost exclusively of formal, religious texts, whereas the MHC is much larger, drawn from numerous texts across several decades, and aims at approximating the vernacular. This asymmetry in size, source diversity, and register makes direct cross-period comparison non-trivial, so researchers should be cautious about conflating language change with corpus composition. A further limitation concerns granularity: each record links back to its source through a corpus and document identifier together with the stored concordance, but not through a sentence- or token-level pointer. Finer locators were not added, since the re-tokenized source texts cannot be redistributed and a pointer into them would not help users lacking access to the corpora. These limitations reflect the historical record and the current state of Hungarian historical corpora rather than the dataset itself.

AI DECLARATION

No Generative AI tool was used at any stage of this project.

ACKNOWLEDGEMENTS

I would like to thank the Ministry of Culture and Innovation of Hungary for funding my project, and the Institute for Language Technologies and Applied Linguistics at the ELTE Research Centre for Linguistics for granting me direct access to the source files of the two corpora.

FUNDING STATEMENT

This project has been supported by the OTKA PD project No. 142317 funded by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the PD 22 funding scheme.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR CONTRIBUTIONS

Ágnes Kalivoda: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Software, Validation, Writing – original draft, Writing – review & editing.

AUTHOR AFFILIATIONS

Ágnes Kalivoda orcid.org/0000-0003-2520-5523
Institute for Lexicology, ELTE Research Centre for Linguistics, Budapest, Hungary

REFERENCES

É. Kiss, K. (2014). Az ómagyar igeidőrendszer [The tense system in Old Hungarian]. In K. É. Kiss (Ed.), *Magyar generatív történeti mondattan* (pp. 60–72). Akadémiai Kiadó.

Kalivoda, Á. (2022). PrevDistro: An open-access dataset of Hungarian preverb constructions. *Acta Linguistica Academica*, *69*(4), 549–563. https://doi.org/10.1556/2062.2022.00578

Novák, A., Gugán, K., Varga, M., & **Dömötör, A.** (2018). Creation of an annotated corpus of Old and Middle Hungarian court records and private correspondence. *Language Resources and Evaluation*, *52*(1), 1–28. https://doi.org/10.1007/s10579-017-9393-8

Novák, A., Siklósi, B., & **Oravecz, Cs.** (2016). A New Integrated Open-source Morphological Analyzer for Hungarian. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16) (pp. 1315–1322). European Language Resources Association (ELRA). https://aclanthology.org/L16-1209/

Simon, E. (2014). Corpus Building from Old Hungarian Codices. In K. É. Kiss (Ed.), *The evolution of functional left peripheries in Hungarian syntax* (pp. 224–236). Oxford University Press.

TO CITE THIS ARTICLE:
Kalivoda, Á. (2026). PrevHist: A Dataset of Old and Middle Hungarian Preverb Constructions. *Journal of Open Humanities Data*, 12: 95, pp. 1–5. DOI: https://doi.org/10.5334/johd.582

Submitted: 30 April 2026
Accepted: 08 June 2026
Published: 16 July 2026

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.