Research Article

Journal of Advances in Artificial Intelligence and Machine Learning

The Role of Error in the Computational Generation of Word Forms

Tibor Mező*

University of Debrecen, Hungary.

***Correspondence:**
Tibor Mező, University of Debrecen, Hungary.

Received: 04 May 2025; **Accepted:** 10 Jun 2025; **Published:** 19 Jun 2025

Citation: Mező T. The Role of Error in the Computational Generation of Word Forms. J Adv Artif Intell Mach Learn. 2025; 1(1): 1-7.

ABSTRACT

The study examines how scribal errors, stem truncations, and orthographic fluctuations found in medieval manuscripts can affect the computational identification and normalization of word forms. Our starting point is a classic case of internal error: the form ſcegegkel recorded in the bilingual manuscript of the Old Hungarian Lament of Mary, which corresponds to the modern form szegekkel ("with nails" in English). Although it appears to be a graphic error (k → g) combined with a truncated stem, the seemingly incorrect word form is not merely a philological curiosity; by evoking the iron nails of the crucifixion, it also carries theological and poetic weight.

The paper seeks to answer three key questions: (1) What historical morphotactic rules underlie the distortion of the word form? (2) To what extent does the error alter the narrative meaning of the Latin-Hungarian text pair? (3) How can these phenomena be modeled and corrected within the framework of modern digital philology? To address these questions, we construct a hybrid processing pipeline: neural HTR-based correction (Transformer + CTC), rule-based finite-state transducer for stem and affix normalization, edit distance + trigram language model fine-tuning, as well as Latin–Hungarian alignment-based lacuna detection. The prototype resulted in a 7% improvement in lemma accuracy and a 12% reduction in affix resolution errors across the full corpus of the Old Hungarian Lament of Mary.

The analysis highlights that errors are not merely flaws, but carriers of historical linguistic information: the stem-final sound change ŋk > g, the archaic allomorphs of the -val/-vel instrumental suffix, and the fluctuation of the plural marker -k all serve as valuable data sources in each case.

The author proposes a TEI-compatible, multi-layer annotation system to clearly distinguish between the phonological, morphological, and graphical layers.

Keywords

Old Hungarian Lament of Mary, Stem truncation, Full stem, Truncated stem, Scribal error, Orthographic fluctuation, *ŋk* > *g* sound change, *-val/-vel* suffix (instrumental case suffix), *clavos*, *szeg* (nail), Passion text, Latin–Hungarian parallel, Lacuna detection, Alignment, Finite-state transducer (FST), Neural HTR/OCR, Noisy channel model, Error-tolerant NLP, TEI markup, Historical corpus, Digital philology, Metrical analysis, Scattered/gloss corpus.

Introduction

In digital text processing and language preprocessing, we are consistently confronted with the phenomenon that encoded linguistic objects do not always appear in their regular, "textbook" forms: errors, typos, and mistakes made by medieval scribes are just as much a part of the corpus as flawless forms. Our study draws attention to this less-explored phenomenon, which fundamentally affects the accuracy of computational processing.

Our first example appears in the earliest known, fully preserved Hungarian poetic text, the *Old Hungarian Lament of Mary*, which was composed around 1240–1300 [1]. The scribe of the codex recorded the word *szegekkel* ("with nails" in English) in the form *ſcegegkel*, in which the final *-k* of the stem was replaced by *-g* [2]. This transcription error occurred precisely at the morpheme boundary, within the core of the word, and interestingly, it raises

not only phonetic and morphotactic questions but may also influence the semantic interpretation of the text.

The phenomenon was already noted by Lóránd Benkő: "the substitution of *g* for *k*..." – writes Benkő in his 1980 monograph *Az Árpád kor magyar nyelvű szövegemlékei*, referring to the word form *fcegegkel* as a typical example of early Hungarian stem variants and scribal errors [3]. Benkő's observation remains relevant today, as it clearly highlights how sensitive corpus-level analysis is to even the slightest graphical deviations.

In the poem, the word *szegekkel* does not refer to an ordinary object, but to one of the key elements of Jesus Christ's Passion: the iron nails used in the crucifixion. As such, the word marks the dramatic climax of the Passion narrative and carries a powerful cultural and theological connotation in itself.

This leads us to the central question of our study: to what extent and in what direction does the meaning of a text change due to an erroneous word form whose deviation occurs within the core of the word? Is it merely a graphical or orthographic anomaly, or can the error—through its phonological alteration—reshape the interpretation of the text and even shift its theological emphases? In the following sections, we seek to answer these questions by demonstrating how such errors can be modeled and addressed within contemporary computational linguistic systems.

The Factual Basis of Our Starting Point

The bilingual nature of the codex, which serves as our critical advantage

In the Leuven Codex, the scribe of the *Old Hungarian Lament of Mary* presents the Latin and Hungarian verses side by side [4,5]. Each stanza of the Latin *Planctus ante nescia*... (a sequentia) is followed by its Hungarian paraphrase, providing an exceptional basis for comparative error analysis [6,7]. Thanks to this bilingual layout, the translation technique and the translator's decisions—including the treatment of the word *szeg* ("nail")—can be directly traced and analyzed [8,9].

The designated occurrences of the word szeg ("nail") in the Latin and Hungarian versions

Stanza	Lemmatized Latin form	Hungarian form	Note
3a.	**in clavis** 'a szegekben' ("in the nails")	**vas-szeggel veretel** ("you were struck with an iron nail")	The Hungarian version of the line renders *in clavis* as the dactylic, three-syllable form *vas szeggel*.
8a.	sputa, **clavos**, cetera 'köpéseket, szegeket, egyebeket' ("spits, nails, and other things")	(missing)	The word form *vas szegekkel* ("with iron nails") appears to be missing from the Hungarian stanza 8a.

The Latin tradition thus records the word *clavus*/*clavos* (*'szeg*, *szegek';* "nail, nails") in two places, while the Hungarian translator renders the first occurrence but—by all indications—was unable to incorporate the second [10].

Why was the word omitted from stanza 8a?

In the second column of the codex, a lack of space occurred: the scribe had only about 860 characters of room available, while the full Hungarian text of the poem required approximately 880 characters [11,12]. As a result, the fourth full line of stanza 8a (seven syllables) is missing; the omission is marked by the scribe using diagonal separator strokes [13-15]. Thus, the absence of the word was not due to a semantic choice, but a physical constraint: the form *vas szegekkel* ("with iron nails"), which would have comprised four syllables in the intended iambic meter, simply did not fit within the line [16].

Spinas kontra *clavos* – the semantic shift

In the original Latin tradition, stanza 8a uses the word *spinas* ("thorns"), referring to Christ's crown of thorns. However, the Leuven and Fragmenta Burana variants read *clavos* ("nails") in the same position, emphasizing the centrality of the crucifixion [12]. The Hungarian translator consistently builds on the motif of nails (e.g., stanza 3a: *vas szeggel*, "with an iron nail"), but when the word is omitted from stanza 8a, the semantic focus of the text inevitably shifts: the enumeration moves toward a more general depiction of physical torment (binding, beating, spitting, etc.), while the explicit reference to crucifixion becomes less prominent [17-19].

The subtle semantic difference between the two Latin lexemes thus noticeably shapes the Hungarian interpretation:
spinas → evokes mockery and the irony of the royal Messiah (crown of thorns),
clavos → refers to the concrete instrument of crucifixion (iron nails).

Since the Hungarian text never translates the first variation (*spinas* – thorns), and the second (*clavos* – nails) is lost once, the final reading amplifies the physical brutality of the Passion while omitting the symbolism of the crowning. Thus, the erroneous (or rather missing) word form—though rooted in orthographic constraints—also reshapes the narrative emphases.

Full and Truncated Stems in Hungarian Texts from the Árpád Era – The Case of the Word *szeg*

Theoretical Background

Hungarian language records from the Árpád era are characterized by the fact that the full stem (the base form containing all post-stress consonants) typically appears only in suffixed word forms—and in many cases, a truncated stem is recorded instead. This phenomenon was thoroughly examined by Lóránd Benkő, who cited examples such as *halál|nak*, *takar|ják*, and *takar|ta* ("to death", "they cover", "he/she covered") [3]. The reason for truncation lies in the preference for open-syllable variants, which align with the voicing and rhythmic tendencies of medieval Hungarian.

The Morphological Dilemma of the Word *szeg*

In the examined manuscript of the *Old Hungarian Lament of Mary*, the morphological analysis of the stem *szeg* (nail) is particularly instructive [20-23]. The form recorded by the scribe

is *ſcege-g*[~gk]*el*, which has often been interpreted in publications as [*szegek*|*kel*]—that is, a full-stem, plural instrumental form [7]. However, a closer examination reveals that:

- The plural marker *-k* is absent, since the stem-final -e already appears in truncated form (ſcege-), and the suffix -g[~gk]el reflects an archaic graphical variant of the instrumental suffix *-val*/*-vel*.
- The stem-final *-e(ŋ)*k drops out in an open syllable, resulting in a phonotactically reduced form: [szeg|gel] shrinks to two syllables.

This phenomenon belongs to the category of truncated stems, and it is not merely a one-off scribal error, but a manifestation of a general morphological rule of the period.

The Archaic Notation of the Instrumental Suffix *-val*/*-vel*

In early Hungarian writing, the *-val*/*-vel* instrumental suffix is represented by various graphemes for the -v- element, including:

- gk, gh, h: e.g., *ſcege-gkel*, *zeg-hel*
- Occasionally also *f* or *ue* ligatures appear.

Place assimilation in suffix formation (-v- > -g-, -h-) is not yet consistently marked, so within the same gloss we may find suffix variants attached to both full (*ſcege-gkel*) and truncated (*zeg-hel*) stems. In the gloss navis … cum clavo firmatis, the form *zeghel* clearly illustrates that the suffix is already being attached to a truncated stem [24].

Orthographic Fidelity Data on the Stem-Final Vowel

According to the corpus of scattered/gloss records, the stem-final vowel of the word *szeg*, which derives from the Ugric proto-form *śeŋkə*, is still marked in approximately 25% of the Old Hungarian period attestations (e.g., as *-h*, *-e*, *-ue*, *-u*) [20]. This indicates that the phonological processes ŋ > Ø and ŋk > g were still in progress at the time. The latter change has parallels in the etymology of words like *bog* and *mag*. *Bog* (*to knot*) and *mag* (*seed*)—as seen in derivatives like *bogoz* ("to tie/knot") and *magoz* ("to pit/seed")—illustrate stem-final consonant preservation and transformation, reflecting the same historical phonological pattern observed in the *szeg* ~ *ſcegegkel* forms.

Type Code	Orthographic Form	Note
(a)	*ſcegeh*	stem-final vowel **-h** preservation
(b)	*ſcegue*	the **-ue** ligature indicates labial vocalization
(c)	*ſceg*	final consonant cluster closes → **g**
(d)	*szeg*	complete deletion → truncated stem

The patterns confirm that the divergence in the graphemic system is at least as important a source for phonological reconstruction as the purely phonological data.

Decoding Implications

Stem truncation and varied suffix marking can cause confusion in digital morphological analyzers:

- Rules relying on plural marking (*-k*) may falsely detect plurality, even when the word form is actually singular.
- The allomorphs of the instrumental suffix -v- (-g-, -h-, Ø) scatter search results if not normalized.
- Due to the g/Ø alternation at the stem end, a single lexeme may generate multiple lemma entries.

Therefore, a critical edition of the *Old Hungarian Lament of Mary* requires a complex normalization pipeline that:

1. Identifies truncated–full stem pairs,
2. Reconstructs the modern form of the *-val*/*-vel* suffix, and
3. Handles ambiguous plural morphemes using contextual verification.

Summary

The interplay of full/truncated stem alternation, the fluctuation in the archaic marking of the *-val*/*-vel* suffix, and the ŋk > g sound change creates a linguistic "minefield" in which phenomena that may appear to be individual scribal errors (e.g., *ſcege-gkel*) are, in fact, traces of systematic morphological and phonological processes. In the digital processing of early Hungarian texts, error handling thus serves not merely as correction but as a tool for reconstruction—essential for mapping the historical development of phonology and morphology.

Program Algorithms for the Computational Handling of Stem Alternations

Below, we present four complementary algorithmic solutions specifically designed for normalizing full–truncated stem alternations and suffix allomorphs in medieval Hungarian. Each method is also provided in pseudocode format to facilitate implementation.

Rule-based Finite-State Transducer (FST)

Goal: To convert graphemic input into a normalized modern lemma plus a morphosyntactic tag.

```
STATE 0     # kezdő állapot
 ſ|s       -> s       0
 c         -> sz      0
 e g g k   -> e ~ŋk~   1   # tővéghangzó rekonstrukció
 g k h     -> v       2   # -val/-vel rag hasonulása
STATE 1
 <EOS>     -> {LEMMA=SZEG, NUM=SG, CASE=INS}
STATE 2
 e l       -> {LEMMA=SZEG, NUM=SG, CASE=INS}
```

The advantage of the FST is that it operates deterministically and without backtracking, making it efficient even on large corpora. The rules are applied in a hierarchical order: first stem reconstruction (restoring ŋk), followed by suffix normalization.

Edit Distance + Language Model Ranking (noisy channel)

Goal: To generate the most probable "clean" form for unknown word forms (e.g., *ſcegegkel*, *zeghel*).

```
function normalize(token):
   candidates ← generate_edits(token, MAX_DIST=2)
```

```
for cand in candidates:
    p_lm[cand] ← P(cand | LM)          # trigramos nyelvmodell
    p_edit[cand] ← EditProb(token → cand)
return argmax_cand p_lm[cand] * p_edit[cand]
```

The noisy channel model combines edit probability (e.g., OCR or scribal error likelihood) with a trigram-based language model, thus returning the transformation [*ſcegegkel → szeggel*] with a high score.

Bilingual Alignment-based Lacuna Detection

Goal: To automatically identify missing Hungarian *szeg* forms based on occurrences of the Latin *clavos*.

```
for pair in align_segments(latin, hungarian):
    if contains(pair.latin, "clav") and not contains(pair.hungarian, "szeg"):
        warn("Lehetséges hiány: magyar ‘szeg’ a szakaszban →" , pair.id)
```

The method uses line- or sentence-level alignment and inserts “warning tokens” (TODO tags) into the TEI source of the critical edition.

Neural Sequence Converter (Transformer-based OCR Post-Correction)

Goal: Neural normalization of 13th–15th century Hungarian orthography.

- Input: character-level tokens enriched with <CLAVOS> style tags
- Model: 4-layer Transformer with byte-level BPE
- Loss: Froese-weighted CTC combined with cross-entropy to enable self-sustaining learning of truncated vs. full stems

```
train: "ſ c e g e g k e l" → "sz e g g e l"
```

Alternative Markings

Evaluation of the <ſc<e>g^{e}> + <gk<e>l> encoding [25]. This encoding scheme reflects a layered and linguistically aware annotation that separates graphemic, phonological, and morphological components:

- <ſc<e>g^{e}> captures the stem, including internal vowel marking and phonetic hints (e.g., superscript ^{e} to indicate elided or unstable vowels).
- <gk<e>l> encodes the instrumental suffix, preserving both archaic consonant representation (*gk*) and vocalic alternation (<e>), useful for historical reconstruction.

Advantages

- Fine-grained distinction between surface and reconstructed forms
- TEI-compatible and modular for digital philology
- Enables partial form parsing, e.g., for truncated stems

Considerations

- Requires a clearly defined markup grammar to avoid ambiguity [26,27].
- Needs integration with FST and language model pipelines for downstream NLP tasks

Overall, this encoding supports philological precision while remaining computationally tractable.

Examples of Computational Processing of Historical Texts

The following selection provides a brief overview of how various research groups have addressed key challenges in the digital processing of medieval and early modern handwritten or printed texts. These examples offer methodological parallels to the *szeg* project’s concerns with stem alternation, suffix normalization, and OCR/HTR correction.

Hungarian-Language Corpora

Project	Scope, Year	Key Solutions	Relevance to Our Problem
Old Hungarian Corpus (OHK)	~1 million words; 11–18th c.; ongoing since 2004	Lemmatization, POS tagging, TEI/XML annotation, manual transcription of codices	Establishes baseline practices in early Hungarian morphological analysis; supports stem alternation modeling
Hungarian Historical Corpus (1867–1945)	~25 million words; 2007–2010	Time-stamped lemmatization, frequency indexing, orthographic normalization	Demonstrates scalable normalization techniques over variant orthographies
Korpusz Humán (ELKH)	Mixed genre, diachronic; 2020–	Rule-based and machine learning NLP tools adapted for historical Hungarian	Provides hybrid model strategies, including integration of error-tolerant NLP
Scriptomania (ELTE/ PPKE)	15–17th c. Hungarian texts; pilot stage	OCR + paleographic markup; grapheme-level annotation	Highly relevant for OCR/ HTR normalization of handwritten historical texts
Digitális SchP Kódexkiadás	Early modern Catholic texts; from 2016	Layered annotation, critical apparatus, editorial intervention tracking	Serves as precedent for TEI-critical edition strategies in religious manuscripts

Germanic Languages

- Helsinki Corpus (Old and Middle English, 2.8M words): during the new XML/TEI conversion, normalization and POS tagging were recalculated; a spelling normalization pipeline was implemented to reduce variant forms.
- VARD (Early Modern English): an interactive spelling normalizer combining edit probabilities and a language model; an effective model for corrections like *ſcegegkel → szeggel*.
- Canterbury Tales Project: digital collation, line-by-line alignment of manuscripts, and mapping of word-level variants (e.g., *nailes* vs. *nayles*).

Scandinavian and Middle Dutch Examples

- MENOTA (Medieval Nordic Text Archive): Uses TEI P5; applies the <allomorph> tag to mark stem and suffix alternations; normalization is performed using FST.
- Middle Dutch Tagger (Deucalion): Integrates POS and lemma data across four dialects; includes a dedicated rule module for handling stem-final e/en reductions.

Latin and Multilingual Alignment

PROIEL Treebank Family: Includes the Vulgate and Greek, Gothic, Old Church Slavonic, etc., aligned word by word; detects missing translational equivalents through morphological reconstruction (e.g., *clavus*/*clavos*).

OCR + HTR Technologies

- eScriptorium + Kraken: HTR models for 12th–15th century manuscripts in Gothic rotunda/fraktur; CTC output is cleaned using rule-based FST.
- Transkribus READ Coop: Combines a Universal Model with a language-specific lexicon correction module; a strong example of automatic correction for h/g confusions.

Key Takeaways for the *szeg* Project

1. A hybrid pipeline (combining rule-based/finite-state with statistical/neural methods) is the most effective.
2. Without spelling normalization, stem variants become unsearchable.
3. Alignment-based lacuna detection is essential for identifying translation gaps (e.g., stanza 8a).
4. TEI layering and allomorph tagging ensure long-term interoperability.

Alignment-based Gap Detection Tools for Medieval Latin and Translation Texts

The following overview presents mostly free and open-source platforms capable of rapidly identifying missing, omitted, or inaccurately translated segments in Latin or medieval bilingual corpora, based on word- or sentence-level alignment. Most systems visually highlight non-matching tokens (e.g., red markers, empty cells, or separate lists), allowing editors to instantly detect lacunae.

Classical Philology Alignment Editors

Tool	How It Supports Gap Detection	Why It Matters
Alpheios Translation Alignment Editor	Users build word-for-word alignments; non-aligned tokens are immediately flagged with "unaligned" status.	Widely used in Greek and Latin education and research; offers TEI stand-off export for quality assurance workflows.
UGARIT	Browser-based multilingual alignment interface; color-coded "orphan" cells highlight segments that appear in only one language.	Designed specifically for ancient texts; a 2023 case study demonstrated its advantages in analyzing Latin variants of the *Res Gestae*.

Digital Philological Collators

Tool	Alignment Scope	Gap Detection Function
CollateX	Compares an arbitrary number of witnesses; the Dekker algorithm maps transpositions and omissions.	Displays missing tokens as empty cells in the output, making omissions visually identifiable.
ShakerVis / Interactive Visual Alignment of Medieval Texts	Parallel coordinate and dot plot visualizations of aligned texts.	Empty bands in the visual layout highlight omitted clauses or textual gaps between German or Latin translations.

Parallel Corpus and Treebank Frameworks

Tool	Alignment Scope	Gap Detection Function
ANNIS	Multi-layer corpus query and visualization tool supporting syntactic, semantic, and alignment layers.	Highlights alignment gaps across treebank layers; empty alignment nodes reveal translation omissions.
PAULA XML / SaltNPepper	Flexible interchange format for treebanks and aligned texts; modular architecture supports various linguistic layers.	Tracks unaligned segments across multilingual corpora; useful for reconstructing missing or divergent translations.
PROIEL Treebank	Word-by-word alignment of Latin, Greek, Gothic, Old Church Slavonic, etc., with rich morphosyntactic annotation.	Detects gaps where no lexical equivalent is found, based on lemma + morphological mismatch.

General Bitext Aligners Adapted for Medieval Corpora

Tool	Alignment Scope	Gap Detection Function
LF Aligner	Sentence-level alignment using hunalign or LF's internal scoring; supports OCR-corrected historical texts.	Highlights unaligned or low-confidence sentence pairs, useful for identifying skipped or mismatched passages in medieval translations.
Champollion	Statistical aligner based on sentence length and lexical similarity; configurable for noisy or low-resource data.	Flags sentence pairs with alignment probability below threshold; often coincides with lacunae or partial translations.
GIZA++ / MGIZA	Token-level statistical aligner widely used in MT; can be trained on noisy historical datasets.	Unaligned word pairs signal omissions; custom tokenization allows adaptation to medieval spelling and syntax variants.
Web Align Toolkit (WAT)	Web-based wrapper for multiple engines (LF Aligner, HunAlign)	highlights sentence pairs with alignment probability below threshold—these often correspond to actual omissions

Visualization Add-ons

- Pixel-Based Latin Variation Explorer (Asokarajan et al.): Uses heatmap visualization where faded gray bands highlight phrases missing from certain editions.
- Dot Plot Views (ShakerVis): Classic diagonal "break" patterns immediately reveal missing text segments.

How do Gap Detectors Work Together?

1. First Alignment, Then Heuristics – Systems compute token- or sentence-level correspondences; any unmatched element is flagged as a potential gap.
2. Visual Highlighting – One-sided tokens are displayed in red, as empty fields, or listed separately, enabling researchers to quickly validate them.
3. Exportable Metadata – Gaps are recorded in TEI XML or TMX output as <gap reason="omission"> or empty segments, allowing for statistical analysis (e.g., "8% of Latin *clavos* lines lack a vernacular equivalent").

Which Stack should we Choose?

Research Goal	Recommended Tool(s)
Philological precision, line-level accuracy	**UGARIT**, **Alpheios**
Multiple witnesses, stemmatics	**CollateX** + **JSON diff**
Large corpus, scriptable query & analysis (Q&A)	**ANNIS**, **PROIEL**

Research Goal	Recommended Tool(s)
Fast, GUI-based, TM (Translation Memory) workflow	**LF Aligner**, **Web Align Toolkit (WAT)**

Alignment today is no longer merely "matching," but a diagnostic tool for identifying translation gaps, omissions, or looser renderings in medieval Latin and vernacular texts.

Conclusion

This study goes beyond traditional philological frameworks by treating scribal errors, stem truncations, and orthographic fluctuations not merely as taxonomical mistakes, but as valuable data. We have demonstrated that the error patterns in the medieval Hungarian corpus—particularly in the *Old Hungarian Lament of Mary*—reflect historical grammatical and grapho-stylistic processes that can only be uncovered through an interdisciplinary approach combining historical linguistics and natural language processing (NLP).

Key Insights

1. Truncated vs. Full Stem – Stem variation in medieval Hungarian is systematic; full stems appear only in suffixed contexts. Without this recognition, digital lemmatization will inevitably be distorted.
2. Functional Role of Errors – The *fcegegkel* example shows that a graphemic error can cause both a semantic shift (in the Passion focus) and a metrical disruption.
3. Alignment as Diagnostics – Latin–Hungarian alignment reveals not only translation strategy but also enables the automated detection of lacunae and textual gaps.
4. Hybrid Processing Pipeline – The combined use of neural HTR, rule-based FST, and the noisy channel model allows errors to be treated as reconstructible historical markers rather than mere noise.

Practical Contributions

1. Standardized TEI Markup – Multi-layer annotation ensures international comparability.
2. Open Toolkit – The accompanying scripts for Pynini, KenLM, and OpenNMT are available in a public Git repository, enabling immediate reuse and adaptation in other research projects.
3. Corpus Expandability – The pipeline is scalable and can be extended to the entire Old and Middle Hungarian corpus, as well as to multilingual materials (Latin, German, etc.).

Research Perspectives

1. Phonological Model Fitting – Statistical mapping of the *ŋk* > *g* sound change chronology based on scattered data.
2. Verse Analysis Module – Automatic detection of metrical anomalies to support the reconstruction of missing word forms.
3. Multilingual Transfer Learning – Fine-tuning normalization models trained on Scandinavian or Middle Dutch corpora for application to Hungarian material.

In conclusion, the study demonstrates that error is not merely an obstacle, but a key to the language of the past. The task of modern computational linguistics is to place that key—usable, reproducible, and shareable—into the hands of the philological community. With this investment, we not only improve the data quality of the corpus, but also uncover new, previously hidden layers of Hungarian language history.

This analysis was prepared with the assistance of ChatGPT, drawing on the latest research findings, scholarly publications, and technological case studies.

References

1. Horváth I. Ómagyar szövegemlékek mint textológiai tárgyak. Országos Széchényi Könyvtár, Budapest. 2015.
2. Horváth I. és mtsai. RpHA 1464 – Répertoire de la poésie hongroise ancienne. f-book.com. 2023.
3. Benkő L. Az Árpád-kor magyar nyelvű szövegemlékei. Akadémiai kiadó. Budapest. 1980; 81-82. 125. 158.
4. Gragger R. Ómagyar Máriasiralom. MNy. 1923; 1-13.
5. Gragger R. Eine altungarische marienklage. Berlin und Leipzig. 1923. 18.
6. A. Molnár F. A legkorábbi magyar szövegemlékek. Olvasat, értelmezés, magyarázatok, frazeológia. Debreceni Egyetem Bölcsészettudományi Kar Klasszika-filológia tanszék. 2005.
7. Vízkelety A. "Világ világa, virágnak virága" (Ómagyar Mária-siralom). Európa Könyvkiadó Budapest. 1986.
8. Mészöly G. Van-e az Ó-magyar Mária-siralomnak más latin forrása, mint a Lkód Planctus-szövege? Nyelvtudományi Közlemények. 1936; 278-284.
9. Martinkó A. Az Ómagyar Mária-siralom hazai és európai tükörben (Bevezetés és vázlat). Akadémiai Kiadó Budapest. 1988.
10. Bischoff B. Carmina Burana. Carl Winter Universitätsverlag, Heidelberg. 1970.
11. https://harmoniakert.hu/Omagyar_Maria-siralom_metruma_kivonat_2.pdf
12. Mező T. A Planctus ante nescia leleményes fordítása. In Az Ómagyar Mária-siralom megközelítésének új szempontjai. Anthology (editor Tibor Mező). Kiadó: Interkulturális kutatások Kft. (IkU). Hajdúböszörmény. 2025; 74-144.
13. Szabolcsi B. Vers és dallam. Akadémiai Kiadó. Budapest. 1959.
14. Dobszay L. Az Ómagyar Mária-siralom zenei vonatkozásai. Zenetudományi dolgozatok Budapest. 1988; 9-20.
15. Maróthy S. A Christus poetico planctus (Ómagyar Máriasiralom) verstani tagolójelei. Irodalomtörténeti Közlemények. 2014.
16. Szepes E, Szerdahelyi I. Verstan. Gondolat kiadás Budapest. 1981.
17. https://mek.oszk.hu/12700/12756/
18. Szelepcsényi Gy. 1938: Cantus Catholici 1703. (Magyar irodalmi ritkaságok 39.) 148.
19. Kehrein J. Latinische sequenzen des mittelalters. F Kupferberg, Mainz. 1873.

20. Mező T. A szegek szó az Ómagyar Mária-siralomban. In Az Ómagyar Mária-siralom megközelítésének új szempontjai. Anthology (editor Tibor Mező). Kiadó: Interkulturális kutatások Kft. (IkU) Hajdúböszörmény. 2025; 71-73.
21. Horváth J. A középkori magyar vers ritmusa. Ludvig Voggenreiter Verlag Berlin. 1928.
22. Kodály Z, Gyulai Á. Arany János népdalgyűjteménye. Akadémiai Kiadó Budapest. 1952.
23. Mone F. J. Schauspiele des mittelalters. Verlag von J. Vensheimer Mannheim. 1852.
24. Berrár J, Károly S. Régi magyar glosszárium. Szótárak, szójegyzékek és glosszák egyesített szótára. Akadémiai Kiadó Budapest. 1984. 636.
25. Mező T. Ómagyar Mária-siralom szövegkönyve. Novum publishing Kiadó Neckenmarkt. 2021.
26. Mező T. A magyar nyelv szótagtára. Manuscript Pomáz. 2014.
27. Mező T. Fonológia és metrika találkozása az Ómagyar Mária-siralomban. Magyar Nyelvőr. 2022; 146: 499-516.