Turkic Studies Journal Special Issue,
2 (2026) 229-241
Journal homepage: https:tsj.enu.kz

Article

Resolving High-Vowel Ambiguity (⟨ï⟩ / ⟨i⟩ / ⟨ı⟩) in OCR-Derived Old Turkic Editions: An Edition-Metadata-Driven Disambiguation Layer

E. Uçar

Friedrich-Schiller-Universität, Jena, Germany
ORCID ID: 0000-0002-0039-9619
(E-mail: erdem.ucar@uni-jena.de)

ARTICLE INFO

Keywords:
Old Turkic, OCR, text normalization, TEI-P5, high vowels, graphemic ambiguity, digital philology, historical corpora, Turcology, disambiguation.

IRSTI 16.21.27

DOI: *https://doi.org/10.32523/2664-5157-2026-2SI-229-241*

ABSTRACT

This study addresses the problem of high-vowel ambiguity in OCR-derived Old Turkic texts, in which the graphemic distinction between ⟨ï⟩, ⟨i⟩, and ⟨ı⟩ is frequently neutralized to ⟨i⟩. This issue arises from the limitations of existing normalization approaches, which fail to adequately capture edition-specific orthographic conventions or the variability across the editorial traditions. As a result, a significant portion of the original graphemic information is lost during digitization, thereby reducing the reliability of subsequent linguistic and philological analysis. To resolve this problem, the paper proposes a dedicated disambiguation layer integrated into a TEI-P5 two-layer encoding framework (orig/reg). The proposed layer operates strictly at the representation level and does not attempt to reconstruct phonology or modify the original OCR output. Instead, it combines edition-specific metadata with rule-based linguistic cues, including lexical allow-lists, morphological constraints, vowel harmony patterns, and loanword profiles. The model follows a deterministic priority structure, ensuring that each ambiguous case is resolved in a transparent, consistent, and reproducible manner. By design, the framework avoids probabilistic inference and prioritizes philological accountability over statistical generalization. The model was evaluated on a dataset of 4,485 tokens drawn from thirteen Old Turkic editions published between 1919 and 2023. Among these, 1,837 tokens exhibited unresolved high-vowel ambiguity after normalization. The proposed method successfully disambiguates approximately 88-93% of these cases, depending on the edition, while preserving unresolved forms through explicit TEI <unclear> annotation. This approach ensures that ambiguity is not obscured but remains visible and accessible for further philological evaluation and interpretation. Comparative analysis across editions further confirmed the stability of the method under varying orthographic conventions. The results demonstrate that a deterministic,

edition-aware approach can significantly improve accuracy compared to baseline methods, without sacrificing transparency, reversibility, or interpretability. The study further highlights the importance of editorial traditions in shaping graphemic representation and shows that ambiguity can be effectively managed through structured metadata and constrained rule-based systems. The proposed framework offers a reproducible and extensible solution for enhancing the quality, consistency, and interoperability of OCR-derived historical corpora in Turcology and digital philology. Future work will focus on extending the framework to additional Turkic editions and integrating it with corpus-level search and annotation tools.

Э. Учар
Фридрих Шиллер атындағы университет, Йена, Германия
ORCID ID: 0000-0002-0039-9619
(E-mail: erdem.ucar@uni-jena.de)

OCR арқылы цифрланған көне түркі мәтіндері басылымдарындағы қысаң дауыстылардың (⟨ï⟩ / ⟨i⟩ / ⟨ı⟩) көпмәнділігін басылым метадеректеріне сүйене отырып ажырату: дизамбигуация қабаты

Аннотация. Бұл зерттеу OCR арқылы цифрланған көне түркі мәтіндеріндегі қысаң дауыстылардың көпмәнділігі мәселесін қарастырады, мұнда ⟨ï⟩, ⟨i⟩ және ⟨ı⟩ арасындағы графемалық айырмашылық көбінесе ⟨i⟩ түріне дейін бейтараптанады. Бұл мәселе қолданыстағы нормализация әдістерінің шектеулерінен туындайды, олар жекелеген басылымдардың орфографиялық ерекшеліктерін және редакциялық дәстүрлердің вариативтілігін жеткілікті түрде ескермейді. Соның салдарынан цифрландыру барысында бастапқы графемалық ақпараттың елеулі бөлігі жоғалып, кейінгі лингвистикалық және филологиялық талдауларға әсер етіп, шынайылығын төмендетеді. Бұл мәселені шешу үшін зерттеу жұмысымызда TEI-P5 екі деңгейлі кодтау жүйесіне (orig/reg) интеграцияланған арнайы дизамбигуация қабаты қоса ұсынылады. Ұсынылған қабат қатаң түрде тек репрезентация деңгейінде ғана жұмыс істейді және фонологияны қайта қалпына келтіруге немесе OCR нәтижесін өзгертуге келмейді. Оның орнына ол нақты басылымдардың метадеректерін лингвистикалық белгілерге негізделген ережелермен біріктіреді, соның ішінде рұқсат етілген лексикалық тізімдер, морфологиялық шектеулер, дауысты дыбыстар үндестігінің үлгілері және кірме сөздер профилін де қоса алады. Модель қатаң белгіленген ережелер жүйесіне басымдық береді, әрбір көпмәнді жағдайдың ашық, бірізді және қайталанатын түрде шешілуін қамтамасыз етеді. Құрылымы бойынша модель ықтималдық әдістерін қолданбайды және статистикалық

Received 27 February 2026. Revised 29 March 2026. Accepted 30 April 2026. Available online 30 June 2026.

For citation: E. Uçar Resolving high-vowel ambiguity (⟨ï⟩ / ⟨i⟩ / ⟨ı⟩) in OCR-derived Old Turkic editions: an edition-metadata-driven disambiguation layer // Turkic Studies Journal. 2026. V. 8. № 2SI. P. 229-241. DOI: http://doi.org/10.32523/2664-5157-2026-2SI-229-241

жалпылаудан гөрі филологиялық есеп берушілікке басымдық береді. Модель 1919–2023 жылдар аралығында жарияланған көне түркі мәтіндерінің он үш басылымынан алынған 4485 токеннен тұратын деректер жиынында бағаланды. Олардың ішінде 1837 токен нормализациядан кейін қысаң дауыстылар бойынша көпмәнділікті сақтаған. Ұсынылған әдіс басылымға байланысты осы жағдайлардың шамамен 88–93%-ын сәтті ажыратады, ал шешілмеген формалар TEI (unclear) аннотациясы арқылы сақталады. Бұл тәсіл көпмәнділікті жасырмай, оны айқын күйде қалдырып, кейінгі филологиялық талдау мен интерпретация үшін қолжетімді етеді. Басылымдар арасындағы салыстырмалы талдау әдістің әртүрлі орфографиялық дәстүрлер жағдайында тұрақтылығын растады. Нәтижелер қатаң белгіленген және басылым ерекшеліктерін ескеретін тәсілге негізделген базалық әдістермен салыстырғанда дәлдікті едәуір арттыра алатынын, сонымен бірге ашықтықты, кері қайтарылымдылықты және интерпретацияланушылықты сақтайтынын көрсетеді. Зерттеу сондай-ақ редакциялық дәстүрлердің графемалық репрезентацияны қалыптастырудағы маңызын атап өтеді және көпмәнділікті құрылымдалған метадеректер мен шектелген ережелер жүйесі арқылы тиімді басқаруға болатынын көрсетеді. Ұсынылған фреймворк түркология мен цифрлық филологияда OCR арқылы цифрланған тарихи корпустардың сапасын, бірізділігін және интероперабельділігін арттыруға арналған қайта жаңғыртылатын және кеңейтілетін шешім ұсынады. Болашақ жұмыс фреймворкты қосымша түркі басылымдарына кеңейтуге және оны корпус деңгейіндегі іздеу мен аннотация құралдарымен біріктіруге бағытталады.

Кілт сөздер: көне түркі тілі, OCR (оптикалық мәтінді тану), мәтінді нормализациялау, TEI-P5, қысаң дауыстылар, графемалық көпмәнділік, цифрлық филология, тарихи корпустар, түркология, дизамбигуация.

Э.Учар
Университет имени Фридриха Шиллера, Йена, Германия
ORCID ID: 0000-0002-0039-9619
(E-mail: erdem.ucar@uni-jena.de)

Разрешение неоднозначности гласных верхнего подъёма (⟨ï⟩ / ⟨i⟩ / ⟨ı⟩) в OCR-оцифрованных изданиях древнетюркских текстов: слой дизамбигуации, основанный на метаданных издания

Аннотация. Данное исследование рассматривает проблему неоднозначности гласных верхнего подъёма в OCR-оцифрованных древнетюркских текстах, в которых графемное различие между ⟨ï⟩, ⟨i⟩ и ⟨ı⟩ часто нейтрализуется до ⟨i⟩. Эта проблема обусловлена ограничениями существующих методов нормализации, которые недостаточно учитывают орфографические особенности отдельных изданий и вариативность редакционных традиций. В результате значительная часть исходной графемной информации утрачивается в процессе цифровизации, что снижает надёжность последующего лингвистического и филологического анализа. Для решения данной проблемы в работе предлагается специализированный слой дизамбигуации, интегрированный в двухуровневую систему кодирования TEI-P5 (orig/reg). Данный слой функционирует строго на уровне представления и не предполагает восстановления фонологии или изменения исходного OCR-вывода. Вместо этого он сочетает метаданные конкретных

изданий с лингвистически обусловленными правилами, включая лексические списки разрешённых форм, морфологические ограничения, модели гармонии гласных и профили заимствованных слов. Модель использует детерминированную систему приоритетов, обеспечивая прозрачное, последовательное и воспроизводимое разрешение каждой неоднозначности. По своей конструкции она исключает вероятностные методы и отдаёт приоритет филологической проверяемости над статистическим обобщением. Модель была протестирована на корпусе из 4485 токенов, извлечённых из тринадцати изданий древнетюркских текстов, опубликованных в период с 1919 по 2023 год. Из них 1837 токенов сохранялинеоднозначностьгласныхверхнегоподъёмапосленормализации. Предложенный метод успешно разрешает приблизительно 88-93% таких случаев в зависимости от издания, при этом неразрешённые формы сохраняются посредством явной TEI-аннотации <unclear>. Такой подход не скрывает неоднозначность, а сохраняет её в явном виде, обеспечивая доступность для последующего филологического анализа и интерпретации. Сравнительный анализ различных изданий подтвердил устойчивость метода при различных орфографических традициях. Результаты показывают, что детерминированный, ориентированный на особенности издания подход может существенно повысить точность по сравнению с базовыми методами, не снижая при этом прозрачности, обратимости и интерпретируемости. Исследование также подчёркивает важность редакционных традиций в формировании графемной репрезентации и демонстрирует, что неоднозначность может эффективно управляться с помощью структурированных метаданных и ограниченных систем правил. Предложенный фреймворк представляет собой воспроизводимое и расширяемое решение для повышения качества, согласованности и интероперабельности OCR-оцифрованных исторических корпусов в тюркологии и цифровой филологии. Дальнейшие исследования будут направлены на расширение фреймворка на дополнительные тюркские издания и его интеграцию с инструментами корпусного поиска, и аннотирования.

Ключевые слова: древнетюркский язык, OCR (оптическое распознавание текста), нормализация текста, TEI-P5, гласные верхнего подъёма, графемная неоднозначность, цифровая филология, исторические корпусы, тюркология, дизамбигуация.

Introduction

Over the past century, various Old Turkic editions have been published, each employing different conventions for the representation of high vowels. These systems are neither fully standardized nor mutually compatible. For example, German, Russian, French, and Japanese editions generally preserve a clear distinction between ⟨ï⟩ (back unrounded high vowel), ⟨i⟩ (front high vowel), and in some cases ⟨ı⟩. On the other hand, Turkish editions based on modern Latin orthography retain only the ⟨i⟩ / ⟨ı⟩ opposition and omit ⟨ï⟩ entirely. For transliteration conventions in Turkish editions, see Uçar (2020; 2021). In Japanese and early Chinese editions, diacritics at are absent, and all three vowels may be merged into a single grapheme ⟨i⟩.

When such editions are processed through OCR, the problem becomes even more pronounced Even relatively accurate OCR systems frequently remove the diaeresis from ⟨ï⟩ and normalize the dotless ⟨ı⟩ to ⟨i⟩. As a result, the original three-way difference is reduced to a single graphemic form. To address this issue, a normalization layer proposed by Uçar (forthcoming) introduces a TEI-P5 parallel encoding framework. Within this framework, the

uncorrected OCR string is preserved in the orig layer, while a normalized representation is provided in the reg layer. At the same time, the framework explicitly identifies high-vowel collapse as the main error type that cannot be resolved solely at the representation level.

There have been a few past efforts to gather electronic collections of Old Turkic texts, with VATEC *(= Vorislamische Alttürkische Texte: Elektronisches Corpus)* standing out as the most significant. This project, led by Marcel Erdal, Jost Gippert, Klaus Röhrborn, and Peter Zieme, was a collaborative effort between the Goethe-Universität Frankfurt, the Georg-August-Universität Göttingen, and the Berlin-Brandenburgische Akademie der Wissenschaften, supported by the DFG. In 2001, VATEC released a pilot CD-ROM while in 2003 a VATEC website that included re-editions of texts written in Runiform, Old Uyghur, Manichaean, Sogdian, and Syriac scripts, complete with transliteration, normalized transcription, morphological parsing, and translation (Erdal et al. 2003). However, it's worth noting that VATEC didn't start with OCR-derived material. Instead, the editions were carefully re-keyed by hand from the original print versions, and the high-vowel distinction was maintained through careful editorial choices rather than relying on a standardized normalization process. Because of this, the differences in orthography between different editorial traditions and the systematic disappearance of ⟨ï⟩, ⟨i⟩, and ⟨ı⟩ that often happens in OCR-based systems didn't fall within the project's goals (Erdal et al. 2003). Other digital resources, like Wilkens's open-access Old Uyghur dictionary (2021), the ongoing Uigurisches Wörterbuch by the Göttingen Academy, and the digitized Turfan collection from the Berlin-Brandenburg Academy, also use editorially curated input and don't tackle the graphemic uncertainties that can arise during OCR. To our knowledge, no existing Old Turkic database offers a clear, edition-aware way to handle high-vowel collapse in OCR-derived texts while keeping the original text intact. This project is designed to address exactly that need.

This paper proposes a dedicated disambiguation layer designed to address this residual ambiguity. The layer is applied after representation-level normalization and prior to subsequent linguistic annotation. The method combines edition-specific metadata with a consistent system of linguistically motivated constraints. Rather than attempting phonological reconstruction, the model operates exclusively at the graphemic level by analyzing distributional patterns within individual editions. In doing so, it preserves the integrity of the OCR output while avoiding intervention in the original textual data.

Materials and research methods

The disambiguation layer proposed in this study is intentionally narrow in scope. It does not replace representation-level normalization. It operates directly above it and does not substitute for philological analysis. Its primary function is to render residual ambiguity machine-tractable without concealing it. The behaviour of the layer is constrained by five core principles.

Non-destructiveness: The original layer remains unchanged throughout the process. All disambiguation decisions are encoded in the reg layer and where necessary, are marked using TEI <unclear> or @cert attribute attribute to indicate residual uncertainty.

Edition-awareness: Each source edition is associated with a dedicated profile describing its treatment of high vowels, OCR related behaviour, and preferred normalization conventions within the reg layer. These profiles are created once and subsequently reused across the corpus.

Deterministic priority: Disambiguation clues are applied according to a fixed hierarchical order. The first rule capable of resolving a given token is accepted, while unresolved ties are handled through edition – specific fallback strategies.

Edition conformity: The system strictly adheres to the orthographic conventions of each edition. No rule may introduce graphemic contrasts that are absent from the source tradition. For example, a Turkish edition that consistently employs the ⟨i⟩ / ⟨ı⟩ opposition will never generate ⟨ï⟩ in the reg layer, even if such a reconstruction might appear philologically plausible.

Reversibility: Since the original text is preserved unchanged and each reg token remains linked to its source through the TEI <choice> structure, all disambiguation decisions remain fully reviewable and revirsible at the token level.

The layer was evaluated using the same 4,485-token dataset employed in earlier representation-level normalization research (Uçar, forthcoming), comprising thirteen editions published between 1919 and 2023 (Le Coq, 1919; Bang, 1923; Arat, 1965; Hamilton, 1971; Röhrborn, 1971; Geng, 1989; Dietz et al., 2015; Zieme et al., 2022; Kaya, 2023; among others). Of these tokens, 1,837 (approximately 41%) contained at least one high vowel that remained ambiguous after normalization. To establish a reference interpretation, two annotators independently evaluated the tokens using the glossaries and indices of the corresponding editions. Disagreements were subsequently resolved through discussion and consensus.

Research background

The graphemes ⟨ï⟩, ⟨i⟩, and ⟨ı⟩ correspond to distinct categories within the reconstructed Old Turkic vowel system. In different editorial traditions, ⟨ï⟩ and ⟨ı⟩ represent the same phonological category, namely the back unrounded high vowel, whereas ⟨i⟩ encodes the front high vowel. The variation is therefore orthographic rather than phonological in nature. Editorial systems map this phonological space onto the Latin alphabet in incompatible ways, as summarized in Table 1.

Table 1. High-vowel encoding across editorial traditions.
1-кесте. Редакциялық дәстүрлердегі қысаң дауыстылардың кодталуы.
Таблица 1. Кодирование гласных верхнего подъёма в различных редакционных традициях.

Tradition	Back high	Front high	Additional notes
Early German	ï	i	Diaeresis is contrastive; ⟨ı⟩ is not used
Russian (Soviet)	ï	i	Follows the German convention
French	ï	i	Occasional use of ⟨ï⟩ with trema
Modern Turkish	ı	i	Employs the orthographic ⟨ı⟩ / ⟨i⟩ opposition
Japanese / early Chinese	i (undiff.)	i (undiff.)	No orthographic distinction between back and front high vowels

OCR processing may further obscure graphemic distinctions present in the source editions. In particular, diacritics are frequently lost during recognition, while the Turkish grapheme ⟨ı⟩ is often misidentified as ⟨i⟩ due to the minimal visual difference between the two characters at

low resolution. Similar challenges in the OCR processing of historical and low-resource scripts have been documented in recent research (Carlson et al., 2023; Özateş et al., 2025).

Analysis

1. The Edition Profile

Each edition is associated with a concise profile containing the information necessary for resolving collapsed high vowels. Functionally, these profiles serve as compact philological guides, specifying which graphemic distinctions are preserved or neutralized in a given edition, how loanwords are treated, and how OCR processes affect the original orthography in practice.

At minimum, each profile contains six components:

1. the inventory of high-vowel sounds employed by the edition (e.g., {ï, i} in German conventions, {ı, i} in Turkish conventions, or {i} in systems lacking a back/front distinction;

2. OCR transformation patterns affecting these graphemes, such as the frequent normalization of ⟨ï⟩ to ⟨i⟩;

3. the default fallback grapheme used when OCR ambiguity cannot be resolved deterministically;

4. loanword handling conventions for lexical items of Sogdian, Sanskrit, Chinese, or Tocharian origin, including whether vowel representation follows the donor language or the adapted Old Turkic form;

5. a lexical allow-list of common forms whose high-vowel representation is fixed within the edition, typically consisting of 150-400 items extracted from the edition's glossary or index and cross-checked against standard Old Turkic etymological references such as Clauson (1972);

6. a parameter indicating whether the edition applies vowel harmony from the initial syllable or permits the lexical stem to determine affix vowel realization.

The profiles are implemented as compact, manually curated YAML files distributed together with the TEI corpus. They are not inferred automatically from statistical patterns in the data. By maintaining explicit and human-readable profile definitions, the framework preserves philological transparency and allows individual disambiguation decisions to be traced directly to their underlying editorial assumptions rather than to opaque probabilistic inference.

2. The Disambiguation Pipeline

If the representation-level normalized form of a token still contains unresolved high-vowel ambiguity, the disambiguation layer applies four cues in a fixed hierarchical sequence grounded in the morphological and vowel-harmonic regularities of Old Turkic. The first cue that produces a determinate outcome is accepted, and no subsequent cues are consulted.

Table 2. Cue priority in the disambiguation layer.
2-кесте. Дизамбигуация қабатындағы белгілердің басымдық тәртібі.
Таблица 2. Приоритет признаков в слое дизамбигуации.

Cue	Mechanism	Example
Lexical allow-list	Direct lookup in the edition-specific fixed high-vowel lexicon.	OCR tingri → reg täŋri (allow-listed lemma).

Morpheme rule	Application of finite rules governing Old Turkic derivational and inflectional morphemes with fixed high-vowel realization.	OCR {-lig} / {-lik} → reg {-lïg} / {-lig} depending on stem class.
Vowel harmony	Propagation of harmonic class from an unambiguous front or back vowel within the stem to the unresolved high vowel.	OCR arig → reg arïg (back-harmonic stem ⟨a⟩).
Loanword profile	Application of the edition-specific loanword policy when the token matches a recognized donor-language pattern.	Skt. sūtra > > OCR sudur → reg sudur (edition keeps donor vowel).
Fallback	Emission of the edition-defined fallback grapheme, accompanied by TEI <unclear> annotation.	OCR ir → reg ir + <unclear>.

In rare cases where multiple cues are simultaneously applicable (e.g., lexical and harmonic evidence), the fixed priority hierarchy ensures deterministic resolution. For example, if a token matches both a lexical entry and a harmonic pattern, the lexical cue takes precedence.

The pipeline is intentionally designed to remain shallow, as the majority of philologically relevant cases can be resolved using the four core cues alone. Additional mechanisms, such as edition-wide statistical models, neural language models, or cross-edition voting strategies, may provide modest improvements in recall; however, such gains would come at the expense of interpretability, transparency, and philological verifiability.

3. TEI Encoding

Disambiguation is implemented as an additional layer within the existing TEI-P5 <choice> structure. Once a determinate interpretation has been established, the resolved token is emitted directly in the normalized representation.

<choice>
<orig>arig</orig>
<reg source="#harmony" cert="high">arïg</reg>
</choice>

When a token cannot be resolved, the system applies the edition-specific fallback grapheme together with explicit TEI <unclear> annotation and an associated @cert value. This ensures that downstream users can seeidentify the form as unresolved ambiguity rather than as a silently normalized representation.

<choice>
<orig>ir</orig>
<reg cert="low"><unclear>ir</unclear></reg>
</choice>

The @source attribute associated with the <reg> element records the cue responsible for the disambiguation decision, such as #lexicon, #morph, #harmony, #loanword, or #fallback. As a result, each token preserves a transparent and traceable path from the original OCR input to its normalized representation.

4. Evaluation

The disambiguation layer was evaluated using the same 4,485-token dataset employed in earlier representation-level normalization research (Uçar, forthcoming), comprising thirteen

editions published between 1919 and 2023 (Le Coq, 1919; Bang, 1923; Arat, 1965; Hamilton, 1971; Röhrborn, 1971; Geng, 1989; Dietz et al., 2015; Zieme et al., 2022; Kaya, 2023; among others). Of these tokens, 1,837 (approximately 41%) contained at least one high vowel that remained ambiguous after normalization. To establish a reference interpretation, two annotators independently evaluated the ambiguous forms using the glossaries and indices of the corresponding editions. Cases of disagreement were subsequently resolved through discussion and consensus.

Across the ambiguous subset, the disambiguation layer correctly resolved approximately 88-93% of high-vowel assignments, depending on the edition. The highest performance was observed in recent German and Turkish editions with extensive and well-structured glossaries, whereas lower performance was recorded for early twentieth-century editions such as Le Coq (1919) and Bang (1923), whose inconsistent typography and orthographic variation complicated stem identification. Table 3 presents the contribution of each disambiguation cue to the overall resolution process.

Table 3. Per-cue share of decisions and accuracy on ambiguous tokens.
3-кесте. Әрбір белгі бойынша шешімдердің үлесі және көпмәнді токендердегі дәлдік.
Таблица 3. Доля решений по каждому признаку и точность на неоднозначных токенах.

Cue	Share of decisions	Accuracy	Effect on recall
Lexical allow-list	∼46%	∼99%	High
Morpheme rule	∼21%	∼96%	High
Vowel harmony	∼18%	∼88%	Medium
Loanword profile	∼5%	∼92%	Low
Fallback (<unclear>)	∼10%	n/a	Preserved

For comparison, a naive baseline that assigns all ambiguous vowels to ⟨i⟩ achieves an accuracy of approximately 52-58%, depending on the edition. This result highlights the substantial improvement provided by the proposed disambiguation layer.

Two observations are particularly noteworthy. First, the lexical allow-list functions not merely as a supplementary heuristic, but as a major source of disambiguation evidence It accounts for nearly half of all successful resolutions while maintaining near-perfect precision. This reflects a broader characteristic of the Old Turkic lexicon, in which a relatively small set of highly frequent lexical items constitutes a substantial proportion of the attested corpus across editions. Second, vowel harmony proved less reliable as an independent cue. This does not indicate weakness in the harmonic system of Old Turkic itself; rather, harmonic propagation becomes unstable when OCR corruption affects the vowels serving as harmonic anchors within a token. Such errors are most common in forms where several complementary cues are simultaneously unavailable, precisely the cases in which the fallback strategy based on TEI <unclear> annotation becomes particularly valuable.

Inter-annotator agreement was high (approximately Cohen's $\kappa = 0.92$), indicating strong consistency in gold-standard assignment. Remaining disagreements were resolved through joint review with reference to the glossaries and indices of the corresponding editions.

All evaluation scripts and annotation guidelines are available upon request in order to support reproducibility.

Results

The results support the central hypothesis of the study: the residual high-vowel ambiguity that remains unresolved after representation-level normalization can, in most cases, be effectively managed through a simple and transparent disambiguation layer. High levels of accuracy can be achieved without the use of neural or statistical modelling techniques. The remaining unresolved cases do not constitute unstructured noise, rather, they are systematically preserved in orig, represented through used fallback forms in reg layer, and explicitly marked as uncertain.

Among the non-fallback cues, the loanword component demonstrates the lowest accuracy. This is primarily attributable to cross-edition variability in the treatment of donor-language vowels, as well as inconsistencies in adaptation strategies within individual editions.

Several limitations should be noted. The proposed layer remains confined to representational normalization and does not address script-level transliteration for Runiform, Old Uygur, or Manichaean sources. Its performance depends in part on the quality of the edition profile: editions lacking a substantial glossary yield smaller lexical allow-lists and therefore lower recall. In addition, the loanword cue remains the weakest component because donor-language treatment varies not only across editions and sometimes even within a single edition. Finally, the vowel-harmony cue presupposes that at least one stem vowel remains recoverable after OCR processing; when this condition is not met, unresolved forms appropriately remain marked with TEI <unclear> annotation.

Project Scope and Hosting: This paper outlines a framework created by a single researcher at the Friedrich-Schiller-Universität Jena. Right now, it's not part of a bigger group, but it's designed to work with and enhance existing resources in Turcology, especially VATEC and the *Uigurisches Wörterbuch* from the Göttingen Academy.

The initial dataset mentioned above includes 4,485 tokens from thirteen Old Turkic editions published between 1919 and 2023. In its complete form, the database is expected to expand over the next few years to include a much broader range of editions, including the main Old Uyghur, Manichaean, and Runiform editions from the 20th and 21st centuries. We're aiming for a total volume of several tens of thousands of tokens, depending on the availability of suitable editions and the creation of edition profiles. Each new edition will follow the same process described above, so adding more editions won't require changing the main disambiguation system.

We haven't yet decided on a permanent hosting solution. The disambiguation layer, edition profiles, and the TEI-P5 corpus are planned for open release under a permissive license, with a long-term open-access repository like Zenodo, and, if possible, integration with an existing Turcological system. If you're interested, please reach out to the author to learn more about the current status of the resources and how you can help develop more edition profiles.

Conclusion

When we look at earlier projects like VATEC, which needed people to manually re-key everything, it couldn't handle the OCR-induced high-vowel collapse issue. But the framework we're proposing has four cool benefits. First, it works directly with noisy OCR output, which

is great for all the digitised editions out there that can't be re-keyed by hand. Second, it puts editorial knowledge into clear, human-readable edition profiles, so everyone can see the decisions made. Third, by using a TEI-P5 ⟨choice⟩ structure with clear ⟨source⟩ and ⟨cert⟩ tags, every disambiguation step can be undone at the token level, which is something we didn't have before. Fourth, the framework is easy to add to: you can just add a profile without changing the main process. All these features make OCR-derived Old Turkic corpora more like what we need for digital philology today, which is all about being reproducible and working together. Plus, they're a nice addition to the depth of curated resources like VATEC (Erdal et al., 2003) and the Uigurisches Wörterbuch (Wilkens, 2021).

High-vowel collapse has remained a persistent problem in the normalization of OCR-derived Old Turkic texts. This study argues that high-vowel ambiguity in OCR-based Old Turkic corpora is best understood as a structurally constrained problem that can be addressed through explicit, edition-aware modelling. To this end, the study introduces declarative edition profiles, a strictly ordered four-cue disambiguation pipeline, and a TEI-P5 extension capable of preserving both provenance and uncertainty information. Together these components make it possible to recover the majority of lost graphemic distinctions without compromising the integrity of the source edition. The proposed framework is incremental, reversible at the token level and fully transparent in its operation. When combined with representation-level normalization, it substantially improves the consistency and interoperability of OCR-derived Old Turkic corpora and thereby contributes to the broader development of large-scale digital philology resources.

Abbreviations

OCR = Optical Character Recognition.
orig = original (TEI element: the unaltered source reading).
reg = regularized (TEI element: the normalized / standardized form).
Skt. = Sanskrit.
TEI-P5 = Text Encoding Initiative, Proposal 5.
YAML = YAML Ain't Markup Language.

Reference

Arat R.R., 1965. Eski Türk Şiiri. Ankara: Türk Tarih Kurumu Yayınları.

Bang W., 1923. Manichaeische Laien-Beichtspiegel. Le Muséon. 36. P. 137–242.

Carlson J. et al., 2023. Efficient OCR for building a diverse digital history. arXiv preprint arXiv:2304.02737.

Clauson G., 1972. An Etymological Dictionary of Pre-Thirteenth Century Turkish. Oxford: Clarendon Press.

Dietz S. et al., 2015. Die alttürkische Xuanzang-Biographie V: Nach der Handschrift von Leningrad, Paris und Peking sowie nach dem Transkript von Annemarie v. Gabain (Hrsg., Übers., Komm.). Wiesbaden: Harrassowitz Verlag.

Erdal M., Gippert J., Röhrborn K., Zieme P., Nevskaya I., Knüppel M., Özertural Z., Taube J., 2003. Vorislamische Alttürkische Texte: Elektronisches Corpus. [Electronic resource]. Available at: https://vatec2.fkidg1.uni-frankfurt.de.

Geng S., 1989. A study of one newly discovered folio of the Uighur Abhidharmakośaśāstra. Central Asiatic Journal. 33. P. 36–45.

Hamilton J.R., 1971. Le conte bouddhique du bon et du mauvais prince en version ouïgoure. Paris: Klincksieck.

Kaya C., 2023. Uygurca Altun Yaruk: Belgeler. Ankara: Türk Dil Kurumu Yayınları.

Le Coq A. von., 1919. Kurze Einführung in die uigurische Schriftkunde. Mitteilungen des Seminars für Orientalische Sprachen an der Friedrich-Wilhelms-Universität zu Berlin, Westasiatische Studien. 22. P. 93–109.

Özateş Ş. et al., 2025. Building foundations for natural language processing of historical Turkish: Resources and models. arXiv preprint arXiv:2501.04828. (Accessed: 20.04.26)

Röhrborn K., 1971. Eine uigurische Totenmesse: Text, Übersetzung, Kommentar. Berliner Turfantexte 1. Berlin: Akademie Verlag.

TEI Consortium, 2024. TEI P5: Guidelines for Electronic Text Encoding and Interchange (Version 4.7.0, tei-c.org). (Accessed: 20.04.26)

Uçar E., 2020. Türkiye'deki Eski Uygurca Metin Neşirleri İçin Kullanılacak Harfçevrim ve Yazıçevrim Kılavuzu. Journal of Old Turkic Studies. 4(1). P. 231–250.

Uçar E., 2021. Türkiye'deki Manihey Harfli Eski Uygurca Neşirler İçin Harfçevrim ve Yazıçevrim Kılavuzu. Journal of Old Turkic Studies. 5(1). P. 161–194.

Uçar E., 2026. A Normalization Layer for Old Turkic Text Editions in OCR-Based Workflows. (forthcoming).

Wilkens, J. 2021. Handwörterbuch des Altuigurischen, Altuigurisch-Deutsch-Türkisch. Göttingen: Universitätsverlag Göttingen, 930 p.

Zieme P. et al., 2022. Avalokiteśvara-Sūtras: Edition altuigurischer Übersetzungen nach Fragmenten aus Turfan und Dunhuang. Berliner Turfantexte 50. Turnhout: Brepols.

Reference

Arat R.R., 1965. Eski Türk Şiiri [Old Turkic Poetry]. Ankara: Türk Tarih Kurumu Yayınları [Ankara: Turkish Historical Society Publications]. [in Turkish].

Bang W., 1923. Manichaeische Laien-Beichtspiegel [Manichaean Lay Confessional Mirror]. Le Muséon. 36. P. 137–242. [in German].

Carlson J. et al., 2023. Efficient OCR for building a diverse digital history. arXiv preprint arXiv:2304.02737.

Clauson G., 1972. An Etymological Dictionary of Pre-Thirteenth Century Turkish. Oxford: Clarendon Press.

Dietz S. et al., 2015. Die alttürkische Xuanzang-Biographie V: Nach der Handschrift von Leningrad, Paris und Peking sowie nach dem Transkript von Annemarie v. Gabain (Hrsg., Übers., Komm.) [The Old Turkic Xuanzang Biography V: Based on the Manuscripts from Leningrad, Paris and Beijing and the Transcript by Annemarie von Gabain (ed., trans., comm.)]. Wiesbaden: Harrassowitz Verlag. [in German].

Erdal M., Gippert J., Röhrborn K., Zieme P., Nevskaya I., Knüppel M., Özertural Z., Taube J., 2003. Vorislamische Alttürkische Texte: Elektronisches Corpus. [Electronic resource]. Available at: https://vatec2.fkidg1.uni-frankfurt.de

Geng S., 1989. A study of one newly discovered folio of the Uighur Abhidharmakośaśāstra. Central Asiatic Journal. 33. P. 36–45.

Hamilton J.R., 1971. Le conte bouddhique du bon et du mauvais prince en version ouïgoure [The Buddhist Tale of the Good and the Bad Prince in Uighur Version]. Paris: Klincksieck. [in French].

Kaya C., 2023. Uygurca Altun Yaruk: Belgeler. Ankara: Türk Dil Kurumu Yayınları [Uighur Altun Yaruk: Documents. Ankara: Turkish Language Association Publications]. [in Turkish].

Le Coq A. von., 1919. Kurze Einführung in die uigurische Schriftkunde [A Short Introduction to Uighur Paleography]. Mitteilungen des Seminars für Orientalische Sprachen an der Friedrich-Wilhelms-Universität zu Berlin [Proceedings of the Seminar for Oriental Languages at the Friedrich Wilhelm University of Berlin, West Asian Studies]. Westasiatische Studien. 22. P. 93–109. [in German].

Özateş Ş. et al., 2025. Building foundations for natural language processing of historical Turkish: Resources and models. arXiv preprint arXiv:2501.04828. (Accessed: 20.04.26)

Röhrborn K., 1971. Eine uigurische Totenmesse: Text, Übersetzung, Kommentar [A Uighur Funeral Mass: Text, Translation]. Berliner Turfantexte 1. Berlin: Akademie Verlag. [in German].

TEI Consortium, 2024. TEI P5: Guidelines for Electronic Text Encoding and Interchange (Version 4.7.0, tei-c.org). (Accessed: 20.04.26)

Uçar E., 2020. Türkiye'deki Eski Uygurca Metin Neşirleri İçin Kullanılacak Harfçevrim ve Yazıçevrim Kılavuzu [Grapheme and Transliteration Guide for Old Uighur Text Editions in Turkey]. Journal of Old Turkic Studies. 4(1). P. 231–250. [in Turkish].

Uçar E., 2021. Türkiye'deki Manihey Harfli Eski Uygurca Neşirler İçin Harfçevrim ve Yazıçevrim Kılavuzu [Grapheme and Transliteration Guide for Manichaean Script Old Uighur Editions in Turkey]. Journal of Old Turkic Studies. 5(1). P. 161–194. [in Turkish].

Uçar E., 2026. A Normalization Layer for Old Turkic Text Editions in OCR-Based Workflows. (forthcoming).

Wilkens, J. 2021. Handwörterbuch des Altuigurischen, Altuigurisch-Deutsch-Türkisch. Göttingen: Universitätsverlag Göttingen, 930 p.

Zieme P. et al., 2022. Avalokiteśvara-Sūtras: Edition altuigurischer Übersetzungen nach Fragmenten aus Turfan und Dunhuang [Avalokiteśvara Sutras: Edition of Old Uighur Translations Based on Fragments from Turfan and Dunhuang]. Berliner Turfantexte 50. Turnhout: Brepols. [in German].

Information about the author:

Erdem Uçar, Doctor of Philology, Professor, Institute of History, Friedrich Schiller University, 13 Fürstengraben Str., 07743, Jena, Germany.

Researcher ID: 1000090124289

Автор туралы мәлімет:

Эрдем Учар, филология докторы, профессор, Тарих институты, Фридрих Шиллер атындағы университет, Fürstengraben көш., 13, 07743, Йена, Германия.

Researcher ID: 1000090124289

Сведения об авторе:

Эрдем Учар, доктор филологии, профессор, Институт истории, Университет имени Фридриха Шиллера, ул. Fürstengraben, 13, 07743, Йена, Германия.

Researcher ID: 1000090124289

Conflict of Interest.

There is no conflict of interest related to this article.

Мүдделер қақтығысы.

Мақалаға байланысты мүдде қақтығысы жоқ.

Конфликт интересов.

Нет конфликта интересов, связанного со статьей.