

From Clay to Code: Transforming Hittite Texts for Machine Learning

Emma Yavasan and Shai Gordin

Ariel University, Dept. of Land of Israel and Archaeology, Ariel, Israel
emma.yavasan@msmail.ariel.ac.il, shaigo@ariel.ac.il

Abstract

This paper presents a comprehensive methodology for transforming XML-encoded Hittite cuneiform texts into computationally accessible formats for machine learning applications. Drawing from a corpus of 8,898 texts (558,349 tokens in total) encompassing 145 cataloged genres and compositions, we develop a structured approach to preserve both linguistic and philological annotations while enabling computational analysis. Our methodology addresses key challenges in ancient language processing, including the handling of fragmentary texts, multiple language layers, and complex annotation systems. We demonstrate the application of our corpus through experiments with T5 models, achieving significant improvements in Hittite-to-German translation (ROUGE-1: 0.895) while identifying limitations in morphological glossing tasks. This work establishes a standardized, machine-readable dataset in Hittite cuneiform, which also maintains a balance with philological accuracy and current state-of-the-art.

1 Introduction

This paper builds on the advancements in corpus technologies and computational linguistics, contributing to the evolution of corpus linguistics for Hittite studies, making Hittite cuneiform texts accessible for data analysis and machine learning. A corpus-based approach in the area of Ancient Language Processing (ALP) is used to create a dataset of Hittite documents converted primarily into CSV format, with plans to extend to additional formats such as JSON and YAML in future releases.

Hittite is the oldest attested Indo-European language of the Anatolian family written in cuneiform script from the 17th to the 12th centuries BCE. All Hittite documents have been structured according to content and genre in the *Catalogue des textes hittites* (CTH) by Laroche, updated in the digital CTH

(see Fig. 1).¹ A more practical way to classify Hittite documents is suggested by van den Hout (2008) who divided the Hittite documents into "prescriptive" (copied over a period of several generations, having a long-term purpose) and "descriptive" (mostly daily economic and administrative texts) categories. This approach, however, is not a formalized one, and there are many exceptions in both groups (van den Hout, 2002; Gordin, 2015).

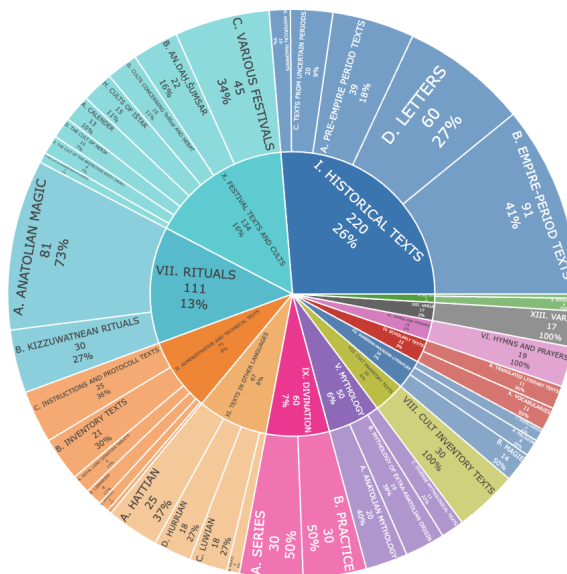


Figure 1: Distribution of texts in CTH

The main digital resource for Hittite is The *Hethitologie Portal Mainz* (HPM), which uses XML mark-up for raw text edition files. By the end of 2024, digital editions of state treaties, laws, myths, prayers, magic rituals and festivals (partly), cult inventories and some administrative texts (mostly inventories) have been published there. Additionally, a searchable annotated ritual and festival corpus of raw transliterations of Hittite documents (**not**

¹Originally published by Emanuelle Laroche in 1972, this resource has been adopted and updated as part of the CTH online: S. Kořak – G.G.W. Müller – S. Görke – Ch.W. Steitler, hethiter.net/: CTH (2025-01-28).

based on published editions) has been released in 2023 under the *Corpus der Hethitischen Festrituale* (HFR), and in 2024 the *Thesaurus Linguarum Hethaeorum digitalis* (TLH^{dig}) was released, which aims to cover eventually the entire known Hittite text corpus. These are the main sources of data for our research (see [Acknowledgments](#)).

Corpus linguistics evolved in the early aughts from a narrow methodology primarily concerned with the digitization of printed texts into a cornerstone of linguistic research and applications ([Lüdeling and Kytö, 2008](#)). This transformation has been driven by advancements in digital technologies: corpus-derived models use such methods as the fine-tuning of large language models (LLMs) for specific linguistic tasks. This paper aims at creating a dataset for the development of a Hittite corpus, transforming the existing annotated XML data into formats specifically optimized for machine learning applications.

2 Background

Research into ancient Near Eastern languages, particularly Hittite, faces unique challenges due to the nature of the source materials. Unlike modern languages, which benefit from vast, well-documented corpora, Hittite studies contend with limited digital resources designed specifically for computational analysis. So far, to our knowledge, two approaches to corpus studies of Hittite have been pursued since 2014: *Goottite* (Digital search of Hittite texts) by D. Frantikova and *Hittitecorpus* (Annotated Corpus of Hittite Clauses) by M. Molina. Both were developed with specific research objectives in mind that differ substantially from our current approach. While these resources allow contextual searches within their text collections, neither was designed to function as a comprehensive, computationally accessible corpus.

The *Hethitologie Portal Mainz* (HPM) represents the most extensive digital resource for Hittite, with richly annotated XML texts primarily optimized for philological accuracy and scholarly reference. However, HPM’s complex XML structure, while excellent for digital editions, presents significant challenges for systematic computational processing or machine learning applications. The critical limitation across all these existing resources is that none provides a standardized, machine-readable dataset that researchers can readily extract, manipulate, and process at scale. This rep-

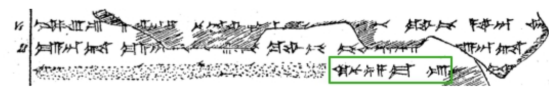
resents the fundamental advancement of our approach—transforming philologically rich but computationally challenging materials into structured formats that preserve scholarly annotations while enabling corpus-wide linguistic analysis and computational methods.

Using HPM corpora XML-marked-up material, we are planning to cover a much bigger amount of documents, as well as propose automated parsing and annotation of Hittite texts, taking as a first approach the dataset previously created for fine-tuning a German T5 model for the tasks of glossing and machine translation ([Yavasan and Gordin, 2024](#)).

The first problem that emerged in the creation of our Hittite corpus is the convertibility of the annotated data. We worked directly with XML files from TLH^{dig} that incorporate SimTex conventions within their structure². These files are traditionally dense with philological remarks and notations as an addition to grammatical information.

Another significant question is the way to represent all different languages contained in every Hittite document. Traditionally, transliterated texts in Hittite use three types of formatting: italic small caps, italic capital letters, and normal capital letters for Hittite words, Akkadian and Sumerian logograms, accordingly; unfortunately, this textual approach cannot be easily supported in the corpus that makes focus on linguistic analysis rather than on philologically rich digital editions.

There is also the problem of fragmented, often damaged, primary texts (see Fig. 2).



Obv. 28': mán=ta mán DUMU.MUNUS=pat ŪL kuwapi peḫḫun mán=ta [...]

Probable reconstruction in the lacuna (Edel 1994; Hoffner 2009: 284):
[NAM.RA.MEŠ GU₄.HI.A UDU.HI.A memahḫun]

Obv. 29': [...] kinun=ma ŪL [...]

"(If I had not at any time (sincerely) given my own daughter to you, would I have you [promised the civilian captives, cattle, and sheep?]) [...] But now not [...]"

Figure 2: KUB 21.38 (NH/NS; CTH 176) obv. 28'-29' - Letter of Queen Puduheba to Pharaoh Ramses II ([Edel, 1994](#); [Hoffner, 2009](#))

Several scholars have proposed solutions for dealing with fragmented texts ([Zemánek, 2007](#); [Inglese, 2016](#); [Molina, 2016](#); [Molina and Molin,](#)

²For the SimTex format description, see [HPM Guide](#).

2016). In our approach to the Universal Dependencies (UD) treebank, we previously proposed a syntactic annotation method in which every fragmented block is treated as dependent on a verb and marked as FRGM (Yavasan and Molina, 2024). However, this approach introduces ambiguity in the linguistic analysis of Hittite syntax. Therefore, outside the dependency grammar framework, we need to identify an alternative solution that preserves the integrity of the information.

3 Methodology & Implementation

3.1 Data Sources and XML Encoding

For this research we chose to create a dataset out of a subset available to us from the existing repository of annotated texts, called *Thesaurus Linguarum Hethaeorum digitalis* (TLH^{dig}). It is an open-access digital repository that provides structured linguistic and philological annotations in XML format for Hittite cuneiform manuscripts. The data within TLH^{dig} ensures a precise representation of the original inscriptions and at the same time preserves information critical for scholarly research. Note, however, that it does not faithfully represent published text editions.

The dataset chosen for the transformation consists of 8,898 XML files, each corresponding to a unique text ID, and encompasses 145 CTH entries (see Fig. 3). The majority of the texts belong to ritual and festival genres, which are the most represented, accounting for 115 entries (107 entries under festival and cultic texts, 8 entries under rituals). This includes such texts as the Kizzuwatna rituals and seasonal festivals, with other genres significantly less represented. Foreign-language texts in Hattic, Hurrian, Luwian, and Palaic account for 24 entries, while cult inventories, administrative texts, mythology, and divination are represented by 1 entry each. Additionally, miscellaneous texts are categorized under Varia comprising 2 entries. These files serve as the raw material for transformation, requiring extensive processing to extract and structure the information for further linguistic, philological, and computational analysis, including applications in machine learning and deep learning.

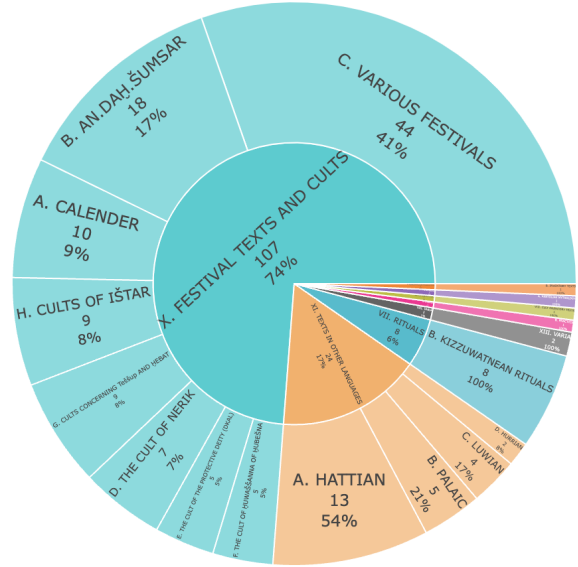


Figure 3: Distribution of texts in the dataset

The XML format captures multiple layers of information essential for Hittitological studies. Both transliteration and transcription (also known as normalization in the literature) are included, allowing for a comprehensive analysis of the texts. About 50% of the texts are glossed, while 16% are completely broken, making glossing impossible. Of the glossed texts, 15% (8% of the total dataset) have been manually validated. Instead, a large number of morphological glossing possibilities has been generated through a rule-based system (Rieken, 2021). These glossing possibilities include multiple grammatical interpretations for individual words, often structured in a format where different cases, numbers, and forms are suggested (see Fig. 4). This variation is a direct consequence of the ambiguities inherent in cuneiform writing, where the same sign can represent multiple sounds or words depending on context (Weeden, 2011). The lack of explicit vowel notation and the polyvalence of signs require multiple possible readings to be considered in the glossing process.

This challenge of multiple possible interpretations and the need for disambiguation is precisely what led us to consider glossing as a task for LLM fine-tuning. Given that traditional rule-based approaches generate numerous possibilities but lack contextual decision-making capabilities, a large language model (LLM) fine-tuned on Hittite data could assist in predicting the most probable gloss based on broader linguistic patterns. By leveraging machine learning, we aim to improve the efficiency of annotation and enhance consistency in glossing,

addressing the inherent uncertainties in cuneiform interpretation and at the same time incorporating philological insights.

```

'LUGAL'-u$ {'lg': 'Hit',
'mrp1': 'LUGAL=u-@König( a → FNL(u).NOM.SG.C) { b → ACC.PL.C}@28.3.1.1e',
'mrp2': 'LUGAL=ma-@Sarrumma( a → DN.STF) { b → DN.HURR.ABS}@36.1.1.1 += u$@PPRO.3PL.C.ACC@e',
'mrp3': 'LUGAL=ma-@Sarrumma( a → DN.STF) { b → DN.HURR.ABS}@36.1.1.1 += u$@PPRO.3PL.C.ACC@e',
'mrp4': 'LUGAL=u-@König( a → NOM.SG(UNM)) { b → ACC.SG(UNM)) { c → NOM.PL(UNM))
{ d → ACC.PL(UNM)) { e → GEN.SG(UNM)) { f → GEN.PL(UNM)) { g → D/L.SG(UNM))
{ h → D/L.PL(UNM)) { i → ALL(UNM)) { j → ABL(UNM)) { k → INS(UNM))
{ l → VOC.SG(UNM)) { m → VOC.PL(UNM))@28.3.1.1.1 += u$@PPRO.3PL.C.ACC@e'

da-ra-i1 {'lg': 'Hit',
'mrp1': 'da-/d-@nehen@35G.PRS@II.2e',
'mrp2': 'taye/a-@setehen@25G.IMP@7.7.5e',
'mrp3': 'da1-/te-/ti(ya)-@setzen@ a → 35G.PRS) { b → 25G.IMP}@II.6.1e',
'mrp4': '@()@HURR@e'

GA.KIN.AG {'lg': 'Hit',
'mrp1': 'GA.KIN.AG=@Kase( a → NOM.SG(UNM)) { b → ACC.SG(UNM)) { c → NOM.PL(UNM))
{ d → ACC.PL(UNM)) { e → GEN.SG(UNM)) { f → GEN.PL(UNM)) { g → D/L.SG(UNM))
{ h → D/L.PL(UNM)) { i → ALL(UNM)) { j → ABL(UNM)) { k → INS(UNM))
{ l → VOC.SG(UNM)) { m → VOC.PL(UNM))@29.1.1e'

HUR.SAG-nli {'lg': 'Hur',
'mrp1': '① HUR.SAG@31d eines Berges@FNL(n).D/L.SG@28.14.2e',
'mrp2': '① HUR.SAG-n=1@Berg( a → FNL(-i).HURR.ABS.SG) { b → STF}@30.10.4.1e'

{{LÜ}}SANGA {'lg': 'Hit',
'mrp1': 'SANGA=@Priester @ { a → NOM.SG(UNM)) { b → ACC.SG(UNM)) { c → NOM.PL(UNM))
{ d → ACC.PL(UNM)) { e → GEN.SG(UNM)) { f → GEN.PL(UNM)) { g → D/L.SG(UNM))
{ h → D/L.PL(UNM)) { i → ALL(UNM)) { j → ABL(UNM)) { k → INS(UNM)) { l → VOC.SG(UNM))
{ m → VOC.PL(UNM)) @ 28.1.1 @ (LÜ)',
'mrp2': 'SANGA=@Priester @ { a → NOM.SG(UNM)) { b → ACC.SG(UNM)) { c → NOM.PL(UNM))
{ d → ACC.PL(UNM)) { e → GEN.SG(UNM)) { f → GEN.PL(UNM)) { g → D/L.SG(UNM))
{ h → D/L.PL(UNM)) { i → ALL(UNM)) { j → ABL(UNM)) { k → INS(UNM)) { l → VOC.SG(UNM))
{ m → VOC.PL(UNM)) @ 28.2.1.1 @ (LÜ)',
'mrp3': 'SANGA=@Priester @ { a → NOM.SG(UNM)) { b → ACC.SG(UNM)) { c → NOM.PL(UNM))
{ d → ACC.PL(UNM)) { e → GEN.SG(UNM)) { f → GEN.PL(UNM)) { g → D/L.SG(UNM))
{ h → D/L.PL(UNM)) { i → ALL(UNM)) { j → ABL(UNM)) { k → INS(UNM))
{ l → VOC.SG(UNM)) { m → VOC.PL(UNM)) @ 28.1.1.1 @ (LÜ)'

```

Figure 4: A word-by-word annotation of a Hittite text made by a rule-based algorithm

Besides linguistic glossing, the XML data also encodes a range of philological annotations that provide critical context for text interpretation. Elements such as Sumerian and Akkadian logograms are explicitly marked, preserving distinctions between phonetic and logographic writing. Additional annotations track features such as scribal corrections, erasures, and textual additions, including elements that were likely intended by the scribe but are missing, as well as those that appear in the text but may not belong based on scholarly evaluation. These details are crucial for reconstructing the original meaning of the texts, reflecting both the complexities of the writing system and the interpretative challenges faced by modern researchers.

Additionally, the morphological glossing contains references to Hoffner and Melchert (2024), which is the most up-to-date Hittite reference grammar. Annotations often include language identifiers such as Hittite, Hattian, Hurrian, Luwian, and Palaic, along with a set of grammatical possibilities for each term. An additional field is included where one or more glossing options are marked as preferable. In cases where only one option is selected, it is typically human-verified, but for a portion of the material, selections have been made automatically without direct manual confirmation.

XML provides a structured and detailed encoding format, yet, it is not always suitable for compu-

tational analysis. Many statistical and corpus-based research methods require a tabular structure, such as CSV, to efficiently process and compare large datasets. Transforming XML into CSV allows for easier searching, filtering, and querying of linguistic features and at the same time makes the data more accessible for machine learning models and text analysis tools. The structured format also facilitates cross-document comparisons, ensuring that the rich philological and linguistic information in TLH^{dig} can be efficiently analyzed and used by other researchers, both in computer and data science, as well as ancient language scholars.

3.2 Reframing Text and Annotation

The transformation of XML-encoded Hittite texts required careful consideration of both segmentation practices and annotation preservation. Unlike modern languages, Hittite cuneiform is commonly written on clay tablets³ and lacks sentence level punctuation, which requires setting up additional algorithms for sentence boundaries mark-up. According to standards in the field, our dataset is primarily segmented at the cuneiform tablet line level, rather than the sentence or clause level. While some genres and text types contain explicit clause divisions (e.g. rituals and festivals), many do not, making line-based segmentation the most consistent and practical approach (Gordin, 2015). Additionally, the fragmentary nature of many sources further complicates sentence segmentation, because missing portions often obscure syntactic structure at the sentence level.

Although the source dataset is organized around line divisions, the transformation process ultimately operates at the word level. We extract and process individual words from the XML structure, so that each token retains its full set of grammatical, lexical, and philological annotations. At the same time, we preserve metadata from the original line structure, including line numbers, obverse and reverse distinctions (Vs./Rs. in the German annotation, or obv./rev. in the English one), and other positional markers, allowing for alignment with the

³While clay tablets were the primary medium for Hittite cuneiform writing, several other materials were also used. Of special importance were metal tablets, particularly bronze (exemplified by the unique Bronze Tablet containing the treaty between Tudhaliya IV and Kuruntiya, Bo 86/99), where wedges were incised rather than impressed. Stone was used for monumental inscriptions in Hieroglyphic Luwian, and wooden writing boards played a significant role in Hittite administration, economy, and cult practices, though few examples survive due to their perishable nature (Camarasano, 2024).

manuscript layout.

The transformation process was designed to maintain a clean primary text representation while storing all linguistic and philological annotations as additional structured fields. Initially, we assumed that achieving both readability and full annotation retention would require compromises. However, as the transformation progressed, it became evident that all linguistic and philological annotations could be preserved as additional fields. This method yields a structured format, in which the core text remains readable, and every nuance documented in the original annotations is retained.

The processed text uses Hittite transcription conventions, including broken marks, determinatives, and other editorial notations for scribal practices, complying with HPM standards. Meanwhile, all philological and linguistic metadata—such as glossing, language identification, restorations, erasures, uncertain readings, *mater lectionis*, and editorial comments—are preserved separately in structured fields. This approach, which is the key methodological insight of this paper, enables researchers to work both with the text without annotations and with its full scholarly annotations, ensuring that no interpretative detail is lost.

3.3 Data Processing and Transformation

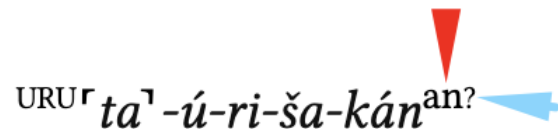
The transformation of XML-encoded Hittite texts into a structured tabular format follows a multi-step pipeline designed to extract, normalize, and organize linguistic and philological data. This process was implemented using Python, utilizing lxml for XML parsing (Shipman, 2014), pandas for data handling, and regular expressions (re) for text cleaning and refinement.

Each XML file was processed to extract and structure relevant linguistic and philological information. Using lxml's XPath functionality, the script identified line markers to track text segmentation, tokenized words with all their attributes. Additionally, it distinguished between different language layers, identifying content in Hittite, Akkadian, Sumerian, Hurrian, Luwian, Palaic, and Hattic. Once extracted, the data was mapped to a structured format, preparing it for subsequent normalization, parsing, and computational analysis.

As for the annotations directly within the text, they were preserved as independent fields. This step helps maintain that each annotation type was correctly mapped. The structured parsing of linguistic and philological data served as the founda-

tion for normalization.

An essential part of our approach was to expand the annotation structure by introducing additional fields that retained philological and linguistic information separately from the core text. These fields included annotations for subscript markings, *matres lectionis*, numerical markers, sign-based annotations, corrections, erased text, editorial insertions, *rasura* and uncertain *rasura*, missing text markers, editorial comments, and references to other texts or glossaries (see example in Figs. 5 and 6).



URUṛ ta¹-ú-ri-ša-kánan?

Figure 5: KBo 51.127+ (CTH 615) (Frg. 1+2) Rs.? III 7' / 3'



{URUṛ}ta¹-ú-ri-ša-kán kán<materlect c="an"/> <corr c="?" />

Figure 6: An example of a word's annotation as represented in XML and in CSV.

Since the XML format includes glossing generated by a rule-based algorithm, producing up to 40 possible glossing variations for a single word, parsing requires identifying and extracting these multiple interpretations. In addition to preserving all algorithmically generated glossing possibilities, the parsing process searched for and isolated the human-validated selection whenever available. This step allowed us to distinguish between computationally generated glosses and those verified by scholars.

In cases where no human-validated gloss was available, the dataset retained all generated possibilities without assigning a default selection, which allowed for future verification and computational processing. We preserved these alternative interpretations specifically to support future research efforts, ensuring that subsequent scholars would have access to the complete range of potential readings rather than being limited by our preliminary assessments.

One of the primary challenges in the parsing process was establishing an optimal parsing sequence to prevent data loss or unintended modification. Due to the complexity of the XML structure, exe-

cutting transformations in an incorrect order risked removing or altering certain elements before they could be fully extracted.

The final stage of data processing involved the integration of Unicode representations to enhance the dataset’s interoperability with computational tools and digital cuneiform research frameworks. Transliteration sequences were systematically mapped to their corresponding cuneiform Unicode characters which allows the use of Unicode in further analysis.

3.4 Format selection

The selection of an appropriate data format was a crucial consideration in ensuring both computational accessibility and philological integrity. Given the structured nature of the dataset and its diverse applications, three primary formats were evaluated: CSV, JSON, and YAML. Each format presents distinct advantages depending on the intended mode of analysis and data processing requirements.

The dataset was initially released in CSV format, prioritizing simplicity, interoperability, and compatibility with statistical analysis tools, machine learning frameworks, and database management systems. The tabular structure of CSV facilitates efficient numerical and textual data processing, making it well-suited for corpus-based linguistic research. However, CSV lacks the ability to encode hierarchical relationships, requiring additional strategies to represent nested linguistic annotations.

In contrast, JSON and YAML provide hierarchical and flexible data structures, making them more appropriate for storing multi-layered annotations, glossing alternatives, and complex linguistic metadata. JSON, widely used in computational linguistics and NLP applications, supports structured querying and integration with automated processing pipelines, while YAML offers a human-readable alternative for philological research (Wang, 2022).

CSV was selected as the primary output format for the initial dataset release, future expansions will incorporate JSON for structured annotation storage and YAML for enhanced interpretability in philological studies.

4 Analysis and Insights

The processed dataset consists of 558,349 tokens, structured with detailed linguistic and philological annotations. The data was analyzed to assess the distribution of glossed words, the extent of human validation, and the proportion of broken or fragmentary text (see Table 1).

	Glossed	Validated	Broken
True	297,095	47,908	87,782
False	261,159	510,346	470,472

Table 1: Distribution of glossed, validated, and broken tokens.

Of the total tokens, 297,095 (53.2%) were assigned glosses through rule-based annotation. However, only 47,908 glosses (16.1%) of those annotated received human validation, confirming the need for further refinement in automatic glossing methods. Text integrity analysis showed that 87,782 tokens (15.7%) were identified as broken or fragmentary, limiting their potential for linguistic annotation.

These findings highlight both the strengths and limitations of the dataset, particularly regarding the reliance on rule-based glossing and the importance of human validation in refining automatic annotation strategies. We are, however, postponing enhancing glossing accuracy through machine learning to future research, where manually validated glosses would create a probabilistic glossing model.

The additional philological and linguistic annotations are not as widely represented across the dataset, but are still retained due to their significance for the analysis. Various elements of markup, such as subscript markings, determinatives, corrections, and editorial interventions, appear in relatively small proportions, with some features occurring in only a few thousand or even hundred instances. Despite their lower frequency, these annotations provide critical insights into scribal practices, textual transmission, and linguistic variation.

The presence of so many glossing possibilities for a single word highlights the morphological ambiguity inherent in the corpus. This is particularly evident in polysemous words, homographs, and inflected forms, where multiple interpretations arise due to overlapping grammatical or lexical functions. Despite the extensive output of the rule-based glossing system, only 16.1% of glossed words received

txtid	lnr	cth_number	word	translit	gloss	trans_de
IBoT 1.30+	Vs. 1	821	LUGALuš	ʽLUGALʽ ₁ -uš	FNL(u).NOM.SG.C	König
IBoT 1.30+	Vs. 1	821	kuapi	ku-wa-pí	CNJ	sobald als
IBoT 1.30+	Vs. 1	821	DINGIRaš	DINGIR(MEŠ)-aš	D/L.PL	Gottheit
IBoT 1.30+	Vs. 1	821	aruazi	a-ru-wa-a-ez-zi	3SG.PRS	sich verneigen
IBoT 1.30+	Vs. 1	821	GUDU ₁₂	(LÚ)GUDU ₁₂	NOM.SG(UNM)	Gesalbter
IBoT 1.30+	Vs. 1	821	kišan	kiš-an	DEMadv	in dieser Weise

Figure 8: First lines of the published dataset

$$\text{ROUGE-1} = \frac{\sum_{unigram \in \text{Reference}} \text{Count}_{\text{match}}(unigram)}{\sum_{unigram \in \text{Reference}} \text{Count}(unigram)} \quad (1)$$

The original pre-trained model showed very poor results, with ROUGE-1 at 0.0255 and ROUGE-2 at 0.02, indicating that it failed to generate meaningful translations. However, after fine-tuning, the instructed model demonstrated a substantial improvement, achieving **ROUGE-1 at 0.895** and **ROUGE-2 at 0.27**, reflecting a significant gain in translation accuracy.

These results suggest that while T5 was ineffective for glossing, it can be successfully fine-tuned for translation tasks in a structured linguistic dataset. This highlights the importance of task selection in NLP applications for low-resource languages. Future work could explore alternative transformer-based architectures specialized for glossing, such as morphology-aware models, or integrate linguistic priors to improve the accuracy of morphological annotation in Hittite and other ancient languages.

```

Input: ekuzi
Expected translation: trinken
Generated translation: (Gefäß)

Input: QA-TAM-Māpat
Expected translation: ebenso
Generated translation: ebenso

Input: DINGIRnana
Expected translation: Gottheit
Generated translation: (Priesterin)

```

Figure 9: Examples of translation by instructed model

6 Conclusion

This study has outlined the creation and implementation of a computationally annotated corpus of Hittite texts, leveraging XML-encoded linguistic and philological data for structured analysis. The research contributes to the evolving field of Ancient Language Processing (ALP) by providing a standardized and machine-readable dataset, facilitating advanced linguistic inquiries and computational methodologies for Hittite studies.

Through the transformation of XML-based textual data into structured formats such as CSV, this work ensures accessibility for both traditional philological research and modern computational applications. The challenges inherent to Hittite corpus development—such as the complexity of XML annotations, the representation of multiple linguistic layers, and the integration of fragmented texts—demand a methodological approach that preserves philological accuracy. This transformation from XML to more computationally accessible formats represents not just a technical conversion but an essential paradigm shift for ancient language processing, moving from formats optimized for philological documentation toward those that enable computational analysis at scale.

The study also underscores the limitations of current transformer-based language models, such as T5, for morphological glossing in low-resource ancient languages, highlighting the need for hybrid approaches that integrate rule-based linguistic knowledge with probabilistic modeling.

Certain questions that have not been considered in this paper are postponed for future research. These include: refining syntactic annotation through dependency-based models, improving neural network performance for gloss prediction via fine-tuning on enriched datasets, and expanding

the corpus to include a broader range of Hittite textual genres. In this way, the current study provides solid foundation for these tasks.

Our data is available as supplementary information to this paper via the [following link](#).

7 Acknowledgments

This work would not have been possible without the constant collaboration of Daniel Schwemer and Gerfrid Müller (JMU Würzburg), who provided early access to the XML files, as well as all the contributors to the HPM projects, whose work allowed us to release this new dataset. All data is made available under a [CC BY-SA 4.0](#).

References

- Michele Cammarosano. 2024. Writing on wood in hittite anatolia. In Marilina Betrò, Michael Friedrich, and Cécile Michel, editors, *The Ancient World Revisited: Material Dimensions of Written Artefacts*, volume 37 of *Studies in Manuscript Cultures*, pages 165–205. De Gruyter, Berlin-Boston.
- Elmar Edel. 1994. *Die Ägyptische-hethitische Korrespondenz aus Boghazköi*, volume 2. Westdeutscher Verlag, Opladen.
- Shai Gordin. 2015. *Hittite Scribal Circles. Scholarly Tradition and Writing Habits*. StBoT 59. Harrassowitz Verlag, Wiesbaden.
- Harry Hoffner. 2009. *Letters from the Hittite Kingdom*. Society of Biblical Literature, Atlanta.
- Harry Hoffner and H. Craig Melchert. 2024. *A Grammar of Hittite Language. Part 1. Reference Grammar*. Penn State Press.
- Guglielmo Inglese. 2016. [Annotating the syntax of fragmentary texts: The case of hittite](#). Presentation at the Workshop "Formal Representation & Digital Humanities: text, language and tools", University of Verona.
- Chin-Yew Lin. 2004. [Rouge: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Anke Lüdeling and Merja Kytö, editors. 2008. *Corpus Linguistics: An International Handbook*. De Gruyter, Berlin-New York.
- Maria Molina. 2016. Syntactic annotation of the hittite corpus: Problems and principles. In *Proceedings of the Workshop on Computational Linguistics and Language Science*, CEUR Workshop Proceedings.
- Maria Molina and Alexei Molin. 2016. [In a lacuna: Building a syntactically annotated corpus for a dead cuneiform language \(on the basis of hittite\)](#). In *Proceedings of the International Conference "Dialogue-2016"*.
- Elisabeth Rieken. 2021. [hethiter.net/: Hfr-annotation \(2021-12-31\)](#). Online resource.
- John W. Shipman. 2014. [Python xml processing with lxml](#). Technical report, New Mexico Tech Computer Center.
- Theo van den Hout. 2002. Another view of hittite literature. In S. de Martino and F. Pecchioli Daddi, editors, *Anatolia Antica. Studi in Memoria di Fiorella Imparati*, pages 857–878. Florence.
- Theo van den Hout. 2008. A classified past: Classification of knowledge in the hittite empire. In R. D. Biggs, J. Myers, and M. T. Roth, editors, *Proceedings of the 51st Rencontre Assyriologique Internationale held at the Oriental Institute of the University of Chicago, July 18-22, 2005*, pages 211–219. Chicago.
- Blair Wang. 2022. [Programming for qualitative data analysis: Towards a yaml workflow](#). *ACIS 2022 Proceedings*.
- Mark Weeden. 2011. *Hittite Logograms and Hittite Scholarship*. StBoT 54. Harrassowitz Verlag, Wiesbaden.
- Emma Yavasan and Shai Gordin. 2024. [Glossed hittite texts with german translation for machine learning](#). Zenodo.
- Emma Yavasan and Maria Molina. 2024. [Universal dependencies for the queen puduheba](#). Presentation at Digital Humanities and Social Sciences (DHSS) in Israel, Tel Aviv University.
- Petr Zemánek. 2007. A treebank of ugaritic. annotating fragmentary attested languages. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*, NEATL, pages 212–218, Bergen.

Appendix: List of CTH entries in the newly released dataset

See Fig. 10

topic	category	cth_list	cth_count	txt_count	total_word_count
II. ADMINISTRATIVE AND TECHNICAL TEXTS	B. INVENTORY TEXTS	CTH 231 Lists of administrators (LUAGRIG)	1	2	188
	IX. DIVINATION	CTH 581 Letters about oracles	1	1	86
	V. MYTHOLOGY	CTH 330 Ritual for the Storm-god of Kuliwišna	1	45	5781
VII. RITUALS	B. KIZZUWATNEAN RITUALS	CTH 475 Ritual of Paliya, king of Kizzuwatna, CTH 480 Ritual of Šamuḫa, CTH 481 Expansion of the cult of the goddess of the night, CTH 482 Reform of the cult of the goddess of the night of Šamuḫa by Mursili II, CTH 488 Ritual referring to Ḫamrišhara, CTH 494 Ritual of the queen and her sons for the goddess NIN.GAL, CTH 500 Fragments of Kizzuwatnaean festival and magical rituals	8	579	39543
VIII. CULT INVENTORY TEXTS	VIII. CULT INVENTORY TEXTS	CTH 523 Provisions (melqātu) for local festivals	1	18	1188
X. FESTIVAL TEXTS AND CULTS	A. CALENDER	CTH 591 Festival of the Month, CTH 592 Spring and Herbst festival in Zippalanda, CTH 593 Spring festival on Mt. Tapala, CTH 594 Spring festival at Tippiwa, CTH 595 Spring festival fragments, CTH 596 Autumn festival fragments, CTH 597 Winter festival for the Sun-goddess of Arinna, CTH 599 Journey of the sacred hunting bag in winter, CTH 600 New year's festival	10	181	18097
	B. AN.DAḪ.ŠUMSAR	CTH 604 AN.DAḪ.ŠUMSAR, outline tablets, CTH 605 AN.DAḪ.ŠUMSAR, day 1, CTH 606 AN.DAḪ.ŠUMSAR, day 2, CTH 608 AN.DAḪ.ŠUMSAR, days 7-8, CTH 609 AN.DAḪ.ŠUMSAR, day 11, CTH 610 AN.DAḪ.ŠUMSAR, days 12-13: temple of Ziparwa, CTH 611 AN.DAḪ.ŠUMSAR, day 14-15: for the Sun-goddess of the earth, CTH 612 AN.DAḪ.ŠUMSAR, day 16: temple of Zababa, CTH 613 AN.DAḪ.ŠUMSAR, days 18-19: for the Storm-god of lightning, CTH 614 AN.DAḪ.ŠUMSAR, day 21: for the deity IBURAS, CTH 615 AN.DAḪ.ŠUMSAR, days 22-25: for Ištar of Hattiana, CTH 616 AN.DAḪ.ŠUMSAR, day 26: for Ea and his circle, CTH 617 AN.DAḪ.ŠUMSAR, day 32: for the protective deity of Tauris, CTH 618 AN.DAḪ.ŠUMSAR, day 33-34: on Mt. Puškurunuwa, CTH 619 AN.DAḪ.ŠUMSAR, day 38: rain festival, CTH 620 AN.DAḪ.ŠUMSAR in Ankuwa for the goddess Kataḫḫa, CTH 621 unassigned (formerly 'AN.DAḪ.ŠUMSAR, "first tablet"; see CTH 608), CTH 625 AN.DAḪ.ŠUMSAR, day 38: rain festival, CTH 626 Festival of haste (EZEN, nuntarriyašhaš), CTH 627 KILAM festival, CTH 628 (ḫ)išuwā- festival, CTH 629 Regular festival (EZEN, SAG.ÜŠ), CTH 630 Moon and thunder festival, CTH 631 Thunder festival, CTH 632 Festival for the ancestors?, CTH 633 Festival of the investiture of royal successor (EZEN, ḫuššamaš), CTH 634 Great festival of Arinna, CTH 635 Fragments of the festival of Zippalanda and Mt. Dajpa, CTH 636 Festival of Sarissa, CTH 637 Festival for the God of ḫiššaišhapa, CTH 638 Festival for Teliḫni, CTH 639 Fragments of the festival for Tiḫwatti, CTH 640 Fragments of festivals for Luwian deities, CTH 641 Cult of Išḫara, CTH 642 Festival fragments referring to the vegetation god Zihkurwa, CTH 643 Festival fragments referring to the god Ziparwa, CTH 644 Festival or ritual fragments referring to Pirinkir, CTH 645 Fragments of festivals for the netherworld deities, CTH 646 Fragments of festivals celebrated by the queen, CTH 647 Festivals celebrated by the Prince (DUMU.LUGAL...)	18	346	37606
	D. THE CULT OF NERIK	CTH 671 Offering and prayer to the Storm-god of Nerik, CTH 672 Monthly festival at Nerik, CTH 674 Fragments of the puḫulliya- festival of Nerik, CTH 675 Fragments of the festival in the ḫešta- house, CTH 676 Fragments of a purifications ritual in Nerik, CTH 677 Ration lists (lamnati), CTH 678 Festival fragments concerning the cult of Nerik	7	135	12517
E. THE CULT OF THE PROTECTIVE DEITY (DKAL)	CTH 681 Festival of Karabḫa, CTH 682 Festival for the protective deities, CTH 683 Renewal of the hunting bag for the protective deities, CTH 684 Festival for the protective deities of the river, CTH 685 Fragments of festivals for the protective deities	5	101	9988	
F. THE CULT OF HUWAŠŠANNA OF ḪUBEŠNA	CTH 690 List of festivals for Huwaššanna, CTH 691 The witašš(i)jaš festival, CTH 692 Fragments of the witašš(i)jaš festival, CTH 693 The šaḫḫan festival, CTH 694 Fragments of festivals for Huwaššanna	5	255	22050	
G. CULTS CONCERNING TEŠŠUP AND ḪEBAT	CTH 698 Cults of Teššup and Ḫebat of Aleppo, CTH 699 Festival for Teššup and Ḫebat of Lawazantiya, CTH 700 Enthronement ritual for Teššup and Ḫebat, CTH 701 Drink offering for the throne of Ḫebat, CTH 702 Ritual after the renewal of a temple of a temple of Ḫebat, CTH 703 Rituals of Muwalanni, priest of Kummanni, for Teššup of Manuzziya, CTH 704 Lists of Hurrian Gods in festivals, CTH 705 Lists of Hurrian Gods in festivals, CTH 706 Fragments of festivals for Teššup and Ḫebat	9	492	51473	
H. CULTS OF IŠTAR	CTH 711 Autumn festival for Ištar of Šamuḫa, CTH 712 Festival for Ištar of Šamuḫa, CTH 713 Ritual for Ištar of Tanninga, CTH 714 Festival for Ištar of Nineveh, CTH 715 Winter festival for Ištar of Nineveh, CTH 719 Festival for Ištar, ḫu(r)dumana, Aruna, CTH 720 Fragments of festivals for Ištar, CTH 721 Festival for Ištar of Mt. Armana, CTH 722 Festival for the Great Sea and the Armana- Sea	9	90	10576	
XI. TEXTS IN OTHER LANGUAGES	A. HATTIAN	CTH 733 Invocation of Hattian deities: language of gods, language of men, CTH 735 Hattian prayers or incantations, CTH 736 Song of the zintuḫi-women for the Sun-goddess, CTH 737 Festivals of Nerik (with Hattian recitations), CTH 738 Festival for the goddess Teššaišhapi, CTH 739 Festivals of the city of Tuḫumiya, CTH 740 Hattian litanies, CTH 741 Hattian songs of the women of Tiššaruliya, CTH 742 Hattian songs (SIR), CTH 743 Hattian antiphonal songs, CTH 744 Festival fragments with Hattian recitations, CTH 745 Hattian fragments, CTH 746 Hattian strophic songs	13	495	29617
	B. PALAIC	CTH 750 Festival for Ziparwa, CTH 751 Festival for the Palaic pantheon – bread-, meat- and drink-offerings in Palaic, CTH 752 Festival for the Palaic pantheon – ritual for the disappearing and returning deity, CTH 753 Festival with Palaic recitations, CTH 754 Palaic fragments	5	58	4723
	C. LUWIAN	CTH 770 Luwian ritual fragments, CTH 771 Tablet of Lallupiya (with Luwianisms), CTH 772 Festival(ritual)s of Išanuwa, CTH 773 Songs of Išanuwa	4	145	11669
	D. HURRIAN	CTH 785 Ritual for Mt. Ḫazzi, CTH 786 Hurrian deity lists	2	25	3364
XIII. VARIA	CTH 821 Kingship and divine authority, CTH 832 Hittite fragments with diverse content	2	652	15167	

Figure 10: Table of CTH entries