A Small Universal Dependencies Treebank for Hittite

Erik Andersen
Brandeis University
erikandersen@brandeis.edu

Benjamin Rozonoyer
Brandeis University
brozonoyer@brandeis.edu

Abstract

We present the first Universal Dependencies treebank for Hittite. This paper expands on earlier efforts at Hittite corpus creation (Molina and Molin, 2016; Molina, 2016) and discussions of annotation guidelines for Hittite within the UD framework (Inglese, 2015; Inglese et al., 2018). We build on the expertise of the above works to create a small corpus which we hope will serve as a stepping-stone to more expansive UD treebanking for Hittite.

1 Introduction

Hittite is an extinct language of the Anatolian sub-branch of the Indo-European language family. It was the main language of the Hittite kingdom (16th-13th centuries B.C.E.), and is recorded from the 18th to the 12th centuries B.C.E. (Molina and Molin, 2016; Molina, 2016). Knowledge of Hittite reached beyond the boundaries of the Hittite kingdom as far as Egypt. As the earliest attested Indo-European language, it remains vital to Indo-European studies and our understanding of the rest of the Anatolian sub-branch, all of whose languages – Luwian, Palaic, Lycian, Lydian, and Carian – are extinct (Dalby, 2004; Hoffner and Melchert, 2008a; Collins, 2012).

The Hittite empire left behind a wide range of texts, which can be classified according to linguistic time periods – Old Hittite, Middle Hittite, and New Hittite – with Middle Hittite acting more as a transitional period between the two (Melchert, 2007).

Hittite's fragmentary corpus of cuneiform tablets with extensive borrowing of signs from both Akkadian and Sumerian make the language challenging for treebank creation (Molina and Molin, 2016). As a dependable source of unfragmented text, we annotated original Hittite sentences presented in Hoffner and Melchert's tutorial (Hoffner and Melchert, 2008b), and which we had analyzed in a lecture setting. The sentences are drawn from a variety of texts, spanning legal, religious, and mythological, from the three linguistic periods. Despite minor diachronic developments in morphology and syntax (see Section 2), scribal recopying of older texts occasionally obscures a definitive chronological classification for the surviving texts. These considerations swayed us in favor of a single corpus for the three linguistic periods. We include the dating, whenever possible, in the sentence's metadata.

1.1 Grammatical sketch

Hittite is an SOV language that "shows the typical features of an older Indo-European language" in that it is synthetic and suffixing in its derivational and inflectional morphology (Hoffner and Melchert, 2008a). The language employs a rich noun case system and appears to display split-ergativity (see §3.2).

Hittite verbs display two main tenses, present and preterite, but they can be augmented with auxiliary verbs *ḫar(k)-* and *ēš-* to create more complex tenses, such as the analytic perfect. Verbs also display two basic moods: imperative and indicative.

Aspect marking is more complex. The three verbal suffixes *-ške-, -anna/i-, -šša-* appear to act as imperfect markers on verbs (Hoffner and Melchert, 2008a). However, not all verbs displaying incomplete action require them, and it is debatable whether they always act as imperfectives (Inglese, 2015) (see §3.2).

Even more unclear is the exact function of Hittite clausal connectives *nu*, *šu*, and *ta* and the topicalizing or contrasting particle *-(m)a* (see §3.1).

Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020), pages 1–7
Barcelona, Spain (Online), December 13, 2020

Category	**Example**
Hittite Word	*pár-ku-iš*
Hittite Word with Determinative	$^{\text{d}}$*A-la-lu*
Sumerogram	MUŠEN
Multiple Character Sumerogram	TA.ÀM
Akkadogram	*EL-LA-AM*
Hittite Word with Sumerian Plural	*up-pé-eš-šar*$^{\text{MEŠ}}$

Table 1: Examples of Hittite in narrow transcription

2 Previous work

Inglese (2015) discusses an annotation schema for a Hittite Universal Dependencies treebank. His work primarily concerns sentences originating from the Old Hittite *Zalpa's text*, so our treebank requires some other rules to account for grammatical conventions reflected only in later texts. These include use of the *-za* particle in nominal and "to be" sentences with 1st or 2nd person subject, and pronominal clitic repetition (see §28.32-42 and §30.19 of Hoffner and Melchert, 2008a).

Furthermore, some of Inglese's work must be updated to adapt it to the current Universal Dependencies 2.0, which was released subsequently to his paper.

Inglese et al. (2018) introduce a Hittite treebank in the PROIEL (Pragmatic Resources in Old Indo-European Languages) framework, whose treebanks for Old Church Slavonic and Latin have been mapped into the UD format.

Molina (2016) has done previous work on a large constituency MsSQL corpus of Hittite texts; we do not use it as a guide to our annotation.

3 Annotation

3.1 Orthography and tokenization

Hittite was initially written in cuneiform, with each word separated by space (Hoffner and Melchert, 2008a; Inglese, 2015). There are two common methods for modern transcription: narrow and broad. In narrow transcription, the boundary of each character is clearly delineated, whereas broad transcription more closely reflects the probable pronunciation. In the example below, the top line is provided in narrow transcription, and the second line in broad transcription:

(1)	zi-ik	am-me-el	É-na	le-e	ú-wa-ši
	zik	ammel	É-na	lē	uwaši
	you	my	house:ALL;SG	PROHIB	come:PRS;2SG [1]

You shall not come to my house. (from KUB 29.1 i 19-20 (OH/NS))

Hittite borrowed extensively from both Sumerian and Akkadian. Sumerograms are represented by non-italicized capital letters, as in the word MUŠEN "bird" or the word for "house" É above, and Akkadograms are represented using capital italic text, such as *EL-LA-AM* "free (ACC)." Hittite words are written using lowercase letters (Hoffner and Melchert, 2008a). Multiple adjacent Sumerograms are separated by a dot.

In Hittite, the determinative, featured only in the written language, is placed before or after a word to codify it as part of a category, and is transcribed with a superscript. For example, the determinative MUŠEN is used to indicate birds, and the determinative *d* (short for DINGIR) is used for deities. This behavior is also seen in both Sumerian and Akkadian. The Akkadian UD treebank (Kopacewicz, 2018), which uses narrow transcription, does not treat determinatives as separate from the words they qualify but attaches them using hyphens in the position they were found originally. Unlike the Akkadian treebank,

[1] We use the Leipzig conventions for our glosses in this paper.

we treat determinatives as separate words but include them in a multiword token with the noun that they qualify.

While we use hyphens, as in narrow transcription, to reflect word-internal cuneiform boundaries, we adopt the "=" from broad transcription to signal clitic boundaries. This hybrid approach allows the reader to immediately recognize clitics in the transcription, while being backwards-compatible with the writing system.

The absence of punctuation in Hittite made sentence splitting decisions non-trivial, since a significant number of "sentences" in Hoffner and Melchert's tutorial were comprised of at least two consecutive independent clauses without conventional coordinating conjunctions. Following Inglese (2015) and Molina and Molin (2016), we took the phrase connectors *nu*, *ta*, and *šu* (the last of which does not appear in our corpus) to delineate sentence boundaries whenever they stand at the beginning of an independent clause. Similarly, we exploited the non-emphatic clitics *-wa*, *-(m)a*, *-kan*, *-šan*, *-za*, *-ašta*, *-an*, *-apa* (the last two of which do not appear in our corpus). Whenever these appear at the start of an independent clause which is not the beginning of quoted speech introduced by a verb of saying, they signal a new sentence. All these discourse particles and clitics may be seen as connectives and give us a relatively clean heuristic for sentence tokenization. We did not split independent clauses which were strung together without such discourse connectives, and opted instead to use parataxis.

Employing this method of sentence tokenization, and treating determinatives and clitics as distinct words (including in complex Sumerogram multiword expressions such as DUMU.NAM.LÚ.U$_{19}$.LU-*(l)a-*, which corresponds to Hittite *dandukišnaš* DUMU-*(l)a-* "human being (*lit.* child of mortality)"), the statistics of our corpus come out to 136 sentences, 1309 words, and 970 (whitespace-separated) tokens.

3.2 Morphology and lemmatization

Hittite has the following cases: nominative, accusative, genitive, dative-locative, instrumental, ablative, ergative, allative and vocative. The ergative case appears when a neuter noun is the subject of a transitive verb (Hoffner and Melchert, 2008a). A neuter noun appears in the "absolutive" when it functions as the subject of an intransitive verb or as the direct object of a transitive verb. Hoffner and Melchert package this behavior under the name "nominative-accusative". The ergative does not occur in pronouns or common-gender nouns. Although the ergative case (Erg) does not occur in our corpus, we have annotated neuter nominative-accusative nouns as absolutive (Abs). Melchert (2011) provides a clause where the neuter subject (and corresponding neuter demonstrative) appears in the ergative case:

(2) maḫḫan=ta kāš tuppianza anda wemiyazzi
when=you this tablet:ERG reach:PRS;3SG
When this tablet reaches you (HKM 14:3-5)

In our annotation, we do not include the aspect feature as per Inglese's (2015) proposal. While Hittite uses the imperfective verbal suffixes *-ške-, -šša-,* and *-anna/i-*, they are not always used when the aspect is imperfective (see §24 of Hoffner and Melchert, 2008a). These suffixes perform a variety of functions, mostly iterative and durative in nature. In contrast, adverbs such as *kuitman* "while" can sometimes indicate an incomplete action without any contribution from the verb:

(3) nu ku-it-ma-an A-NA $^{\text{LÚ}}$-SANGA pa-a-an-zi
CONN while to priest go:PRS;3PL
And while they go to the priest (from KUB 5.6 i 39-41 (NH))

In (3), the subordinating conjunction *ku-it-ma-an* appears with the verb *pa-a-an-zi*, which does not use an imperfective suffix.

Out of the official Universal Dependency part of speech tagset, we used all values except for SYM and PUNCT, as Hittite does not make use of special symbols or punctuation like in English. We display the part of speech tags we used, together with the raw counts and percentages in Table 2.

Tag	Count	Percentage
NOUN	387	29.56%
VERB	208	15.89%
PART	164	12.53%
PRON	146	11.15%
CCONJ	88	6.72%
ADV	73	5.58%
PROPN	73	5.58%
ADP	49	3.74%
SCONJ	34	2.60%
NUM	31	2.37%
ADJ	21	1.60%
DET	16	1.22%
AUX	14	1.07%
X	4	0.31%
INTJ	1	0.08%

Table 2: Hittite UPOS tag statistics

Feature	UPOS	Values
Case	NOUN, VERB, PRON, PROPN, NUM, ADJ, DET	Nom, Acc, Gen, Dat, Abl Ins, All, Erg, Voc, Abs
Definite	NOUN	Cons
Gender	NOUN, VERB, PRON, PROPN, NUM, ADJ, DET	Com, Neut, Masc, Fem
Number	NOUN, VERB, PRON, NUM, ADJ, DET, AUX	Sing, Plur
NumType	ADV, NUM	Card, Ord
Person	VERB, PRON, AUX	1, 2, 3
Poss	PRON	Yes
PronType	PRON, DET	Dem, Ind, Int, Prs, Rel, Tot, [Neg]
Mood Tense VerbForm Voice	VERB, AUX	Ind, Imp Pres, Past Fin, Inf, Part, Sup, Vnoun Act, Mid
Language	*any (except* X*)*	Akk, Sum

Table 3: Feature values for Hittite grouped by UPOS

We show the morphological features we used, mostly adapted from Inglese, in Table 3. The *PronType=Neg* feature is included in Inglese (2015), but does not appear in our corpus.

Lemmas are taken from the stem as provided in Hoffner and Melchert's tutorial, and are always in broad transcription (Hoffner and Melchert, 2008b). While using the stem for a lemma is a convention in Hittitology, this results in different verbs being covered by the same lemma in some cases. To avoid this problem, we add *-#1-* or *-#2-* after the stem, following the example of the Hittite PROIEL annotation team (Inglese et al., 2018).

3.3 Dependency Relations

In our treebank, we introduced the following language-specific dependency relations:

acl:relcl – used to introduce relative clauses, subordinates the predicate of a relative clause to nominal that is modified.

advmod:emph – used for the emphatic particle *-pat* (and *-ila*, which is not in our corpus), which depends on the noun or pronoun it is attached to. For example, in *a-pu-un=pát* "that very one," *-pát* depends on the distal demonstrative *a-pu-un*.

Relation	Count	Relation	Count
root	136	dislocated	13
obj	126	expl:pass	13
nmod	107	parataxis	13
clf	97	xcomp	11
nsubj	96	orphan	9
obl	96	appos	8
advmod	89	aux	7
cc	87	cop	7
discourse	67	vocative	7
case	49	compound	6
conj	45	acl:relcl	4
advmod:loc	34	advmod:emph	4
mark	34	dep	4
advcl	30	ccomp	2
discourse:conn	26	csubj	2
nummod	26	expl	2
iobj	22	acl	1
det	15	flat	1
amod	13		

Table 4: Hittite dependency relation statistics

advmod:loc – used to subordinate the local particles *-šan, -kan, -ašta*, *-an, -apa* (where the last two do not appear in our corpus) to the predicate in the clause (see §28.43-47 of Hoffner and Melchert, 2008a). Inglese (2015) notes the complexity of these motion particles.

discourse:conn - used for the special phrasal connectives *nu*, *šu* and *ta* when they occur (typically sentence-medially) as discourse clause connectors that are neither subordinating nor coordinating. As per Inglese (2015), **cc** is used when these connectives occur sentence-initially (and act as coordinating conjunctions).

expl:pass – used for reflexive particle *-za* when it embodies a reflexive meaning, rather than change-of-state or first/second-person subject of a copular sentence (see §28.17-31 of Hoffner and Melchert, 2008a). This represents the current UD 2.6 version of Inglese's suggestion to use **auxpass:reflex**.

The **discourse** relation, often used for interjections, is very common. We annotated *-(m)a* as **discourse**, following Inglese (2015). We also used the discourse relation when *-za* acts as a 1st or 2nd person subject indicator in a copular sentence. In (4), the *-za* particle does not act reflexively, but indicates that the subject of the sentence is in the 2nd person:

(4) zi-g=a-a=z $^{\text{GIŠ}}$-ḫa-tal-ki-iš-na-aš
You=but=*za* hawthorn:GEN;SG
You are like the hawthorn
(from KUB 33.54 ii 13-14 (OH/NS) [restored version])

We display the count of each dependency relation in Table 4. Out of the official dependency relations, we used all the universal relations except for *list*, *goeswith*, *punct*, *reparandum*, and *fixed*.

While we used Inglese (2015) as a guide, we needed some other relations for distinct phenomena which occur more exclusively in New Hittite. For example, some New Hittite texts sometimes repeat pronouns within the same clause, possibly as a form of emphasis.

(5) nu=wa-r=a-an=za=an $^{\text{LÚ}}$-MU-TI$_4$=YA i-ya-mi
CONN=QUOT=him=REFL=him husband=my make:PRS;1SG
(I do not want to take my servant) and make him my husband. (from KBo 5.6 iv 6-7 (NH))

In (5), the accusative common-gender 3rd person clitic *-an* is repeated twice, though it refers to the same entity. Clitic doubling is not exclusive to Hittite; it can be seen in Bulgarian, whose UD treebank uses the relation **expl** for this phenomenon (Simov et al., 2015).

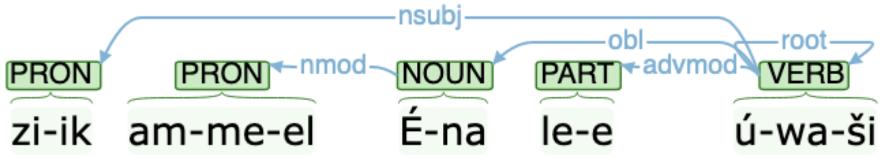

Figure 1: A Hittite UD Annotation of Sentence (1) using WebAnno

4 Implementation of some features proposed by Inglese (2015)

Inglese (2015) underlines the challenge of including philological information within the UD framework, and proposes the following extensions to reflect Hittitological information:

Language=Hitt, Akk, Sum. This feature indicates what language the written word is in. We use the Akk and Sum features, and leave Hittite words unmarked for this feature.

Determinative=1-16. Inglese introduces a Determinative feature which can take a value between 1 and 16, but does not treat determinatives as separate words in his annotation. Since these determinatives most often act as hypernyms, with a grammatical function of categorizing or classifying, we decided that it would be more consistent to have them relate to the head noun with **clf** (similar to numerical measure words in Chinese).

Hlemma (Hittite lemma for a word). We do not use this feature since our lemmatization is Hittite by default. If the underlying Hittite word for a Sumerogram or Akkadogram is not known, we resort to the Akkadian or Sumerian lemma, while retaining as much non-inflectional Hittite morphology in the lemma as possible. For instance, while the full Hittite lemma for the Sumerogram DUMU "son" is unknown, the uninflected stem ends in *-(l)a-*, so we record the lemma as DUMU-*(l)a-*.

Ntrans (Narrow transcription). Since we are working in narrow transcription by default, we do not include a feature for this.

5 Scribal peculiarities

Scribal idiosyncrasies occasionally complicate the annotation process:

(6) da-aš-šu-š=a-a=š-ši ^{d}-A-nu-uš ... pé-ra-an=še-et ar-ta
mighty=but=DAT;3SG;MASC Anu:NOM;SG ... before=GEN;3SG;MASC stand:PRS;3SG;MID
while Anu the mighty (foremost of the gods) stands before him. (from KUB 33.120 i 8-10 (NS))

The original cuneiform of sentence (6) uses NS (New Script), but clearly contains Old Hittite elements. Old Hittite prefers the genitive with postpositions to the dative-locative, sometimes appending the genitive possessive clitic to the postposition, as in *pé-ra-an=še-et* (Hoffner and Melchert, 2008a). However, this sentence also includes the extra New Hittite dative-locative element *-š-ši*, which could be the result of a scribal "correction". In this instance, the postposition *pé-e-ran* appears to govern two co-referential pronominal objects: *=še-et* and *-š-ši* (see §20.23 and §20.26 of Hoffner and Melchert, 2008a). This situation is similar to the clitic doubling, as in Bulgarian (Simov et al., 2015). For consistency, in (6) we have decided to mark the pronoun *-š-ši* as **expl** with respect to the verb in the clause, instead of making use of the **reparandum** relation, because there is no way to prove whether the extra pronoun is indeed a **reparandum**.

6 Conclusion

We have provided a seedling Universal Dependencies corpus for the earliest attested Indo-European language. With a view on making our work as extensible as possible, we have attempted to incorporate both the practical and the philological concerns voiced by Inglese (2015) for treebanking Hittite in the UD framework. We hope that this treebank will serve as a stepping stone for increased computational analysis of Hittite and typological research.

7 Acknowledgements

We would like to extend our gratitude to David Wright, Professor of Bible and Ancient Near East at Brandeis University, for offering his excellent course on Hittite last Spring.

References

Billie Jean Collins. 2012. *The Hittites and Their World.* SBL Press.

Andrew Dalby. 2004. *Dictionary of Languages*. Columbia University Press, Revised edition.

Harry A. Hoffner, Jr. and H. Craig Melchert. 2008a. *A Grammar of the Hittite Language. Part 1: Reference Grammar*. Eisenbrauns.

Harry A. Hoffner, Jr. and H. Craig Melchert. 2008b. *A Grammar of the Hittite Language. Part 2: Tutorial*. Eisenbrauns.

Guglielmo Inglese, Maria Molina, and Hanne Eckhoff. 2018. Incorporating Hittite into PROIEL: a pilot project. In Andrew U. Frank, Christine Ivanovic, Francesco Mambrini, Marco Passarotti, and Caroline Sporlede, editors, *Proceedings of the Second Workshop on Corpus-based Research in the Humanities*, pages 95–104.

Guglielmo Inglese. 2015. Towards a Hittite Treebank. Basic Challenges and Methodological Remarks. In M. Passarotti, F. Mambrini, and C. Sporleder, editors, *Proceedings of the Workshop on Corpus-Based Research in the Humanities (CRH).*

Kamil Kopacewicz. 2018. UD Akkadian PISANDUB.

H. Craig Melchert. 2007. Middle Hittite revisited. In A. Archi and R. Francia, editors, *VI Congresso Internazionale di Ittitologia*, pages 525–531.

H. Craig Melchert. 2011. The Problem of the Ergative Case in Hittite. In M. Fruyt, M. Mazoyer, and Dennis Pardee, editors, *Grammatical case in the languages of the Middle East and Europe. Acts of the international colloquium Variations, concurrence et evolution des cas dans divers domaines linguistiques*, pages 161–167.

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 *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016"*.

Maria Molina. 2016. Syntactic Annotation for a Hittite Corpus: Problems and Principles. In *Proceedings of the Workshop on Computational Linguistics and Language Science*.

Kiril Simov, Petya Osenova, and Martin Popel. 2015. UD Bulgarian BTB.