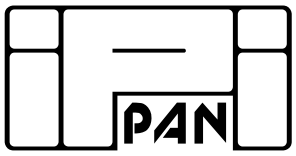
Proceedings of the Workshop on Corpus-Based Research in the Humanities (CRH)

10 December 2015
Warsaw, Poland

Editors:
Francesco Mambrini
Marco Passarotti
Caroline Sporleder

Sponsors



ISBN: 978-83-63159-19-1

Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5
01-248 Warszawa, Poland

`http://ipipan.eu/`

Preface

The workshop on *Corpus-Based Research in the Humanities* (CRH) is a direct descendant of the workshop on *Annotation of Corpora for Research in the Humanities* (ACRH), which was held three times: in Heidelberg (5.1.2012), Lisbon (29.11.2012), and Sofia (12.12.2013).
All three editions were co-located with the international workshop on *Treebanks and Linguistic Theories* (TLT), a tradition which we continue with CRH.

The new name was motivated by the wish to change the focus slightly, towards corpus-based research in the humanities in general. While the earlier editions focused on questions related to annotation and a number of papers in the current proceedings do so as well, we wanted to visibly broaden the scope of the workshop, as even the earlier editions of the workshop had attracted submissions that did not centre on the question of annotation. In fact, there are many scholars in the humanities who use textual corpora in their everyday work but are not interested in or just do not need to deal with annotation issues. This is partly due to the fact that many corpora still lack linguistic annotation at all, thus requiring scholars to use just the raw text for their research purposes. As our original motivation for initiating the ACRH workshop series was to bring together the often separate communities of (digital) humanities and computational linguistics and to foster communication and collaboration between them, we felt that the focus on annotation in the name of the workshop was undermining our intention by discouraging humanities researchers working with corpora to submit papers.

In addition to changing the name of the workshop, we made several smaller adjustments. First, we included several scholars from digital humanities in the programme committee. While such a mixed committee is not entirely without problems due to different reviewing cultures in digital humanities and computational linguistics, we still believe this a step in the right direction for bringing both communities closer together and assessing submissions from both areas fairly. Second, this year's call asked for long abstracts (up to six pages) rather than full papers. This reflects common practices in the digital humanities better and did help to attract more proposals. Finally, we decided to organise the workshop on a biannual basis instead of an annual one in order to reduce the workload of the organisers and reviewers and avoid competing with too many similar workshops too frequently.

In total we received 17 long abstracts by authors from 12 different countries in Europe and South and North America. Each submission was reviewed independently by three members of the programme committee in a double-blind fashion. After the reviewing process, we accepted 11 submissions. One further submission was moved from TLT to CRH because it was a better fit to the topics of CRH than those of TLT. The overall acceptance rate was 70.6%. This reflects the fact that the average quality of the abstracts was high and most of them received favourable reviews. Another positive observation is that a number of the workshop speakers are promising young scholars.

We hope you will enjoy the workshop and the proceedings and wish to thank all authors who submitted papers, the 19 members of the programme committee, Reinhard Förtsch, who kindly agreed to give the invited talk, and last but not least the local and non-local organisers of TLT-14 and in particular the chair of the local organisation committee, Adam Przepiórkowski.

The CRH Co-Chairs and Organisers
Francesco Mambrini (Deutsches Archäologisches Institut, Berlin, Germany)
Marco Passarotti (Università Cattolica del Sacro Cuore, Milan, Italy)
Caroline Sporleder (University of Göttingen, Germany)

Program Committee

Chairs:
Francesco Mambrini (Deutsches Archäologisches Institut, Berlin, Germany)
Marco Passarotti (Università Cattolica del Sacro Cuore, Milan, Italy)
Caroline Sporleder (University of Göttingen, Germany)

Members:
Monica Berti (Germany)
Federico Boschetti (Italy)
David Bouvier (Switzerland)
Neil Coffee (USA)
Lonneke van der Plas (Malta)
Dag Haug (Norway)
Neven Jovanovic (Croatia)
Mike Kestemont (Belgium)
John Lee (Hong Kong)
Alexander Mehler (Germany)
Roland Meyer (Germany)
Willard McCarty (UK)
John Nerbonne (The Netherlands)
Bruce Robertson (Canada)
Neel Smith (USA)
Uwe Springmann (Germany)
Melissa Terras (UK)
Sara Tonelli (Italy)
Martin Wynne (UK)

Towards a Hittite Treebank. Basic Challenges and Methodological Remarks

Guglielmo Inglese
IUSS Pavia\Università di Pavia\Università di Bergamo
guglielmo.inglese01@ateneopv.it

Abstract

The creation of a Hittite treebank constitutes quite a challenging task for computational linguists, as texts require a certain amount of preliminary work on philological issues before linguistic annotation can be effectively implemented. The aim of this paper is to survey a number of problems in laying the foundation of a resource which complies both with current digital annotation standards, as provided by UD, and with the philological practices established in the field of Hittitology.

1 Introduction

In this paper, I outline the first steps towards the creation of a treebank of the Hittite language, built within the framework of Universal Dependencies (UD). Hittite is the most anciently attested Indo-European language, and as such it is of primary importance for Indo-Europeanists, as well as for scholars interested in the puzzling linguistic scenario of the Ancient Near East.

Despite the compilation of grammars of the language (cf. Hoffner & Melchert [3]), and the development of electronic resources dedicated to Hittite texts (cf. Giusfredi [1]), many linguistic issues still remain open, so that a Hittite treebank is nowadays a *desideratum* of research. Still, the peculiarities of the sources pose a number of problems to computational linguists. First, unlike modern languages with digital-born texts, such as English, for which a number of NLP tools has been developed, Hittite texts must be manually annotated. Second, up-to-date linguistic annotation following current trends in NLP should be paired with the encoding of philological notes. The importance of the interaction between these two components cannot be underestimated in building a resource able to reach an audience as wide as possible. On the one hand, the design from scratch of a Hittite treebank constitutes an interesting case study for digital humanists, as it provides important clues as to how to deal with the digital encoding of cuneiform languages.[1] On the other hand, working with UD allows one to build a resource valuable for computational linguists as well. In what follows, I present the basic issues in the annotation of Hittite, taking as a case study the so-called 'Zalpa's text' (120 sentences, 1270 words).

2 Cuneiform script and philological problems

Hittite texts are written in cuneiform script, a syllabic script native of ancient Mesopotamia, which is exemplified in figure 1.

[1] Projects currently working on similar issues are the Ugaritic corpus by Zemánek [5], the *Annotated Cuneiform Luwian Texts* (http://web-corpora.net/LuwianCorpus/search/).

Figure 1: autography of the first line of KBo 22.2, from *HPM* [2]

This script displays a certain degree of complexity, as it contains both syllabic signs and logograms. In Hittite texts, logograms consist of Sumerian and Akkadian words used as graphic shortcuts for underlying Hittite words. To make Hittite texts available to non-specialists, cuneiform signs must be first transliterated, either in 'narrow transliteration' or in 'broad transcription' (cf. Hoffner & Melchert [3]). The former matches closely the script, with each sign transliterated as a syllable, as in *e-eš-zi* 'he is', whereas the latter consists of a phonological interpretation of words, as in *ešzi*. In addition, syllabic signs and logograms are graphically kept distinct. Hittite syllabic signs are written in lowercase italics letters, as in *e-eš-zi* 'he is', Sumerograms are transliterated in uppercase, as in MUNUS 'woman', and Akkadograms in uppercase italics, as in *TUP-PI* 'tablet'. Moreover, Sumerograms occur either in their root-form or bearing phonetic complements, i.e. final syllables marking Hittite case endings, as in DUMU-*an* 'son (acc.)'. Finally, a handful of Sumerograms, the so-called 'determinatives', were graphically preposed to nouns, and indicated the semantic class that nouns referred to. For instance, the sign URU in ^{URU}*Ne-e-ša* indicates that the noun *Ne-e-ša* belongs to class of city names.

In the treebank, words are given in broad transcription, thereby allowing users lacking a philological training to easily look into the corpus. Narrow transliteration and determinatives are stored as philological features. Also, both Sumerograms and Akkadograms are temporarily transcribed in uppercase.

Finally, one must consider the conservation status of tablets. As a matter of fact, most tablets are not entirely preserved, but rather broken or otherwise damaged. As a result, scholars usually need to reconstruct missing parts to assemble readable texts, either referring to less damaged copies, or drawing upon their expertise of the language. The conservation status of tablets brings about at least two practical issues: first, the integration status of each word should be properly annotated; second, it is necessary to develop a schema dedicated to the syntactic annotation of incomplete sentences.

3 Tokenization and philological features

Once transliteration and transcription have been performed, the first task to attend to is the tokenization. Luckily, we possess a great clue as to how to segment Hittite texts, as Hittite scribes separated words through blank spaces.

Still, tokenization cannot be restricted to the observation of blank spaces. First, one needs to split off clitics. For instance, the graphic word *nu-wa-aš-ša-an* must be split up as *nu=wa=šan*, that is, as the connective *nu* plus the particles *wa* and *šan*. Second, one needs to normalize texts by resolving elisions and assimilations, in order to improve the searchability of the treebank. Elision and assimilation often take place within clitic chains. For instance, the sequence

[2] URL: http://www.hethport.uni-wuerzburg.de/hetkonk/hetkonk_abfrage.php

an-da-ma-pa should be split as *anda=m=apa*, with *m* being the enclitic pronoun *mu* 'me', which undergoes elision before the particle *apa*. Assimilation takes place in sequences such as *n=at=ši*, in which the pronoun *at* 'it' assimilates to the pronoun *ši* 'him', thus yielding the graphic word *na-aš-ši*. Once normalized, the raw text is converted to CoNLL-U and further annotated.

The CoNLL-U format employed includes ten fields for each word line: ID, FORM, LEMMA, CPOSTAG, POSTAG, FEATS, HEAD, DEPREL, DEPS, MISC. Under FORM, I put words in broad transcription. Within this field, the two operations of clitic splitting and normalization are performed thanks to the token vs. word indexation. Each multiword is tokenized as one word, and it is indexed with integer ranges, whereas words are indexed with simple integers.

Moreover, philological data discussed in the previous section are associated to each word in the form of philological features, which are stored in the MISC field. First, I add three language-specific close-ranged philological features: Integration=0,1,2; Language=Hitt, Sum, Akk; Determinative=1-16. The Integration feature takes three values, and indicates whether a word is actually attested on the tablet (0), or, if restored, whether it is restored after other copies (1), or by the editor himself (2). The Language feature indicates whether cuneiform signs should be read as Hittite (Hitt), Sumerian (Sum), or Akkadian (Akk). Furthermore, determinative signs are handled thanks to the feature Determinative, which takes as value a numeric code corresponding to a specific determinative sign, based upon the list in Hoffner & Melchert [3]. Note that, since determinatives are stored as philological features, they do not visually surface in the treebank as tokens. Second, I add two open-ranged features, that is, Ntrans and HLemma. The former takes narrow transliteration of words and multiwords as value, whereas the latter indicates the corresponding Hittite lemma of logograms, if available. It should be stressed that these features do not aim at a full coverage of all issues in Hittite philology, but merely at providing users with notes essential for a proper reading of Hittite texts.

4 Morphological annotation

UD employs a three-layered model of morphological annotation, whereby each entry is lemmatized, assigned a POS label, and tagged for its morphological features. First, each word is given a LEMMA, based on dictionaries such as Tischler [4]. Note that I lemmatize Sumerograms and Akkadograms with Sumerian and Akkadian words, and store available Hittite equivalents through the Hlemma feature. Also, since forms can instantiate different lemmas, as in the case of *iyanzi*, which is a form of either *iya1*- 'go' or *iya2*- 'make', the LEMMA field can host multiple lemmas, separated by simple comma. POS tagging will be done in accordance with UD guidelines. Of the POS tags featured in UD, only PUNCT is left out, as Hittite texts display no punctuation.

The finer-grained morphological analysis is carried out through the use of morphological Feature=Value pairs, encoding both lexical and inflectional features of words. A general problem one is faced with in morphological annotation is that a number of features cannot be inherently assigned to single

tokens. As an example, let us consider the feature Aspect. Hittite displays the well-known IE verbal derivational suffix *-šk-*, which is often associated with imperfective aspect (Hoffner & Melchert [3]). Still, not every *šk*-suffixed verb is imperfective, nor unsuffixed verbs are always perfective. As a result, aspectual interpretation of verbs cannot simply rely on morphological marking. In the treebank, I prefer to exclude such borderline features, as they ultimately depend on the linguist's judgment on specific tokens, and their annotation is thus liable to a high degree of inconsistency.

I adopt the following UD lexical features: NumType=Card, Ord; Poss=Yes; PronType=Prs, Int, Rel, Dem, Tot, Neg, Ind. Note that though lexical features are inherent features of lemmas, this is not always the case, as for the pronoun *kuis* 'who', which is either interrogative, relative, or indefinite, depending on the context. To these, I add the language-specific feature Clitic, which indicates whether a token constitutes an independent word or a clitic item. Finally, I leave out the Reflexive feature, as reflexivity is not morphologically marked.

UD includes both nominal and verbal inflectional features. Nominal features adopted for Hittite are: Gender=Com, Neut, Masc, Fem; Number=Sing, Plur; Case=Nom, Acc, Dat, Gen, Voc, Inst, Abl, Dir, Erg; Definiteness=Red. Remarkably, not all UD features are relevant for Hittite. For instance, neither Animacy nor Degree have been included, as neither of them is morphologically coded. As for Gender, to the bipartite Com vs. Neut Hittite system, I added the Masc and Fem values, in order to account for the distinction in Akkadian pronouns such as *=ŠU* 'his' vs. *=ŠA* 'her'. Among the cases, it is much disputed whether the *-anza* ending on neuter nouns is an ergative case ending (Goedegebuure [2]). I do not take a stand in this scholarly debate, but for our purposes, it is more parsimonious to treat this ending as an ergative case, since treating it as a derivational suffix would artificially increase the amount of lemmas in the treebank. Sumerograms lacking phonetic complements are not tagged for Case. Definiteness of noun phrases is not explicitly marked in Hittite, but the Red value is retained to mark the reduced state of Akkadian nouns.

The verbal features adopted include VerbForm=Fin, Inf, Sup, Part; Mood= Ind, Imp; Tense=Pres, Past; Voice=Act, Mid; Person=1,2,3. Not every verbal features available in UD has been employed, as for Aspect, and some features display a reduced range of values. For instance, only indicative vs. imperative, and present vs. past distinctions are morphologically encoded. As for voice, only a two-fold opposition active vs. middle is marked by verbal endings.

5 Syntactic annotation

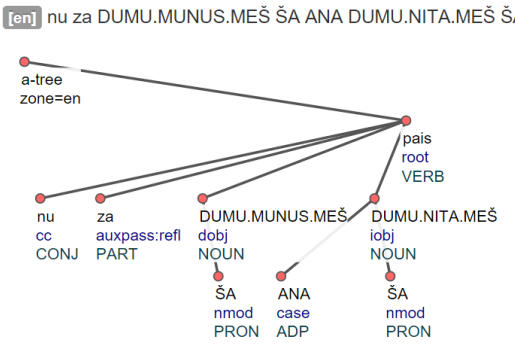
In UD, syntactic relations are represented as dependency relations between words, and are stored in the HEAD and DEPS fields. UD's universal set of dependency relations has been expanded with the following language-specific relations: *acl:relcl*, *advmod:emph*, and *auxpass:refl*. Moreover, I introduce the newly created relation *advmod:loc* to annotate Hittite so-called 'local particles'.

For reasons of space, I focus here on the annotation of complete sentences only, and leave the annotation of broken sentences for further research.

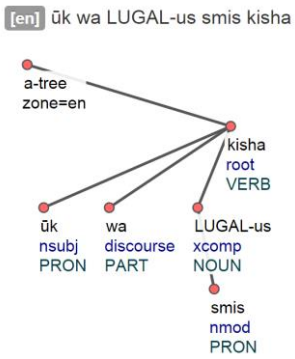
5.1 Clausal predicates and core arguments

Predicates constitute the head of the predication, and as such receive the *root* relation. When predicates are omitted, the *root* label is conventionally assigned to the first word in the sentence. In nominal predications, the complement of the copula takes the *root* relation, and the verb 'be' is tagged as *cop*.

Non-clausal core arguments of the predicate are tagged as *nsubj*, *dobj*, and *iobj*, as in (1) and (2). In my corpus, clausal arguments are not attested. Secondary predicates of predicative verbs are tagged as *xcomp*, as in (2).



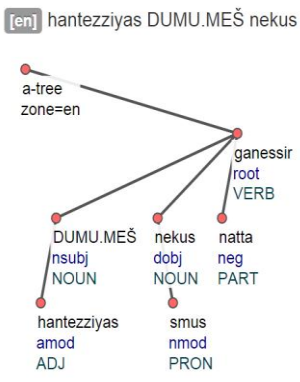
(1) *Core arguments of clausal predicates*
"And she gave her own daughters to her own sons."



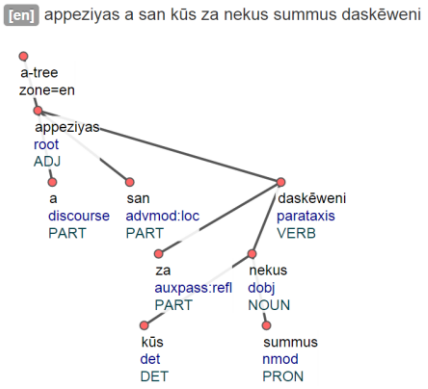
(2) *Secondary predicates*
"I have become your king."

5.2 Noun dependents

In the treebank, noun dependents include adjective (*amod*) and determinatives (*det*), as in (3) and (4). Numeral modifiers (*nummod*) are attested as well. Nouns can be modified by nominal modifiers (*nmod*) as well (cf. sec. 5.3).



(3) *Adjectives*
"The older sons did not recognize their sisters."

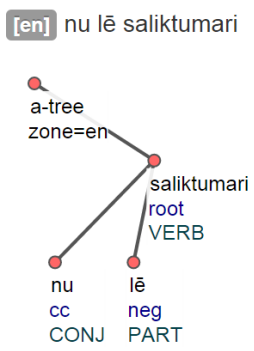


(4) *Determinatives*
"But the first one (said): we are taking these sisters of ours (in marriage)."

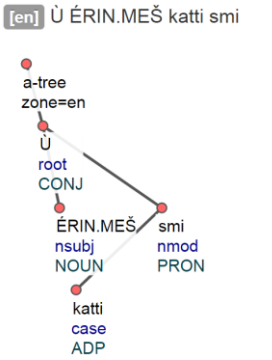
5.3 Non-core predicate dependents

Predicates can be modified by several non-core dependents, such as negation (*neg*), as in (5), or other adverbs (*advmod*). Note that preverbs are consistently tagged as *advmod*. Bare nominal modifiers are tagged as *nmod*, and adpositions

depend on their nominal or pronominal head through the *case* relation, as in (6). The same annotation is adopted for Akkadian prepositions.

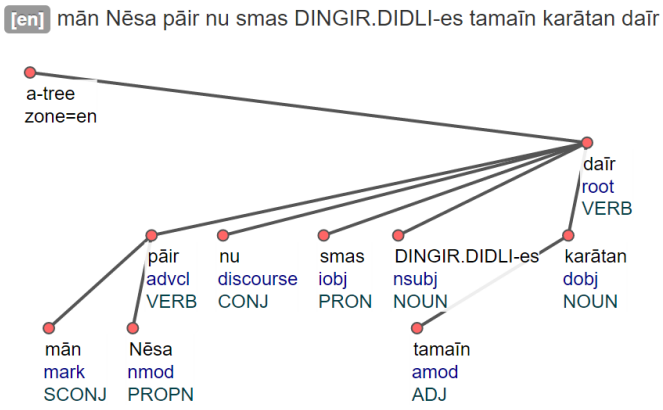


(5) *Negation*
"Now do not commit an outrage!"



(6) *Nmods and adpositions*
"And the troops (are) with me."

Clausal modifiers include adverbial subordinate clauses. In UD, subordinators depend on the predicate of the dependent clause as *mark*, which in turn depends on the predicate of the main clause as *advcl*, as in (7).



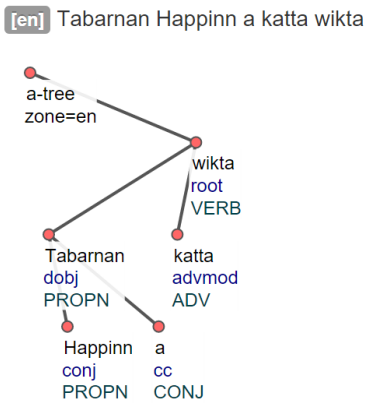
(7) *Adverbial clauses*
"When they went to Nesa, the gods gave them a different appearance."

5.3.1. Clause-linking devices

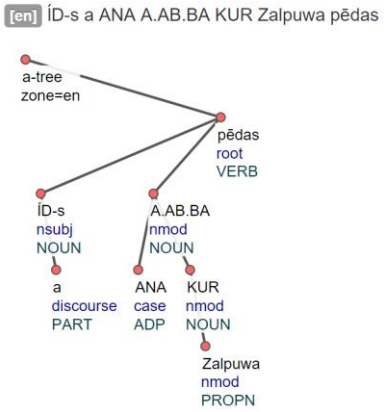
Hittite displays a number of non-subordinating connective devices, that is, sentence initial connectives *nu*, *šu*, and *ta*, and enclitic *=(y)a* and *=(m)a*.

The only proper coordinative conjunction is *=(y)a*, which links both sentences or phrases, as in (8). In UD, the first coordinand is annotated as the head of the coordination, on which the second one depends as *conj*. The conjunction depends on the first coordinand as well, and takes the *cc* relation.

The function of enclitic *=(m)a* is disputed. It arguably serves either as a topic-switching device or as a generic adversative connective. This difference is reflected in the annotation, as *=(m)a* depends via the *discourse* relation either on a topicalized noun, as in (9), or on the main predicate.



(8) *Coordinating* =(y)a
"And he demanded Tabarnas and Happis."

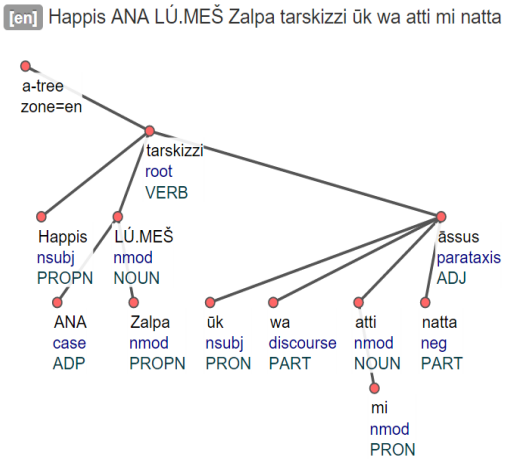


(9) *Enclitic* =(m)a
"But the river took (them) to the sea at Zalpuwa."

Unlike coordinating *=(y)a*, sentence initial connectives *nu*, *šu*, and *ta* establish additive links between sentences. The annotation mirrors these semantic and syntactic peculiarities. Sentences featuring connectives are always treated as independent from each other, and connectives are annotated as paragraph-initial conjunctions, that is, they depend upon the main predicate via *cc*, as in (5). In addition, connectives can be placed at the juncture between a preposed subordinate clause and its main clause. When this is the case, connectives depend on the predicate of the main clause as *discourse*, as in (7).

5.4 Direct speech

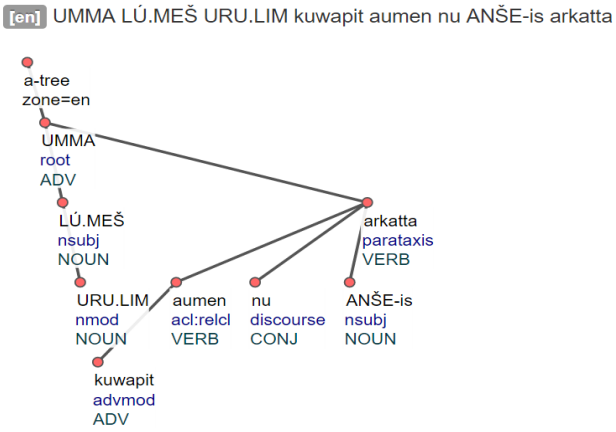
Hittite makes extensive use of reported direct speech, marked by the clitic particle *=wa(r)*. The particle depends as *discourse* on the predicate of the reported speech, which in turn depends via *parataxis* on the predicate of the main clause, as in (10).



(10) *Direct speech*
"Happis says to the men of Zalpa: «I (am) not dear to my father»."

5.5 *Relative clauses*

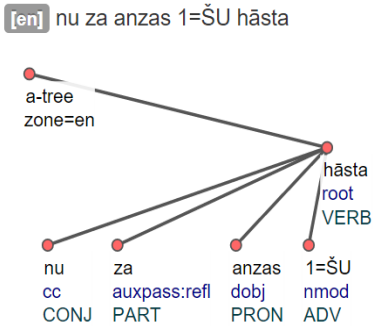
UD's *acl* relation is not suitable for the annotation of most Hittite relative clauses, as they usually do not modify a lexical head. Therefore, they are better treated as parallel to adverbial clauses, with their predicate depending on the predicate of the main clause via the *acl:relcl* relation, as in (11). Relative pronouns and adverbs depend on the predicate of the relative clause. In my corpus, only relative clauses introduced by the adverb *kuwapit* 'wherever' are attested, but this schema can also be extended to other kinds of relative clauses.



(11) *Relative Clauses*
"The men of the city (speak) as follows: «Wherever we have looked, a donkey will *arkatta*»."

5.6 *Reflexive* =za

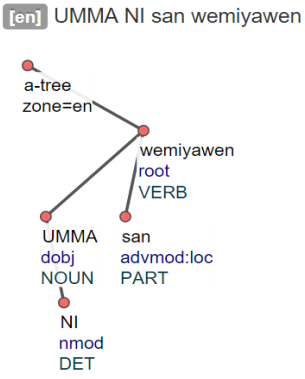
The so-called reflexive particle *=za* occurs in a number of contexts. First, it occurs with reflexive middle verbs and with transitive verbs indicating some sort of subject involvement. Moreover, in some cases it slightly modifies the semantics of the main predicate, or it occurs in nominal predications adding no detectable semantic contribution (Hoffner & Melchert [3]). Given the uncertainty in assigning *=za* a semantic and syntactic role, I conventionally annotate it as depending on the main predicate via the language-specific *auxpass:refl* relation, as in (12).



(12) *Reflexive* =za
"And (she) bore us at one time."

5.7 Local particles

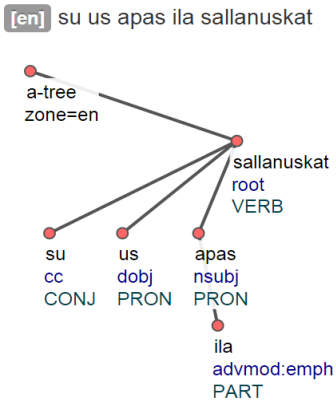
The cover term 'local particles' is employed to refer to the clitic particles *=an*, *=apa*, *=ašta*, *=kan*, and *=šan*. These particles arguably modify either the verb or some other local expression by adding spatial information, though this is much disputed. Therefore, I take them as depending either on the predicate or on adverbial modifiers via the newly created *advmod:loc* relation, as in (13).



(13) *Local particles*
"We found our mother."

5.8 Emphatic particles

Enclitic particles *=pat* and *=ila* give emphasis of some sort to nouns and pronouns which they are attached to. Both particles are annotated as depending on their phonological hosts as *advmod:emph*.



(14) *Emphatic particles*
"And she brought them herself."

6 Metadata

CoNLL-U format licenses the insertion of metadata in the form of comments to sentences. This proves extremely useful to store a number of textual information. So far, I have included sentence ID, reference to the text and the tablet that the sentence belongs to, place of retrieval of the tablet, dating of both

the tablet and the text, and possible translations. Note that these data differ in scope from the philological features discussed in section 2, as they concern the text in general, and it is consequently more parsimonious to store them as comments to sentences. I am well aware that these metadata could be more fruitfully stored by editing Hittite raw texts adopting TEI guidelines,[3] but this constitutes a long-term task which goes beyond the initial stage of this project.

7 Conclusion and future work

In this paper, I have addressed some preliminary issues in designing a treebank for the Hittite language built within the framework of Universal Dependencies. Crucially, in order to grasp all relevant linguistic and textual features of Hittite, UD's template needs to be enriched with a number of language-specific dependency relations and morphological features. Also, the need to add a set of new features dedicated to the encoding of philological data has emerged.

This paper constitutes only the first step of a larger project, and much work still needs to be done. First, a schema for the syntactic annotation of broken sentences should be worked out, along the lines of what discussed by Zemánek [5]. A simple solution would be to treat gaps in tablets as instances of ellipsis, but the relation *remnant* employed to annotate ellipsis in UD is seemingly not suitable for the purpose. An alternative option could be the insertion of empty nodes in the expected position of missing words, but the insertion of empty nodes would go against the very tenets of UD, and should be avoided.

Finally, it should be stressed that the annotation outlined throughout this paper is based upon a text written in the oldest variety of the language, that is, Old Hittite. It would be thus intriguing to test whether this annotation holds for more recent phases of the language, that is, Middle and New Hittite, in which a number of significant morphological and syntactic changes have occurred.

References

[1] Giusfredi, F. 2014. Web resources for Hittitology. *Bibliotheca Orientalis* 71: 358-362.

[2] Goedegebuure, P. 2013. Split-ergativity in Hittite. *Zeitschrift für Assyriologie und vorderasiatische Archäologie*, 102 (2): 270-303.

[3] Hoffner, H. A. & Melchert, C. H. 2008. *A Grammar of the Hittite Language. Part I: Reference Grammar*. Winona Lake: Eisenbrauns.

[4] Tischler, J. 2001. *Hethitisches Handwörterbuch. Mit dem Wortschatz der Nachbarsprachen*. Innsbruck: Institut für Sprachwissenschaft.

[5] Zemánek, P. 2007. A Treebank of Ugaritic. Annotating Fragmentary Attested Languages. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*, De Smedt, K, Hajič, J & Kübler, S. (eds.), 212-218. NEATL: Bergen.

Universal Dependencies, <https://universaldependencies.github.io/docs/>

[3] URL: http://www.tei-c.org/index.xml