Syntactic Annotation for a Hittite Corpus: Problems and Principles

Maria Molina

Institute of linguistics, Russian Academy of Sciences, Moscow, Russia

maria.molina@me.com

Abstract. The aim of the paper is to present a project of a syntactically annotated corpus of Hittite, a dead cuneiform language (Anatolian family), the oldest Indo-European language attested in writing, that was spoken in 18-12 cc. BC on the territory of present-day Turkey. No publicly available corpus of Hittite with syntactic annotation exists so far, meanwhile Hittite syntax proves to be more and more interesting for the researchers, so the need of an online annotated corpus for this language is more and more compelling.

There are certain problems arising in development of such a corpus. Some of them are specific to the language itself, like 2P clitic chains, their position in the clause in terms of generative linguistics, and constituency structure of the Hittite clause. Others are connected to sociolinguistic peculiarities of Hittite system of writing: Akkadian and Sumerian logograms had been widely used by the Hittite scribes, and should be properly marked up in a Hittite corpus. Another problem is lacunae — clay tablets had been heavily broken in the last 3000–3500 years. What should be principles of phrase structure annotation when half the sentence is gone? The paper discusses these and others problems and principles on the material of the presented project.

Keywords. Hittite · Syntax · Corpus linguistics · Cuneiform · Transliteration · Anatolian family · Indo-European group of languages · Treebanks · Constituency trees · Phrase structure

1 Introduction

Syntactically annotated corpora have been developed for a wide variety of languages and grammatical frameworks and become a significant tool of linguistic research. Hittite is one of rare Indo-European languages for which no corpus with syntactic annotation has been developed so far. A project of such a corpus started at the Institute of linguistics, Russian Academy of Sciences (Moscow). The paper aims to present the project and to discuss the problems arising during the development of a syntactically annotated corpus for a dead cuneiform language.

As for syntax, Hittite is a language with second position and Wackernagel's clitics. Clitics are a problem for any syntactic annotation, as they are generated under the

rules of their own clitic syntax. Hittite, moreover, features long clitic chains that might consist of 3–5 enclitics in a row. They are commonly written together with their host, which demands algorithms for automated separation of clitics and their hosts. Hittite syntax is yet poorly understood, and syntactic annotation of Hittite material, in fact, demands some additional amount of corpus studies while the resources for corpus studies are limited. Recent research, including works based on our corpus material, improved situation (see Sideltsev 2015, Sideltsev & Molina 2015) and allowed us at last to start working on a treebank.

Hittite morphology shows quite a number of homonyms and variants, and there is no complete dictionary for the language to use in the development of a parser (CHD 1980-, the most recent and comprehensive dictionary of Hittite, is far from being finished). Hittite orthography does not include punctuation marks, which complicates the task of separating sentences in the text. To make parsing even more difficult, Hittite texts contain words in three different languages, namely, Sumeran, Akkadian and proper Hittite. The same word can be written in three ways: as a Sumeran or Akkadian logogram, as an Akkadian or Hittite word, as a mixture of them with addition of phonetic complements. The texts are written in cuneiform on clay tablets, most of which are broken during the last 35 centuries. There are multiple gaps in almost every text.

All mentioned above makes it hard to use standard text mining tools to parse Hittite texts automatically and to annotate it syntactically using previously known methods.

Web resources for Hittitology, currently available on the Internet, consist of two main initiatives. First, The Hethitologie Portal of the Akademie der Wissenschaften und der Literatur in Mainz, henceforth HPM, is the largest resource hosting a number of projects from several universities and research centers. It includes, among others, the project of text corpora — philological editions of Hittite texts with grammar annotation and organized search of word forms. Second, TITUS, gives limited access to its publication of Indo-European texts and includes several Hittite texts. None of these resources offers syntactic or information structure annotation. Enclitics there are being analysed as one phonological word with its host. Separate mark-up of enclitics that is realised in our project (the Hittite Corpus), is not realised in both publicly available corpora of Hittite texts.

2 The Hittite Corpus

The Annotated Hittite Corpus of letters, instructions and prayers is now in the process of being developed and published at *http://hittitecorpus.ru*. It is based on two publications, Middle and New Hittite letters published by Hoffner 2009, Beckman, Bryce, & Cline, 2011 and Hittite instructions by Miller 2013. Hittite prayers are to be included as well, based on a recently published online corpus by Philipps Universitaet Marburg's project, which is a part of HPM: Gebete der Hethiter, available at: http://www.hethport.uni-wuerzburg.de/txhet_gebet/textindex.php?g=gebet&x=x.

All letters and instructions have been analysed and digitalized according to the project guidelines, and a database has been built using FileMakerPro 13 shell. Then, quite recently, a MsSQL relational database system was developed for the corpus, including four indexed and cross-linked databases, online search interface with an organized search through all the fields in all the databases in all possible combinations. An option is to search matching or containing phrases.

The basic element of the corpus is a clause (a simple sentence). This means that raw material (Hittite texts) is being processed first into separate clauses, and each clause has its own annotation: positive/negative/interrogative; word order (SOV/OSV/SV/OV/V/postverbal) and phrase structure. Having certain difficulties in automated separation of Hittite words, as described above, we had little problems in parsing texts for clauses. The Hittite clause has its obvious boundary markers (phrase connectors, Wackernagel's clitics and clause final verbs, for more details see Molina & Sideltsev 2014). On the first stage, therefore, texts were transformed into a bank of clauses, making it a starting point for our corpus. Every clause then was being automatically parsed into word forms using a built-in algorithm on site. The amount of clauses reached 3,897 entries on the material of Hittite letters and instructions (roughly 19,500 word forms), aiming at 5,000 (25,000 word forms) when all the Hittite instructions known so far are processed. Hittite prayers make promise of another 3,000 clauses (roughly 15,000 word forms).

Tokenization process has just started on the corpus material. One token is one word form. Each word form, including clitics, is annotated for part-of-speech (PoS) and other grammatical information, and, on the limited number of texts, linked to a lemma. Search of lemmas is available online, with the output in the form of a list of clauses containing the word forms linked to the lemma.

A phrase structure treebank is being built basing on Penn Treebank standards; tree structure is visualized as a graph using built-in algorithms on site (see below more on phrase structures in the corpus).

3 Basic Elements and Principles of Parsing

When the aim is to investigate syntax properties, a corpus developer usually starts with providing grammatical information for each word in the sentence. Most syntactically annotated corpora in the world use lexeme/word form as a basic element (a token), it is widely accepted as a standard of corpus building. Our corpus would probably have been just the same, if not for the problem of lacunae. The ratio of non-broken to broken clauses in Hittite is so low that one should either exclude half of the material from the analysis or try to get sentence structure out of broken fragments. Frequent lacunae complicate determination of boundaries for many words, not to mention that often a text in a lacuna is not to be recovered at all. Boundaries of a clause could be recovered with much more certainty — for my corpus only about 1% of clauses are really tricky. This is why it is best first to parse text into clauses, and only then divide it into word forms (tokens), and then, if possible, annotate its grammar.

Lack of human resources would have made it hard to properly mark up a really big amount of word forms. Meanwhile, with a clause as the basic element, a text can be processed half-automatically.

A text-mining algorithm for a Hittite text is based on the principle of clause boundaries markers. As I said, there are three markers (for details see Molina & Sideltsev 2014), and three basic commands with IF…THEN operator to spot the clause boundaries:

- phrase connector *nu*

  IF *nu* (*nu=*, *n=a..*, *n=e*) THEN tag <next clause> BEFORE *nu*

- Wackernagel's clitics

  IF CliticChain (list of all possible variants of clitic chains) THEN tag <next clause> BEFORE CliticChain

- Verb final clause

  IF Verb THEN ?tag <next clause> AFTER Verb

If the first aim of a corpus is to make a description of syntax and information structure, as it is in our case, and to provide users with a tool of looking for certain syntactical peculiarities, rather than to index morphological structures, grammatical annotation of each word takes second place. Our goal is to build up an instrument for Hittitologists which would help to solve certain theoretical problems with a corpus study.

There are a number of puzzles in Hittite syntax yet waiting for their answer. For that sort of a purpose it is best to make main focus on annotating clauses, rather than word forms. For example, there are no quantitative data on negation and word order in Hittite. In our corpus each clause in the text is marked up as positive/negative/interrogative, and a certain type of word order. After that a quantitative study of negation and word order, for one, is possible. Annotation of negation markers as word forms would not help to make such a study. Annotation on the clause level helps to distinguish complicated cases, such as (1) and (2), where (1) is positive, and (2) is negative. Negation marker in both cases has value NEG in the field PoS, and if you search for negative constructions on the morphology level, the output would be irrelevant:

(1) CTH 181 NH/NS KUB 14.3 i 11

nu=za	*ŪL*	*mema-š*
CONN=REFL	NEG	say-2SG.PST

'And said "no"'

(2) CTH 181 NH/NS KUB 14.3 i 15

nu=wa	*ŪL*	*uwa-mi*
CONN=QUOT	NEG	come-1SG.PRS

‘I will not come (to Ḫatti)’

4 Treebank: phrase structures vs. dependencies

All syntactically annotated historical corpora, to the best of my knowledge, use dependency grammar for syntactic annotation. Universal Dependencies project (UD, http://universaldependencies.org), indeed, promises to become a standard for syntactic annotation in linguistic corpora. In this case sentence structure is built up automatically, using mark-up of word forms. The problem with dependencies, though, is that one does not really get information about phrase structure and word order in general. If a corpus study includes any questions on word ordering in the clause, the corpus should have a phrase structure treebank.

Certain theoretical explanations are in order here. Phrase structure grammar is a type of generative grammar based on constituent relation, as opposed to dependency grammars. Basic clause structure is understood in terms of binary division of a clause into phrases. A theory of phrase structure under Minimalist Program (one of transformational generative grammatical theories) implies that sentence is built prior to movement (bare phrase structure, BPS), which means that we should copy the parts that experienced movements. BPS permits only binary branching, which is supported in the Hittite corpus.

Phrase structures in the Hittite corpus are first built manually for each clause, in the form of a line of characters structured with curly brackets. This format has been chosen following the formats of Standford Tregex software, a shell for analyzing phrase structures, to ease analysis of the corpus material with the help of this software. A value for the field ‘Tregex annotation’ is demonstrated in Ex. (3):

(3) (ForceP (Spec (NP (N LÚ.MEŠ) (NP (N *ṬE$_4$-ME*) (Pron *-KA*)))) (Force’ (Force =*wa*) (FinP (Conj *kuwapi*) (TP (T’ (NP(N [LÚ.MEŠ]) (NP(N [*ṬE$_4$-ME*]) (Pron [*-KA*]))) (VP (Broken […])(V *uēr*)))))))

It results in the following graph online (Fig. 1):

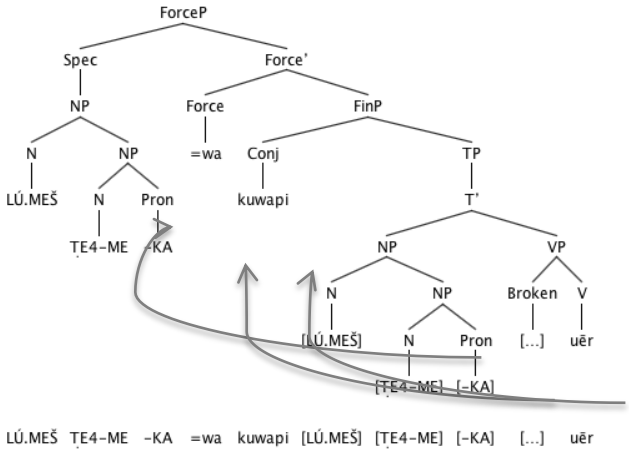


Fig. 1. Tregex visualization for Ex. (3)

The code for online transformation of the bracketed line of characters into the graph is revealed in **Appendix**.

Originally, the constituency treebank for the corpus was planned to avoid a theory-biased annotation. Still, as we tried to make sense of our annotation, and not only to build an abstract treebank, the end version makes use of phrase structure in terms of generative linguistics. As Stanford Tregex has been chosen for visualizing clause structure, tagset is based on the principles of Penn Treebank, with addition of tags for Anatolian languages specific features, like clitic chains, and tags for additional projections, like SpeechActP and split CP (ForceP/FocP/FinP), that are necessary for a proper analysis of the Hittite clause structure.

Second position clitics are analyzed as standardly being positioned in C^0, while their host is raised to Spec,CP (or $Force^0$ and ForceP, consequently, in case of split CP). Hittite features both Wackernagel's clitics hosted purely phonologically, to the first phonological word in the clause, and non-Wackernagel's clitics placed syntactically in the second non-Wackernagel's position (for detailed analysis of the Hittite clause architecture see Sideltsev, 2015). For the latter certain function words do not count as first position. If both types of clitics are presented in a clause, we analyze them as placed separately in two projections, SpeechActP for phonological Wackernagel's clitics and CP for non-Wackernagel's clitics (see Fig. 2 for the exemplar phrase structure).

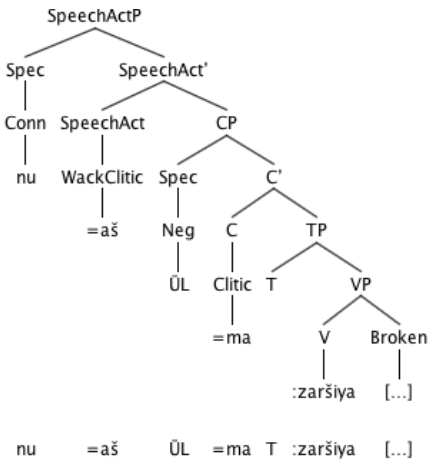


Fig. 2. Example of phrase structure for a clause with two types of clitics

5 Corpus Structure

The MsSQL relational database used for the Hittite Corpus consists of four linked tables: table of texts, table of clauses, table of word forms, table of information structure (see Fig. 3).

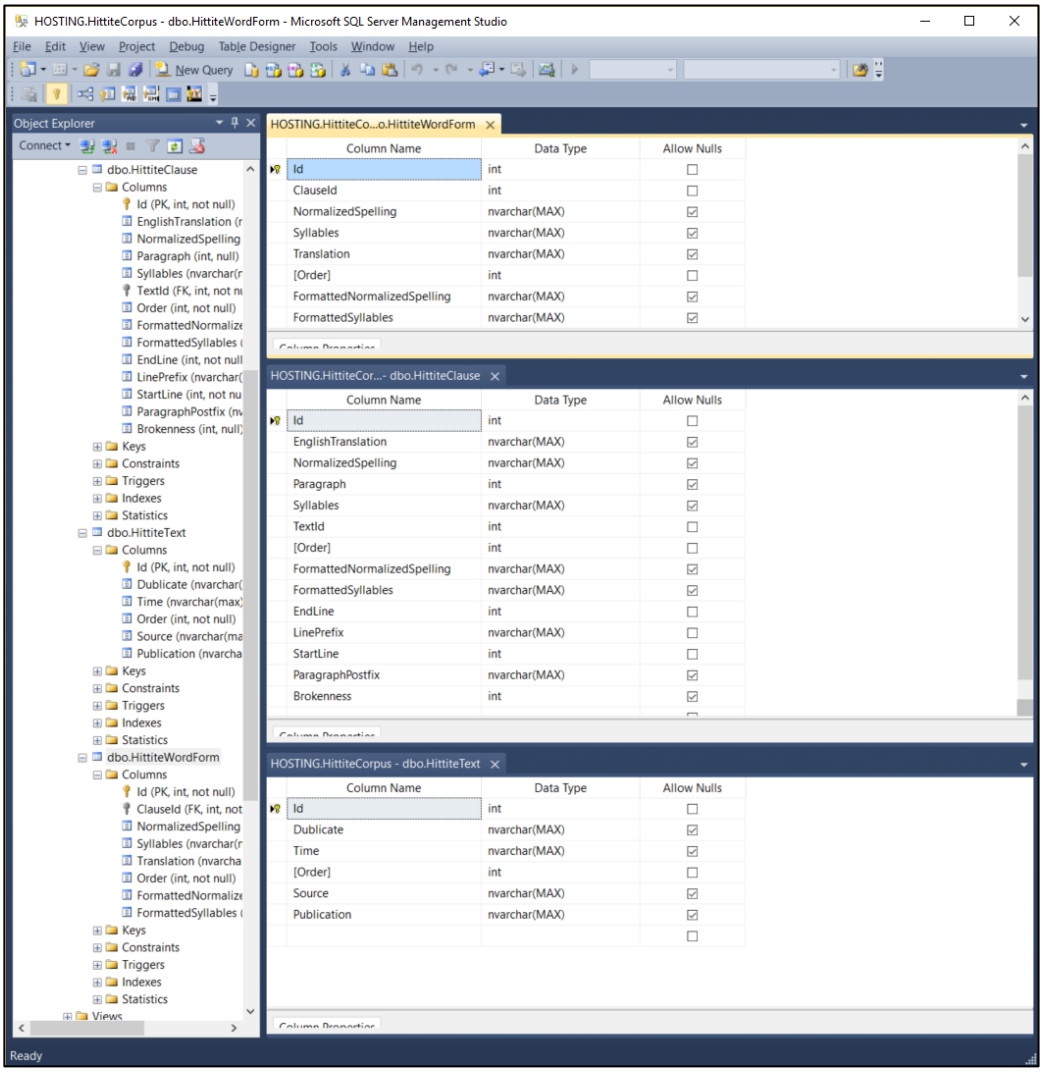


Fig. 3. MsSQL Database structure

Text metadata comprises publication, duplicate, time and text index. Publication index follows the traditional system of text indexation in Hittitology: KUB, KBo, StBoT, SMEA etc.; number of series; number of text. If a duplicate is not null value, it means that there are several copies of the text, every clause can be cited only from one of them, according to what had been chosen by a source of corpus data. Index for the duplicate is a capital letter: A, B, C, D, E, F, G etc. Time means the period when the text was composed and the period of the available copy (ductus of the text): OH/OS (Old Hittite/Old Script), OH/MS (Old Hittite/Middle Script), OH/NS (Old Hittite/New Script), MH/MS, MH/NS, NH/NS. Text index is a unique for the corpus

and obligatory value, along with the other similar indexes aimed at linking basic elements in due order, repeated for every clause.

Clause metadata comprises level of brokenness (numeric, in the range 1–5, see below details on the principles of brokenness annotation). Paragraph value is just a number of paragraph, the same as in publication or just a number of a section on the clay tablet, it is repeated for every clause in the paragraph. Lines means position of the clause on the clay tablet. Clause index is unique for the text, obligatory value, along with the other similar indexes aimed at linking basic elements in due order.

Word form metadata includes text index, clause index (to link the word form to its clause), and word form index, unique in the clause, obligatory value. It also includes lines information to point at the position of the word form on the clay tablet.

All clauses and word forms are provided with two variants: the one in syllable transliteration matching cuneiform writing, the other in narrow transliteration close to how the Hittite words were pronounced, with spelling normalization for the search purposes. All the data are also provided with translation into English (glosses for word forms, written according to the Leipzig Glossing Rules, with additions relevant for the Hittite material).

Clause syntax annotation comprises the following features (Table 1):

	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Clause annotation: syntactic features											
2	word order	question	wh.position	negation	neg.position	neg.scope	focus.type	focus.contra	focus.scope	focus.positio	particles	verb
3	V	no	fronted	no	fronted	wide	information	replacing	thetic	fronted	MA.foc	basic
4	SV	yes/non-rhet	internal	natta	internal	narrow	id/contrast	restricting	predicate	internal	MA.top	topical
5	OV	yes/rhet	preverbal	le-e	preverbal	unclear	id/verum	selecting	argument	preverbal	MA.both	fronted
6	SOV	yes/unclear	postverbal	nawi	postverbal		id/scalar	rejecting	verb	postverbal	YA.narrow	internal
7	OSV		other	nu-u-uman	other		parallel	expanding		other	YA.wide	postverb
8	V--			nekku				not contrast			YA.both	
9	Other			NU.GAL							MA-YA	
10											PAT	
11											NU	
12											TA	
13												

Table 1. Clause syntax annotation

Word form syntactic annotation comprises:

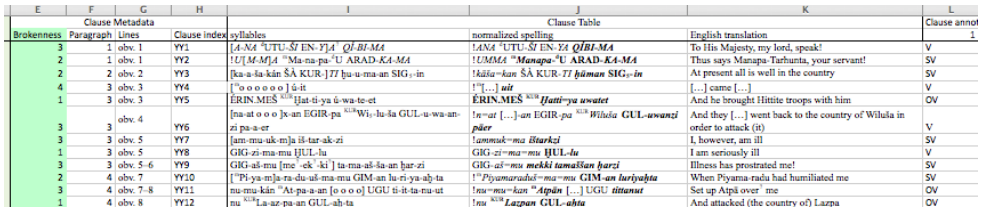
1) Lemma — a dictionary form of the Hittite word; in case the word is written in Sumerian or Akkadian, the lemma field contains a relevant word in Hittite;

2) PoS (part-of-speech) — N, V, ADJ, PART, CONJ etc. (if the word form is broken, then it is marked up as “fragment”);

3) Constituent — a phrase containing the word form: NP, VP, AdvP etc (if the word form is broken, then it is marked up as ‘fragment’, for clitic chains tag ‘Clitic’ is used);

4) Syntactic role in the sentence — S, O, IO, V, Adv, Cl (for Clitic), fragment (for broken).

Raw data are processed and annotated in Excel (Table 2) and then xls-files are uploaded into the database using uploading algorithm on site. This makes it easy to involve unqualified human resources for routine procedures and to upload annotated data into the database via the Internet.



E	F	G	H	I	J	K	L
Clause Metadata				Clause Table			Clause annot
Brokenness	Paragraph	Lines	Clause index	syllables	normalized spelling	English translation	1
3	1	obv. 1	YY1	[*A-NA* ^{d}UTU-*ŠI* EN-*Y*]*A*$^{?}$ *QÍ-BI-MA*	!*ANA* ^{d}UTU-*ŠI* EN-*YA* ***QÍBI-MA***	To His Majesty, my lord, speak!	V
2	1	obv. 1	YY2	!*U*[*M-M*]*A* ^{m}Ma-na-pa-^{d}U ARAD-*KA-MA*	!*UMMA* ^{m}***Manapa-^{d}U ARAD-KA-MA***	Thus says Manapa-Tarḫunta, your servant!	SV
2	2	obv. 2	YY3	[ka-a-ša-kán ŠÀ KUR-]*TI* ḫu-u-ma-an SIG$_5$-*in*	!*kāša=kan* ŠÀ KUR-*TI* ***ḫūman SIG$_5$-in***	At present all is well in the country	SV
4	3	obv. 3	YY4	[^{m}o o o o o o] ú-it	![…] ***uit***	[…] came […]	V
1	3	obv. 3	YY5	ÉRIN.MEŠ ^{KUR}Ḫat-ti-ya ú-wa-te-et	**ÉRIN.MEŠ** ^{KUR}***Ḫatti=ya uwatet***	And he brought Hittite troops with him	OV
3	3	obv. 4	YY6	[na-at o o o]x-an EGIR-pa ^{KUR}Wi$_{5}$-lu-ša GUL-u-wa-an-zi pa-a-er	!*n=at* […]-*an* EGIR-*pa* ^{KUR}*Wilušа* **GUL-*uwanzi pēr***	And they […] went back to the country of Wilula in order to attack (it)	V
3	3	obv. 5	YY7	[am-mu-uk-m]a íš-tar-ak-zi	!*ammuk=ma* ***ištarkzi***	I, however, am ill	SV
1	3	obv. 5	YY8	GIG-zi-ma-mu ḪUL-lu	GIG-*zi=ma=mu* **ḪUL-*lu***	I am seriously ill	V
3	3	obv. 5–6	YY9	GIG-aš-mu [me-ek-ki] ta-ma-aš-ša-an ḫar-zi	GIG-*aš=mu* ***mekki tamaššan ḫarzi***	Illness has prostrated me!	SV
2	4	obv. 7	YY10	[^{m}Pí-ya-m]a-ra-du-uš-ma-mu GIM-an lu-ri-ya-aḫ-ta	!^{m}*Piyamaraduš=ma=mu* **GIM-*an luriyaḫta***	When Piyama-radu had humiliated me	SV
3	4	obv. 7–8	YY11	nu-mu-kán ^{m}At-pa-a-an [o o o] UGU ti-it-ta-nu-ut	!*nu=mu=kan* ^{m}***Atpān*** […] UGU ***tittanut***	Set up Atpā over me	OV
1	4	obv. 8	YY12	nu ^{KUR}La-az-pa-an GUL-aḫ-ta	!*nu* ^{KUR}***Lazpan GUL-aḫta***	And attacked (the country of) Lazpa	OV

Table 2. Clause annotation of a letter

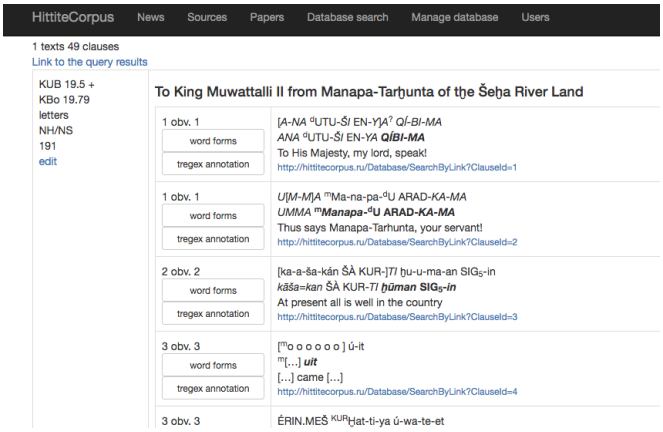


Fig. 4. The same letter on site

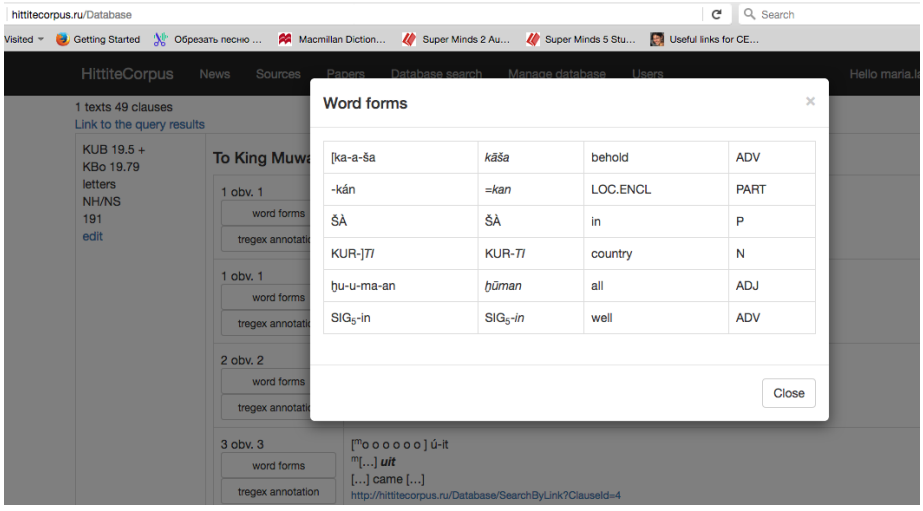


Fig. 5. Word forms on site

6 Hittite Writing System and Principles of Its Online Representation

The big problem for a Hittite corpus is transliteration. We do not represent cuneiform spelling online, as it is in fact already perfectly done by HPM. But information of the Hittite syllabic and narrow transliteration is kept in due order in the Hittite corpus.

Traditionally Hittite texts are transliterated from cuneiform as follows:

- Hittite words and complements: **italics, small letters**
- Akkadian words: **italics, all caps**
- Sumerograms: **plain text, all caps**
- Determinatives: **superscript**
- Numeric indices of Sumerian and Akkadian cuneiform signs: **subscript**
- **Diacritics** are also used for homophone signs discretion and also for some signs in syllable transliteration

This means that text mark-up should be saved in the corpus to keep linguistic information relevant for the understanding of Hittite texts. The best way would have been to use HTML tags, but this would have taken too many resources for manual mark-up of texts. Therefore, all text mark-up is done with the tools of Microsoft Word and Excel (which makes it much easier, as we partly use pre-processed docx-files with fully formatted texts). A special algorithm works during the uploading annotated data to the online database, where all text mark-up transforms into HTML tags: <b>, <i>, <sup>, <sub>. This helps to represent data on site in proper way. The tagged clauses are placed into database fields <FormattedSyllables>; <FormattedNormalizedSpelling>, and the search engine makes use of fields without tags.

7 Brokenness, or Is There Life in Lacunae

As was mentioned above, the problem of broken fragments is highly important for a syntactically annotated Hittite corpus. The clauses that are hard-broken, cannot be tagged properly. They still can represent important linguistic information, but we suggest that they are not involved into a treebank.

In order to build a treebank, we suggested that there are 5 stages of brokenness (for details, see Molina & Molin 2016):

1. *completely good*, e.g.:

   (4) CTH 181 NH/NS KUB 14.3 iv 49 (Hoffner 2009:312)
   SAG.DU-*an* *ku-ra-an-du*
   head.ACC.SG cut.3PL.IMP

"Let them cut off his head!"

2. *broken, but fully restorable,* e.g.:

(5) CTH 181 NH/NS KUB 14.3 iv 50–51 (Hoffner 2009:312)
[SAG.DU-*an-m*]*a* *ku-in* *ku-ra-an-zi*
head.ACC.SG=PTCL which.ACC.SG cut.3PL.PRS
"And the head that they cut off…"

The context is closely parallel to the one in (Ex. 2), the same noun has been used several lines above, the meaning of the sentence is obvious.

3. *clause boundaries are obvious, S, O, V and other vital constituents could be restored*, their order obvious, but some of not-so-vital constituents are missing, e.g.:

(6) CTH 182 NH/NS KUB 19.55 obv. 3–4 (Hoffner 2009:317)
[*A-BU-KA-ma* ...] [Z]AG.MEŠ-*YA* *i-la-liš-ke*-[*et* ...]
your.father.NOM.SG=PTCL my.border.territories desire.3SG.PST.IMF
"But your father [...] was coveting my border territories (had always desired my border territories)…"

There must be something in between "your father" and "my border territories", but the subject, the object and the verb are preserved or restored with certainty. The context makes it improbable that there is anything in the postverbal position.

4. [word order is not obvious], [vital constituents are missing], [clause boundaries are not quite obvious], but *the meaning of the sentence is obvious from the context*, e.g.:

(7) CTH 182 NH/NS KUB 19.55 obv. 22 (Hoffner 2009:317)
[x-x]x-mu-za le$^{?}$-e$^{?}$ i[-la$^{?}$-li$^{?}$-ya$^{?}$-ši$^{?}$…]
X=PRON.GEN.SG=REFL NEG desire.2SG.PRS
You shall not desire? my (land...)

5. hard-broken case, the sentence is good only for attesting word form usage, spelling and statistical purposes like "what's the ratio of *nu*-clauses", e.g.:

(8) CTH 182 NH/NS KUB 19.55 obv. 24 (Hoffner 2009:317)
[…] *A-BU-KA* ku-w[a-pí …]
X my.father when X
"[…] your father when […]"

Every sentence in a corpus can be annotated for the level of brokenness. After this job is done, the clauses of the first 3 levels can be syntactically annotated, as described above. Fully broken fragments are marked as […] and are considered whole constituents ("null constituents"). Null constituents at level 3 might be indirect objects and adverbs and be dependents of the verb, and analyzed as such in the phrase structure. They might rarely be link-verbs, subjects, or objects, in case it is obvious from the context, and then the null constituent is analyzed as such. Restored fragments are considered unbroken, the same as originally unbroken material. In case a researcher needs information of what was broken in the context, a syllable transliteration is always available in the corpus, conserving all features of the text according to the sources of publication.

E.g. for Ex. (5) the following sentence is recorded in narrow transliteration:

SAG.DU-*an=ma kuin kuranzi;*

for Ex. (6):

ABU-KA=ma [...] ZAG.MEŠ-*YA ilališket* [...]

For the last two levels of brokenness only a general count of clauses is allowed, as well as statistical and some grammatical information for separate word forms.

This algorithm is already implemented in the corpus and in the search machine: you can choose your level of brokenness in a search request and, therefore, work only with unbroken fragments or all clauses if aims of your research allow it.

Broken fragments are the first problem that should be solved when we talk about Hittite phrase structures (see discussion above). We can never know what was originally in lacunae. Therefore, we can only draw a constituency structure for clauses with 1–3 levels of brokenness. The 3rd level implies that we are able to reconstruct the nature of broken fragments. If not, then it is not the 3rd level, but the 4 or the 5th one.

8 Conclusions

Thus, there are three general issues that should be considered when developing a syntactically annotated corpus of Hittite. The first is principles of parsing a text, and we propose that for Hittite (and, possibly, other Anatolian cuneiform languages) one should start with a clause, which then can be parsed into words if possible, but annotated syntactically on the clause level as well. A half-automated parsing can be developed for speeding up the text mining, and this work is already in process for our corpus, based on the principles of clause boundaries markers and strict word order of the Hittite clause. The next problem concerns broken fragments and their syntactic annotation. We rate material for 5 levels of brokenness, and we give reliable syntactic annotation (phrase structures) only for sentences with ratings 1–3. For the level 3 additional consideration is needed concerning broken parts, which could be considered as dependents of Verb (IO, Adv) or, if it is clear from the context, as Object, Subject or Verb. If their syntactic role is obvious, we can build a phrase structure for the sentence.

The third main problem is building phrase structures for Hittite clauses, as there are certain theoretical issues connected to clause architecture questions; and we propose to use splitCP projections and additional layer (SpeechActP) for syntactic annotation when it is conditioned by the information structure of the clause.

References

1. Beckman G, Bryce T. & Cline E (2011) The Ahhiyawa Texts. Atlanta: Society of Biblical literature.
2. CHD: Chicago Hittite Dictionary (1980-) Chicago: Oriental Institute.
3. Giusfredi F (2014) Web resources for Hittitology. In: Bibliotheca Orientalis, Vol. 71, pp. 358–362.
4. Giusfredi F (2015) Phrase Structure and Ancient Anatolian Languages. Methodology and challenges for a Luwian syntactic annotation. In: Proceedings of CLiC-it, available at: http://clic.humnet.unipi.it/proceedings/Proceedings-CLICit-2014.pdf.
5. Gebete der Hethiter (2017). http://www.hethport.uni-wuerzburg.de/txhet_gebet/textindex.php?g=gebet&x=x
6. Hoffner H (2009) Letters from the Hittite Kingdom. Atlanta: Society of Biblical Literature.
7. Leipzig Glossing Rules. https://www.eva.mpg.de/lingua/resources/glossing-rules.php
8. Luraghi S (1990) Old Hittite Sentence Structure. Routledge, London/New York.
9. Marcus M, Santorini B, Marcinkiewicz MA (1993) Building a large annotated corpus of English: the Penn Treebank. In: Computational Linguistics, Vol. 19.
10. Marcus M, Kim G, Marcinkiewicz MA, MacIntyre R, Bies A, Ferguson M, Katz K, Schasberger B (1994) The Penn Treebank: annotating predicate argument structure. https://www.cis.upenn.edu/~treebank/.
11. Miller J (2013) Royal Hittite Instructions and Related Administrative Texts. Atlanta: Society of Biblical Literature.
12. Molina M, Molin A (2016) In a Lacuna: Building a Syntactically Annotated Corpus for a Dead Cuneiform Language (on the Basis of Hittite). In: Proceedings of DIALOG 2016. Online articles. http://www.dialog-21.ru/media/3476/molinammolina.pdf
13. Molina M, Sideltsev A (2014) Corpus research of information structure and the clause boundaries in Hittite (on the basis of the Middle Hittite letters). In: Indo-European linguistics and classical philology — XVIII. Proceedings of the 18th Conference in Memory of Professor Joseph M. Tronsky, St. Petersburg, pp. 657–666.
14. Sideltsev A (2015) Hittite Clause Architecture. In: Revue d'Assyriologie et d'archéologie orientale, vol. 109, pp. 79–112.
15. Sideltsev A, Molina M (2015) Enclitic -(m)a, clause architecture and the prosody of focus in Hittite. In: Indogermanische Forschungen, Bd. 120, pp. 209–254.
16. Standford Tregex. Tregex, Tsurgeon and Semgrex. The Stanford Natural Language Processing Group. https://nlp.stanford.edu/software/tregex.shtml
17. HPM (2001-) The Hethitologie Portal of the Akademie der Wissenschaften und der Literatur in Mainz. http://www.hethport.uni-wuerzburg.de
18. TITUS. http://titus.uni-frankfurt.de/indexe.htm