Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2016”

Moscow, June 1–4, 2016

IN A LACUNA: BUILDING A SYNTACTICALLY ANNOTATED CORPUS FOR A DEAD CUNEIFORM LANGUAGE (ON THE BASIS OF HITTITE)

Maria Molina (maria.molina@me.com)
Institute of linguistics (Russian Academy of Sciences), Moscow, Russia

Alexei Molin (plint@ngs.ru)
PlintTech, Novosibirsk, Russia

The paper presents a new corpus of an ancient language—Hittite, a dead cuneiform language of Anatolian family attested on clay tablets of 18–12 cc. BC. Hittite is the oldest attested Indo-European language, and it remains practically the only major Indo-European language with a significant body of texts that does not have an online corpus with a syntactic annotation. Hittite syntax proves to be more and more interesting for the researchers, so the need of an online syntactically annotated Hittite corpus is more and more compelling.

There are several major problems of building an annotated corpus for a dead cuneiform language. The vast majority of cuneiform tablets has been broken up into pieces during the last 3,000–3,500 years. All texts have lacunae in almost every sentence, therefore it is rather difficult to apply standard methods of parsing and syntactic annotation. The main problem concerns marking up syntax in broken parts. Most treebanks are based on morphosyntactic annotation performed at the level of a word form. But having half of a sentence lost, we can hardly build a tree out of broken fragments. Instead, we suggest annotation of a whole clause. The paper discusses the principles of syntactic annotation of broken parts and basic structure of a new syntactically annotated Hittite corpus (available at http://hittitecorpus.ru), the on-going project of the Department of Anatolian and Celtic Studies at the Institute of linguistics, Russian Academy of Sciences.

Key words: Hittite, ancient languages, syntax, corpora, annotation, phrase structure, clause, parsing, broken fragments

В ЛАКУНЕ: ПРОБЛЕМЫ ПОСТРОЕНИЯ СИНТАКСИЧЕСКИ РАЗМЕЧЕННОГО КОРПУСА ДЛЯ МЕРТВОГО КЛИНОПИСНОГО ЯЗЫКА (НА ПРИМЕРЕ ХЕТТСКОГО)

Мария Молина (maria.molina@me.com)
Институт языкознания (РАН), Москва, Россия

Алексей Молин (plint@ngs.ru)
PlintTech, Новосибирск, Россия

Ключевые слова: хеттский язык, синтаксис, корпусная лингвистика, разметка, мертвые языки, дерево составляющих, клауза, лакуны

1. Introduction to the Hittite Corpus

Syntactically annotated corpora have been developed for a wide variety of languages and grammatical frameworks and are gradually becoming a significant tool for linguistic research. Hittite is one of rare Indo-European languages for which no such corpus has been developed so far. A project of a syntactically annotated corpus of Hittite started at the Institute of linguistics, Russian Academy of Sciences, in 2015. The paper aims to present the project and to discuss the problems arising in a development of a corpus for a dead cuneiform language.

The corpus includes Hittite letters published by [Hoffner 2009; Beckman et al. 2011] and instructions published by [Miller 2013] in syllabic transliteration (one cuneiform sign correlates with one syllable) and provided in the corpus with so called narrow transliteration—when all syllables are split into what might be considered a word form. The current amount of the corpus is 3,861 clauses, 2,212 clauses with the Middle Hittite material, and 1,649 clauses in the New Hittite part (an average length of a clause being 5 words, it amounts to approx. 19,300 words). By the end of June 2016 the corpus will be enlarged to the amount of 5,000 clauses (approx. 25,000 words) on the material of letters and instructions, and we plan to further develop it on the material of the Hittite prayers, with the final volume of around 10,000 clauses (approx. 50,000 words).

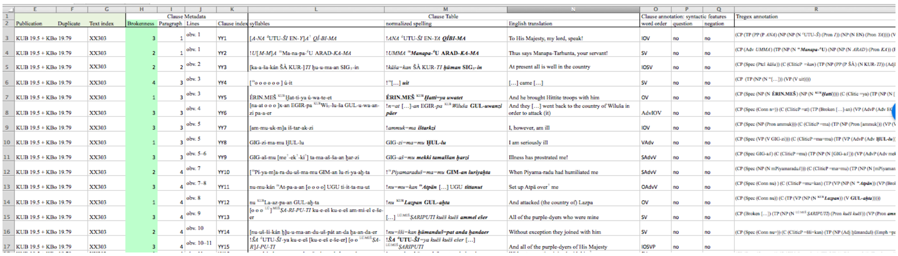
The Hittite corpus is built as a MsSQL relational database with online uploader and search engine. The basic element of the corpus is a clause: every entry in the main table of the database contains one clause, all relevant annotation so far concerns clause features, contra usual practice to base a linguistic corpus on a word form tagging. A clause means a simple sentence containing one finite verb, including relative clauses. All texts in the corpus were manually parsed for clauses. Each clause was then automatically parsed into word forms using an algorithm embedded to the online uploader. A limited access to the corpus material was opened online at *http://hittitecorpus.ru* in December 2015—January 2016.

Syntactic annotation means a constituency treebank. Phrase structure is drawn manually for every database entry (i.e. for every clause) and is then visualized online as a graph by a built-in tree generator. Certain algorithms have been developed for keeping linguistic information safe while automatically converting texts.

Firstly, a list of traditional rules of transliteration is used for Hittite and other Anatolian languages. Those rules are supported in the corpus by a simple HTML tagset applied to the text automatically during the uploading to the database.

Secondly, the algorithm for parsing clauses into words was developed. It particularly deals with the problem of many lacunae in the Hittite texts. The parser built into the uploader might be easily used in other corpora of ancient texts with broken fragments—all data are provided in an unified XSLX form. The uploader, the database structure and all online applications might well be used for any other ancient language.

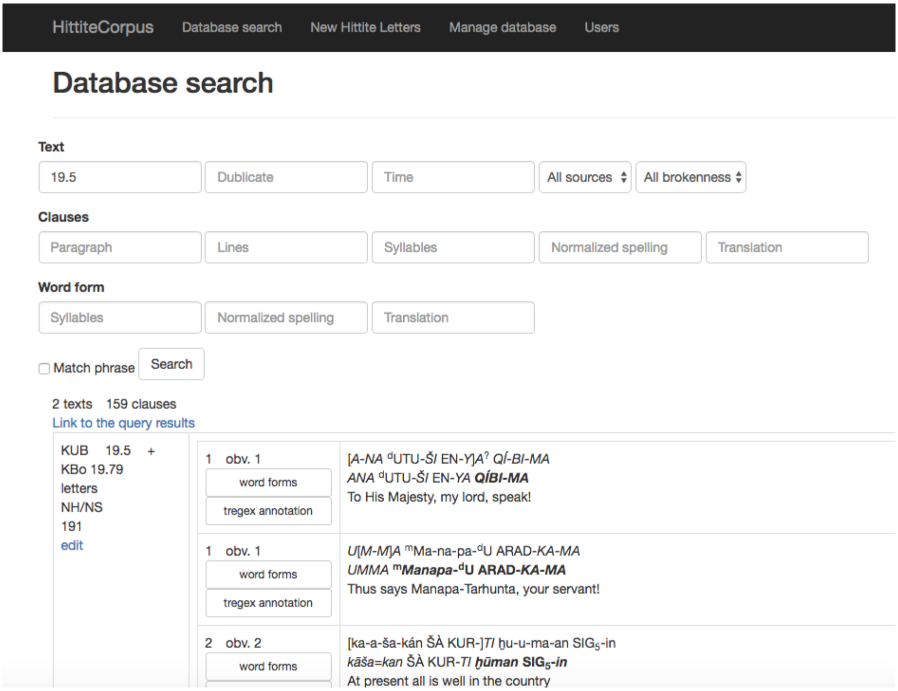
The raw material ready for uploading is structured in XSLX format as shown on Fig. 1.



	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1				Clause Metadata					Clause Table		Clause annotation: syntactic features			Tregex annotation
2	Publication	Duplicate	Text index	Brokenness	Paragraph	Lines	Clause index	syllables	normalized spelling	English translation	word order	question	negation	
3	KUB 19.5 + KBo 19.79		XX303	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!	IOV	no	no	(CP (TP (PP (P *ANA*) (NP (NP (N ^{d}UTU-*Š*) (Pron *I*)) (NP (N EN) (Pron *YA*)))) (V
4	KUB 19.5 + KBo 19.79		XX303	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-Tarhunta, your servant!	SV	no	no	(CP (Adv *UMMA*) (TP (NP (N ^{m}***Manapa***-^{d}**U**) (NP (NP (N *ARAD*) (Pron *KA*)) (I
5	KUB 19.5 + KBo 19.79		XX303	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	IOSV	no	no	(CP (Spec (Ptcl *kāša*)) (C (CliticP =*kan*) (TP (NP (PP (P ŠÀ) (N KUR-*TI*)) (Adj
6	KUB 19.5 + KBo 19.79		XX303	4	3	obv. 3	YY4	[^{m}o o o o o o] ú-it	!^{m}[…] ***uit***	[…] came […]	SV	no	no	(CP (TP (NP (N ^{m}[…])) (VP (V uit))))
7	KUB 19.5 + KBo 19.79		XX303	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	no	no	(CP (Spec (NP (N **ÉRIN.MEŠ**) (NP (N ^{KUR}***Ḫatti***)))) (C (Clitic =ya) (TP (NP (N [
8	KUB 19.5 + KBo 19.79		XX303	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}*Wiluša* **GUL-*uwanzi*** ***pāer***	And they […] went back to the country of Wiluša in order to attack (it)	AdvIOV	no	no	(CP (Spec (Conn n=)) (C (CliticP =at) (TP (Broken […]-an) (VP (AdvP (Adv EG
9	KUB 19.5 + KBo 19.79		XX303	3	3	obv. 5	YY7	[am-mu-uk-m]a iš-tar-ak-zi	!*ammuk=ma* ***ištarkzi***	I, however, am ill	IOV	no	no	(CP (Spec (NP (Pron ammuk))) (C (CliticP =ma) (TP (NP (Pron [ammuk])) (VP (V
10	KUB 19.5 + KBo 19.79		XX303	1	3	obv. 5	YY8	GIG-zi-ma-mu ḪUL-lu	GIG-*zi=ma=mu* **ḪUL-*lu***	I am seriously ill	VAdv	no	no	(CP (Spec (VP (V GIG-zi))) (C (CliticP =ma=mu) (TP (VP (AdvP (Adv **ḪUL-*lu***))
11	KUB 19.5 + KBo 19.79		XX303	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!	SAdvV	no	no	(CP (Spec GIG-*aš*) (C (CliticP =mu) (TP (NP (N [GIG-*aš*])) (VP (AdvP (Adv mek
12	KUB 19.5 + KBo 19.79		XX303	2	4	obv. 7	YY10	[^{m}Pi-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	SAdvV	no	no	(CP (Spec (NP (N mPiyamaraduš))) (C (CliticP =ma=mu) (TP (NP (N [mPiyamar
13	KUB 19.5 + KBo 19.79		XX303	3	4	obv. 7–8	YY11	nu-mu-kán ^{m}At-pa-a-an [o o o o] UGU ti-it-ta-nu-ut	!*nu=mu=kan* ^{m}***Atpān*** […] UGU ***tittanut***	Set up Atpā over$^{?}$ me	OAdvV	no	no	(CP (Spec (Conn nu)) (C (CliticP =mu=kan) (TP (VP (NP (N ^{m}***Atpān***)) (VP (Bro
14	KUB 19.5 + KBo 19.79		XX303	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	no	no	(CP (Spec (Conn nu)) (C (TP (VP (NP (N ^{KUR}***Lazpan***)) (V **GUL-*aḫta***)))))
15	KUB 19.5 + KBo 19.79		XX303	3	4	obv. 9	YY13	[o o o ^{LÚ}]^{MEŠ}*ṢA-RI-PU-TI* ku-e-eš ku-e-eš am-mi-el e-še-er	[…] $^{LÚ.MEŠ}$*ṢARIPUTI kuēš kuēš* ***ammel ešer***	All of the purple-dyers who were mine	SV	no	no	(CP (Broken […]) (TP (NP (N $^{LÚ.MEŠ}$ *ṢARIPUTI*) (Pron *kuēš kuēš*)) (VP (Pron ***amm***
16	KUB 19.5 + KBo 19.79		XX303	2	4	obv. 10	YY14	[nu-uš-ši-kán ḫ]u-u-ma-an-du-uš-pát an-da ḫa-an-da-er	!*nu=šši=kan* ***ḫūmanduš=pat anda ḫandaer***	Without exception they joined with him	SV	no	no	(CP (Spec (Conn nu=)) (C (CliticP =šši=kan) (TP (NP (Adj ḫūmanduš) (Emph =pa
17	KUB 19.5 + KBo 19.79		XX303	3	4	obv. 10–11	YY15	!*ŠA* ^{d}UTU-*ŠI*-ya ku-e-eš [ku-e-eš e-še-er] [o o $^{LÚ.MEŠ}$*ṢA-R*]*I-PU-TI*	!***ŠA*** ^{d}**UTU-*ŠI*=*ya*** *kuēš kuēš ešer* […] $^{LÚ.MEŠ}$*ṢARIPUTI*	And all of the purple-dyers of His Majesty	IOSVP	no	no	

Fig. 1. The letter to King Muwattalli II from Manapa-Tarḫunta of tḫe Šeḫa River Land (NH, KUB 19.5 + KBo 19.79)

A screenshot of the search interface and a query results is shown on Fig 2.



Text: 19.5

All sources

All brokenness

□ Match phrase

2 texts 159 clauses

KUB 19.5 +
KBo 19.79
letters
NH/NS
191

1 obv. 1
[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!

1 obv. 1
U[M-M]A ^{m}Ma-na-pa-^{d}U ARAD-*KA-MA*
UMMA ^{m}***Manapa-***^{d}**U ARAD-*KA-MA***
Thus says Manapa-Tarhunta, your servant!

2 obv. 2
[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

Fig. 2. The results for texts with 19.5 in publication number

2. Principles of dealing with the fragmentary nature of the material

The biggest problem of syntactic annotation of a dead cuneiform language is lacunae. Before we start applying any digital methods to the Hittite texts, or, for that matter, to any other texts written thousands of years ago on clay tablets (on any material that can be broken or torn apart), principles of dealing with broken fragments should be developed. Clay cuneiform tablets, the main source of the Hittite texts, have been broken into many pieces during the last 3000–3500 years, and a joining procedure is applied to them to read the text. Still, almost all texts in Hittite (or any cuneiform language) have lacunae. There is no procedure to know for certain what was in the broken fragments. But if we limit the corpus to only completely unbroken sentences, the amount of linguistic material would be too small. We are convinced that all existing material should be available in a relevant corpus of any low resource language, which means that every single word or clause that can be reconstructed from copies of the text or from context should be thoroughly taken into account.

The main principles of dealing with the fragmentary nature of our material are based on the philological approach traditionally applied to ancient cuneiform languages, which implies reconstruction of text in the broken parts. Standardly, reconstruction is based on the other known copies of the text; on the understanding of patterns (such as greetings or rituals); on the analysis of the cuneiform remains on the tablet; etc.

The reconstructed text is usually transliterated with square brackets, with unrestored parts presented as dots; if the reconstruction is supported with readings from other copies of the text, round brackets are added. Our first task is to keep this textual information safe in the corpus, and keep it when automatically parsing the clause into words. The parsing algorithm mentioned above recognises square brackets with dots as a broken fragment and reads it as a separate word in the following form: […].

Ex. (1) is just one example of a clause with hard broken fragments. It cannot be translated otherwise as “*Broken*”, but can be used for statistical purposes:

(1) syllabic transliteration: [x-x-x-x-]x-šar-ra-aš ^{m}*ŠU*$^{?}$-*MI*-^{d}A.A[…]
narrow transliteration:
[…]-*šarraš* ^{m}*ŠU*-*MI*-^{d}A.A […]
X-sarra.NOM.SG PN [verb]
(KBo 18.2 rev. 3’, Hoffner 2009)

The procedure of automated parsing the clause is presented online in the following table:

[x-x-x-x-]x-šar-ra-aš	[…]-*šarraš*	X-sarra.NOM.SG
^{m}*ŠU*$^{?}$-*MI*-^{d}A.A	^{m}*ŠU*-*MI*-^{d}A.A	PN
[…]	[…]	[verb]

3. Syntactic annotation of a broken text

It seems that syntactic annotation (a phrase structure) cannot be drawn for clauses broken as hard as Ex. (1) (see below principles of markup for lacunae). If a clause is not that broken, a phrase structure can be analyzed even if there are broken fragments in it. In fact, drawing constituency trees for the clauses with broken fragments is the main problem of syntactic annotation for any texts with lacunae in any language with closed corpus of texts and only written material. The following principles, therefore, can be easily accepted for the texts in all languages that face the problem of lacunae, but should be especially relevant for ancient languages.

We suggested that all clauses should be rated for the level of brokenness, depending on whether all words could be reconstructed in the context or not. Hard-broken clauses cannot be tagged properly, but they still can represent important linguistic information. The hard-broken material is not involved into the treebank, but can be searched in the database anyway. Users can choose a relevant level of brokenness for their own research.

There are 5 stages of brokenness in the corpus:

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

(2) 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.:

(3) 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.:

(4) 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.IMPF

“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 unbelievable 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.:

(5) 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.:

(6) 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 normal constituents, the same as 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. 2) the following sentence is recorded in narrow transliteration:

SAG.DU-*an=ma kuin kuranzi;*
for (Ex. 3):
ABU-KA=ma [...] ZAG.MEŠ-*YA ilališket* [...]

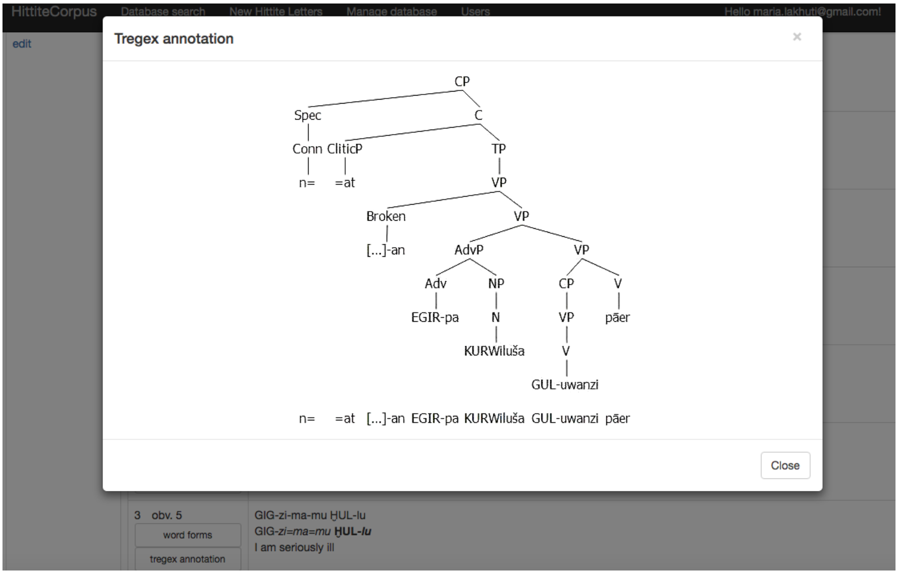
For the last two levels of brokenness only a general count of clauses is allowed, as well as statistical and some grammatical information for separate word forms. Ex. (2) is a better conserved sentence with just one broken fragment, and we are going to demonstrate the procedure of parsing and tree generating on site.

(2) [na-at o o o]x-an EGIR-pa $^{\text{KUR}}$Wi5-lu-ša GUL-u-wa-an-zi pa-a-er
n=at […]-*an* EGIR-*pa* $^{\text{KUR}}$*Wiluša*
CONN=PRON.3PL [ADV] back land.of.Wilusa
GUL-*u-wanzi* *pā-er*
attack-INF go-3PL.PST
And they […] went back to the country of Wiluša in order to attack (it)
(KUB 19.5 + KBo 19.79 obv. 4, Hoffner 2009)

The parsing table for the Ex. (2) is the following:

[na-	*n=*	CONN
-at	*=at*	they.PRON.NOM.3PL.ENCL
o o o]x-an	[…]-*an*	[…]
EGIR-pa	EGIR-*pa*	back
$^{\text{KUR}}$Wi5-lu-ša	$^{\text{KUR}}$*Wiluša*	country of Wiluša
GUL-u-wa-an-zi	GUL-*uwanzi*	attack.INF
pa-a-er	*pāer*	go.3PL.PST

The phrase structure for the clause is shown at Fig. 2. The broken fragment is analyzed as an adjunct (an adverb) on the basis of a context analysis. Penn Treebank tagset adapted for Stanford Tregex tree generator (see details in [Marcus et al. 1994]), with some language specific additions, was used for drawing phrase structures of the Hittite clauses.



n= =at [...]-an EGIR-pa KURWiluša GUL-uwanzi pāer

3 obv. 5

GIG-zi-ma-mu ḪUL-lu
GIG-*zi=ma=mu* **ḪUL-*lu***
I am seriously ill

Fig 3. Phrase structure for Ex. (2)

4. Basic elements and principles of parsing

The Hittite Corpus does not follow one universal principle of corpora for natural languages: as was mentioned above, it is not built from a word form (when morphosyntax of a word form helps to build syntactic annotation of the clauses), but from a clause. The reason for doing this is the following. Lacunae make it difficult to apply standard methods of parsing—it is difficult to automatically analyze phrase structure if words are missed, to gather a sentence from broken parts of words. The boundaries of a clause could be recovered with a big certainty—for my corpus only about 1% are really tricky.

If only the Hittites had any punctuation marks in their writing, the task would have been even easier. But while we do not have any orthographic markers of the beginning or ending of a clause in cuneiform texts, we could rely on indirect markers, such as conjunctions, verbs (that in Hittite as a SOV language go clause final) and Wackernagel clitics (that in Hittite always clitisize to the first phonological word in a clause). A development of text mining algorithm is planned for the Hittite corpus project that would help to mark up basic elements—clauses—half-automatically (about 50% of the clause boundaries can be marked up with the algorithm). There are some clause boundaries markers in Hittite, that are partly discussed in [Molina, Sideltsev 2014] and can be formally described as:

- phrase connector *nu*
  IF *nu* (*nu=*, *n=a..*, *n=e*) THEN tag <next clause> BEFORE *nu*
- Wackernagel's 2P clitics
  IF CliticP (list of all possible variants of clitic chains)
  THEN tag <next clause> BEFORE CliticP
- Verb final clause
  IF Verb THEN ?tag <next clause> AFTER Verb

We hope to be able to parse at least half of all volume of clauses with help of these markers. Previous research has shown that phrase connector nu is attested in about 49% of clauses, and with 2P clitics and verb recognition we can achieve even more.

If a corpus is built with a clause as the basic element, and the boundaries of a clause are therefore defined throughout the corpus, an automatic glossing based on text mining probabilistic technologies can be introduced at least for verbs, which are clause final in the majority of clauses in Hittite, and phrase connectors which are clause initial; some other rules for the parser are also planned for trying on our corpus material.

5. Conclusions

Summing up, lacunae are the main issue that should be solved before annotating syntax for any ancient language with closed corpus of texts and broken materials. The authors offered a systematic approach to syntactic analysis of texts with broken fragments: rating material on 5 levels of brokenness for the Hittite material and tagging only those with ratings 1–3. For the level 3 additional annotation is needed for the broken parts, which are considered as dependents of Verb (IO, Adv.) or, if it is clear from the context, as Object, Subject or Verb. A tree generator was built in the corpus and is now available online for the clauses annotated against phrase structure according to Stanford Tregex tagset, with additional language specific tags. Systematic annotation of the level of brokenness allowed users to choose the most relevant search options for their research. A clause was made the basic element of the corpus, which helped processing broken texts, and a new parsing algorithm was developed for separating words in a context while keeping linguistic information including that of brokenness. The principles of an automated text mining algorithm (a clause parser), that can be used for speeding up the work on the corpus volume, were also discussed in the paper. The Hittite corpus, published online at *http://hittitecorpus.ru*, is a long-awaited tool for Hittitologists and Anatolists studying syntax, as well as typologists working in the formal framework. The MsSQL database structure, online search engine, online word form parser and tree generator, developed for the corpus, could be further used for syntactic annotation of other cuneiform and hieroglyphic Anatolian languages as well as cuneiform languages of other language families.

References

1. *Beckman G., Bryce T. and Cline E.* (2011), The Ahhiyawa Texts, Society of Biblical literature, Atlanta.
2. *Giusfredi F.* (2014), Web resources for Hittitology, Bibliotheca Orientalis, Vol. 71, pp. 358–362.
3. *Giusfredi F.* (2015), Phrase Structure and Ancient Anatolian Languages. Methodology and challenges for a Luwian syntactic annotation, Proceedings of CLiC-it, available at: http://clic.humnet.unipi.it/proceedings/Proceedings-CLICit-2014.pdf.
4. *Hethiter.net:* Hethitologie Portal Mainz (HPM), available at: http://www.hethport.uni-wuerzburg.de/HPM/index.html.
5. *Hoffner H. A. Jr.* (2009), Letters from the Hittite Kingdom, Society of Biblical Literature, Atlanta.
6. *Luraghi S.* (1990), Old Hittite Sentence Structure, Routledge, London/New York.
7. *Marcus M., Santorini B., Marcinkiewicz M. A.* (1993), Building a large annotated corpus of English: the Penn Treebank, Computational Linguistics, Vol. 19, 1993.
8. *Marcus M., Kim G., Marcinkiewicz M. A., MacIntyre R., Bies A., Ferguson M., Katz K., Schasberger B.* (1994), The Penn Treebank: annotating predicate argument structure, available at: https://www.cis.upenn.edu/~treebank/.
9. *Miller J.* (2013), Royal Hittite Instructions and Related Administrative Texts, Society of Biblical Literature, Atlanta.
10. *Molina M., Sideltsev A.* (2014), Corpus study of information structure and clause boundaries in Hittite (on the basis of the Middle Hittite letters). [Korpusnoye issledovaniye informacionnoi struktury i granicy klauzy v hettskom yazyke, na materiale srednehettskih pisem.], Indo-European linguistics and classical philology [Indoevropeiskoye yazykoznaniye i klassicheskaya filologiya], St. Peterburg, T. 18, pp. 657–666.
11. *Sideltsev A.* (2015). Hittite Clause Architecture, Revue d’assyriologie et d’archéologie orientale 2015/1 (Vol. 109), p. 79–112.