Computational Perspectives on Ancient Near Eastern Literature, Art and Material Culture

February 19th-21st, 2023

Universal dependencies for Hittite

Maria Molina, Tel Aviv University

mariya.molina@gmail.com

In the recent years universal dependencies (UD) became the standard for digital treebanks, they are an essential part of annotated linguistic corpora, and they are widely used for comparison of language features in linguistic research. Historical corpora develop treebanks slower than corpora of modern languages. However, there have already been published some very good examples of historical treebanks [1], including those with UD annotation, e.g., for Ancient Greek and Latin [2], Old East Slavic, Old and Middle Russian [3], Old French [4]. An attempt to build a UD treebank for Hittite was undertaken already in 2017 by G.Inglese and M.Molina [5] in the framework of the PROIEL [6]. Finally, a small UD-annotated treebank based on examples from [7] was developed and published in 2022 by E.Andersen and B.Rozonoyer [8, 9].

The authors of the [8, 9] treebank based their annotation on Inglese [11], taking into account experience and data of our PROIEL experiment [5]. The concern here is that they did not account for some certain Hittite features, such as second position clitic particles *-(m)a*, *-(y)a*, and *-pat* (using instead the discourse feature after [11]), while in recent years following the publication of [11] there have been extensive research on 2P particles in Hittite, particularly including *-pat* [12, 13, 14]. The treebank [7] does not include lemmas with both a Hittite word and a Sumerian/Akkadian heterogram in cases of variations in writing of the word (that was suggested in [8], but never realized in practice before this work). However, the biggest problem of the existing Hittite UD treebank is that it is completely taken out of context being just a set of sentences out of a tutorial [7].

The Annotated Corpus of Hittite Clauses [ACHC, 10] was first launched in 2015 on the basis of the Hittite letters and instructions (not digitalized before). It was syntactically annotated for the word order (SOV/OSV). It was also annotated as a constituency treebank, with morphological mark-up and the UD prep annotation accomplished on ca. 1500 clauses. Now the time came to develop the UD annotation in the whole corpus of letters and instructions, finally putting the Hittite grammar in context.

This paper describes the UD annotation for ACHC. It is an ongoing project, starting from previously achieved number of UD-annotated clauses. In comparison to [7], we add a layer of mark-up including separate fields for both Hittite and heterogram lemmas, and indexation for clitic chains. All the tokens are provided by glosses and translation into English. The clauses represent Hittite of letters and instructions – the closest possible genre to oral speech (for a dead cuneiform language).

UD distinguishes 17 universal part-of-speech definitions (UPOS) [18:261] – the categories widely attested in the world's languages, and Hittite is not an exclusion, – such as *noun, verb, adjective,* or *adverb*. There are also standard morphological features, like *pronoun*, *numeral*, *possessive,* or *gender* types (cf. in [18:263]). Additional features in UD may be defined in language-specific documentation for use in individual languages. The latter is highly relevant for Hittite, as there must be language-specific tags not only for clitic chains and heterograms, but also for ergativity features and subject expressed with *-za*. There are also grammatical relations, including syntactic and semantic roles, that connect a head of phrase and a dependent word. In UD standard 37 types are defined for the universal use. In my paper I discuss the Hittite specific set of grammatical relations. In general, it is strongly recommended to keep universal tags as much as possible, to support the comparability of the languages, but Hittite certainly demands particular solutions discussed in this work, as well as in [8] and [15], which are planned for realization in ACHC.

References

1. Eva Pettersson and Beáta Megyesi (2018). The HistCorp Collection of Historical Corpora and Resources. // Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference, Helsinki, Finland, March 7-9, 2018.
2. The Ancient Greek and Latin Dependency Treebank (AGLDT). http://perseusdl.github.io/treebank_data
3. UD for Old East Slavic. https://universaldependencies.org/orv/index.html
4. Sophie Prévost, Achim Stein. 2013. Syntactic Reference Corpus of Medieval French (SRCMF) [version number]. ENS de Lyon/ILR Stuttgart.
5. Guglielmo Inglese, Maria Molina, and Hanne Eckhoff. 2018. Incorporating Hittite into PROIEL: a pilot project. // Andrew U. Frank, Christine Ivanovic, Francesco Mambrini, Marco Passarotti, and Caroline Sporlede, eds., Proceedings of the Second Workshop on Corpus-based Research in the Humanities, pp. 95–104.
6. PROIEL. https://proiel.github.io
7. Harry A. Hoffner, Jr. and H. Craig Melchert. 2008. A Grammar of the Hittite Language. Part 2: Tutorial. Eisenbrauns.
8. Erik Andersen, Ben Rozonoyer. 2020. A Small Universal Dependencies Treebank for Hittite. Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020), pages 1–7. Barcelona, Spain (Online), December 13, 2020
9. UD Hittite HitTB. 2022. https://universaldependencies.org/treebanks/hit_hittb/index.html
10. Maria Molina. The Annotated Corpus of Hittite Clauses (ACHC). 2015-. http://hittitecorpus.com
11. Guglielmo Inglese. 2015. Towards a Hittite Treebank. Basic Challenges and Methodological Remarks. // M. Passarotti, F. Mambrini, and C. Sporleder, editors, Proceedings of the Workshop on Corpus-Based Research in the Humanities (CRH).
12. Molina, Maria (2016). "Emphatic enclitic =pat in Hittite: function analysis and semantics of foci". In: Indo-European linguistics and classical philology — XX. Proceedings of the 20th Conference in Memory of Professor Joseph M. Tronsky 20–22 June 2016. St. Petersburg, 2016, 739–754. https://tronsky.iling.spb.ru/static/tronsky2016_01.pdf
13. Molina, Maria (2017). "Identificational foci in Hittite marked by =pat". Talk at the 50th Meeting of Societas Linguistica Europea (SLE 2017), 10–13 September 2017, Zurich.
14. Molina, Maria (2018). Word order in Hittite: corpus methods and analysis from typological perspective. PhD Diss. Moscow: Institute of Linguistics, Russian Academy of Sciences.
15. Maria Molina and Alexei Molin. 2016. In a Lacuna: Building a Syntactically Annotated Corpus for a Dead Cuneiform Language (on the basis of Hittite). // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016".
16. Universal Dependencies. https://aclanthology.org/2021.cl-2.11.pdf,
17. UD 2.0 Specification. https://arxiv.org/pdf/2004.10643.pdf

Maria Molina
Tel Aviv University

ANNOTATED CORPUS OF HITTITE CLAUSES

Computational
Perspectives on
Ancient Near Eastern
Literature, Art and
Material Culture

19th-21th February 2023

ACHC

1 Hittite corpora: why?

Anatolian languages being the first to leave Indo-European community, Hittite (and Luwian) reconstruction remains one of the most important steps to the Proto-Indo-European reconstruction. And we need a corpus to understand the language! It is the only way to work with a dead language, - we cannot just go and ask a Hittite native speaker.

Hittite

The most ancient Indo-European language attested in writing

Spoken in 18-12 cc. BC on the territory of Central Anatolia (modern Turkey)

The official language of the archives of the Hittite Empire, one of the Great Kingdoms of the Bronze Age

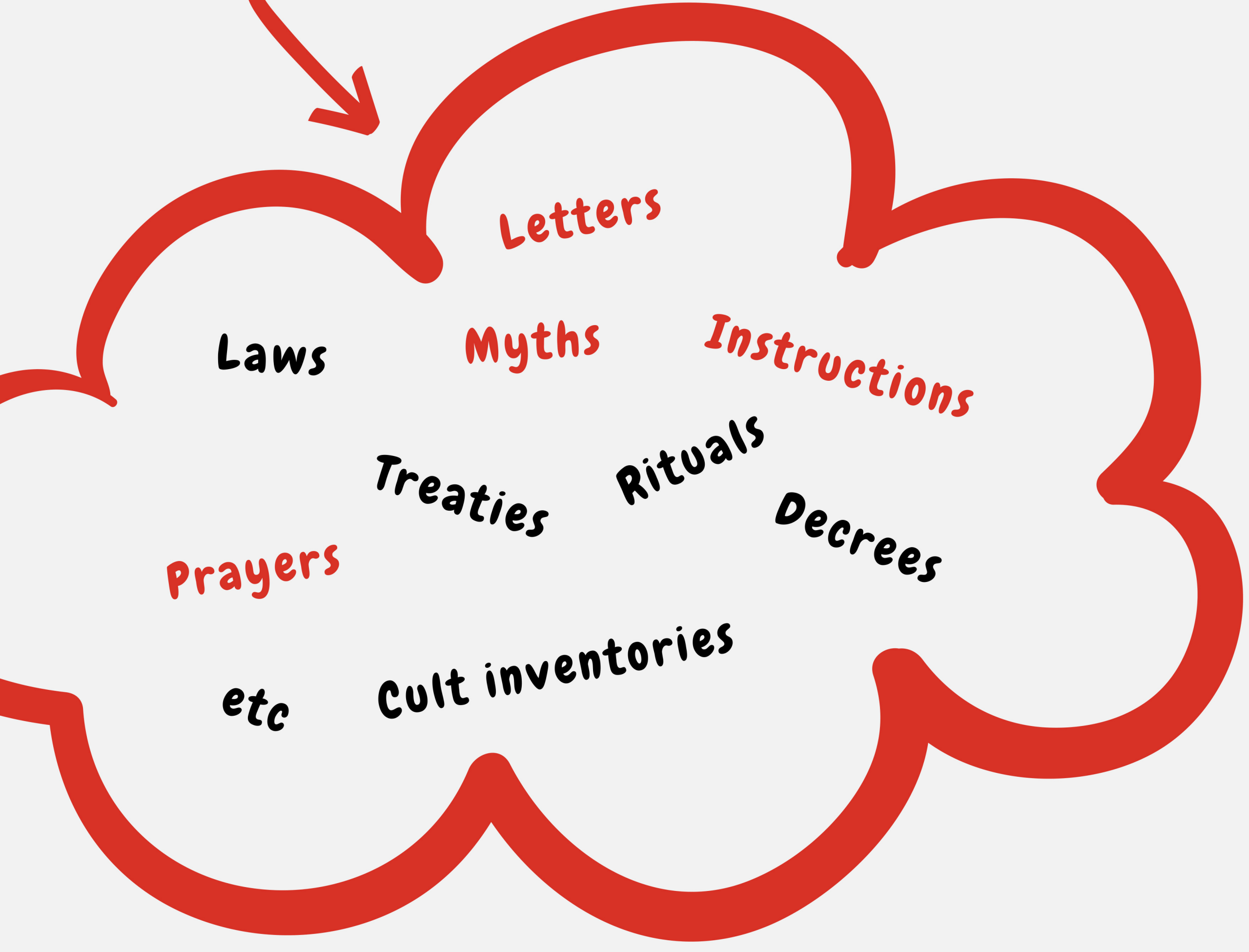
Hittite corpora: other projects

Textcorpora at Hethitologie Portal Mainz

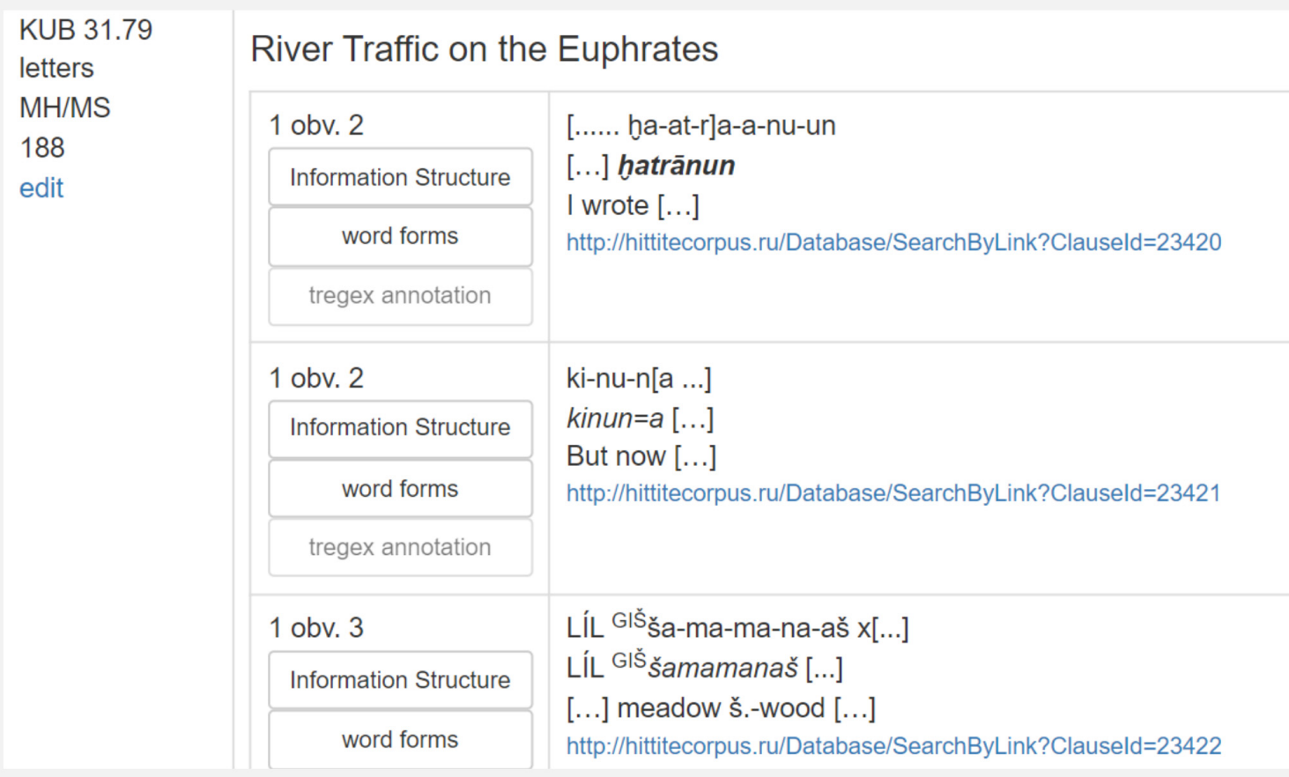
Hittitetexts.com

Goottite.com

2 Documents in Hittite



3 Annotation



KUB 31.79
letters
MH/MS
188
edit

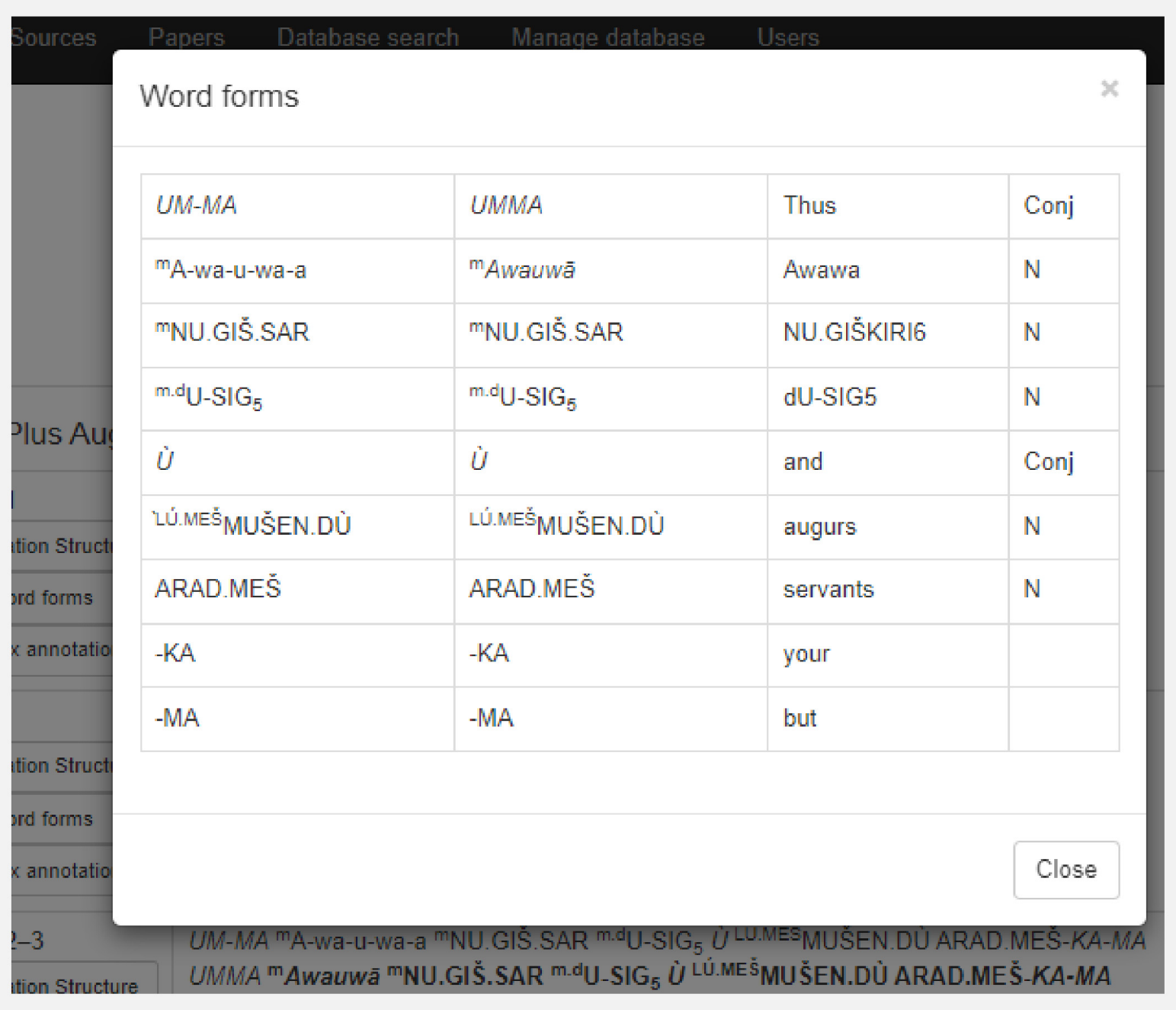
River Traffic on the Euphrates

1 obv. 2
[...... ḫa-at-r]a-a-nu-un
[…] ***ḫatrānun***
I wrote […]
http://hittitecorpus.ru/Database/SearchByLink?ClauseId=23420

1 obv. 2
ki-nu-n[a ...]
kinun=a […]
But now […]
http://hittitecorpus.ru/Database/SearchByLink?ClauseId=23421

1 obv. 3
LÍL ^{GIŠ}ša-ma-ma-na-aš x[...]
LÍL ^{GIŠ}*šamamanaš* [...]
[…] meadow š.-wood […]
http://hittitecorpus.ru/Database/SearchByLink?ClauseId=23422

Morphology



Word forms

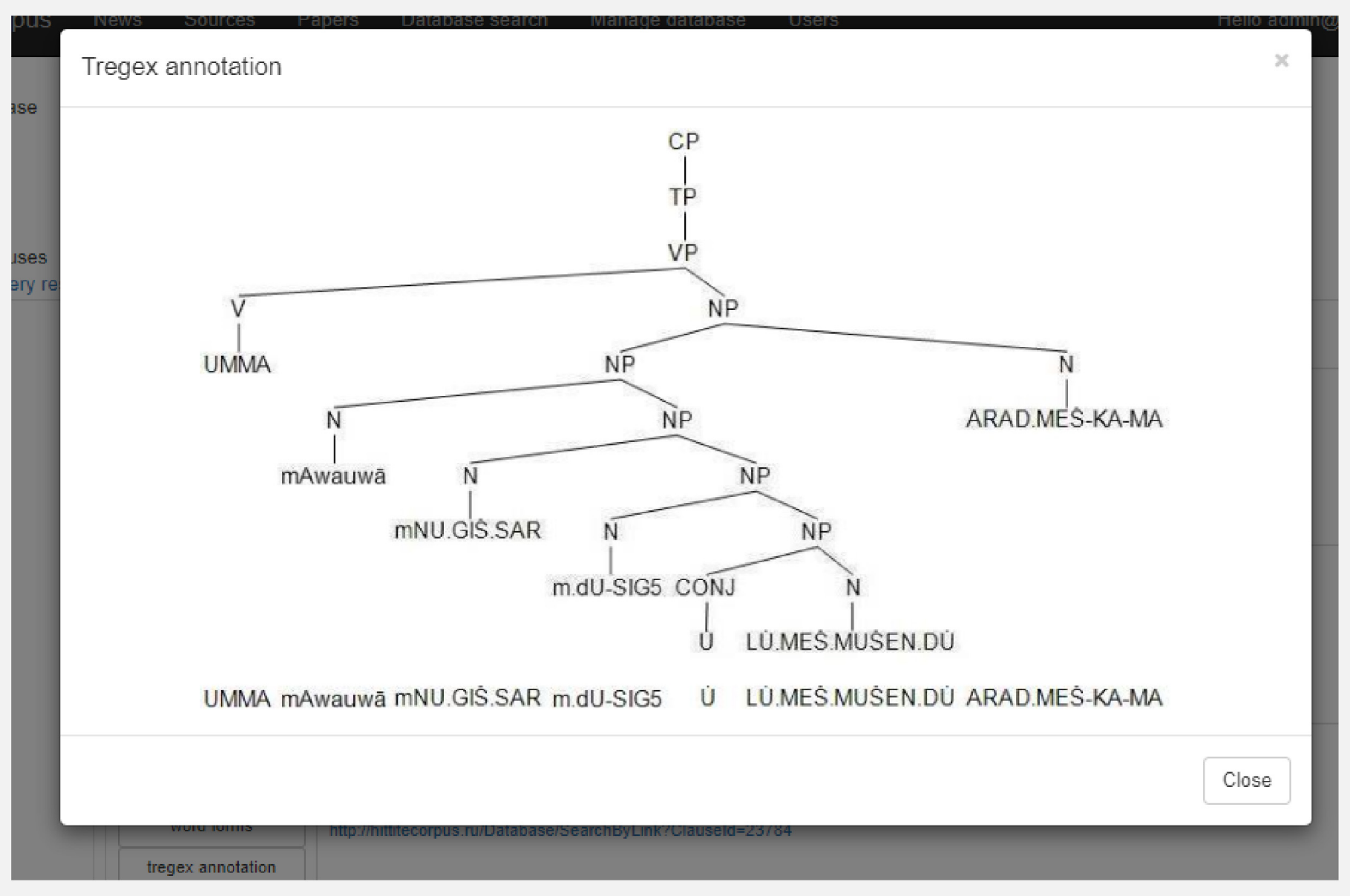
UM-MA	*UMMA*	Thus	Conj
^{m}A-wa-u-wa-a	^{m}*Awauwā*	Awawa	N
^{m}NU.GIŠ.SAR	^{m}NU.GIŠ.SAR	NU.GIŠKIRI6	N
$^{m.d}$U-SIG$_5$	$^{m.d}$U-SIG$_5$	dU-SIG5	N
Ù	*Ù*	and	Conj
$^{LÚ.MEŠ}$MUŠEN.DÙ	$^{LÚ.MEŠ}$MUŠEN.DÙ	augurs	N
ARAD.MEŠ	ARAD.MEŠ	servants	N
-KA	-KA	your	
-MA	-MA	but	

UM-MA ^{m}A-wa-u-wa-a ^{m}NU.GIŠ.SAR $^{m.d}$U-SIG$_5$ *Ù* $^{LÚ.MEŠ}$MUŠEN.DÙ ARAD.MEŠ-*KA-MA*
UMMA ^{m}*Awauwā* ^{m}NU.GIŠ.SAR $^{m.d}$U-SIG$_5$ *Ù* $^{LÚ.MEŠ}$MUŠEN.DÙ ARAD.MEŠ-*KA-MA*

- Tokenization
- Lemmatization
- Translation
- Glosses
- PoS
- UD index

ACHC

Syntax

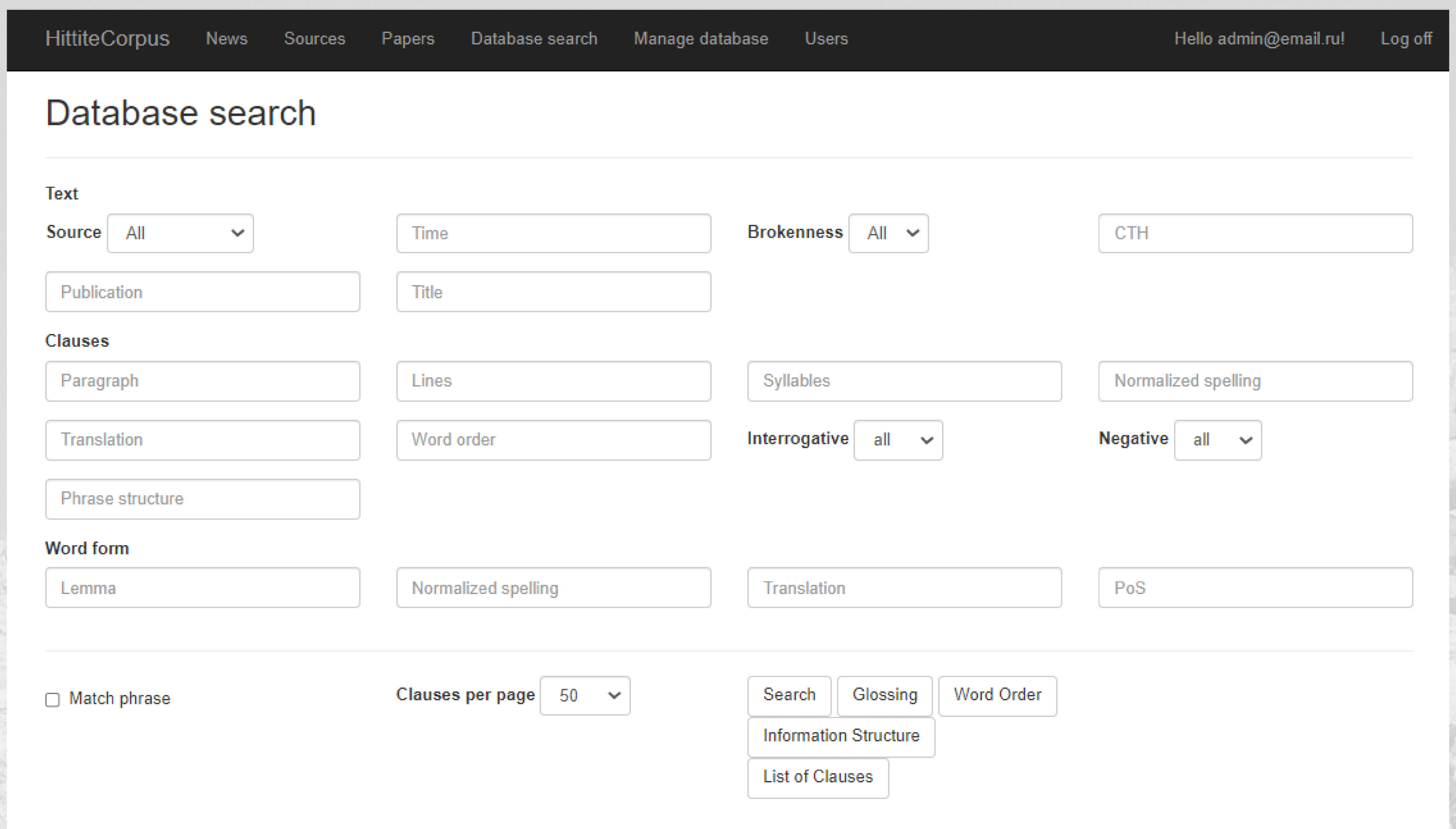


Tregex annotation

- SOV/OSV
- negation mark-up
- question mark-up
- constituency
- Universal Dependencies (under construction)

4 Practice!

Every skill
needs practice
You can do it!



Database search

Text
Source: All
Time
Brokenness: All
CTH
Publication
Title

Clauses
Paragraph
Lines
Syllables
Normalized spelling
Translation
Word order
Interrogative: all
Negative: all
Phrase structure

Word form
Lemma
Normalized spelling
Translation
PoS

☐ Match phrase
Clauses per page: 50

www.hittitecorpus.com