Morphological analysis and lemmatisation for Hittite

Innokentii Smirnov
Moscow State University
Philipp's University of Marburg
innokenty.smirnoff@gmail.com

Abstract

The development and evaluation of a morphological tagger and contextual lemmatiser for the Hittite language is described.

1 Introduction

In this paper, I describe the neural tagger and lemmatizer developed to assist the morphological annotation of Hittite texts. The tagger and lemmatiser were developed for two ongoing corpus annotation projects:

- The Corpus of Hittite Festival Rituals
- The Hittite Corpus of Divinatory Texts

It must be noted that both corpora comprise texts of very specific genres and application of the models to texts of other genres (e. g., history, law) would require additional training.

The morphological analysis of Hittite texts for these two corpora has been being performed by a dictionary-based morphological analyzer developed by researchers at the universities of Marburg and Mainz. Since the analyzer does not take context into account, its output requires manual disambiguation for ambiguous words, which make up 62 % of all tokens in the Festival Rituals corpus.

Therefore, it is desirable to develop a *context-aware* morphological analyzer and lemmatiser for Hittite. Since no large language models are available for Hittite, I have used convolutional and recurrent neural networks (LSTM) for that purpose.

Firstly, I describe the difficulties accompanying the morphological analysis of Hittite texts.

- Enclitic clusters
- Syncretism
- Cross-paradigmatic homonymy

Afterwards, I describe the architecture devised to overcome a part of these complications. Finally, the performance of the system is reported.

2 The problem

2.1 Enclitics

In Hittite texts, tokens are delimited by whitespace (on cuneiform tablets as well as in transliterations). However, not all tokens are words - the minimal autonomous and syntactically indivisible units of language occupying the nodes of the syntactic tree. Some tokens are *clitic groups* - combinations of a word and following *enclitics*.[1] A clitic is a linguistic sign which is syntactically indivisible but not autonomous (i. e., it cannot constitute an utterance on its own). It is prosodically dependent on some full word, which is called its *host*. For an enclitic, this is the nearest full word on the left. The enclitics attached to some word constitute a *clitic cluster*.[2]

In Universal Dependencies, clitic groups are usually represented as multiword tokens. One should, however, bear in mind that enclitics are not words. Yet, they occupy separate nodes of the syntactic tree, which justifies this practice.

In Hittite texts, enclitics are *not* delimited by whitespace either from one another or from the host.

Since some enclitics (the argument pronouns) inflect for case, number and gender, they require their own morphological analyses. Consequently, the morphological analysis of a text in Hittite could be accomplished in two ways.

1. Firstly, each clitic group is segmented (split) into a word and accompanying enclitics. Afterwards, enclitics are processed as if they were normal words: the sequence input to

[1] The notion of clitic group as a prosodic structure distinct from the phonological word is controversial (Spencer and Luis, 2012). I use the term to refer to clitic-host combinations irrespective of their possible status as phonological words or phrases.

[2] Not to be confused with the clitic group, which consists of the host and the clitic cluster.

Table 1: Positions of the enclitic cluster

Conj.	Quot.	Pl.: 3 p. Dat., 1-2 p.	3 p. Nom.-Acc.	Sg.: 3 p. Dat., 1-2 p.	Refl.	Local
ma 'but' *ya* 'and'	*wa* *war*	*nnaš* 1PL.DAT/ACC *šmaš* 2PL.DAT/ACC *šmaš* 3PL.DAT	*aš*, *an*, *at* *e*, *uš*, *e*	*mu* 1SG.DAT/ACC *tta* 2SG.DAT/ACC *šši* 3PL.DAT	*za*	*kkan* *šša*n etc.

the tagger consists of full words and enclitics as separate tokens. It must be noted that the segmentation of clitic groups is context-dependent: Some clitic clusters are homonymous with each other, and some enclitics are homonymous with inflectional endings.

This is the approach taken by (Brusilovsky and Tsarfaty, 2022) for Hebrew and Arabic. (In those languages, proclitic prepositions and articles are not delimited by whitespace from their hosts.)

2. The morphological analysis is performed directly, and no segmentation (splitting) of clitic groups is performed. Each clitic group is assigned a complex morphosyntactic description which takes into account the word as well as the enclitics.

Since annotations in the corpus do not include segmented representations for clitic groups, only the second method is straightforwardly applicable.[3] It also has the advantage of recovering explicitly the morphosyntactic properties required to disambiguate between homonymous clitic clusters or enclitics and endings.

An immediate question is how exactly the morphosyntactic description of a clitic group should be organised and how it can be generated by the tagger.

If the concatenation of a word's morphosyntactic description and the tags of all adjacent enclitics was treated as a single indivisible label, the amount of labels would reach 10^6, since words can have several hundred distinct morphosyntactic descriptions and there are about 2,500 distinct clitic clusters. The majority of these composite labels simply do not occur in the corpus.

Therefore, we apply the multiclass multilabel classification (McMl) model (Tkachenko and Sirts, 2018), i. e. we use a separate decoder (linear classifier with softmax activation) for each position of the clitic cluster, while the encoder (a bidirectional LSTM) is the same.

In total, there are seven positions in the enclitic chain. They are represented in table 1 following (Hoffner and Melchert, 2008).

[3] One could, however, attempt splitting the clitic groups automatically.

Table 2: Paradigm of the adj. *kunna-* 'straight, correct, fortuitous'

	SG	PL
NOM.C	kunna-š	kunn-eš
ACC.C	kunna-n	kunn-uš
N/A.N	kunna-n	kunn-a
VOC	kunna	(=NOM)
GEN	kunn-aš	kunn-aš
D/L	kunn-i	
ALL	kunn-a	
ABL	kunn-az	
INS	kunn-it	

2.2 Syncretism

A striking pattern of syncretism in the Hittite language is the coincidence of nom. sg. and gen. sg. in the most productive declension - the stems in *-a*, also known as the thematic stems (table 2). The syncretism arises due to the deletion of the stem-final theme vowel before vocalic endings.

Since genitival modifiers precede their syntactic head in most functions, the ambiguity can be resolved in context. However, the homonymy of the gen. pl. and dat.-loc. pl. endings with the gen. sg., which is *not* restricted to the *a*-stems, further complicates the disambiguation.

2.3 Cross-paradigmatic homonymy

Some extremely frequent irregular verbs are homonymous in certain grammatical forms.

Thus, the 3 pl. *tianzi* belongs either to *dai-/te-/ti(ya)-* 'put' or to *tiya-* 'step' (ex. 1-2). The 3 sg. *tai* is a form either of *dai-/te-/ti(ya)-* 'put' or of *dā-/d-* 'take' (ex. 3-4).

It is important for further discussion that the verb forms in question have *the same set of morphosyntactic properties* associated with them, and the ambiguity cannot be resolved by the morphological

Figure 1: Ambiguous word-forms *tianzi* and *tai*

(1) $^{LÚ.MEŠ}$MUḪALDIM ÚKUŠ *kakkap-an=na* *ti-anzi*
cook[NOM.PL] cucumber[ACC.PL] partridge-ACC.SG=and put-3PL.PRS
'The cooks put cucumbers and partridge (on the table).'
CTH 609, IBoT 3.1, Rs. 76'

(2) $^{LÚ.MEŠ}$MUḪALDIM=ya *ḫantezi* *ti-anzi*
cook[NOM.PL]=and forward step-3PL.PRS
'And the cooks come forward.'
CTH 609, IBoT 3.1, Rs. 70'

(3) *nu=šši=kan* ^{LÚ}ḪAL ^{GIŠ}EREN *kiššar-i=šš-i* *ta-i*
CONN=3SG.DAT=LOCP priest[NOM.SG] cedar[GEN.SG] hand-D/L.SG=his-D/L.SG put-3SG.PRS
'The priest puts the sceptre of cedar into his hand (=gives him the sceptre of cedar).'
CTH 712, KBo 35.168, Vs. I 9'-10'

(4) LUGAL*u-š=za* GÍR ZABAR *ta-i*
king-NOM.SG=REFL knife[ACC.SG] bronze[GEN.SG] take-3SG.PRS
'The king takes a knife of bronze.'
CTH 712, KUB 27.1, Rs. III 20

tagger alone, nor by a lemmatiser of the type described in (Malaviya et al., 2019). Our solution to this problem will be described below.

3 The models

3.1 A note on BERT

State-of-the-art results on morphological tagging and lemmatisation have been achieved by using pretrained embeddings from transformer language models such as BERT, as demonstrated by SIGMORPHON-2019 shared task 2 results (McCarthy et al., 2019); see especially the two winning systems - CHARLES-SAARLAND (Kondratyuk, 2019) and UDPipe (Straka et al., 2019).

This method would be suboptimal for Hittite, however. The amount of data available does not seem sufficient to train a Hittite BERT, and the complete lack of Hittite data in the multilingual BERT training corpus makes its applicability questionable. Therefore, convolutional and recurrent neural networks appear to be the reasonable solution.

3.2 Morphological tagger

Our neural morphological tagger is essentially that of (Heigold et al., 2017) as far as the encoder is concerned. It is a two-level network with a character encoder[4] and a word-level bidirectional LSTM. I have experimented with word embeddings as well,

[4]Either recurrent or convolutional. For Hittite, the latter has shown better results on unknown word-forms.

Table 3: Determinatives and their meaning

Det.	Mean.	Example
KUR	country	^{KUR}ELAM 'Elam'
URU	city	^{URU}KÁ.DINGIR.RA 'Babylon'
DUG	vessel	^{DUG}*kangur* 'mug'

but, giving a slight increase in the accuracy on ambiguous word-forms, it led to a significant decrease in the ability of the network to handle unknown words.

Apart from the character-level representation, the word-level encoder receives the embeddings of *determinatives* as an additional feature. Determinatives are special cuneiform signs that were used especially with proper names to specify the semantic category (mountains, rivers, cities, metals etc.) of the following noun. A small selection is given in table 3. Taking determinatives into account has led to an increase in tagging accuracy of approximately 2 %.

The decoder consists of several classification layers with softmax activation, each being responsible for some positional class of enclitics or the main word-form. Architecturally, this is the McMl model devised by A. Tkachenko and K. Sirts (Tkachenko and Sirts, 2018). An important difference is that the multiple labels correspond not to different morphosyntactic properties but rather to distinct elements of the clitic group.

3.3 Lemmatiser

The lemmatiser is a sequence-to-sequence character-level LSTM transducer with soft attention. Before generating the next symbol of the lemma, the output of the attention layer is concatenated with a context vector. The context vector is built by a two-level encoder similar to the one used for morphological tagging which is trained together with the transducer.

4 Evaluation

4.1 Train-test split

The Corpus of Hittite Festival Rituals is an online publication of ancient texts and not a corpus designed specifically for training and evaluating neural taggers. A great deal of texts in the corpus have copies or duplicates - they were either written in several exemplars or copied by later scribes. In both cases minor orthographic, lexical and morphological variation is possible. This means we cannot simply apply the standard train-test split routine to an array of clauses, even after merging clauses which are completely identical.

Fortunately, each text in the corpus is assigned one of the CTH-numbers (fr. *Catalogue des textes Hittites*), which refer to thematic text groups. Copies of a particular text belong to the same group. The reverse is not necessarily true but, ideally, a CTH-group should contain copies and variants (which have minor differences in their content) of the same text.

For evaluation, we randomly selected 10 CTH-groups in such a way that the test dataset comprises no less than 10 % of all clauses in the corpus. No text or clause from these groups has been included in the training dataset.

4.2 Metrics

The accuracy is given in table 4. It was computed on all tokens except for digits (which are rendered by arabic numerals in transliterations).

Table 4: Accuracy

		Neural network only	With dict. analyzer
Token	Morph. tag	81.51 %	83.9 %
	Lemma	89.53 %	94.51 %
Clause	Morph. tag	50.24 %	55.04 %
	Lemma	67.02 %	80.72 %

As is evident from table 4, joint application of the neural networks with the rule-based morphological analyzer leads to a significant increase in accuracy. If the lemma or morphological tag produced by the network was not among the alternative analyses given by the analyzer, the lemma or tag with greatest frequency[5] was selected from those given by the analyzer.

This result has led me to consider another method of lemmatisation: lemmatisation can be treated as a sequence classification problem, where each lemma corresponds to a separate class. The lemmatiser is thus architecturally analogous to a tagger and predicts a probability distribution over the entire lexicon. This allows us to select the most probable lemma from those suggested by the rule-based analyzer. This method achieved the accuracy of 95.17 % (against 94.51 % above). Its additional advantage is that a tagger is much faster to train than a sequence-to-sequence transducer. However, the obvious problem of ambiguity between out-of-vocabulary lemmata remains thus far unsolved.

4.3 Out-of-domain test

Since only a small part The Hittite Corpus of Divinatory Texts has been annotated so far, we have not used it for training and reserved the divinatory texts for an out-of-domain test. The results show that the models do not generalize well to texts of other genres: the tagging accuracy was only 59.17 %, and the lemmatiser achieved 72.20 %.[6] This is partly attributable to the abundance of logograms in divinatory texts (I. Yakubovich, p. c.) and can also be explained by the fact that many vocabulary items which are frequent in divinatory texts (e. g., most bird names) do not occur in the training dataset.

5 Conclusion

The models described above have been used to lemmatise and tag the remaining unannotated texts in The Corpus of Hittite Festival Rituals. As the digitalisation of other text genres will proceed, we will be able to find out whether their applicability is indeed restricted to the domain of Festival Rituals or the divinatory texts are rather an exception.

[5] The most frequent tag or lemma for the word-form in question, if it the word-form had occurred in the training dataset; the one with gratest absolute frequency otherwise. The lemmata were additionally filtered according to the tags the analyzer associated with them.

[6] Without the rule-based analyzer.

Acknowledgments

The author is grateful to prof. Elisabeth Rieken for providing the training data and for the advice to take demonstratives into account.

References

Idan Brusilovsky and Reut Tsarfaty. 2022. Neural token segmentation for high token-internal complexity. *arXiv preprint arXiv:2203.10845*.

Georg Heigold, Guenter Neumann, and Josef van Genabith. 2017. An extensive empirical evaluation of character-based morphological tagging for 14 languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 505–513, Valencia, Spain. Association for Computational Linguistics.

H.A. Hoffner and H.C. Melchert. 2008. *A Grammar of the Hittite Language*. Eisenbrauns.

Dan Kondratyuk. 2019. Cross-lingual lemmatization and morphology tagging with two-stage multilingual BERT fine-tuning. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 12–18, Florence, Italy. Association for Computational Linguistics.

Chaitanya Malaviya, Shijie Wu, and Ryan Cotterell. 2019. A simple joint model for improved contextual neural lemmatization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1517–1528, Minneapolis, Minnesota. Association for Computational Linguistics.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

A. Spencer and A.R. Luis. 2012. *Clitics: An Introduction*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Milan Straka, Jana Straková, and Jan Hajic. 2019. UDPipe at SIGMORPHON 2019: Contextualized embeddings, regularization with morphological categories, corpora merging. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 95–103, Florence, Italy. Association for Computational Linguistics.

Alexander Tkachenko and Kairit Sirts. 2018. Modeling composite labels for neural morphological tagging. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 368–379, Brussels, Belgium. Association for Computational Linguistics.

A CTH-groups included in the test dataset

A.1 The Corpus of Hittite Festival Rituals

The CTH groups which have been used as the main test dataset can be seen in table 5.

Table 5: Main test dataset: The Corpus of Hittite Festival Rituals

CTH 482	Reform of the cult of the goddess of the night of Šamuḫa by Muršili II
CTH 488	Ritual referring to Ḫamrišhara
CTH 609	AN.DAḪ.ŠUM^{SAR}, day 11
CTH 615	AN.DAḪ.ŠUM^{SAR}, days 22–25: for Ištar of Ḫattarina
CTH 626	Festival of haste (EZEN$_4$ nuntarriyašhaš)
CTH 642	Festival fragments referring to the vegetation god Zinkuruwa
CTH 694	Fragments of festivals for Ḫuwaššanna
CTH 699	Festival for Teššup and Ḫebat of Lawazantiya
CTH 700	Enthronement ritual for Teššup and Ḫebat
CTH 712	Festival for Ištar of Šamuḫa

A.2 The Hittite Corpus of Divinatory Texts

The texts in table 6 have been used for the out-of-domain test.

Table 6: Out-of-domain test: The Hittite Corpus of Divinatory Texts

CTH 532	Lunar eclipse
CTH 549	Liver omens: “position” (KI.GUB)
CTH 561	Oracles concerning the king’s campaigns in the Kaška region
CTH 563	Oracles concerning the overwintering of the king
CTH 573	Bird (MUŠEN) oracles
CTH 581	Letters about oracles