Distantly Supervised POS Tagging of Low-Resource Languages under Extreme Data Sparsity: The Case of Hittite

Maria Sukhareva[†], Francesco Fuscagni[‡], Johannes Daxenberger[†],
Susanne Görke[¶], Doris Prechel[‡] and Iryna Gurevych[†]

[†] Ubiquitous Knowledge Processing Lab (UKP)
Department of Computer Science, Technische Universität Darmstadt
www.ukp.tu-darmstadt.de
sukhareva@ukp.informatik.tu-darmstadt.de

[‡] Altorientalische Philologie, Institut fr Altertumswissenschaften
Johannes Gutenberg-Universität Mainz

[¶] Akademie der Wissenschaften und der Literatur Mainz
Philipps-Universität Marburg

Abstract

This paper presents a statistical approach to automatic morphosyntactic annotation of Hittite transcripts. Hittite is an extinct Indo-European language using the cuneiform script. There are currently no morphosyntactic annotations available for Hittite, so we explored methods of distant supervision. The annotations were projected from parallel German translations of the Hittite texts. In order to reduce data sparsity, we applied stemming of German and Hittite texts. As there is no off-the-shelf Hittite stemmer, a stemmer for Hittite was developed for this purpose. The resulting annotation projections were used to train a POS tagger, achieving an accuracy of 69% on a test sample. To our knowledge, this is the first attempt of statistical POS tagging of a cuneiform language.

1 Introduction

Natural Language Processing (NLP) for historical languages is a challenging task. The mere digitization of historical texts can take several years as the original data vary from ancient manuscripts to clay tablets which only a trained historical linguist can read and transliterate. The manual morphosyntactic annotation of the digitized historical resources demands a rare expertise and is a slow and painstaking process (Bennett et al., 2010). It is frequently impossible to annotate the amount of data sufficient for training a supervised part-of-speech (POS) tagger. Thus, NLP for historical languages frequently uses distantly supervised methods to compensate for the lack of training data (Piotrowski, 2012).

Traditionally, historians and historical linguists apply manual qualitative methods to the data. Such work usually involves a narrow expertise that focuses on a particular phenomenon or a time period. For example, presently, Hittite texts can only be read and understood by trained cuneiform specialists whose scope of interests is confined to certain texts, diachronic periods or linguistic phenomena. Statistical machine translation (SMT) and information retrieval (IR) methods would make these texts available to a wider public, including historians and sociologists (Daxenberger et al., 2017). The automatic methods are also applicable to whole corpora and have a much wider coverage than qualitative analysis. However, for optimal performance, SMT and IR need basic linguistic annotation such as POS tags and syntactic parses that are currently not available for Hittite. Thus, we propose a distantly supervised tagger and an unsupervised stemmer for Hittite which can be the first milestone in creating more advance NLP tools for cuneiform languages.

Performance of distantly supervised methods such as annotation projection or cross-lingual tool adaptation depends on the diachronic relatedness between the source and the target languages. For example, annotation projection from modern English into middle English gives better results than into old English because middle English grammatically and lexically resembles modern English much more than Old English (Sukhareva and Chiarcos, 2014). Annotation projection is thus typically applied to related languages (Tiedemann

Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. Proceedings, pages 95–104, Vancouver, BC, August 4, 2017. ©2017 Association for Computational Linguistics

and Agic, 2016).

In this paper we show that our annotation projection method is robust enough to reach decent performance on a highly inflectional language that has been extinct over millennia and does not have any modern relatives. Also, the data sparsity caused by multilingualism and rich Hittite morphology poses additional challenges for statistical NLP methods. On a small parallel corpus of Hittite and German, we use character-based alignment to create an unsupervised stemmer for Hittite and word-based alignment as a basis for annotation projection from POS tagged German translations. The resulting POS projections are used as training data for a POS tagger. Our evaluation shows that stemming Hittite and German texts prior to annotation projection largely improves POS tagging accuracy for Hittite as compared to a POS tagger trained on unstemmed projections.

The paper is structured as follows: Section 2 introduces the data used in this research and outlines linguistic characteristics of Hittite that affect the performance of our method. It also describes the manually annotated evaluation dataset for Hittite that was created for the sake of this study. Our main contributions, the unsupervised Hittite stemmer and annotation projection approach to Hittite POS tagging, are described in Section 3. The evaluation of the presented approach is in Section 4. Section 5 discusses related work and the state-of-the-art of NLP for cuneiform languages. Finally, we discuss the results and outline future work in Section 6.

2 Data

Hittite texts pose such challenges as developed inflectional morphology, non-standardized orthography, diachronic variations and multilingualism. Given a relatively small amount of data available for Hittite, direct application of state-of-the-art NLP approaches leads to sub-optimal results. Also, modern machine learning techniques are not directly applicable because of the limited amount of data. With data sparsity being the main obstacle, we see the solution in understanding the linguistic reasons for data sparsity and based on them to exploit means of data sparsity reduction.

2.1 Hittite language

Hittite is an extinct language spoken between 16 and 12 c.c. BCE in the territories of modern Turkey and Northern Syria. It is an inflectional synthetic Indo-European language. Hittite belongs to a dead Anatolian branch of Indo-European languages along with Luwian and Palaic. Hittite as well as its closely related languages do not have any modern descendants. This poses an additional challenge to the application of distantly supervised methods to our data as their performance depends on diachronic relatedness (Section 1).

There are three chronological periods of the Hittite language: old Hittite (OH, 1650-1500 BCE), middle Hittite (MH, 1500 - 1350 BCE) and new Hittite (NH, 1350 - 1180 BCE). Diachronic orthographic variations are strongly pronounced between the time periods: The shapes of many cuneiform signs differ in these three periods. Also, the so called plene writing occurs when a vowel already present in a cuneiform sign is expressed by a further unnecessary vocal. Plene writing is a typical feature of OH and MH texts, disappearing progressively with NH and is practically absent in late NH.

During all periods Hittite was a highly inflectional language with a wide variety of word forms. For example, the nominal declension included inflectional paradigms determined by two genders, nine cases and two numbers (van den Hout, 2011). Also, adjectives had a rich inflectional paradigm as they agreed with nouns in gender, case and number. As for the verbal inflectional paradigm, it was relatively simple and was determined by only two tenses, two moods and two voices. Though Hittite in all periods did not have any grammatical definiteness marking (e.g. articles), it had determiners that would indicate the class of the nouns (e.g. city, land, woman, bread, etc.) and were expressed in writing by unpronounced Sumerograms (e.g. $^{\text{URU}}$*ḫatti, "the land of ḫatti"*; $^{\text{GIŠ}}$*natḫi, "(wooden) bed"*)

To sum up, rich inflectional morphology, spelling variations and diachronic variations in Hittite greatly increase the data sparsity making the automatic statistical processing of Hittite texts extremely challenging. The key to successful automatic annotation of Hittite is the reduction of the data sparsity by normalizing diachronic variations and reducing the word form paradigm to a single stem or lemma. While we leave the problem of normalization open, the paper will further discuss the reduction of word forms and propose a method for data sparsity reduction through stemming.

(1) Types of transliteration used in the DPHT and multilingualism.

a.	*nu*	*ma-ah-ha-an*	*A-NA*	$^{\text{GIŠ}}$GU.ZA	*A-BI-IA*	*eš-ha-ha-at*	(Syllabic transliteration)
	nu	*mahhan*	*ANA*	$^{\text{GIŠ}}$GU.ZA	*ABI=IA*	*ešhahat*	(Bound transcription)
	HIT	HIT	AKK	SUM	AKK	HIT	(Language)
	and	as soon as	on	throne	father-my	sit	

And as soon as I sat down on the throne of my father

2.2 Corpus of Hittite Texts

The Digitale Publikation Hethitischer Texte corpus (DPHT) is available via the Hittitology Portal Mainz (HPM).[1] It covers more than 30,000 mostly fragments of clay tablets that have been archived in Ancient Anatolia, nowadays Turkey, during the later half of the second millennium BCE. Most of the texts were found in Hittite capital Hattusa, only smaller archives came to light in other towns of the Hittite Empire. Therefore, Hittite texts used in this research do not have dialectal variations which contribute to the data sparsity and negatively influence the performance of the NLP pipeline.

The DPHT is relatively small as compared to modern corpora and has only 60,058 tokens. An additional challenge for NLP processing of Hittite texts is posed by their extreme multilingualism. Several languages are found in the texts: Hittite, Luwian and Palaic are Indo-European languages, Hattic, Hurrian and Sumerian are isolated agglutinating languages and Akkadian is a Semitic language. Sumerian and Akkadian words are particularly frequent in Hittite texts (see ex. 1). Some words can be written both with sumerograms and with akkadograms or in syllabic Hittite. For example, *"god"* is often written by the sumerogram DINGIR. Furthermore, the akkadogram *ILU(M)* and the Hittite word *iu(na)* can be found in the corpus.

Texts cover various genres; most of them belong to a religious sphere, like festival descriptions or magic rituals, but also historic documents like treaties, annals, etc. have been found. As every genre is associated with genre-specific vocabulary and syntactic constructions, this genre variety can negatively affect the performance of the POS tagger. Furthermore, diachronic variations in spelling, morphology and syntax can have a negative impact on the tagging accuracy. The texts cover the whole of Hittite history, from OH throughout MH to NH. More than two thirds of all Hittite texts in our data were written in NH.

Hittite texts are transliterated in accordance with the syllabic and logographic structure of their signs. The transliteration conventions are compatible with generally recognized rules of transliteration of cuneiform languages.[2] The DPHT provides syllabic transliteration which is a syllable-wise literal transliteration of the original texts. Furthermore, a bound transcription is given which focuses on word transcription and is closer to the way the words were most likely pronounced (ex. 1). In our experiments, we used bound transcription as it has less diachronic spelling variations.

2.3 POS Annotation of Hittite

In order to evaluate our pipeline, a hittitologist and co-author of this paper annotated selected documents with Universal POS tagset (Petrov et al., 2012). These were only used for the evaluation. As the pipeline was trained on a diachronic corpus containing various genres, we balanced the evaluation set and included texts that represent all the time periods. Table 1 shows the list of the texts included in the evaluation set. It totals 969 tokens and has proportionally balanced texts from NH, MH and OH. The complexity of the annotation process varied based on the period. While MH and NH are well-researched and there are many available texts in MH and NH, OH is very complicated and has words whose translation is not known.

We decided to create a balanced evaluation set rather than creating three evaluation sets for various periods due to practical reasons. First, annotation of this test set was a painstaking task that demanded a rare expertise. It was practically impossible to annotate large enough evaluation sets for all the three periods. Second, we could not split the training data into time periods as there would not be enough data to train a classifier for

[1] `https://www.hethport.uni-wuerzburg.de`

[2] `http://www.hethport.uni-wuerzburg.de/HPM/hpm.php?p=hpmguide`

Title	Period	Tokens
Purification Ritual for the Royal Couple	OH	113
Instructions for Bodyguard	MH	144
Military Instructions of Tutḫaliya I	MH	137
Ten Years Annals of Muršili II	NH	390
Prayer of Muršili II	NH	127
Apology of Hattušili III	NH	58

Table 1: POS annotated evaluation set

each period. Thus, the POS tagger (see Section 3) was both trained and tested on data from various periods.

3 NLP Pipeline for Hittite

Automatic morphosyntactic annotation of Hittite is a non-trivial task. As discussed in Section 2, the Hittite texts are affected by diachronic variations in the lexicon, morphosyntax and orthography. Additionally, Hittite is a highly inflectional language with the immediate consequence of high type-token ratio. All of these factors lead to a data sparsity that is the key obstacle for statistical NLP processing of the data.

We present an approach that builds a NLP pipeline for automatic morphosyntactic annotation of Hittite. The pipeline (Figure 1) consists of four modules: preprocessing, data sparsity reduction, annotation projection and POS tagging. The initial data are just primary texts that are neither tokenized nor linguistically annotated. The transliteration and translation texts are clause-wise aligned which makes it possible to create word-based and character-based alignment. The morphosyntactic annotations are then projected into the Hittite texts from their German translations.

The quality of the annotation projection imminently depends on the quality of the alignment which is strongly affected by the data sparsity. Nevertheless, some of the data sparsity is relatively easy to reduce. For example, German, though by far not as inflectionally rich as Hittite, still has a relatively rich inflectional morphology. Thus, a noticeable improvement on the annotation projections can already be reached by stemming the German texts. Hittite stemming is also beneficial for word alignment quality though it is a more challenging task as there are no off-the-shelf Hittite stemmers or lemmatizers. Thus, this approach also proposes an unsupervised method for stemming of Hittite.

The final element of the pipeline is the POS tagging of Hittite texts. The annotation projections are used as training data for a supervised POS tagger. Presently, there are no POS-annotated datasets for Hittite available. We manually annotated several text excerpts to evaluate the output of the Hittite NLP Pipeline (see Section 2.3).

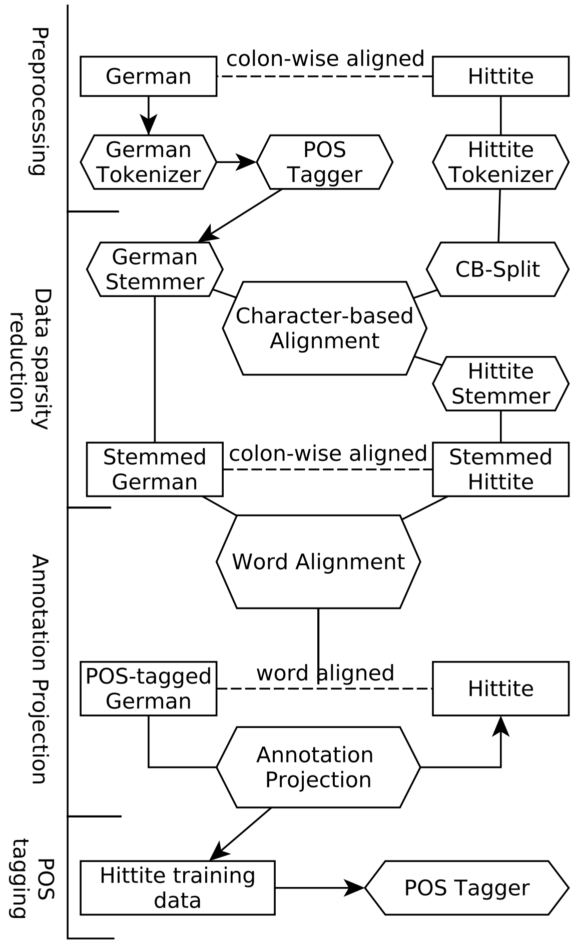

Figure 1: Morphosyntactic NLP pipeline for Hittite

3.1 Data Preprocessing

The input to the pipeline are the initial digitized Hittite transliterations and their German translations provided in a XML format. As modern principles of text segmentation into clauses, sentences and phrases appeared only a few centuries ago, the original Hittite texts do not have any text segmentation nor any punctuation. During transliteration, the texts were split in paragraphs and *colons*. Colons in most cases correspond to clauses which

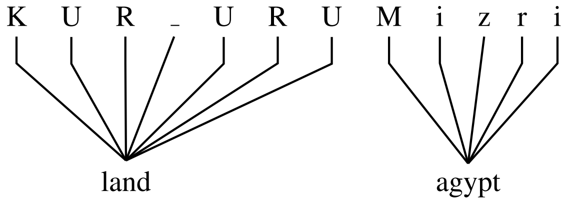

Figure 2: An example of character-based alignment of the Hittite phrase *"the land of Egypt"* with German stems.

start with an introductory particle *nu* or with a conjunctive adverb (e.g. *mahhan* "when"). Each colon (with rare exceptions) has a verb in the final position which is the standard word-order in Hittite. Colons as well as paragraphs are aligned to German translations.

The NLP pipeline for Hittite was built on the basis of the uimaFIT library and, more specifically, DKPro Core libraries (Eckart de Castilho and Gurevych, 2014). In the preprocessing stage, the colon alignments were extracted from the XML files. We used the off-the-shelf OpenNLP tokenizer[3] trained on Tiger corpus (Brants et al., 2004) to tokenize German text. As there is no available tokenization for Hittite, it was decided to use white spaces and equals sign "=" as token separators. It is important to mention that as there is no punctuation in Hittite transliterations, the whitespace tokenization worked quite well but is not sufficient as many function morphemes are bound. The bound function morphemes include, for example, location affixes or possessive pronominal suffixes. Such bound morphemes are usually suffixes marked in the transliteration by the equals sign (e.g. *ABU=IA "father=my"*).

3.2 Reduction of data sparsity

One of the most straight-forward ways to reduce the data sparsity is through lemmatization. Nevertheless, there are no off-the-shelf lemmatizers for Hittite neither is there a machine readable Hittite dictionary with sufficient coverage to create a dictionary-based lemmatizer. An alternative approach is to use stemming. Hittite has developed paradigms of outer flection and does not demonstrate many cases of inner flection, thus, root morphemes do not have a large variance and the separation of inflectional morphemes is likely to suffice. We implemented a character-based stemmer for Hittite that relies on character-based alignment. The purpose of the stemmer is to separate all the affixes from the root morpheme. Affixes are bound morphemes that include both suffixes and prefixes. Affixes can be both derivational and inflectional. The Hittite stemmer splits a word into three parts: prefixes, root and suffixes. As the main purpose of the stemmer is to reduce data sparsity rather provide a morphological analyzer the stemmer does not split prefixes or suffixes e.g. if a word has several suffixes the stemmer treat it as one suffix. Further on, the paper will refer to such complex affixes as "prefix" or "suffix".

First, the parallel German texts were stemmed. For this purpose, we used Snowball stemmer for German.[4] Then, we split all the Hittite words into characters and word boundaries were marked with a special character. To create a character-based (CB) alignment we used Phrasal ITG Aligner (Neubig et al., 2012).

Figure 2 shows a character alignment of Hittite phrase "KUR $^{\mathrm{URU}}$*Mizri*" to the German stems *land* and *agypt*.[5] Both the Hittite noun KUR meaning *land* and the Hittite determiner URU are aligned to the German stem *land* while *Mizri* is aligned to the German stem *agypt*. This example shows the basic principals of how the stemmer works: Hittite substring aligned with German stems are likely to be stems themselves. It is particularly effective in Hittite because of the abundance of noun determiners that are frequently translated by a separate German word.

The resulting CB alignment was processed as follows. First, all the character sequences aligned to a single German stem were extracted. The sequences were split by word boundaries. Thus, a German stem could be mapped to several Hittite character sequences. Each character sequence would be assigned with the corresponding frequency of its co-occurrence with the German stem. In order to detect prefixes and suffixes, we would treat each Hittite sequence with co-occurrence frequency over 15 as a potential root. This high threshold for such a small corpus was chosen empirically to ensure that the stemmer is initialized with high quality alignments. Lower thresholds

[3] https://opennlp.apache.org

[4] http://snowballstem.org

[5] Full German word form is *Ägypten*.

allowed too many low quality alignments and a higher thresholds did not have enough alignments to initialize the training. The assumption behind this is that more frequently aligned sequences tend to be root morphemes as they co-occur with the German stem more often. In other words, any Hittite character sequence that is aligned to the same German stem more than 15 times is treated as a potential root. We split all the other aligned sequences by this potential root morpheme r and collected the associated counts $c(r, w_a)$, frequency of r aligned to the German word w_a and $c(\cdot|w_a)$, the total of all the alignments to w_a, and end up with two other sequences: prefix(es) pr and suffix(es) suf. We create a map of prefix and suffix co-occurrences with the initial l_{first} and final l_{last} letters of the root and save the corresponding frequencies $c(pr, l_{first})$ and $c(suf, l_{last})$. Thus, we can define five initialization scores S:

$$S(r) = P(r|w_a) = c(r, w_a)/c(\cdot|w_a) \tag{1}$$

$$S(pr) = P(pr|l_{first}) = c(pr, l_{first})/c(\cdot|l_{first}) \tag{2}$$

$$S(suf) = P(suf|l_{last}) = c(suf, l_{last})/c(\cdot|l_{last}) \tag{3}$$

$$P(pr) = c(pr)|c(\cdot) \tag{4}$$

$$P(suf) = c(suf)|c(\cdot) \tag{5}$$

The initial root score $S(r)$ (eq. 1) is the translation probability $P(r|w_a)$ of a Hittite character sequence r and aligned German stem w_a. There are four affix scores: conditional probabilities of a prefix and a suffix occurring with the first and the last letter of a root respectively (eq. 2, 3) and the overall probabilities of observing a certain affix (eq. 4 and 5) in the corpus. Originally, the prefix and suffix probabilities were conditioned on the root rather than on the first and the last letters but due to the data sparsity, it was not possible to collect reliable statistics. Empirical observations showed that conditioning on the first and last letter improves stemming. This can be explained by the fact that there are phonetic assimilations in Hittite such as regressive assimilation of *n* by *š* into *šš*.

The initialization scores are calculated based on the CB-alignment and are further updated in the training phase. In the training phase, the stemmer iterates over all the words in the corpus. It considers all possible segmentations of a word under the following conditions: a root cannot be shorter than two letters, a prefix cannot be longer than fives letter and a suffix cannot be longer than five letters. Words are allowed not to have suffixes or prefixes but any word must have a root. This might seem inefficient but as we are dealing with a small amount of data and Hittite words are seldom longer than six letters, the algorithm is not time consuming. If it encounters an unaligned root, $S(r)$ is set to a smoothing value 10^{-4}. $S(pr)$ and $S(suf)$ are also set to 10^{-4} if counts $c(pr, l_{first})$ and $c(suf, l_{last})$ are 0. The affix scores are updated in a straight-forward way by updating the counts with every segmented word. Updating the root scores is more complicated as in case of the unaligned root morphemes there is no $P(r|w_a)$. Nevertheless, the aligned roots provide important clues for segmentation and should not be abandoned. Thus, each time a root is assigned by the stemmer, its score is increased by 10%. We empirically tried various increase values but 10% delivered optimal results for POS tagging. Nevertheless, we recommend future work to look into ways of learning the increase value from the data. Though this method loses its probability-like elegance, it forces the stemmer to choose aligned roots over unaligned roots unless the unaligned roots were assigned frequently enough. Thus, the overall score assigned by the stemmer is:

$$S = S(r) * S(pr) * S(suf) * P(pr) * P(suf) \tag{6}$$

3.3 Annotation Projection

The core element of the annotation projection module is the word alignment. The word alignment is created automatically with GIZA++. As we have a limited amount of data and are only interested in one-to-one word alignments and lexical translation probabilities, we used the IBM Model-2 to produce word alignments.

The parallel German translations were tagged with OpenNLP POS Tagger using the German model that was provided with the tagger.[6] It is worth mentioning that the performance of the POS Tagger was not affected by the fact that the source Hittite texts do not have sentence marking. The

[6] http://opennlp.sourceforge.net/models-1.5/de-pos-maxent.bin

parallel translation was done for each Hittite colon and followed modern conventions of text segmentation. Thus, though the sentence segmentation is not available in Hittite, they were introduced in the translation for the purpose of readability. Furthermore, despite the fact that the source Hittite texts did not have any punctuation, their German translations follow the modern punctuation rules.

As we were primarily interested in one-to-one word alignment, we had to eliminate all the German words and symbols that cannot be aligned to a Hittite word before applying GIZA++ to the parallel data. First of all, it involved deleting all the punctuation from the German texts. As the Hittite language does not have any articles, we also eliminated all the German words that were assigned a coarse POS tag "DET". The Hittite texts were stemmed as described in Section 3.2. As the approach cannot differentiate between inflectional and derivational morphemes, we kept the Hittite root and eliminated all the affixes.[7]

Training a POS tagger demands unambiguous POS annotation of the training data, therefore, we had to resolve one-to-many alignments. For this purpose, assuming that f is a source German word and e is the aligned Hittite word, the lexical translation probabilities $P(f|e)$ and $P(e|f)$ were consulted and the alignment with the higher overall probability $P(f|e) * P(e|f)$ was preferred.

3.4 POS Tagging

In order to train a POS tagger we used the annotation projections from German into Hittite. Annotation projection creates rather noisy data and can be unreliable in cases when the word alignment quality is low. Some related work suggests to only use projections based on high confidence alignment to train a tagger. Unfortunately, this approach would not be applicable to our data as the Hittite corpus is relatively small and further reducing the amount of training data would have a negative affect on the tagger's performance.

Also, not all the Hittite sentences were fully annotated. This is not surprising as GIZA++ allows null alignments. A null alignment is not necessarily an error as sometimes there is no corresponding word in the translation (e.g. Hittite determiners described in Section 2.1). Therefore, we had to eliminate all the Hittite sentences with partial POS annotations which are 30% of all the sentences. Alternatively, it was possible to introduce dummy tags but this would introduce additional noise in already noisy projected data. The amount of fully annotated sentences is sufficient for training a POS tagger and, thus, no dummy tags are needed. Finally, we trained OpenNLP POS Tagger on 11,704 Hittite colons.[8]

stemming	POS Accuracy
None (majority class)	25.4%
None (projection)	39.4%
Hittite only	65.7%
German only	65.1%
Hittite+German	**69.1%**

Table 2: Tagging accuracy of POS taggers trained on annotation projection

4 Evaluation

We evaluated the tagger on the data described in Section 2.3. The taggers' performance was measured as tagging accuracy, a conventional measure that counts the percentage of correctly tagged tokens. The evaluation was done in three set-ups which tested the effect of the data sparsity reduction through stemming on the tagging accuracy. The most straightforward baseline was to tag all the words with the majority class NOUN. This baseline reached only 25.4% tagging accuracy. To create a more elaborated baseline, GIZA++ was directly applied to the parallel data and the data sparsity reduction step was fully omitted. The POS tagger trained on the resulting annotation projection managed to reach 39.4% of accuracy. The low tagging accuracy can be easily explained by the low quality of the word alignment. The performance of statistical word alignment applied to a small parallel corpus of two highly inflectional languages will inevitably be harmed by data sparsity. The data sparsity in the corpus of Hittite texts is very high: For instance, only 1% of all the trigrams and 0.02% of 5-grams in the corpus occur more than five times. Thus, the baseline results confirm that data sparsity is the major problem for distantly supervised POS tagging of Hittite.

As it has been previously discussed, the major

[7] The usage of affixes as additional features for training a POS tagger is possible and at the moment remains in the scope of future work.

[8] The average "sentence" (colon) is quite short (often less than six words), which explains the relatively high number of colons, compared to the overall number of tokens in DPHT.

source of the data sparsity in Hittite are the rich inflectional paradigms of Hittite words. In Section 3.2, we propose our CB-based method for stemming of Hittite that reduces the variety of Hittite word forms to the associated stem. Currently, there is no evaluation data available to test the quality of the Hittite stemmer so its usefulness can only be evaluated indirectly by examining the results of POS tagging.

Thus, in the second experimental setup, the Hittite texts were stemmed and then aligned to non-stemmed German texts. The POS tagger trained on the resulting projections showed a large 26,3% improvement over the non-stemming baseline (Table 2). The stems were, however, used only for word alignment and the POS tagger was trained and tested on the original word forms. Similarly, when the non-stemmed Hittite texts were aligned to stemmed German texts, the POS tagger showed a slightly minor improvement of 25,7% over the baseline. The fact that the Hittite stemming leads to better results is actually consistent with the fact that Hittite is morphosyntactically richer than German and, thus, has greater impact on the data sparsity. Finally, we stemmed German and Hittite parallel texts and trained the POS tagger on the annotation projections. The improvement over the baseline is almost 30% and almost 4% over the setup with only Hittite stemming.

All in all, the evaluation results show that our stemming approach to data sparsity reduction improves tagging accuracy by a large margin. While both German and Hittite stemming had a positive effect on the performance of the POS tagger, the best results were achieved through stemming of both Hittite and German translations which lead to the 30% improvement of tagging accuracy over the non-stemming baseline.

5 Related Literature

Despite the fact that low resource and historical languages have been steadily attracting attention of NLP researchers, hardly any NLP methods have been applied to the cuneiform languages. So far, most works have focused on resource building. For example, the Cuneiform Digital Library Initiative (CDLI)[9] is a large project that aims to digitize cuneiform resources. CDLI maps images of original clay tablets with transliterated texts and their translations. CDLI also constructs digitized machine-readable dictionaries for cuneiform languages. The majority of CDLI data are in Sumerian or Akkadian.

A related project that builds on the CDLI data is the Open Richly Annotated Cuneiform Corpus (ORACC).[10] ORACC includes corpora building projects that cover a variety of cuneiform resources. ORACC corpora have varying levels of annotation though most of the corpora are comprised of transliterated texts aligned with their translations. The transliterated words are annotated with a normalized form and a POS tag. However, ORACC does not contain annotated Hittite texts that could be used for training a POS tagger.

While Sumerian and Akkadian are the best researched cuneiform languages, there are also several notable resources in Hittite. Various resources and tools are provided by the Hittitology Portal Mainz (HPM), including the data that were used in this research (see Section 2.2). An important lexicographic resource is Chicago Hittite Dictionary.[11] Unfortunately, as the available digital version covers words for only five initials, we could not use it for our purpose. Daxenberger et al. (2017) describe a method to enable semantic search in translations on the DPHT. Giusfredi (2014) gives a comprehensive overview of further digital resources for Hittite. Despite the availability of digitized resources, there is hardly any NLP research on cuneiform languages other than corpus building. A reason is that many state-of-the-art NLP methods use supervised classifiers such as POS taggers, syntactic parsers etc. but the available digital resources for cuneiform do not provide enough annotated data to train a supervised classifier.

This holds for most historical languages. The only exception are the ancestors of modern world languages (e.g. Latin, historical Germanic dialects). For example, several diachronic annotated corpora have been recently released for historical varieties of modern Germanic languages. The Penn Parsed Corpora of Historical English (PPCHE)[12] covers all the historical stages of English and PPCHE's sister projects on PTB-style annotation of other historical Germanic languages, e.g. Icelandic (Rögnvaldsson et al., 2012) or Early

[9] http://cdli.ucla.edu

[10] http://oracc.museum.upenn.edu

[11] https://hittitedictionary.uchicago.edu/page/chicago-hittite-dictionary

[12] http://www.ling.upenn.edu/hist-corpora

New High German.[13]

Because of the lack of training data, historical NLP frequently uses unsupervised or distantly supervised methods. For example, annotation projection has been successfully applied to a wide variety of low-resource and historical data. Agić et al. (2016) used multilingual annotation projections to train POS taggers for 30 languages. Sukhareva and Chiarcos (2016) trained a neural network on multilingual annotation projections to create rich POS annotations for Middle Low German. Das and Petrov (2011) presented a graph-based approach where high confidence annotations are projected from the target into the source texts and are further propagated within a bilingual co-occurrence graph. They build vertices of the graph by computing trigram cooccurrence using PMI. The drawback of the approach is that it demands a large amount of parallel data which is not available for Hittite. It is not possible to utilize any of these approaches for the task presented in this study because the data sparsity of Hittite texts does not allow this: Only 1% of all the trigrams in Hittite texts occur more than 5 times. Rogati et al. (2003) uses word-based alignment to train an unsupervised Arabic stemmer. It utilizes a small parallel corpus and guesses root morphemes and and affixes by finding common substrings in Arabic words that are aligned to the same English word. This approach inspired our character-based method for Hittite stemming.

6 Conclusion

This paper describes a distantly supervised POS tagging method for Hittite. The proposed method uses a small parallel corpus of Hittite texts and its German translations as a basis for annotation projection. The annotation projections are used as training data for a POS tagger. The small amount of parallel data and developed inflectional morphology of both Hittite and German inevitably lead to data sparsity that had a drastic impact on the quality of the word alignment and, consequently, on the tagging accuracy. In order to reduce the data sparsity, we proposed an unsupervised method for Hittite stemming. The method is based on character-based alignment from which it learns morphological segmentation of Hittite words. Reduction of data sparsity using stemming had a large impact on the tagging accuracy, improving it by 30%.

To our knowledge, this is the first attempt of statistical morphosyntactic annotation of a cuneiform language. We presented a POS tagger for Hittite trained on annotation projection from German translations. We also created an unsupervised character-based stemmer for Hittite. Additionally, we annotated diachronic Hittite text fragments for evaluation. While this approach can be easily portable to other low-resource languages irrespective of the script, cuneiform Latin transcription has features that are not found in conventional phonetic writing. For example, Sumerograms and Akkadograms are transliterated based on their cuneiform sign but the actual pronunciation can differ, additionally, they are frequently followed by phonetic complements that would remind the reader of the correct Hittite word. For example, Sumerian `ŠU` "*Hand*" is disambiguated by a phonetic complement `-it` and is written as `ŠU-it` but is pronounced as *keššarit*.

Tagger, stemmer and evaluation data are freely available.[14] We are confident that our approach can be transferred to other cuneiform and low-resource languages. Though Hittite is an inflectional language, the method of data sparsity reduction and annotation projection is very likely to yield similar if not better results on agglutinating languages. The method is also portable to other cuneiform languages. Applying this method to the agglutinating Sumerian language is in the scope of the future work.

Acknowledgments

The first and third author were supported by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 01UG1416B (CEDIFOR).

References

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Alonso Martínez, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics* 4:301–312.

Paul Bennett, Martin Durrell, Silke Scheible, and Richard J Whitt. 2010. Annotating a historical corpus of German: A case study. In *Proceedings*

[13] `http://enhgcorpus.wikispaces.com`

[14] `https://github.com/UKPLab/latech-clfl2017-hittitenlppipeline`

of LREC 2010 Workshop on Language Resource and Language Technology: Standards - state of the art, emerging needs, and future developments. Paris, France, pages 64–68.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. Tiger: Linguistic interpretation of a German corpus. *Research on Language and Computation* 2(4):597–620.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Portland, OR, USA, pages 600–609.

Johannes Daxenberger, Susanne Görke, Darjush Siahdohoni, Iryna Gurevych, and Doris Prechel. 2017. Semantische Suche in ausgestorbenen Sprachen: eine Fallstudie für das Hethitische. In *Proceedings of the DHd 2017*. Bern, Switzerland, pages 196–200.

Richard Eckart de Castilho and Iryna Gurevych. 2014. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*. Dublin, Ireland, pages 1–11.

Federico Giusfredi. 2014. Web resources for hittitology. *BiOr* 71:358–361.

Graham Neubig, Taro Watanabe, Shinsuke Mori, and Tatsuya Kawahara. 2012. Machine translation without words through substring alignment. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Jeju Island, Korea, pages 165–174.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*. Istanbul, Turkey.

Michael Piotrowski. 2012. Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies* 5(2):1–157.

Monica Rogati, Scott McCarley, and Yiming Yang. 2003. Unsupervised learning of Arabic stemming using a parallel corpus. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Sapporo, Japan, pages 391–398.

Eirı́kur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurdsson, and Joel Wallenberg. 2012. The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of the Eight International Conference on Language Resources and Evaluation*. Istanbul, Turkey.

Maria Sukhareva and Christian Chiarcos. 2014. Diachronic proximity vs. data sparsity in cross-lingual parser projection. a case study on Germanic. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*. Dublin, Ireland, pages 11–20.

Maria Sukhareva and Christian Chiarcos. 2016. Combining ontologies and neural networks for analyzing historical language varieties. a case study in middle low German. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*. Portoroz, Slovenia, pages 1471–1480.

Jörg Tiedemann and Zeljko Agic. 2016. Synthetic treebanking for cross-lingual dependency parsing. *J. Artif. Intell. Res.(JAIR)* 55:209–248.

Theo van den Hout. 2011. *The Elements of Hittite*. Cambridge University Press.