Machine Translation and Automated Analysis of the Sumerian Language

Émilie Pagé-Perron[†], Maria Sukhareva[‡], Ilya Khait[¶], Christian Chiarcos[‡],

[†] University of Toronto
e.page.perron@mail.utoronto.ca

[‡] University of Frankfurt
sukhareva@em.uni-frankfurt.de
chiarcos@em.uni-frankfurt.de

[¶] University of Leipzig
ges12bry@studserv.uni-leipzig.de

Abstract

This paper presents a newly funded international project for machine translation and automated analysis of ancient cuneiform[1] languages where NLP specialists and Assyriologists collaborate to create an information retrieval system for Sumerian.[2]

This research is conceived in response to the need to translate large numbers of administrative texts that are only available in transcription, in order to make them accessible to a wider audience. The methodology includes creation of a specialized NLP pipeline and also the use of linguistic linked open data to increase access to the results.

1 Context

The project Machine Translation and Automated Analysis of Cuneiform Languages (MTAAC)[3] focuses on the application of NLP methods to Sumerian, a Mesopotamian language spoken in the 3rd millennium B.C. Assyriology, the study of ancient Mesopotamia, has benefited from early developments in NLP in the form of projects which digitally compile large amounts of transcriptions and metadata, using basic rule- and dictionary-based methodologies.[4] However, the orthographic, morphological and syntactic complexities of the Mesopotamian cuneiform languages have hindered further development of automated treatment of the texts. Additionally, digital projects do not necessarily use the same standards and encoding schemes across the board, and this, coupled with closed or partial access to some projects' data, limits larger scale investigation of machine-assisted text processing.

The history and society of ancient Mesopotamia are mostly known to the general public through works that draw on myths and royal inscriptions as primary sources, texts which are mostly translated and readily available. Among these works the Sumerian texts and their translations form a perfect testbed for distantly supervised NLP methods such as annotation projection and cross-lingual tool adaptation. However, the aforementioned translated texts make up only around 10% of the total amount of transcribed Sumerian data. The majority of the Sumerian texts are administrative

[1] The Cuneiform script was invented in Ancient Iraq more than 5000 years ago. Signs were drawn, and later impressed, onto a tablet-shaped fresh lump of clay using a reed stylus. This script was in use for 4000 years to record texts in different languages such as Sumerian, Akkadian and Elamite. See figure 1 in section 1b for an example.

[2] We would like to thank the reviewers, and Robert K. Englund and Heather D. Baker, for their insightful comments and suggestions.

[3] The project is generously funded by the Deutsche Forschungsgemeinschaft, the Social Sciences and Humanities Research Council, and the National Endowment for the Humanities through the T-AP Digging into Data Challenge. See the project website at https://cdli-gh.github.io/mtaac.

[4] Among others, the Cuneiform Digital Library initiative (CDLI) http://cdli.ucla.edu/ and the Open Richly Annotated Cuneiform Corpus (ORACC) http://oracc.museum.upenn.edu/ are two examples of such endeavors.

Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. Proceedings, pages 10–16, Vancouver, BC, August 4, 2017.

and legal in nature. The manual annotation and translation of these texts is hardly possible, owing to the large volume of the data and the need for an extremely rare expertise in Mesopotamian languages. However, having a parallel corpus, the solution to automatic processing of these texts lies in using machine translation (MT) techniques: Sumerian texts can be automatically translated and information extraction methods can be applied to the resulting translations.

In this paper we present a newly funded international project that will apply state-of-the-art NLP methods to Sumerian texts. We seek to create a pipeline for cuneiform languages with three major components: NLP processing, machine translation, and information extraction. The NLP tools for Sumerian created in the framework of the project will also be applicable to other cuneiform languages. The resource interoperability will be achieved through linking the annotation with linguistic linked open data ontologies (LLOD).

2 Data

The data for this project takes the form of unannotated raw transliterations of almost 68,000 Sumerian texts of the Ur III period (21st century B.C.) comprising 1.5 million transliteration lines. Around 1600 of these texts have also been translated. Each text entry is augmented with a set of metadata which describes the medium of the text, its context, and some elements of internal analysis. These texts are restricted in style and topic, and include a large proportion of numero-metrological elements. They are also repetitive, brief, and formulaic. As the inscribed medium comes in varied sizes and shapes, structural elements in the transliterations indicate on which surface of the artifact the text appears. Figure 1 shows an example of an ASCII transliteration and translation of a cuneiform text, accompanied with a picture of the obverse and reverse of the artifact. [5]

3 NLP Pipeline for Sumerian

State-of-the-art statistical NLP widely uses supervised classifiers to produce automatic linguistic annotation. Although some Sumerian and Akkadian corpora have been annotated through the ORACC platform in the form of various sub-projects,[6] manual annotation of large enough training sets to train a supervised classifier is not possible as it demands a rare expertise and is time-consuming. We thus propose a pipeline that uses distantly supervised methods (e.g. annotation projection) to create automatic linguistic annotation of Sumerian. Figure 2 shows the workflow of the NLP module. The majority of the data at hand comprises untranslated Sumerian texts. The distantly supervised methods will be applied to Sumerian texts and their English translations. The core of the pipeline is the annotation projection module that will produce morphosyntactically and syntactically annotated training data for supervised NLP tools. This section will further discuss in detail each module of the NLP pipeline.

3.1 Data Preprocessing

After verifying the uniformity in the standardization of the texts, we will convert the data to a machine readable format and sign readings will be verified against our digital syllabary. Transliterations and translation of our gold standard will be tokenized, lemmatized, and morphologically analyzed. The error rate of the corpus transliterations will be calculated against the curated gold standard.

3.2 Morphological analysis

Our morphological analyzer will be partly based on existing tools such as Tablan et al. (2006)'s rule-based morphology and Liu et al. (2015)'s algorithm to identify named entities. We will design a custom parser for numero-metrological content for the occasion. Since Sumerian affixes are ambiguous, we will build on previous work on the disambiguation of morphologically rich languages, such as Sak et al. (2007)'s neural methods for Turkish and Rios and Mamani (2014)'s conditional random fields used to disambiguate Quechua morphology. Morphological tags assigned following rule-based algorithms will be reranked using different machine learning (ML) approaches. The disambiguated morphology will be used for syntactic parsing, MT, and information extraction. We plan to develop a lemmatizer that will exploit a high-coverage dictionary. The available off-the-shelf lemmatizer for Sumerian[7] was

[5] Cuneiform text of the Ur III period from the settlement of Garshana, Mesopotamia (Owen, 2011, no. 851) and its transliteration as stored in the Cuneiform Digital Library Initiative (CDLI) database http://cdli.ucla.edu/P322539 (picture reproduced here with the kind permission of David I. Owen)

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

[7] http://oracc.museum.upenn.edu/doc/help/languages/sumerian/sumerianprimer/

(1) P322539 = CUSAS 03, 0851.

tablet.

obverse.

1. `1(disz) kusz udu niga`
   1 hide, grain-fed sheep;

2. `1(disz) kusz masz2 niga`
   1 hide, grain-fed goat;

3. `kusz udu sa2-du11`
   sheep hides, regular offerings,

4. `ki {d}iszkur-illat-ta`
   from Adda-illat,

reverse.

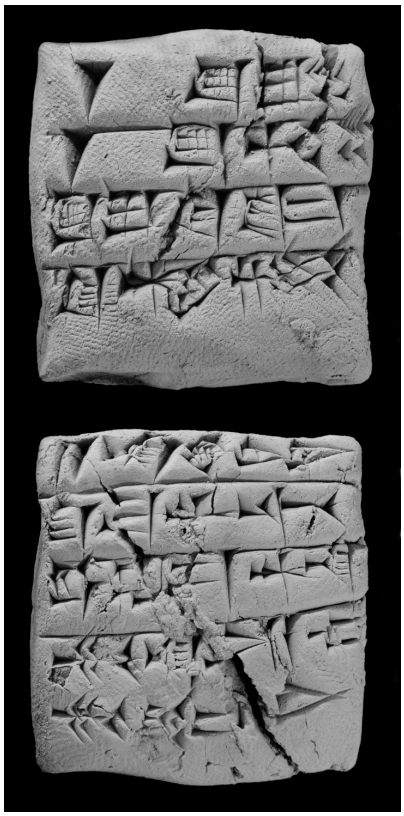
1. `a-na-ah-i3-li2`
   Anah-ili;

2. `szu ba-an-ti`
   did receive.

3. `iti ezem-an-na`
   Month: An-festival,

4. `mu na-ru2-a-mah mu-ne-du3`
   Year: He erected the great stele for them.

(a) ASCII transliteration and English translation



(b) Example of a Sumerian source text

Figure 1: Artifact and its digitization

applied to our corpus during the preparation of this project and it was revealed that its coverage and accuracy are not sufficient for our needs since headwords are assigned to tokens without taking into account the textual context, although part of this software might be reused.

3.3 POS tagging

An important part of the NLP pipeline is the distantly supervised POS Tagging. As the corpus is currently unannotated, a supervised approach to POS tagging would not be applicable as it demands annotated training data. The creation of such training data through manual POS annotation of the data would demand an extremely rare expertise and is a time-consuming process. Therefore, we have to turn our attention to distantly supervised methods.

As we are in possession of parallel English translations of Sumerian texts, an annotation projection (Tiedemann, 2014) approach would be a most suitable distantly supervised method. English texts can be tokenized, stemmed, lemmatized, POS tagged and parsed by off-the-shelf freely available NLP tools. Using an off-the-shelf word-alignment tool Giza++ (Och and Ney, 2003), we can produce word alignment between English and the Sumerian texts. After we automatically tag English parallel texts, the assigned POS will be projected onto the aligned Sumerian words. The general assumption behind the annotation projection based on the word alignment is that translated words are likely to have the same POS as the source words. It is quite clear that this is a very bold assumption and there are a number of exceptions. Thus, both manual and automatic POS correction will be needed. However, the distantly supervised solution is temporary as there are parallel efforts to annotate the texts manually to produce training data for a supervised classifier.

3.4 Syntactic parsing

In order to facilitate MT and information extraction from our source texts, we will syntactically parse the corpus. In a similar manner to POS tags, dependency labels can be projected into Sumerian texts. Annotation projections of both POS tags and dependency labels need to be manually corrected. Using an adapted scheme for Sumerian, we will annotate a gold standard composed of a

total of 10,000 sentences with dependencies and POS tags to train a supervised dependency parser and POS tagger. The rest of the data will be tagged and parsed automatically. The quality of the dependency parses will be estimated by labeled and unlabeled attachment score (UAS and LAS), and different parsing toolkits will be evaluated (Chen and Manning 2014, Nivre 2003, etc.).

4 Machine Translation

As MT for cuneiform languages is a novel task and there is no prior research, we will have to experiment with several approaches in order to establish the one most suitable for these languages. The standard phrase-based translation will form a good baseline.

Currently, there are over 1600 parallel Sumerian and English texts which are aligned sentence-wise. The baseline will be created by the Moses SMT toolkit (Koehn et al., 2007). It will be trained on these parallel texts and applied to the rest of the data to create automatic translations.

Nevertheless, due to the spelling variations and morphological richness of the language, data sparsity is inevitable. Thus, the baseline will be compared with a character-based MT system based on Phrasal ITG Aligner (Pialign) (Neubig et al., 2012) but tailored towards cuneiform data. Pialign uses synchronous context-free grammars and substring prior probabilities to produce many-to-many character alignment; it can thus efficiently capture mid-distance dependencies, as required for dealing with rich morphology and ideosyllabic writing systems without explicit word separators (e.g., Japanese).

In addition to this state-of-the-art SMT system, we will also apply innovative neural techniques to the translation of Sumerian cuneiform text. Neural Machine Translation (NMT) (Bahdanau et al., 2014) has been applied to various language pairs in the past few years, with successful applications for translating structurally different languages: Eriguchi et al. (2016) applied an attention-based neural network on Japanese and English that we will take as our point of departure, as the writing system of Japanese is structurally similar to cuneiform (using both ideographic and syllabic components), and they demonstrated that their approach is capable of generalizing over smaller amounts of training data than normally required by NMT systems. Following their syntax-based extension of the traditional sequence-based encoder-decoder approach, we will integrate syntactic dependency annotation.

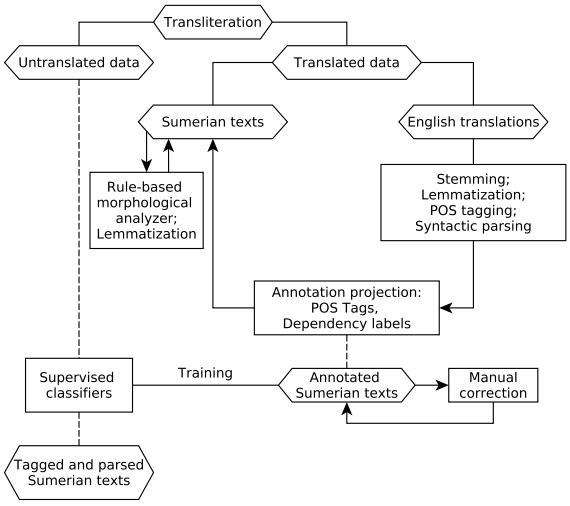


Figure 2: NLP pipeline for Sumerian

5 Information Extraction

In this project, we intend to go beyond automatic (morpho)syntactic annotation and collocation analysis to create an information extraction system for the Sumerian language. This system will aim to identify concepts and relations in the text, and the results will be available in human-readable format to integrate into the interface, alongside their labels and definitions.

The main objective of this step is to prepare data for prosopographical[8] research into the Ur III historical period. The Sumerian texts have an abundance of individuals' names but tracing them throughout various texts is not a trivial task. For example, the proper name ^{d}iškur-illat can be found 1212 times in 1092 texts. Additionnaly, it can be used as a toponym, a road name, or as a personal name that occurs in at least five different cities. Thus, it is impossible to tell at first sight whether these occurrences represent a single individual or five or more different people. In order to mitigate this problem, we will apply automatic collocation-based classification of proper names into specific entity categories (people, places, gods, etc.). A prosopographical study of extracted names will include a profiling of the individuals, which will entail identifying an individual's activities, titles,

[8] Prosopography is the study of past individuals and their relationships through sparse sources that give clues concerning their activities as groups.

properties and other pertinent information. We will then build a graph representation of the social connections of an individual. These structured features will be used for the disambiguation of individuals' profiles.

5.1 Research in Social History

Until now, prosopographical studies in Sumerian have focused on specific private or institutional archives (e.g. Dahl 2007) or have been based on a specific topic of inquiry but restricted to a region and period: their scope has always been limited. This selective nature of prosopographical study in Sumerian is largely due to the fact that the creation of the individuals' profiles involves a significant amount of manual work. The automatic translations and annotations produced by our NLP pipeline will enable us to automatically extract descriptions of individuals which will in turn enable us to perform large-scale social inquiries using a full prosopographical network based on the corpus at hand.

The main question that will be researched in this context is looking at is the dynamics of social mobility in the Ur III period. Administrative texts of the Ur III period are often dated with a ruler name, year name and month, sometimes days. This makes it possible to trace individuals through time when they appear in the archives. With our social network graph in place, it will be possible to identify clues to social mobility such as displacement, change in role, responsibility level, ego network variations, changes in property status, name and title. We will also take into account the influence of political and environmental changes through time. Such unprecedented large-scale prosopographical study can reveal important social and political trends that will shed light on the processes that enable or limit social mobility at that period.

6 Direct Applications

6.1 Linked Open Data

As an integral part to the project, all the manual and automatic linguistic annotations will be mapped to the *Ontologies of Linguistic Annotation* (Chiarcos, 2008), a reference terminology of Linguistic Linked Open Data (LLOD) which will ensure the interoperability of our annotation scheme and greatly increase the reusability of our data. As part of our prosopographical project, we will map our results onto the Standards for Networking Ancient Prosopographies (SNAP:DRGN)[9] that we will augment with a custom extension if needed. Pleiades[10] and Periodo[11] will be considered for mapping places and periods respectively. A machine readable interface will be developed to share all these prepared linked data.

6.2 Interface

One way of interfacing with the data generated will be through a new web facade designed with known audiences in mind, but also applying principles of universal design in order to increase the accessibility of the data and interface to a wider public of knowledge drawn from cuneiform sources. Translations will be easily retrievable and researchers will benefit from an advanced search engine. Concepts and entities present in the texts and metadata will be interlinked to permit navigation through the texts. Other visualization tools will be available, from dependency visualization to automated plotting of network analysis graphs, as well as traditional graphs to display statistics concerning a chosen group of texts. Data will also be available for download in full and in part, in different open formats.

6.3 Future Applications

Following this project, we expect to extend the scope of our pipeline to process other genres and periods of Sumerian texts and then work with the Akkadian language and other cuneiform languages. We expect that parts of our work can serve as test cases for other languages such as Basque and Turkish that share agglutinative and split-ergative characteristics and also logo-syllabic languages such as Japanese.

Having these tools available will foster future research into Ur III texts since they will be more accessible, including to machines. There is already a renewed interest in the study of Ur III texts because Assyriologists are starting to employ statistical methods to study larger groups of texts, so our project will also open doors for these interested scholars.

Because we will be using LLOD, new studies across languages, including Sumerian, will become possible. This will enrich the pool of varia-

[9] https://snapdrgn.net/
[10] https://pleiades.stoa.org
[11] http://perio.do/

tions of language morphology, especially because Sumerian is an isolate.

7 Challenges and Risks

The automatic processing of Sumerian texts is not a trivial task. Among others, we will face most of the traditional challenges of historical corpora. First of all, data sparsity will be inevitable. Sumerian is an agglutinative language with productive affixation which leads to an extremely high number of word forms, but we will significantly reduce data sparsity by means of lemmatization as explained in section 3.2. Regional variations can also increase the data sparsity: words can have different meanings, different readings, and different spellings depending on contextual factors such as the type of text, the period, the archive, and the region. Code-switching and foreign words pose an additional challenge for the morphological analyzer, but the texts have been marked with a structural language switch for ease of processing since Sumerian texts can be sprinkled with Akkadian words, for example verbs and personal names.

Other difficulties arise in the annotation projection and machine translation from the fact that the Sumerian language does not have any modern descendants. This is particularly important for the annotation projection, as previous studies have shown that diachronic relatedness is an important factor that affects the quality of annotation projection (Sukhareva and Chiarcos, 2014). Thus, we plan to conduct our pilot experiments on modern languages such as Turkish and Basque that are grammatically similar to Sumerian (agglutinative, split-ergative) to guarantee the scalability of our implementation, but more importantly to be able to conduct experiments in parallel with the morphological and syntactic annotation of the Sumerian texts.

8 Conclusion

Even though some basic NLP methods are already being employed in cuneiform studies, the use of modern computer science methods is still in its infancy and such powerful methods as ML and statistics are yet to be properly introduced into the field. Our MT and information extraction project, based on a practical research need in the Humanities, will contribute a methodology, its implementation, and a body of translated and analyzed texts. It will also assist in the processing of a host of related datasets, as well as setting an example for ML and MT in the Humanities. Moreover, it will facilitate studies of grammar and semantics of a language that is still not fully understood. The project will provide a unified access to a highly representative corpus of early writing, and will foster an unprecedented scholarly cooperation among researchers in a variety of disciplines. We think this is a unique opportunity to make a leap forward in natural language processing for ancient languages that will at the same time open up to a networked public the heritage of ancient civilizations.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473. http://arxiv.org/abs/1409.0473.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *EMNLP*. pages 740–750.

Christian Chiarcos. 2008. An ontology of linguistic annotations. *LDV Forum* 23(1):1–16.

Jacob L Dahl. 2007. *The ruling family of Ur III Umma: a prosopographical analysis of an elite family in Southern Iraq 4000 years ago*. Nederlands Instituut voor het Nabije Oosten.

Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2016. Tree-to-sequence attentional neural machine translation. *CoRR* abs/1603.06075. http://arxiv.org/abs/1603.06075.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '07, pages 177–180. http://dl.acm.org/citation.cfm?id=1557769.1557821.

Yudong Liu, Clinton Burkhart, James Hearne, and Liang Luo. 2015. Enhancing sumerian lemmatization by unsupervised named-entity recognition. In *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies (NAACL HLT 2015)*.

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. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '12, pages 165–174. http://dl.acm.org/citation.cfm?id=2390524.2390548.

Joakim Nivre. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*. pages 149–160.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19–51.

David I. Owen. 2011. *Garsana studies*. CDL Press.

Annette Rios and Richard Castro Mamani. 2014. Morphological disambiguation and text normalization for southern quechua varieties. *COLING 2014* page 39.

Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2007. Morphological disambiguation of turkish text with perceptron algorithm. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer Berlin Heidelberg, pages 107–118.

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*. Association for Computational Linguistics and Dublin City University, Dublin, Ireland, pages 11–20. http://www.aclweb.org/anthology/W14-5302.

Valentin Tablan, Wim Peters, Diana Maynard, Hamish Cunningham, and K Bontcheva. 2006. Creating tools for morphological analysis of sumerian. In *5th Language Resources and Evaluation Conference (LREC), Genoa, Italy*.

Jörg Tiedemann. 2014. Rediscovering annotation projection for cross-lingual parser induction. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pages 1854–1864. http://www.aclweb.org/anthology/C14-1175.