When Hieroglyphs Meet Technology: A Linguistic Journey through Ancient Egypt Using Natural Language Processing

Ricardo Muñoz Sánchez
Språkbanken Text, University of Gothenburg
ricardo.munoz.sanchez@svenska.gu.se

Abstract

Knowing our past can help us better understand our future. The explosive development of NLP in these past few decades has allowed us to study ancient languages and cultures in ways that we couldn't have done in the past. However, not all languages have received the same level of attention. Despite its popularity in pop culture, the languages spoken in Ancient Egypt have been somewhat overlooked in terms of NLP research. In this survey paper we give an overview of how NLP has been used to study different variations of the Ancient Egyptian languages. This not only includes Old, Middle, and Late Egyptian but also Demotic and Coptic. We begin by giving a short introduction to these languages and their writing systems, before talking about the corpora and lexical resources that are available digitally. We then show the different NLP tasks that have been tackled for different variations of Ancient Egyptian, as well as the approaches that have been used. We hope that our work can stoke interest in the study of these languages within the NLP community.

Keywords: Ancient Egypt, Ancient Languages, Coptic, Demotic, Historic Languages, Literature Review, Low-Resource Languages

1. Introduction

Ancient Egyptian culture has been called one of the cradles of western civilization (Maisels, 1998). However, there is still much that we do not know about it. The Egyptian people left behind vast amounts of primary textual sources, which the dry weather of the desert helped preserve even if it was in a fragmentary manner. As an example of this, we can take the Oxyrhynchus papyri, a collection of over 500,000 papyri containing fragments of texts, currently housed at the University of Oxford.[1] All of these documents can give us invaluable insights into the lifestyles that these people led and the state of the world at that time. It also can provide unique insights into how technology, science and religion have evolved over time. Developing computational approaches can help us better understand the languages within these documents and how they connect to their environment, while helping preserve them for future generations.

Some issues are quick to appear when attempting to use NLP for Ancient Egyptian. First and foremost is that there are no longer any native speakers left. This means that we cannot know how the language was pronounced[2] or clarify any doubts we may have about the documents. As for making linguistic annotations and translations, it will often take much longer than for living languages (Polis et al., 2015). Furthermore, some of the subtleties of the text might be missed due to lack of the relevant sociocultural context.

Another major issue is that the Ancient Egyptian language was used for over 3,000 years. The vast expanse of the Ancient Egyptian empire and the lack of quick and inexpensive media of transportation lead to major variations in the language (Bard, 2005). More details on the language and on these variations will be discussed in section 2.

Finally, even though a lot of documents survived, most of them are at least partly damaged due to weather conditions, human intervention or just the passage of time. This means that, even if we can extract the whole meaning of the sentence, some nuances or regional variations can be lost to history.

All of these issues mean that the different variations of Ancient Egyptian are considered low-resource languages (Zeldes and Schroeder, 2016; Nederhof and Rahman, 2015). This means that most of the cutting-edge strategies such as transformers (Vaswani et al., 2017) cannot be used for these languages, as those often require vast amounts of data.[3]

For this literature review, we made a survey of the Natural Language Processing (NLP) techniques that have been used recently to study the Ancient Egyptian language. This includes not only the actual implementations, but also some of the difficulties they faced, how they were able to overcome them and some of the implications of their works.

[1] https://www.ees.ac.uk/papyri

[2] Despite this, at least one paper has attempted to do automated pronunciation mining for several dead languages, including Ancient Egyptian and Coptic (Lee et al., 2020).

[3] It should be noted that this is not necessarily the case for Demotic, as it has parallel corpora that allow it to be used for multilingual approaches, see Choudhary and O'riordan (2023) or Khakhmovich et al. (2020) for examples of this.

LT4HALA 2024@LREC-COLING-2024, pages 156–169
25 May, 2024. © 2024 ELRA Language Resources Association: CC BY-NC 4.0

We looked for papers dealing with computational approaches and Ancient Egyptian in the ACL Anthology[4], the ACM Digital Library[5], and Google Scholar[6]. We then filtered the papers that are related to NLP. More specifically, we decided not to talk about optical character recognition or the digital representation of the characters either, as we consider those to be image recognition and data representation tasks, respectively, as opposed to NLP ones. We have included all works that match our criteria until January 2024.

It is important to note that we focused mainly on Middle and Late Egyptian as barely any NLP work has focused on Old Egyptian and Demotic. We also focus on Coptic, as this language can also be considered a variation of Ancient Egyptian (Bard, 2005) and a good amount of work has been done for it.

As for the organization of the rest of this paper, we first describe the language in Section 2 and make some comments about it in order to showcase common issues that arise when working with the language. We talk about the corpora available in Section 3, including the kinds of annotations they have and the periods over which they have been updated. In Section 4 we talk about the NLP tasks that are relevant for Ancient Egyptian. Finally, we devote Section 5 to the current state of the use of NLP techniques for Coptic. Even though it still can be considered an evolution of the Ancient Egyptian language, it has a completely different writing system and we have a greater amount of well-preserved documents. As a result, the issues faced when dealing with Coptic are different than those that we face with Ancient Egyptian.

2. The Language

Nederhof and Rahman (2015) provide a good overview of the Ancient Egyptian language and its characteristics in their paper. It is the main source of the information in this section, along with the introduction to hieroglyphs given by Kamrin (2004) and the description of the language given by Bard (2005). However, most of the papers that we mention throughout this literature review also have a brief explanation of the language.

Ancient Egyptian is a language in the Afro-Asiatic family. This family includes the Semitic languages (Hebrew, Arabic, etc.). In the languages of this family the vowels are usually not written and Ancient Egyptian is no different[7]. This, coupled with the fact that there are no native speakers alive, means that we cannot really know how Ancient Egyptian sounded like. Some of the approximations we currently have are made taking into account how phonetics work in the other languages of the family, but we should not fall into the trap of considering them how the language actually sounded.

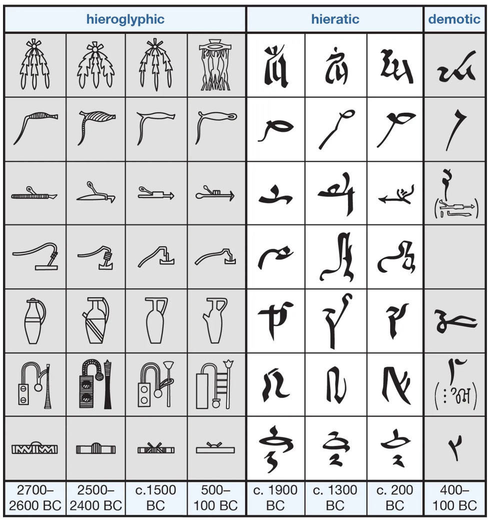

Figure 1: A table illustrating how the Ancient Egyptian scripts evolved over time. It compares seven symbols in hieroglyphic, hieratic, and demotic scripts. Taken from the Encyclopaedia Britannica website,[9] based on the same table by Möller (1919, p. 78).

The writing system was hieroglyphic, but it could also be written in hieratic, a manuscript version of hieroglyphs. An example of how these writing systems evolved over time can be seen in Figure 1. We have included more examples of how these script systems look like in Appendix A. The symbols of this writing system can be divided into logographs, phonographs, determinatives or typographical signs.

Logographs represent either whole words or ideas. That means that a single symbol can represent a complete idea, such as a river or a bird. Phonographs, on the other hand, represent sounds. Each phonograph can correspond from one to three consonants, depending on the symbol. Determinatives help clarify the meaning of the word or disambiguate between otherwise identically written words. Finally, typographical signs are used to give semantic meaning to the word or as fillers.

There are some important considerations that must be taken into account when trying to parse these symbols. Some words can be written either using logograms, just phonograms or a combina-

[4]https://aclanthology.org/

[5]https://dl.acm.org/

[6]https://scholar.google.com/

[7]With the exception of Coptic, where vowels are written.

[9]https://commons.wikimedia.org/wiki/File:Leaves_from_a_Coptic_Manuscript_MET_sf21-148-1as3.jpg (Accessed March 30, 2024)

tion of the two (like in Japanese). Also, some symbols can have more than one function and there are neither end-of-word nor end-of-sentence markers. Furthermore, scribes took into account the aesthetic value of their work, adding or removing symbols as they deemed appropriate. Along the same vein, while the language was written from top to bottom, it could be written from left to right or from right to left and the orientation of the text could be either vertical or horizontal. This means that there is no standardized way of writing the language.

The language also had important variations throughout its history. The Ancient Egypt empire lasted for around 3,000 years and is usually divided into the Old, Middle and New Kingdoms. Between these kingdoms there were periods of great unrest, which lead to big cultural changes. Because of that, the Ancient Egyptian language can be divided into these same stages, with Old and Middle Egyptian being sometimes grouped into Classical Egyptian due to their similarity. However, Late Egyptian does show important differences when compared to Middle Egyptian, both grammatical and morphological, and is often considered as a different language.

Finally, Demotic and Coptic can also be considered later stages of Ancient Egyptian, even though they do not use neither hieroglyphs nor the hieratic script any longer (Bard, 2005). They can also have bigger variations in terms of morphological and grammatical variation, as evidenced by the greater amount of usage of suffixes and the lack of repetition of phonemes in Coptic (Zeldes and Schroeder, 2016).

It is because of all these reasons that most papers just focus on one of the stages of the language instead of trying to focus on all of its history at the same time.

3. Corpora and Lexical Resources

An important first step in order to do any kind of NLP is to have corpora available. However, when studying ancient languages we have the major issue that there are no longer any native speakers to annotate sentences or documents. This in turn means that it takes much longer for them to be annotated (Polis et al., 2015). Here we present the most recent and most comprehensive corpora for the different stages of Ancient Egyptian that we mentioned in Section 2.

3.1. Middle Egyptian

While there were attempts at making corpora of annotated Middle Egyptian, it was until 2017 when Nederhof and Rahman (2015) annotated a corpus for hieratic transliteration that also included the function of each symbol. Taking into consideration that the current NLP approaches do not use the spatial relations of the script, they linearized the text. They also removed variations of symbols, considering that they would do more harm than to help training the models. The corpus currently consists of only two texts. Due to how some words tend to be often repeated throughout each text, its creators suggest to train it on one of them and test it in the other. They argue that, even though mixing both texts allows for more training data, doing so would skew the results of machine learning models and give a false sense of confidence due to data leakage. The corpus is available as part of the larger St. Andrews corpora.[10]

3.2. Late Egyptian

The Ramses project is the most ambitious project regarding Ancient Egyptian corpora, as it is an attempt to build a comprehensive annotated corpus of all available texts in Late Egyptian (c. 1350-700 BC). The project began in 2008, and a first version of their software was first made publicly available in 2013 by Polis et al. (2013). A beta of an online version was released in 2015 (Polis et al., 2015). At the time of its presentation, the corpus had already more than 1350 texts, which amount to over a million words. When the website was announced, it already had over 4000 texts and, during a presentation in 2017 (Polis and Razanajao, 2017), it was announced that the corpus was nearing 5000 texts.

An important feature of this corpus is that from its inception, it included the documents that are considered the most useful for studying the language, along with other texts considered to be relevant for linguistic analysis. The corpus's annotations focus heavily on inflections, lemmata, and spellings, but also include all of the relevant metadata for each text, along with annotations on the state of preservation of the documents (or sections of them) and on alterations or editings of the texts. It also allows the annotators to include comments or criticism on their choices, with references that justify them. Their original paper (Polis et al., 2013) also includes a small tutorial on how to use their software and a list of ways to further expand the project, one of which was including syntactic analysis of the texts.

The online version is currently available at the project website.[11] However, this is only the beta version of the website, which is only available in French and provides access to only a small portion of the corpus. Another issue is that the last update to the website was made in 2016, though Polis and Razanajao (2017) noted in 2017 that the project

[10]https://mjn.host.cs.st-andrews.ac.uk/egyptian/texts/

[11]http://ramses.ulg.ac.be/

was still alive.

3.3. Demotic

The Chicago Demotic Dictionary (Johnson, 2001) is one of the few lexica available for Demotic. It was maintained and updated from 1972 to 2012 and includes not only the words themselves, but also scans of the actual documents. The 2002 edition can be found on the project's website as a PDF document.[12]

3.4. Coptic

A comprehensive corpus of Coptic was created in 2013 and released in 2016. This corpus, called the Coptic Scriptorium (Schroeder and Zeldes, 2016), was designed to be used to study a wide variety of subjects, from linguistics to biblical studies, and consists of eleven smaller corpora. At the time of its release, it had a little less than 60 thousand manually annotated words. This corpus can be used for a wide variety of NLP tasks, most of which can be consulted at the project's website.[13] Most notably, it covers a wide variety of annotations, from tokenization (i.e. identifying the words in a document) all the way to parts-of-speech tagging and a treebank which follows the universal dependencies notation. This is an ongoing project that currently has around 850 thousand annotated words and the documents have enough metadata to tell whether these annotations were made automatically or whether they were either made or revised by humans. Their most recent release was on October 2023 and the current status of the project can be found at their blog.[14]

Several other lexicons for Coptic have been created through time. There is also the Database and Dictionary of Greek Loanwords in Coptic[15], which contain Coptic Lemmas that were adopted from Ancient Greek lemmas. The Marcion project[16] is another lexicon freely available online, with over 11 thousand head words and over 87 thousand items. Both of these lexicons were based on an already existing dictionary (Crum, 1939).

In return, both of these lexicons along with the Coptic section of the TLA were used to create both an online dictionary (Feder et al., 2018) and WordNet (Slaughter et al., 2019). Both of these have been incorporated into the Coptic Scriptorium and its other resources.

Some multilingual collections of corpora contain data in some of the variations of Ancient Egyptian. The Coptic Scriptorium corpus mentioned previously forms part of the Universal Dependencies framework (Zeldes and Abrams, 2018; de Marneffe et al., 2021), a project whose aim is to create a framework for consistent grammatical annotations across different languages. Finally, the OPUS corpora (Tiedemann, 2016) contains parallel data for translation, one of the languages included being Coptic.

3.5. Various Time Periods

The Thesaurus Linguae Aegyptiae (TLA) (Seidlmayer, 2011) was a corpus released in 2004 and was updated until 2012. It contains a wide variety of texts, ranging all the way from the Old Kingdom to the Roman times, including the oldest pyramid texts. This amounts to almost a million and a half words, containing texts in Old, Middle and Late Egyptian, Demotic, and Coptic. It is one of the few annotated Old Egyptian and Demotic corpora. The corpus only has lemmatization and morpho-syntactic annotation and most of their website, including the handbook on how to access and use the database, is in German. The corpus is freely available online.[17]

The Thot Sign List (TSL) (Polis et al., 2021) is a collection of graphemes that have been attested in hieroglyphic or hieratic texts. Its first release contains 1,203 signs, 4,842 functions, and 21,834 tokens. The TSL is freely available on the project website,[18] but a (free) account is necessary to access all of its features.

Nordhoff and Krämer (2022) created a dataset with morpheme annotation for several low-resource languages. It contains examples in Old and Late Egyptian, as well as in Coptic. However, they do not mention the corpus size for any of the languages included.

4. NLP for Middle and Late Egyptian

Rosmorduc (2015) gives a quick overview of some of the main tasks that have been tackled from the 90s to 2015. He notes that, other than some attempts in the 90s, most of the work up until recently had been geared towards creating a standard Unicode representation of hieroglyphs. The most recent updates in this regard were in 2019 and 2021 (Nederhof et al., 2019; Glass et al., 2021), when some control characters to signal some spatial properties of the characters were introduced.

[12] https://oi.uchicago.edu/research/projects/chicago-demotic-dictionary-cdd-0
[13] https://copticscriptorium.org/tools
[14] https://blog.copticscriptorium.org/
[15] https://www.geschkult.fu-berlin.de/en/e/ddglc/index.html
[16] http://marcion.sourceforge.net/dictionary/coptic.html
[17] http://aaew.bbaw.de/tla/
[18] http://thotsignlist.org

4.1. Transliteration

We currently have a very good understanding of how Ancient Egyptian script works, even going as far as having developed standardized methods of transliteration to Latin script and designed Unicode symbols for hieroglyphic script (Nederhof et al., 2019). However, most of these methods require human annotators to work on the text due to the lack of standardization in how the language was written (see section 2). This means that transliteration is still an open problem in the Ancient Egyptian machine learning field.

As mentioned in Section 3, an important issue is that annotation of Ancient Egyptian is a slow process. Because of this, any major breakthrough would mean that more manpower would be available for other tasks in Egyptology.

One of the latest approaches for transliteration is the one by Nederhof and Rahman (2017). They made a probabilistic automaton that can transliterate a text in Middle Egyptian hieratic (i.e. manuscript hieroglyphs) to its phonetic values. For this, they created the Middle Egyptian corpus mentioned in Section 3. It has annotations for the functions of each symbol so as to help the model learn. They consider that the innovation of their system is that it does more than just doing a simple transliteration, it also makes notes on semantic elements of the text. Due to the scarcity of annotated texts from that era, they compare n-gram models (with n varying from 1 to 3) and Hidden Markov Models (HMM). They were able to reach recall and precision scores of approximately 0.95 when interpolating the results from the 3-gram and HMM models. The authors mention that, even though the model used was very basic, this is an important stepping stone for transliterating documents from this era.

In a previous work, Nederhof (2009) notes that alignment could be another possible way to approach transliteration. The proposed model assumes that the signs in the text can only be either phonograms or determinatives, thus ignoring logographs and typographical signs. Moreover, it also assumes that the text can be read without skipping signs or repeating phonograms. In order to make the model more robust, it assigns a penalty to words that could break these rules. The word boundaries are then chosen as the configuration that minimizes this penalty through the use of beam search. When using a simpler text he got an accuracy of 0.98 while experimenting with variations of the model, while a more complicated text got an accuracy of 0.97. He does note, however, that the model might struggle with unseen and/or more complex texts due to things such as unusual ways that words might be written.

Rosmorduc (2009) tried another approach to transliteration. He derived a set of rules on how words are formed and created a series of transducers, that is, finite-state automatons that parse the words and use these rules to verify whether a word is valid or not. The validation set was one of the same texts that Nederhof and Rahman (2015, 2017) used for their corpus and his model achieved a precision of around 0.91. However, this was the same set from which the rules were derived. When using another text as a test set, the precision dropped to 0.82. He justifies his results by claiming that they were due to some small technical errors. Finally, he tried to use the same model on a Late Egyptian text. Even though the precision score for this test is not reported and the author notes that it is quite bad, he mentions that it is on par with what he would expect for a student that has only studied Middle Egyptian but not any of its latter variants.

A later paper by Barthélemy and Rosmorduc (2011) compares two kinds of transducers, but does not report performance scores for either of the models.

Similarly, Bédi et al. (2022) present a multimodal system for transcribing or transliterating endangered and extinct languages (depending on whether the modality is audio or text, respectively). They tested their model on Ancient Egyptian inscriptions, but do not report any quantitative results. A later paper shows how this system would work with a sample text (Bédi et al., 2022), which is also available online.[19]

Finally, Wiesenbach and Riezler (2019) use transcription and part-of-speech tagging as an intermediate step towards translation into German. They used encoders and decoders to achieve these joint tasks. Given that they do not report results for the transliteration, we will talk about their approach in the following section.

4.2. Translation and Part-of-Speech Tagging

Even though translation and part-of-speech (POS) tagging are completely separate tasks, the only paper (to the best of our knowledge) that tackles these tasks in Ancient Egyptian does it in tandem. It should be noted that only the results for the translation task are reported.

Wiesenbach and Riezler (2019) compare different approaches for translating Middle Egyptian into German. These model several tasks jointly under the assumption that it would help with the small amount of data available. They compare using hieroglyphs and their transcription for translation (the many-to-one approach); using hieroglyphs to translate, transcribe, and extract POS tags at the

[19] https://c-lara.unisa.edu.au/lara_legacy/hieroglyphics1avocabpages/_hyperlinked_text_.html

same time (the one-to-many approach); and using both hieroglyphs and their transcription to translate, transcribe, and extract the POS tags (the many-to-many approach). As a baseline with which to compare these approaches they use a system that directly translates hieroglyphs to German.

Their models have an encoder for each type of input and a decoder for each type of output (depending on the approach). These are based on a GRU[20] architecture with attention. They experimented both with a more shallow network of one layer and a deeper one of four layers. For the learning process they compare different schedules to determine whether to lend more weight to the main task (translation) or to the assistance tasks. The data they used was a subset of the Thesaurus Linguae Aegyptiae (TLA) (Seidlmayer, 2011) mentioned in Section 3.

The best performance of their baseline system is a BLEU score of 19.86 points. This score is improved for the best many-to-one system to 21.61 points and to 22.79 points for the best one-to-many system. Meanwhile, the many-to-many system showed no improvement over the baseline, with a BLEU score of 18.07. Thus they conclude that jointly translating, transliterating, and doing POS tagging yields better results than doing a direct translation. It is of note that they do not report results neither on the transcription task nor on the POS tagging task.

4.3. Text Classification

Automatic text classification is another important task in NLP, as it can help document organization and management, text filtering or sense disambiguation. This is particularly useful for ancient languages as it allows us to study them without having to sift through and manipulate the original documents.

Gohy et al. (2013) mention that doing text classification can also give us insights into the registers used for different kinds of texts, which in turn should help improve the performance of machine learning techniques in other NLP tasks. They further claim that this is an important endeavor in the case of dead languages such as Late Egyptian.

In their paper Gohy et al. (2013) did genre classification. The genres they chose were letters, judicial documents, oracular questions, educational texts, monumental inscriptions, hymns and administrative texts. The authors argue that, while assuming that different genres do not overlap is an oversimplification, when chosen carefully they should be relatively independent from each other. They also note that another strong assumption that they are making in their paper is that each genre will have one and only one register and that each register will be exclusive to one genre, which is not true in general. Finally, as they are only interested in the registers, their models use mainly just semantic and morpho-syntactic features, while mostly ignoring the metadata and the structure of the texts.

The models that they used were a naïve Bayes classifier, an SVM, and a segment and combine method (which learns from each syntactic property of the document and then combines what it learnt to get further insights). Their best performing model was the naïve Bayes classifier, which achieves a recall of slightly over 0.84 in general and of over 0.97 with both letters and monumental inscriptions. They consider that in the case of the monumental inscriptions this is due to the more rigid structure used for the language and in the case of the letters it is due to the higher volume of training data available. On the other hand, this model gets a recall of only 0.66 with oracular texts. The authors consider that this is because oracular questions were usually very short (usually one or two sentences) and dealt with daily life matters thus being mostly misclassified as letters. Therefore, they created a modified naïve Bayes classifier which takes into account the length of the texts. This new model improved the recall of oracular questions to over 0.9 and got a general recall improvement of approximately 3%. Their SVM model got similar, but slightly worse results, while the segment and combine model got much more extreme results, with letters, judicial and educational documents, and monumental inscriptions getting a recall of over 0.9, but oracular questions and administrative texts having a recall lower than 0.3.

4.4. Text Retrieval

One of the NLP tasks that would be the most useful for egyptologsts is text retrieval. This task allows to create systems capable of searching and querying indexed documents. Using these kinds of systems would save researchers the effort of sifting through piles of useless data. They also function as a cultural preservation tool, by diminishing the amount of manipulation suffered by the actual physical documents.

In their paper, Iglesias-Franjo and Vilares (2020) created a text information retrieval system for Middle Egyptian. They consulted several egyptologists in order to determine the needs of such a system, most of which were either simplicity of use, flexibility and adhering to the current standard practices of the field. The system first preprocesses and normalizes the text of the documents. The normalization step refers to the way the hieroglyphs are tokenized into "sign groups" as opposed to each symbol being taken separately. After this, an index

[20]GRU stands for gated recurrent unit, a kind of recurrent neural network (Cho et al., 2014).

is created and stored. Once the index is in place, queries can be made. These can be made in latin script, hieroglyphs or a combination of the two. The text is then normalized as in the indexing stage, with the difference that a query using hieroglyphs can specify whether the symbols are the only ones appearing or if the user is looking for words that contain those symbols. Then, a list is selected and ranked according to a Boolean model and a vector space representation of the documents. The authors note that this is a first release and that there is still much work to be done. The system is freely available at their GitHub page.[21] Another approach that they proposed was using a method similar to those used for Japanese dictionaries, where words can be searched by using a combination of kanji (ideograms) and kana (syllabary). However, this query method was considered too unintuitive by the authors. They also note that completion of the Ramses or the Thesaurus Linguae Aegyptiae corpora mentioned in Section 3 could be a great boon to these kinds of systems.

4.5. Semantic Representations

Even though Ancient Egyptian lacks the amount of text needed to create embeddings (either contextual or non-contextual), that does not mean that useful semantic representations cannot be made.

Semantic maps (Georgakopoulos and Polis, 2018) are graphs of meanings such that two meanings are connected to each other if there is a language in which the same linguistic item is used for both meanings. These maps not only help visualize how meanings vary across languages, but can also be used to determine how languages vary across time. Thus, Georgakopoulos and Polis (2021) created diacronic semantic maps both for Ancient Egyptian and Ancient Greek. They argue that these maps properly reflect the expected semantic changes that happened during the chosen period of time.

5. NLP for Coptic

Even though Coptic can be considered a later stage of Ancient Egyptian, it has important differences with respect to Classical and Late Egyptian (Bard, 2005). This leads to a different set of problems when using NLP techniques with the language. One of these differences is that Coptic is no longer written in hieroglyphs, as it uses a modified version of the Greek alphabet instead. This leads to transliteration no longer being an issue, as there is a one-to-one correspondence between symbols and phonemes.

[21] http://github.com/estibalizifranjo/hieroglyphs

Another factor is that the morphology of the language went through several major changes. One example of this is the difference in the usage of affixes along with a huge influx of loanwords from Greek, which did not always adapt to the Coptic morphology (Kramer, 2006; Zeldes and Schroeder, 2016). An example on how this affects the design of NLP tools is with segmentation, especially when attempting to detect the language origin of a word.

A lot of documents from early Christianity were written in Coptic and the Coptic Orthodox Church still uses the language during mass. This means that there are more well-preserved texts in Coptic than in Ancient Egyptian. Thus, the contents of these texts tend to attract more attention from a wider variety of scholars such as those in Christian theology and related fields.

5.1. Morphological Analysis

Smith and Hulden (2016) did morphological analysis on Sahidic Coptic, one of the dialects of Coptic. They consider that a good model could be a transducer as it is mainly a prefixing language save for a few notable exceptions. Their testing set was composed of over a hundred words and had a recall slightly lower than 0.95. They think that their work could be useful for teaching the Coptic grammar and note that it could help study the larger Coptic texts. However, they make no mention on whether their model would need major modifications to consider other dialects, only stating that increasing the coverage of their analyser would need more lexicographical work.

Meanwhile, Ashton (2012) use a combination of a context-free grammar and transducer to model a smaller-scale morphological phenomenon, namely, second position clitics in Sahidic Coptic. They base the rules for their grammar in the linguistic literature. They do not provide any implementation or experimental results, as they note that an actual implementation of their system would be complicated from a technical point of view.

5.2. Named Entity Recognition

Yousef et al. (2023) combined out-of-the-box named entity recognition (NER) systems with transformer-based architectures for text alignment. Their system worked reasonably well for Ancient Greek and Latin versions of the Bible. However, they note that this approach did not work when dealing with Coptic versions of the same texts.

On the other hand, Khakhmovich et al. (2020) propose to use cross-lingual transliteration with transformer-based models as a way to tackle out-of-vocabulary terms, using Coptic as an example among other languages.

5.3. The Coptic Scriptorium and Universal Dependencies

As was mentioned in Section 3, the Coptic Scriptorium (Schroeder and Zeldes, 2016) is a corpus that had at its release a little less than 60 thousand words available. Several tools have been developed to be used along with it, which we will talk about in the rest of this section.

Zeldes and Abrams (2018) considered that the creation of a treebank compatible with the Universal Dependency (UD)[22] (de Marneffe et al., 2021) annotation scheme would be an important addition to the study of Coptic in general. They decided to work with the Coptic Scritptorium corpus due to it being freely available and also that the automatic segmentation achieves a very high precision score, which means that it can be considered a gold standard. They mainly decided to follow two main principles: when possible their notation should be compatible with the previous literature in the field and they would try to keep the notation in line with the practices in Hebrew and Arabic, which come from the same language family. When testing their treebank against expert human annotators, they got an agreement of over 95%. The agreement dropped to slightly over 85% when compared to undergraduate students. This was the first treebank built for the Egyptian language subfamily.

Another tool for the Coptic Scriptorium came in the form of a pipeline for NLP analysis. Zeldes and Schroeder (2016) created an online tool that automates several tasks, namely segmentation, normalization, tagging and lemmatization, detection of language of origin, and parsing.

For the segmentation task they selected around 180 rules and created a model that determined the priority order of the rules through 10-fold cross-validation. The accuracy of this model was slightly higher than 0.9. In the normalization stage, they had to consider the use of diacritics, spelling variations, and abbreviations. For this task, they used a combination of a predetermined list of common variations and a learnt list of the use of diacritics and capitalization. This model had an accuracy of 0.98. For part-of-speech tagging and lemmatization, they used an algorithm called TreeTagger (Schmid, 1999) and achieved accuracies of 0.95 and 0.97, respectively. As for determining whether the language of the text was Coptic, they had an accuracy of over 0.93. Finally, the parsing section has a preliminary version of the model of the paper from Zeldes and Abrams (2018) mentioned previously in this Section, which achieves an accuracy of 0.87.

Each of the components on the paper by Zeldes and Schroeder (2016) can be used either on their own or as part of a pipeline and can be accessed both at the author's website[23] or as part of the Coptic Scriptorium project[24].

As part of UD, the Coptic Scritorium has also been used for other projects. One of these was the the second shared task of SIGMORPHON 2019 (McCarthy et al., 2019), which was on morphological analysis given a word's context. The winning team (Straka et al., 2019) used an ensemble of nine LSTM (Hochreiter and Schmidhuber, 1998) models using BERT (Devlin et al., 2019). They also joined subcorpora from different languages. Their model achieved the highest performance on the Coptic subcorpus, with a lemma accuracy of 0.97 and a morpheme accuracy of 0.96.

Other projects in which the UD version of the Coptic Scriptorium has been used are multilingual dependency parsing (Dehouck and Denis, 2019; Choudhary, 2021; Choudhary and O'riordan, 2023), morphological tagging (Chakrabarty et al., 2019), studying the order of cosisters[25] (Dyer, 2018), studying information-theoretic locality properties of trees (Futrell, 2019), developing a multilingual categorical grammar (Tran and Miyao, 2022), as well as studying whether quantitative laws of language hold (Berdicevskis, 2021). We don't go into technical details of these approaches as Coptic is not a central part of any of these papers.

Finally, it has also been used as part of a study on the quality of the different treebanks of UD (Kulmizev and Nivre, 2023). While the Coptic treebank scores well in most of the metrics investigated in that paper, the authors note that it is one of the bottom three treebanks in terms of variability as defined by Swayamdipta et al. (2020).

6. Summary & Conclusion

The use of NLP methods on Ancient Egyptian is useful as it can help us gain insights both from a linguistic and from a historical standpoint. However, the advances in this field of research have been sparse through time. Polis et al. (2013) and Nederhof and Rahman (2015) consider that this has been in good part due to the lack of annotated text. They also note that most attempts are trying to generalize over large periods of time even when taking into account divisions such as Middle and Old Egyptian.

Another notable thing is that most papers have focused on Coptic. This is understandable as its inclusion in the UD project means that it has access to a wide array of tools that are being developed

[22]https://universaldependencies.org/

[23]https://corpling.uis.georgetown.edu/coptic-nlp/

[24]https://copticscriptorium.org/

[25]Defined in that paper as "sister constituents of the same syntactic form on the same side of their head".

with this project in mind. However, this tends to shift attention from the other stages of Ancient Egyptian, with Demotic being the most affected.

In their 2017 talk, Polis and Razanajao (2017) note that more interaction between projects could be useful, not only in the field of computational linguistics, but in Egyptology as a whole. This is especially important as most projects use either the same datasets or the same objects, but end up having their own systems and annotation schemes that are not compatible with each other. An example they give is that of a statue with inscriptions. The artifact itself has value for some researchers, while the kind of object or its inscriptions might be of interest to others. They also note that, while some researchers might be interested in the location and the layout of the text, some others might be just interested in the text itself or even in just the content. They mention that there is a current collaborative project called THOT (Dils et al., 2018) that aims to be a bridge for these areas of study. While the project does not have any sort of connection to the actual databases, their website has a roadmap to show how it will grow in the future.

This area of research appears to be approached by a very limited amount of researchers. However, some of these research groups appear to be growing, such as the one dedicated to the Ramses corpus, the evolution of which can be seen in Polis et al. (2013), Polis et al. (2015), and Polis and Razanajao (2017). We hope that this work will bring about a larger interest and allow for fruitful collaborations between the fields of NLP and Egyptology.

As a final note, an interesting thing would be to compare and contrast the NLP advances that have been done in other ancient languages, such as Sumerian, Ancient Greek, Sanskrit, etc. This could show how the advances in these different languages have affected or influenced each other. Even though some of the papers that we have mentioned so far did show this, most did not. A development in this direction comes from an NLP package called The Classical Language Toolkit (Johnson et al., 2021). It has tools for several ancient languages and even provides access to corpora for several of them, including the Coptic Scriptorium corpora mentioned in Section 3. This package could help encourage more research on these languages, which will help in turn gain important insights into our past.

7. Bibliographical References

Neil Ashton. 2012. Second position clitics and monadic second-order transduction. In *Proceedings of the Workshop on Applications of Tree Automata Techniques in Natural Language Processing*, pages 31–41, Avignon, France. Association for Computational Linguistics.

Kathryn A. Bard. 2005. Egyptian language and writing. In *Encyclopedia of the Archaeology of Ancient Egypt*, pages 325–328. Routledge. Google-Books-ID: MH7sAgAAQBAJ.

François Barthélemy and Serge Rosmorduc. 2011. Intersection of multitape transducers vs. cascade of binary transducers: The example of egyptian hieroglyphs transliteration. In *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*, pages 74–82. Association for Computational Linguistics.

Aleksandrs Berdicevskis. 2021. Successes and failures of menzerath's law at the syntactic level. In *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pages 17–32, Sofia, Bulgaria. Association for Computational Linguistics.

Branislav Bédi, Belinda Chiera, Cathy Chua, Brynjarr Eyjólfsson, Manny Rayner, Catherine Orian Weiss, and Rina Zviel-Girshin. 2022. Using LARA to create annotated manuscripts and inscriptions for museums: an initial feasibility study. In Birna Arnbjörnsdóttir, Branislav Bédi, Linda Bradley, Kolbrún Friðriksdóttir, Hólmfríður Garðarsdóttir, Sylvie Thouësny, and Matthew James Whelpton, editors, *Intelligent CALL, granular systems and learner data: short papers from EUROCALL 2022*, 1 edition, pages 18–23. Research-publishing.net.

Abhisek Chakrabarty, Akshay Chaturvedi, and Utpal Garain. 2019. Neumorph: Neural morphological tagging for low-resource languages—an experimental study for indic languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(1).

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Chinmay Choudhary. 2021. Improving the performance of UDify with linguistic typology knowledge. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 38–60, Online. Association for Computational Linguistics.

Chinmay Choudhary and Colm O'riordan. 2023. Multilingual end-to-end dependency parsing with linguistic typology knowledge. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 12–21, Dubrovnik, Croatia. Association for Computational Linguistics.

Mathieu Dehouck and Pascal Denis. 2019. Phylogenic multi-lingual dependency parsing. 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 192–203, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William Dyer. 2018. Integration complexity and the order of cosisters. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 55–65, Brussels, Belgium. Association for Computational Linguistics.

Richard Futrell. 2019. Information-theoretic locality properties of natural language. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 2–15, Paris, France. Association for Computational Linguistics.

Thanasis Georgakopoulos and Stéphane Polis. 2018. The semantic map model: State of the art and future avenues for linguistic research. *Language and Linguistics Compass*, 12(2):e12270. E12270 LNCO-0727.R1.

Thanasis Georgakopoulos and Stéphane Polis. 2021. Lexical diachronic semantic maps: Mapping the evolution of time-related lexemes. *Journal of Historical Linguistics*, 11(3):367–420.

Stéphanie Gohy, Benjamin Martin, and Polis Stéphane. 2013. Automated text categorization in a dead language. the detection of genres in late egyptian. In *Texts, Languages & Information Technology in Egyptology. Selected papers from the meeting of the Computer Working Group of the International Association of Egyptologists (Informatique & Égyptologie), Liège, 6-8 July 2010*, Aegyptiaca Leodiensia. Presses Universitaires de Liège. Backup Publisher: F.R.S.-FNRS - Fonds de la Recherche Scientifique.

Sepp Hochreiter and Jürgen Schmidhuber. 1998. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Janice Kamrin. 2004. *Ancient Egyptian Hieroglyphs: A Practical Guide - A Step-by-Step Approach to Learning Ancient Egyptian Hieroglyphs*. Harry N. Abrams. Google-Books-ID: JsWZQgAACAAJ.

Aleksandr Khakhmovich, Svetlana Pavlova, Kira Kirillova, Nikolay Arefyev, and Ekaterina Savilova. 2020. Cross-lingual named entity list search via transliteration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4247–4255, Marseille, France. European Language Resources Association.

Ruth Kramer. 2006. Root and pattern morphology in coptic: Evidence for the root. In *Proceedings of the 36th Annual Meeting of the North East Linguistic Society*, volume 2. University of Massachussets Amherst.

Artur Kulmizev and Joakim Nivre. 2023. Investigating UD treebanks via dataset difficulty measures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1076–1089, Dubrovnik, Croatia. Association for Computational Linguistics.

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.

C.K. Maisels. 1998. *Near East: Archaeology in the 'Cradle of Civilization'*. The experience of archaeology. Routledge.

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.

Georg Möller. 1919. Die buchschrift der alten Ägypter. In *Zeitschrift des Deutschen Vereins für Buchwesen und Schrifttum*, volume 2, pages 73–79. Verlag des Deutschen Vereins für Buchwesen und Schrifttum.

Mark Jan Nederhof. 2009. Automatic alignment of hieroglyphs and transliteration. In *Information Technology and Egyptology in 2008: Proceedings of the meeting of the Computer Working Group of the International Association of Egyptologists*, 2. Gorgias Press. Accepted: 2011-01-07T14:05:02Z ISSN: 1943-9369.

Mark-Jan Nederhof and Fahrurrozi Rahman. 2015. A probabilistic model of ancient egyptian writing. In *Proceedings of the 12th International Conference on Finite-State Methods and Natural Language Processing 2015 (FSMNLP 2015 Düsseldorf)*. Association for Computational Linguistics.

Mark-Jan Nederhof and Fahrurrozi Rahman. 2017. A probabilistic model of ancient egyptian writing. *Journal of Language Modelling*, 5(1):131–163.

Serge Rosmorduc. 2009. Automated transliteration of egyptian hieroglyphs. In Nigel Strudwick, editor, *Information Technology and Egyptology in 2008*, pages 167–182. Gorgias Press.

Serge Rosmorduc. 2015. Computational linguistics in egyptology. *UCLA Encyclopedia of Egyptology*, 1(1).

Daniel Smith and Mans Hulden. 2016. Morphological analysis of sahidic coptic for automatic glossing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2584–2588. European Language Resources Association (ELRA).

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.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9275–9293, Online. Association for Computational Linguistics.

Tu-Anh Tran and Yusuke Miyao. 2022. Development of a multilingual CCG treebank via Universal Dependencies conversion. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5220–5233, Marseille, France. European Language Resources Association.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Philipp Wiesenbach and Stefan Riezler. 2019. Multi-task modeling of phonographic languages: Translating middle Egyptian hieroglyphs. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Tariq Yousef, Chiara Palladino, Gerhard Heyer, and Stefan Jänicke. 2023. Named entity annotation projection applied to classical languages. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 175–182, Dubrovnik, Croatia. Association for Computational Linguistics.

8. Language Resource References

Branislav Bédi, Hakeem Beedar, Belinda Chiera, Nedelina Ivanova, Christèle Maizonniaux, Neasa Ní Chiaráin, Manny Rayner, John Sloan, and Ghil'ad Zuckermann. 2022. Using LARA to create image-based and phonetically annotated multimodal texts for endangered languages. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 68–77, Dublin, Ireland. Association for Computational Linguistics.

Walter E. Crum. 1939. *A Coptic Dictionary*. Oxford University Press.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Peter Dils, Silke Grallert, Ingelore Hafemann, Stéphane Polis, Lutz Popko, Vincent Razanajao, Simon Schweitzer, and Daniel Werning. 2018. Thot - thesauri and ontology for ancient egyptian resources.

Frank Feder, Maxim Kupreyev, Emma Manning, Caroline T. Schroeder, and Amir Zeldes. 2018. A linked Coptic dictionary online. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 12–21, Santa Fe, New Mexico. Association for Computational Linguistics.

Andrew Glass, Jorke Grotenhuis, Mark-Jan Nederhof, Stephane Polis, Serge Rosmorduc, and Daniel A Werning. 2021. Additional control characters for ancient egyptian hieroglyphic texts. Accessed: 2024-02-15.

Estíbaliz Iglesias-Franjo and Jesús Vilares. 2020. Searching four-millenia-old digitized documents: A text retrieval system for egyptologists. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 22–31. Association for Computational Linguistics.

Janet H Johnson, editor. 2001. *The Demotic Dictionary of the Oriental Institute of the University of Chicago*. The Oriental Institute, University of Chicago.

Kyle P. Johnson, Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. 2021. The Classical Language Toolkit: An NLP framework for pre-modern languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 20–29, Online. Association for Computational Linguistics.

Mark-Jan Nederhof, Stéphane Polis, Serge Rosmorduc, and Simon Schweitzer. 2019. Unicode control characters for ancient egyptian. In *12th International Congress of Egyptologists*. IFAO, Cairo, Egypt.

Sebastian Nordhoff and Thomas Krämer. 2022. IMTVault: Extracting and enriching low-resource language interlinear glossed text from grammatical descriptions and typological survey articles. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 17–25, Marseille, France. European Language Resources Association.

Stéphane Polis, Luc Desert, Peter Dils, Jorke Grotenhuis, Vincent Razanajao, Tonio Sebastian Richter, Serge Rosmorduc, Simon D. Schweitzer, Daniel A. Werning, and Jean Winand. 2021. The thot sign list (tsl). an open digital repertoire of hieroglyphic signs. *Égypte nilotique et méditerranéenne*, 14.

Stéphane Polis, Anne-Claude Honnay, and Jean Winand. 2013. Building an annotated corpus of late egyptian. the ramses project: Review and perspectives. In *Texts, languages & information technology in egyptology: selected papers from the meeting of the Computer Working Group of the International Association of Egyptologists*, Aegyptiaca Leodiensia, pages 25–44. Presses Universitaires de Liège. OCLC: 843421912.

Stéphane Polis and Vincent Razanajao. 2017. Ancient egyptian philology: The digital turn. current projects and future perspectives for the study of ancient egyptian texts. In *Global Philology Open Conference*. Mondes anciens.

Stéphane Polis, Serge Rosmorduc, and Jean Winand. 2015. Ramses goes online. an annotated corpus of late egyptian texts in interaction with the egyptological community. International Congress of Egyptologists XI.

Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to german. In Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, Text, Speech and Language Technology, pages 13–25. Springer Netherlands.

Caroline T. Schroeder and Amir Zeldes. 2016. Coptic SCRIPTORIUM: A corpus, tools, and methods for corpus linguistics and computational historical research in ancient egypt. In *White Paper*. University of the Pacific.

Stephan J. Seidlmayer. 2011. Handbuch zur benutzung des thesaurus linguae aegyptiae (TLA). Berlin-Brandenburg Academy of Sciences and Humanities.

Laura Slaughter, Luis Morgado Da Costa, So Miyagawa, Marco Büchler, Amir Zeldes, and Heike Behlmer. 2019. The making of Coptic Wordnet. In *Proceedings of the 10th Global Wordnet Conference*, pages 166–175, Wroclaw, Poland. Global Wordnet Association.

Jörg Tiedemann. 2016. OPUS – parallel corpora for everyone. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia. Baltic Journal of Modern Computing.

Amir Zeldes and Mitchell Abrams. 2018. The coptic universal dependency treebank. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 192–201. Association for Computational Linguistics.

Amir Zeldes and Caroline T. Schroeder. 2016. An NLP pipeline for coptic. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 146–155. Association for Computational Linguistics.

A. Writing Systems

In this appendix we illustrate what the writing systems of the different variations of Ancient Egyptian looked like through a few examples.

Figure 2: An example of hieroglyphs from the Temple of Kom Ombo in Egypt. Picture taken from Encyclopaedia Britannica. This temple was built during the Ptolemaic Dynasty from 180 to 47 BC.
Copyright: Icon72/Dreamstime.com.
https://www.britannica.com/topic/hieroglyph#/media/1/265009/118144 (Accessed March 30, 2024)

Figure 3: A sheet in hieratic from the Papyrus D'Orbine. It contains part of the Tale of Two Brothers. This document was written during the 19th Dynasty, circa 1185 BC.
Copyright: Image in the public domain.
https://commons.wikimedia.org/wiki/File:Tale_of_two_brothers.jpg (Accessed March 30, 2024)

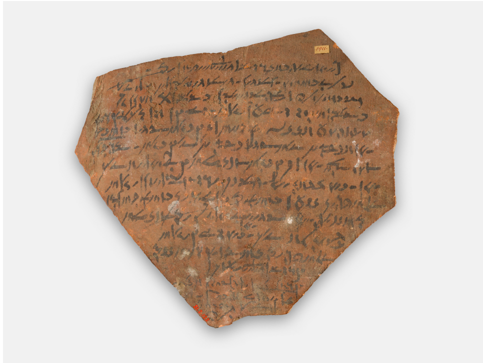

Figure 4: A text written in demotic script, from the Ptolemaic period (127 BC). It is an oath to the god Hathor denying the author's involvement in a cloths-theft.
Copyright: Rogers Fund, 1921. Image available under the Creative Commons CC0 1.0 Universal Public Domain Dedication.
https://commons.wikimedia.org/wiki/File:Demotic_Temple_Oath_MET_LC-21_2_122_EGDP023779.jpg (Accessed March 30, 2024)

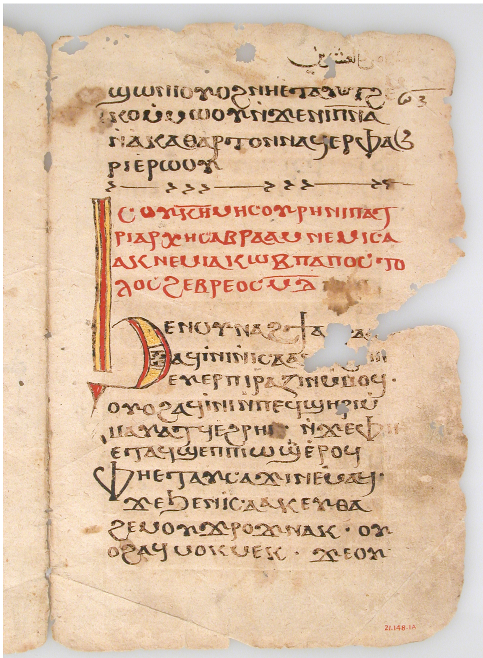

Figure 5: A page from a manuscript in Coptic. It is from sometime between the 6th and 14th centuries.
Copyright: Rogers Fund, 1921. Image available under the Creative Commons CC0 1.0 Universal Public Domain Dedication.
https://commons.wikimedia.org/wiki/File:Leaves_from_a_Coptic_Manuscript_MET_sf21-148-1as3.jpg (Accessed March 30, 2024)