#+TITLE: Computational Papyrology
#+AUTHOR: David Mimno and Hanna Wallach

The arid conditions of Ptolemaic and Roman Egypt preserved a remarkable number of papyrus documents. While these documents provide an unparalleled window into the culture of Egypt following the conquest of Alexander the Great, they can be difficult to work with. The quality of preservation varies: some documents are nearly complete while others are highly fragmentary. Furthermore, the language and cultural context of the papyri are unfamiliar even to scholars with a strong background in Greco-Roman antiquity. It is therefore these oldest documents that can benefit the most from cutting-edge text processing technologies. In this paper, we use advanced text processing algorithms to enhance the reconstruction, searchability and analysis of an existing online corpus of papyri.

The Duke Databank of Documentary Papyri (DDBDP) consists of over 50,000 non-literary texts, such as letters, contracts, and tax records. The corpus is primarily in Greek, with small amounts of Latin and Egyptian, in its Demotic and Coptic forms. Each document consists of an XML formatted text along with a find location and an estimate of its creation date. In most cases a papyrologist has manually entered reconstructions for missing and unclear text, to the extent possible. All gaps and reconstructions are marked.

We apply two techniques, developed within the natural language processing community, to the DDBDP. The first, statistical language modeling, has been widely used in many applications, including speech recognition and spelling correction. Given a "context" consisting of a short sequence of words, a statistical language model predicts the next word in the sequence. Such models can therefore be used to reconstruct papyri where words have been nibbled away by rodents or otherwise destroyed. In general, the more formulaic the language, the better the predictions. Documentary papyri frequently consist of highly formulaic language, thereby facilitating good predictions. Although statistical language models cannot replace the work of trained papyrologists, they can augment papyrologists' work by providing a broad range of hypothetical reconstructions based on statistical patterns in the entire corpus, which is too large for any one person can be familiar with. Furthermore, associating a probability with each hypothetical reconstruction means that the variability of reconstructions can be more easily explored.

The second technique, statistical topic modeling, involves clustering words within documents into "topics" or groups of semantically related words. Given a corpus, a statistical topic model automatically infers the words that comprise each topic cluster, as well as the topic clusters that occur in each document. Inference requires no human intervention. The topic clusters inferred for the DDBDP include words for Hellenistic rulers (Alexander, Arsinoe, Ptolemy), agricultural products (wheat, barley, beans) and sums of money (obols, drachmas). Topic clusters can be used to develop enhanced search interfaces that allow researchers to find not only those documents that contain particular query terms, but also documents that contain semantically related terms. Additionally, topic clusters can provide a broad semantic overview of an entire corpus and can be used to analyze the temporal and spatial evolution and distribution of language.