

First Results in a Study Evaluating Pre-annotation and Correction Propagation for Machine-Assisted Syriac Morphological Analysis

Paul Felt, Eric Ringger, Kevin Seppi, Kristian Heal[†], Robbie Haertel, Deryle Lonsdale[‡]

Dept. of Computer Science, [†]Neal A. Maxwell Institute, [‡]Dept. of Linguistics
Brigham Young University, Provo, Utah 84602 USA
{paul_felt, eric_ringger, kseppi, kristian_heal, robbie_haertel, lonz}@byu.edu

Abstract

Manual annotation of large textual corpora can be cost-prohibitive, especially for rare and under-resourced languages. One potential solution is *pre-annotation*: asking human annotators to correct sentences that have already been annotated, usually by a machine. Another potential solution is *correction propagation*: using annotator corrections to dynamically improve to the remaining pre-annotations within the current sentence. The research presented in this paper employs a controlled user study to discover under what conditions these two machine-assisted annotation techniques are effective in increasing annotator speed and accuracy and thereby reducing the cost for the task of morphologically annotating texts written in classical Syriac. A preliminary analysis of the data indicates that pre-annotations improve annotator accuracy when they are at least 60% accurate, and annotator speed when they are at least 80% accurate. This research constitutes the first systematic evaluation of pre-annotation and correction propagation together in a controlled user study.

Keywords: Annotated Corpora, Annotation, User Study

1. Introduction

The current success and widespread use of data-driven techniques for processing human language make annotated corpora an essential language resource. For instance, many popular natural language processing (NLP) algorithms require significant amounts of high quality annotated training data in order to perform effectively. Also, annotated text can be useful in its own right as a means of exploring the language and the culture that produced it. For example, one might use syntactic annotations to study discourse patterns, or topical annotations to track the movement of important ideas through time and space.

Scholars at the Center for the Preservation of Ancient Religious Texts (CPART) of the Neal A. Maxwell Institute for Religious Scholarship at BYU and at the Oriental Institute at the University of Oxford are jointly working on a project called the Syriac Electronic Corpus, with the goal of creating a comprehensive, labeled corpus of classical Syriac. Classical Syriac (‘kthobonoyo’) is an under-resourced Semitic language of the Christian Near East and a dialect of Aramaic. It was largely replaced by Arabic as a spoken language by the end of the ninth century, and is now primarily a liturgical language. Many prolific authors wrote in Syriac. The goal of the Syriac Electronic Corpus project is to annotate all of these texts with morphological information to facilitate systematic study of Syriac by historians, linguists, and language learners.

Morphological analysis of Syriac involves segmenting a word into its constituent morphemes and labeling each according to its grammatical form(s). For our purposes, a word token consists of a prefix, a suffix, and a *stem*, which we define as the remaining text. The dictionary citation form (or baseform) and, where applicable, the root are identified from the stem (Figure 1).

In contrast to English, where searching for a few forms of a word is often sufficient for discovering patterns reflecting the word’s usage and meaning, in Semitic languages search and discovery are not so straightforward. If we could search Syriac texts on citation forms or even on roots, we could search for and discover patterns as easily as in English; however, Syriac roots are altered by extensive inflectional and derivational morphological processes such that numerous surface forms correspond to any given root. As a result, searching Syriac text is ineffective since one must either limit one’s query to a single inflected surface form or use heuristics to expand the query, buying higher recall at the price of lower precision.

A morphologically annotated digital corpus of a lesser studied language lends itself to search and therefore to careful study in a way that formerly only experts could attempt based on long years of familiarity. Such annotated corpora enable scholars to study and discover the contributions of and trends in historical documents. One outstanding example of such a corpus is the Dead Sea Scrolls Electronic Library, assembled by CPART scholars (Tov, 2007). The Syriac Corpus will be an artifact of similar value to linguists, Syriac students, and scholars of Syriac, the Near East, and Eastern Christianity.

Unfortunately, creating annotated corpora can be extremely time-consuming. The Way International Foundation, a Biblical research, teaching, and fellowship ministry, spent 15 years labeling the Syriac New Testament with morphological annotations (Kiraz, 1994). The Syriac New Testament consists of approximately 100,000 words.

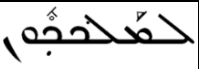
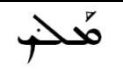
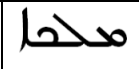

			
token	stem	citation form	root

Figure 1. The Syriac word token LMaLK’K,uON “to your king” and its related forms.

