World Academy of Science, Engineering and Technology
International Journal of Humanities and Social Sciences
Vol:12, No:02, 2018

The Automatisation of Dictionary-Based Annotation in a Parallel Corpus of Old English

Authors : Ana Elvira Ojanguren Lopez, Javier Martin Arista

Abstract : The aims of this paper are to present the automatisation procedure adopted in the implementation of a parallel corpus of Old English, as well as, to assess the progress of automatisation with respect to tagging, annotation, and lemmatisation. The corpus consists of an aligned parallel text with word-for-word comparison Old English-English that provides the Old English segment with inflectional form tagging (gloss, lemma, category, and inflection) and lemma annotation (spelling, meaning, inflectional class, paradigm, word-formation and secondary sources). This parallel corpus is intended to fill a gap in the field of Old English, in which no parallel and/or lemmatised corpora are available, while the average amount of corpus annotation is low. With this background, this presentation has two main parts. The first part, which focuses on tagging and annotation, selects the layouts and fields of lexical databases that are relevant for these tasks. Most information used for the annotation of the corpus can be retrieved from the lexical and morphological database Nerthus and the database of secondary sources Freya. These are the sources of linguistic and metalinguistic information that will be used for the annotation of the lemmas of the corpus, including morphological and semantic aspects as well as the references to the secondary sources that deal with the lemmas in question. Although substantially adapted and re-interpreted, the lemmatised part of these databases draws on the standard dictionaries of Old English, including The Student's Dictionary of Anglo-Saxon, An Anglo-Saxon Dictionary, and A Concise Anglo-Saxon Dictionary. The second part of this paper deals with lemmatisation. It presents the lemmatiser Norna, which has been implemented on Filemaker software. It is based on a concordance and an index to the Dictionary of Old English Corpus, which comprises around three thousand texts and three million words. In its present state, the lemmatiser Norna can assign lemma to around 80% of textual forms on an automatic basis, by searching the index and the concordance for prefixes, stems and inflectional endings. The conclusions of this presentation insist on the limits of the automatisation of dictionary-based annotation in a parallel corpus. While the tagging and annotation are largely automatic even at the present stage, the automatisation of alignment is pending for future research. Lemmatisation and morphological tagging are expected to be fully automatic in the near future, once the database of secondary sources Freya and the lemmatiser Norna have been completed.

Keywords : corpus linguistics, historical linguistics, old English, parallel corpus
Conference Title : ICHLL 2018 : International Conference on Historical Linguistics and Literature
Conference Location : Barcelona, Spain
Conference Dates : February 27-28, 2018