Menota and research infrastructures for medieval languages

Christian-Emil Ore, University of Oslo

The Medieval Nordic Text Archive, Menota, is a network of leading Nordic archives, libraries and research departments working with medieval texts and manuscript facsimiles. The aim of Menota is to preserve and publish medieval texts in digital form and to adapt and develop encoding standards necessary for this work. In addition, Menota has established a text archive, a text corpus and tools for lemmatization and annotating of texts. The archive is open for texts in the (medieval) Nordic languages as well as in Latin. There are currently 17 members of the Menota network including all major lexicographical projects for the Nordic Medieval Languages.

The current Menotic mark-up scheme is a slight extension of the encoding scheme defined in the TEI P5 guidelines. The TEI P5 basis of Menota enables the transcribers to establish texts with full mark-up according to modern text edition standards.

The Medieval texts are written in non standardized languages and demonstrate a very large degree of orthographical variation as well as a frequent use of abbreviations of letters, groups of letters and words. This makes a linguistic analysis challenging because of the difficulty in searching for words on the basis of a lemma. In Menota a text is seen as a series of graphic words. This is compatible with the text view in the Stuttgart Corpus Workbench which is used by a large number of text corpus projects. The Menotic extensions to TEI allow a three level view on (the words of) a running text: the facsimile, the diplomatic and the normalized levels. Each word can be annotated with a lemma and morpho-syntactical information. On the basis of a Menoticly encoded text, editions and text corpora can be established. An example of the latter can found at www.menota.org.

There are several extensions to the current Menota text collection. The lemmatisation is not straight forward since the variations, even within Old Norse (Old Norwegian and Old Icelandic), are very large and since the texts span a period of several hundred years. Even though the Menota mark-up scheme opens for a normalized level, the normalisation should be established as a derivative from the lemmatisation and the morpho-syntactical analysis. A tool for assisting the scholars in this kind of analysis will be the planned form thesaurus or meta dictionary connecting the word forms to the lemmas.

The Menotic mark-up scheme is currently not designed to encode information about syntactic structures, i.e., trees. Whereas the categories of morphological annotation are more or less given by the grammar of a language, this is much less true for syntax, where there are several competing theoretical models. The annotation software also needs to be more complex. The Menotic view of a text as an ordered series of words enables separate or stand off encoding of syntactic structures.

The medieval texts are important sources for a wide spectrum of scholarly disciplines spanning from corpus linguistics and text philology to onomastics, paleography and history. Thus the CLARIN network initiative can be seen as a generalization of the Menota network. It has similar objectives and meets the same challenges with respect to formats, perseveration, distributed resources, access and intellectual rights.

On the other hand, a network like Menota is in many ways more ambitious than a general collaborative network like CLARIN. In Menota there is a requirement that the source material should be based on transcriptions directly from the primary sources (not from existing editions) and encoded according to the well defined encoding standard given in the Menota handbook. This makes the archive a most important database from an editorial and text critical point of view, since it makes it possible to check textual witnesses in existing synthetic editions (i.e. editions based on more than one source and for which the critical apparatus does not allow the reconstruction of individual witnesses).