Gradual expansion in the use of the definite article
Checking a theory against the Old Hungarian Corpus

Barbara Egedi & Eszter Simon
Research Institute for Linguistics, Hungarian Academy of Sciences
egedib@yahoo.com, eszter@nytud.hu

Our paper aims to explore the possibilities and the limits of testing linguistic hypotheses made for the reconstruction of certain syntactic phenomena of an ancient synchronic system as well as to support the model we established for various grammaticalization processes and language change with the statistic data that can be gained from digital corpora. Both the theoretical and the computational work presented here make part of an ongoing research project set to the task, among others, of building a corpus of the Old Hungarian codices and minor records in addition to some selected texts from the Middle Hungarian period.

The hypothesis to be checked against a larger amount of data is the following. The definite article as a fully grammaticalized category to encode definiteness already exists in Old Hungarian, but is still absent in various contexts in which Modern Hungarian makes extensive use of it. Our careful classification of articleless noun phrases with definite interpretation revealed that this early definite article appears only in the constructions where the referent of the noun phrase is not anchored in another way. This investigation has been carried out by a manual search on a closed uniform text: the Gospel of Matthew in the so called *Munich Codex* which is one of the first extant continuous manuscripts from the first half of the Late Old Hungarian period (1370- mid 15th c.). Here, the definite article is absent with proper names and with a group of lexemes that describe entities with a prototypically unique referent. Unlike in Modern Hungarian, the article cannot co-occur with demonstratives, and is missing in case of a generic reading of the noun phrase. We do not find the definite article in the presence of a possessor expression either. Lastly, the lack of an article can also be due to the fact that the given expression is an adjunct rather than an argument of the verb and may remain unspecified with respect to definiteness. To sum it up, in the Old Hungarian system of determination the article has a more restricted use than it does in the subsequent language phases.

Expansion in the use of the article, however, does not take place simultaneously in all the possible contexts. A preliminary research (a contrastive analysis checking the collected articleless noun phrases against the corresponding loci in a parallel gospel text of a later date, but still in the same period) showed that the definite article spread into the generic function, and, at the same time, also started to appear before possessive pronouns, while its co-occurrence with demonstratives and nominal possessors is still not attested, these latter constructions being more characteristic of the Middle Hungarian period.

Studying various texts that follow each other in a diachronic order thus may reveal the way of the gradual spreading in article use, and the digitized corpus of the Old and Middle Hungarian sources have a crucial role in this task. Firstly, with the help of the searchable corpus, we can demonstrate that the proportion of the article increased gradually also within the Old Hungarian period. Secondly, the hypothesis outlined above can be checked against a larger amount of texts by means of a multi-leveled corpus query tool. Thirdly, by executing separated queries according to minor time-spans within the historical corpus, the micro-steps of the change can also be detected.

The Old Hungarian Corpus constantly and dynamically develops and will contain approx. 2 million tokens. At the moment, about 770000 tokens are already available and searchable.

Furthermore, a minor part of the texts has already been normalized and morphosyntactically annotated. Because of the specificity of the syntactic contexts we are working with, annotated corpora are particularly needed to confirm most of the assumptions concerning the grammatical encoding of referential properties. Some of the contexts are easily tested in any of the chosen manuscripts (such as deictic or possessive environments) and also a proposed list of inherently unique nouns can be checked with respect to determiner-resistance. Still, a corpus query has its limits since there will remain special contexts where expansion in article use can hardly be tested automatically. For instance, the absence of the article in Old Hungarian may mark the generic reading of a noun phrase, but generics are difficult to individualize without involving local semantic and pragmatic factors into the interpretational process. In most of the cases, however, our historical corpus is special enough, and thus suitable for testing well-defined syntactic contexts. The fact that the digitized sources come from different periods within a longer era under investigation makes the computational work and normalization more difficult, but, at the same time, considerably advances our endeavors to track the degrees of grammaticalization of the Hungarian definite article.