SEMATIA: Linguistic Annotation of Greek Papyri

Marja Vierros

Greek papyri from Egypt have preserved bigger and smaller entities of Greek as it was written by the ancient speakers from ca. 300 BCE to 700 CE. There are different registers, styles and vernaculars. Therefore the papyrological corpus forms an important direct source for Greek linguists, which has been exploited very little due to the lack of good tools for searching linguistic structures. The papyrological corpus is freely available in digital form in the Papyrological Navigator platform (http://papyri.info/) which also allows users to search both text strings and metadata (such as date and provenance). The search possibilities do not, however, easily yield to searching linguistic structures or variation in spelling or morphosyntax.

It has been my goal for several years to develop a solid system for getting at least parts of the papyrological corpus linguistically annotated so that they would better serve scholars studying variation and change in Greek in the postclassical period. There were two bigger questions to be solved.

First, what kind of linguistic annotation would be most useful? Because there already are two linguistically annotated treebank corpora for Ancient Greek, the Ancient Greek Dependency Treebank (374,490 words of Homer, Hesiod, tragedies) and the PROIEL treebank (181,364 words of the New Testament and Herodotus), the most reasonable solution was to follow the same framework (they both follow the Dependency Grammar) in order to gain maximal synergy between the corpora. Since a new annotation environment, Arethusa, was developed for the AGDT within the Perseids project (http://sites.tufts.edu/perseids/), my preliminary annotations of the papyri have followed the AGDT system.

The second question was more fundamental: How to deal with the fragmentary nature of the papyrological material within the annotation framework? The papyri are full of gaps and uncertain readings, abbreviations and nonstandard language which is not recognised by the automatic tools. For this purpose we have developed a layering tool called SEMATIA (http://sematia.hum.helsinki.fi) which filters two different layers from the EpiDoc XML format in which the texts are in the Papyrological Navigator. These layers are then ready for the linguistic annotation, one presenting only what was originally written and preserved in the papyrus and the other presenting the same text as a standardized version. By annotating both versions, it is possible to find the differences between standard and nonstandard language and search phenomena from both layers or only from the original layer. In this poster, I will present the layering tool and the current stage of annotations as well as discuss the querying possibilities and future developments we have been planning. Especially a third layer for marking up linguistic variation, and means for detecting and separating idiolectal features from scribal language are on the agenda.

References: