Representing Information Structure in Corpora of Ancient Languages

Hanne Eckhoff Dag Haug Eirik Welo

In the PROIEL project we study lexical and morpho-syntactic means of expressing pragmatic categories in Ancient Greek, Latin, Gothic, Old Church Slavic and Armenian. To this end we have developed a parallel corpus consisting of the New Testament texts in these languages. Since the corpus is relatively small, it is possible to add rich annotation, also concerning information structure which is at the core of the project. In this paper we describe the annotation system used and discuss some challenges and applications based on a pilot study containing 655 annotated Greek NPs.

The annotation focuses on the accessibility of discourse referents, using the familiar trepartite given/accessible/new distinction (Prince 1981). In addition, the syntactic annotation is exploited to render the 'anchoring' of a new referent to an old or accessible one. This simple scheme gives good results on interannotator agreement measures (average kappa-values of .859 between three annotators) and allows us to avoid the problematic notion of 'topic' while still being able to reconstruct it by combining the accessibility annotation with morphosyntactic features such as definiteness and word order.

Accessibility predicts the morphosyntactic realization of referents well, but to add more granularity to our scheme without subdividing the categories, we add anaphoric links to all given referents. This permits us to look at intervening material (number of subjects, referents, words) between anaphor and antecedent, which can be shown to correlate with a nominal realization even of old referents. We can also study the syntactic relationship between anaphor and antecedent, e.g. to extract conditions on the binding of reflexives, which are know be of a pragmatic nature in Ancient Greek. Finally, by combining the anaphoric links with the reconstruction of topic we can study phenomena such as topic shifting.