The Uppsala Corpus of Student Writings - Corpus Creation, Annotation, and Analysis
Beáta Megyesi, Jesper Näsman and Anne Palmér

The Uppsala Corpus of Student Writings consists of Swedish texts produced as part of a national test of students ranging in age from nine (in year three of primary school) to nineteen (the last year of upper secondary school) who are studying either Swedish or Swedish as a second language. National tests have been collected since 1996. The corpus currently consists of 2,500 texts containing over 1.5 million tokens. Parts of the texts have been annotated on several linguistic levels using existing state-of-the-art natural language processing tools. In order to make the corpus easy to interpret for scholars in the humanities, we chose the CoNLL format instead of an XML-based representation. Since spelling and grammatical errors are common in student writings, the texts are automatically corrected while keeping the original tokens in the corpus. Each token is annotated with part-of-speech and morphological features as well as syntactic structure. The main purpose of the corpus is to facilitate the systematic and quantitative empirical study of the writings of various student groups based on gender, geographic area, age, grade awarded or a combination of these, synchronically or diachronically. The intention is for this to be a monitor corpus, currently under development. The work has been previously presented at LREC 2016.