Learner corpus research methods in Latvia and Lithuania

Learner corpora research is a new but increasingly popular field in the world. There are various kinds of learner corpora, and they offer different possibilities. Every corpus must be designed accordingly to the kind of data one wants to research in it. In order to build a suitable corpus, one should know what kinds of methods they are going to use and what is therefore required of the prospective corpus.

The aim of this paper is to shortly describe the methods used and the technical requirements therefore imposed on learner corpora in Latvia and Lithuania as well as elsewhere in the world. In order to do that, relevant publications were studied and the methods were aligned with the structure and tagging of the learner corpora used. It was especially attempted to identify the main tendencies and problems in building and using learner corpora in Latvia and Lithuania.

Most of the learner corpora that are currently in use are built on written texts produced by language learners. They are often error-tagged. This helps one research the different kinds of errors students make in specific phases of language learning. While such corpora give a great insight, other kinds of corpora would make other approaches possible.

Written learner corpora that are tagged for parts of speech or other kinds of morphologically annotated learner corpora allow researchers to focus not on errors but on usage tendencies of specific parts of speech or forms. Syntactic annotation allows the same approach on syntactic level. A reference corpus of native speakers is often used to find out the differences and possibly spot the influence of native language and/or other languages.

Spoken language differs from written language greatly, and this is true for language learners as well as native speakers. In order to research it, a spoken learner corpus is required. Such corpora are more challenging to make than written corpora in terms of both collecting data and preparing it for use. Yet such corpora can provide other research possibilities, e.g., research on phonetic level such as pronunciation errors that cannot be found in written corpora. Information about such phenomena as filled pauses and backchannelling can also only be found in spoken corpora. All of these approaches require a specific kind of annotation in the transcription.

In Latvia and Lithuania, both written and spoken learner corpora are used. Most researchers work with error-tagged or untagged texts using concordance lines and word frequency lists. Tagging is a major challenge since it must be done manually and the corpora are often made by a single person with limited resources.