Eiríkur Rögnvaldsson, Department of Icelandic, University of Iceland

Corpus of Spoken Icelandic (ÍS-TAL)

The corpus of spoken Icelandic (ÍS-TAL) is a joint project with participants from three institutions: Icelandic University of Education (Kennaraháskóli Íslands), University of Iceland (Háskóli Íslands), and the Institute of Lexicography (Orðabók Háskólans). The project leader is Þórunn Blöndal, Assistant Professor at the Icelandic University of Education. The project has received generous grants from the Icelandic Research Council.

Studies of regional differences in pronunciation have a long tradition in Iceland, but apart from that, research on the spoken language has been very little, and no corpora of spoken Icelandic have been available to researchers. This led seven researchers from three different institutions to embark on the task of building a corpus of spoken Icelandic. These researchers have different background and different interests, comprising fields such as phonetics, phonology, morphology, syntax, lexicography, sociolinguistics, discourse studies, conversational analysis, language acquisition, corpus linguistics, and psychology.

The goal of the project is to establish a corpus of spoken Icelandic containing a precise transcription of ca. 200.000 words, or approximately 20 hours of spontaneous natural conversations. This material will hopefully lay a foundation for future research on the spoken language and also for comparative research on the written and spoken modes of the Icelandic language. Furthermore, it is expected that the corpus can be of use in speech technology. Obviously, this is a rather small corpus, so this must be regarded as a pilot project. However, the corpus has already proved its usefulness in several areas of research and teaching.

The project started in 1999, and the recordings were finished in 2000. The first phase of the transcription is now well under way, in which the material is transcribed using standard Icelandic orthography. Overlapping, interruption and latching is shown, and several types of comments have been inserted. In designing the corpus and the transcription system, we have especially looked at descriptions of the Swedish Spoken Language Corpus in Gothenburg and the British National Corpus, but neither of those has been closely followed.

By the end of this year, the first phase of the transcription will be completed. Lemmatization of the corpus material is also under way, and pilot studies of the vocabulary and syntax of the material have been presented publicly. The corpus material has already been used extensively in teaching, especially in courses on discourse analysis.

More detailed information on the corpus (in Icelandic) can be found at http://www.hi.is/~eirikur.