Eiríkur Rögnvaldsson
University of Iceland

The Icelandic Speech Recognition Project *Hjal*

1. The project

In the end of 2002, the University of Iceland and four leading companies in the telecommunication and software industry decided to join their efforts to build the first Icelandic speech recognizer. This project, which was called *Hjal* (‘babble’), was sponsored by the Icelandic Language Technology Fund. The goal of the project was to collect sufficient material to train a speaker-independent isolated word recognition system. Since the Language Technology Fund is government funded, the products of the projects that it supports are supposed to be public domain. This means that anyone who wants to develop a speech recognizer for Icelandic can get access to this material.

The project partners established a steering group with one member from each participant. Sæmundur Þorsteinsson from Icelandic Telecom served as Chairman of the steering group. The project leader was Helga Waage, MS, of Hex Software. Professor Eiríkur Rögnvaldsson at the University of Iceland was responsible for the linguistic preparations. We cooperated with ScanSoft, Inc. Their role was to train the speech recognizer on the basis of the material that we prepared. ScanSoft is a well established company in the ASR industry, and they have already developed speech recognizers for almost 50 languages.

2. Linguistic preparations

ScanSoft uses the SAMPA phonetic alphabet for phonemic transcription, so our first task in the linguistic preparations was to develop a SAMPA transcription standard for Icelandic. The next task was to make a detailed description of the phoneme inventory of Icelandic, including an exhaustive list of all possible diphones and a list of the most common triphones. No such list was available, so this took considerable effort. There turned out to be almost 800 different diphones in Icelandic.

The main task in the preparatory phase was to design caller sheets containing words, phrases and sentences for the participants to read. ScanSoft sent us rough guidelines as to the structure and content of these sheets. They were to include words and phrases that are likely to be used in ASR applications; a certain number of person names, place names, company names, numerals, numbers (money amounts etc.), commands, and meaningful fillers (*OK*, *please*, etc.). Furthermore, each sheet should contain five "phonetically rich" sentences and three strings of isolated letters.

In designing our sheets, we had to take into account the inflectional nature of Icelandic. Names, like other nouns, inflect for four cases, and some numbers (including all ordinal numbers) inflect for both case and gender (a few numbers even inflect for number as well). Hence, we felt it necessary to include more examples of these categories than proposed in the guidelines we got from ScanSoft.

The most difficult part was to construct the complete sentences. They were to be composed in such a way as to get enough samples of all occurring diphones and common triphones in Icelandic. The largest publishing house in Iceland, Edda Publishing, gave us access to the text of more than 100 recent novels (approximately 64 megabytes of text). From this corpus, all sentences containing 5-12 words were extracted automatically. This gave us almost 90,000 sentences. Then we made a frequency list of all the diphones occurring in these sentences. This list was used to select 3,000 sentences containing a sufficient number of all occurring diphones and common triphones.

After we had gone through all these sentences and removed all sentences which contained foreign words (especially names) or some potentially offensive material, we ended up with 1433 different sentences that were used in the caller sheets. 1,000 different sheets were then generated by randomly extracting a fixed number of items from each of the different lists (names, numbers, sentences, etc.).

The final task of the linguistic preparations was to make a word frequency list for Icelandic. This list was compiled from various sources; the newspaper *Morgunblaðið,* recent novels, and the Icelandic spoken language corpus *Ístal*. ScanSoft set the minimum size of the list to 30,000 word forms, but due to the inflectional character of Icelandic, we concluded that a considerably larger list would be feasible, and so we ended up with a list of almost 50,000 word forms.

3. Recordings and transcriptions

In order to be able to train a speech recognizer for Icelandic, ScanSoft needed to have speech data from at least 2,000 native speakers. In collecting these data, people were first asked to register as participants. These volunteers were then contacted and asked to call a toll free number. When they called in, they first had to answer a few questions and then were asked to read the caller sheets that had been sent to them. As mentioned above, 1,000 different sheets were generated, so on average, each sheet was read by two callers.

When the project was officially launched, we got good media coverage and used the opportunity to ask people to volunteer to call in and participate. We also hired Gallup Iceland to assist in recruiting volunteers to call in. By the end of the data collection phase, almost 3,000 people had volunteered to call in. Since the population of Iceland is about 285,000, this amounts to 1% of the whole population. When we had reached our goal of 2,000 valid recordings, sufficiently well distributed with respect to gender, age groups, regional dialects, and type of telephone (mobile vs. fixed line), we stopped contacting volunteers.

The recordings were distributed over 90,000 sound files. They were transcribed using normal Icelandic orthographic conventions. The wordlist, on the other hand, was transcribed using the SAMPA phonetic alphabet. The transcriptions were done by students in Language Technology at the University of Iceland, under the auspices of Professor Eiríkur Rögnvaldsson.

4. Results – and beyond

Linguistic preparations for the project started in February 2003, but most of the work was carried out in April and May. The recordings started in the end of May, and were completed by the middle of August. The transcribers began their work in early June, and finished in the end of August. In the beginning of September, all the recordings and the transcriptions had been sent to ScanSoft. The training of the Icelandic language model was completed by the end of October. The project was finished on budget and on schedule.

Since the speech recognizer was delivered in the beginning of November, comprehensive testing has been carried out. Although these tests are still ongoing, it seems safe to say that the results are quite satisfying; the recognition rate appears to be at least 97%. The system is able to tell apart very similar words, for instance, different inflectional forms of the same lexeme where the only difference lies in a vowel in an unstressed syllable (*hestur* 'horse' vs. *hestar* 'horses'). We expect that the speech recognizer will be integrated into several commercial applications in the course of the next few months.

The partners in *Hjal* feel that the cooperation between the University and the commercial companies has been very successful, and we are confident that this project can serve as a model for similar cooperation in other language technology projects. At present, the partners in *Hjal* are negotiating plans for developing a better speech synthesizer for Icelandic, and we hope that work on that project can start in the spring of 2004.