ALMANNARÓMUR: AN OPEN ICELANDIC SPEECH CORPUS

Jón Guðnason[⋆]*, Oddur Kjartansson*[⋆]*, Jökull Jóhannsson*[⋆]*,*
Elín Carstensdóttir[⋆]*, Hannes Högni Vilhjálmsson*[⋆]*, Hrafn Loftsson*[⋆]*,*
Sigrún Helgadóttir[†]*, Kristín M. Jóhannsdóttir*[‡]*, Eiríkur Rögnvaldsson*[‡]

[⋆] Reykjavik University, Menntavegur 1, 101 Reykjavik.
[†] The Árni Magnússon Institute for Icelandic Studies, Suðurgata, 107 Reykjavik.
[‡] University of Iceland, Sæmundargotu 2, 107 Reykjavik.

ABSTRACT

The purpose of the Almannarómur project is collecting data for a speech corpus (database) for Icelandic. Its main aim is creating an open source speech project to enable research and development for Icelandic language technology. The database is particularly suitable for acoustic modelling for speech recognition but it could also be used for other purposes, such as to develop a speaker recognition system or to analyze prosody. The project is run by Reykjavik University and the Icelandic Centre for Language Technology in cooperation with Google who provided technical support. The number of participants achieved in this effort was 563, providing, on average, around 219 read sentences each. This paper gives a short introduction to Icelandic language technology, describes how the text corpus was constructed for the database, and presents how the recording effort was organized as well as its main results.

***Index Terms*—** Icelandic, Speech Recording, Corpus Creation, Automatic Speech Recognition

1. INTRODUCTION

The effort for developing a basic language resource kit (BLARK) for Icelandic has been going on for over a decade and its main focus has been on text-based corpora and applications. From a spoken language perspective the only project during this time, Hjal, was carried out in cooperation with Iceland Telecom. Almannarómur[1] is a continuation of the effort of recording spoken Icelandic. The paper gives a short introduction to Icelandic language technology and presents how the recording effort was organized, as well as its main results.

The Almannarómur project is a part of an open source speech project, hosted by Google. The aim of the project is to enable small language communities to generate an open source speech corpus that can be used for research. The main aim of Almannarómur is to create a database of spoken sentences to aid development of automatic speech recognition for Icelandic. However, the database can be used for many other types of spoken language technologies. These design parameters are reflected in the method for generating the text corpus, i.e. there was neither a specific emphasis on having the corpus phonetically balanced, nor that it would contain all possible Icelandic phones.

The Almannarómur database was recorded using ten Google Android G1 smart-phones. A client-software was set up on the smart-phones that enabled downloading of Icelandic utterances and the uploading of speech recordings. The advantages of using smartphones for the data recording were that it was easy to reach participants (they did not need to turn up to specific locations) and the phones were easy to maintain and configure for the volunteers running the effort. The result is a database recorded on smart-phones in variable acoustic environments. This makes the database suitable for applications that are meant to work in such situations.

[1]The word is Icelandic and literally means "what is said in public."

2. ICELANDIC LANGUAGE TECHNOLOGY

2.1. Main Language Characteristics

Icelandic is a highly inflected (synthetic) language, with cases, numbers, genders, persons, tense and mood, as well as weak and strong declension. It is also fusional, in such a way that a single ending usually stands for more than one morphological category. The inflectional system is further complicated by a great number of inflectional and conjugational classes, such that the same morphological category, or combination of categories, is represented by a number of different endings depending on the stem.

The pronunciation of Icelandic is fairly transparent as the projection from the spelling to the pronunciation obeys almost unexceptional rules. This means that a speaker who knows the rules should be able to pronounce fairly accurately any new words that (s)he reads – as long as it is relatively clear where morpheme boundaries lie, as they can affect the pronunciation of some letter combinations. This is quite useful, for instance, in the making of speech synthesizers and speech recognizers. Furthermore, the stress rule in Icelandic is quite simple as the main stress always falls on the first syllable of a word, usually with additional stress on every second syllable after.

2.2. Language Technology

At the turn of the century, Icelandic Language Technology (LT) was virtually non-existent [1]. A relatively good spell checker had been developed and a speech synthesizer existed that served the blind community for the most part. In 1998 the minister of Education, Science and Culture appointed a joint committee of experts to investigate the situation of LT in Iceland and to make proposals for further advancement for Icelandic LT. This resulted in the launch of a special government-sponsored LT program with the aim of supporting institutions and companies in creating basic resources for Icelandic LT. The main direct products of the LT program are: a full-form morphological database of modern Icelandic inflections; a balanced morphosyntactically tagged corpus of 25 million words; a training model for data-driven PoS taggers; a text-to-speech system; an isolated word speech recognizer; an improved spell checker. At the end

of the LT program in 2004, researchers from three institutes (University of Iceland, Reykjavik University, and the Árni Magnússon Institute for Icelandic Studies) joined forces in a consortium called the Icelandic Centre for Language Technology (ICLT), in order to follow up on the tasks of the program. ICLT researchers have initiated several new projects, which have been partly supported by the Icelandic Research Fund and the Icelandic Technical Development Fund. Detailed descriptions of all these resources are given in [1] and references therein.

One of the most important product of these projects is the IceNLP package. IceNLP is a toolkit for processing and analyzing Icelandic text [2]. It contains components such as sentence segmentizer/tokenizer, part-of-speech tagger [3] and a syntactic parser [4].

In 2009, the ICLT received a relatively large three-year Grant of Excellence from the Icelandic Research Fund for the project "Viable Language Technology beyond English – Icelandic as a test case"[2]. Within that project, three types of LT resources are being developed: a database of semantic relations [5]; a prototype of a shallow-transfer machine translation system [6], and a treebank with a historical dimension [7]. The ICLT takes at present part in the European project META-NET [1].

Four speech synthesizers have been developed for Icelandic. A formant-based speech syntheziser was originally made around 1990 and another one, based on bi-phone techniques, around 2000. These synthesizers were used mostly by the blind and visually impaired, as their quality was far from satisfactory for use in commercial applications for the general public. In 2005, a new text-to-speech system was made in cooperation between the University of Iceland, Iceland Telecom and the now defunct Hex Software. The system was trained by Nuance and uses their technology. This system has hitherto not been used in commercial applications, and many users do not find its voice quality satisfactory. As the existing TTS-systems are lacking in quality for their main users, the Icelandic Organisation of Blind and Partially Sighted is now developing a new TTS-system in cooperation with the University of Iceland, Reykjavik University, and the Polish Ivona software company. This system is expected to be ready later this year.

An isolated word speech recognizer for Icelandic was developed in 2003 (the Hjal project), but the software was not widely used, partly due to limited access. An Icelandic student at the Tokyo Institute of Technology has developed a prototype of a system for automatic continuous speech recognition for Icelandic [8]. This system reached up to 67.5% word accuracy. Neither of these systems has been put to use in commercial applications.

3. TEXT CORPUS

The text corpus was based on sentences from news stories obtained from the online edition of the newspaper Morgunblaðið, named mbl.is. The news sentences were augmented with other sources which are described in this section.

3.1. Generating the Text Corpus

The text sources used in generating the Text Corpus are listed below. The percentages show the ratio of the individual text source in the final corpus.

- News stories: 50%
- Rare tri-phones: 10%
- Names of streets: 10%
- Names of people: 10%
- Miscellaneous: 10%
- Countries and capitals: 5%
- URLs: 5%

The total number of sentences in the database was fixed at twice the number of the sentences from the news stories source, or a total of 55,000. The other lists were multiplied appropriately in order to obtain the set ratio. For example, the number of sentences in the rare tri-phone list is 1432 but they occur 5500 times in the corpus, giving a repetition of circa 4 for each sentence in that list. The following subsections describe each sentence source and its weight in the corpus. The order of the sentences was randomized to ensure proportional sampling from the corpus during the recording process.

3.2. The news stories sentences

The online team at Morgunblaðið provided an access to about 55,000 news stories spanning a one year time period between June of 2010 and 2011. The main advantage of obtaining the data from this source is that it has significant amount of data which is sufficient to generate a large corpus. The headlines were easily extracted but the text needed to be processed by the IceNLP sentence segmentizer in order to obtain a complete sentence list.

The length of each sentence was limited to 6 words, in order to make reading easier and to ensure that the sentence would fit on the screen of the Android G1 device. Each sentence was checked for spelling, using the Database of Modern Icelandic Inflections [9]. Any sentences containing words not found in the dictionary were deleted from the final list.

3.3. Rare tri-phones

A list of sentences, containing both phonetically and grammatically significant words for the Icelandic language, was obtained from the Department of Icelandic at the University of Iceland. These sentences were hand-picked and added to the corpus. It was ensured that the rare bi- and tri-phones were included in the recorded data. This was important since the composition of the bi- and tri-phones in the rest of the corpus was not checked.

3.4. Locational word lists

Names of streets and historical places in Iceland were selected and added to the corpus. The names were presented in the dative case since that form is the most common occurrence of addresses in Icelandic. The idea was that the data collected might be used for navigational purposes such as in geographical positioning systems while driving. **Countries and capitals** of 190 countries were included in the corpus for the same reasons.

3.5. Names of people

List of proper names of people was added to the corpus. The list was available at the national registry website[3]. This was done in order to increase the ratio of proper nouns in the database.

[2]See: `http://iceblark.wordpress.com`

[3]See: `http://www.skra.is/`

3.6. Universal Resource Locators and Miscellaneous word lists

A list of the 100 most visited websites in Iceland was collected and added to the corpus. This list was obtained from the website `modernus.is`, which collects data on the most visited websites in Iceland.

A list containing numbers, dates, times of day, names of days and months, simple questions, and common greetings was included in the corpus. The inclusion of these items was considered important for applications such as spoken dialogue systems.

4. RECORDING

4.1. Combining and creating packages

To avoid fatigue and ill will, it was decided that the amount of time each participant would be asked to spend on the recording would not exceed 30 minutes. After preliminary tests, the target for the average number of utterances per participant was set to 250.

4.2. Recording the data

As discussed in the introduction the data was recorded using Android G1 phones using the mobile application Datahound [10]. No external microphone was used for the recording of the data. The recordings were sampled at 16 kHz in mono and no compression was used to store the samples.

The people donating their voice were non-paid participants of the project. They were asked to read for as long as they could, up to a maximum of 30 minutes and/or 350 utterances. Some were willing to read for the full 30 minutes while others contributed less. Participants signed a participation agreement allowing open use of the speech data.

The people who contributed to the project by recording their voices will hereafter be referred to as participants; the people who oversaw the recording will be referred to as volunteers or instructors.

4.3. Instructions given to participants

The following is a description of the instructions given to the participants before they started recording. This was done prior to recording in order to prevent any degradation quality of the data due to user fault.

Placement of recording device: The participants were free to place the recording device in whatever way they saw fit, as long as the voice was suitably recorded. Some participants placed the device on a table in front of them; some held the device at a comfortable distance; while others talked directly into the device. The instructors showed the participants where the microphone was located on the device so that they would not obstruct it while recording, which would create muffled and unusable recordings.

Tone of recording: The users were told that they should read the utterances in a natural way. The intention was to create a natural sounding dataset.

Hard to pronounce utterances: Some utterances can be considered hard to pronounce, examples of such are some names of countries or capitals which were included in the corpus. In order to make the process easier and faster the participants were given instructions to either skip the utterance or just read it as they believed to be correct. That is, they should not spend too much time on each utterance in order to get it right.

URLs: A special attention was given to URLs, since they cannot contain special Icelandic characters. Special characters are represented by their ASCII counterparts. Letters containing a diacritical mark, for example, are represented by the letter without the diacritic (ú becomes u, í becomes i, etc.). Such URLs are still read as if the diacritics were present. For example:

`ja.is` is commonly pronounced "já.is"

`visir.is` is commonly pronounced "vísir.is"

Similarly, abbreviations can be read differently. For example:

`mbl.is` is commonly read as m-b-l-punktur[4]-is, while

`ruv.is` is commonly read as rúv-punktur-is

Participants were instructed to read the utterances as they would normally do when communicating with others. The same rule applied here as for the hard to pronounce words, that they were instructed to skip them, or read them to the best of their knowledge, if they were in doubt.

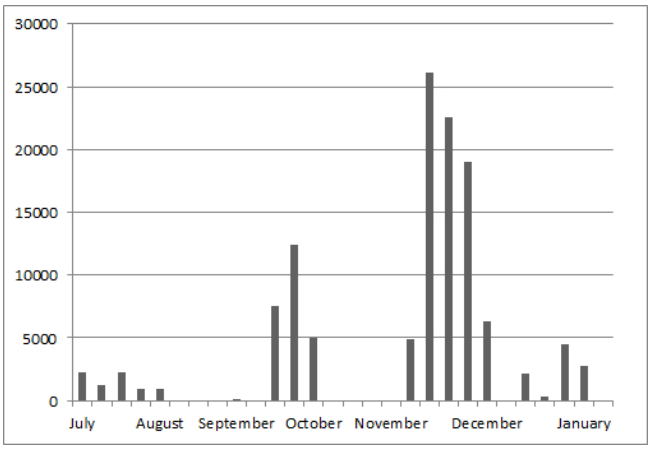

Fig. 1. The figure shows the number of sentences recorded for each week during the data recording effort.

5. DATA RECORDING RESULTS

5.1. Collection process

The data collection effort was carried out entirely on volunteering basis. It began on July 15, 2011 and was completed by January 31, 2012. The number of utterances recorded each week can be seen in Figure 1. Three approaches were tried in organizing the data collection, which is apparent in the figure. The first approach lasted until September. It was based on distributing the 10 smart-phones to volunteers, each of whom had the responsibility of getting participants. This approach was not as effective as anticipated. It turned out to be hard to get people to volunteer. The volunteers that did help out also had a hard time getting participants. The total number of people participating in this phase was 59.

The second approach was based on organized events around the data collecting effort. Series of events were advertised within the universities, where two to three volunteers collected speech from participants, using all 10 phones. This approach lasted for 4 weeks and was considerably more effective than the first approach, as 104 people participated in the project. However, this phase turned out to be very straining on the few volunteers that organized the effort.

[4]Punktur means dot in Icelandic.

Table 1. The number of participants and sentences.

	Male	Female	Total
Participants	303 (53.8%)	260 (46.2%)	563
Sentences	63,215 (51.3%)	60,012 (48.7%)	123,227
Average Sent.	208.6	230.8	218.9

Furthermore, it got progressively harder to find willing participants within the universities. A bigger outreached was therefore needed.

The most successful approach to collecting data was based on organized visits to companies and institutions. The preparation for this phase took some time as key individuals in the workplace were identified and approached. Typically, the manager of human resources or marketing helped with the organization within each workplace. Two to five volunteers were recruited and the duration of the collection was deliberately kept low, usually three to four days. This ensured a maximum impact from each workplace, as the volunteers knew that the effort only lasted for a short time, but the recommended target for the number of participants was high. This phase started in November and continued through January, but the highest volume was achieved in November, after a period of intense recruiting of workplaces in October. The total number of workplaces visited was 19 and the total number of participants in this phase was 430.

5.2. Database composition

The total number of participants was 563 and the total number of sentences in the database was 123,227. The breakdown of these numbers by gender is shown in Table 1. There were slightly fewer female participants than male participants in the collection. The table also shows the average number of sentences per speaker. This shows that on average female participants provided greater number of sentences than male participants.

Figure 2 shows the number of sentences provided by each age group and each gender. For example, the male participants in the age group between 30 and 39 provided 23,582 sentences and the female participants of the same age group provided 16,907 sentences. As can be seen in the figure, the distribution of sentences read by male participants is more concentrated in the younger age groups of 20 to 39 year olds while the distribution of sentences read by female participants is more even. Participation from people younger than 18 years was not permitted, which explains the low collection number for the youngest age group.

6. SUMMARY AND FUTURE WORK

The data recording effort of Almannarómur lasted from July 2011 to January 2012. During this time 113,547 sentences were recorded by 563 participants. This database will be made available to the public in order to develop spoken language technologies for the Icelandic language. For example, the database will be particularly suitable for short utterances in a mobile environment.

The development of this database facilitates the development of an Icelandic speech recognizer. The immediate steps include the training of an acoustic model, compilation of a pronunciation dictionary and a setup of a suitable language model for the task at hand.

Further development of spoken language resources for the Icelandic language is needed in order to extend this work to other forms of speech. The development of dictation systems or conversation analyzers need different forms of recordings.

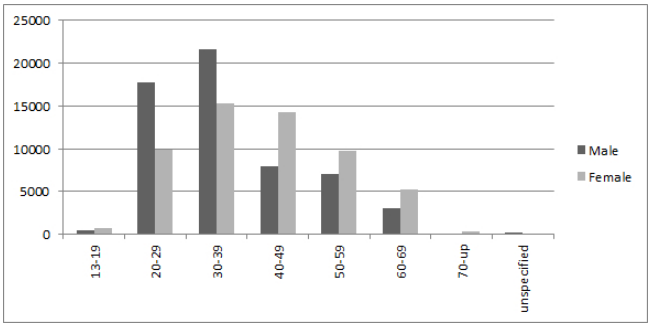

Fig. 2. The age distribution of the participants.

7. ACKNOWLEDGEMENTS

The Almannarómur project was partially realized because of the generous help received from Google and its employees. Google provided the smart-phones for the data recording effort and the server technology used to host the database.

8. REFERENCES

[1] E. Rögnvaldsson, "Icelandic Language Technology: An Overview," in *Language, Languages and New Technologies: ICT in the Service of Languages. Contributions to the Annual Conference 2010 of EFNIL in Thessaloniki*, Gerhard Stickel and Tamas Varadi, Eds. 2011, vol. 87 of *Duisburger Arbeiten zur Sprach- und Kulturwissenschaft*, pp. 187–195, Lang, Frankfurt am Main.

[2] H. Loftsson and E. Rögnvaldsson, "IceNLP: A Natural Language Processing Toolkit for Icelandic," in *InterSpeech 2007*, Antwerp, Belgium, 2007.

[3] H. Loftsson, "Tagging Icelandic text: A linguistic rule-based approach," *Nordic Journal of Linguistics*, vol. 31, no. 1, pp. 47–72, 2008.

[4] H. Loftsson and E. Rögnvaldsson, "IceParser: An Incremental Finite-State Parser for Icelandic," in *Proceedings of the* 16^{th} *Nordic Conference of Computational Linguistics (NoDaLiDa 2007)*, Tartu, Estonia, 2007.

[5] A. Nikulásdóttir and M. Whelpton, "Extraction of Semantic Relations as a Basis for a Future Semantic Database for Icelandic," in *7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages (Workshop 22 of 7th Language Resources and Evaluation Conference)*, Valletta, Malta., May 2010, pp. 33–39.

[6] M. D. Brandt, H. Loftsson, H. Sigurþórsson, and F. M. Tyers, "Apertium-icenlp: A Rule-based Icelandic to English Machine Translation System.," in *The 15th Annual Conference of the European Association for Machine Translation (EAMT-2011)*, Leuven, Belgium., 2011.

[7] E. Rögnvaldsson, A. K. Ingason, E. F. Sigurič£sson, and J. Wallenberg, "Creating a dual-purpose treebank," *Journal for Language Technology and Computational Linguistics*, vol. 26, no. 2, pp. 141–152, January 2012, Proceedings of the ACRH Workshop.

[8] A. Jensson, K. Iwano, and S. Furui, "Language model adaptation using machine-translated text for resource-deficient languages," *EURASIP Journal on Audio, Speech, and Music Processing*, 2008.

[9] K. Bjarnadóttir, "Modern Icelandic Inflections," in *Nordisk Sprogteknologi 2005*, H. Holmboe, Ed. Museum Tusculanums Forlag, Copenhagen, 2005.

[10] T. Hughes, K. Nakajima, L. Ha, A. Vasu, P. J. Moreno, and M. LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages," in *INTERSPEECH*, T. Kobayashi, K. Hirose, and S. Nakamura, Eds. 2010, pp. 1914–1917, ISCA.