Hjal Speech Corpus, HJAL corpus

1. BASIC INFORMATION

1.1. Corpus composition

Synchronized text and speech, sampled from 2005 individuals, text is transcribed and recorded in SAMPA standard.

1.2. Representation of the corpora (flat files, database, markup)

Sound and text files.

1.3. Character encoding

The characters have been encoded in UTF8.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: Eiríkur Rögnvaldsson
Affiliation: Íslensku- og menningardeild Háskóla Íslands.
Address: Árnagarði, 101, Reykjavík, Iceland
E-mail: eirikur@hi.is

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

Available for download from own webpage.

2.3. Copyright statement and information on IPR

CLARIN PUB license.

3. TECHNICAL INFORMATION

3.1. Directories and files

883 subdirectories. Each subdirectory contains data from one speaker, both sound files and a text file. - usually 47 sound files for each speaker, each file containing one utterance, and one text file containing the transcription of all the sound files.

3.2. Data structure of an entry

The sound files are in .wav format - usually 47 sound files for each speaker, each file containing one utterance. The transcription of all the sound files for each speaker is contained in one text file.

3.3. Corpora size (nmb. of tokens, MB occupied on disk)

42,000 sound files, 883 text files, 1,4 GB.

4. CONTENT INFORMATION

4.1. *Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)*

Monolingual, annotated?.

4.2. *The natural language(s) of the corpus*

Icelandic.

4.3. *Domain(s)/register(s) of the corpus*

Individual words.

4.4. *Annotations in the corpus (if an annotated corpus)*

4.4.1. *Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)*

Not applicable.

4.4.2. *Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),*

SAMPA phonetic alphabet.

4.5. Intended application of the corpus

For training speech recognizers.

4.6. Reliability of the annotations (automatically/manually assigned) – if any

Transcription and phonetic transcription performed manually.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Rögnvaldsson, Eiríkur (2004). The Icelandic Speech Recognition Project Hjal. In Holmboe, Henrik Ed.), *Nordisk Sprogteknologi. Årbog 2003*. Museum Tusculanums Forlag, University of Copenhagen, pp. 239-242.