Parliament Speech Corpus

1. BASIC INFORMATION

1.1. Corpus composition

About 20 hours of unprepared speeches in parliament 2004-2005, synchronized speech and text files. The corpus consists of sound files, transcribed speech (output of the software Transcriber) and TEI-conformant xml-files with morphosyntactic tags and lemmas.

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

The corpus consists of sound files (mp3), transcribed speech (output of the software Transcriber) and TEI-conformant xml-files with morphosyntactic tags and lemmas. The corpus is also available for online search.

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: Ásta Svavarsdóttir
Affiliation: The Arni Magnusson Institute for Icelandic Studies
Adress: Neshagi 16, 107, Reykjavík, Iceland
E-mail: asta@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

One directory, set of files (mp3, trs, xml) from 12 periods.

3.2. Data structure of an entry

The corpus is provided for download as (1) TEI-conformant xml-files. Each file contains a header with bibliographic information. The text is segmented into sentences and each sentence into tokens where each token is equivalent to a word or a named entity. Each token (running word) is accompanied by a morphosyntactic tag and a lemma. (2) Transcriber files where text and sound is synchronized. (3) Sound files (mp3-files).

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

About 190 thousand running words, 20 hours of speech.

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*

Parliamentary talk.

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)*

Sentence mark-up, morphosyntactic mark-up, synchronization of text and sound.

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

Morphosyntactic tags, lemmas

4.4.3. *Alignment information (if the corpus contain aligned documents: level of alignment, how it was achieved)*

Not applicable.

4.4.4. *Attributes and their values (if annotated)*

For the xml-files the following holds: The *s* tag has one attribute: *n* which identifies the sentence in the text. The *w* tag is present at the token level and can have two attributes: *type* whose value is the morphosyntactic tag and *lemma* whose value is the dictionary form of the word form.

4.5. Intended application of the corpus

Web-version will be used for linguistic investigations. Download version will be used for LT projects such as speech analysis, speech recognition, speech synthesis, automatic speech recognition.

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

Synchronization performed manually, PoS tagging and lemmatization performed automatically.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Ásta Svavarsdóttir. 2007. Talmál og málheildir — talmál og orðabækur. *Orð og tunga* 9:25-50.

Höskuldur Thráinsson, Ásgrímur Angantýsson, Ásta Svavarsdóttir, Thórhallur Eythórsson and Jóhannes Gísli Jónsson. 2007. The Icelandic (Pilot) Project in ScanDiaSyn. *Nordlyd*. *Tromsø University Working Papers on Language & Linguistics*, Vol. 34, Nr. 1: 87-124. (Sérhefti um Scandinavian Dialect Syntax 2005). Vefrit á slóðinni http://www.ub.uit.no/baser/nordlyd/viewissue.php?id=11.