Icelandic Parsed Historical Corpus, IcePaHC

1. BASIC INFORMATION

1.1. Corpus composition

Treebank

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

61 texts from 1150 to 2008. For each text there are three files: raw text; PoS tagged text (word, tag, lemma) in a flat file, one word per line; the texts in labelled bracketing format; With the corpus comes *Corpald*, a cross-platform graphical user interface to search corpora in labelled bracketing format. *Corpald* calls *CorpusSearch* by Beth Randall on the command line to execute search queries. (http://corpussearch.sourceforge.net/).

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 web page.

2.3. Copyright statement and information on IPR

Open Source, LGPL license

3. TECHNICAL INFORMATION

3.1. Directories and files

The package comes with several directories. The corpus itself is in the directory 'corpora' which has four subdirectories, one for bibliographic information for each of the 61 files ('info'), one for the raw text ('txt'), one for the tagged text ('tagged') and one for the parsed text in labelled bracketing format ('psd'). There are also directories for information needed to install the software for searching the corpus ('Corpald').

3.2. Data structure of an entry

Each text sample is one text file. The parsed files are annotated according to the Penn Treebank format with certain modifications. A detailed description of the annotation is found in http://www.linguist.is/icelandic_treebank/Icelandic_Parsed_Historical_Corpus_(IcePaHC)#Annotation_guidelines.

3.3. Corpora size (num. of tokens)

1 million running words

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*

Narratives (sagas, fiction), religious texts (bible, sermons),
science (linguistics, natural sciences, history),
formal texts (law, formal letters), biographical material (biographies, travelogues)

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, syntactic mark-up

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

POS and syntactic tags.

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)*

Not applicable.

4.5. Intended application of the corpus

The corpus is intended for use both within language technology and in syntactic research, synchronic and diachronic.

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

Mark-up automatically assigned, manually checked

5. RELEVANT REFERENCES AND OTHER INFORMATION

Wallenberg, Joel C., Anton Karl Ingason, Einar Freyr Sigurðsson og Eiríkur Rögnvaldsson. 2011. *Icelandic Parsed Historical Corpus (IcePaHC).* Version 0.9. http://www.linguist.is/icelandic_treebank

Eiríkur Rögnvaldsson, Anton Karl Ingason og Einar Freyr Sigurðsson. 2011. Coping with Variation in the Icelandic Parsed Historical Corpus (IcePaHC). Johannessen, Janne Bondi (ritstj.): *Language Variation Infrastructure. Papers on selected projects*, s. 97-111. Oslo Studies in Language 3.2. University of Oslo, Osló.