Icelandic Frequency Dictionary Corpus, IFD Corpus

1. BASIC INFORMATION

1.1. Corpus composition

Contemporary Icelandic texts from the years 1980-1990. The corpus contains 100 texts of about 5000 running words each. Texts were collected from printed books containing literary works for adults (20 texts written in Icelandic, 20 translated) and children (10 written in Icelandic, 10 translated), biographies (20 texts) and informative writings (10 from the humanities, 10 from natural sciences). The corpus will be made available in three ways: 1) most of the corpus will be available for online search. 2) 61 files (original Icelandic texts) are available for download under a special license. 3) Ten different disjoint pairs of files where in each pair there is a training set containing about 90% of running words from the corpus and a test set containing about 10% of running words from the corpus. The test sets are independent of each other whereas the training sets overlap and share about 80% of the examples. All words in the texts except proper nouns start with a lower case letter. These sets are compiled from all the texts in the corpus (100 texts). The files contain words and tags.

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

Collection of TEI-conformant XML-files.

1.3. Character encoding

The characters are UTF8 encoded.

2. ADMINISTRATIVE INFORMATION

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

Name: Sigrún Helgadóttir
Affiliation: The Arni Magnusson Institute for Icelandic Studies
Adress: Neshagi 16, 107, Reykjavík, Iceland
E-mail: sigruhel@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

Freely open for search, own license needed for download.

3. TECHNICAL INFORMATION

3.1. Directories and files

44 TEI conformant xml-files for download.

3.2. Data structure of an entry

The corpus is provided for download as 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.

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

About 300 thousand tokens for download, about 550 thousand for online search.

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*

Fiction both for adults and children, biographies, informative writings. The search interface will also provide access to translated fiction. Only texts written originally in Icelandic will be available for download.

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)*

Each file (text) has a header containing bibliographic data. Text is segmented into sentences and sentences into words. Each word is assigned a morphosyntactic tag (MSD), and lemmata.

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

The texts in the corpus are annotated with morphosyntactic tags and lemmata. The tagset was developed for this corpus that was originally used to make a Frequency Dictionary for Icelandic. (Pind et al., 1991). The corpus was part-of-speech tagged by semi-automatic means, all morphosyntactic tags and lemmas were manually corrected.

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)*
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
The corpus has been used to train PoS taggers for Icelandic. The online search will be useful for teaching purposes, both for Icelanders and foreigners. This will give guidance on language use and morphosyntactic analysis. 2) The downloadable corpus will be useful for various LT projects. 3) The ten disjoint pairs will be used for training PoS taggers.

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

5. RELEVANT REFERENCES AND OTHER INFORMATION

Jörgen Pind (ed.), Friðrik Magnússon and Stefán Briem. 1991. *Íslensk orðtíðnibók*. Orðabók Háskólans, Reykjavík. [Referred to as the Icelandic Frequency Dictionary, IFD.]

2004h. Sigrún Helgadóttir. Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic. In H. Holmboe (ed.), *Nordisk Sprogteknologi 2004*. Museum Tusculanums Forlag.

Sigrún Helgadóttir. Mörkun íslensks texta *Orð og tunga* 9:75-107. Reykjavík. 2007.

HRAFN LOFTSSON. 2006. Tagging a morphologically complex language using heuristics. In T. Salakoski, F. Ginter, S. Pyysalo and T. Pahikkala (eds.), *Advances in Natural Language Processing, 5th International Conference on NLP, FinTAL 2006, Proceedings*. Turku, Finland.

HRAFN LOFTSSON. 2007. Tagging Icelandic Text using a Linguistic and a Statistical Tagger. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the ACL*. Rochester, NY, USA.

HRAFN LOFTSSON. 2009. Correcting a POS-Tagged Corpus Using Three Complementary Methods. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Athens, Greece.