The Saga Corpus

1. BASIC INFORMATION

1.1. Corpus composition
Texts from 44 Sagas.

1.2. DanNet

1.3. Representation of the corpora (flat files, database, markup)
Collection of TEI-conformant XML-files.

1.4. Character encoding
The characters are UTF8 encoded.

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

3. TECHNICAL INFORMATION

3.1. Directories and files
44 TEI conformant xml-files.

3.2. Data structure of an entry
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)
1,659,385 tokens.

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*
The language of the Icelandic Sagas, transliterated to modern spelling.

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 automatically annotated with morphosyntactic tags and lemmata. The tagset was developed for the IFD Corpus (Pind et al., 1991). Morphosyntactic tags and lemmas were not 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

For researchers to study Old Icelandic. The corpus will be available for download and web search.

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

Annotations are automatically assigned and not manually checked. Tagging accuracy has been estimated as 92.7% (Rögnvaldsson and Helgadóttir, 2011).

5. RELEVANT REFERENCES AND OTHER INFORMATION

Eiríkur Rögnvaldsson and Sigrún Helgadóttir. 2011. Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change. Sporleder, Caroline, Antal P.J. van den Bosch og Kalliopi A. Zervanou (ritstj:): *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, s. 63-76. Springer, Berlín.