The Balanced Tagged Corpus of Icelandic

Sigrún Helgadóttir

The Árni Magnússon Institute for Icelandic Studies

sigruhel@hi.is

Nordic Language Variation:

Grammatical, Sociolinguistic and Infrastructural Perspectives

Institute of Linguistics, University of Iceland, Reykjavík

7-9 October 2010

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM

The Balanced Tagged Corpus of Icelandic
The texts

A balanced morphosyntactically tagged corpus of 25 million words

Both written and spoken text (transcribed speech)

Texts written during the period 2000–2009 by native speakers of Icelandic

No translated texts were sampled

Texts from published books are shortened by about 20%

Other texts are included unabridged

Only electronically available texts were sampled

The Balanced Tagged Corpus of Icelandic
Copyright issues

Permission was sought for the use of all texts

Copyright holders were contacted with a draft of an end-user licence, declaration to be signed and information about the project

Contact through e-mail and letters to copyright owners of printed books

One reminder was sent

Agreement with the Icelandic Publishers Association about supplying the texts

Divison of texts in the Tagged Icelandic Corpus	%
Blog	7.6
Printed periodicals	36.6
Adjudications	1.2
Printed books	22,3
Web media	8.3
Websites	12.6
Radio news scripts	1.1
Law	2.9
Written-to-be-spoken	0.9
Student essays	2.6
Miscellaneous	0.8
Spoken	3.2
Total	100.0

The Balanced Tagged Corpus of Icelandic

Text cleaning

Texts come in various formats – “clean” text was extracted from all formats

Text from columnar pdf-files was rearranged

Texts cleaned by automatic methods and manually

Many difficult characters had to be handled, e.g. quotation marks -> « and »

In some texts headings had to be identified

…….

Texts will be delivered in codepage UTF8

The Balanced Tagged Corpus of Icelandic

Tokenization, sentence segmentation, tagging, lemmatization

Tokenization and sentence segmentation is performed with a sentence segmentizer and tokenizer which are a part of the IceNLP toolkit (Hrafn Loftsson)

Combined tagging using *CombiTagger:*

Individual taggers (in descending order of accuracy when tagging Icelandic text): *IceTagge*r (part of the IceNLP toolkit), *Bidir* (Dredze and Wallenberg), *TnT* (Brants), *fnTBL* (Ngai and Florian) and *MXPOST* (Ratnaparkhi)

Accuracy has been estimated at 88-95%, depending on text type

Lemmatization is performed with a lemmatizer for Icelandic developed by a member of our LT-team (Anton K. Ingason)

Not much known about accuracy, estimated at 90% for one text type (newspaper Morgunblaðið)

The Balanced Tagged Corpus of Icelandic

Availability of the corpus

The corpus will be made available in two forms:

1. Search interface on the website of the Arni Magnusson Institute for Icelandic Studies (demonstration to follow)
2. The corpus files will be made available as XML-files. The corpus will be encoded according to the Guidelines of the Text Encoding Initiative (TEI) to represent both tags and lemmas. Full classification and contextual and bibliographic information is also included with each text in the form of a TEI-conformant header. Users will be able to obtain these files against signing an end-user license.

The Balanced Tagged Corpus of Icelandic

Search Interface

The search interface is based on Glossa (http://www.hf.uio.no/tekstlab/glossa.html) from the University of Oslo

Glossa uses the corpus search engine IMS Corpus Workbench (CWB) from the University of Stuttgart (http://cwb.sourceforge.net/)

Preliminary search interface is available:

- Search options have been adjusted to the tagging of Icelandic
- All labelling used in the web interface has been translated
- Bibliographic information will be available later
- One million words of text available for search (text from www.visindavefur.is)