A combined presentation and demonstration of the INESS treebanking infrastructure

Who? Petter Haugereid
Department of Linguistic, Literary and Aesthetic Studies
University of Bergen

When? DELPH-IN meeting
St. Wendel, July 29 – August 2, 2013

INESS – *Infrastructure for the Exploration of Syntax and Semantics*

- Carried out by
  - The University of Bergen (Norway)
    - Helge Dyvik, Victoria Rosén, Koenraad De Smedt, Gunn Inger Lyse , Gyri Smørdal Losnegaard , Martha Thunes, Petter Haugereid
  - Uni Computing (a division of Uni Research, Bergen)
    - Paul Meurer
- Funded by
  - The Research Council of Norway
  - The University of Bergen
- Duration: 2010–2016

INESS – Main activities

Main activities:

- The development of a large, deep LFG parsebank for Norwegian
- Implementation and operation of a comprehensive open treebanking environment for building, hosting and exploring treebanks

Hosting treebanks

- INESS currently provides the most comprehensive web-based treebanking services available
- A normal browser sufficient for:
  - accessing, searching and downloading treebanks
  - annotation of LFG-based parsebanks:
    - computer-aided manual disambiguation
    - text cleanup
    - handling of unknown words

Treebanks hosted by INESS

- The Icelandic Parsed Historical Corpus (IcePaHC)
  - 73,014 sentences
- The German Tiger treebank
  - 50,472 sentences with dependency annotation
  - 9,221 with LFG annotation
- The dependency part of the Bulgarian BulTreeBank
  - 11,900 sentences
- HPSG treebanks
  - WeScience: 48,000 sentences
  - DeepBank: 37,000 sentences
- and other smaller corpora

INESS – Parallel corpora

The INESS infrastructure offers tools for the development and exploitation of parallel corpora

- Excerpts of the novel *Sofies verden* (*Sophie's World*)
  - 26 aligned language pairs
- A document from the *Acquis communautaire* (EU legislation)
  - 21 aligned language pairs

The INESS Norwegian treebank – Development

- Obtained by parsing automatically with the LFG grammar NorGram on the XLE parsing platform
  - `http://iness.uib.no/redmine/projects/inesspublic/wiki/NorGram_documentation`
- Consists of a number of different text types in fiction and non-fiction
- Part of the treebank is manually disambiguated with the LFG Parsebanker (Rosén et al., 2009)

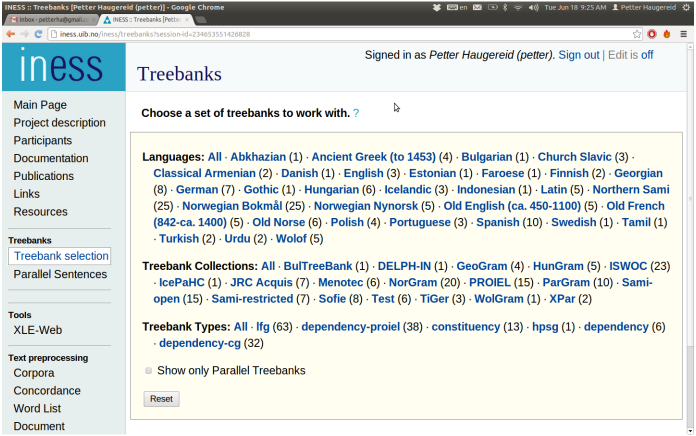
The INESS Norwegian treebank – Data

- Manually (at least partially) disambiguated:
  - 8,405 sentences (81,934 words)
- Fully disambiguated:
  - 6,966 sentences (65,599 words)

Stochastic disambiguator

- A stochastic disambiguator is implemented based on the manually disambiguated sentences
- Used to rank the analyses of a given sentence
- Corpus searches are limited to the top ranked analyses in the parsebank
- The manually selected discriminants are saved and reused in case the corpus is reparsed

Treebank selection page



INESS :: Treebanks [Petter Haugereid (petter)] - Google Chrome
en
Tue Jun 18 9:25 AM
Petter Haugereid
Inbox - petterha@gmail.cc
INESS :: Treebanks [Petter
iness.uib.no/iness/treebanks?session-id=234653551426828
iness

Signed in as *Petter Haugereid (petter)*. Sign out | Edit is off

Treebanks

Main Page
Project description
Participants
Documentation
Publications
Links
Resources

Treebanks
Treebank selection
Parallel Sentences

Tools
XLE-Web

Text preprocessing
Corpora
Concordance
Word List
Document

Choose a set of treebanks to work with. ?

Languages: All · Abkhazian (1) · Ancient Greek (to 1453) (4) · Bulgarian (1) · Church Slavic (3) · Classical Armenian (2) · Danish (1) · English (3) · Estonian (1) · Faroese (1) · Finnish (2) · Georgian (8) · German (7) · Gothic (1) · Hungarian (6) · Icelandic (3) · Indonesian (1) · Latin (5) · Northern Sami (25) · Norwegian Bokmål (25) · Norwegian Nynorsk (5) · Old English (ca. 450-1100) (5) · Old French (842-ca. 1400) (5) · Old Norse (6) · Polish (4) · Portuguese (3) · Spanish (10) · Swedish (1) · Tamil (1) · Turkish (2) · Urdu (2) · Wolof (5)

Treebank Collections: All · BulTreeBank (1) · DELPH-IN (1) · GeoGram (4) · HunGram (5) · ISWOC (23) · IcePaHC (1) · JRC Acquis (7) · Menotec (6) · NorGram (20) · PROIEL (15) · ParGram (10) · Sami-open (15) · Sami-restricted (7) · Sofie (8) · Test (6) · TiGer (3) · WolGram (1) · XPar (2)

Treebank Types: All · lfg (63) · dependency-proiel (38) · constituency (13) · hpsg (1) · dependency (6) · dependency-cg (32)

□ Show only Parallel Treebanks

Reset

DEMO

http://iness.uib.no/iness/main-page