A new treebank interface for the digital humanities

Tinna Frímann Jökulsdóttir
Anton Karl Ingason

UNIVERSITY OF ICELAND

Overview

1. Introduction
2. PaCQL
3. The software
   - The search engine
   - The web interface
4. Example
   - How to use
   - Processing the results
5. Next up

Introduction

- IcePaHC
  - The Icelandic Parsed Historical Corpus.
  - Contains just over a million manually corrected words.
  - Icelandic texts from every century between the 12th and the 21st centuries inclusive.
  - The annotation scheme mostly compatible with the one used in the Penn Parsed Corpora of Historical English.

- But how can we efficiently utilize those treebanks?

Introduction

- Why not use existing software?
  - Tgrep
  - Corpus Search
  - NLTK (Python)
  - And so on…
- Pros and cons everywhere – we are trying to collect and maximize the pros.
- Especially those that are useful in linguistic research, notably in historical syntax.

PaCQL

- The basic syntactic relationships:
  - **`idoms`**: immediately dominates
  - **`idomsonly`**: immediately dominates x and nothing else
  - **`idomsfirst`**: immediately dominates the leftmost child x
  - **`idomslast`**: immediately dominates the rightmost child x
  - **`doms`**: dominates at an arbitrary depth
  - **`sprec`**: sisterwise precedence
  - **`precedes`**: precedence regardless of embedding
  - **`hassister`**: sisterhood
  - **`sameindex`**: A has the same index as B

PaCQL

- The special relationships:
  - **`haslabel`**: match node label
  - **`domswords`**: match nodes dominating N orthographic words
  - **`domswords<`**: match nodes dominating less than N words
  - **`domswords>`**: match nodes dominating more than N words
  - **`idomslemma`**: POS-tag has child that has a specific lemma

PaCQL

- Text level metacoding:
  - **`text textid:`** id of the text
  - **`text year:`** (estimated) year the text was written
  - **`text century:`** century the text was written
  - **`text genre:`** main genre of the text
  - **`text subgenre:`** subgenre of the text
  - **`text postnt:`** 0 if written before New Testament translation, 1 otherwise
  - **`text texttrees:`** total number of trees in the text
  - **`text meantreewords:`** mean number of words per tree in the text
  - **`text mediantreewords:`** median number of words per tree in the text
  - **`text meanwordletters:`** mean number of letters per word in the text
  - **`text lexicaldiversity:`** type frequency of word forms divided by the totalnumber of words in the text

PaCQL

- Tree level meta coding:
  - **`tree treeid:`** unique id for the tree
  - **`tree treewords:`** number of words in the tree

- Node level meta coding:
  - **`node label A:`** the label matched by A
  - **`node nodestring A:`** the string of leafs dominated by A
  - **`node nodewords A:`** the number of words dominated by A

The software

- The search engine: Python.
- Employs a fast in-memory index that cuts down waiting time.
- The server: Pyro 4
- The web interface:

www.treebankstudio.org

Example

- The evolution from object-verb (OV) word order to the verb-object (VO) order in Icelandic.
- An example in English:
  a. She will [the bread eat] - OV
  b. She will [eat the bread] - VO
- The same example in Icelandic:
  a. Hún mun [brauðið borða] - OV
  b. Hún mun [borða brauðið] - VO

Example

- Tutorial (www.treebankstudio.org)
  - Documentation
  - Syntax
  - Results
  - Processing the resaults
  - Etc.

Next up

- Make the system available to the users of other treebanks.
- Release the PaCQL search engine under a free and open source software license.
- The output:
  - Offer more visualized and interactive output types.
  - Provide tools for more sophisticated analysis that now is dependent on other software, like R or Excel.

Next up

- The query language:
  - Negation.
  - Equality testing across matched nodes and quantification.
- The user interface:
  - Re-design with software quality metrics in mind.
  - More visual input possibilities (drag&drop).
- The search engine:
  - Of course, we are always seeking to improve the speed.

...et cetera.

References

Further information about the project and references can be found here:

Ingason, A. K. (2016). PaCQL: A new type of treebank search for the digital humanities. *Italian Journal of Computational Linguistics, 2*(2), 51-66.

IcePaHC can be accessed here:

Wallenberg, J. C., Ingason, A. K., Sigurðsson, E. F., & Rögnvaldsson, E. (2011). Icelandic Parsed Historical Corpus (IcePaHC). Version 0.9. http://www.linguist.is/icelandic_treebank