Annotating and querying the Icelandic Parsed Historical Corpus and closely related cross-linguistic counterparts

Anton Karl Ingason
University of Iceland

www.linguist.is

Outline

- Introduction to the Icelandic Parsed Historical Corpus (IcePaHC)
- Our annotation process and software.
- PaCQL query language and online search engine
  - A new type of treebank search for the Digital Humanities.
  - Ingason, A. K. (2016). PaCQL: A new type of treebank search for the digital humanities. *Italian Journal of Computational Linguistics*, 2(2), 51-66.
  - Google or look up on: **www.linguist.is/papers**

Introduction to IcePaHC

- IcePaHC is a treebank, annoted according to the annotation scheme of the Penn Parsed Corpora of Historical English (for quantitative diachronic syntax)
  - Phrase structure annotation. A growing family of similar treebanks.
  - Minimum changes for Icelandic-specific properties.
  - Often the same unmodified query works well across treebanks in this tradition.
- Joel Wallenberg, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson and Anton Karl Ingason.
- Ca. 1.000.000 words of manually corrected parses.
- Spans the period 12th-21st centuries
  - All those centuries are included.
  - Includes narratives and religious texts from throughout this period.
- All raw data freely available under an open source license.
  - The annotation itself was carried out in an open Github repository.

Example tree

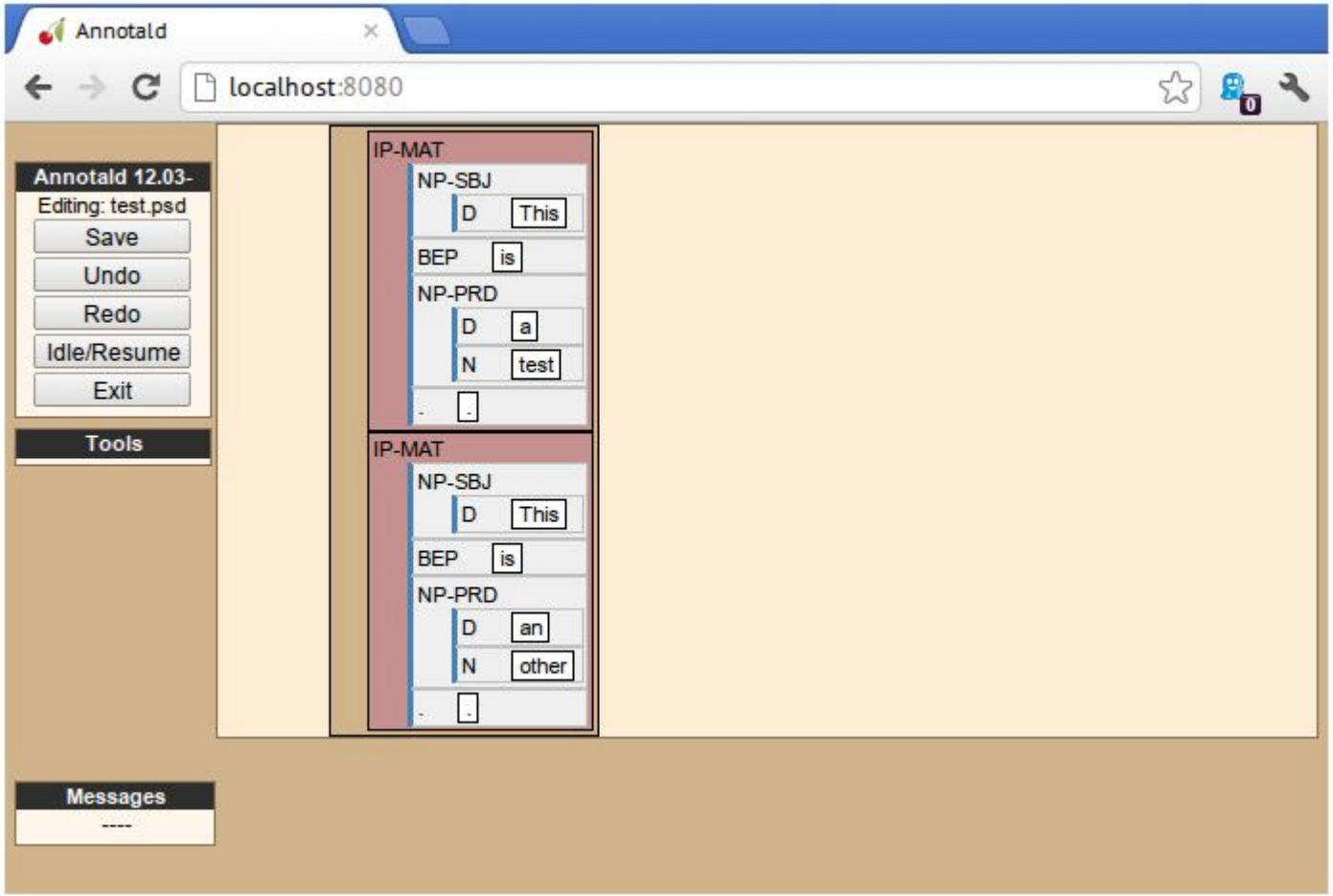
- Format: Labeled bracketing, UTF-8 plain text.
- Documentation: http://www.linguist.is/icelandic_treebank/

```
( (IP-MAT (NP-SBJ (PRO-N hann-hann))
          (VBDI tók-taka)
          (NP-OB1 (PRO-D okkur-ég))
          (ADVP (ADV smám-smám) (ADV saman-saman))
          (RP inn-inn)
          (PP (P í-í)
              (NP (NP-POS (PRO-A sinn-sinn))
                  (ADJS-A innsta-innri)
                  (N-A hring-hringur))))
  (ID 2008.OFSI.NAR-SAG,.16))
```

Annotald annotation software

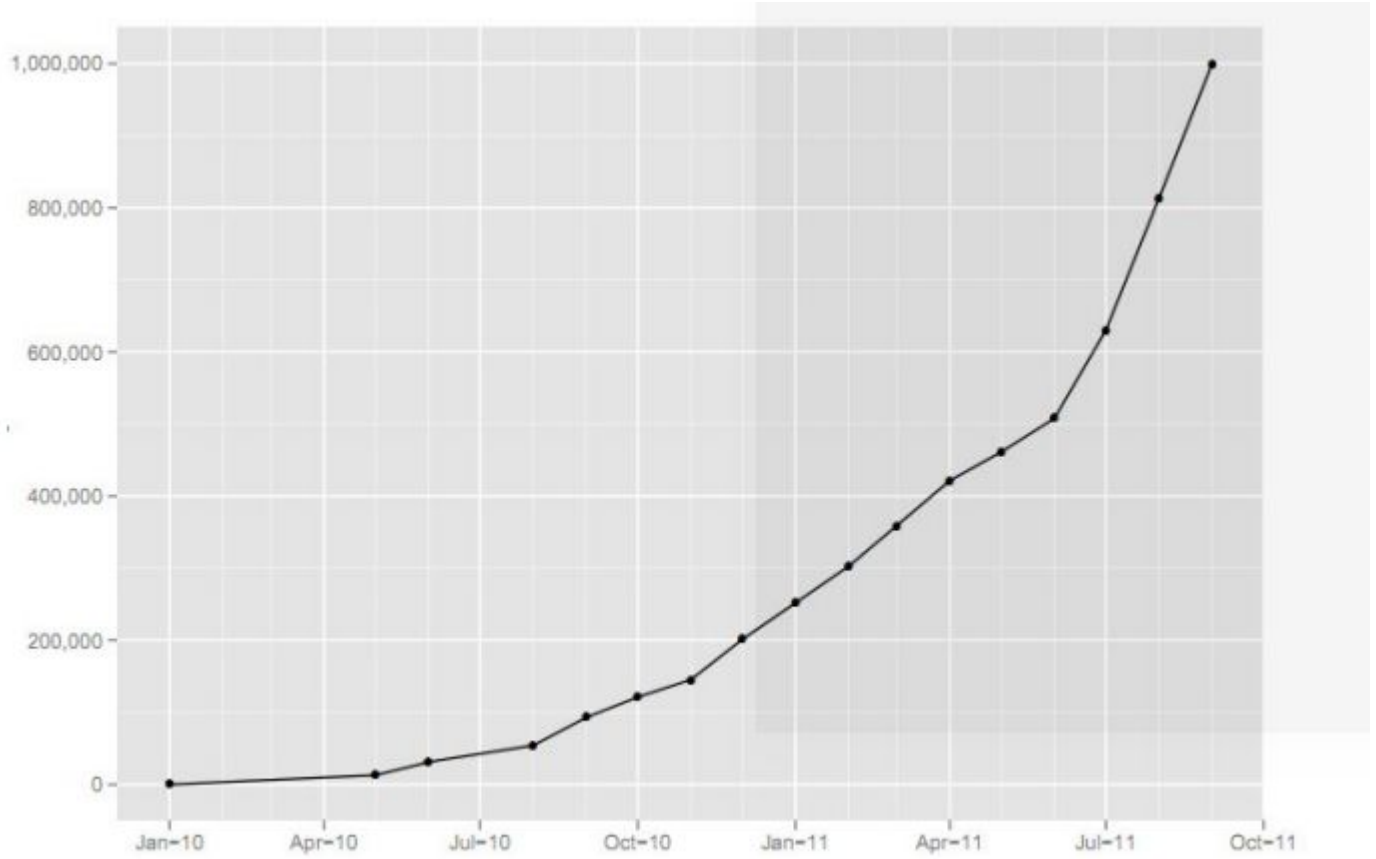
- Website: https://annotald.github.io/
- Annotald was originally developed by AKI as part of the IcePaHC project but has since been improved considerably and maintained by Aaron Ecay.
- We initially used software that displayed trees like trees and had a more traditional graphical user interface.
  - This turned out to slow us down so we wrote our own system.
- Design:
  - The hierarchy extends from left to right (not top down).
  - Left hand never leaves the keyboard. All shortcuts are on the left side of the keyboard.
  - Right hand never leaves the mouse. The mouse is used to select and move things.
- License: GPL. Code available on Github.

Screenshot


Annotald
localhost:8080
Annotald 12.03-
Editing: test.psd
Save
Undo
Redo
Idle/Resume
Exit
Tools
Messages

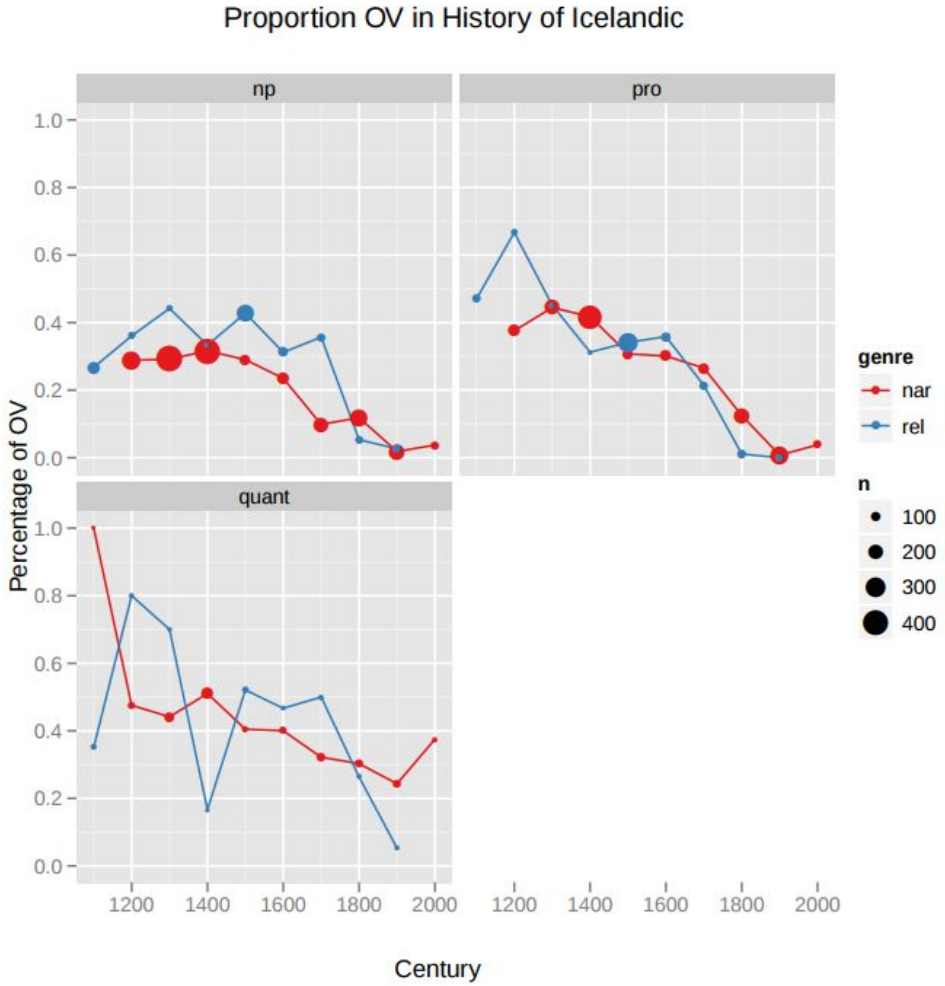
IP-MAT
NP-SBJ
D This
BEP is
NP-PRD
D a
N test
. .
IP-MAT
NP-SBJ
D This
BEP is
NP-PRD
D an
N other
. .

Annotation speed



PaCQL - Parsed Corpus Query Language

- Most recent addition to our tools.
- Why not use existing tools?
  - There are many useful tools out there that you should use if you like them.
  - We wanted the right combination of a **fast indexed search engine** and **powerful coding queries** as typically used in quantitative diachronic syntax.
  - The language should make sense to historical syntacticians -- the way CorpusSearch does.
- Emphasis on output for syntacticians when using web search:
  - Practical visual features (color coding etc.)
  - Can download coding results as a .tsv file (for R, SPSS, Excel, ...)
  - Automatic plotting of the dependent variable over time.
  - Summary reports per centuries and per individual texts.


Proportion OV in History of Icelandic
np
pro
quant
Percentage of OV
1.0
0.8
0.6
0.4
0.2
0.0
1200
1400
1600
1800
2000
Century
genre
nar
rel
n
100
200
300
400

PaCQL - 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 - 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 - 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 meta coding

- **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 is written in Python
- Fast in-memory index cuts down waiting time.
- Server: Pyro 4
- Web interface (uses Django/JQuery etc.):
  - www.treebankstudio.org

Example

- Evolution from object-verb (OV) to verb-object (VO) word order in Icelandic.

(1) a. She will the bread eat. (OV)

b. She will eat the bread. (VO)

See **treebankstudio.org**:

- Documentation
- Syntax
- Results (export to .tsv for R/SPSS/Excel etc.)
- Summary reports
- Stability

Plans

- Make the system available to the users of other treebanks.
  - Let us know if you are interested!
- 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.
- More advanced search functionality.
- Improve user interface.