Historical Treebanks

The Penn Historical Corpora and the Icelandic Historical Parsed Corpus

The Penn Historical Corpora

- Consist of:

  - the Penn-Helsinki Parsed Corpus of Middle English, 2nd edition (PPCME2) (1150-1500)

  - the Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME) (1500-1710)

  - the Penn Parsed Corpus of Modern British English (PPCMBE) (1700-1914)

  - the Parsed Corpus of Early English Correspondence (PCEEC)

People

Tony Kroch

(Beatrice) Santorini

And Ann Taylor, Susan Pintzuk, the people behind the Helsinki corpus among others

Icelandic Parsed Historical Corpus (IcePaHC)

Wallenberg, Joel C., Anton Karl Ingason, Einar Freyr Sigurðsson and Eiríkur Rögnvaldsson. 2011.

Version 0.5. http://www.linguist.is/icelandic_treebank

Joel

Anton

IcePaHC

- Guidelines are based on and supplement the Penn historical corpora guidelines
- Texts range in time from the 12th century to modern times
- Fewer really old texts; these are covered in full. Later texts are sampled partially.
- Begins with: Fyrsta málfræðiritgerðin (The first grammatical treatise) from the 12th century

Philosophy and Goals 1

- to create an annotation system that facilitates automated searches, not to give a correct linguistic analysis of each sentence.

- if a construction can be found unambiguously through a combination of properties of a bracketed sentence, our annotation may not contain all of the structure that a full phrase structure diagram of the sentence would have.

Philosophy and Goals 2

- information is to be added in a monotonic way.
- future revisions of the bracketed structures should always add information, never change it.
- Hence avoid subjective judgments since they are extremely error-prone:

  - no distinguishing adjectival from verbal passive participles

  - no argument-adjunct distinction.

Philosophy and Goals 3

- As many categories as possible should have clear meanings so that unclear cases can be relegated to a small number of categories of residual cases.
- The price of making most categories homogeneous is that these residual categories will not be.
- Future revisions of the corpus may make it possible to divide some of these residual categories into homogeneous subcategories.

Philosophy and Goals 4

- avoid making decisions that would be controversial, whether with regard to text interpretation or to linguistic theory.
- In doubtful cases, either avoid specifying structure, or use default rules to decide the case for search purposes.

  - VPs are normally not indicated in the corpus, since VP boundaries are normally indeterminate.

  - PP attachment. Whenever it is unclear where a PP attaches, attach it by default as high as possible.

Icelandic and English treebanks

- The Icelandic treebank guidelines try to hew to the Penn Historical Treebank guidelines and overall decisions concerning the organization of the tree bank, with appropriate cross-linguistic diversions.

- This allows for an easy way to identify and document crosslinguistic comparisons.

Layout

Each text in the corpus comes in three different formats, each with a characteristic filename extension:

- text (.txt)
- part-of-speech (POS) tagged (.pos)
- parsed (.psd)

The .txt file

<P_2>
<heading>
I . (CMMALORY,2.3)

Merlin (CMMALORY,2.4)
</heading>

HIT befel in the dayes of Uther Pendragon , when he was kynge of all Englond and so regned , that there was a myghty duke in Cornewaill that helde warre ageynst hym long tyme . (CMMALORY,2.6)

and the duke was called the duke of Tyntagil . (CMMALORY,2.7)

And so by meanes kynge Uther send for this duk chargyng hym to brynge his wyf with hym . (CMMALORY,2.8)

The .pos file

<P_2>_CODE

<heading>_CODE
I_NUM ._. CMMALORY,2.3_ID

Merlin_NPR CMMALORY,2.4_ID
</heading>_CODE

HIT_PRO befel_VBD in_P the_D dayes_NS of_P Uther_NPR Pendragon_NPR ,_, when_P he_PRO was_BED kynge_N of_P all_Q Englond_NPR and_CONJ so_ADV regned_VBD ,_, that_C there_EX was_BED a_D myghty_ADJ duke_N in_P Cornewaill_NPR that_C helde_VBD warre_N ageynst_P hym_PRO long_ADJ tyme_N ,_. CMMALORY,2.6_ID

and_CONJ the_D duke_N was_BED called_VAN the_D duke_N of_P Tyntagil_NPR ._. CMMALORY,2.7_ID

And_CONJ so_ADV by_P meanes_NS kynge_NPR Uther_NPR send_VBD for_P this_D duk_N chargyng_VAG hym_PRO to_TO brynge_VB his_PRO$ wyf_N with_P hym_PRO ,_. CMMALORY,2.8_ID

The .psd file

Parsed have the extension .psd. Each token is enclosed with its ID in a set of unlabelled parentheses.

```
( (CODE <P_2>))
( (CODE <heading>))
( (NUMP (NUM I)
      (. .))
  (ID CMMALORY,2.3))

( (NP (NPR Merlin))
  (ID CMMALORY,2.4))

( (CODE </heading>))
( (IP-MAT (CONJ and)
        (NP-SBJ-1 (D the) (N duke))
        (BED was)
        (VAN called)
        (IP-SMC (NP-SBJ *-1)
              (NP-OB1 (D the) (N duke)
                    (PP (P of)
                       (NP (NPR Tyntagil)))))
        (. .))
  (ID CMMALORY,2.7))
```

Tags and Dash Tags

- Tags: ADJP, ADVP, CP, FOREIGN, IP, NP, NUMP, PP, QP, W*P
- Dash Tags:

  CP-CLF, CP-DEG, CP-EOP, CP-EXL, CP-QUE, CP-REL, CP-THT, CP-TMC

  IP-ABS, IP-INF, IP-MAT, IP-PPL, IP-SMC, IP-SUB

  NP-OB1, NP-OB2, NP-SBJ, NP-VOC, NP-TMP

Empty Categories

- 0 – empty operator
- *arb* - arbitrary PRO
- *con* - subject elided under conjunction
- *exp* - expletive subject
- *pro* - pro subject
- *ICH* - trace of movement that's not A or A'
- *T* - trace of A-bar movement
- * - trace of A-movement

_# - indicates co-indexation between XP and empty categories

English vs. Icelandic

- Case information is not marked for the most part in English.
- Case information is represented explicitly in Icelandic at the word level but not at the phrase-level:

(NP-SBJ (PRO-D þér-þú))

- Case information is marked on nouns, determiners, adjectives and participial verbs.

CorpusSearch

http://corpussearch.sourceforge.net/

- a Java program for searching annotated corpora
- find and count lexical and syntactic configurations of any complexity
- can also be used for corpus development
- uses syntactic annotation in Penn-Treebank format

CorpusSearch

The Penn Historical Corpora and IcePaHC bundle together CorpusSearch.

There is also a web-interface that comes with the DIGS_WORKSHOP demo.

CorpusSearch

node: IP-SUB
query: IP-SUB idoms NP-OB*

NP-OB* matches anything that begins with NP-OB.

node: IP*
query: (IP* idoms NP-SBJ) AND (NP-SBJ idoms *T*)

Traces are marked by * (e.g. *T*) but * is a special character and hence must be `escaped' by \.

CorpusSearch

Naming in CorpusSearch: search patterns are treated like names e.g. if you re-use NP*, then all uses refer to the same element.

query: (IP* idoms NP*) AND (NP* idoms D)

node: IP*

query: (IP* idoms NP-OB*)

AND (IP* idoms NP-SBJ)

AND (NP-SBJ precedes NP-OB*)

CorpusSearch

Naming nodes:

node: IP*

query: (IP* idoms [1]NP-*)

AND (IP* idoms [2]NP-*)

AND ([1]NP-* precedes [2]NP-*)

CorpusSearch

Negation in CorpusSearch: !

- added after relation symbol

node: IP*

query: IP* idoms V*

AND V* iPrecedes !NP-OB1

means V* does not immediately precede NP-OB1 (and precedes something else).

node: IP-SUB

query: IP-SUB idoms !NP-OB*

Case Studies

- Historical Stability of Dative Subjects in Icelandic (Ingason, Wallenberg & Sigurdsson)

- The analysis of Heavy NP shift and Auxiliary contraction (Ingason & MacKenzie)