PROGRESS IN POS TAGGING THE CEECE

(CORPUS OF EARLY ENGLISH CORRESPONDENCE EXTENSION)

TANJA SÄILY

UNIVERSITY OF HELSINKI
PARSED CEEC (1410?-1681)
• Letters mainly written in 1680–1800, c. 2.2 million words
• Standardised-spelling version: Hakala, Palander-Collin & Nevala 2012
• Part-of-speech (POS) tagging clearly of interest (e.g. Säily et al. 2011) → a (non-funded) tagging project set up in 2013
  • Team: Terttu Nevalainen, Tanja Säily, Mikko Hakala
  • Assistants: Emanuela Costea, Anne Kingma
• Which tagset to use?
  • Penn Treebank → comparability with PCEEC, Penn historical corpora
  • CLAWS → comparability with present-day English corpora
but _CONJ continue _VB when _P My _PRO$ youth _N likewise _ADJ+N and _CONJ Greatest _ADJS vigour _N is _BEP past _VBN , __.

• Problems:
  • Accuracy of trained tagger c. 80–90% (PCEEC) → extensive manual post-editing needed
  • POS tagging seen merely as a necessary step before parsing
  • Goal “to create an annotation system that facilitates automated searches, not to give a correct linguistic analysis of each sentence” (Taylor & Santorini 2006)
  • Grammaticalisation, long diachrony → many adverbs conservatively tagged as nouns (likewise _ADJ+N = gentleman _ADJ+N), etc.
but _CCB continue _VV0 when _RRQ My _APPGE youth _NN1 likewise _RR and _CC Greatest _JJT vigour _NN1 is _VBZ past _RL , _.

• Based on present-day English (likewise _RR ≠ gentleman _NN1)
• Accuracy on present-day English c. 96–97%
  • Also works on standardised-spelling EModE (Hiltunen & Tyrkkö 2013)
• Output available in many formats (horizontal, vertical, pseudo-XML)
• Able and willing collaborators: Paul Rayson, Turo Hiltunen
• Problems:
  • Cannot handle spelling variation, does not understand CEEC coding
All Sober people here are inclined & p=r=p~ing to go to West Gursey.

- Frequency cut-off in standardisation: rare types not standardised → many **abbreviations** left in text
- Superscript marked with `==` signs in the CEEC corpora, abbreviations with ~ (p^ping → p=r=p~ing → preparing)
- These need to be expanded prior to tagging
  - Searched for using WordSmith Tools, expanded in Excel (Emanuela),
    global search & replace in corpus files using a script (Turo)
PARAMETER CODING

<l austen_001>
<q c 1796? fn jausten>
<x jane austen>
<p 1>

- Header information at the beginning of each letter
- Format based on the Helsinki Corpus of English Texts
- Needs to be removed for tagging (Paul)
TEXT-LEVEL CODING

[^...^] Our comment, e.g. [^ADDRESS^]
[\...\] Editor’s comment, e.g. [\SECOND LEAF OF LETTER MISSING\\]
[{...{ Emendation, e.g. req[{uired{]
[]...}] Heading, e.g. [] [3. TO CASSANDRA AUSTEN TUESDAY 23 AUGUST 1796]] (Most headings are editorial, hence the double coding.)
(\...\) Foreign language, e.g. (\tete a tete\)
(^...^) Font other than basic (mostly italics), e.g. (^they^) do not know

• More difficult to remove for tagging and put back afterwards
  • Paul is working on this
CONCLUSION

Process of POS tagging:
1. Choose tagset & tagger
2. Prepare data for tagger (both manually + automatically)
3. Automatic tagging
4. Manual spot checks, estimate error rate

• We are in stage 2
• Hoping to complete the entire process by spring 2014
• Tagged corpus to be deposited in the Norwegian CLARIN repository?

UNIVERSITY OF HELSINKI
REFERENCES

• CLAWS part-of-speech tagger for English. http://ucrel.lancs.ac.uk/claws/
• WordSmith Tools by Mike Scott. http://www.lexically.net/wordsmith/
THANK YOU!

Thank _VV0 you _PPY !_!

0000001 002 ----------------------------------------------------
0000003 010 Thank 93 VV0
0000003 020 you 93 PPY
0000003 021 ! 03 !

<w id="2.1" pos="VV0">Thank</w>
<w id="2.2" pos="PPY">you</w>
<w id="2.3" pos="!">!</w>

UNIVERSITY OF HELSINKI