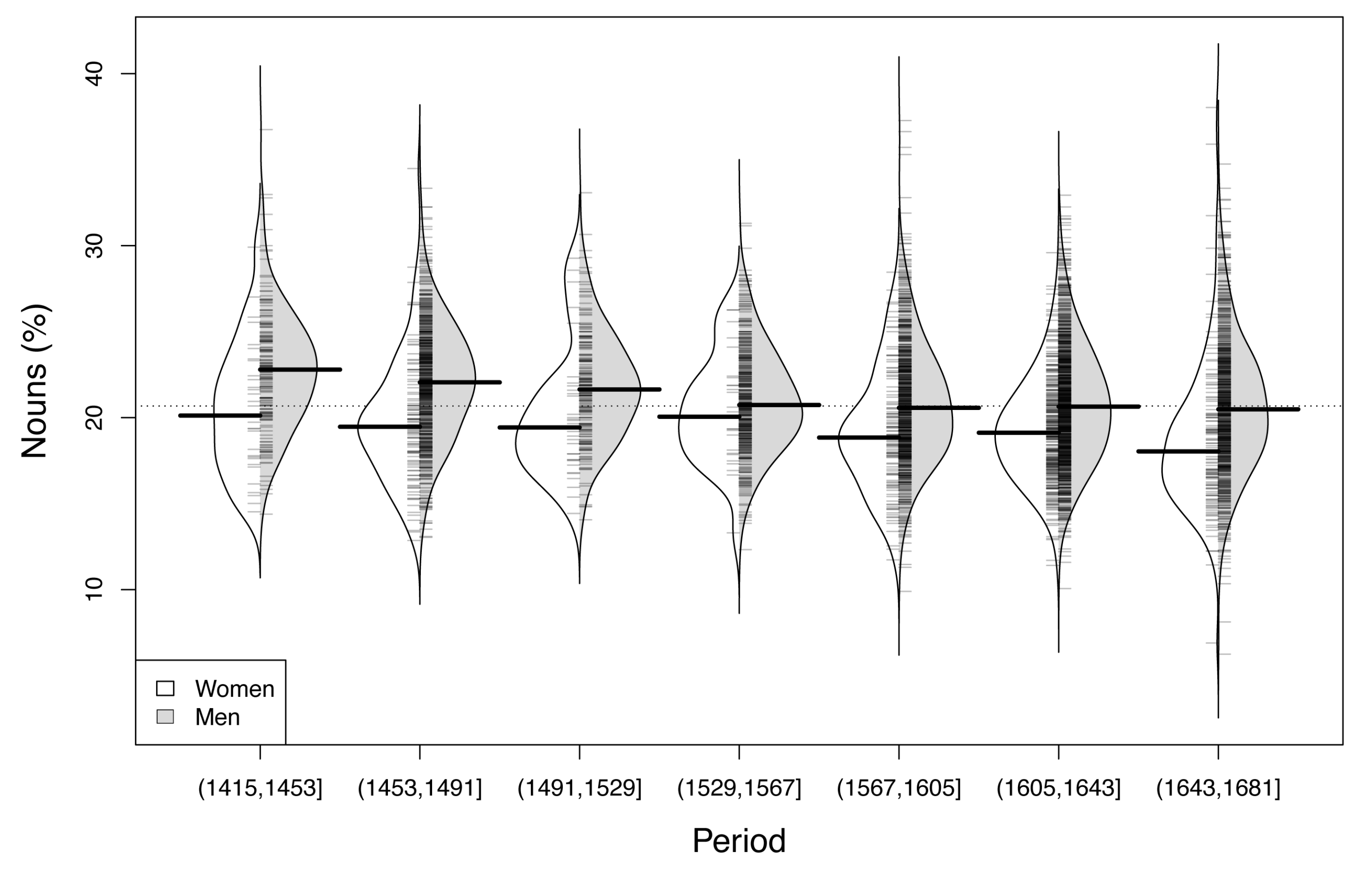
PROGRESS IN POS TAGGING THE CEECE

(CORPUS OF EARLY ENGLISH CORRESPONDENCE EXTENSION)

TANJA SÄILY

UNIVERSITY OF HELSINKI

PARSED CEEC (1410?-1681)



UNIVERSITY OF HELSINKI

CEEC EXTENSION

- 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

UNIVERSITY OF HELSINKI

PENN TREEBANK (BRILL TAGGER)

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.

CLAWS

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

STANDARDISED SPELLING?

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$^{r}\underline{p}$ing → 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

UNIVERSITY OF HELSINKI

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/
- Hiltunen, T. & J. Tyrkkö. 2013. “Tagging Early Modern English Medical Texts (1500–1700).” Presentation, CANS 2013, Lancaster, UK, 22 July.
- Säily, T., T. Nevalainen & H. Siirtola. 2011. “Variation in noun and pronoun frequencies in a sociohistorical corpus of English.” *Literary and Linguistic Computing* 26(2): 167–188.
- *Standardised-spelling Corpora of Early English Correspondence*. 2012. Compiled by T. Nevalainen, H. Raumolin-Brunberg, S. Kaislaniemi, J. Keränen, M. Laitinen, M. Nevala, A. Nurmi, M. Palander-Collin, T. Säily & A. Sairio. Standardised by M. Hakala, M. Palander-Collin & M. Nevala. Department of English / Department of Modern Languages, University of Helsinki. http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/
- Taylor, A. & B. Santorini. 2006. *The Parsed Corpus of Early English Correspondence*. http://www-users.york.ac.uk/~lang22/PCEEC-manual/
- WordSmith Tools by Mike Scott. http://www.lexically.net/wordsmith/

UNIVERSITY OF HELSINKI

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