HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

The CEEC corpora and their external databases

Samuli Kaislaniemi ***20 July 2006***
(***revised 7 August 2006***)

Research Unit for Variation, Contacts and Change in English (VARIENG)
University of Helsinki

The *Corpus of Early English Correspondence* (CEEC)

- The CEEC was designed to test the applicability of sociolinguistic methods on historical data
- The aim was to chart the influence of society on language historically
- Sociolinguistic factors taken into account would include gender, age, geography and social rank
- The desire was for socioregional and quantitative coverage: the aim was representativeness – or at least a balanced corpus

The *Corpus of Early English Correspondence* (CEEC) cont'd

- There were five requirements for the corpus:
  - Its size had to be sufficient for research on morphological variation and change
  - The social background of authors had to be recoverable
  - The texts had to represent private writing and the language had to be close to the spoken register
  - The sources had to be easily accessible and the material readily digitisable
  - The time period covered had to be sufficient for diachronic comparisons

The *Corpus of Early English Correspondence* (CEEC) cont'd

- Personal letters were deemed suitable material; published editions were seen to be convenient sources

- There were problems, but not insuperable ones:
  - Edited letters collections do not always record the authenticity of their sources, nor do they always retain original orthography etc.
  - The illiteracy of the lower ranks and lack of edited material would partly skew the structure of the corpus
  - Compiling a corpus from published sources would require asking for copyright permissions

The CEEC corpora 1: The original CEEC

- The *Corpus of Early English Correspondence* (CEEC or CEEC 1998)
  - c. 2.6 million words (c. 1410–1681)
  - Completed 1998

- Balanced corpus as regards –
  - Social representativeness (e.g. rank, gender, geography)
  - Content types of correspondence (e.g. news, love, business)
  - Relationships between correspondents (e.g. nuclear family, close friends, family servants, business)

The CEEC corpora 2: Extension & Supplement

- Work in progress (nearing completion)

- The *CEEC Extension* (CEECE)
  - Extends the CEEC to 1800
  - Aims to be a balanced corpus
  - c. 2.2 million words (1653–1800)

- The *CEEC Supplement* (CEECSup)
  - Includes material obtained since 1998
  - Fills socioregional gaps in the CEEC 1998: not a balanced corpus in itself
  - c. 440,000 words (1402–1663)

The CEEC corpora 3: Released corpora

- Two excerpts of CEEC 1998 have been released:
  - The *CEEC Sampler* (CEECS)
    - Contains the non-copyrighted material in CEEC 1998
    - 450,000 words (1418–1680)
    - Released 1999
  - The *Parsed Corpus of Early English Correspondence* (PCEEC)
    - 2.2 million words (c. 1410–1681) (copyright permission to some material was unattainable)
    - Released 2006
- Both are available through the Oxford Text Archive

Annotation of the CEEC corpora

- Based on the model of the *Helsinki Corpus of English Texts* (HC), with minor alterations
- The number of parameter codes decreased
  - e.g. text identifiers, author codes, page number
- Text-level codes largely retained
  - e.g. letter headings, editors' comments, marking of differing font or language

Annotation of PCEEC

- Further annotation jointly at the University of Helsinki and the University of York (cf. PCEEC manual)
  - More parameters coded
  - Part-of-speech (POS)-tagged and parsed
  - “Metadata”: searchable sociolinguistic information

- Example sentence:

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

- Example sentence POS-tagged:
  - `and_CONJ the_D duke_N was_BED called_VAN the_D duke_N of_P Tyntagil_NPR ._. CMMALORY,2.7_ID`

Annotation of PCEEC cont'd

- Example sentence parsed:
  - ```
    ( (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)))))
              (E_S .))
        (ID CMMALORY,2.7))
    ```
    ```

Searching the CEEC: textual issues

- Problem: great variety caused by spelling variation & change
  - The CEEC has not been lemmatized or normalized, and searches can be difficult to make
  - An example: how to spell “like” in CEECS:
    - leke, liche, lick, licke, lieke, lik, like, lych, lyck, lycke, lyk, lyke
- Problem: syntactic and semantic variation & change
  - A solution: PCEEC
    - Yet PCEEC was parsed largely manually
  - Further, CEECE and CEECSup remain unparsed

Searching the CEEC: socioregional issues

- Problem: how to search by social variables?
  - *Text Encoding Initiative* (TEI)-headers were felt to be too complex/cumbersome for search tools in 1994
  - HC-style text header of 25 COCOA-coded parameters partly redundant, partly cumbersome

- Solutions for CEEC 1998
  - Two searchable & sortable spreadsheet databases
    - Downside: corpus texts have to be searched separately
  - Division of the CEEC into “personal files” (for writers with more than 2000 words; the rest grouped by period)
    - Yet with 777 individuals, this divides the corpus in c. 250 files

The external databases of the CEEC

- Two separate databases:
  - “Senders database”
  - “Letter database”

- Sources of information:
  - The letter editions
  - The *Oxford Dictionary of National Biography*
  - Archives, registers, etc.
  - The internet (Google)
  - Books on regional/social history
  - Etc.

The CEEC Senders database

- A spreadsheet charting the social representiveness of the writers

- It contains writer-specific information:
  - last and first name
    title
    years of birth & death
    sex
    years of first & last letter
    rank; father's rank
    career details
    migration history
    extra information
    social mobility
    migrant (Y/N)
    education
    religion
    no. of letters in corpus
    no. and kind of recipients
    no. of words in corpus
    letter content types & quality
    name of collection

The CEEC letter database

- A separate spreadsheet charts the authenticity of each letter, and contains:
  - Letter-specific information:
    - date, year & place of writing
      letter authenticity
      sender/rec. relationship
      opening formula (Y/N)
      name of source edition
      no./pp. in edition
      no. of words
      closing formula (Y/N)
  - Information on the sender:
    name, lifespan, rank, father, migration, extra...
  - Information on the recipient (as above)

Present plans and developments

- For the CEEC 1998
  - Digitization of the letter database
  - Creation of a relational database architecture
    - Joining the corpus texts, senders database, and letter database into a freely searchable corpus
  - A project in data analysis:
    - Making the most of bad data: an ongoing data mining project with the *Helsinki Institute of Information Technology* (HIIT)

Present plans and developments cont'd

- For the CEEC *Extension & Supplement*
  - Completion of these (sub-)corpora & their databases
  - Creation of a relational database architecture like with CEEC 1998
  - Further annotation?
    - POS-tagging, parsing, pragmatic tagging?
    - Tools such as VARD would facilitate these projects
  - And finally, release
    - The biggest obstacle remains copyright issues: acquiring copyright for CEEC 1998/PCEEC took four years, and the permissions granted do not cover the entire corpus...

Sources

- CEECS = The Corpus of Early English Correspondence Sampler. 1998. Compiled by the CEEC Project Team: Terttu Nevalainen (leader), Jukka Keränen, Minna Nevala (née Aunio), Arja Nurmi, Minna Palander-Collin and Helena Raumolin-Brunberg. Helsinki: University of Helsinki. <http://khnt.hit.uib.no/icame/manuals/ceecs>.

- PCEEC = The Parsed Corpus of Early English Correspondence. 2006. Annotated by Ann Taylor, Arja Nurmi, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki. <www.eng.helsinki.fi/varieng/team2/1_2_4_projects.htm>

- Nevalainen, Terttu & Helena Raumolin-Brunberg (eds.). 1996. *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. London: Longman.

- Nevalainen, Terttu & Helena Raumolin-Brunberg. 2003. *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. London: Longman.

- Raumolin-Brunberg, Helena & Terttu Nevalainen. 2005. “Historical sociolinguistics: The Corpus of Early English Correspondence”. *Models and Methods in the Handling of Unconventional Digital Corpora vol. 2: Diachronic Corpora* ed. by J. C. Beal, K. Corrigan & H. Moisl. Palgrave.

Thank You!

- My email: samuli.kaislaniemi@helsinki.fi
- VARIENG: http://eng.helsinki.fi/varieng
- CEECS manual: http://khnt.hit.uib.no/icame/manuals/ceecs
- CEECS & PCEEC: *Oxford Text Archive*
  http://ota.ahds.ac.uk