NEOLOGISMS IN EARLY ENGLISH LETTERS: HOW TO FIND THEM AND WHAT THEY CAN REVEAL

Tanja Säily, Eetu Mäkelä & Mika Hämäläinen
INTRODUCTION
• *Interfacing structured and unstructured data in sociolinguistic research on language change* (Academy of Finland, DIGIHUM, 2016–2019)
  • [blogs.helsinki.fi/stratas-project/](blogs.helsinki.fi/stratas-project/)
• NATAS subproject: *Social embedding of neologisms in early English correspondence*
  • Previous research: mostly lexicographical data, bias towards well-known authors
  • *Corpora of Early English Correspondence (CEEC):* wide social spectrum, speech-like genre
CEEC

• Personal letters, c. 1400–1800
  • 1,180 writers, 11,713 letters, 5.2 million words
  • Compiled for historical sociolinguistics: metadata on letters, writers, recipients (e.g. gender, social rank)
• Compiled by T. Nevalainen, H. Raumolin-Brunberg et al. at the University of Helsinki
  • Based on published editions of letters
• SCEEC = Standardized-spelling version using VARD2 (excluding 15th century)

www.helsinki.fi/varieng/CoRD/corpora/CEEC/
1. **Who** are the innovators/early adopters? Which social groups do they represent?
2. **How do the new words spread** socially, geographically and diachronically?
3. **Which semantic domains** do the neologisms represent?
4. **Why** are the neologisms created and established? Can they be linked to:
   - Specific historical events?
   - Changes in culture & society?
   - Social meanings?
BIG-DATA APPROACH TO ANALYSING NEOLOGISMS

• Automatically map each word in the corpus to lexicographical data and contemporary published texts, compare first attestation dates
  • Spelling variation: SCEEC not enough, additional **normalization** required
• Automatic retrieval of related **lexicographical data**
  • *Oxford English Dictionary (OED)*, *Historical Thesaurus (HT)*, *Middle English Dictionary (MED)*
• Automatic retrieval of data from **databases of contemporary published texts**
  • *Early English Books Online (EEBO)*, *Eighteenth Century Collections Online (ECCO)*, *British Library Newspapers (BLN)*, *Burney & Nichols Collections (BN)*
• **Interface** for pruning the possible neologisms found, exploring social factors
CURRENT PIPELINE
INITIAL MAPPING TO OED

- **Prepare corpus**: convert to Unicode, remove most punctuation, tokenize
- Attempt to **lemmatize** with NLTK (based on Princeton WordNet)
- **Map** lemmas to OED (local JSON version)

- Extend already performed VARD2 normalizations to 15th century
- Use MorphAdorner to automatically **normalize further**
- **Map** again
- Successfully mapped: c. 50,000 word forms, unmapped: c. **100,000**
ADDITIONAL NORMALIZATION
(HÄMÄLÄINEN ET AL. 2018)

• Idea: use machine translation!
  • NMT = neural machine translation (OpenNMT)
  • SMT = statistical machine translation (Moses)
  • Use known VARD2 normalizations, MED and OED as input
  • Character-based; language model = BNC
• Levenshtein edit distance + filter by semantic similarity + Soundex pronunciation by edit distance
• Extend VARD2 normalization rules to all words
  • 58 rules, e.g. “u → v anywhere”
• Changing the neural network to bi-directional recurrent neural network
• Looking at the top 10 predictions with OED and Spacy
INTERFACE FOR FILTERING THE NEOLOGISM CANDIDATES

• FiCa (Filtering and Categorization)

• Developed by Eetu Mäkelä; see Säily et al. (2018)
<table>
<thead>
<tr>
<th>Word</th>
<th>Category</th>
<th>Notes</th>
<th>Earliest letter</th>
<th>OED - CEEC</th>
<th>Total DF</th>
<th>Total TF</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>acharya</td>
<td>yes</td>
<td></td>
<td>acharya</td>
<td>1789</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>anthroponomiological</td>
<td></td>
<td></td>
<td></td>
<td>1734</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>anti-democrat</td>
<td>yes</td>
<td>different</td>
<td></td>
<td>1799</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>blueism</td>
<td>yes</td>
<td></td>
<td>blueism</td>
<td>1795</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>bonneted</td>
<td>yes</td>
<td>even if V</td>
<td></td>
<td>1781</td>
<td>43</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>canicule</td>
<td>yes</td>
<td></td>
<td>canicule</td>
<td>1701</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>cardiphonia</td>
<td>yes</td>
<td>actual 1i</td>
<td></td>
<td>1780</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>cast-off</td>
<td>yes</td>
<td></td>
<td>cast-off</td>
<td>1692</td>
<td>48</td>
<td>71</td>
<td>73</td>
</tr>
<tr>
<td>catchy</td>
<td>catchy</td>
<td>yes</td>
<td></td>
<td>1784</td>
<td>20</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>chaplaincy</td>
<td>chaplain</td>
<td>yes</td>
<td>chaplaincy</td>
<td>1741</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>cleverosity</td>
<td>cleveras</td>
<td>yes</td>
<td>cleverosity</td>
<td>1778</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>curtailless</td>
<td>curtailk</td>
<td>yes</td>
<td>curtailless</td>
<td>1789</td>
<td>23</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>delineator</td>
<td>delineat</td>
<td>yes</td>
<td>delineator</td>
<td>1736</td>
<td>38</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>dicky-bird</td>
<td>dicky-bi</td>
<td>yes</td>
<td>dicky-bird</td>
<td>1778</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>double-beded</td>
<td>double-t</td>
<td>yes</td>
<td>double-beded</td>
<td>1798</td>
<td>0</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>double-cross</td>
<td>double-t</td>
<td>yes</td>
<td>double-cross</td>
<td>1754</td>
<td>80</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>embodiment</td>
<td>embodi</td>
<td>yes</td>
<td>embodiment</td>
<td>1777</td>
<td>51</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>envoyship</td>
<td>envoyst</td>
<td>yes</td>
<td>envoyship</td>
<td>1706</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>eschandlion</td>
<td>eschant</td>
<td>yes</td>
<td>eschandlion</td>
<td>1717</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>escripture</td>
<td>escript</td>
<td>yes</td>
<td>escripture</td>
<td>1684</td>
<td>13</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>freshen</td>
<td>fresheni</td>
<td>yes</td>
<td>freshen</td>
<td>1680</td>
<td>17</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>fussy</td>
<td>fussy</td>
<td>yes</td>
<td>fussy</td>
<td>1787</td>
<td>34</td>
<td>72</td>
<td>72</td>
</tr>
</tbody>
</table>

"...your own diamonds &c must be kept, Barton. Mr Collings prize tup is not very capital to handle, but rather catchy to look at. It looks like Mr Wastles judgment weakening Mr Collings says he has let one to George Lax & better..."

This entry has not yet been fully updated (first published 1889).
PILOT STUDY

Social aspects of new 17th-century lexis
NEW 17\textsuperscript{TH}-CENTURY LEXIS IN THE CEEC

- CEEC, socially representative sample from 1640–1660 (\textit{Civil War} period)
  - 54 writers, 107 letters, 36,265 words
- Criteria for new vocabulary:
  - CEEC 1\textsuperscript{st} attestation max 40 years after OED 1\textsuperscript{st} attestation
  - Aim also to test pipeline → filtering done more manually than usual
  - 820 CEEC 1\textsuperscript{st} attestations for human to filter in interface
- Results: \textbf{42 new words, 12 OED antedatings, 3 “actual” neologisms} (seem to antedate both OED and contemporary published texts):
  - \textit{packet-boat} (1641<1642), \textit{statement} (1642<1750), \textit{tea} (1643<1655)
WHO PRODUCED THE 3 NEOLOGISMS?

- Two noblemen, father and son!
  - Royalists, Catholic leanings; art collectors
  - Travelled on the continent
- **Thomas Howard**, Earl of Arundel (1585–1646)
  - Courtier, politician
  - Wrote in his late 50s
- **William Howard**, Viscount Stafford (1612–1680)
  - Younger son, minor role in politics
  - Wrote in his early 30s

William Howard and Thomas Howard (Wikimedia Commons)
just nowe Count Fabroni, & President Cognewe are come unto me from Q: Mother, to entreate very earnestly, that the gentleman cominge alonge wth this called Don Martino Dugaldi may instantly passe to Dunkerke for her Mties especiall service wch depends soe much upon it as upon his retorne or any others sent before by ye Packette Boate …

(Thomas Howard in Dover to Sir John Pennington, 1641; OED3 1642)

I have receaved onely one letter in which there is a statement that the ssouldiers went to Mr John Penneducks house at King berry and ransaked it totally …

(William Howard in Antwerp to Thomas Howard, 1642; OED3 1750)

I have scarce bought any thinge for my selfe but an Indian Brewhouse for tee, which hath beene very good Black Lack worke, but it is all spoyled and rased and yett I payed exceeding deare for it.

(William Howard in Amsterdam to his mother, 1643; OED2 1655)
WHO ARE THE EARLY ADOPTERS?

• Surprisingly many new words compared to number of running words:
  • Harrison (4), parliamentarian army officer
  • L’Estrange (4), royalist politician

• Social networks:
  • Elizabeth & Charles Stuart
  • Charles Nichols & Henry Oxinden
  • Anne Conway & Henry More
  • William & Thomas Howard

<table>
<thead>
<tr>
<th>Name</th>
<th># of new words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harrison, Thomas</td>
<td>4</td>
</tr>
<tr>
<td>L’Estrange, Hamon</td>
<td>4</td>
</tr>
<tr>
<td>Stuart, Elizabeth, Q. Bohemia</td>
<td>4</td>
</tr>
<tr>
<td>Conway, Anne</td>
<td>3</td>
</tr>
<tr>
<td>Cary, Anne</td>
<td>2</td>
</tr>
<tr>
<td>Fleming, Daniel</td>
<td>2</td>
</tr>
<tr>
<td>Howard, William</td>
<td>2</td>
</tr>
<tr>
<td>Nichols, Charles</td>
<td>2</td>
</tr>
<tr>
<td>Oxinden, Henry</td>
<td>2</td>
</tr>
<tr>
<td>Stuart, Charles I</td>
<td>2</td>
</tr>
<tr>
<td>More, Henry</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>
EXAMPLES

To agree (as is alreadie) to act in dearest love expressed to him named Protector, (or Mount Sirion as the Sidonians called Hermon, and David in the spirit followed that faithfully, believingly, undoubtingly, unanimously, that He would retreat in action of undertaking (and soe witnes repentance by condisention) and wee would as willingly repent of or sinfull dissentions) …

(Thomas Harrison to John Jones, 1656; OED3 1631, OED2 1642)

soone after I recd from Mr Playford himselfe a large and fervent letter wth profession of his skill and respects; wch letter (together wth his printed Manifesto or Publicacion thereincluded) I presume to send unto you to peruse …

(Hamon L’Estrange to Thomas Browne, 1653; OED3 1620)
but you must give me Leave to take notice of that part of yoř Letter where you say y’t all coloures are formally in the soul and but causally in the objects & organ, that green and white is not in yoř carpett & papyr, but in yoř mind …  
(Anne Conway to Henry More, 1651; OED2 1635)

• Philosophers, from pupil–tutor relationship to close friends

for first the paper transmitts the colour of white wch is its own motion, and if it should transmitt the motion caused by any other obiect, then why does not everything we Looke upon yeeld the eidolum or representation of something else …  
(Anne Conway to Henry More, 1651; OED2 1619)
Surprisingly few new words compared to number of running words:

- Upper clergy

I keep my self so close an anachoret, and make so few visits, that I am not likely suddainly to meet with Sir John Dingly …

(Brian Duppa to Justinian Isham, 1650; OED 1626)

<table>
<thead>
<tr>
<th>Social rank</th>
<th># of new words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clergy, lower</td>
<td>7</td>
</tr>
<tr>
<td>Genry, lower</td>
<td>7</td>
</tr>
<tr>
<td>Gentry, upper</td>
<td>7</td>
</tr>
<tr>
<td>Professional</td>
<td>7</td>
</tr>
<tr>
<td>Nobility</td>
<td>7</td>
</tr>
<tr>
<td>Royalty</td>
<td>6</td>
</tr>
<tr>
<td>Clergy, upper</td>
<td>1</td>
</tr>
<tr>
<td>Merchant</td>
<td>0</td>
</tr>
<tr>
<td>Other non-gentry</td>
<td>0</td>
</tr>
</tbody>
</table>
• More data from men, more new words used by men
  • More advanced statistics needed

• **More** new words by **older** people (40–70)
  • Previous research on Dutch: highest lexical productivity among highly educated older men (Keune 2012)

• Surprisingly **many** new words in **letters to acquaintances and close friends**, fewer to nuclear family members
  • Consistent with “bulge theory” (Wolfson 1990); less stable relationship triggers more creative language use (cf. Säily 2018, -ity)
DISCUSSION

• New 17th-century vocabulary used by the upper & middling ranks
  • 18th century: surprisingly many new words by the lower ranks
  • Access to education, specialized registers, new things & ideas; social networks
  • Lack of data from the lower ranks
• Age, audience design also factors
  • 18th-century results similar although more focused on close friends, not other acquaintances – changing functions of letter-writing?
• Novelty sometimes indicated in the text
  • Underlining, glossing, part of a list
CONCLUSION

• **Big-data approach**: massive databases & automated pipeline → possible to quickly discover dozens of neologisms in millions of words of running text
  • Letters seem to be a great source for historical lexicography & lexicology
• What are we **missing**? Homonyms, zero derivation, MWUs, …
  • Spelling variation still an issue, could disproportionately affect lower ranks
  • Actual 1st attestations? Innovators vs. early adopters
• **Future work**:
  • Extend analysis to entire time period
  • Ask more focused questions (e.g. who are the coiners of nonce-words, early adopters of vocabulary that comes into general use later)
  • Improve normalization & pipeline; integrate social analysis functionality into pipeline
REFERENCES


THANK YOU!

acrimonious, believingly, candid, candour, causally, compensate, compliance, condescension, coney ground, congregational, covenanting (adj.), crawling (n.), dishearten, dragooner, efficaciously, eminently, endeared, entanglement, helpfulness, hint (v.), idolum, incendiary, incognito, initiatory, joke, landgravine, leveller (sense 2), malignancy, manifesto, oversweetness, packet-boat, plenipotentiary, remind, rickets, sequestrator, statement, Swede, tea, variously, vibrate (v.), visit (n.), voluminous

Special thanks to Oxford University Press and the Middle English Compendium for sharing their lexicographical data (OED, HT; MED)

https://github.com/mikahama/natas
https://github.com/jiemakel/fica