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

Tanja Säily, Eetu Mäkelä & Mika Hämäläinen
NATAS PROJECT: SOCIAL EMBEDDING OF NEW WORDS

• Part of Academy of Finland funded STRATAS project, 2016–2019: Combining text and social information in historical sociolinguistics
  • blogs.helsinki.fi/stratas-project/
• Who creates and adopts new vocabulary in the history of English? Differences across social groups?
  • Previous research focused on published writings by highly educated men
  • We take into account women, lower social ranks
  • Case studies of -ness, -ity, -er: sociolinguistic variation found in productivity (e.g. Säily & Suomela 2017, Säily et al. 2018)
CORPORA OF EARLY ENGLISH CORRESPONDENCE (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/
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
REVISITING NMT FOR ADDITIONAL NORMALIZATION

- Neural Machine Translation: “translating” from historical to modern English at the level of characters
- New: change the neural network (Hämäläinen et al. 2018) to a bi-directional recurrent neural network (Hämäläinen et al. 2019)
- Focus on words found in the OED

<table>
<thead>
<tr>
<th></th>
<th>15th</th>
<th>16th</th>
<th>17th</th>
<th>18th</th>
<th>generic</th>
</tr>
</thead>
<tbody>
<tr>
<td>OED</td>
<td>69%</td>
<td>78%</td>
<td>71%</td>
<td>50%</td>
<td>61%</td>
</tr>
<tr>
<td>+Lemma</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lemma</td>
<td>61%</td>
<td>67%</td>
<td>63%</td>
<td>45%</td>
<td>53%</td>
</tr>
</tbody>
</table>
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 (52)</th>
<th>Earliest letter</th>
<th>OED - CEEC</th>
<th>Total DF</th>
<th>Total TF before</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>acharya</td>
<td>yes</td>
<td></td>
<td>1769</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>anthroponomical</td>
<td>yes</td>
<td></td>
<td>1734</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>anti-democrat</td>
<td>yes</td>
<td>different</td>
<td>1769</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>bluesium</td>
<td>bluesium</td>
<td></td>
<td>1768</td>
<td>48</td>
<td>71</td>
<td>73</td>
<td>2</td>
</tr>
<tr>
<td>bonneted</td>
<td>bonneted</td>
<td></td>
<td>1768</td>
<td>43</td>
<td>17</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>canicule</td>
<td>canicule</td>
<td></td>
<td>1768</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>cardiphonia</td>
<td>cardiph</td>
<td>actual 1r</td>
<td>1760</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>cast-off</td>
<td>cast-off</td>
<td></td>
<td>1762</td>
<td>3</td>
<td>17</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>catchy</td>
<td>catchy</td>
<td></td>
<td>1784</td>
<td>3</td>
<td>35</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>chaplaincy</td>
<td>chaplain</td>
<td></td>
<td>1741</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>cleverability</td>
<td>clevera</td>
<td></td>
<td>1778</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>curtailness</td>
<td>curtail</td>
<td></td>
<td>1769</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>delineator</td>
<td>delineat</td>
<td>yes</td>
<td>1756</td>
<td>38</td>
<td>28</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>dicky-bird</td>
<td>dicky-bi</td>
<td>figurativ</td>
<td>1778</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>double-beded</td>
<td>double-b</td>
<td>yes</td>
<td>1768</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>double-cross</td>
<td>double-c</td>
<td>yes</td>
<td>1754</td>
<td>80</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>embodiment</td>
<td>embodi</td>
<td>yes</td>
<td>1777</td>
<td>30</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>envoyship</td>
<td>envoy</td>
<td></td>
<td>1760</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>eschatantion</td>
<td>eschant</td>
<td></td>
<td>1717</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>eschatire</td>
<td>eschato</td>
<td></td>
<td>1684</td>
<td>13</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>freshe</td>
<td>fresheri</td>
<td></td>
<td>1680</td>
<td>17</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>fussy</td>
<td>fussy</td>
<td></td>
<td>1767</td>
<td>34</td>
<td>72</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. Wastes judgement weakening Mr. Collings says he has not to George Lax &c better.”

They havessbovi et Ben tear a highly commenr, guy, of dule-street, York-buildings, Sur;

Catchy 

Guy, deit off;

“*But they are All of a piece; yet they lie upon the

CATCHY

Catching to Trip up the Heels one of Another. *Prethee will thou make these Things Hang Together, now.

Catched supplied – if for trivalent any double with ref imperfect to the figurative, ‘the packer’s’ lazy may be able to

OED

*Purchased for £250.

This entry has not yet been fully updated (First published 1886).
NEW 17TH-CENTURY LEXIS IN THE CEEC

• CEEC, socially representative sample from 1640–1660 (Civil War period)
• Results: 42 new words, 12 OED antedatings, 3 “actual” neologisms (seem to antedate both OED and contemporary published texts):
  • *packet-boat* (1641<1642), *statement* (1642<1750), *tea* (1643<1655)
• New vocabulary mainly 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
EARLIEST KNOWN ATTESTATIONS OF THE ENGLISH WORDS FOR TEA

I have scarce bought any thinge for my selfe but an Indian Brew-house for tee, which hath bee 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; OED 1655)

Mr Eaton, I pray you buy for me a pot of the best sort of chaw in Meaco, 2 farre [fair] bowes & arrowes, some half a dozen of Meaco guilt boxes, square, for to put in tobacco.

(Richard Wickham at Hirado to William Eaton at Kyoto, 1615; OED 1616)
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 (~60% accuracy in lemmatization. Errors evenly spread across ranks, but not according to sex or relationship)
  • 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
THANK YOU!

acrimonious, believingsly, 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