NEOLOGISM DETECTION IN HISTORICAL CORPORA

Tanja Säily, Mika Hämäläinen & Eetu Mäkelä
INTRODUCTION
• Interfacing structured and unstructured data in sociolinguistic research on language change (Academy of Finland, DIGIHUM, 2016–2019)
  • 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
  • Corpus of Early English Correspondence (CEEC): wide social spectrum, speech-like genre
• 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 15\textsuperscript{th} century)

www.helsinki.fi/varieng/CoRD/corpora/CEEC/
RESEARCH QUESTIONS

1. **Who** are the innovators? 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 NEOLOGISMOS

• 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
• **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 15\textsuperscript{th} century
• Use MorphAdorner to automatically **normalize further**
• **Map** again
• Successfully mapped: c. 50,000 word forms, unmapped: c. 100,000
ADDITONAL 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”
## EXAMPLES FROM THE 18\textsuperscript{TH} CENTURY

<table>
<thead>
<tr>
<th>Word</th>
<th>NMT</th>
<th>SMT</th>
<th>Edit distance</th>
<th>VARD2</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>elyzian</td>
<td>elizian</td>
<td>elysian</td>
<td>elysian</td>
<td></td>
<td>Elysian</td>
</tr>
<tr>
<td>arch-type</td>
<td>archedip</td>
<td>archetype</td>
<td>archetype</td>
<td></td>
<td>archetype</td>
</tr>
<tr>
<td>kindelyer</td>
<td>kindler</td>
<td>kinder</td>
<td>kindlier</td>
<td></td>
<td>kindlier</td>
</tr>
<tr>
<td>supp'd</td>
<td>supped</td>
<td>supply</td>
<td>supply</td>
<td>cupped</td>
<td>supped</td>
</tr>
<tr>
<td>prospirity</td>
<td>prosperity</td>
<td>prosperity</td>
<td>prosperity</td>
<td></td>
<td>prosperity</td>
</tr>
<tr>
<td>affectionett</td>
<td>affection</td>
<td>affection</td>
<td>affectionate</td>
<td></td>
<td>affectionate</td>
</tr>
<tr>
<td>driction</td>
<td>driction</td>
<td>diction</td>
<td>direction</td>
<td></td>
<td>direction</td>
</tr>
<tr>
<td>octvo</td>
<td>ctuocotoo</td>
<td>octal</td>
<td>oct, oct.</td>
<td></td>
<td>octavo</td>
</tr>
<tr>
<td>mic'lemas</td>
<td>miclemas</td>
<td>miles</td>
<td>micklemas</td>
<td></td>
<td>Michaelmas</td>
</tr>
<tr>
<td>midetrenian</td>
<td>midetrenian</td>
<td>mid-on</td>
<td></td>
<td></td>
<td>Mediterranean</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Generic</th>
<th>15th century</th>
<th>18th century</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT</td>
<td>28%</td>
<td>43%</td>
<td>14%</td>
</tr>
<tr>
<td>NMT 18th</td>
<td>28%</td>
<td>43%</td>
<td>15%</td>
</tr>
<tr>
<td>NMT 15th</td>
<td>28%</td>
<td>43%</td>
<td>15%</td>
</tr>
<tr>
<td>NMT no years</td>
<td>46%</td>
<td>55%</td>
<td>25%</td>
</tr>
<tr>
<td>SMT</td>
<td>32%</td>
<td>31%</td>
<td>28%</td>
</tr>
<tr>
<td>SMT 18th</td>
<td>16%</td>
<td>14%</td>
<td>19%</td>
</tr>
<tr>
<td>SMT 15th</td>
<td>32%</td>
<td>28%</td>
<td>31%</td>
</tr>
<tr>
<td>Edit distance</td>
<td>31%</td>
<td>31%</td>
<td>31%</td>
</tr>
<tr>
<td>VARD2 rules</td>
<td>15%</td>
<td>10%</td>
<td>8%</td>
</tr>
<tr>
<td>At least one correct</td>
<td>67%</td>
<td>71%</td>
<td>52%</td>
</tr>
</tbody>
</table>
Overlap
SELECTING THE BEST NORMALIZATION

- **Voting**: the normalization with the most votes from the different methods wins
- **Weighted voting**: methods are weighted based on their accuracy in a test set
- **Markov chain** trained on the BNC

For the 18\textsuperscript{th}-century test set:
- Cases in which at least 1 normalization method is correct: 52%
- Accuracy of methods of picking the best normalization: 22–24%
  - Not good enough to be used yet
  - We are dealing with the most difficult cases; easy cases dealt with in the initial mapping
RETRIEVING RELATED DATA

- **OED**: all fields except pronunciation
- **MED**: years, link to OED
- **EEBO, ECCO, BLN, BN**: term frequency & document frequency before & after CEEC 1st attestation
- **CEEC**: metadata on letter, sender, recipient

- Scripts written by Mika Hämäläinen
- Data currently collected in a massive Excel file…
151,210 distinct word forms originally
Proportion of words that aren’t in dictionaries decreases through time (= spelling standardizes)
34,352 distinct word forms first seen in the long 18th century
31,659 of these are not tagged foreign, proper or superscript
FILTERING STATISTICS

- Of the 31,659, 8,813 directly match a dictionary word
- 1,297 more can additionally be normalized to a dictionary word
- Out of these 10,110, 487 appear in the letter corpus earlier than or in the same year as their earliest dictionary attestation

On the other hand:

- Out of the 21,549 that couldn’t be mapped to a dictionary word, 12,245 appear in 100 or less documents in the comparison corpora before their first letter appearance
- Out of those, 2,540 appear in 100 or more documents later
- Overall, out of the 21,549, 9,115 appear in 100 or less documents in the comparison corpora at any time
INTERFACE FOR FILTERING THE NEOLOGISM CANDIDATES

- FiCa (Filtering and Categorization)

- Developed by Eetu Mäkelä; see Säily et al. (in press)
catchy, adj.

Pronunciation: /ˈkætʃi/ (U.S.) /ˈkʌtʃi/ (Brit.)

Etymology: < catch v.

1. Adapted to catch the attention or fancy; attractive; "taking".
INTERFACE FOR ANALYSING SOCIAL ASPECTS OF NEOLOGISMS

- Currently: combination of FiCa and Excel

- Need to develop an all-in-one interface that also provides visualizations and statistical analyses
PILOT STUDY

Social aspects of 18th-century neologisms
18TH-CENTURY NEOLOGISMS IN THE CEEC

- CEEC, long 18th century (1680–1800)
  - 315 writers, 4,945 letters, 2.2 million words
- Criteria:
  - CEEC 1st attestation ≤ OED 1st attestation
  - Can occur in max 100 contemporary published texts before CEEC 1st attestation
  - Automated procedures → only 220 candidates for human to filter in interface
    - 81 neologisms found
Surprisingly many neologisms compared to number of running words:
- Twining (13), clergyman
- Austen, Bentham, Gray (4), authors
- Jackson, St. Michel (2)
Surprisingly few:
- Lady Mary Wortley Montagu (1)
Social networks:
- Twining & Burneys
- Jackson & St. Michel (related to Pepys)

<table>
<thead>
<tr>
<th>Name</th>
<th># of neologisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twining, Thomas</td>
<td>13</td>
</tr>
<tr>
<td>Austen, Jane</td>
<td>4</td>
</tr>
<tr>
<td>Bentham, Jeremy</td>
<td>4</td>
</tr>
<tr>
<td>Burney, Frances</td>
<td>4</td>
</tr>
<tr>
<td>Gray, Thomas</td>
<td>4</td>
</tr>
<tr>
<td>Cowper, William</td>
<td>3</td>
</tr>
<tr>
<td>Lennox, Sarah</td>
<td>3</td>
</tr>
<tr>
<td>Burney, Charles</td>
<td>2</td>
</tr>
<tr>
<td>Jackson, John</td>
<td>2</td>
</tr>
<tr>
<td>St. Michel, Balthasar</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>2</td>
</tr>
</tbody>
</table>
I must not omit acquainting you, Sir, That upon Opening his Body (which the uncommonness of his Case required of us, for our own satisfaction as well as Publick Good) there was found in his Left Kidney a nest of no less than 7 stones, of the most irregular Figures your imagination can frame, and weighing together 4 1/2 ounces …

(John Jackson to John Evelyn, 1703; OED 1705)

But I don’t recollect any single word in our language but Tune that expresses the intune-ness of an interval. Intonation is rather more scientific & jargonic than I like.

(Thomas Twining to Charles Burney, 1781; OED 1819)

(intune-ness not in OED!)
SOCIAL RANKS

- Surprisingly many neologisms compared to number of running words:
  - Lower clergy (due to Twining)
  - **Other non-gentry** (lowest category)
    - John Jackson (2), farmer's son, upwardly mobile
    - Ignatius Sancho (2), son of a slave, upwardly mobile
    - George Culley (1), farmer
- Surprisingly few:
  - Royalty

<table>
<thead>
<tr>
<th>Social rank</th>
<th># of neologisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional</td>
<td>26</td>
</tr>
<tr>
<td>Clergy, lower</td>
<td>22</td>
</tr>
<tr>
<td>Gentry, lower</td>
<td>13</td>
</tr>
<tr>
<td>Nobility</td>
<td>6</td>
</tr>
<tr>
<td>Gentry, upper</td>
<td>6</td>
</tr>
<tr>
<td>Other non-gentry</td>
<td>5</td>
</tr>
<tr>
<td>Clergy, upper</td>
<td>1</td>
</tr>
<tr>
<td>Merchant</td>
<td>1</td>
</tr>
<tr>
<td>Royalty</td>
<td>1</td>
</tr>
</tbody>
</table>

Royalty: 1
Other non-gentry: 5
Gentry, upper: 6
Gentry, lower: 13
Clergy, lower: 22
Professional: 26
Mr Collings prize tup is not very capital to handle, but rather **catchy** to look at.

(George Culley to Matthew Culley, 1784; OED 1831)

We hope he is well, and enjoys this fine weather unplagued by flies, and **unbitten** by fleas.

(Ignatius Sancho to Mrs. Cocksedge, 1779; OED 1794)

Our dear mother is well but hurried, my sister very **fussy** & agitated, the rest of the family in full trim though **heart full** from the thoughts of so soon being seperated, with laughing faces to keep up one another’s spirits.

(Elizabeth Hanover to the Prince of Wales, 1797; OED 1831)
• More data from men, more neologisms by men
  • More advanced statistics needed

• Fewer neologisms by younger (10–29) and less well educated people (secondary / apprentice)
  • Previous research on Dutch: highest lexical productivity among highly educated older men (Keune 2012)

• Surprisingly many neologisms in letters to 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)
SEMANTICS

• **People**: emotion, mental capacity, attention & judgement, behaviour, manner of action
  - *ill-natured, cleverality, nidgetty, missish, fussy*

• **Society**: communication, trade, work, faith, authority
  - *escritoire, knick-knackatory, wagon-way, chaplaincy, envoyship*

• **World**: action, space
  - *godsend, unstow*

# of neologisms per HT category
- 6 *the world* » action or operation
- 6 *the mind* » emotion
- 5 *society* » communication
- 4 *the world* » space
- 4 *the mind* » mental capacity
- 4 *the mind* » attention and judgement
- 3 *society* » trade and finance
- 3 *society* » occupation and work
- 3 *society* » faith
- 3 *society* » authority

…
CONCLUSION

- **Big-data approach**: massive databases & automated pipeline → possible to quickly discover dozens of neologisms in millions of words of running text
- What are we **missing**? Homonyms, zero derivation, MWUs, …
  - Spelling variation still an issue, could disproportionately affect lower ranks
  - Actual 1\textsuperscript{st} attestations? Are these innovators or just early adopters?
- **Future** work:
  - Expand analysis to all research questions, entire time period; also analyse words not in the OED, check instances in contemporary published texts
  - Improve normalization & pipeline; develop methods & interface for analysing social aspects & spread of neologisms
REFERENCES


THANK YOU!

acharya, anthroponomical, anti-democrat, blueism, bonneted, canicule, cardiphonia, cast-off, catchy, chaplaincy, cleverality, curtainless, delineator, dicky-bird, double-bedded, double-cross, embodiment, envoyship, eschantillon, escritoire, freshen, fussy, godsend, grumpy, guimpe, hummily, hydrogenate, idlish, impracticability, incomed, incontestably, inexact, inside-outness, internment, intrepid, jargonic, jumpable, keyless, kibitka, knick-kackatory, letteret, malformation, marituriend, mevrouw, missish, monotonous, moon, moonery, nidgetty, non-papist, pacifist, paperless, pheasantry, pushery, rishi, schoolmasterishness, scratch-back, shockingly, silentious, slushy, spidery, sprawly, squeezy, stiffish, sweet-hearted, tawdrily, trickster, truantism, unailing, unbitten, unclassed, uncommonness, undefeat, unenjoyable, uneventful, ungown, uninsured, uninteresting, unstowed, wagon-way, yester-evening

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