BETWEEN THE OED AND HISTORICAL TEXT COLLECTIONS:

THE CASE OF DISCOVERING NEW WORDS

Tanja Säily, Eetu Mäkelä & Terttu Nevalainen
OUTLINE

VARIENG, HISTORICAL TEXT COLLECTIONS AND THE OED

NATAS: SOCIAL ASPECTS OF NEW WORDS IN EARLY ENGLISH LETTERS

WORKFLOWS: BACK AND FORTH BETWEEN THE OED AND HISTORICAL TEXT COLLECTIONS
HISTORICAL TEXT COLLECTIONS COMPILED AT VARIENG

- **VARIENG** = Research Unit for the Study of Variation, Contacts and Change in English
- Community of linguists who study language variation and change
- Special focus on designing and compiling digitized historical text collections (corpora)
  - [http://www.helsinki.fi/varieng/](http://www.helsinki.fi/varieng/)
- **Helsinki Corpus of English Texts** (HC)
- **Helsinki Corpus of Older Scots** (HCOS)
- **Corpus of Early English Correspondence** (CEEC)
- **Corpus of Scottish Correspondence** (CSC)
- **Corpus of Early English Medical Writing** (CEEM)
- **Helsinki Corpus of British English Dialects**
THE OED AS A TOOL AND A REPOSITORY IN CORPUS-BASED RESEARCH

• The OED is studied in its own right as a source for vocabulary developments over time
• The OED is a baseline of comparison for studies on other levels of language
• Current VARIENG Digital Humanities projects referring to the OED:
  • Reassessing Language Change
  • STRATAS
  • https://www.helsinki.fi/en/researchgroups/varieng/research

• OED antedatings from VARIENG research publications, e.g.:
  • linguist, in the sense ‘interpreter, translator’ – 1st attestations
    • OED2 (1989) 1711
    • Kaislaniemi (2011) 1610
    • OED3 (2013) 1612
• Data from:
  • OED2 & 3: printed book (3: EEBO)
  • Kaislaniemi: Hakluyt Society edition
_RATE OF CHANGE: COMPARING CORPUS FINDINGS WITH THE MIDDLE ENGLISH INTAKE OF ROMANCE LOANS

Rate of change: incoming features in the HC

Fig. 2.6 Loanwords from French, Latin, and French and/or Latin as a proportion of all new words, as reflected by parts of OED3 so far completed.

NATAS: SOCIAL ASPECTS OF NEW WORDS IN EARLY ENGLISH LETTERS
WORKFLOW

- **Research question**: discover social differences in the coining of new words
- **Approach**: take a collection of texts with social metadata. Find who there uses words not found elsewhere before. Discover differences by group, location, gender…
- **How to find words not found elsewhere before?**
  - Compare against **dictionaries**: Oxford English Dictionary, Middle English Dictionary
  - Compare against **other text collections**: British Library Newspapers, Burney & Nichols Collections, Eighteenth Century Collections Online, Early English Books Online
PILOT STUDY: NEW WORDS IN 18TH-CENTURY LETTERS

- Corpora of Early English Correspondence (CEEC), long 18th century (1680–1800)
  - 315 writers (social backgrounds known), 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
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, hummingly, hydrogenate, idlish, impracticability, incomed, incontestably, inexact, inside-outness, internment, intrepid, jargonic, jumpable, keyless, kibitka, knick-knackatory, letteret, malfunction, mariturient, mevrouw, missish, monotonous, moon v., moonery, nidgety, 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
WHO ARE THE INNOVATORS?

• Surprisingly many new words compared to number of running words:
  • Lower clergy (Thomas Twining: 13)
  • 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>
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)
OED ANTEDATINGS: CATCHY

• OED3 catchy, adj., sense 2b:
  ‘That is designed or intended to draw attention; attractive; eye-catching.’

• Current 1st attestation:
  1831 Fraser’s Mag. July 679/2 A catchy, stage-like effect.

• Straightforward antedating:
  1784 G. Culley 1 Oct. Mr Collings prize tup is not very capital to handle, but rather catchy to look at.
OED ANTEDATINGS: ANTI-DEMOCRAT

• OED2 anti-democrat, n.: ‘One who is opposed to democracy or to the principles of the Democratic party in the United States.’

• Current 1st attestation:
  1802 (title) The Republican; or, Anti-Democrat [cited in D.A.].

• Antedating:
  1799 S. LENNOX 25 Mar. My husband you know is a decided Royalist & a supporter of monarchy, consequently a decided anti-Democrat & anti-Republican (as this term is now understood, for the Government of England properly speaking is a Republican form of Govt): this agrees perfectly with my own disposition.
OED ANTEDATINGS: PACIFIST

• OED3 pacifist, adj.: ‘Of, relating to, supportive of, or characterized by pacifism.’

• Current 1st attestation:
  1906    Times 8 Aug 3/3    M. Paul Fribourg, waxes eloquent to-day in the Petite République in his reproaches of his German comrades for their opposition to the ‘Pacifist League’.

• Antedating?
  1717    T. YORKE 30 May    My mother..has an honest attorney, that will not hurry her into vexatious lawsuits out of any self-interest; and we take all opportunitys to incline her to a pacifist disposition.
OED ANTEDATINGS: HUMMINGLY

- OED2 hummingly, adv.: ‘With a humming sound.’

- Current 1st attestation:
  1908  Daily Chron. 21 Jan 6/5  He..endeavours to follow the service and join hummingly in the hymns.

- Antedating?
  1775  D. GARRICK 7 Nov.  – the Little Gypsy* goes on Hummingly, & rises nightly in repute –

  *Musical farce, words by Garrick

- From OED2 humming, adj. 2a ‘of extraordinary activity’ (1654)
OED ANTEDATINGS: PAPERLESS

- OED3 **paperless**, adj.: ‘Lacking official documentation; computerized’

- Current 1\textsuperscript{st} attestation:
  
  1854 T. CONNEAU
  
  *Capt. Canot 245*

  As soon as the cargo was safe.., I resolved to disregard the flagless and paperless craft that bore it.

- Antedating?
  
  1799 F. BURNEY 22 July

  I shall be very anxious for dry weather for them – in our curtainless & **paperless** apartment.

- Different sense – cf. OED2 **curtainless** ‘without a curtain’ (1822)
MEANINGS OF NEW WORDS IN 18TH-CENTURY LETTERS

• **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 new words 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

…
• **Big-data approach**: massive databases & automated pipeline → possible to quickly discover dozens of new words in millions of words of running text
  • Contribution to OED: not only antedatings but also new senses, etc.
• What are we **missing**?
  • Different words/senses spelled in the same way, compounds written separately, …
  • Actual 1\textsuperscript{st} attestations? Are these innovators or just early adopters?
• **Research question**: discover social differences in the coining of new words

• **Approach**: take a collection of texts with social metadata. Find who there uses words not found elsewhere before. Discover differences by group, location, gender…

• How to find words not found elsewhere before?
  ○ Compare against **dictionaries**: Oxford English Dictionary, Middle English Dictionary
  ○ Compare against **other text collections**: British Library Newspapers, Burney & Nichols Collections, Eighteenth Century Collections Online, Early English Books Online
WORKFLOWS

- **Research question**: discover social differences in the coining of new words
- **Approach**: take a collection of texts with social metadata. Find who there uses words not found elsewhere before. Discover differences by group, location, gender…
- **How to find words not found elsewhere before?**
  - Compare against **dictionaries**: Oxford English Dictionary, Middle English Dictionary
  - Compare against **other text collections**: British Library Newspapers, Burney & Nichols Collections, Eighteenth Century Collections Online, Early English Books Online
WORKFLOWS

- **Research question**: discover social differences in the coining of new words
- **Approach**: take a collection of texts with social metadata
  - CEEC: Personal letters, c. 1400–1800. 1,180 writers, 11,713 letters, 5.2 million words.
- How to find words not found elsewhere before?
  - Compare against **dictionaries**:
    - Oxford English Dictionary (OED): ~280,000 words with dates, ~165,000 with earliest attestation before 1800, ~115,000 after.
    - Middle English Dictionary (MED): ~60,000 words from 1000–1500.
WORKFLOWS

- **Research question**: discover social differences in the coining of new words
- **Approach**: take a collection of texts with social metadata
  - CEEC: Personal letters, c. 1400–1800. 1,180 writers, 11,713 letters, 5.2 million words. 
    Contains spelling variation: ~150,000 distinct word forms. ~125,000 not tagged as foreign/proper/abbreviation
- **How to find words not found elsewhere before?**
  - Compare against **dictionaries**:
    - Oxford English Dictionary (OED): ~280,000 words with dates, ~165,000 with earliest attestation before 1800, ~115,000 after. ~375,000 spellings
    - Middle English Dictionary (MED): ~60,000 words from 1000–1500. ~260,000 spellings
- **Result**: ¼, ~36,000 word forms can be mapped to the dictionaries, ¾, ~88,000 cannot
WORKFLOWS

- **Research question**: discover social differences in the coining of new words
- **Approach**: take a collection of texts with social metadata
  - CEEC: Personal letters, c. 1400–1800. 1,180 writers, 11,713 letters, 5.2 million words.
- **CEEC Contains spelling variation**: ~150,000 distinct word forms. ~125,000 not tagged as foreign/proper/abbreviation
  - “... Ships as shall be in *readynesse* for that service, and this matter *requireing* the greatest secrecy,... Your very *affectionett freind*” (Queen Anne, 1704)
  - “Right *worshipfull* and my most *entierly beloude moder*, in the most *louly maner* I *recommaund* me *vnto youre gode moderhode*,” (Elizabeth Poynings, 1459)
WORKFLOWS

- **Research question:** discover social differences in the coining of new words
- **Approach:** take a collection of texts with social metadata
  - CEEC: Personal letters, c. 1400–1800. 1,180 writers, 11,713 letters, 5.2 million words.  
    Contains spelling variation: ~150,000 distinct word forms. ~125,000 not tagged as foreign/proper/abbreviation

- **How to find words not found elsewhere before?**
  - Compare against **dictionaries**:
    - Oxford English Dictionary (OED): ~280,000 words with dates, ~165,000 with earliest attestation before 1800, ~115,000 after. ~375,000 spellings
    - Middle English Dictionary (MED): ~60,000 words from 1000–1500. ~260,000 spellings

- **Result:** ¼, ~36,000 word forms can be mapped to the dictionaries, ¾, ~88,000 cannot
• Proportion not in dictionaries decreases through time (= spelling standardizes)
• ~32,000 non-foreign/non-proper words first seen in the long 18th century
• ~8,800 directly match a dictionary word, ~1,300 more are found through variants
• 487 of the matching words predate their OED attestation
<table>
<thead>
<tr>
<th>Word</th>
<th>Category</th>
<th>Notes (in)</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>1769</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>anthroponomical</td>
<td></td>
<td></td>
<td>1724</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>anti-democrat</td>
<td>yes</td>
<td>different</td>
<td>1769</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>bleuism</td>
<td>bleuism</td>
<td></td>
<td>1795</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>bonneted</td>
<td>bonneted</td>
<td>even if V</td>
<td>1781</td>
<td>43</td>
<td>17</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>canicule</td>
<td>canicule</td>
<td>actual 1t</td>
<td>1760</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>cardiophonia</td>
<td>cardiophonia</td>
<td>yes</td>
<td>1760</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>cast-off</td>
<td>cast-off</td>
<td>yes</td>
<td>1862</td>
<td>48</td>
<td>71</td>
<td>73</td>
<td>2</td>
</tr>
<tr>
<td>catchy</td>
<td>catchy</td>
<td>yes</td>
<td>1864</td>
<td>20</td>
<td>35</td>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td>chaplaincy</td>
<td>chaplain</td>
<td>yes</td>
<td>1741</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>cleverishly</td>
<td>cleverly</td>
<td>yes</td>
<td>1778</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>curtainess</td>
<td>curtain</td>
<td>yes</td>
<td>1769</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>delineator</td>
<td>delineate</td>
<td></td>
<td>1736</td>
<td>1</td>
<td>28</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>dicky-bird</td>
<td>dicky-bird</td>
<td>yes</td>
<td>1778</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>double-decided</td>
<td>double-decided</td>
<td></td>
<td>1709</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>double-cross</td>
<td>double-cross</td>
<td>yes</td>
<td>1754</td>
<td>80</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>embodiment</td>
<td>embodiment</td>
<td>yes</td>
<td>1777</td>
<td>51</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>envoyship</td>
<td>envoy</td>
<td>yes</td>
<td>1706</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>equivocation</td>
<td>equivocating</td>
<td>yes</td>
<td>1717</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>escantion</td>
<td>escantion</td>
<td>yes</td>
<td>1694</td>
<td>13</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>fressen</td>
<td>fressen</td>
<td>yes</td>
<td>1860</td>
<td>17</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>fussy</td>
<td>fussy</td>
<td>yes</td>
<td>1667</td>
<td>34</td>
<td>72</td>
<td>72</td>
<td>1</td>
</tr>
</tbody>
</table>
• ~32,000 non-foreign/non-proper words first seen in the long 18th century
• ~8,800 directly match a dictionary word, ~1,300 more are found through variants
• ~⅔ of the material, ~22,000 not matched

→ cannot yet say anything reliable about social aspects of new word creation as a whole
NORMALIZATION

- **VARD2** normalization rules
  - 58 rules, e.g. “u → v anywhere”
- **Edit distance** against dictionary words + filter by context similarity
- **Character-based machine translation**
  - Neural machine translation (NMT)
  - Statistical machine translation (SMT)
  - Both use CEEC, MED and OED as input
  - Language model = British National Corpus
# 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,cupped</td>
<td>cupped</td>
<td>supped</td>
</tr>
<tr>
<td>prosperity</td>
<td>prosperity</td>
<td>prosperity</td>
<td>prosperity</td>
<td></td>
<td>prosperity</td>
</tr>
<tr>
<td>affectionett</td>
<td>affectiont</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>ctuoctoo</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>Mediterraneant</td>
</tr>
</tbody>
</table>
METHOD ACCURACIES

- NMT: 46%
- SMT: 32%
- Edit distance: 31%
- VARD2 rules: 15%
- *At least one correct:* 67%

- Adding social or temporal metadata to the NMT did not help
METHOD ACCURACIES

- NMT: 46%
- SMT: 32%
- Edit distance: 31%
- VARD2 rules: 15%
- At least one correct: 67%

- Adding social or temporal metadata to the NMT did not help

- Neural lemmatisation + NMT + OED filtering: 61%
- Century-specific accuracies:
  - 15th: 69%
  - 16th: 78%
  - 17th: 71%
  - 18th: 50%

→ Is 18th century variation more individual, earlier spelling variation more general?
METHOD ACCURACIES

- NMT: 46%
- SMT: 32%
- Edit distance: 31%
- VARD2 rules: 15%
- At least one correct: 67%

- Adding social or temporal metadata to the NMT did not help

- Neural lemmatisation + NMT + OED filtering: 61%
- Century-specific accuracies:
  - 15th: 69%
  - 16th: 78%
  - 17th: 71%
  - 18th: 50%

→ Is 18th century variation more individual, earlier spelling variation more general? Or, is this an artifact due to our training data also being mostly from 14th–17th century?
WORKFLOWS

- **Research question**: discover social differences in the coining of new words
- **Approach**: take a collection of texts with social metadata. Find who there uses words not found elsewhere before. Discover differences by group, location, gender...
- **How to find words not found elsewhere before?**
  - Compare against **dictionaries**: Oxford English Dictionary, Middle English Dictionary
  - Compare against **other text collections**: British Library Newspapers, Burney & Nichols Collections, Eighteenth Century Collections Online, Early English Books Online
WORKFLOWS

- Research question: discover social differences in the coining of new words
- Approach: take a collection of text with social metadata
  - CEEC: Personal letters, c. 1400–1800. 1,180 writers, 11,713 letters, 5.2 million words.
- How to find words not found elsewhere before?
  - Compare against other text collections:
    - British Library Newspapers: ~655,000 newspaper issues, ~40 billion running words 1800–1900 (BLN)
    - Burney & Nichols Collections: ~230,000 pamphlets and newspapers, ~2.7 billion running words before 1600–1800 (BN)
    - Eighteenth Century Collections Online: ~210,000 books, ~10 billion running words 1700–1800 (ECCO)
    - Early English Books Online: ~60,000 books, ~1.4 billion running words 1400–1700 (EEBO)
WORKFLOWS

- Research question: discover social differences in the coining of new words
- Approach: take a collection of text with social metadata
  - CEEC: Personal letters, c. 1400–1800. 1,180 writers, 11,713 letters, 5.2 million words.
- How to find words not found elsewhere before?
  - Compare against other text collections:
    - British Library Newspapers: ~655,000 newspaper issues, ~40 billion running words 1800–1900 (BLN)
    - Burney & Nichols Collections: ~230,000 pamphlets and newspapers, ~2.7 billion running words before 1600–1800 (BN)
    - Eighteenth Century Collections Online: ~210,000 books, ~10 billion running words 1700–1800 (ECCO)
    - Early English Books Online: ~60,000 books, ~1.4 billion running words 1400–1700 (EEBO)
- All except EEBO contain a large number of OCR errors
Isw-not- Saint George we Sing of here, 
Nor George, the fatal Duke Villier ;
Nor George a Green, nor Castriot, 
Nor Buchanan the learned Scot q
But us of George the Valiant Monck,
That made Van-Trump in'S Blood deod
and in theseus his Navy snuck. (drunk,
Ok I this is our brave George !
THE
Firste volume of the
Chronicles of England, Scotia
lande, and Irelan:
Conteyning,
The description and Chronicles of England, from the
first inhabiting unto the conquest.
The description and Chronicles of Scotland, from the
first origimall of the Scoces nation, til the yeare
of our Lordes 1571.
The description and Chronicles of Irelande, likewise
from the firste origimall of that Nation, but in the
yeare 1547.
Faithfully gathered and set forth, by
Raphaell Holinshed.

AT LONDON,
Imprinted for John Harrison.
Out of the ~22,000 18th century words that couldn’t be mapped to a dictionary word:

- only 590 appear in over 100 documents in the comparison corpora before
- ~9,000 appear in 100 or less documents at any time
- ~2,500 appear in more than 100 documents later but not before
Because we have so many sources of noise that aren’t uniformly distributed, purely automated general analysis of social aspects is impossible

Need to ask more focused questions that can be answered by combining the tools we do have with a bearable amount of manual labour

Limit by time, filter by appearance in dictionaries and texts before/after

“Who coins new words” →
  • “Who coins new words in the 18th century that come into general use later?”
  • “Who coins nonce words in the 18th century that are used only once and never seen again?”
Currently: combination of FiCa and Excel

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

• Extend analysis to entire time period; also analyse words not in the OED, check instances in contemporary published texts
• Improve normalization, develop discovery pipeline further
• Integrate social analysis functionality to pipeline