

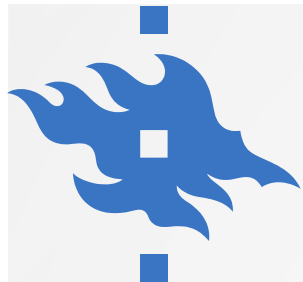


DEVELOPING AN ENVIRONMENT FOR NEOLOGISM DETECTION IN HISTORICAL CORPORA

Tanja Säily, Eetu Mäkelä & Mika Hämäläinen



INTRODUCTION



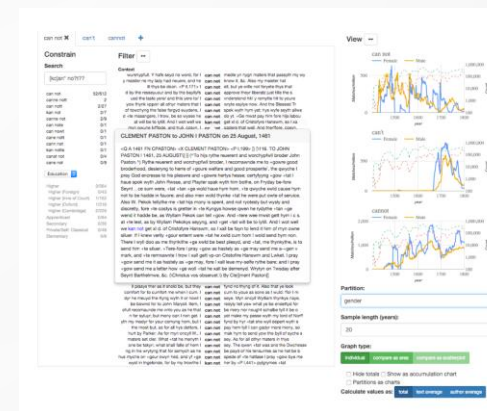
STRATAS PROJECT

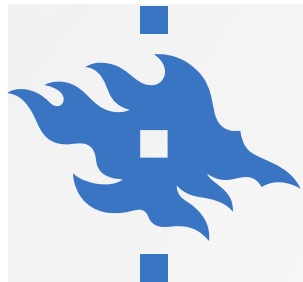
- *Interfacing structured and unstructured data in sociolinguistic research on language change* (Academy of Finland, DIGIHUM, 2016–2019)

- blogs.helsinki.fi/stratas-project/

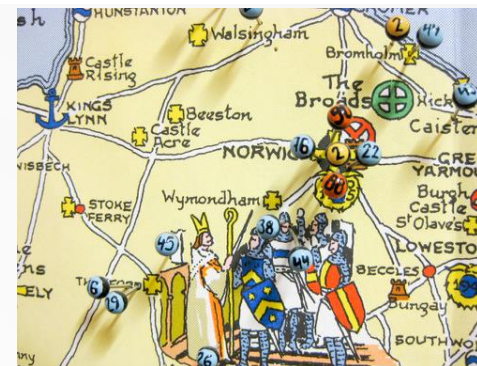
- NATAS subproject:
Social embedding of neologisms in early English correspondence

- Previous research: mostly lexicographical data, bias towards well-known authors
- *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)

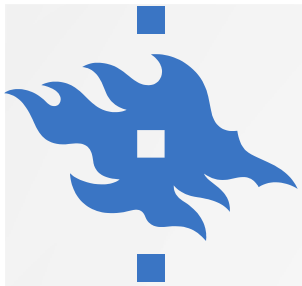
www.helsinki.fi/varieng/CoRD/corpora/CEEC/



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?

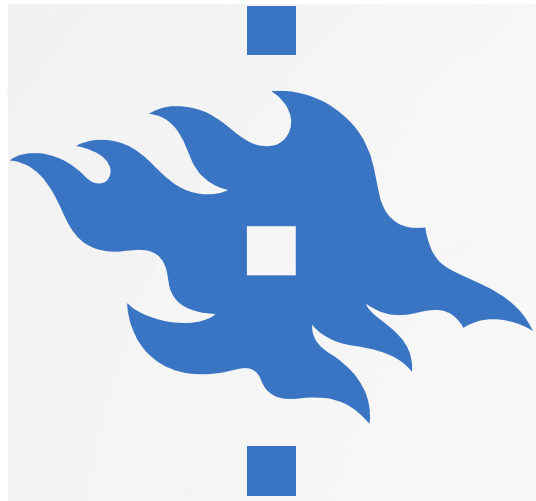




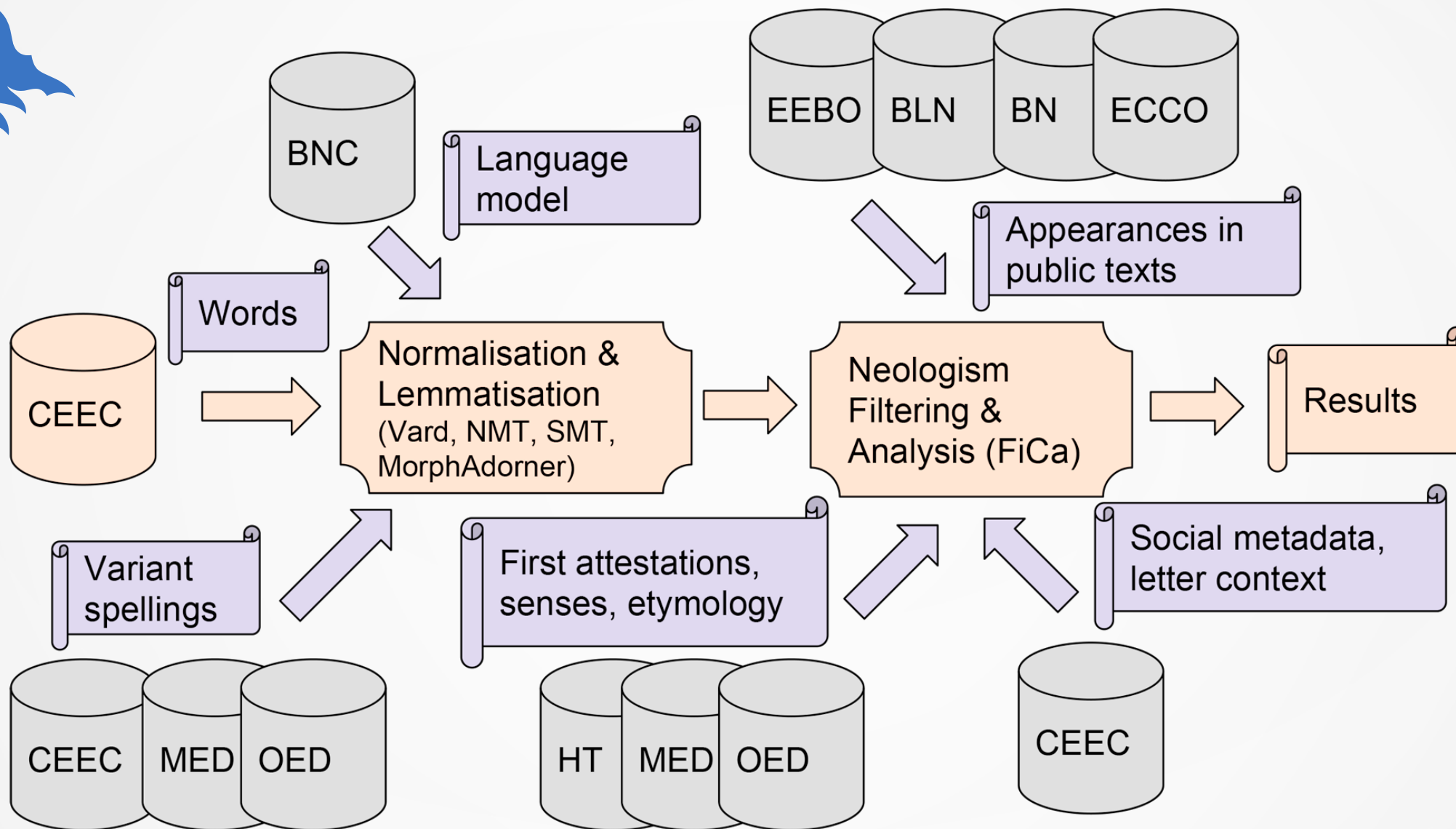
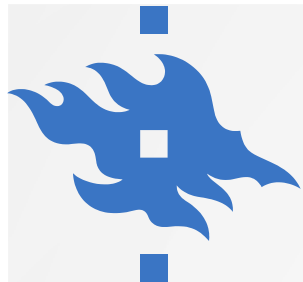
BIG-DATA APPROACH TO ANALYSING NEOLOGISMS

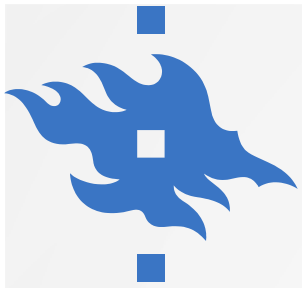


- Automatically map each word in the corpus to lexicographical data and contemporary published texts, compare first attestation dates
 - Spelling variation: SCEEK not enough, additional **normalization** required
- Automatic retrieval of related **lexicographical data**
 - *Oxford English Dictionary* (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



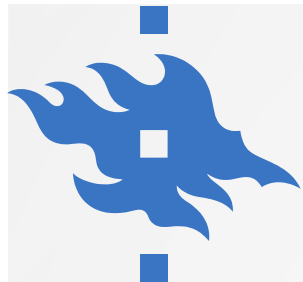


INITIAL MAPPING TO OED

- **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

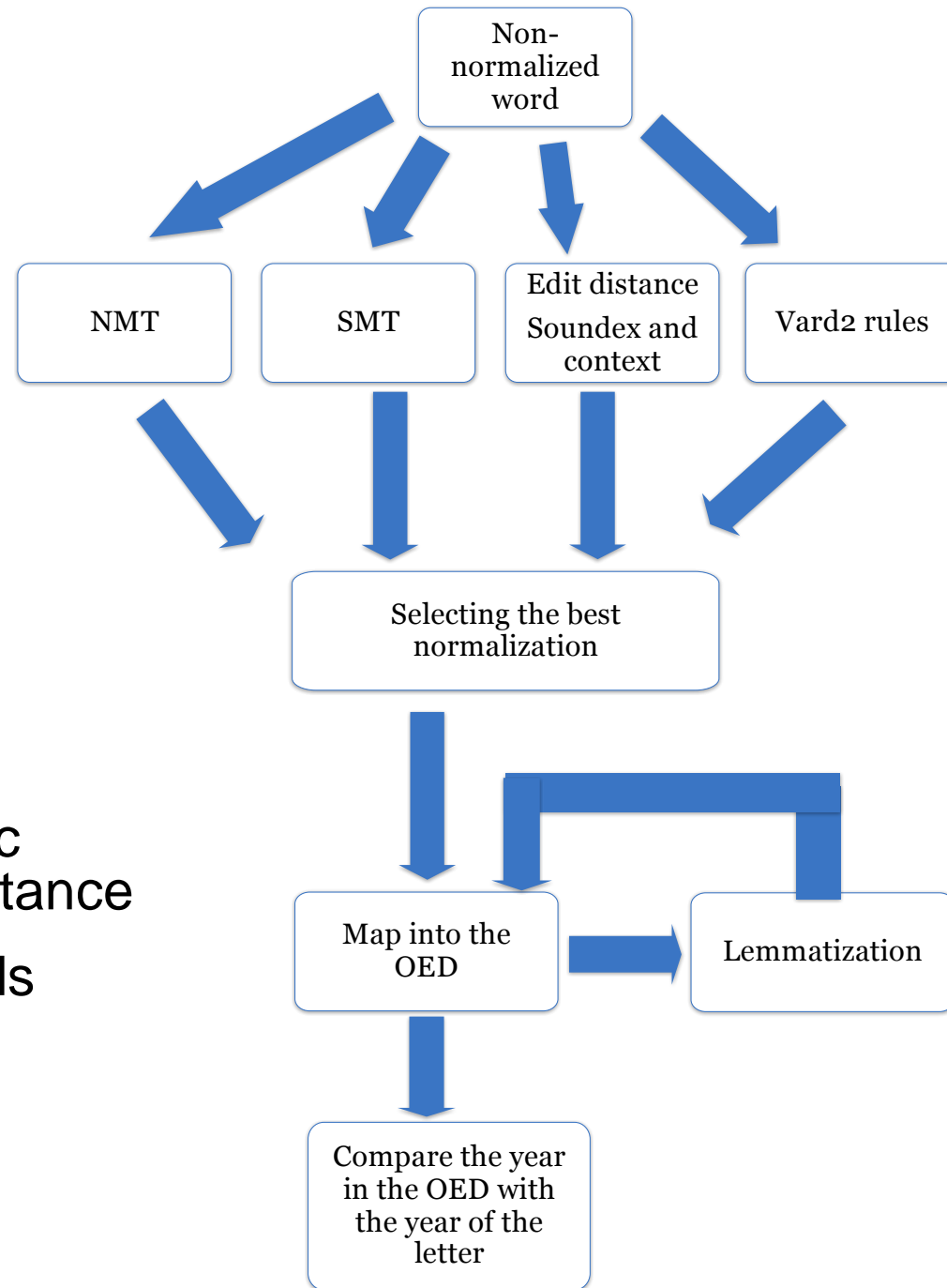
- Successfully mapped: c. 50,000 word forms, unmapped: c. **100,000**



ADDITIONAL NORMALIZATION

(HÄMÄLÄINEN ET AL. 2018)

- Idea: use machine translation!
 - **NMT** = neural machine translation (OpenNMT)
 - **SMT** = statistical machine translation (Moses)
 - Use known VARD2 normalizations 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 18TH CENTURY

Word	NMT	SMT	Edit distance	VARD2	Correct
elyzian	elizian	elysian	elysian		Elysian
arch-type	archedip	archetype	archetype		archetype
kindelyer	kindler	kinder	kindlier		kindlier
supp'd	supped	supply	supply	cupped	supped
prosprity	prosperity	prosperity	prosperity		prosperity
affectionett	affectiont	affection	affectionate		affectionate
driction	driction	diction	direction		direction
octvo	ctuoctoo	octal	oct, oct.		octavo
mic'lemas	micelemas	miles	micklemas		Michaelmas
midetrenian	midetrenian	mid-on			Mediterranean



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 18th-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...



INTERFACE FOR FILTERING THE NEOLOGISM CANDIDATES

- FiCa (Filtering and Categorization)
- Developed by Eetu Mäkelä; see Säily et al. (in press)

OED Lemma	Word	Category	Notes (62)	Earliest letter	OED - CEEC	Total DF before	Total TF before	Frequency
acharya	acharya	yes		1789	3	0	0	1
anthroponomical	anthropi	yes		1734	0	0	0	1
anti-democrat	anti-den	yes	different	1799	3	0	0	1
blueism	bluism	yes		1795	0	0	0	1
bonneted	bonnete	yes	even if V	1781	43	17	17	1
canicule	canicule	yes		1701	0	5	5	1
cardiphonia	cardiph	yes	actual 1s	1780	1	2	2	1
cast-off	cast-off	yes		1692	48	71	73	2
catchy	catchy	yes		1784	20	35	35	1
chaplaincy	chaplair	yes		1741	4	0	0	1
cleverality	cleveral	yes		1778	50	0	0	1
curtainless	curtainle	yes		1799	23	0	0	1
delineator	delineat	yes		1736	38	28	28	1
dicky-bird	dicky-bi	yes	figurative	1778	3	0	0	1
double-bedded	double-l	yes	double-t	1798	0	7	7	1
double-cross	double-	yes	different	1754	80	0	0	1
embodiment	embodii	yes		1777	51	2	2	1
envoyship	envoyst	yes		1706	30	0	0	1
eschantillon	eschant	yes		1717	3	5	5	1
escritoire	escritoir	yes		1694	13	3	3	1
freshen	fresheni	yes		1680	17	1	1	1
fussy	fussy	yes		1797	34	72	72	1

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

<Q A 1784 FN GCULLEY> <X GEORGE CULLEY> <P 179> [] [1..1] [^GEORGE CULLEY TO MATTHEW CULLEY^] Durham 1st October 1784 Mr Gill had so much to say about his tour into the west that I could not get from him last night. He has a nephew a son of Joseph, not of (^Arramathea^)^ but Shildon that he wants to put to us, to work and learn farming, we to find him victuals only, Mr Gill clothes, I told him I could say nothing as I had no house, and was rather cold about it, and said he wo... along. He says he does not kill half the large sheep he used to at this season and his neighbours the same. Bought the other day of the Cook &c near 400 small sheep and near 30 cattle. I think you must buy Mr John Robsons wethers, and may enquire about Mr Wilsons of Eslington. I now think dear as sheep are they will pay till Christmas, but am not fond of keeping all till spring, your own dinmonds &c must be kept. Barmton. **Mr Collings prize tup is not very capital to handle, but rather **catchy** to look at.** It looks like Mr Wastles judgement weakening Mr Collings says he has let one to George Lax &c better than him but not so pleasing to look at, he was gone so did not see him, let at 9 guineas also, Mr Colling will not be a marr trade. He is clear for letting at good prices or none. Mr Colling has some promising lambs, which I am glad off. Wheat a better crop here than I expected, oats a bad one, beans good and barley good, this last grain I am inclined to think has a chance to get...t to expect them untill they come. I was at Burdon, Mr Wastle poorly, Mrs Wastle very well. Mr Wastle says beef must be 2s 8d before March. Northallerton a pretty good shew of cattle, and not quite so high as have been. I send this by your friend Mr John Mason to Alnwick fair. Bob got here with the tups all well by 6 o'clock but I have not seen them. Mr Wastle says Mr Charge sent bad queys into Russia. Tell Nanny Brown I saw her brother Jack, who desired his love, and his wife is still poorly.

They havesabovi eT Sen tear highly commended for making the OED. Common, &c. being much used in letters offt!

Guy, of Duke-dtreet, York-buildings, Sur; **catchy** i geon, deceafed, begs leave to inform the pullie, that he has ,le, chi purchafed of Mr.

NTbhrfdar nert, the 12th lthflant, the O cleb x ted **CATCHY!**

.* But they are All of a piece; st yet they lye upon the **Catchy** to Trip up the Heels one of Another.* Prethee wilt thou make these Things Hang Together, now.

:rpondenta in' moll of the great towns in the ldnngnm, thore **Catchy** fulfilled ;; -for if th y- lhotild have any doubts with refepel to the .vho live in the country may be the fignarureori rbe padke'ts~t laey ma~y be: able to

OED Oxford English Dictionary Quick search: GO

Lost for Words? | [Advanced search](#) | [Help](#)

View full results for 'catchy' Help on Dictionary Entry | [Print](#) | [Save](#) | [Email](#) | [Cite](#)

catchy, *adj.*

View as: [Outline](#) | [Full entry](#) Quotations: [Show all](#) | [Hide all](#) | [Keywords: On](#) | [Off](#)

Pronunciation: Brit. ▶ /'kætʃi/, U.S. ▶ /'ketʃi/

Etymology: < CATCH *v.* + -Y *suffix*'.

colloq.

1. Adapted to catch the attention or fancy; attractive, 'taking'.

Text size: A A

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

[Entry history](#)
[Entry profile](#)

Previous version:
[OED2 \(1989\)](#)

In other

[Thesaurus](#) ▶
[Categories](#) ▶

Browse:

▶ [Dictionary](#) ▶ [Sources](#)

▶ [Categories](#) ▶ [Historical Texts](#)

▶ [Timelines](#)

My entries (13)

My searches (15)

Jump to:

▶

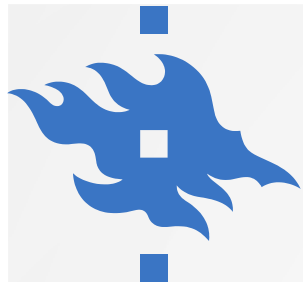
Entry ▼

catchiness, *n.*

catching, *n.*

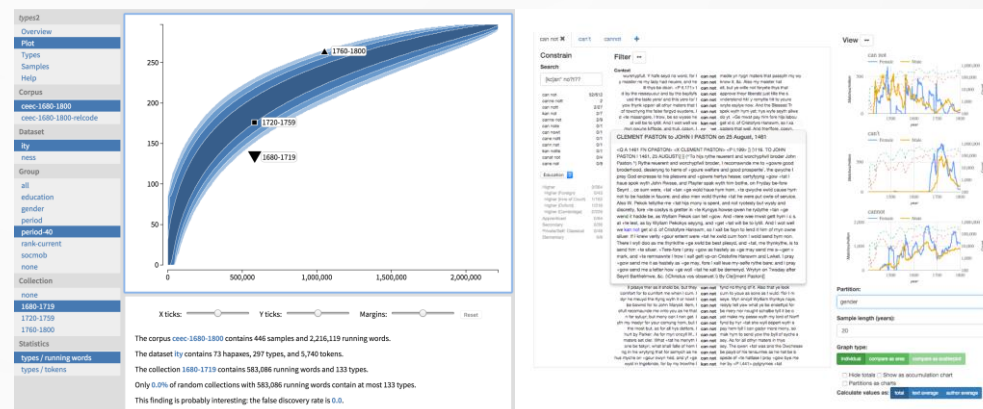
catching, *adj.*

catchment, *n.*



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





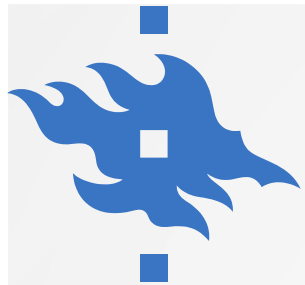
PILOT STUDY

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 \leq OED 1st attestation
 - Can occur in max 100 contemporary published texts before CEEC 1st attestation
- Automated procedures \rightarrow only 220 candidates for human to filter in interface
 - 81 neologisms found



WHO ARE THE INNOVATORS?

- 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)

Name	# of neologisms
Twining, Thomas	13
Austen, Jane	4
Bentham, Jeremy	4
Burney, Frances	4
Gray, Thomas	4
Cowper, William	3
Lennox, Sarah	3
Burney, Charles	2
Jackson, John	2
St. Michel, Balthasar	2
...	2



EXAMPLES

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)



Thomas Twining and Charles Burney
(© NPG; NPG D39475, NPG 3884)

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

Social rank	# of neologisms
Professional	26
Clergy, lower	22
Gentry, lower	13
Nobility	6
Gentry, upper	6
Other non-gentry	5
Clergy, upper	1
Merchant	1
Royalty	1



EXAMPLES

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)

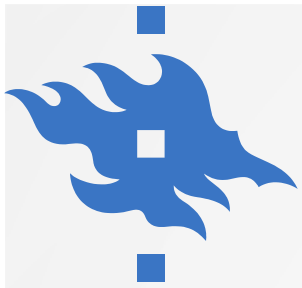


Ignatius Sancho & Princess Elizabeth

(Wikimedia Commons)

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)



GENDER, AGE, EDUCATION; REGISTER

- 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*)



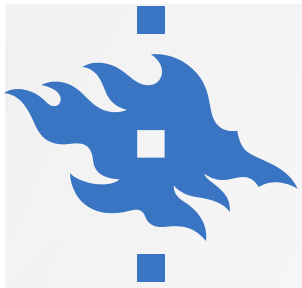
SEMANTICS

- **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

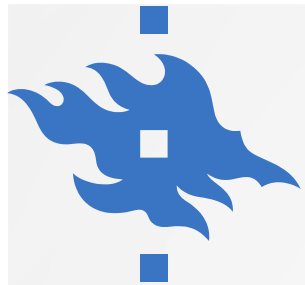
- 6 **the world** » action or operation
- 6 **the mind** » emotion
- 5 **society** » communication
- 4 **the world** » space
- 4 **the mind** » mental capacity
- 4 **the mind** » attention and judgement
- 3 **society** » trade and finance
- 3 **society** » occupation and work
- 3 **society** » faith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

...



CONCLUSION

- **Big-data approach:** massive databases & automated pipeline → possible to quickly discover of 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 1st 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

- Hämäläinen, Mika, Tanja Säily, Jack Rueter, Jörg Tiedemann & Eetu Mäkelä (2018). “Normalizing early English letters to Present-day English spelling”. Beatrice Alex, Stefania Degaetano-Ortlieb, Anna Feldman, Anna Kazantseva, Nils Reiter & Stan Szpakowicz (eds.), *Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (ACL Anthology W18-45), 87–96. Stroudsburg, PA: Association for Computational Linguistics.
- Keune, Karen (2012). *Explaining Register and Sociolinguistic Variation in the Lexicon: Corpus Studies on Dutch*. PhD dissertation, Radboud University Nijmegen.
- Säily, Tanja (2018). “Change or variation? Productivity of the suffixes *-ness* and *-ity*”. Terttu Nevalainen, Minna Palander-Collin & Tanja Säily (eds.), *Patterns of Change in 18th-century English: A Sociolinguistic Approach* (Advances in Historical Sociolinguistics 8), 197–218. Amsterdam: John Benjamins.
- Säily, Tanja, Eetu Mäkelä & Mika Hämäläinen (in press). “Explorations into the social contexts of neologism use in early English correspondence”. *Pragmatics and Cognition*, special issue on the dynamics of lexical innovation.
- Wolfson, Nessa (1990). “The bulge: a theory of speech behavior and social distance”. *Penn Working Papers in Educational Linguistics* 2(1). 55–83.



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, hummingly, hydrogenate, idlish, impracticability, incomed, incontestably, inexact, inside-outness, internment, intrepid, jargonic, jumpable, keyless, kibitka, knick-knackatory, letteret, malformation, mariturient, mevrouw, missish, monotonous, moon, moonery, nidgetty, non-papist, pacifist, paperless, pheasantry, pushery, rishi, schoolmasterishness, scratch-back, shockingly, silentious, slushy, spidery, sprawly, squeezezy, 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)