Variation and change in the productivity of BE *going to* V

in the *Corpus of Historical American English*, 1810–2009

Tanja Säily, Florent Perek & Jukka Suomela

Grammaticalization of BE *going to* V

———

1. *I'm going to the market to buy bananas*
   'motion with intention'
2. *I'm going to read your work tomorrow*
   'motionless intention'; EModE
3. *There's going to be some serious trouble here*
   'prediction'; LModE–PDE
   a. *You're going to* ***feel*** *very foolish* (mental verb; COHA, 1932)
   b. ***It****'s going to rain* (inanimate subject, *it*; COHA, 1811)
   c. *Father Paul was going to* ***be cheated*** *of his share* (passive voice; COHA, 1946)

   (Budts & Petré 2016; Wu et al. 2016)

Research questions

---

1. How is the grammaticalization reflected in the **productivity** of the construction in LModE–PDE?
    - Internal factors: semantics of the verb (including mental verbs), inanimate subject (*it*), passive voice

2. Did the social factor of **gender** play a role in the process?

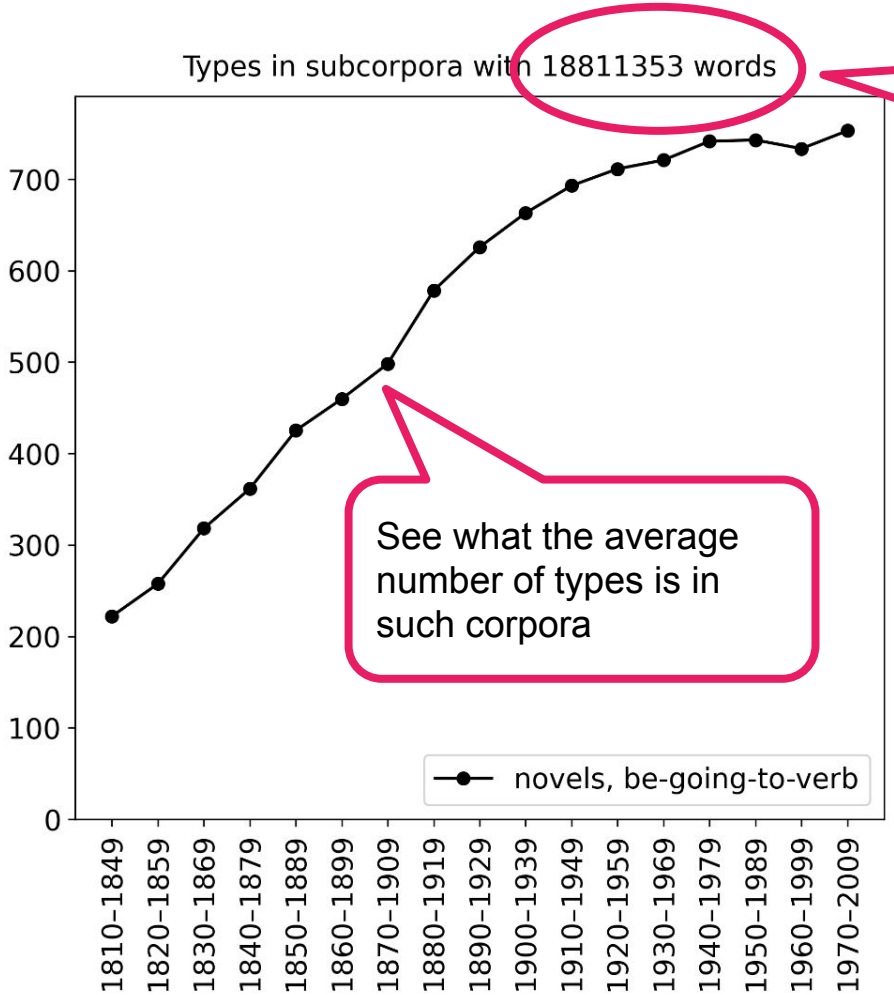
Material

- *Corpus of Historical American English* (COHA)
  - 400 Mw, 1810–2009
- Fiction section: c. 50% of the data
  - **Gender metadata** for authors developed by Öhman et al. (2019)
  - Promising material for sociolinguistic investigation: a more speech-like genre (dialogue)
  - Types of fiction (e.g. short stories, drama, movie scripts) unevenly distributed over time (Säily & Vartiainen forthcoming)
    → restriction to **novels only, c. 150 Mw**
- List of mental verbs from Halliday & Matthiessen (2014: 256–257)

Analysis 1: type frequencies

Methods

- We study the productivity of BE *going to* V by studying **type frequencies**
  - I.e. how many different verbs follow BE *going to* in different time periods
- Key challenges:
  - Different amounts of text from different time periods, different amounts of text from men and women: how to **compare** type frequencies?
  - If we observe trends, are they **statistically significant**?



Choose **random subcorpora** with the same number of words from each time period

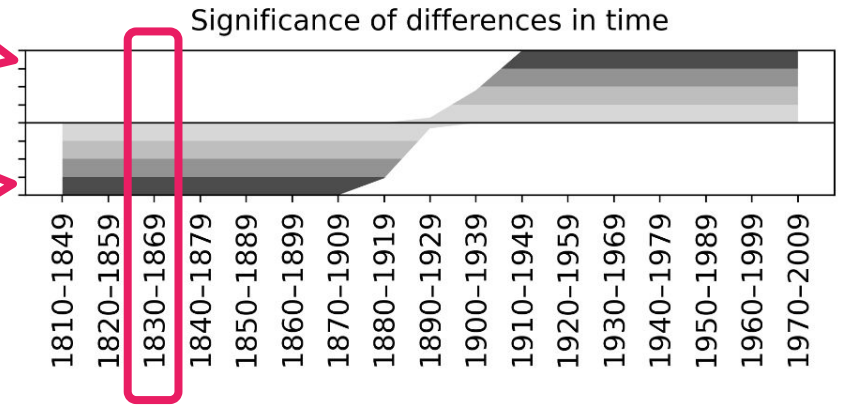
See what the average number of types is in such corpora

Visualizing trends

Assessing statistical significance

These periods have significantly many types

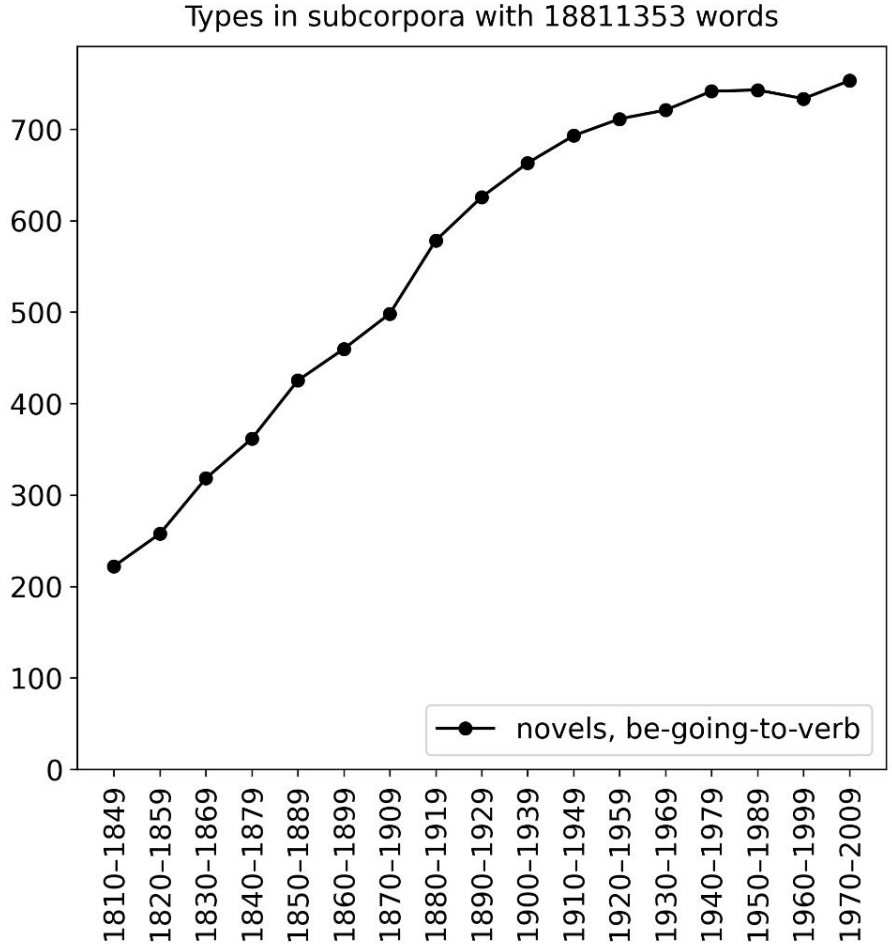
These periods have significantly few types

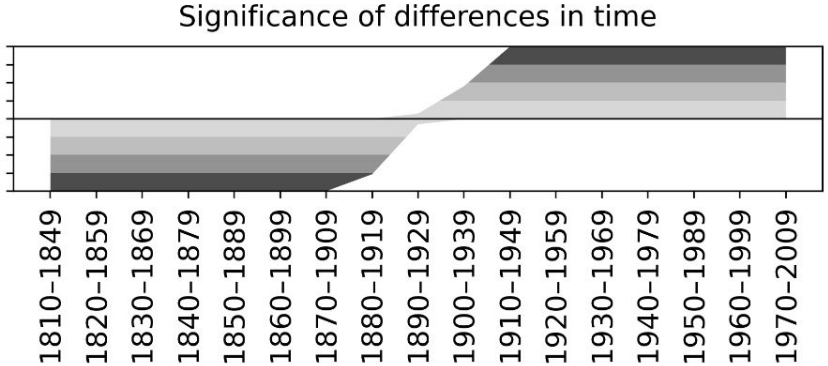


For each period (using **all** of the data):

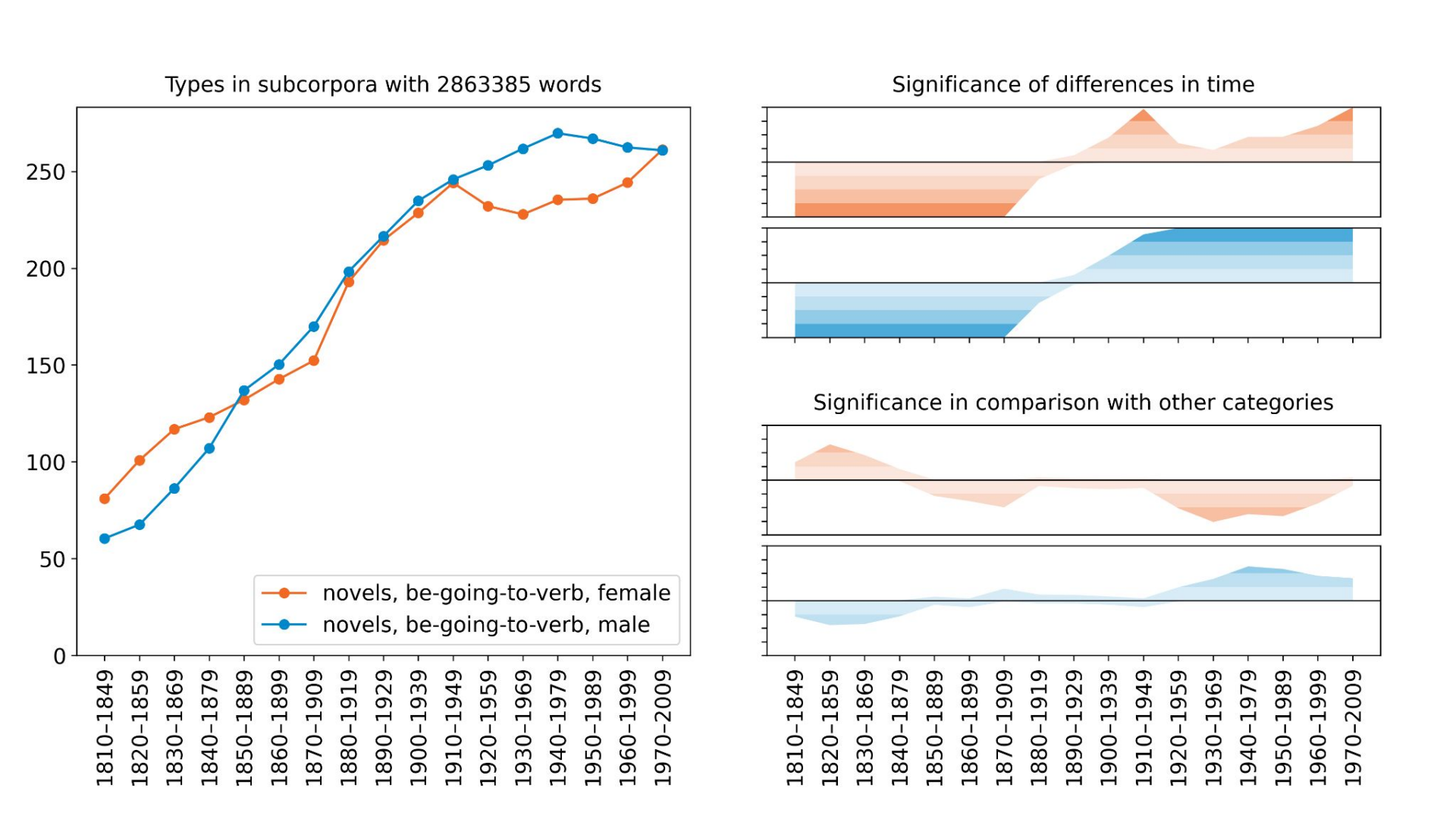
Sample random subcorpora from the whole corpus until you have a subcorpus of a comparable size

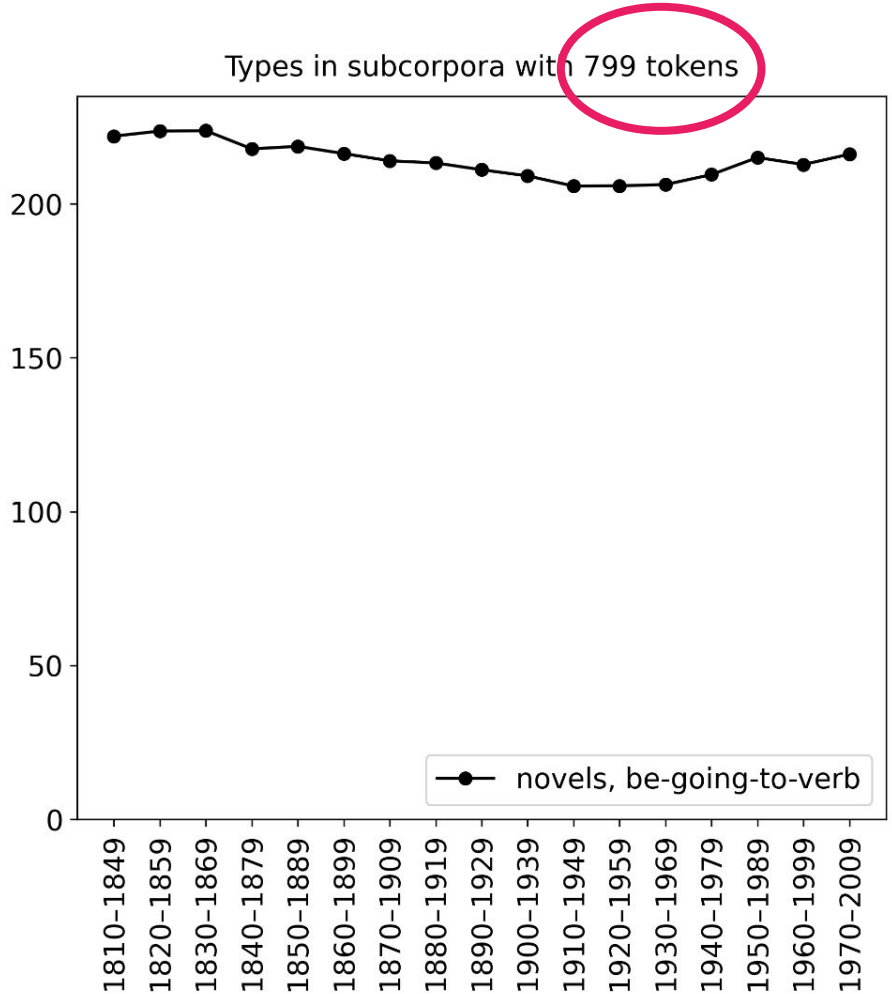
Do you typically get more or fewer types?

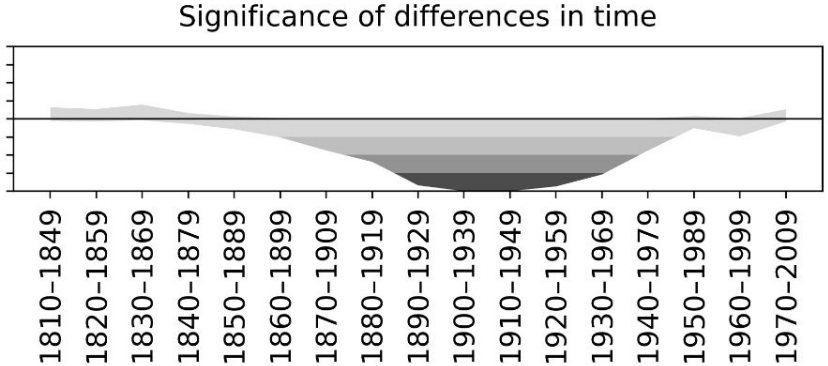




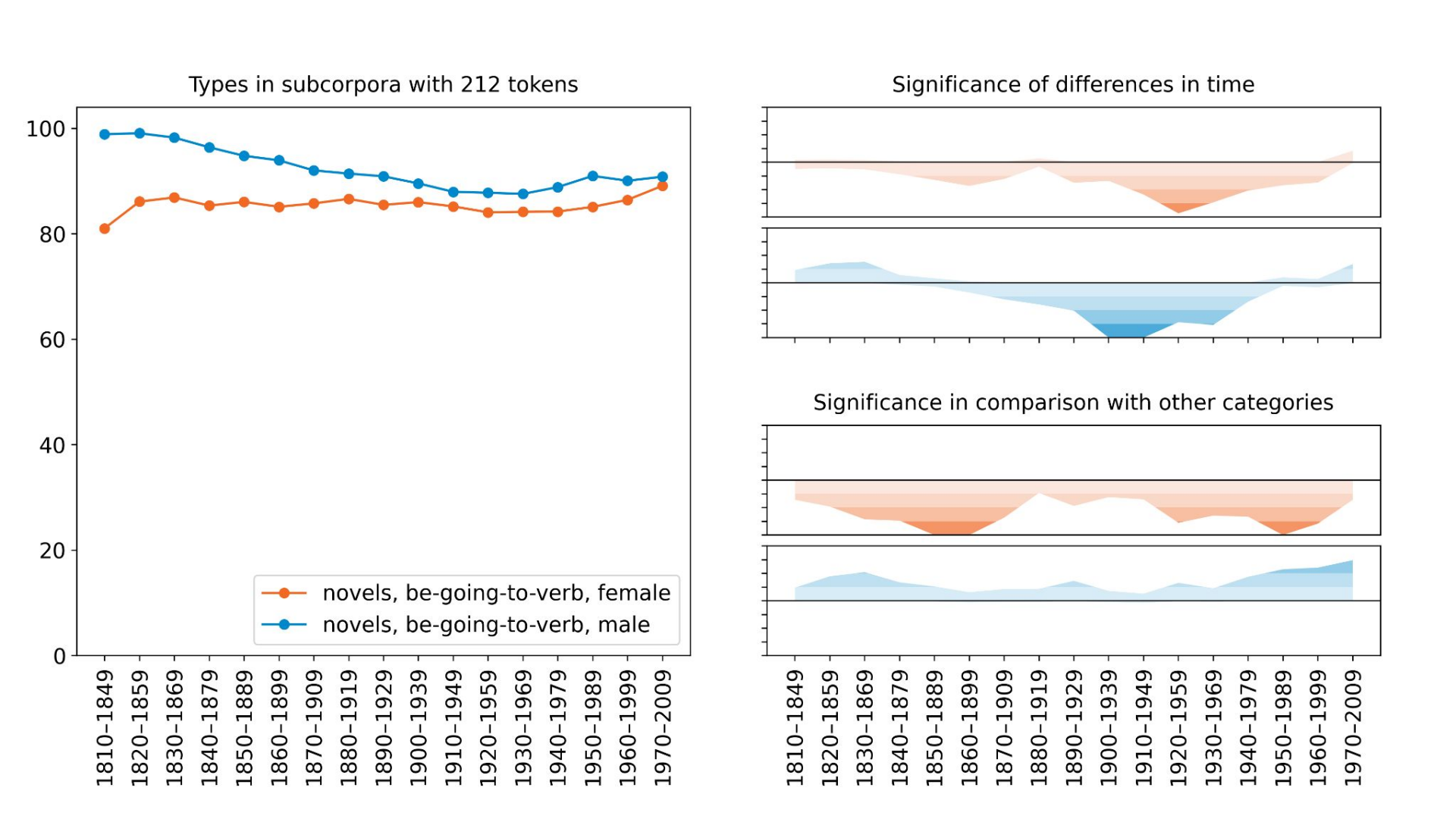
A clear increasing trend that is also statistically significant

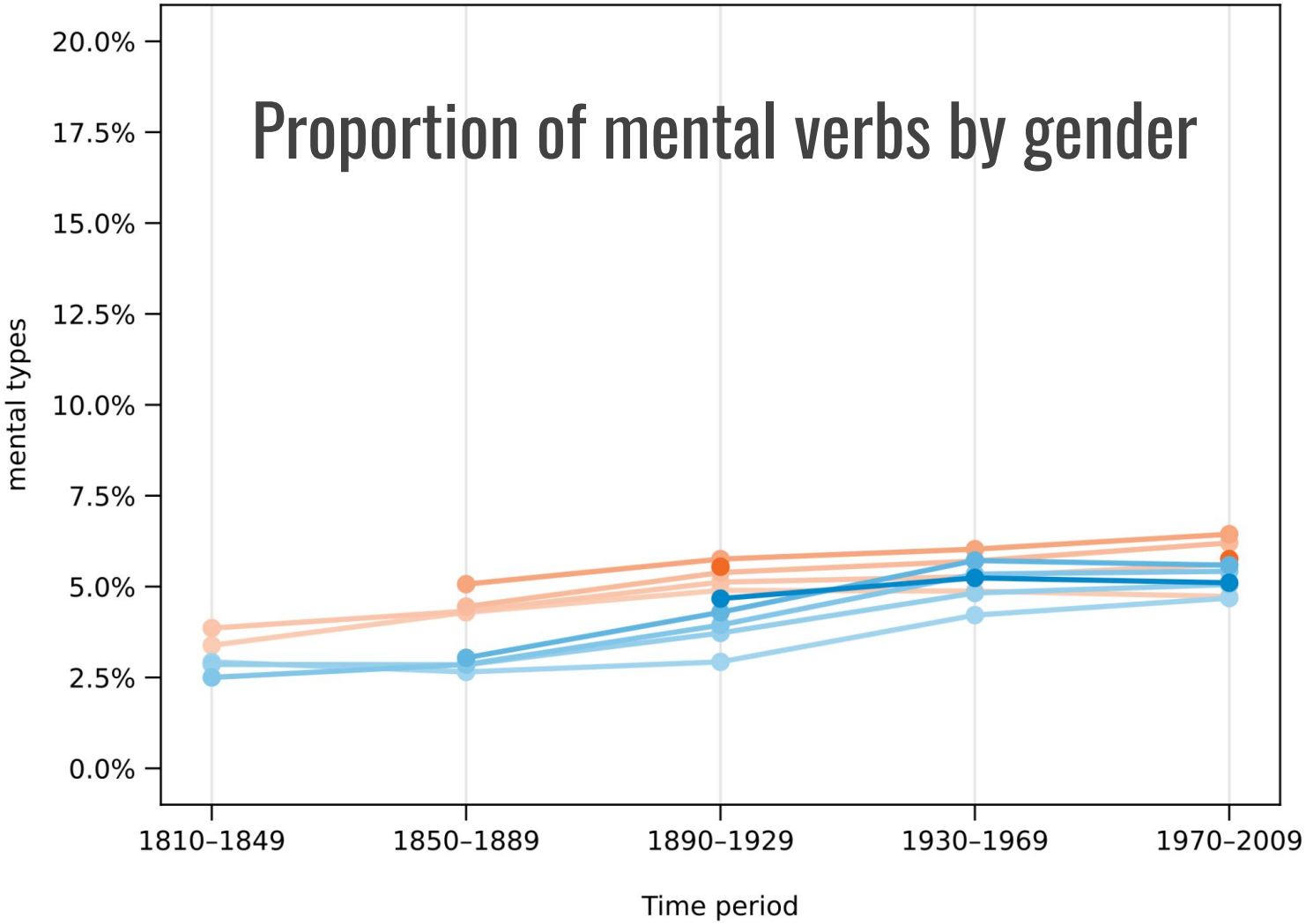

Types in subcorpora with 2863385 words
0
50
100
150
200
250
novels, be-going-to-verb, female
novels, be-going-to-verb, male
1810–1849
1820–1859
1830–1869
1840–1879
1850–1889
1860–1899
1870–1909
1880–1919
1890–1929
1900–1939
1910–1949
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009
Significance of differences in time
Significance in comparison with other categories
1810–1849
1820–1859
1830–1869
1840–1879
1850–1889
1860–1899
1870–1909
1880–1919
1890–1929
1900–1939
1910–1949
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009

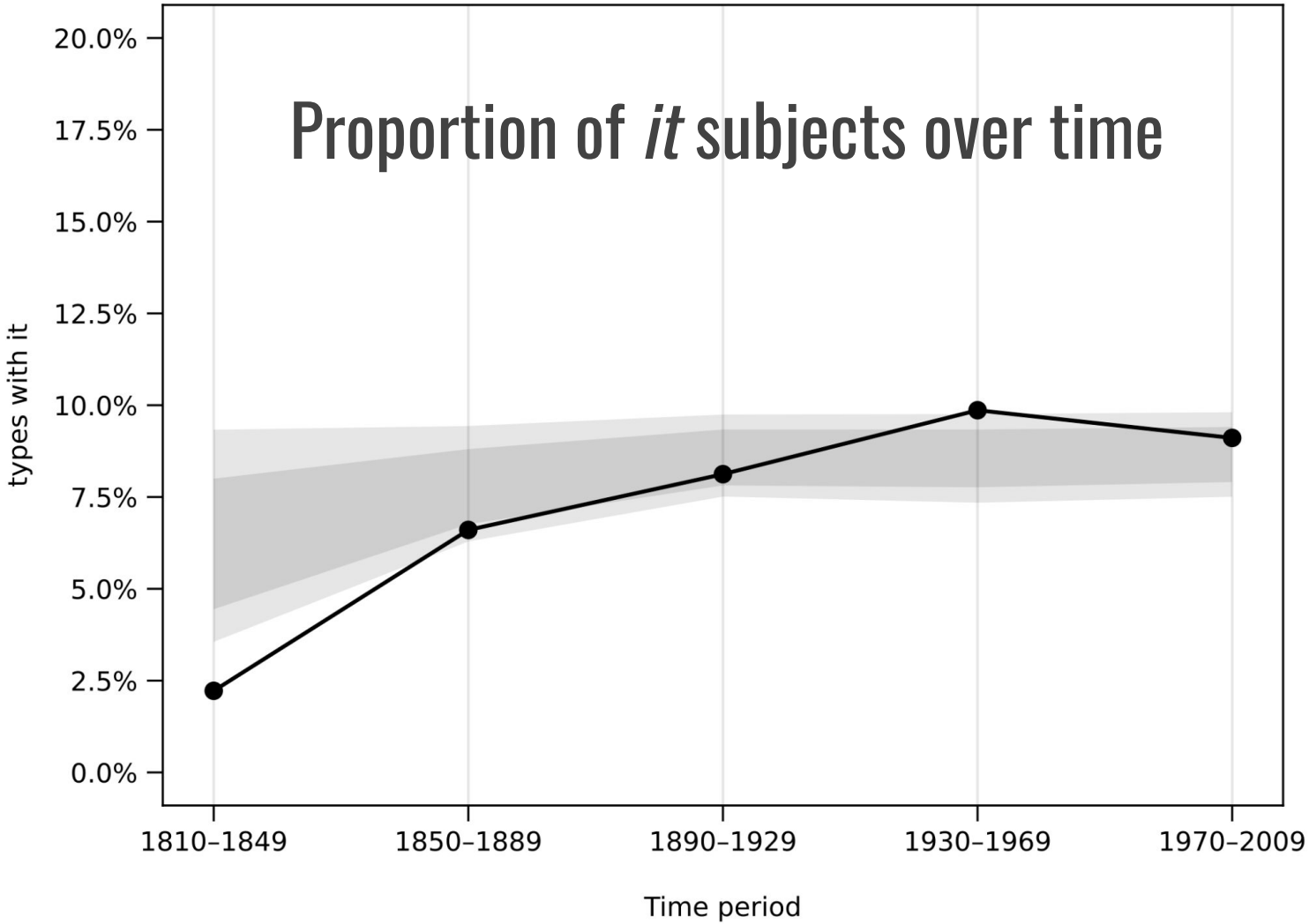


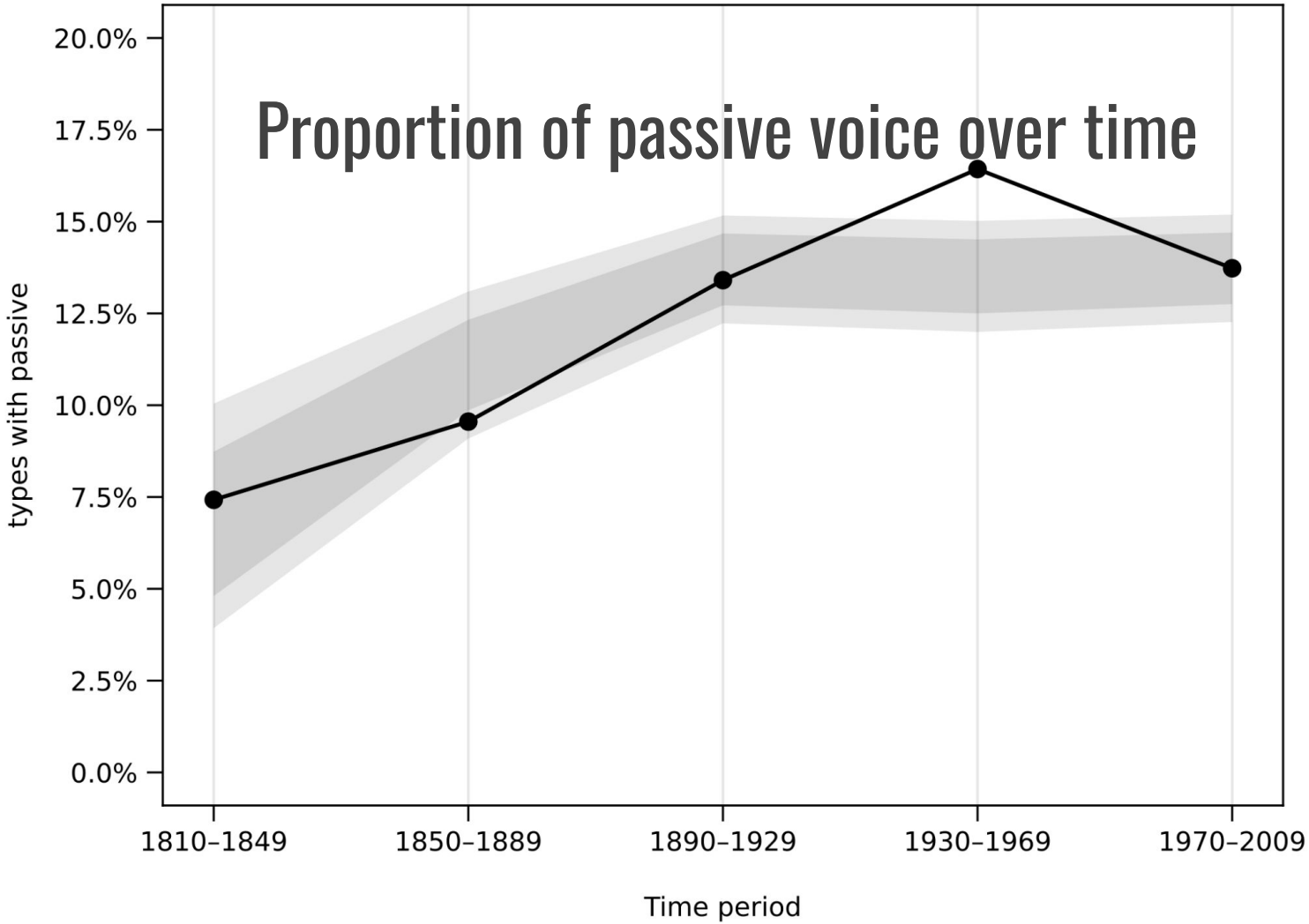


More frequent use or more diverse use?


Types in subcorpora with 212 tokens
100
80
60
40
20
0
novels, be-going-to-verb, female
novels, be-going-to-verb, male
1810–1849
1820–1859
1830–1869
1840–1879
1850–1889
1860–1899
1870–1909
1880–1919
1890–1929
1900–1939
1910–1949
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009
Significance of differences in time
Significance in comparison with other categories
1810–1849
1820–1859
1830–1869
1840–1879
1850–1889
1860–1899
1870–1909
1880–1919
1890–1929
1900–1939
1910–1949
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009


Proportion of mental verbs by gender
20.0%
17.5%
15.0%
12.5%
10.0%
7.5%
5.0%
2.5%
0.0%
mental types
1810–1849
1850–1889
1890–1929
1930–1969
1970–2009
Time period


Proportion of *it* subjects over time
20.0%
17.5%
15.0%
12.5%
10.0%
7.5%
5.0%
2.5%
0.0%
types with it
1810–1849
1850–1889
1890–1929
1930–1969
1970–2009
Time period


Proportion of passive voice over time
20.0%
17.5%
15.0%
12.5%
10.0%
7.5%
5.0%
2.5%
0.0%
types with passive
1810–1849
1850–1889
1890–1929
1930–1969
1970–2009
Time period

Type frequency

- Slight decrease in types over time, especially for men
- Type counts for men and women converge over time
- But what *kind* of types are they?
  - Is the construction undergoing semantic specialization?
  - Do men and women use it in different semantic areas?
- We examine these questions using ***distributional semantics***

Analysis 2: distributional semantics

Distributional semantics

> “You shall know a word by the company it keeps”
> Firth (1957: 11)

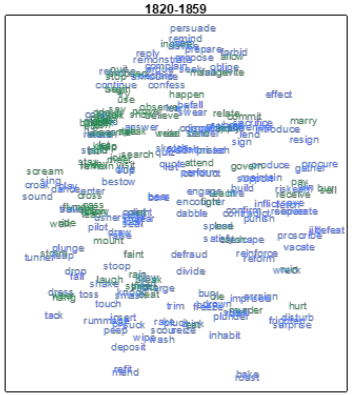
- Aim = capturing word meaning through lexical collocates in large text corpora
- Semantically similar words are expected to have the same collocates
  - e.g. *drink* and *sip* > *wine*, *water*, *coffee*, *cup*, *bottle*, etc.
- Semantic similarity is approximated by similarity in distribution

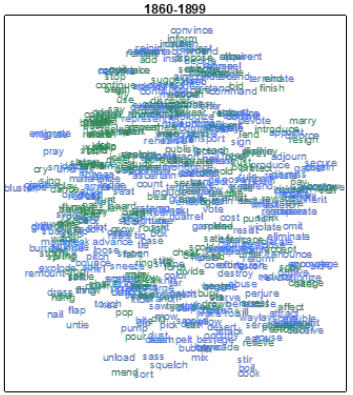
Distributional semantic model

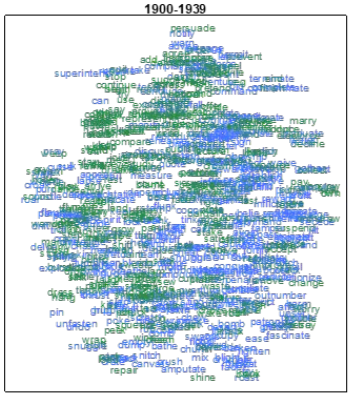
- DSM built with word2vec (SkipGram, cf. Mikolov et al. 2013), using gensim
- Trained on the whole COHA, context window +/- 2 words
- Each word is assigned a “vector”, i.e. array of values
- This quantification of meaning allows us to (*inter alia*):
  - Visualise the semantic distance between a set of words by plotting them in two dimensions (using e.g. t-SNE) (Perek 2016, 2018)
  - Measure and compare the semantic spread of constructions (Hilpert & Perek 2022)

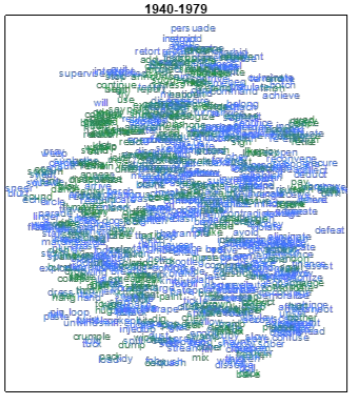
Distributional semantic plots, whole corpus

Men

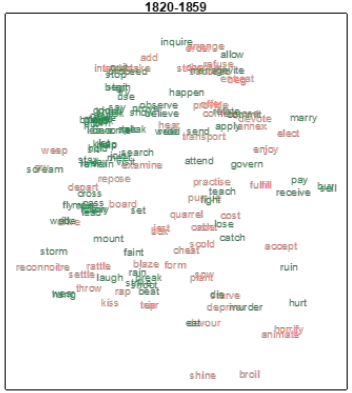


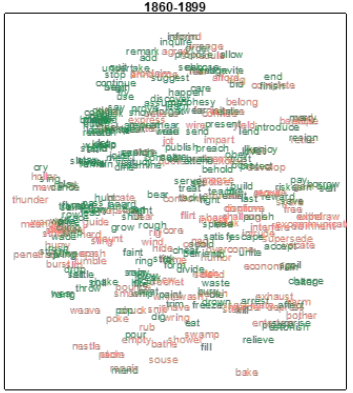


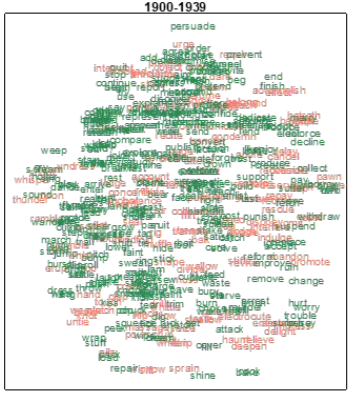


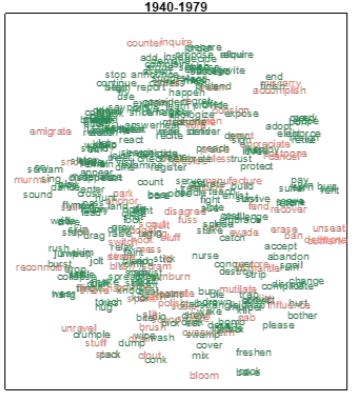


Women



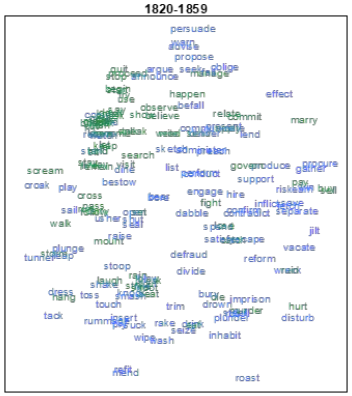


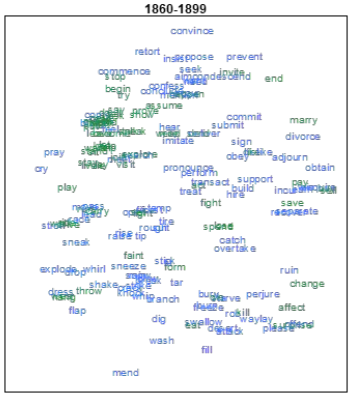


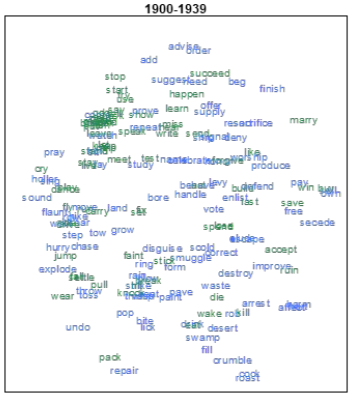


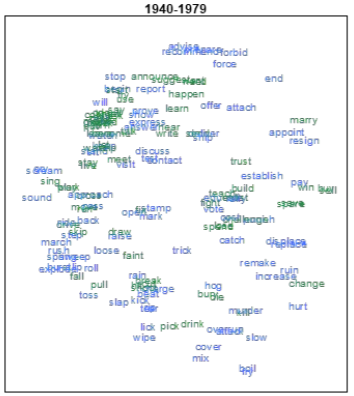
Using random samples (matched for frequency, N = 534)

Men

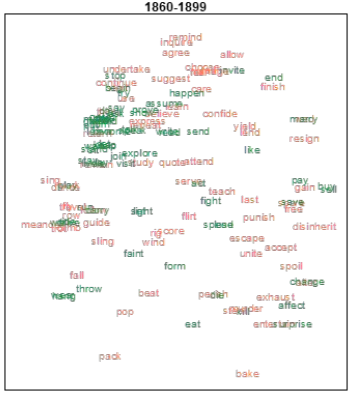






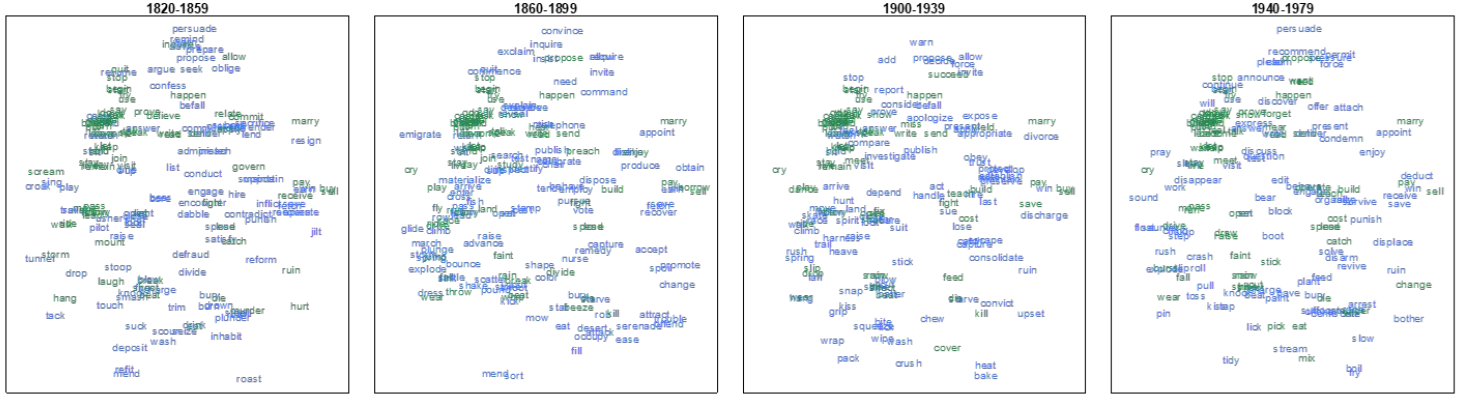


Women

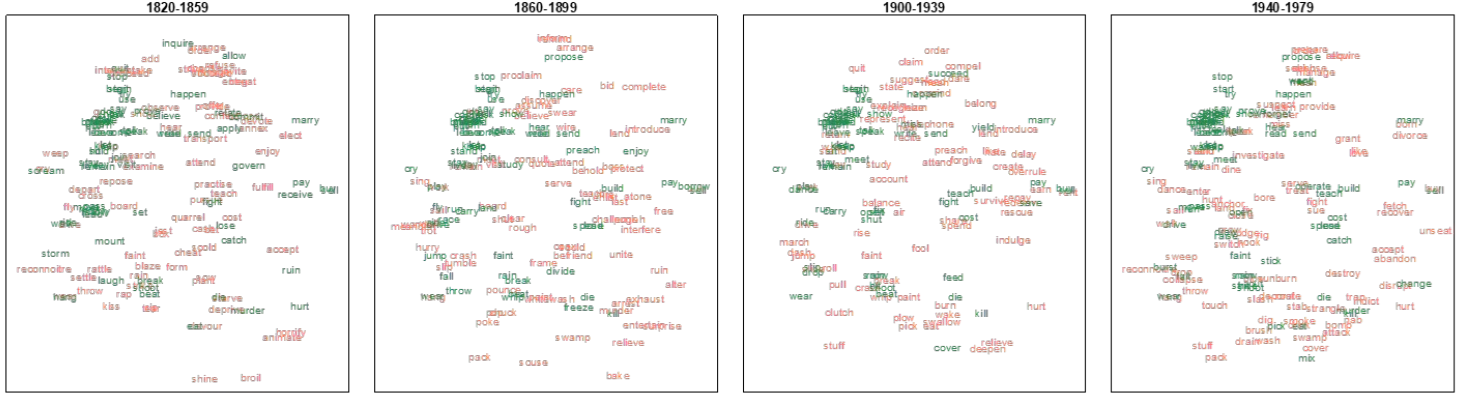


Using random samples #2 (N = 534)

Men

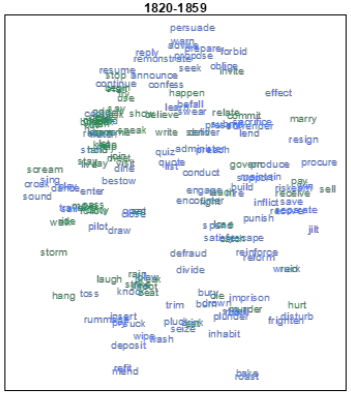


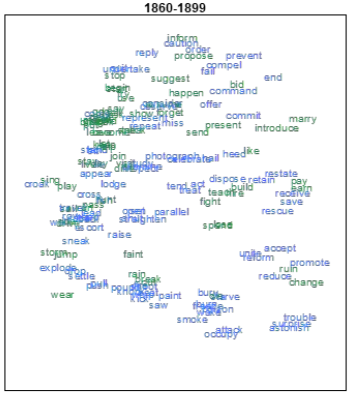
Women

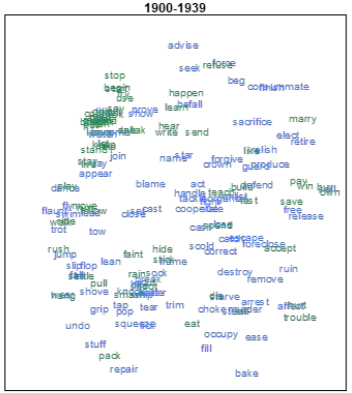


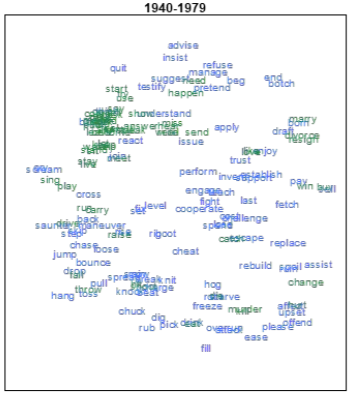
Using random samples #3 (N = 534)

Men

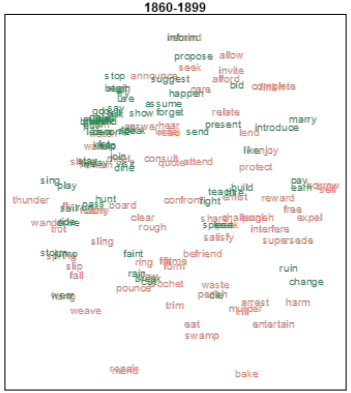


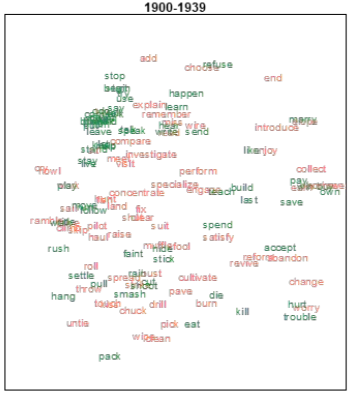






Women





Qualitative type-based analysis

- Type distribution highly variable from sample to sample
- Problem = we cannot average over individual types!
  - But we can average over type counts
  - We just need to add a semantic dimension to type counts
- Idea: types are sorted into discrete semantic categories
  - We can average over type counts in each category across samples
  - This gives us a representation of the average “semantic spread” of the construction

Qualitative type-based analysis

- We collect all types in the random samples (1419 types)
- We extract pairwise semantic similarity scores between these types from the DSM
- We use these scores to automatically group types into semantic categories using cluster analysis (PAM)

Qualitative type-based analysis

- In each cluster, we calculate the average number of types attested in each period across the 1000
- Similarity between type distributions can be measured using Pearson's correlation coefficient (ρ)
  - Between different periods
  - Between genders in the same period

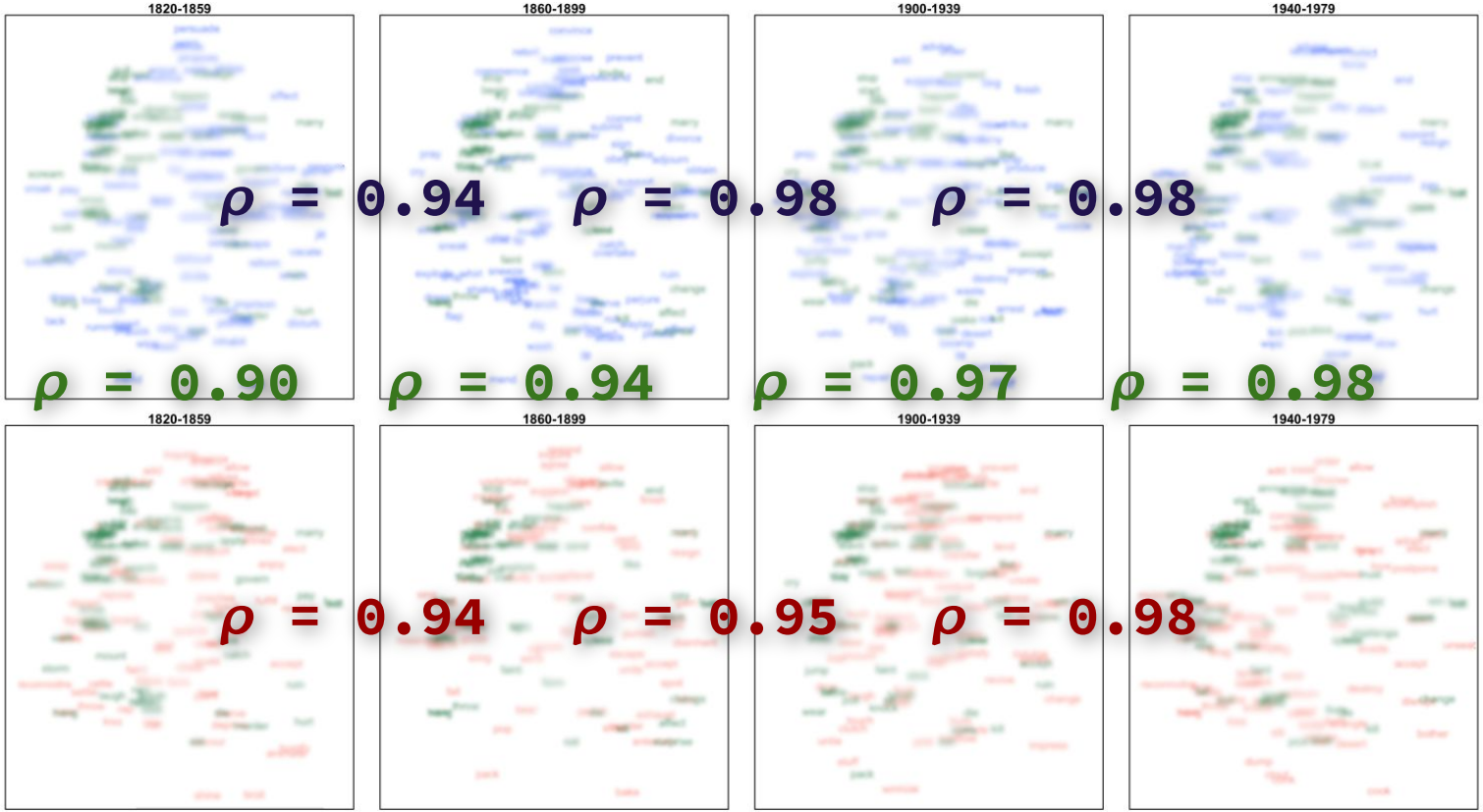
	1820-1859	1860-1899	1900-1939	1940-1979
Cluster 1	11.467	10.841	10.885	10.626
Cluster 2	6.434	5.892	6.958	7.35
Cluster 3	19.707	19.289	18.35	16.016
...	...	...	...	...

$\rho = 0.94$

Variation in semantic spread (30 clusters)

Men

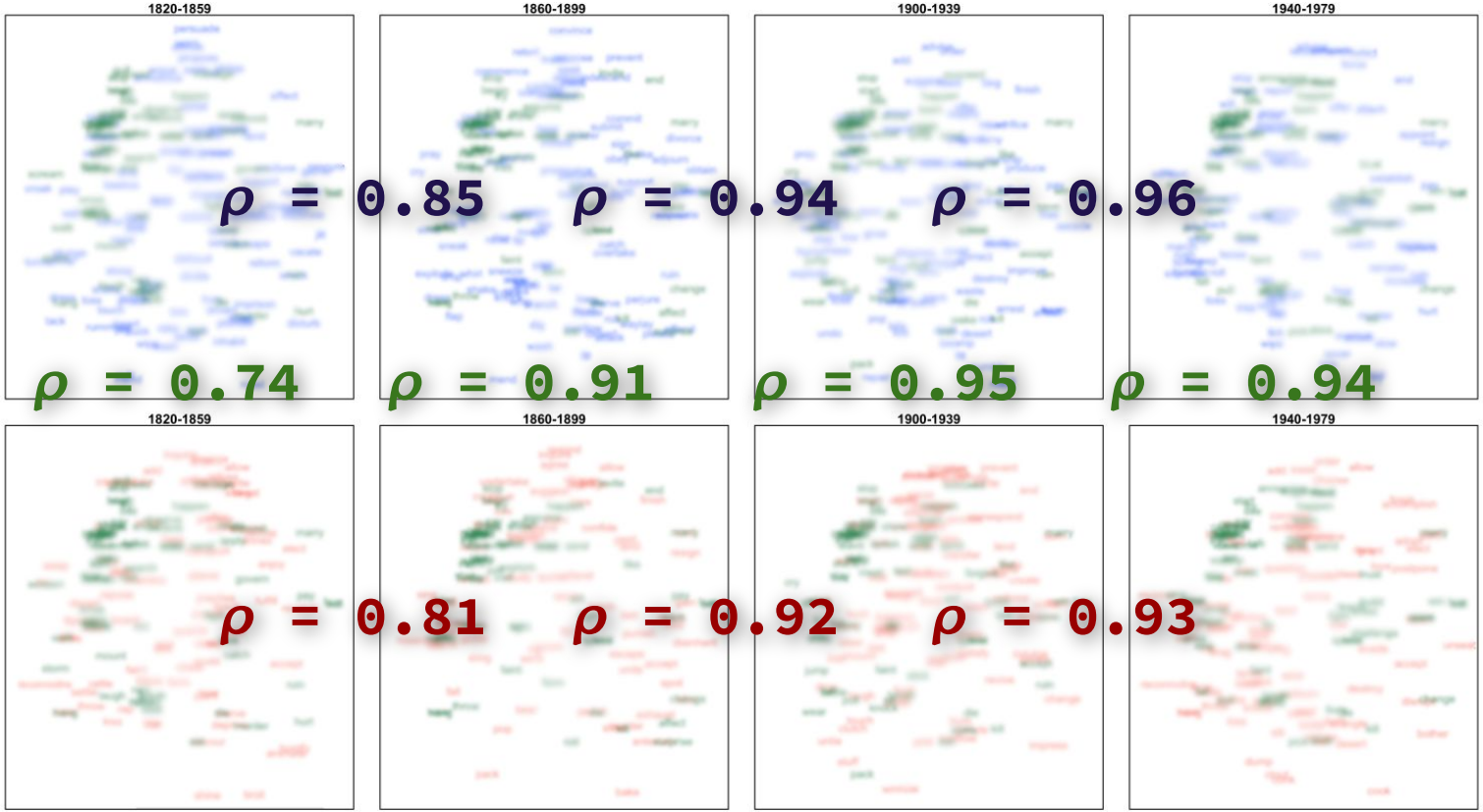
Women

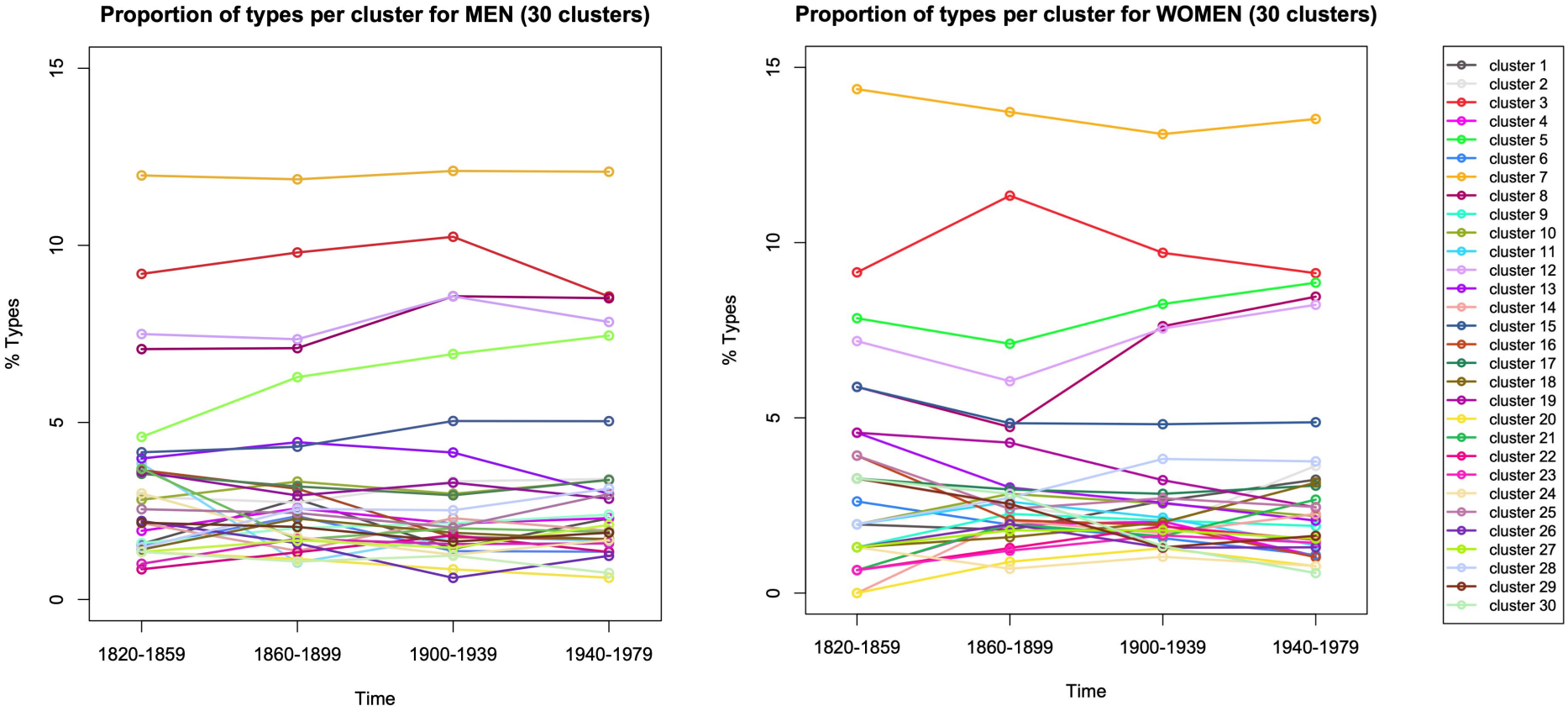


Variation in semantic spread (200 clusters)

Men

Women

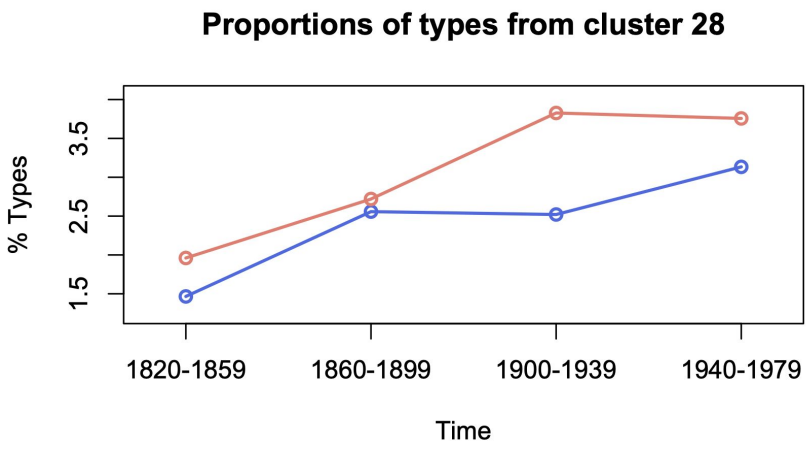


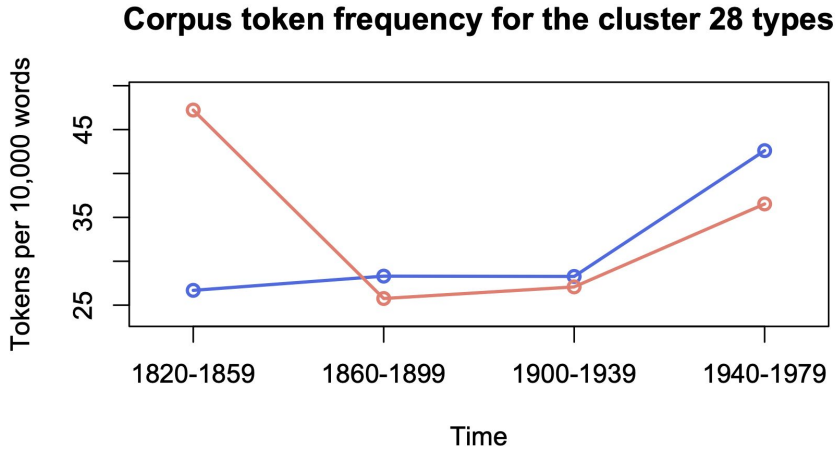

Proportion of types per cluster for MEN (30 clusters)
% Types
0
5
10
15
1820-1859
1860-1899
1900-1939
1940-1979
Time
Proportion of types per cluster for WOMEN (30 clusters)
% Types
0
5
10
15
1820-1859
1860-1899
1900-1939
1940-1979
Time
cluster 1
cluster 2
cluster 3
cluster 4
cluster 5
cluster 6
cluster 7
cluster 8
cluster 9
cluster 10
cluster 11
cluster 12
cluster 13
cluster 14
cluster 15
cluster 16
cluster 17
cluster 18
cluster 19
cluster 20
cluster 21
cluster 22
cluster 23
cluster 24
cluster 25
cluster 26
cluster 27
cluster 28
cluster 29
cluster 30

Cluster 28: mental verbs (cognitive type)

learn accept feel believe forget explain satisfy remember discover suit regret understand ignore mind realize fear penetrate unravel suspect guess recognize appreciate heed interpret describe wonder respect judge notice acquaint suppose

(+ stop mention solve belong materialize exist excuse ...)

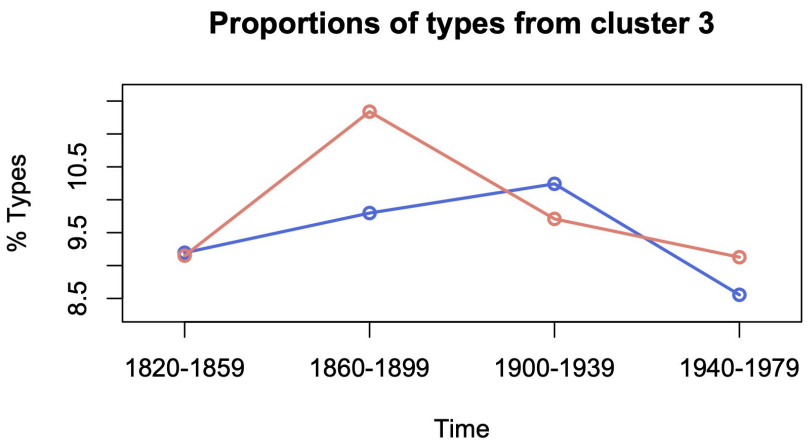


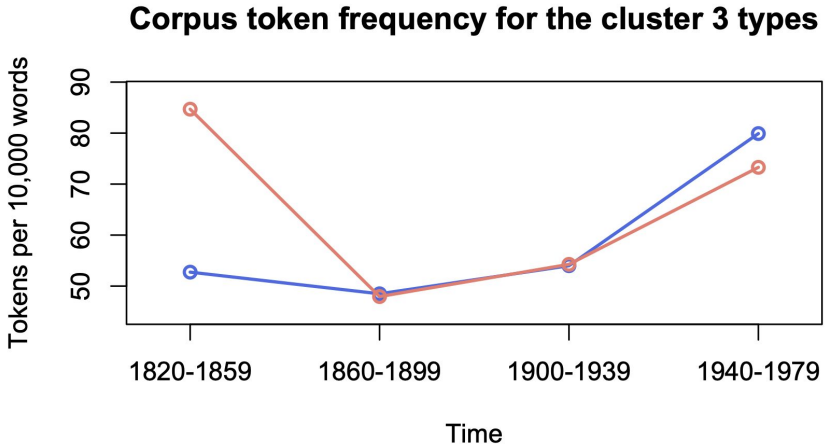


Cluster 3: verbs of motion

leave run drive ride walk follow move dance sail enter cross travel step march head hurry hike wander wade trot sneak swim stumble meander shuffle stride stroll approach descend skate parade trudge ramble circle tread amble paddle stalk saunter sprint

(+ stay live play carry stand pass sit wait lead watch listen row usher face ...)





Discussion

Summary of results

- Overall productivity/type diversity of BE *going to* V doesn't increase in C19-20 AmE, even a slight decrease
  - Men's usage more productive, convergence over time
- Internal factors do indicate increasing productivity
  - Proportion of types with mental verbs, *it* subjects, passive voice
  - Women lead increase in mental verbs
- Type-based semantic analysis identifies areas of growth
  - E.g. mental verbs, motion verbs
  - Points to an increase in grammaticalization
  - Gender differences as well, with women leading the way

Conclusions

- At this stage of grammaticalization, overall type diversity stagnates but **internal factors** linked to grammaticalization indicate increasing productivity
  - Important to take into account
- Consistent **gender** differences – different leaders of change and/or different genres?
  - Gender cannot be ignored as a possible factor
  - Mental verbs could be linked to women's involved writing style (Biber & Burges 2000)
- Future work: analyse hapax legomena / new types

References

Biber, Douglas & Jená Burges. 2000. Historical change in the language use of women and men: Gender differences in dramatic dialogue. *Journal of English Linguistics* 28(1): 21–37.

Budts, Sara & Peter Petré. 2016. Reading the intentions of *be going to*: On the subjectification of future markers. *Folia Linguistica Historica* 37: 1–32.

Halliday, M.A.K. & Christian M.I.M. Matthiessen (2014). *Halliday's introduction to functional grammar*, 4th edition. Routledge.

Hilpert, Martin & Florent Perek. 2022. You don't get to see that every day: On the development of permissive *get*. *Constructions and Frames* 14(1): 14–41.

Öhman, Emily, Tanja Säily & Mikko Laitinen. 2019. Towards the inevitable demise of *everybody*? A multifactorial analysis of *-one/-body/-man* variation in indefinite pronouns in historical American English. 40th Annual Conference of the International Computer Archive of Modern and Medieval English (ICAME 40), Neuchâtel, Switzerland, June 2019. https://tanjasaily.fi/talks/icame40_ohman_et_al_2019.pdf

Perek, Florent. 2016. Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics* 54(1): 149–188.

Perek, Florent. 2018. Recent change in the productivity and schematicity of the way-construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory* 14(1): 65–97.

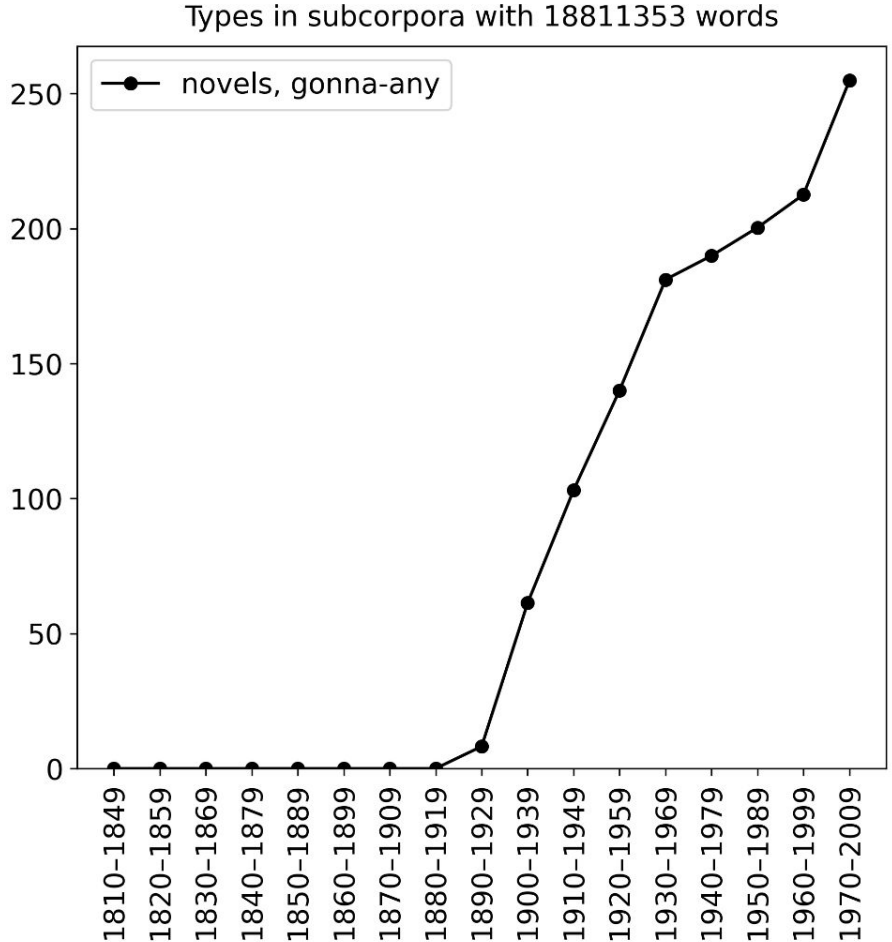
Rodríguez-Puente, Paula, Tanja Säily & Jukka Suomela. 2022. New methods for analysing diachronic suffix competition across registers: How *-ity* gained ground on *-ness* in Early Modern English. *International Journal of Corpus Linguistics* 27(4): 506–528.

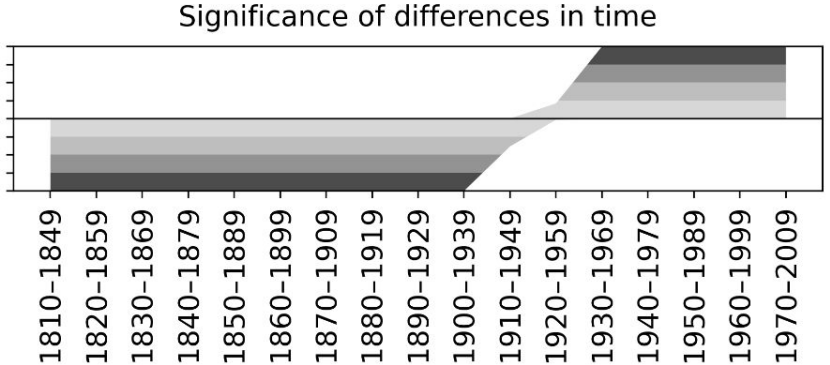
Säily, Tanja, Martin Hilpert & Jukka Suomela. Forthcoming. New approaches to investigating change in derivational productivity: Gender and internal factors in the development of *-ity* and *-ness*, 1600–1800. In Patricia Ronan, Theresa Neumaier, Lisa Westermayer, Andreas Weilinghoff & Sarah Buschfeld (eds.), *Crossing boundaries through corpora: Innovative approaches to corpus linguistics*. Benjamins.

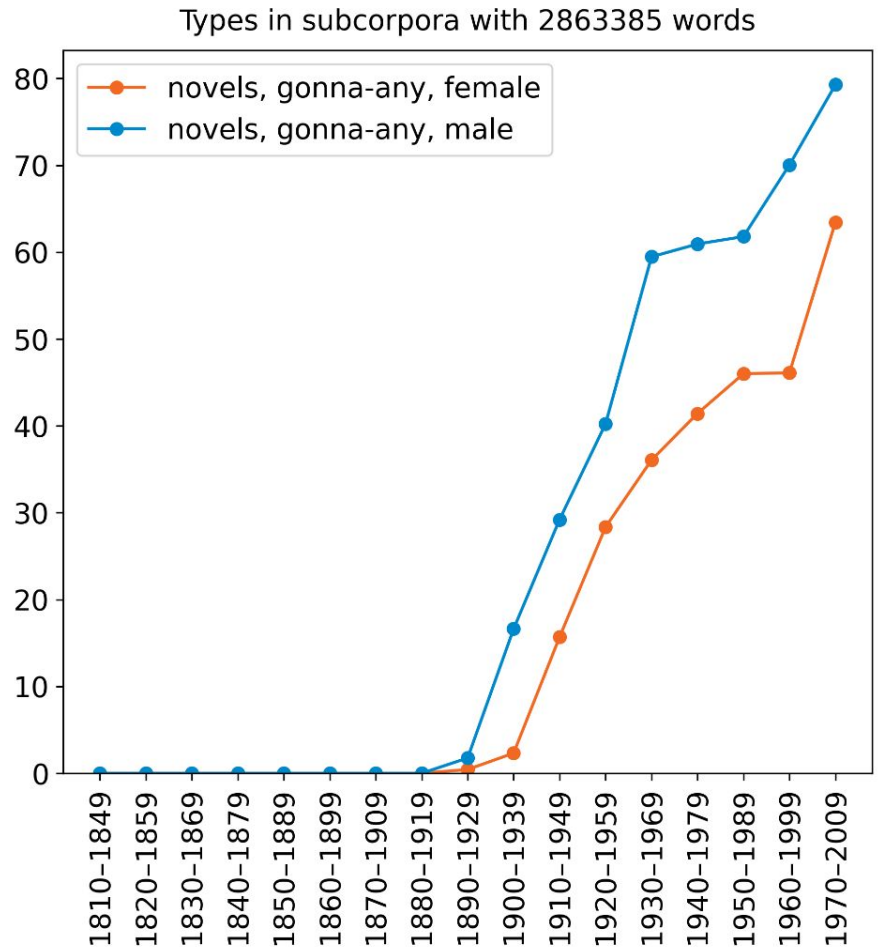
Säily, Tanja & Turo Vartiainen. Forthcoming. Historical linguistics. In Michaela Mahlberg & Gavin Brookes (eds.), *Bloomsbury handbook of corpus linguistics*. Bloomsbury.

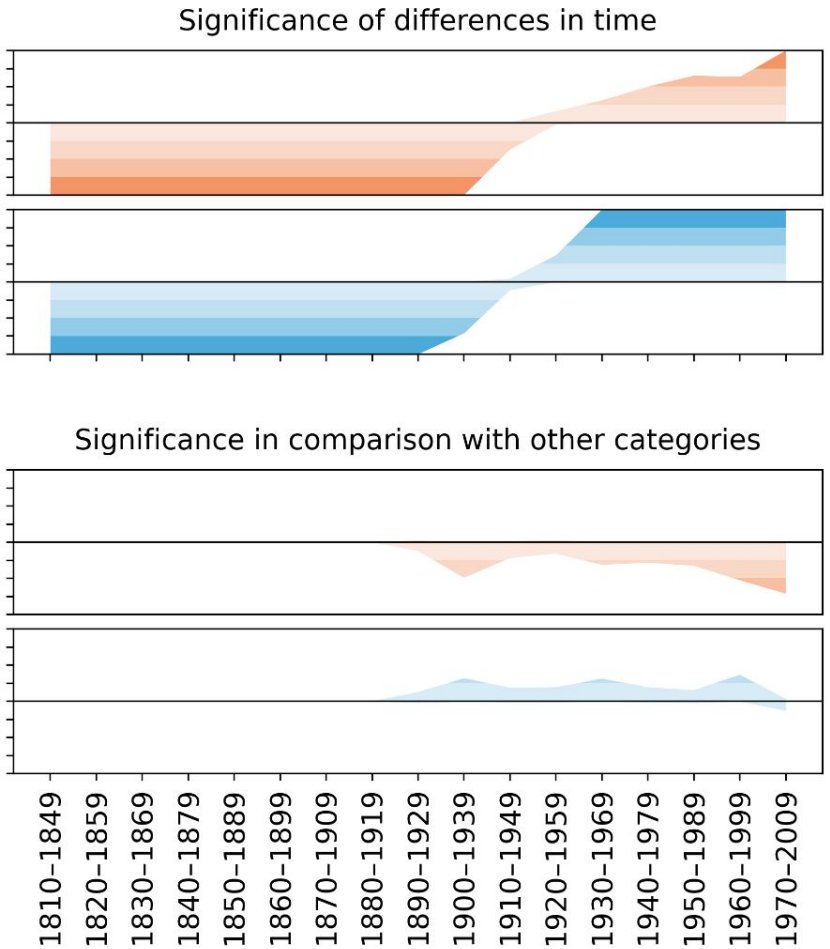
Wu, Junhui, Qingshun He & Guangwu Feng. 2016. Rethinking the grammaticalization of future *be going to*: A corpus-based approach. *Journal of Quantitative Linguistics* 23(4): 317–341.

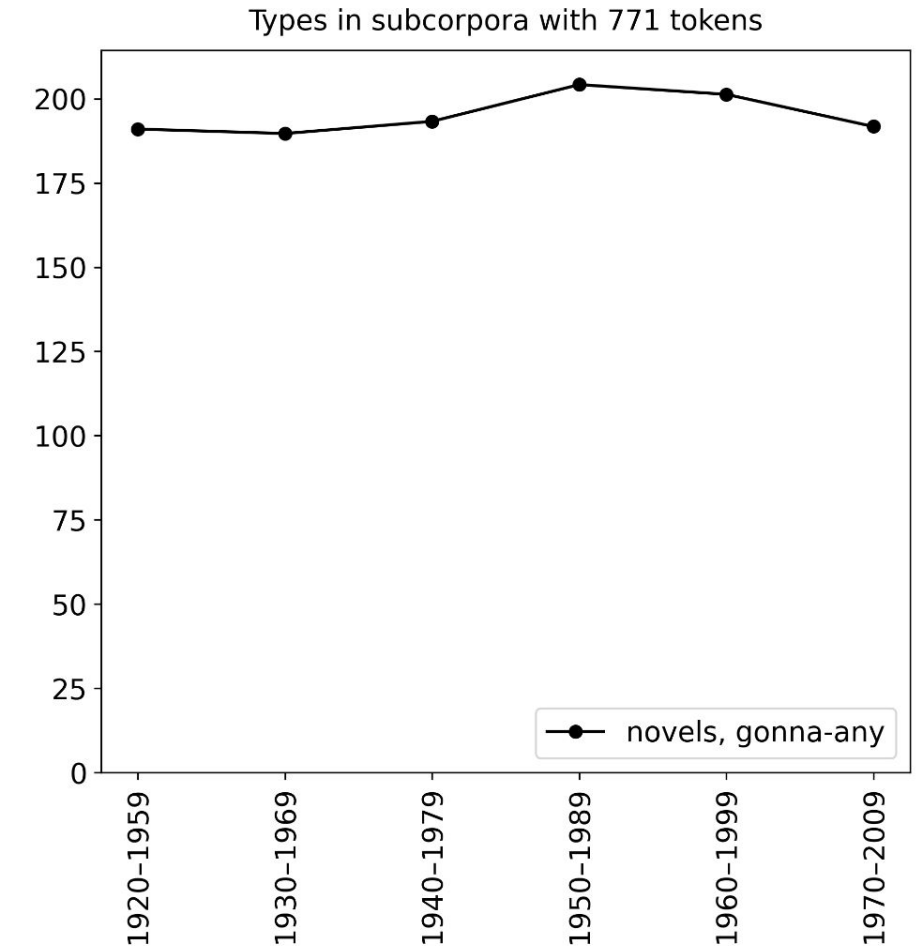
What about *gonna*?

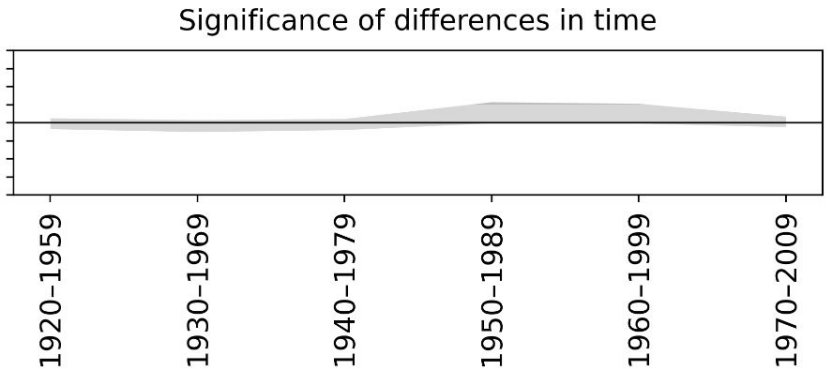

Types in subcorpora with 18811353 words
novels, gonna-any
0
50
100
150
200
250
1810–1849
1820–1859
1830–1869
1840–1879
1850–1889
1860–1899
1870–1909
1880–1919
1890–1929
1900–1939
1910–1949
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009

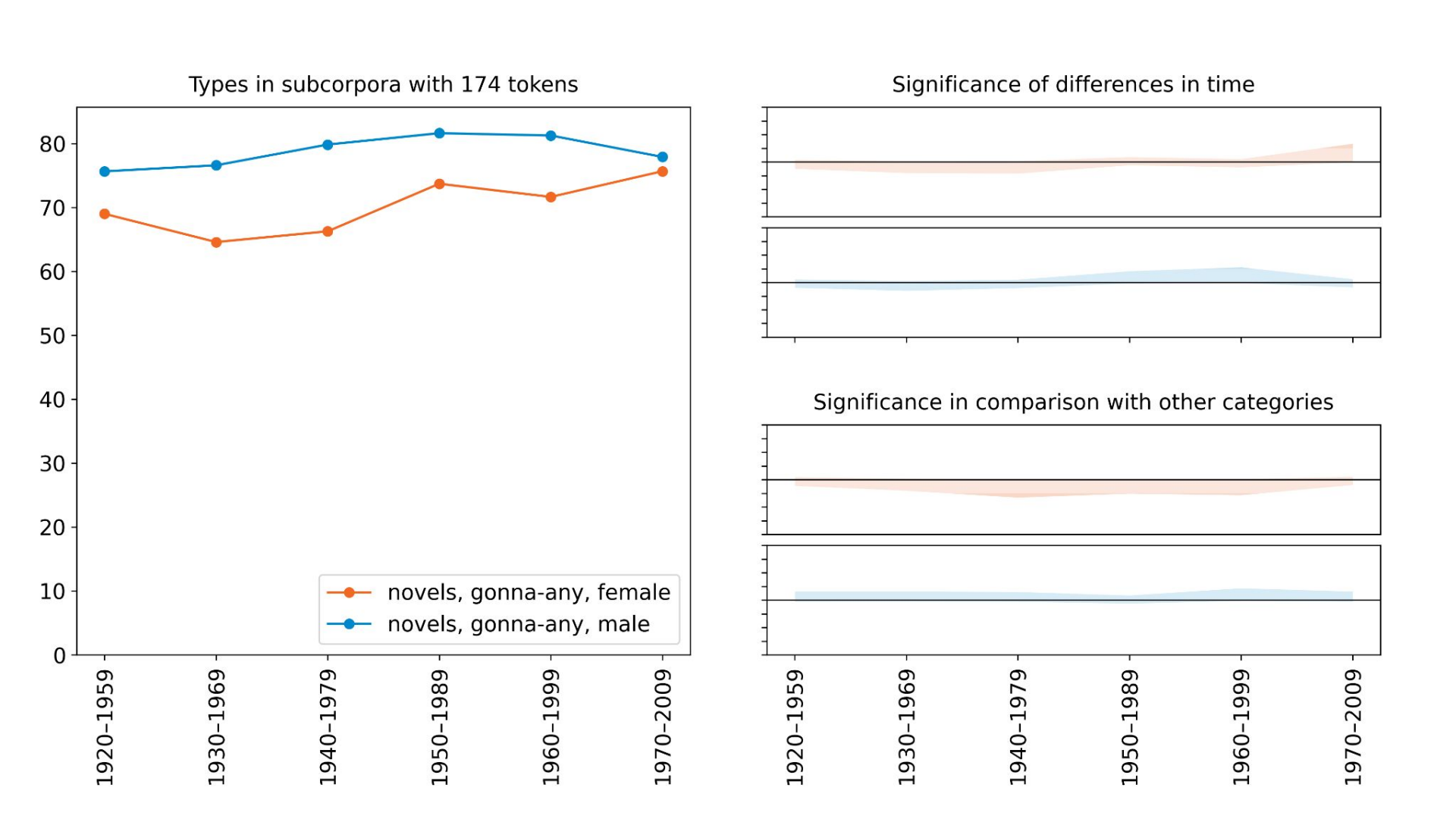

Significance of differences in time
1810–1849
1820–1859
1830–1869
1840–1879
1850–1889
1860–1899
1870–1909
1880–1919
1890–1929
1900–1939
1910–1949
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009

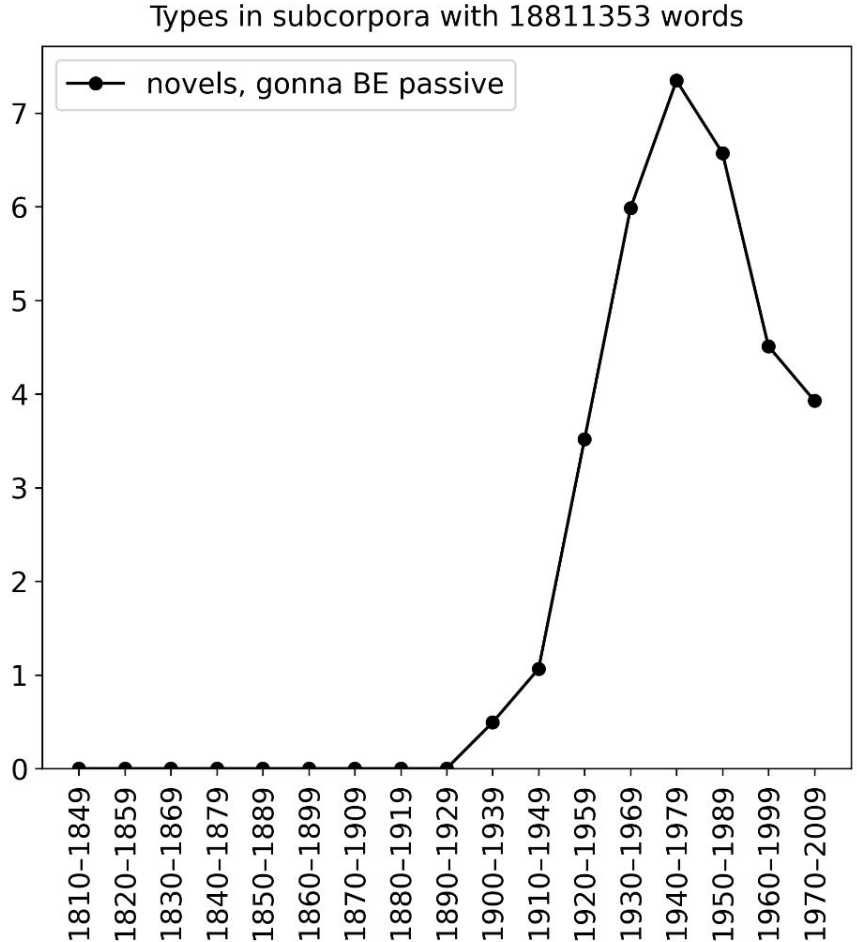

Types in subcorpora with 2863385 words
novels, gonna-any, female
novels, gonna-any, male
0
10
20
30
40
50
60
70
80
1810–1849
1820–1859
1830–1869
1840–1879
1850–1889
1860–1899
1870–1909
1880–1919
1890–1929
1900–1939
1910–1949
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009

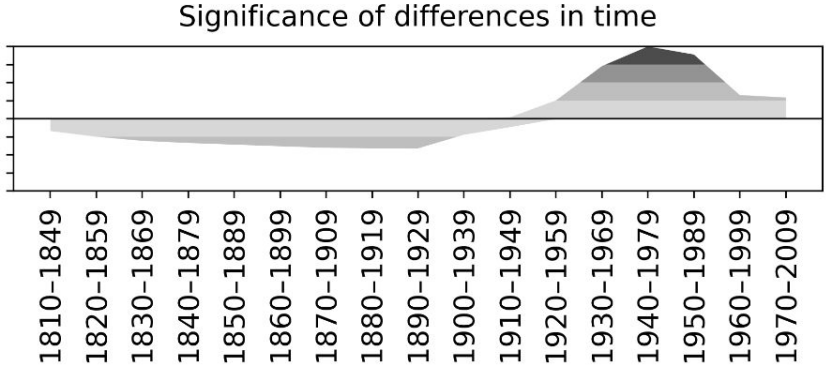

Significance of differences in time
Significance in comparison with other categories
1810–1849
1820–1859
1830–1869
1840–1879
1850–1889
1860–1899
1870–1909
1880–1919
1890–1929
1900–1939
1910–1949
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009

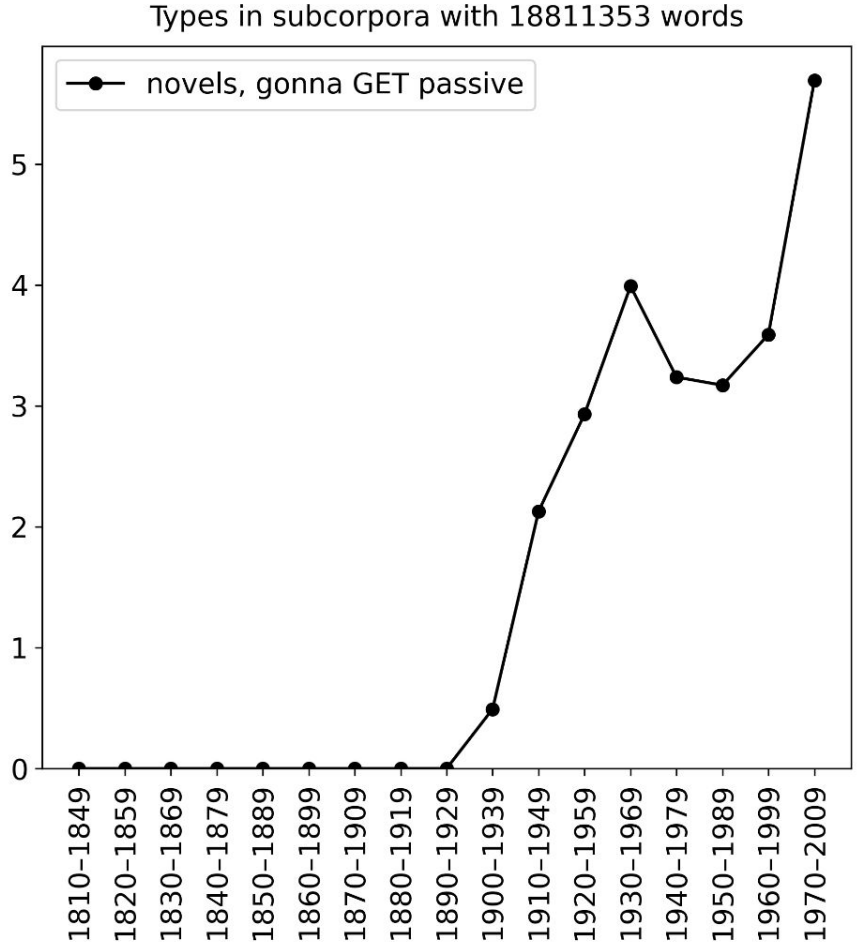

Types in subcorpora with 771 tokens
200
175
150
125
100
75
50
25
0
novels, gonna-any
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009

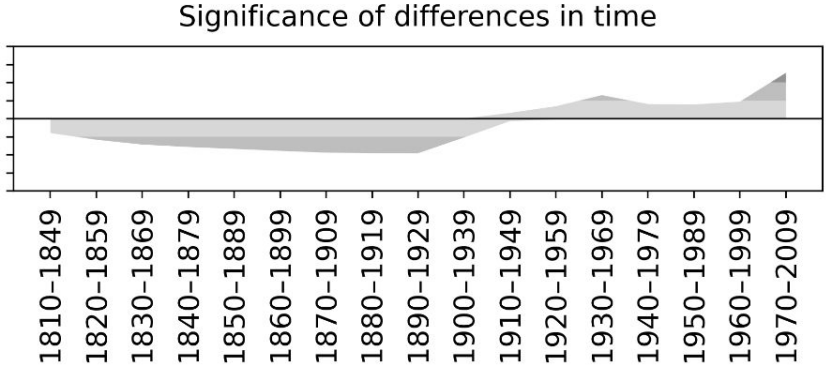

Significance of differences in time
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009


Types in subcorpora with 174 tokens
80
70
60
50
40
30
20
10
0
novels, gonna-any, female
novels, gonna-any, male
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009
Significance of differences in time
Significance in comparison with other categories
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009


Types in subcorpora with 18811353 words
novels, gonna BE passive
0
1
2
3
4
5
6
7
1810–1849
1820–1859
1830–1869
1840–1879
1850–1889
1860–1899
1870–1909
1880–1919
1890–1929
1900–1939
1910–1949
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009


Significance of differences in time
1810–1849
1820–1859
1830–1869
1840–1879
1850–1889
1860–1899
1870–1909
1880–1919
1890–1929
1900–1939
1910–1949
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009


Types in subcorpora with 18811353 words
novels, gonna GET passive
0
1
2
3
4
5
1810–1849
1820–1859
1830–1869
1840–1879
1850–1889
1860–1899
1870–1909
1880–1919
1890–1929
1900–1939
1910–1949
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009


Significance of differences in time
1810–1849
1820–1859
1830–1869
1840–1879
1850–1889
1860–1899
1870–1909
1880–1919
1890–1929
1900–1939
1910–1949
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009