English Lexical Semantic Evolution across Time

Sangpil Youm, He Zhou

Introduction

Human language is undergoing a constant evolution driven by the ongoing change in the real world: new concept, language contact, etc.

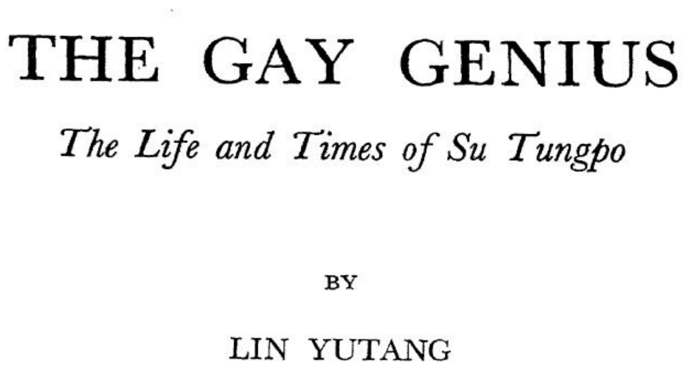
Language change: phonetic&phonological, morphological, syntactic, semantic

Lexical semantic change: diachronic evolution of lexicon usage, which is easier to change compared with other components of language.

“Every word has its own history.”

Theoretically, lexical semantic change mainly includes 3 ways:

1. Broadening 2. Narrowing 3. Shift

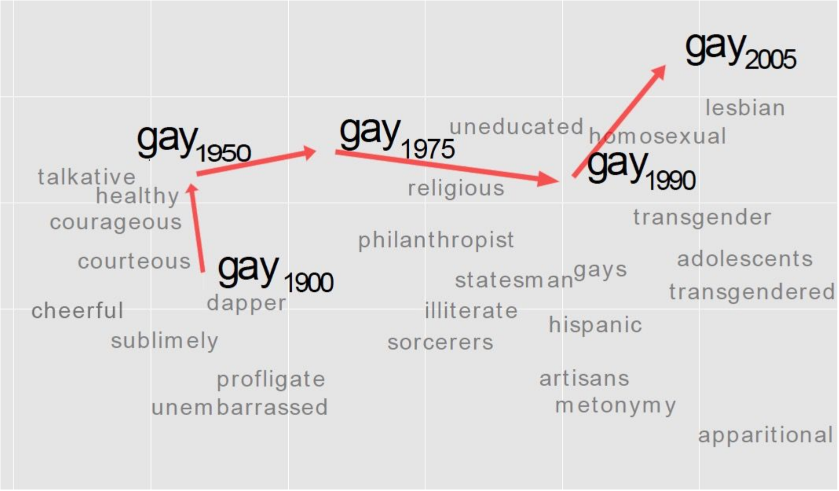


THE GAY GENIUS

The Life and Times of Su Tungpo

BY

LIN YUTANG



Questions:

Can we figure out what words have changed in meaning through time?

If a word has changed its meaning, when and how?

Can we analyze the diachronic change from network science perspectives?

Related Work

Adam Jatwot et al. 2014.

A Framework for Analyzing Semantic Change of Words across Time

A visual analytics framework for discovering and visualizing lexical change at 3 different levels -- individual words, word pairs, sentiment orientation

Data: Google Book 5-gram, COHA

Word Representations:

1) Normal Word Representation
2) Positional Word Representation
3) Latent Semantic Analysis based Representation

Vivek Kulkarni et al, 2015

Statistically Significant Detection of Linguistic Change

Word Evolution Modeling: frequency, syntactic, distributional

Statistical Soundness: use change point detection in time series to assign significance of change scores to each word

Cross-Domain Analysis: books, tweets and online reviews

Data: Google Book N-gram, Tweets, Amazon reviews

William Hamilton et al. 2016

Diachronic Word Embeddings Reveal Statistical Law of Semantic Change

Developed a robust methodology for quantifying semantic change by evaluating word embeddings (PPMI, SVD, word2vec) against known historical changes.

Data: 6 historical corpora spanning 4 languages and 2 centuries.

Two quantitative laws of semantic change:

1) the law of conformity- frequent words change more slowly

2) the law of innovation - polysemous words change more quickly

Data

- English Literature from Gutenberg Project
  - 17th Century 200 books (5,741,155)
  - 19th Century 200 books (141,422,786)
  - 20th Century 200 books (till around 1970s) (86,618,416)
- 21st: Google News Embeddings

Preprocessing:

- Convert all capital letters to lower-case
- Remove all punctuations
- Remove stopwords with NLTK

Method

❖ Word-embeddings (word2vec) for words in each time period

- Default setting, bigram detection
- Analogy test
- ex) “King - Man + Woman = Queen”

❖ Measure distance between words with cosine similarity

- Find out similar meaning words
- Figure out semantic change among different time periods

Method

- Build a word list in which all words show up in both periods
  - 17th vs 20th, 19th vs 20th
  - Make vocab containing words appeared in both time perionds
- Randomly sample 1,000 words from the word list

Method

Network Analysis

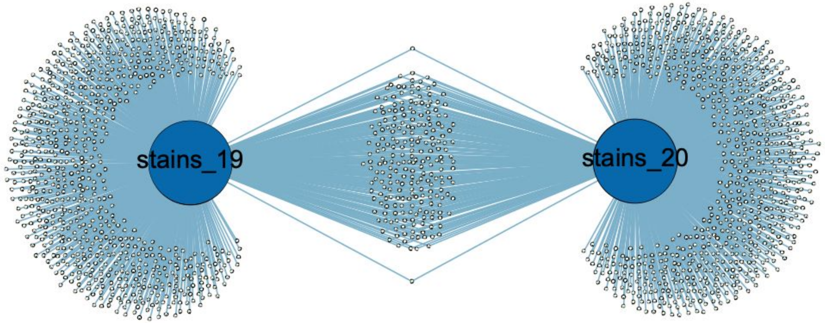
- ❖ Generate graph with selected words and similar words based on cosine similarity
- ❖ Calculate Jaccard Coefficient for each word in different time periods
  - 17th vs 20th, 19th vs 20th
- ❖ Plot the correlation between Jaccard coefficient and frequency

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

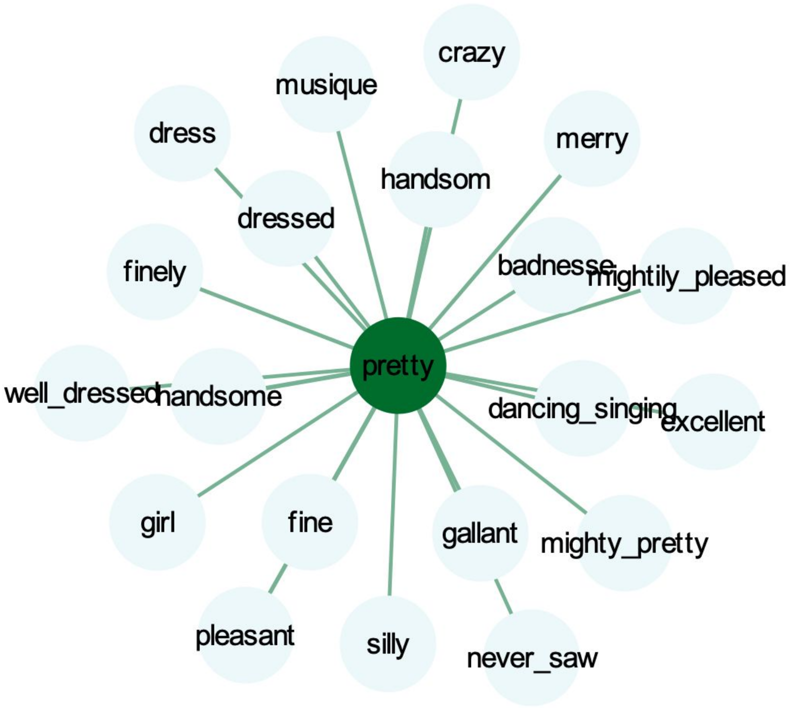
Method

Network Analysis

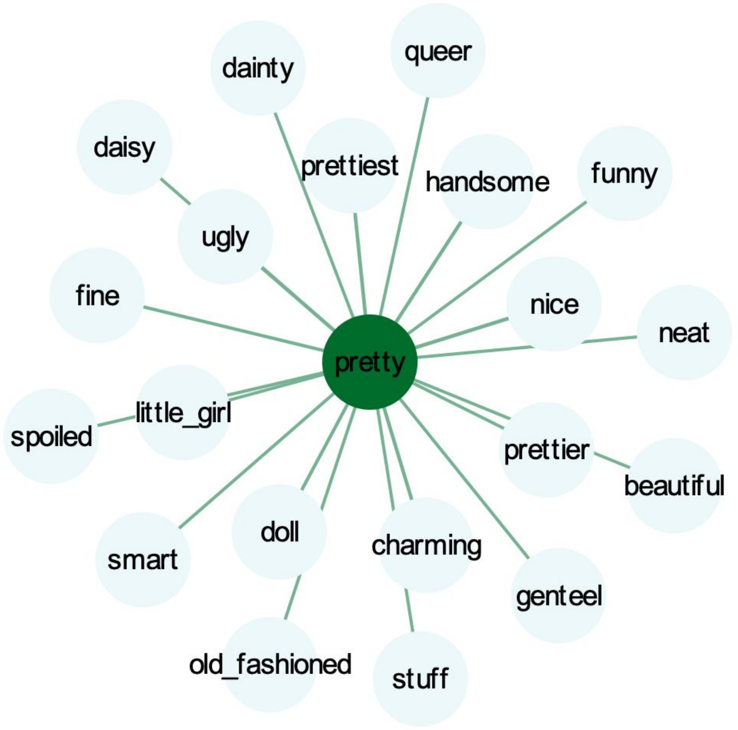
- Cosine similarity threshold 0.8



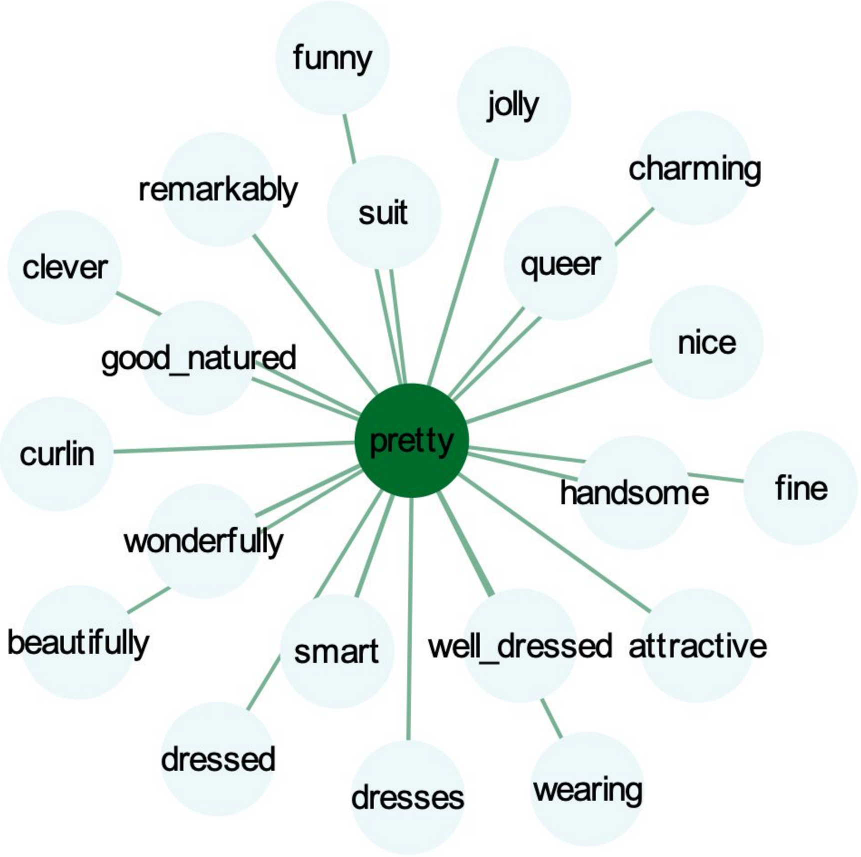
Result pretty (ADJ → ADV)



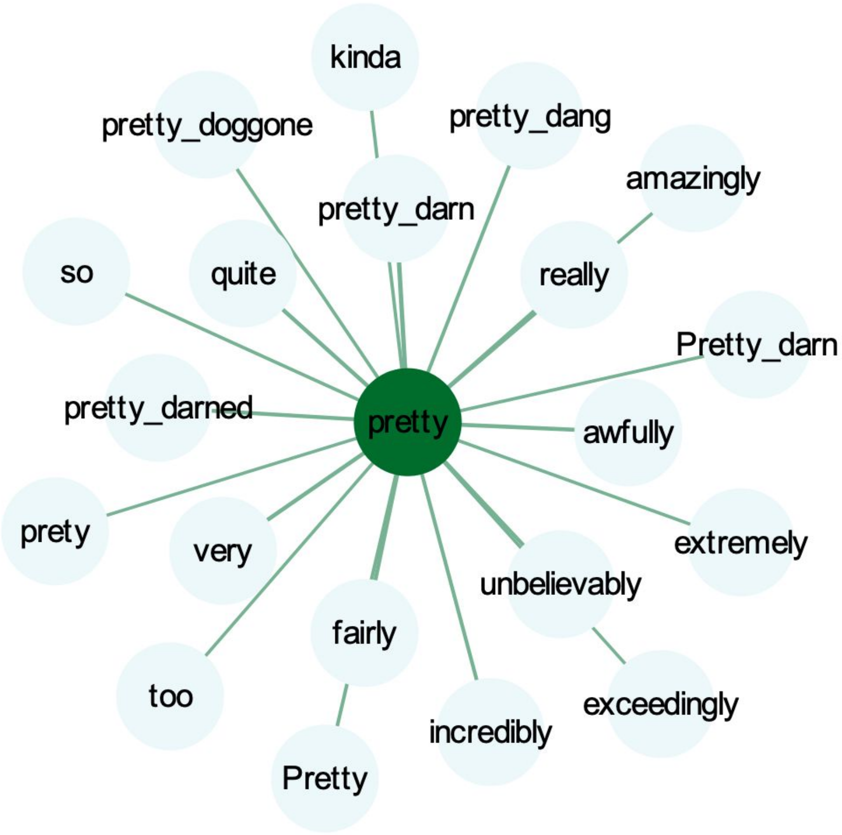
‘pretty’ - 17th century



‘pretty’ - 19th century

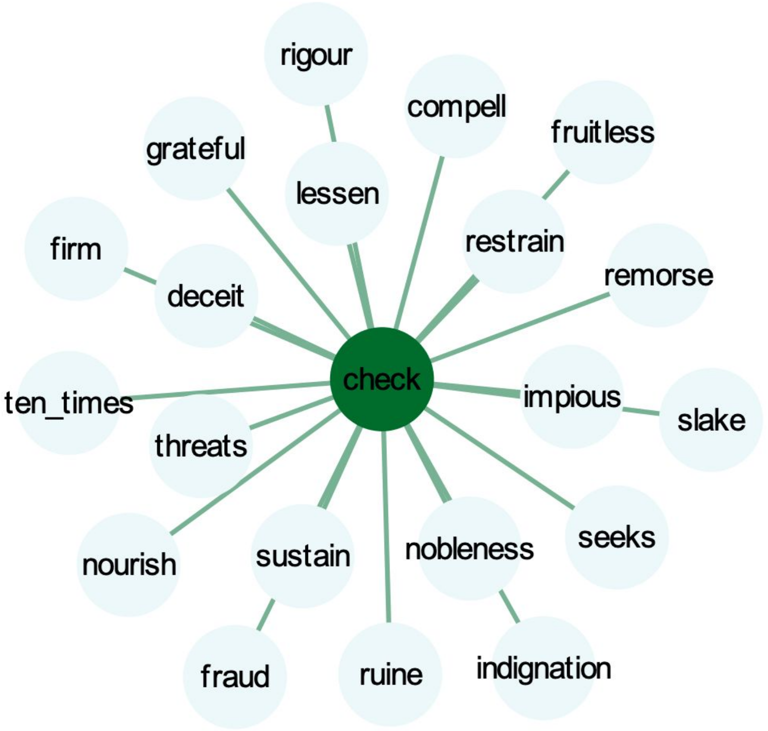


‘pretty’ - 20th century

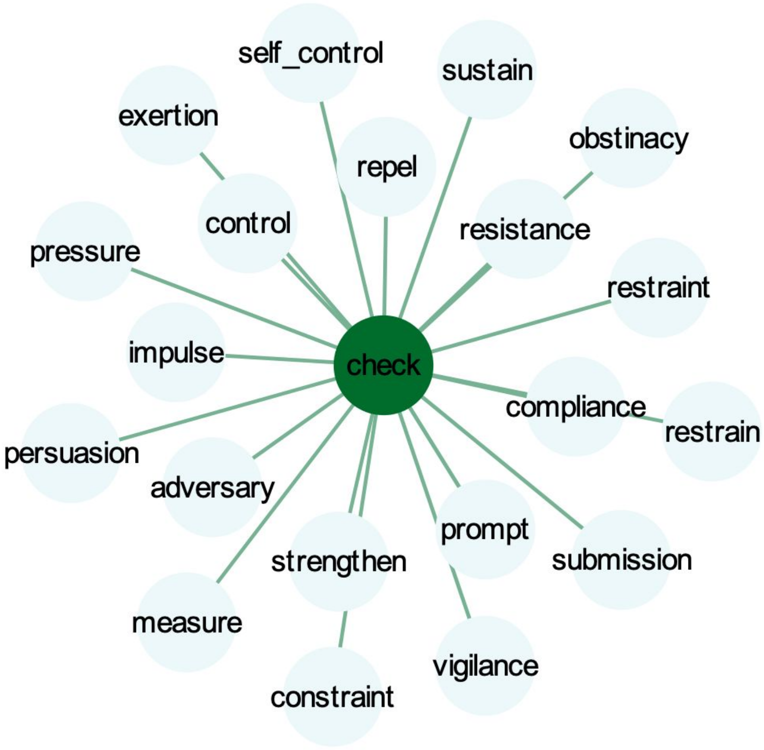


‘pretty’ - 21st century

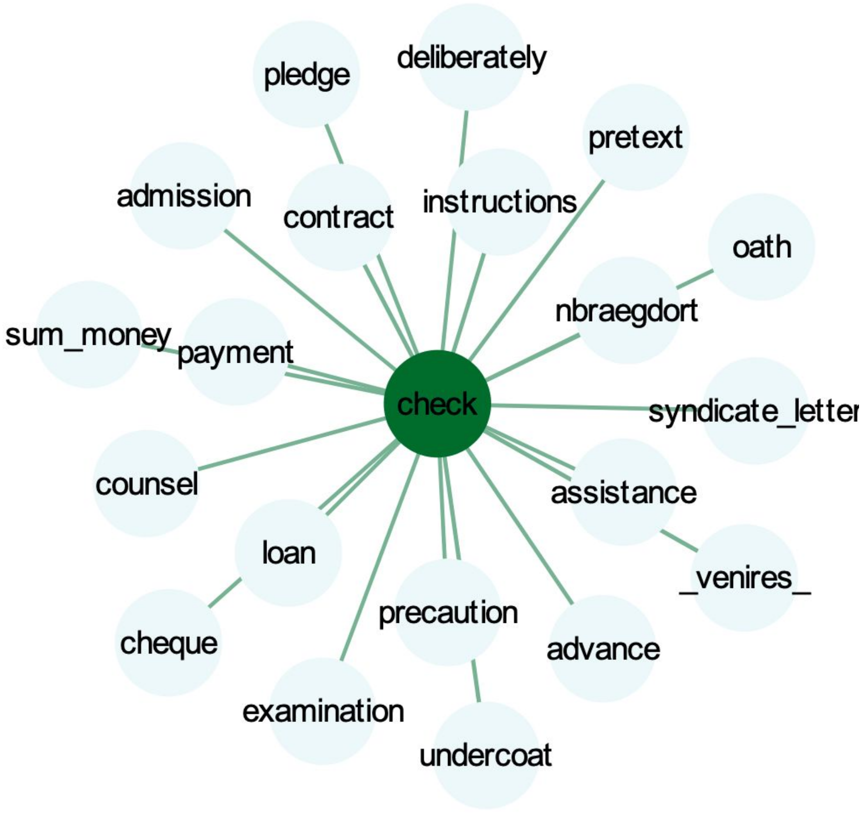
check (VERB → NOUN)



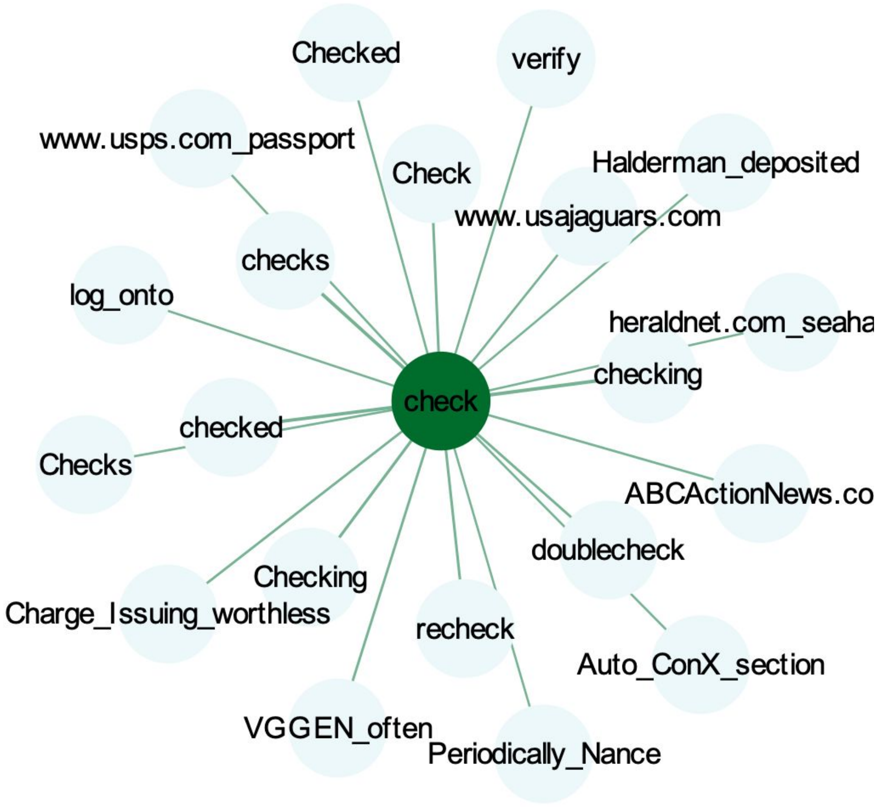
‘check’ - 17th century



‘check’ - 19th century

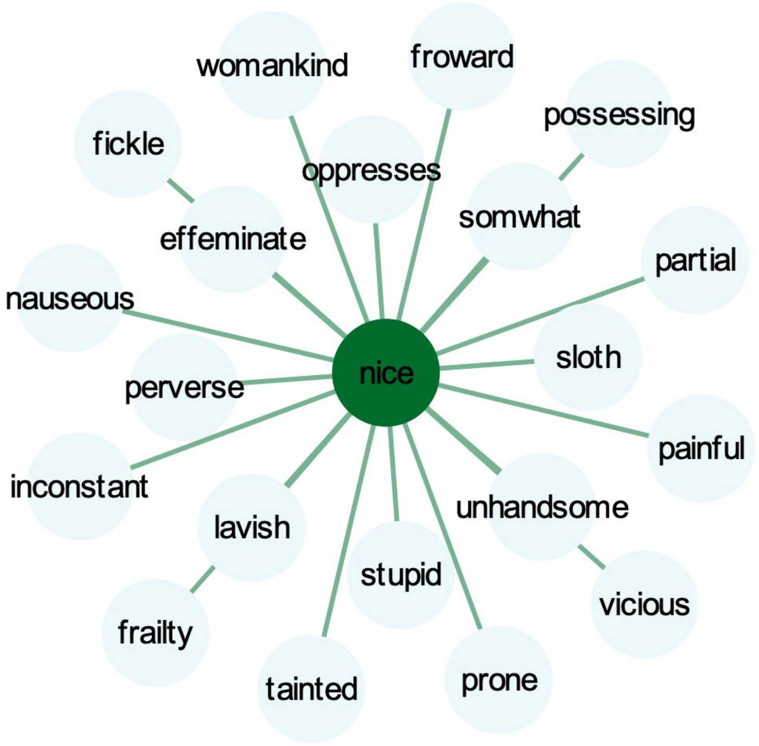


‘check’ - 20th century

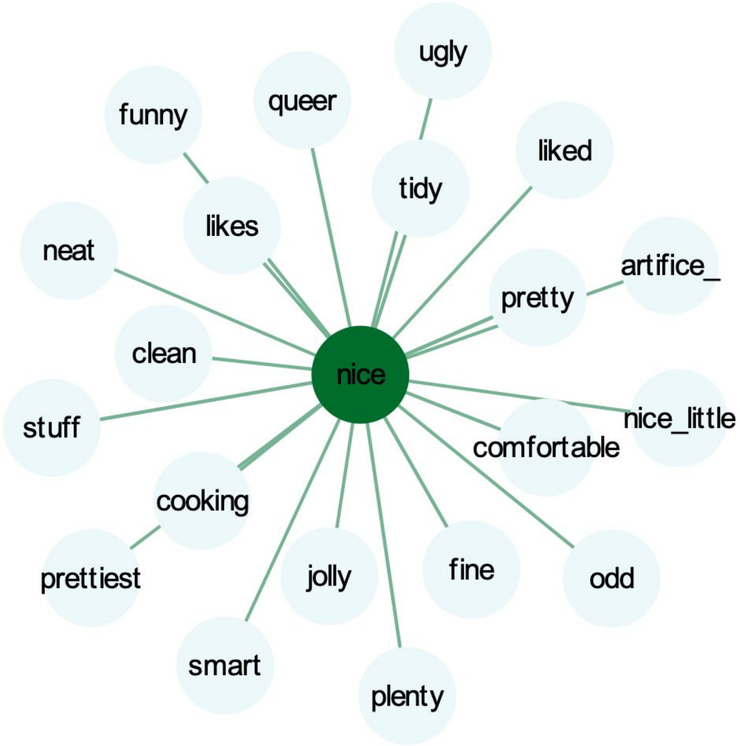


‘check’ - 21st century

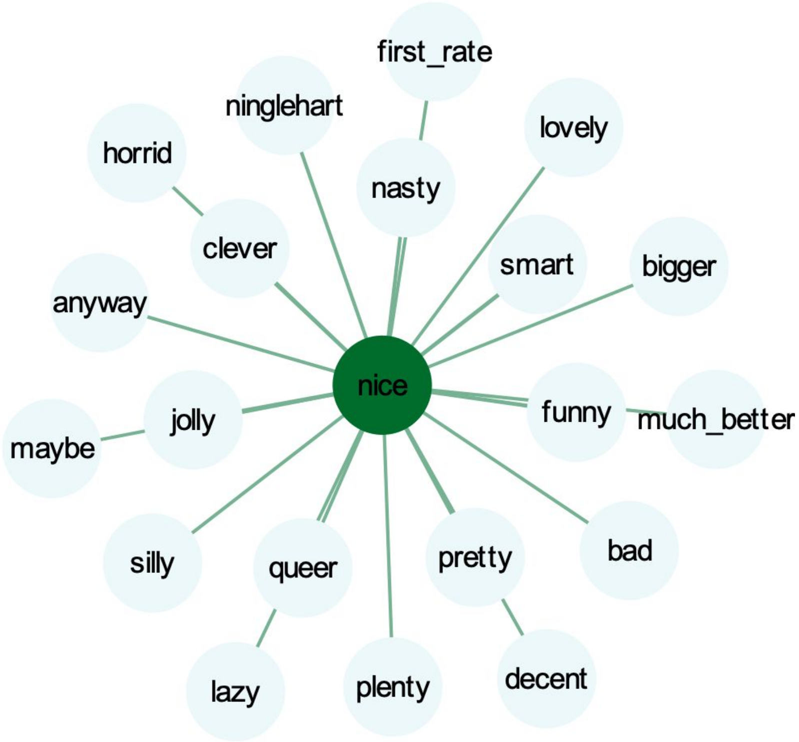
nice (negative → positive)



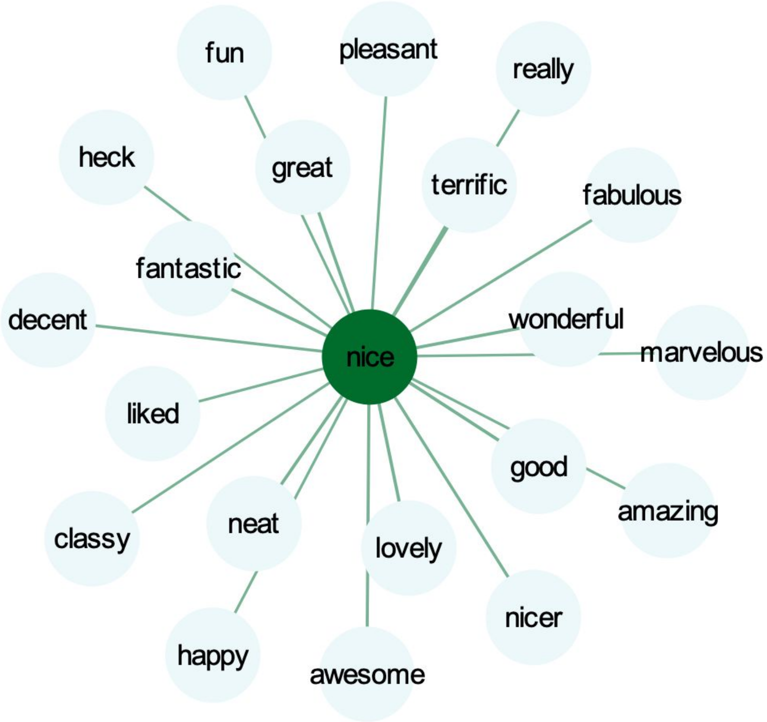
‘nice’ - 17th century



‘nice’ - 19th century

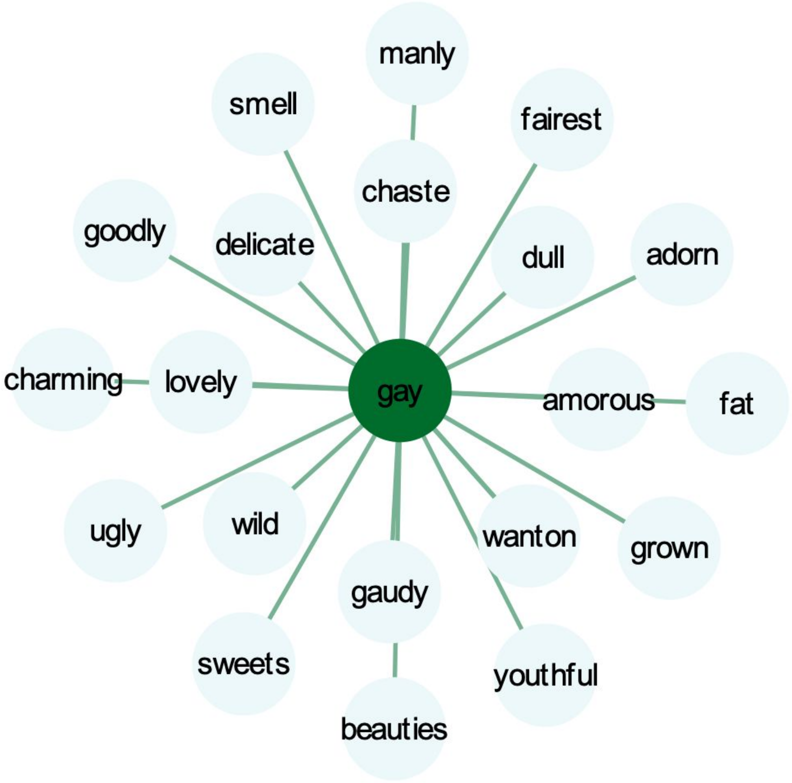


‘nice’ - 20th century

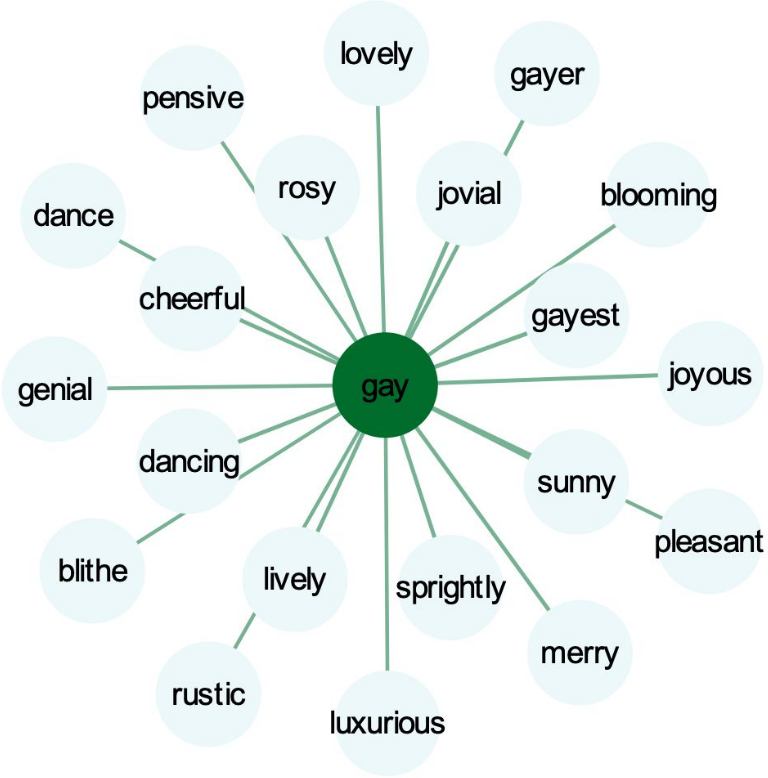


‘nice’ - 21st century

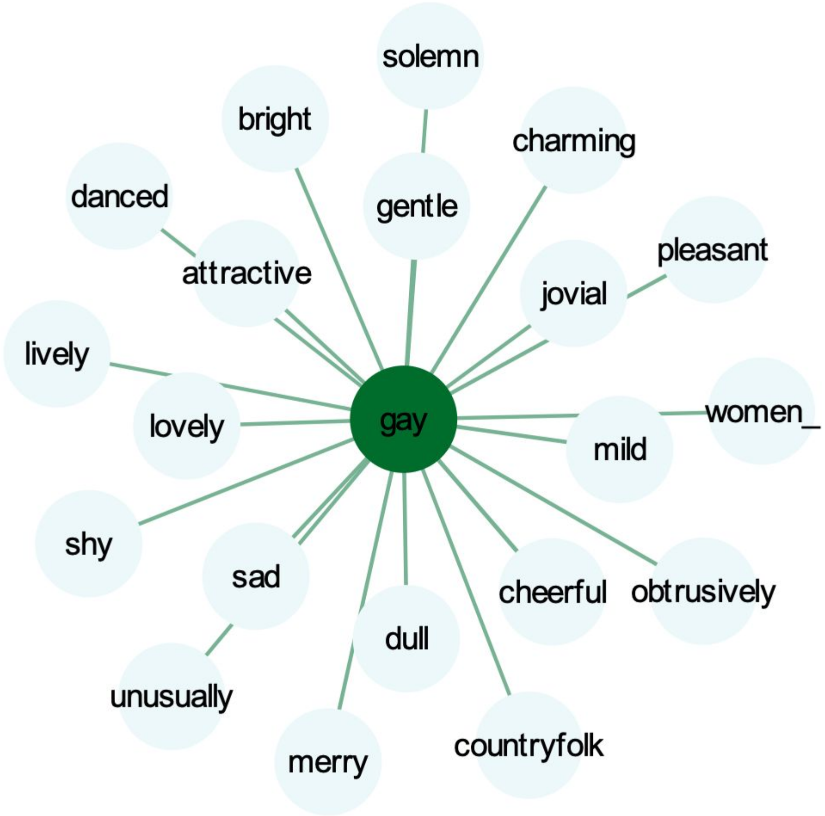
gay



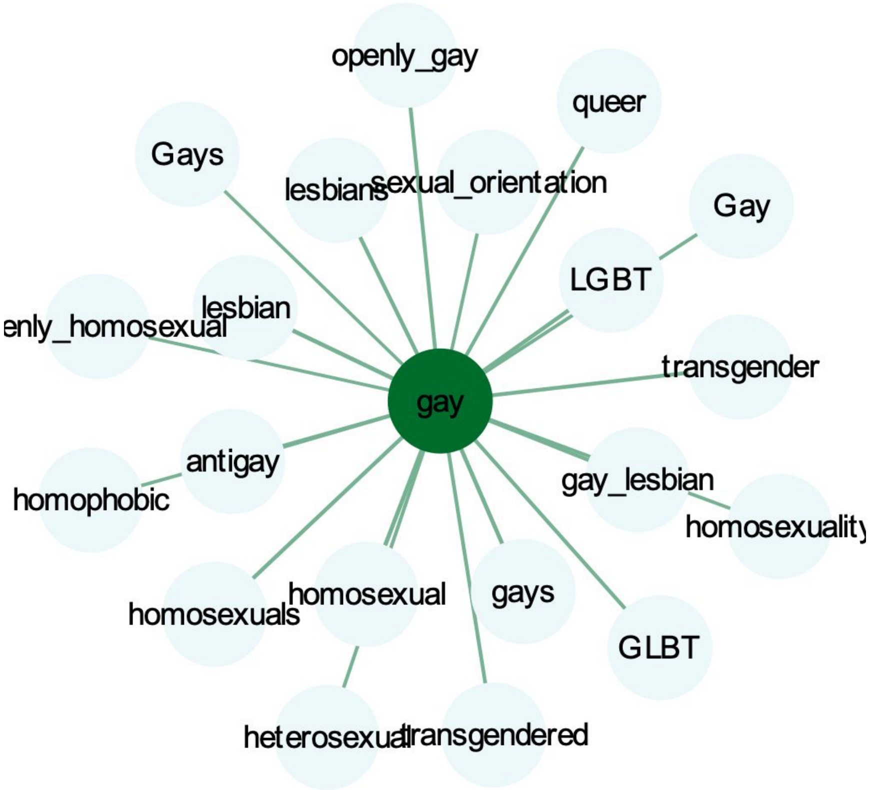
‘gay’ - 17th century



‘gay’ - 19th century

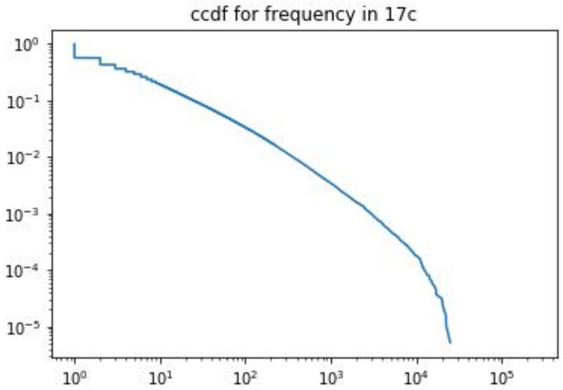


‘gay’ - 20th century

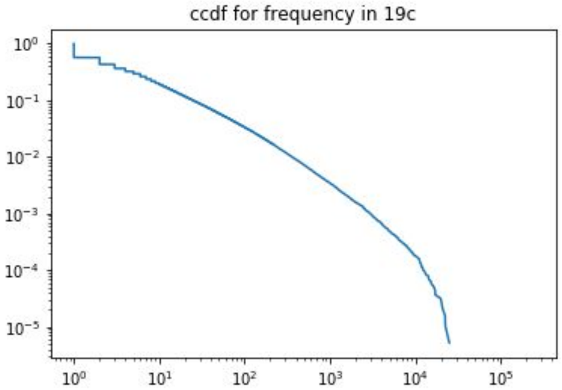


‘gay’ - 21st century

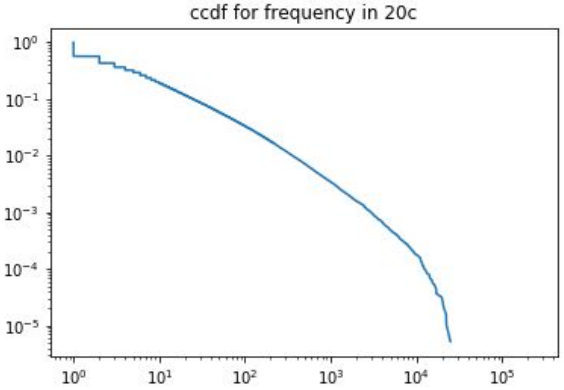
Frequency



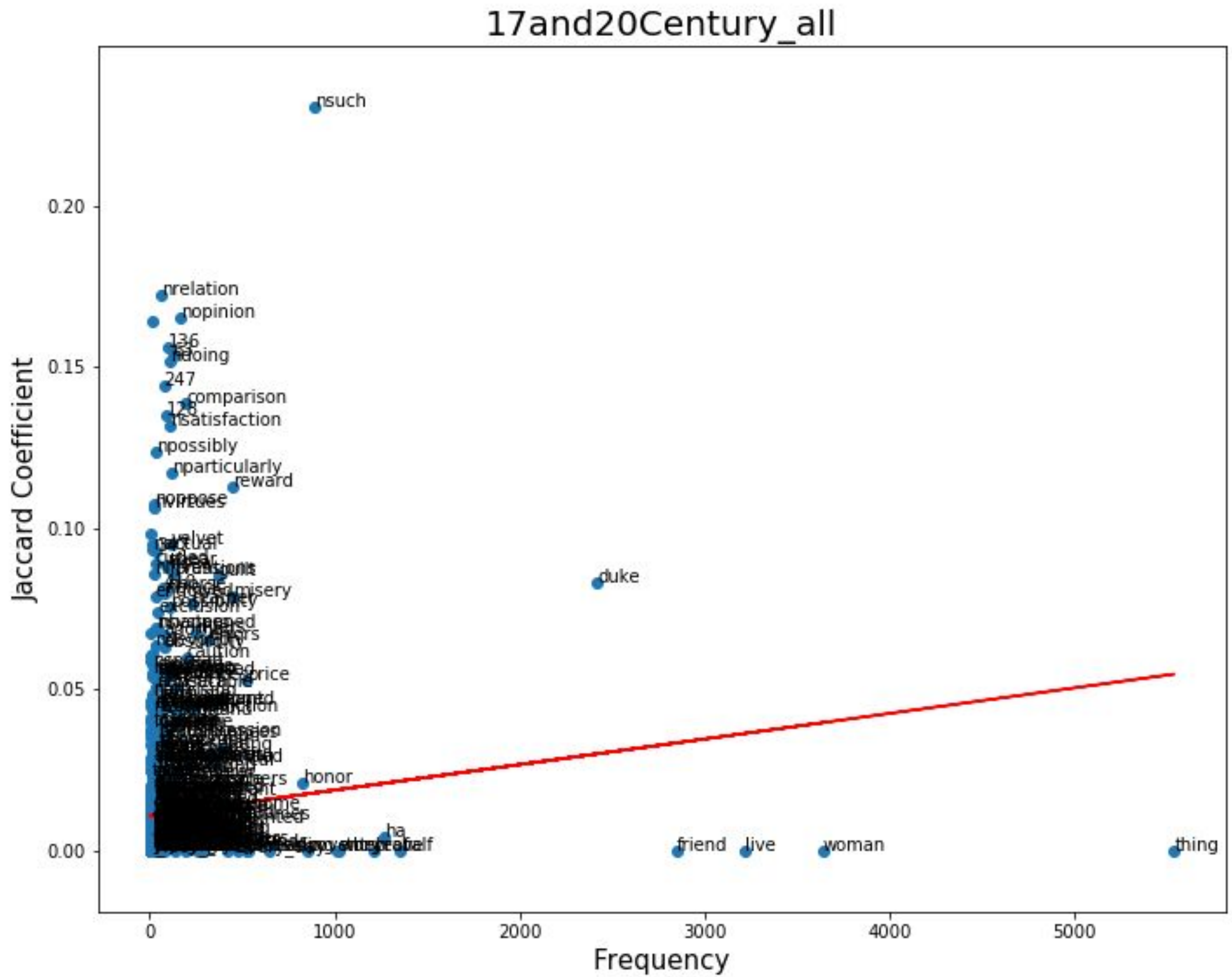
ccdf for frequency in 17c

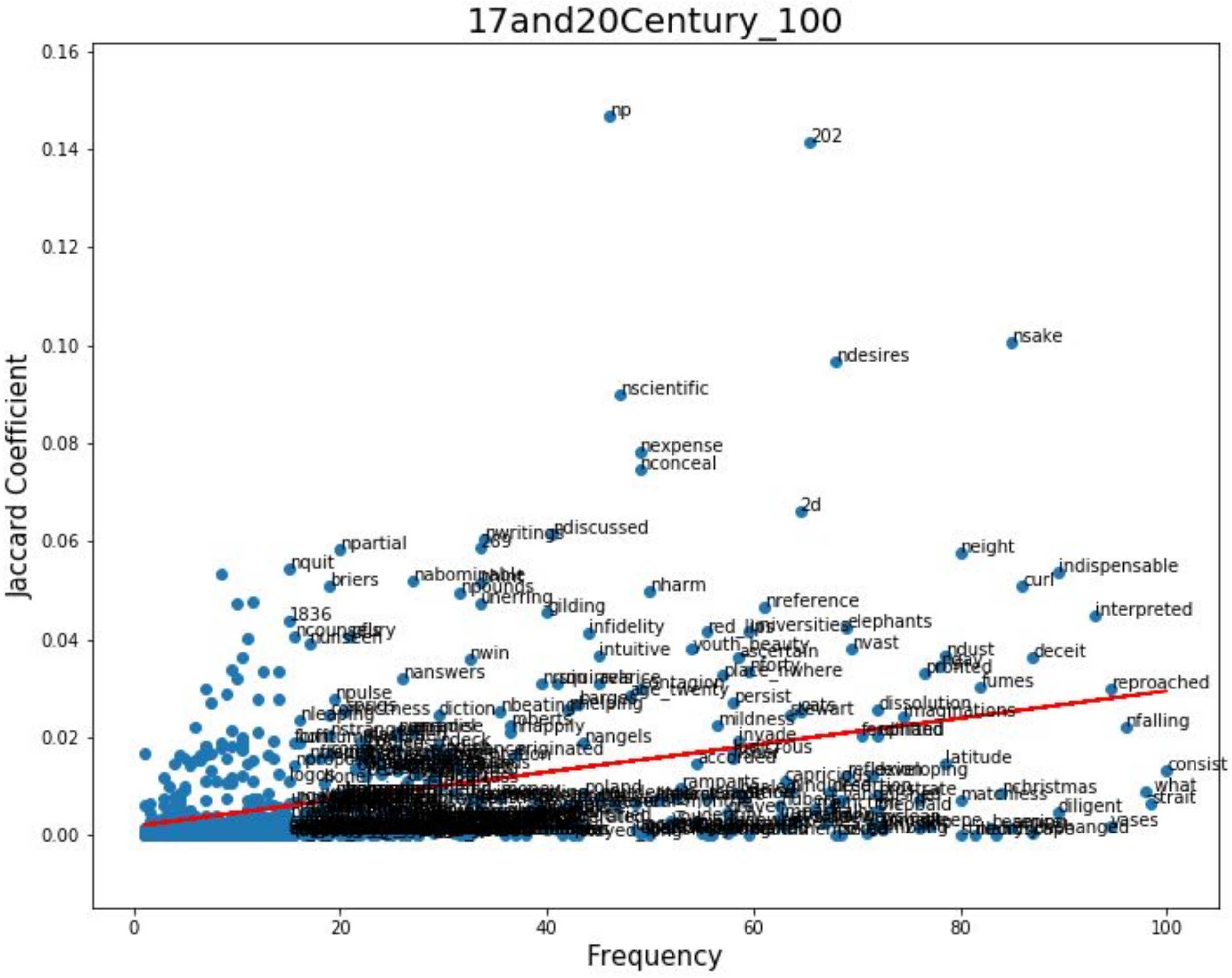


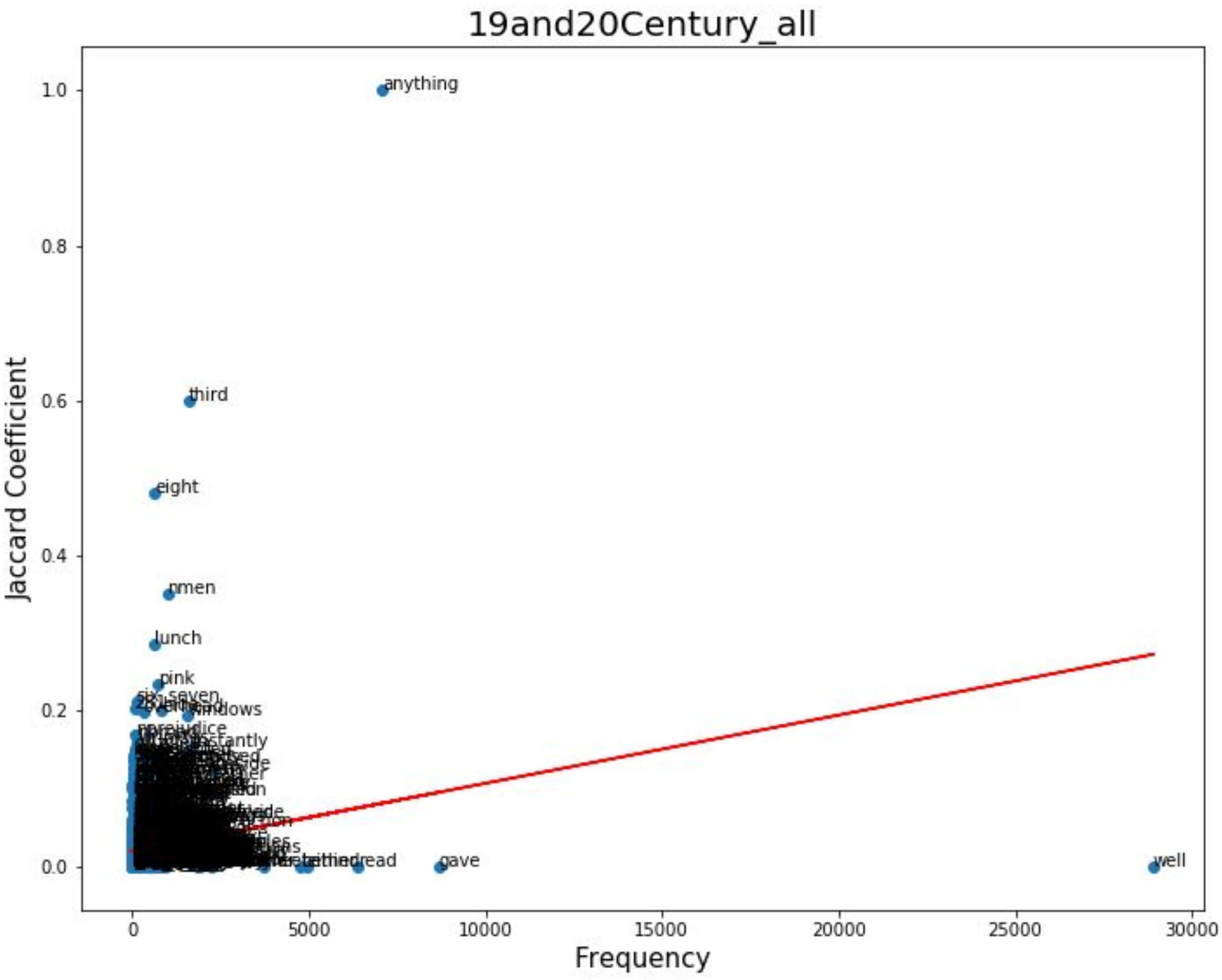
ccdf for frequency in 19c

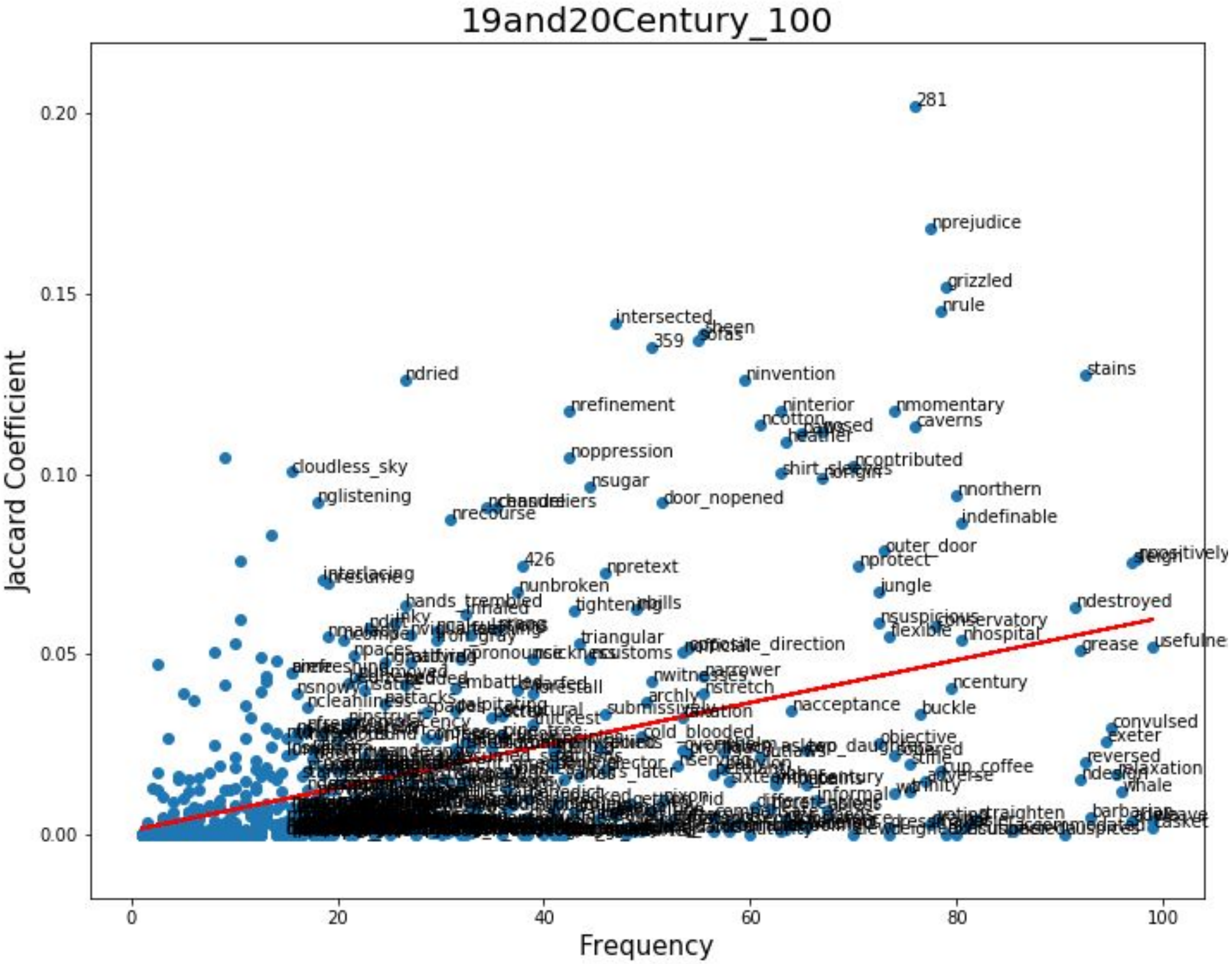


ccdf for frequency in 20c

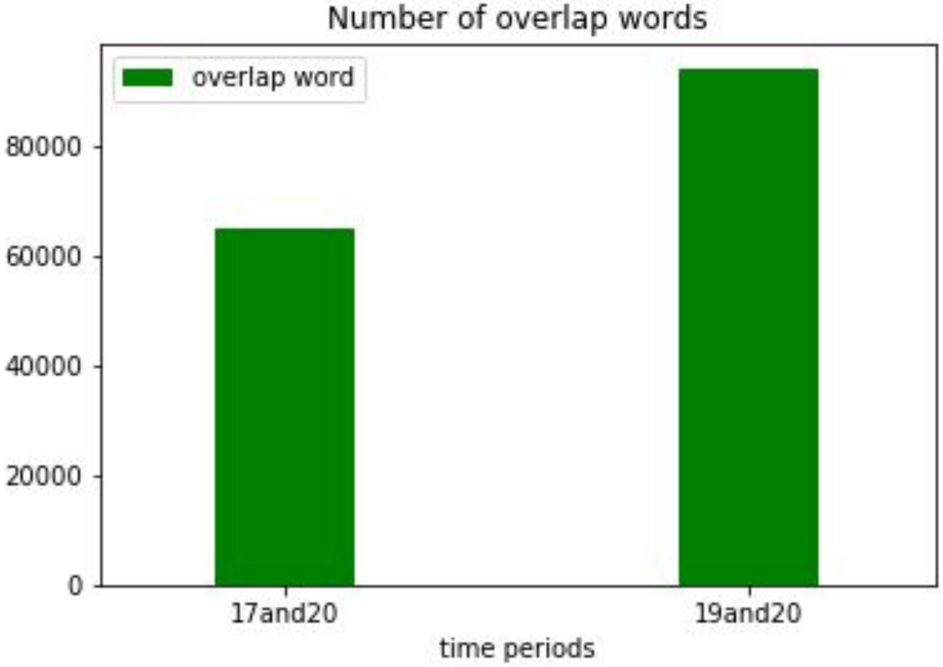

17and20Century_all
Jaccard Coefficient
Frequency
0.00
0.05
0.10
0.15
0.20
0
1000
2000
3000
4000
5000
nsuch
nrelation
nopinion
136
ndoing
247
comparison
128
nsatisfaction
npossibly
nparticularly
reward
velvet
misery
caution
price
duke
honor
ha
friend
live
woman
thing


17and20Century_100
Jaccard Coefficient
Frequency
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
100
np
202
nsake
ndesires
nscientific
nexpense
nconceal
2d
ndiscussed
nwritings
269
npartial
nquit
briers
nabominable
npounds
unerring
nharm
gilding
neight
indispensable
curl
nreference
interpreted
1836
ncounsels
infidelity
red_lips
universities
elephants
intuitive
youth_beauty
ascertain
nvast
ndust
deceit
nwin
nanswers
place_nowhere
nforty
pointed
fumes
contagion
age_twenty
reproached
npulse
persist
diction
nbeating
helping
dissolution
imaginations
nfalling
mildness
invade
nangels
unhappily
originated
accorded
latitude
consist
ramparts
capricious
developing
nchristmas
matchless
what
strait
diligent
vases
poland
logos

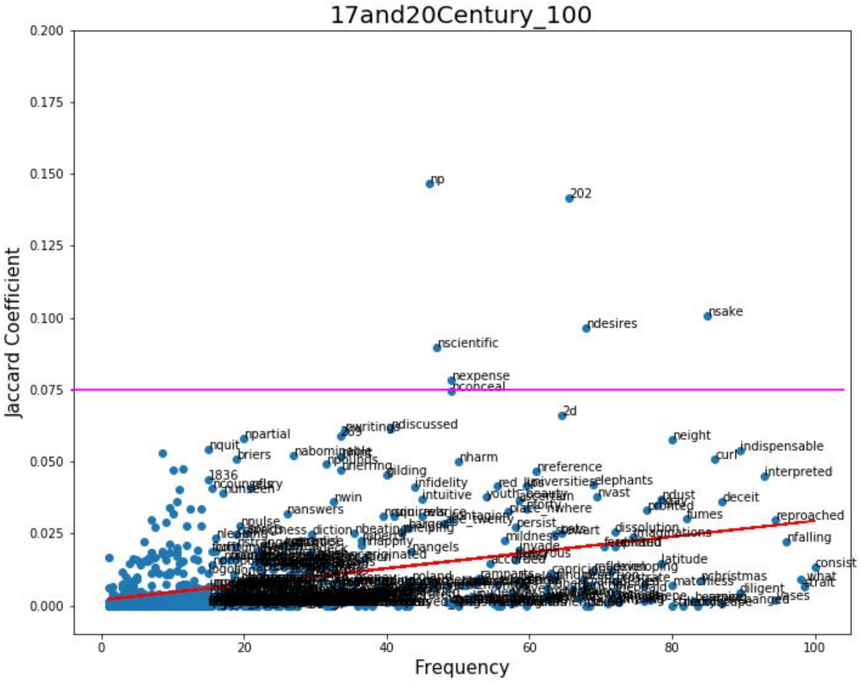

19and20Century_all
Jaccard Coefficient
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
0
5000
10000
15000
20000
25000
30000
anything
third
eight
nmen
lunch
pink
six
seven
windows
prejudice
constantly
read
gave
well


19and20Century_100
Jaccard Coefficient
Frequency
0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
100
281
nprejudice
grizzled
nrule
intersected
sheen
sofas
359
ndried
ninvention
stains
nrefinement
ninterior
nmomentary
ncotton
caverns
heather
noppression
ncontributed
cloudless_sky
nsugar
shirt_sleeves
nglistening
nchandeliers
door_nopened
nnorthern
nrecourse
indefinable
outer_door
positively
426
npretext
nprotect
interlacing
presumes
nunbroken
jungle
hands_trembled
tightening
hills
ndestroyed
inky
inhaled
nsuspicious
conservatory
flexible
nhospital
triangular
opposite_direction
grease
usefulne
npaces
pronounce
customs
narrower
nwitnesses
ncentury
embattled
foresall
nstretch
nsnow
archly
attacks
ncleanliness
palpitating
submissively
nacceptance
buckle
thickest
cold_blooded
convulsed
objective
exeter
reversed
cup_coffee
relaxation
ndesign
adverse
whale
informal
straighten
barbarian
casket
auspices

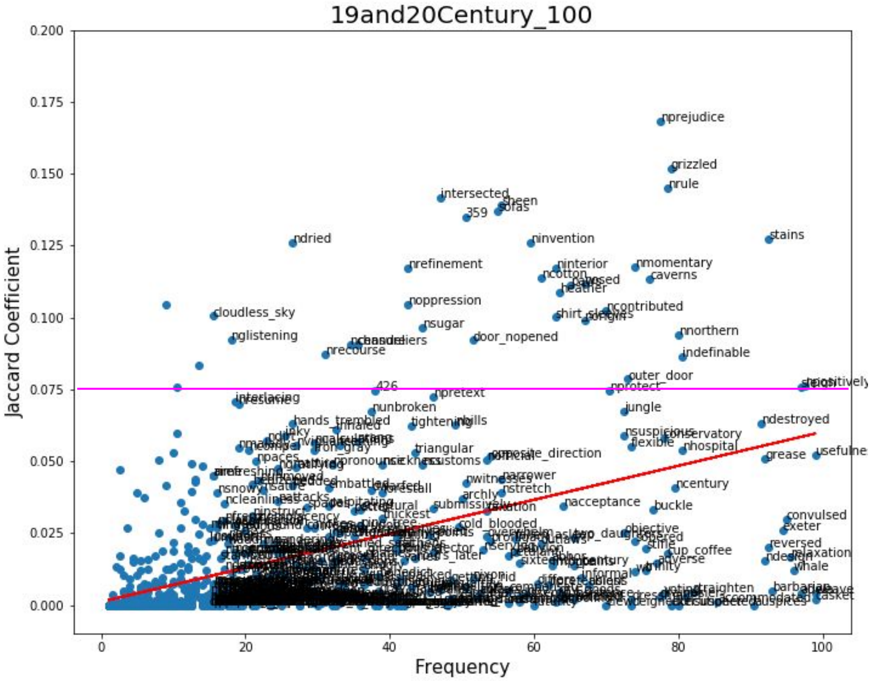
Number of overlap words



Words have changed overtime



863 words



868 words

Conclusion & Future Work

Conclusion:

1. Semantic change can be detected within the limited amount of data.

   - Semantic change over time

2. Relation between semantic change and frequency can be shown with network analysis.

   - Higher frequency words have less semantic change
   - Words have been changed over time

Conclusion & Future Work

Future Work:

1. Data is never enough

2. Better way to represent word meaning, e.g.: transfer learning on Google Book N-gram

3. More measures

Thank you

&

Any feedback is welcomed!