Finding Rising and Falling Words

Erik Tjong Kim Sang
Meertens Institute, Amsterdam
erik.tjong.kim.sang@meertens.knaw.nl

11 December 2016

Meertens Institute, Amsterdam

The Meertens Institute in Amsterdam studies various aspects of Dutch language and culture, like for example regional variances and language change over time

Nederlab

The Nederlab project aims at collecting and making available all texts relevant for the culture and language of The Netherlands

This involves texts of about the year 800 until now

These historical texts are being made available to the research community on a website, together with analysis tools

Extensive meta data are available for the texts: document types and dates, authors and even text analysis

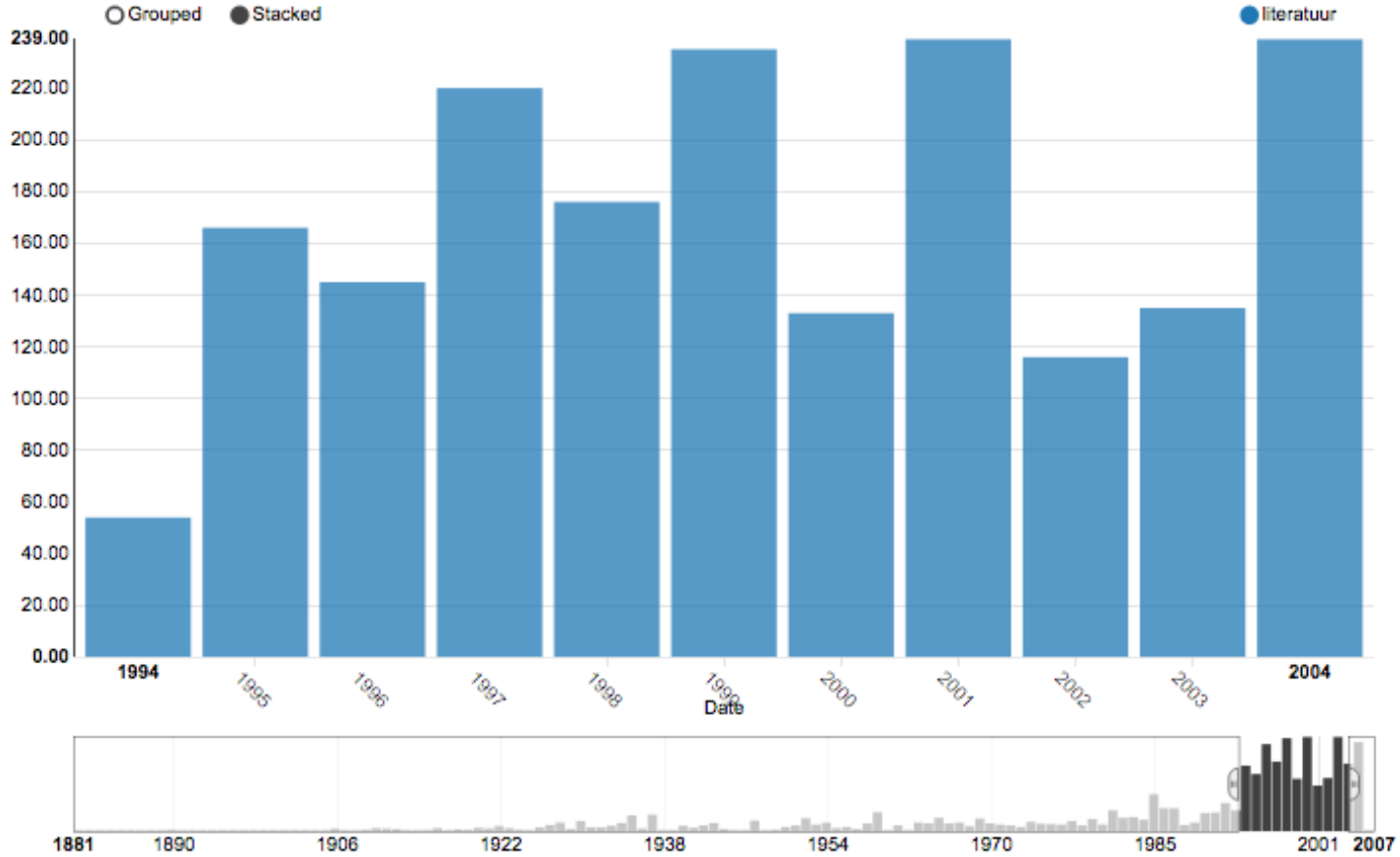
Corpus: Collectie: alle collecties, alle data, metadata van bron: tijdschrifttitel: Tijdschrift voor Nederlandse Taal en Letterkunde (1881-), samenstelling: zelfstandige titels, onzelfstandige titels, koepeltitels

Let op: deze operatie kan voor complexe zoekvragen of veel documenten enkele minuten duren.

Woordentrends over de tijd

Alle trendwoorden voor de tijdlijn (kommagescheiden): literatuur Geef weer JSON

Scoretype: Hits per jaar Gemiddeld per document Hits/woorden in gevonden documenten



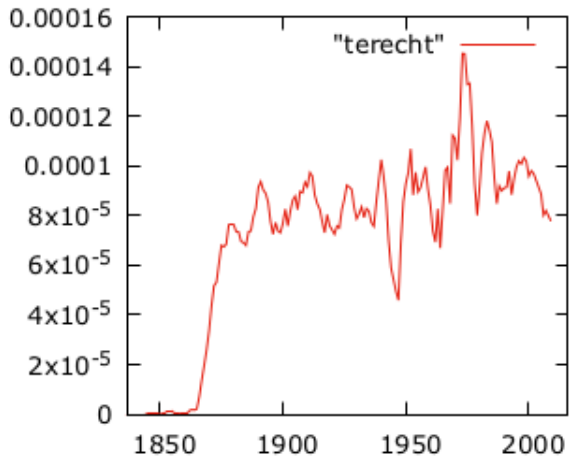
User question: By searching on your website, I found this word “.....” with an interesting behavior over time. Can you send me a list of similar words?

Rising and Falling Words

We are interested in the life cycle of words: when did certain words become more popular and when did they get out of fashion?

This is closely related to finding neologisms (new words) and archaisms (out-of-fashion words)

We will only consider relative word frequency as clue



Methods

We evaluated two methods for finding words with rising or falling frequency:

1. Delta scores: divide the last known relation frequency by the first

2. Correlation coefficients computed using all known frequencies of a word

Both methods result in a score per word.
Higher scores indicate rising frequencies.
Lower scores indicate falling frequencies.

Corpora

We use two text corpora in this study:

1. The Dutch literary magazine *De Gids* (1837-2009):
   169 editions
   88 million tokens
   32,312 unique tokens with frequency larger than 100

2. Dutch Twitter (2011-2016):
   almost six years
   27 billion tokens
   38,230 unique tokens with frequency larger than 10,000

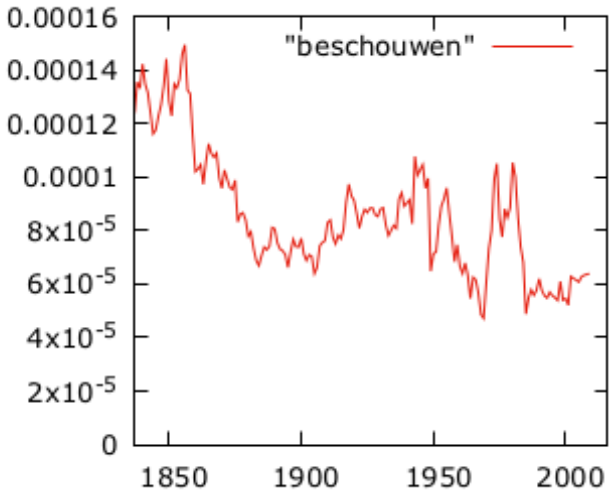
Evaluation

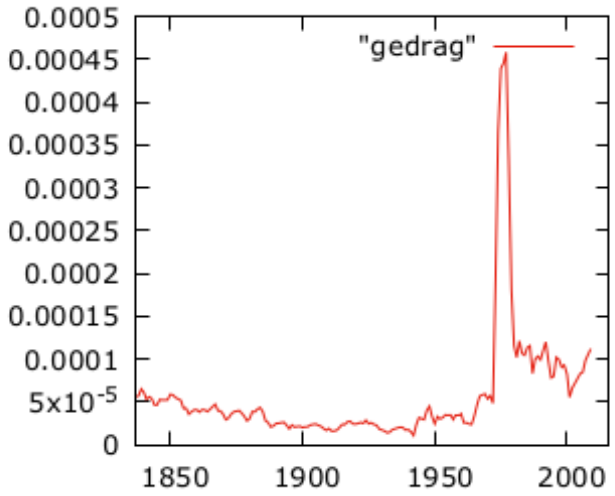
Evaluating the results was a problem: there is no good gold standard data available for most of these tasks

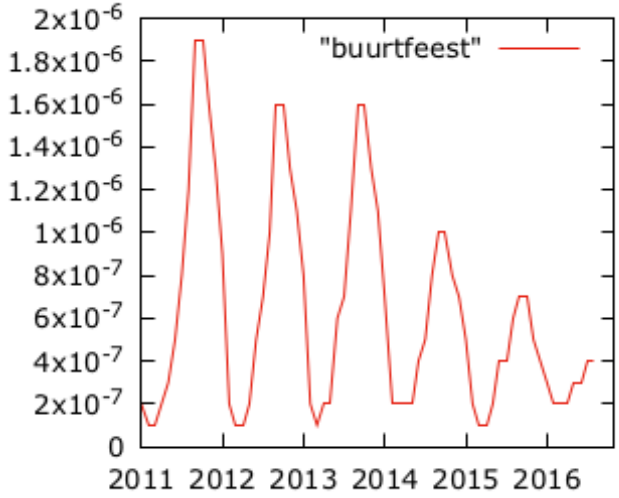
For evaluation, we ranked the words by their scores and inspected the graphs for interesting frequency patterns

We defined interesting frequency patterns as drops or rises with a factor of at least 4, resulting in lasting frequency levels

Examples of rejected frequency graphs





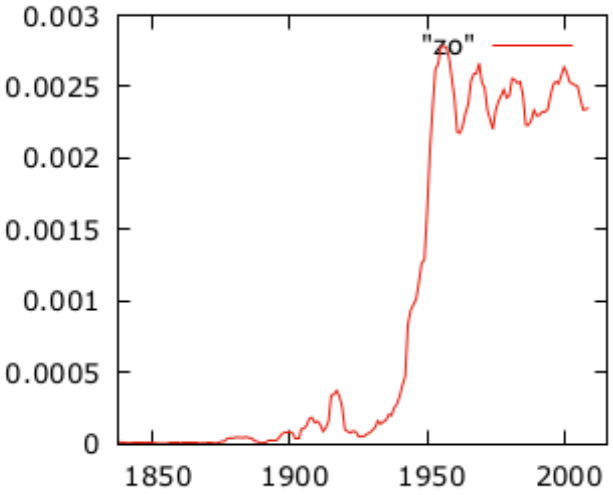


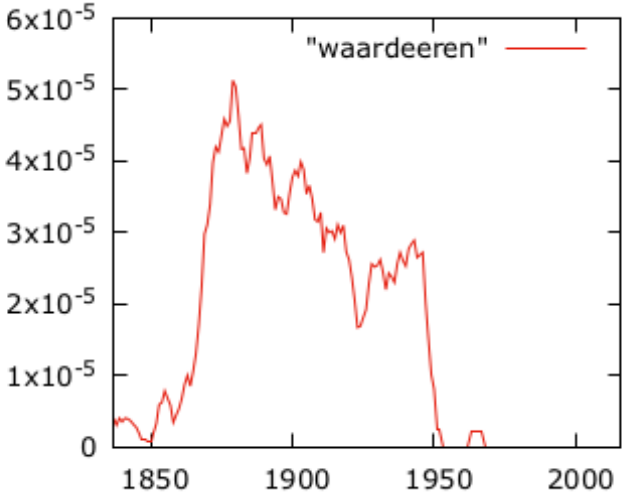
Results

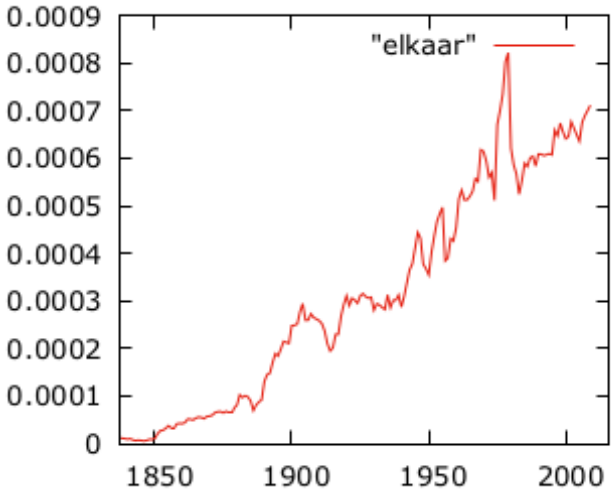
	Magazine text	Tweets
Delta scores	93%	96%
Correlation coefficients	94%	79%

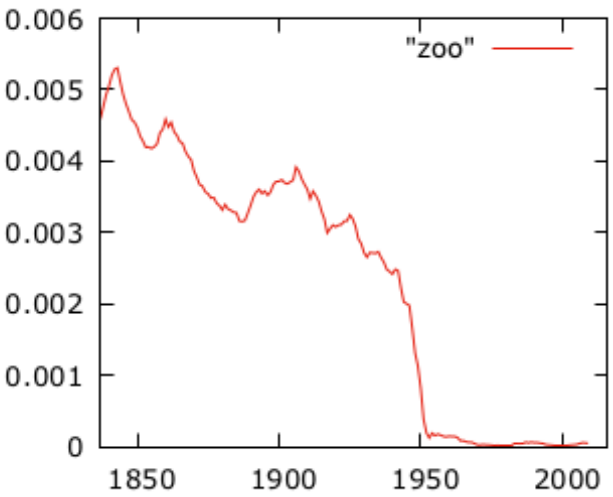
Experiment variants

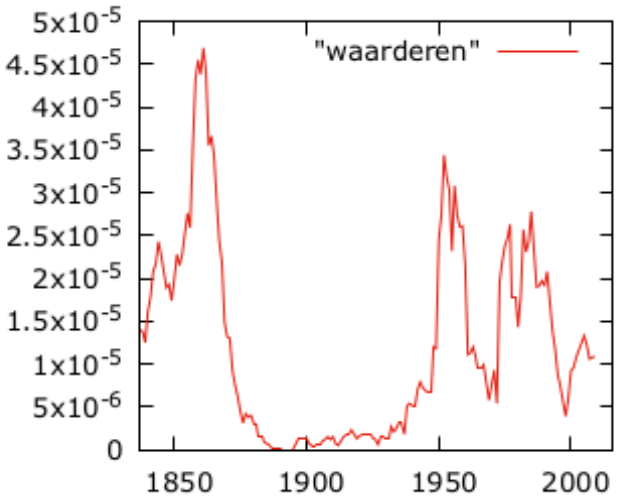
- combining frequencies of years/months for noise reduction
- consider smaller time frames: 10, 20 or 30 years/months

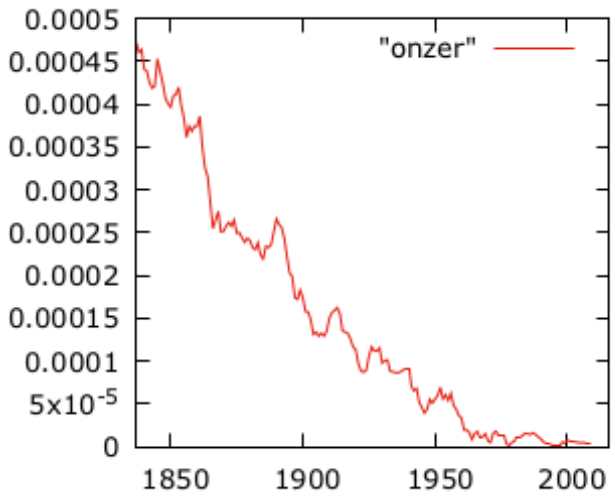

"zo"
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
1850
1900
1950
2000


"waardeeren"
6x10-5
5x10-5
4x10-5
3x10-5
2x10-5
1x10-5
0
1850
1900
1950
2000


"elkaar"
0.0009
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0
1850
1900
1950
2000


"zoo"
0.006
0.005
0.004
0.003
0.002
0.001
0
1850
1900
1950
2000


"waarderen"
5x10-5
4.5x10-5
4x10-5
3.5x10-5
3x10-5
2.5x10-5
2x10-5
1.5x10-5
1x10-5
5x10-6
0
1850
1900
1950
2000


"onzer"
0.0005
0.00045
0.0004
0.00035
0.0003
0.00025
0.0002
0.00015
0.0001
5x10-5
0
1850
1900
1950
2000

Alternative evaluations

Van der Sijs (2001) contains a list of 18,540 Dutch lemmas with their year of first usage

Only a few of the top 100 rising magazine words suggested by delta scores were part of the list: 2 (27 when also counting inflected forms)

Among the top 100 rising words from tweets, we found several neologisms: *drone*, *emoji*, *jihadist*, *koningsdag*, *matchfixing*, *onesie*, *selfie*, *smartwatch* and *yolo*

There were also rising and falling brand names: on one side: *instagram*, *netflix* and *snapchat*, and on the other: *msn* and *xbox*

Concluding remarks

We evaluated two methods for finding words with a rising or falling frequency in diachronic text

Both methods worked well, returning 90+% interesting words in the top-100 and bottom-100 of the ranked results

The top and bottom parts of the ranked lists had little overlap so the two methods can be used in parallel

The two approaches can be used for suggesting candidate neologisms and archaisms

A challenge will be how they will cope with noise in the text corpora

THE END