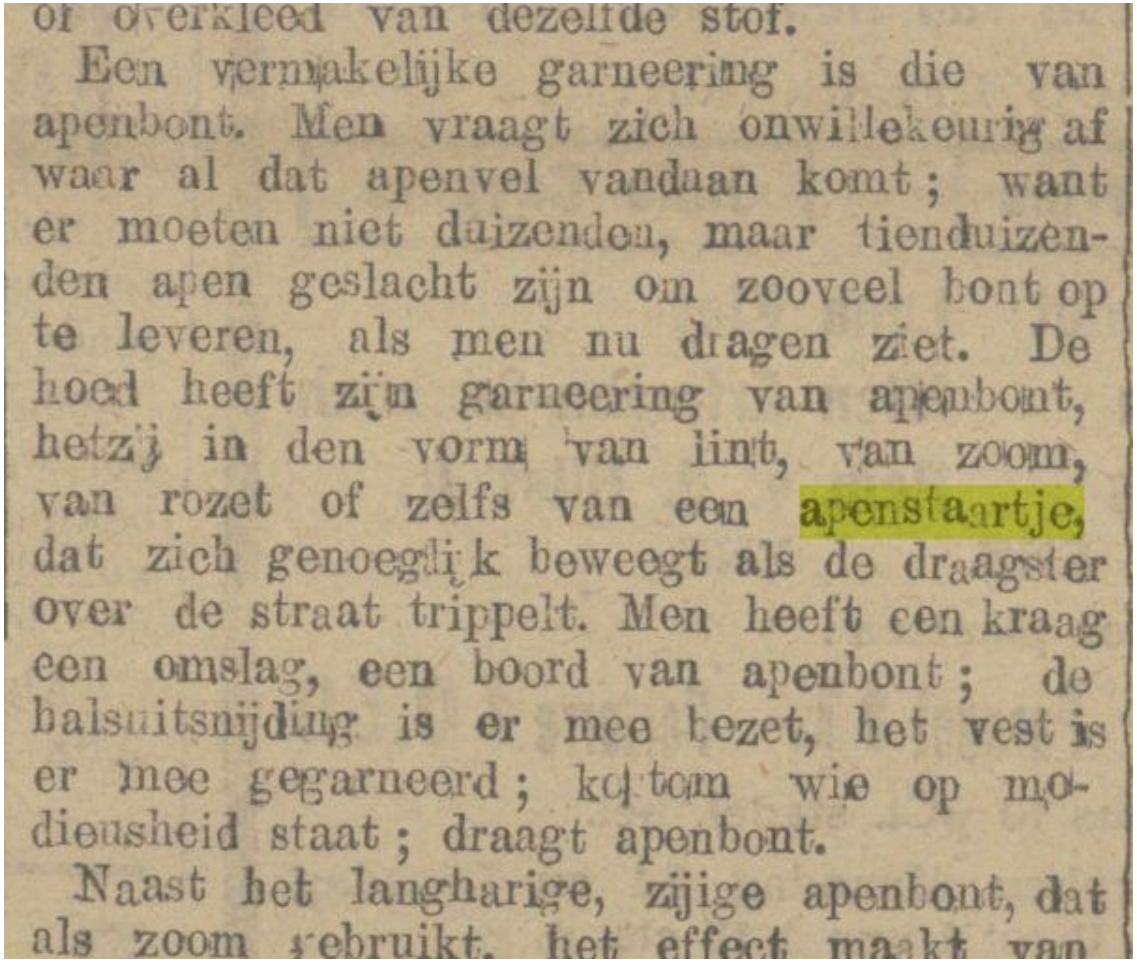
Can we spot meaning shifts in diachronic representations?

Hessel Haagsma and Malvina Nissim
hessel.haagsma@rug.nl, m.nissim@rug.nl

Centre for Language and Cognition, University of Groningen

CLIN 27
Leuven, 10 February 2017

A Monkey's Tail



of overkleed van dezelfde stof.

Een vermakelijke garneering is die van apenbont. Men vraagt zich onwillekeurig af waar al dat apenvel vandaan komt; want er moeten niet duizenden, maar tienduizenden apen geslacht zijn om zooveel bont op te leveren, als men nu dragen ziet. De hoed heeft zijn garneering van apenbont, hetzij in den vorm van lint, van zoom, van rozet of zelfs van een apenstaartje, dat zich genoeglijk beweegt als de draagster over de straat trippelt. Men heeft een kraag een omslag, een boord van apenbont; de halsuitsnijding is er mee bezet, het vest is er mee gegarneerd; kortom wie op modieusheid staat; draagt apenbont.

Naast het langharige, zijige apenbont, dat als zoom gebruikt, het effect maakt van

Tilburgsche Courant, 24-02-1919, *Voor de vrouwen*

A Monkey's Tail


'echo off' een 'apestaartje' (als regel shift 2). Als gevolg hiervan komt het 'echo off'-commando niet in beeld.
De volgende tip biedt een deeloplossing als het gaat om problemen die te maken hebben met het steeds groter worden van harde schijven en het volstampen hiervan met veel (ook steeds groter wordende) programma's etc

Nederlands Dagblad, 20-09-1993, *Apestaart, bat en functietoetsen*

Purpose

- **Ideally:** Real-time, automatic detection of meaning shifts

Purpose

- **Ideally:** Real-time, automatic detection of meaning shifts
- **Here:** detect known shifts, as proof-of-concept

Purpose

- **Linguistic:** investigate the rise of new meanings
- **Comparative:** compare meaning shifts in different corpora, demographics & languages
- **Lexicographic resources:** faster updating of sense inventories

Intuition

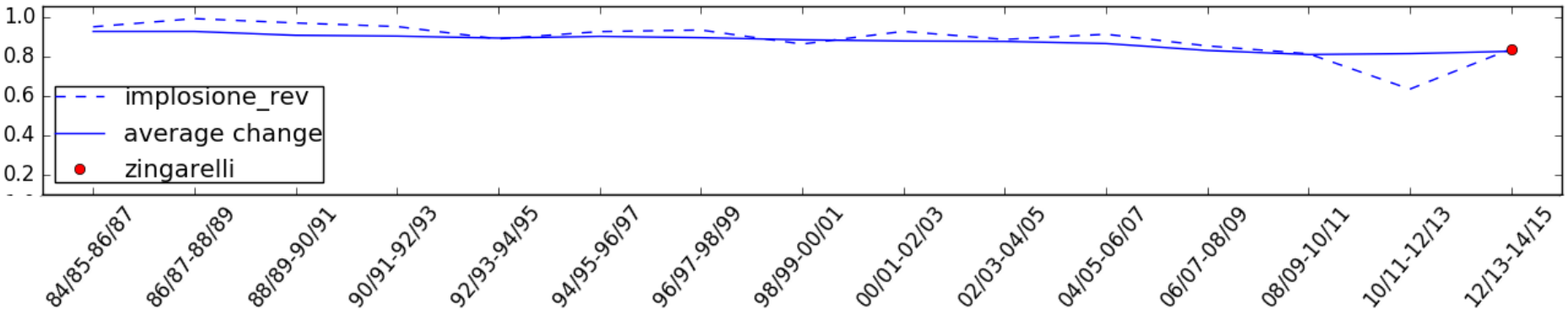
Different meaning representations:
word senses, logical symbols, embeddings.

- If meaning representations faithfully represent meaning,
- and meanings change over time,
- meaning representations should change over time!

Compare meaning representations to detect, identify, or track changes

Previous Work

- Del Tredici et al. (2016) *Tracing metaphors in time through self-distance in vector spaces*



Approach

Follow Del Tredici et al.:

- Train embeddings on per-year-slices of corpus
- Initialize embeddings of year n using embeddings of year $n + 1$
- Quantify change using self-similarity
- Manual, qualitative analysis of selected set of candidate words
- Use dictionary as 'gold standard'

Candidate Words

- laag
- vertrek
- bieden
- podium
- ziek
- breken
- bodem
- bewegen
- nazi
- hoofdrol
- keihard
- architectuur
- terreur
- recept
- management
- slim
- monument
- dodelijk
- vet
- platform
- dictatuur
- bloedbad
- formaat

Source: Van Dale & native speakers' intuitions

Criterion: Minimum Average Frequency > 10 w.p.m.

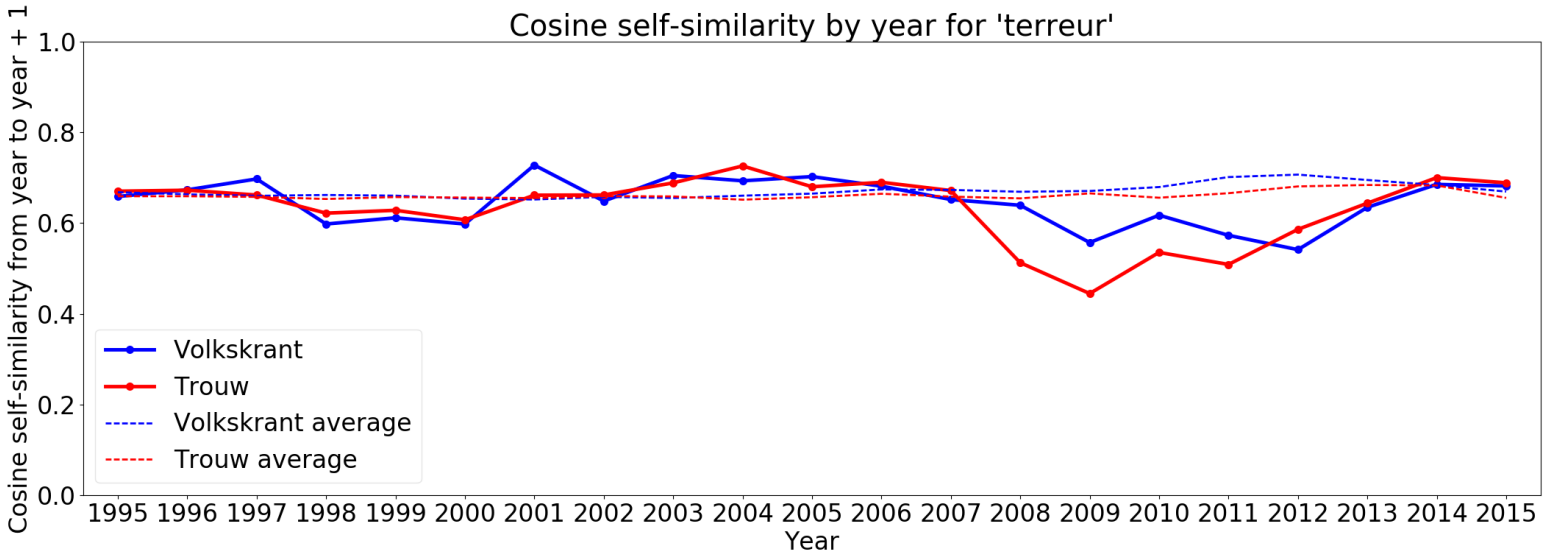
Corpus

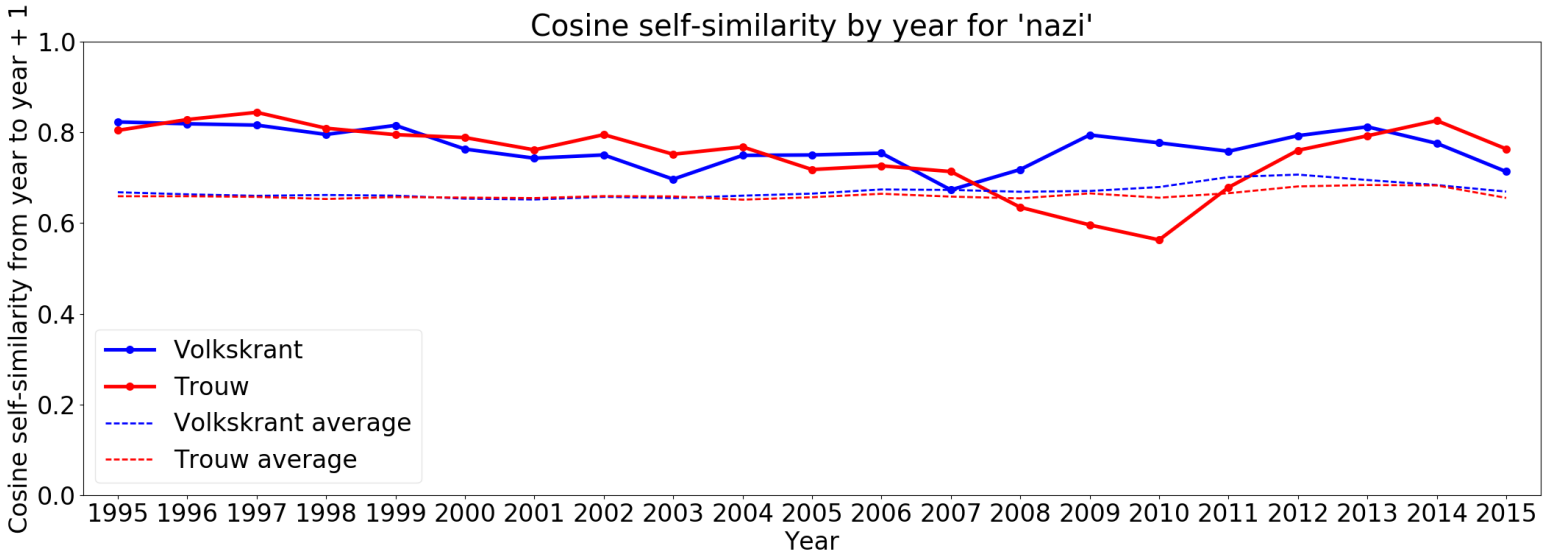
- Scraped text of 2 national Dutch newspapers, *Trouw* & *Volkskrant*
- 485M tokens in *Volkskrant*, 315M tokens in *Trouw*
- Timespan of mid-1994 to end-2016
- Tokenization using `elephant` (Evang et al., 2013)

Representations

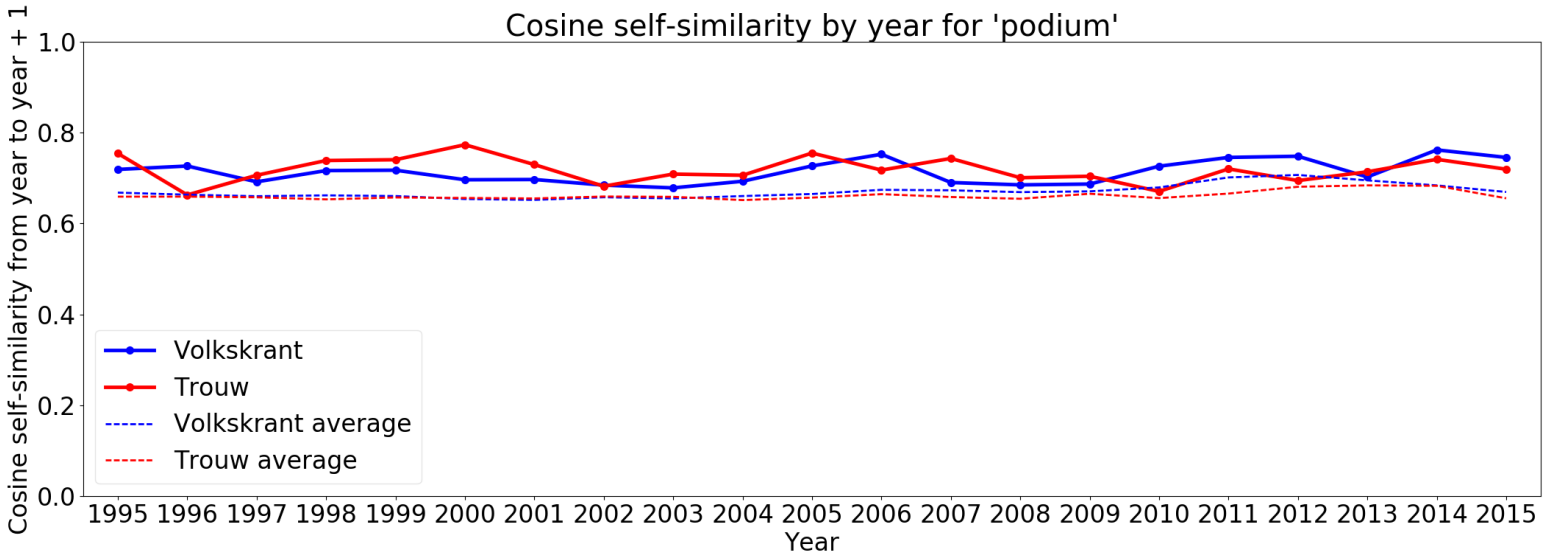
- Separate representations for *Trouw* and *Volkskrant*
- One set of embeddings per year, train **backwards**
- Use `gensim` (Řehůřek & Sojka, 2010) for `word2vec` skip-gram (Mikolov et al., 2013)
- Minimum word frequency `30`, learning rate `0.01`, dimensionality `200`, window size `5`, iterations `20`

Results

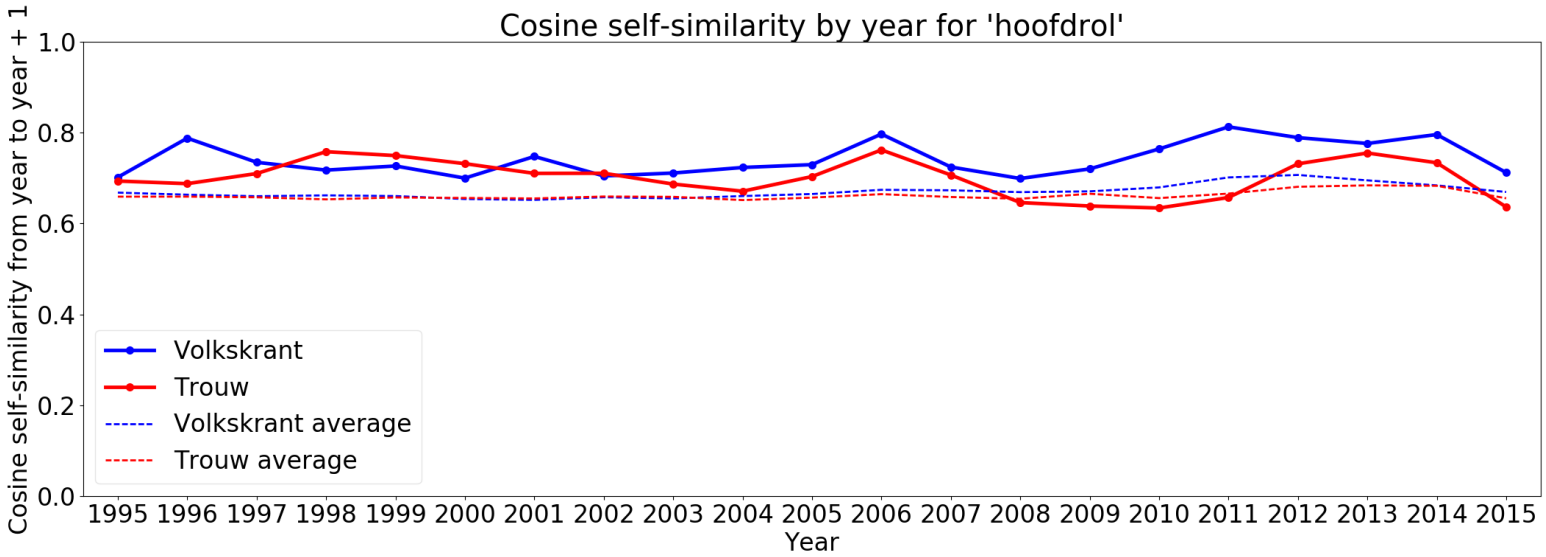

Cosine self-similarity by year for 'terreur'
Cosine self-similarity from year to year + 1
1.0
0.8
0.6
0.4
0.2
0.0
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
Year
Volkskrant
Trouw
Volkskrant average
Trouw average


Cosine self-similarity by year for 'nazi'
Cosine self-similarity from year to year + 1
1.0
0.8
0.6
0.4
0.2
0.0
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
Year
Volkskrant
Trouw
Volkskrant average
Trouw average

Problems


Cosine self-similarity by year for 'podium'
Cosine self-similarity from year to year + 1
1.0
0.8
0.6
0.4
0.2
0.0
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
Year
Volkskrant
Trouw
Volkskrant average
Trouw average

Problems

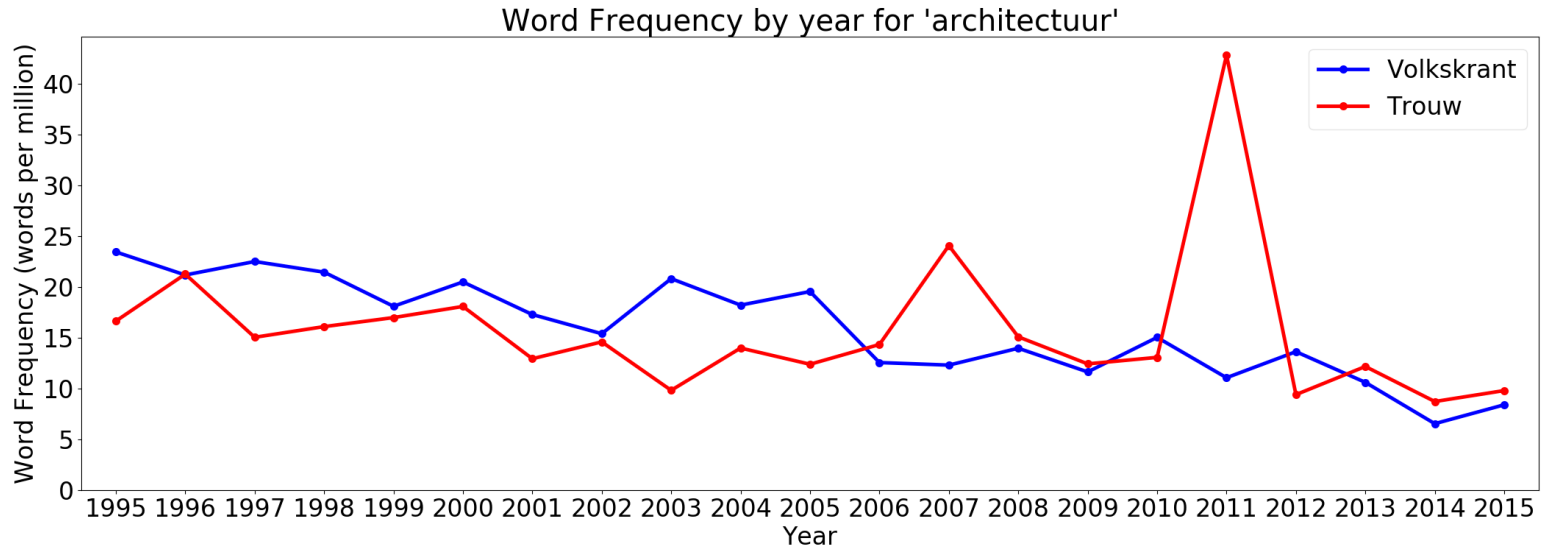

Cosine self-similarity by year for 'hoofdrol'
Cosine self-similarity from year to year + 1
1.0
0.8
0.6
0.4
0.2
0.0
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
Year
Volkskrant
Trouw
Volkskrant average
Trouw average

Problems


Cosine self-similarity by year for 'architectuur'
Cosine self-similarity from year to year + 1
1.0
0.8
0.6
0.4
0.2
0.0
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
Year
Volkskrant
Trouw
Volkskrant average
Trouw average

Problems


Cosine self-similarity by year for 'architectuur'
Cosine self-similarity from year to year + 1
1.0
0.8
0.6
0.4
0.2
0.0
Volkskrant
Trouw
Volkskrant average
Trouw average
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
Year


Word Frequency by year for 'architectuur'
Word Frequency (words per million)
40
35
30
25
20
15
10
5
0
Volkskrant
Trouw
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
Year

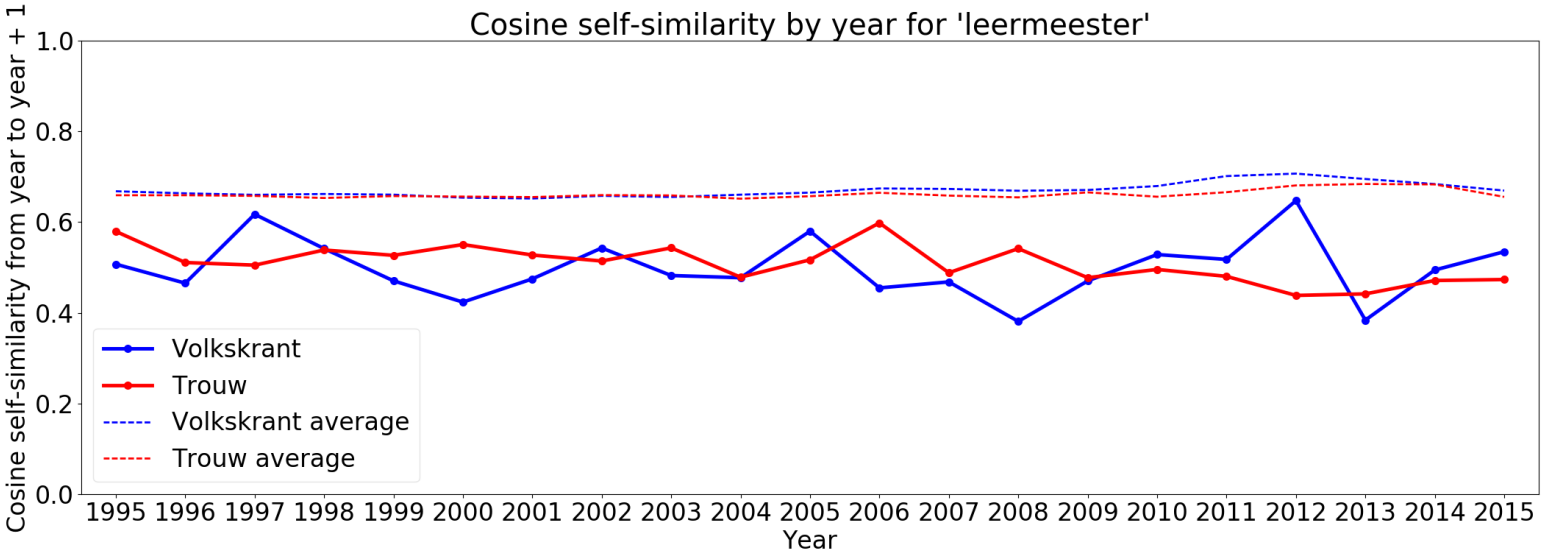
Confounding Factors

Various confounding factors:

- Word frequency affects self-similarity
- Polysemy affects self-similarity
- Low frequency leads to unstable representations

Confounding Factors

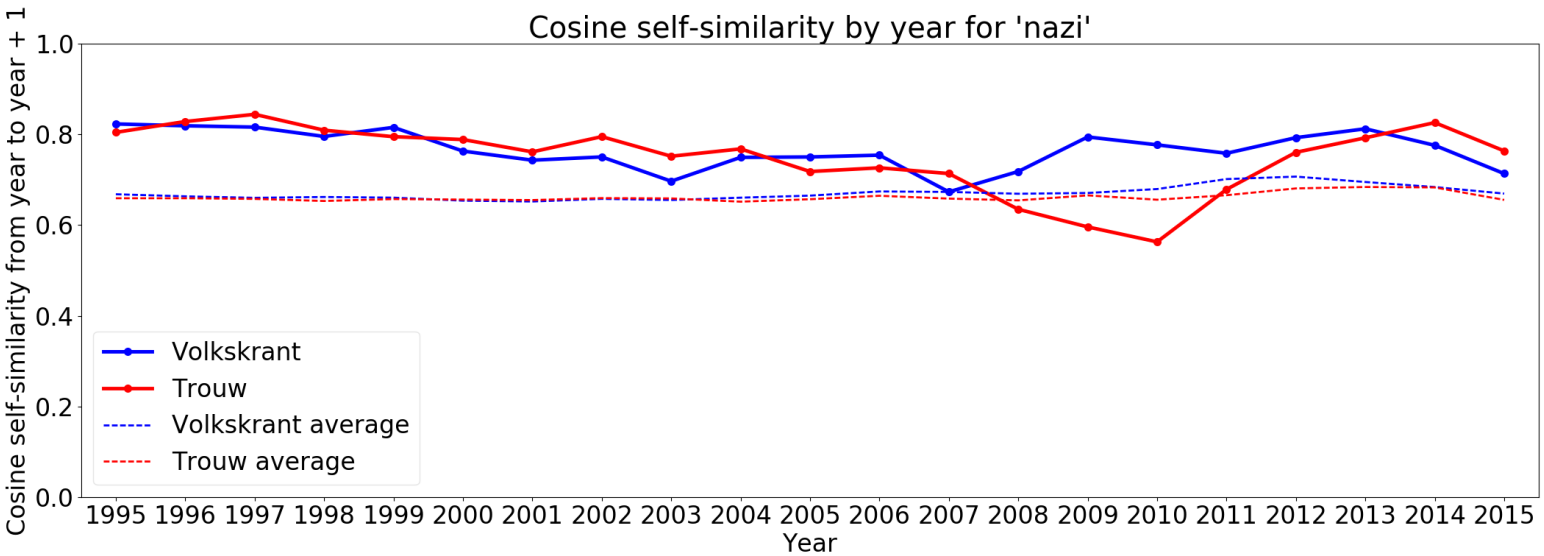
- Low frequency leads to unstable representations



Representation Issues

Problems in representation:

- Different corpora show different results
- Change in context not due to change in sense
- Compounding in Dutch



Room for Improvement

- Discount confounding effects - frequency, polysemy
- Change metric - use similarity to other words, not self-similarity
- Change method - use alignment, not reinitialization
- Change corpus - use a larger corpus (cf. Hamilton et al., 2016)
- Change evaluation - do not use dictionary as proxy
- Change analysis - reverse analysis, use bottom-up approach

The End

Frequency vs. Similarity