Social network analysis of lexical diffusion

Investigating the spread of new words on Twitter

Quirin Würschinger

LMU München

ICAME 2019, Neuchâtel
Workshop ‘Corpus approaches to social media’

1 May, 2019
Outline

- Theoretical background
  - lexical innovation
  - diffusion
- Twitter approach
  - usage intensity
  - social network analysis
  - comparative analyses
What are ‘lexical innovations’?

nonce formations  ICAMERs, Neuchâtelgate

↓

neologisms  microflat, burquini, bediquette, biobag, poppygate, emojinal, toxic, alt-right, bromance, Brexit, selfie, smartphone

↓

conventional lexemes  phone, internet, laptop
The S-curve model of diffusion

Early stages
- prestige of coiner and early adopters
- dense networks (J. Milroy and L. Milroy 1985)
- strong ties

Later stages
- diffusion via weak ties (Granovetter 1977)
How do new words diffuse?

Theoretical framework: The EC Model (Schmid forthc.)

Diffusion:

“Linking the three aspects of speakers, cotexts, and contexts, I define diffusion as a process that brings about a change in the number of **speakers and communities** who conform to a regularity of co-semiotic behaviour and a change in the types of **cotexts and contexts** in which they conform to it.”
Diffusion of lexical innovations

Research questions – two dimensions of diffusion:

1. How do neologisms diffuse across usage contexts?
   - increasing usage intensity (Stefanowitsch and Flach 2017)
   - increasing diversity in text types and semantic domains

2. How do neologisms diffuse across the speech community?
   - increasing number of speakers
   - increasing diversity of speaker communities
Diffusion across usage contexts

Previous empirical studies

- case studies: Hohenhaus 2006
- traditional corpora: Elsen 2004
- web corpora: Gérard 2017; Cartier 2017; Davies 2013; Kerremans, Stegmayr and Schmid 2012
- social media corpora: Grieve, Nini and Guo 2016

→ main focus: usage intensity
Diffusion across the speech community

Using Twitter data to study diffusion

- historical data
- informal and creative language use
- social media as a driving force in lexical innovation
- beyond usage frequency:
  - speaker information
  - sociolinguistic dynamics of diffusion
Collecting and processing Twitter data

Data collection

- Twitter’s APIs
- scraping: twint

Data overview

- neologisms: 87
- timespan: 2006–2019
- number of tweets: 32 Mill.
- number of unique users: 13 Mill.
Focus of analysis

- degree of diffusion
  - usage intensity
  - social networks of diffusion
- diffusion stages
Social network analysis of lexical diffusion

- Degrees of diffusion
- Usage intensity

Advanced diffusion: *shareable*
Unsuccessful diffusion: *microflat*
Topical diffusion: *poppygate*\(^1\)

\(^1\) *poppygate*: scandals around the ritual of wearing artificial flowers for Remembrance Day
Limited diffusion: *alt-right*²

²*alt-right:* short for *Alternative Right* after White Supremacist Richard Spencer
Limited diffusion: \textit{alt-left}

![Graph showing limited diffusion of alt-left with tweets per month from 2016 to 2019.](image)
Degrees of diffusion – Clusters

unsuccessful: *microflat*

limited: *alt-left*

topical: *poppygate*

advanced: *shareable*
Corpus examples

use of *alt-left* in 2016

The 'Alt-Left' (Black Lives Matter, Islam apologists) is far more racist, intolerant and violent than the 'Alt-Right'. Fact.


1.116 Retweets 2.229 „Gefällt mir“-Angaben

use of *alt-left* in 2017

They really hate it when we use the term "alt-left".

It would be a shame if this got 10,000 retweets. 😞

03:43 - 18. Aug. 2017

65.420 Retweets 50.793 „Gefällt mir“-Angaben
Social network analysis

Constructing the network

- Extracting nodes and edges (tidygraph, igraph):
  - Based on: mentions, retweets
  - Data format: from text, from columns (twint)

- Subsetting data: 1,000 interactions per
  1. First stage
  2. Average usage intensity
  3. Maximum usage intensity
  4. Last stage

---

3 all analyses and visualizations were done in R
Network structure\textsuperscript{4}

Nodes

- node centrality: in-degree
- node positioning: Kamada-Kawai algorithm (\textit{ggraph})

Ties

- directionality: directed
- weights: degree

Communities

- clustering: edge betweenness algorithm
- modularity: fraction of ties within vs. between sub-communities

\textsuperscript{4}all metrics and visualizations rely on \textit{tidygraph} and \textit{igraph}
Social network analysis of lexical diffusion

- Degrees of diffusion
- Social network analysis

**microflat**

subset: last (2018-11-14--2011-01-06)
Social network analysis of lexical diffusion

- Degrees of diffusion
- Social network analysis

**shareable**

**shareable**
subset: last (2018-12-31--2018-12-14)
alt-right

subset: last (2018-09-10--2018-09-09)
Social network analysis of lexical diffusion

- Degrees of diffusion
- Social network analysis

alt-left

alt-left
subset: last (2018-12-30--2018-12-19)
microflat
subset: last (2018-11-14--2011-01-06)

alt-right
subset: last (2018-09-10--2018-09-09)

alt-left
subset: last (2018-12-30--2018-12-19)

shareable
subset: last (2018-12-31--2018-12-14)
Comparison: metrics

number of communities in last 1,000 interactions

<table>
<thead>
<tr>
<th>Subcommunities / 1,000 Interactions</th>
<th>alt-left</th>
<th>poppygate</th>
<th>alt-right</th>
<th>shareable</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100</td>
<td>200</td>
<td>300</td>
<td>400</td>
</tr>
</tbody>
</table>

LEMMA
- alt-left
- alt-right
- poppygate
- shareable
Stages of diffusion
First stage

shareable
subset: first (2007-07-19--2009-08-02)
Second stage

shareable
subset: mean (2011-11-02--2011-11-28)
Third stage

shareable

subset: max (2016-03-02–2016-03-18)
Social network analysis of lexical diffusion

- Stages of diffusion
  - shareable

**Fourth stage**

shareable

subset: last (2018-12-31--2018-12-14)
Social network analysis of lexical diffusion

Stages of diffusion

shareable

All stages

shareable

shareable

shareable
subset: max (2016-03-02–2016-03-18)

shareable
subset: last (2018-12-31–2018-12-14)
First stage

alt-left
subset: first (2008-06-19--2016-08-31)
Second stage

**alt-left**

subset: mean (2016-11-02--2016-11-15)
Third stage

alt-left

subset: max (2017-08-02--2017-08-03)
Fourth stage

alt-left

subset: last (2018-12-30--2018-12-19)
Social network analysis of lexical diffusion

- Stages of diffusion

alt-left

All stages

alt-left
subset: first (2008-06-19--2016-08-31)

alt-left
subset: mean (2016-11-02--2016-11-15)

alt-left
subset: max (2017-08-02--2017-08-03)

alt-left
subset: last (2018-12-30--2018-12-19)
Comparison of diffusion stages across lexemes

alt−left
alt−right
shareable
COMM_first
COMM_mean
COMM_max
COMM_last
COMM_first
COMM_mean
COMM_max
COMM_last
COMM_first
COMM_mean
COMM_max
COMM_last
COMM_first
COMM_mean
COMM_max
COMM_last
COMM_first
COMM_mean
COMM_max
COMM_last
Comparison of communities across all lexemes

e.g. covfefe, dotard, birther, Pizzagate

---

e.g. animoji, detweet, man bun, monthversary
Users vs. usage

e.g. Brexit, blockchain, smartwatch

e.g. tweeter, bromance, man bun, ghosting
Conclusion: zooming out again . . .

- Social media data make it possible to go beyond usage intensity to study the sociolinguistic dynamics of diffusion.
- This is particularly important for cases of limited social diffusion.
- Dense networks promote diffusion in earlier stages of diffusion.
- Weak ties are crucial for advanced diffusion to new parts of the speech community at later stages.
Discussion

Legal issues

- scraping
- data vendors
- Twitter authorization
- user privacy
Social network analysis of lexical diffusion

Thanks for your attention!
The NeoCrawler
(Kerremans, Stegmayr and Schmid 2012)

Monitoring diffusion *across usage contexts* on the WWW

- sample: \(\approx 1,000\) candidates
- time window: 2011–2018
- corpus: \(\approx 800,000\) pages
- usage contexts: private forums, blogs, newspaper websites etc.
Increase in speakers for *shareable*
Increase in speakers for *alt-left*
The role of weak ties
Diffusion of *gegenpressing*

![Graph showing the diffusion of gegenpressing from 2014 to 2018.](image)

- The graph plots the number of tweets per month against time.
- The x-axis represents the years 2014 to 2018.
- The y-axis represents the number of tweets, with a range from 0 to 800.
- The data shows a peak around 2016, followed by fluctuations until 2018.
Social network analysis of lexical diffusion

Social network of **gegenpressing**

**gegenpressing**