Diffusion of Lexical Innovations
Investigating the Spread of English Neologisms on the Web and on Twitter

Quirin Würschinger, LMU Munich

FJUEL 2018
Munich
14 September, 2018
Research questions

- Which new words enter the English language?
- How do they diffuse?
- Which factors affect how they diffuse?
Which new words enter the English language?
Urban Dictionary

TRENDING RN - SEPTEMBER 13, 2018

1. fo fo fo
2. gammin
3. Solider
4. Adele Syndrome
5. sleepyboner
6. ingenuine
7. I like your shoes
8. shekels
9. FWB
10. 7/11 Was a Part Time Job
11. kissogram
12. tampico
13. Boin
14. nobber
15. sheckles
16. Throwing Neck
17. sponcon
18. kante
19. Amala
20. sit on it and rotate it
21. Cartier
22. the spongebob
23. Flossed
24. 24/7/365
25. The Crab
26. Dominatrix
27. throw neck
28. Happy 9/11
29. GRU
30. Forelsket
What is a ‘new word’?

- **nonce-formations**: used once, but have not diffused
- **neologisms**: have diffused to some degree, but are still perceived to be ‘new’
- **conventional words**: have successfully diffused and are known to the majority of the speech community
Which words enter the English lexicon?

Morphological productivity
Which words are entering the English language?

NeoCrawler: Discoverer module (Kerremans and Prokic 2018)

- goal: investigating *incipient* diffusion
- method:
  - retrieve sample of web pages
  - dictionary matching
  - semi-manual selection of candidates
  - store in database ($\approx 1,000$ lemmas)
How do new words diffuse and become conventional?
Previous work

- **cultural innovation**: S-curves (Rogers 1962; Rogers and Shoemaker 1971), big data (Kim, McFarland and Leskovec 2017)

- **sociolinguistics and language change**: mainly phonology and syntax, diffusion, early and late adopters (Labov 1980; J. Milroy and L. Milroy 1985; Croft 2000)

- **structural**: lexicalization, institutionalization, establishment (Bauer 1983; Lipka 1992)

- **corpus linguistics**:
  - recent work: large-scale studies, bigger samples (Eisenstein et al. 2014; Grieve, Nini and Guo 2016)
Figure 1: Integration of Milroy’s and Rogers’ model of diffusion stages into an S-curve (Kerremans 2015, p. 65)
How can we model diffusion?

The EC model (Schmid 2015) – a simplified account:

- **coining**: first use
- **usualization**: agreement over communicative function
- **diffusion**: spread to new usage contexts and speakers
- **normation**: establishment of norms about how to use new words
Which factors influence diffusion?

lemma-inherent (type level)

- form
  - transparency
  - productivity of word-formation pattern
  - formal appeal
- meaning
  - semantic domain
  - existing near-synonyms
  - nameworthiness

in usage (token level)

- sociolinguistic
  - density of social network
  - speakers’ prestige
- cognitive
  - formal salience in use
  - metalinguistic uses
- pragmatic
  - type of source
- emotive-affective
  - sentiment
Dimensions of diffusion

new uses bring about . . .

- spread across speakers
- spread across usage contexts

<table>
<thead>
<tr>
<th>speakers</th>
<th>usage contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>hypostatization</td>
</tr>
<tr>
<td>high</td>
<td>electron</td>
</tr>
</tbody>
</table>
How can we measure diffusion empirically?

- detecting candidates: Discoverer
- investigating diffusion
  - on the web: NeoCrawler
  - on social media: Twitter
How do new words spread on the web?
NeoCrawler (Kerremans, Stegmayr and Schmid 2012)

- weekly Google Searches (about 1,000 lemmas)
- download all html pages found
- pre- and post-processing
- corpus compilation

1Google Custom Search API
Results

Word classes

It should be noted that our data are not collected by means of a systematic sampling method, but are based on the Discoverer's capacity to detect new words on the web and on Twitter. To check whether the composition of our sample covers the spectrum of lexicological innovations as investigated by previous lexicographic work, a systematic investigation of new words recorded in the OED was conducted by the PhD candidate working on the project. A quantitative analysis of all neologisms which have entered the OED since 1800 has been found to be largely in line with the composition of word classes and word-formation processes in our sample.

1.4.2 Discussion in the light of initial hypotheses regarding factors affecting diffusion

1.4.2.1 Productivity

The distribution of our data in terms of word class and word-formation are an indicator of the productivity of the different word-formation patterns and word classes on a macroscopic scale. They reflect speakers' tendencies to coin new words and recruit word-formation patterns for lexical innovations. The distribution of word classes in our sample (see Figure 2.1) is dominated by a high percentage of nouns (79%), followed by lower percentages of adjectives (15%), verbs (12%), adverbs (1%) and phrases (1%). This is in line with the expectation that new nouns are particularly useful for naming innovative products, concepts and practices which are salient in public discourse. Our quantitative study of OED data has shown that the distribution of word classes among neologisms that have entered the lexicon since 1800 has remained very stable over time. In the period between 1950 and 2010, a total of 14,796 new nouns (69%), adjectives (22%), verbs (8%) and adverbs (1%) have been entered. In comparison with these data, our sample features a slightly higher proportion of nouns. While our sample cannot claim full representativeness, the numbers indicate that our database of neologisms at least reflects the distribution of new words in terms of word classes mirrored in the OED (with all due reservations regarding the OED's text sampling, lemma inclusion policy, etc.).

Regarding word-formation processes, compounding (37%), blending (31%) and derivation (24%) have given rise to the great majority of new words we detected (see Figure 2.2). The dominance in productivity of these three patterns is in accordance with previous studies (e.g. Bauer 1983). As with the evaluation of word classes, a quantitative comparison with OED data was drawn. While compounds (38%) and derivations (19%) account for a similarly large proportion of new words in the OED in the period between 1990 and 2010, blends (6%) are much rarer than in our data sample, even though this number has increased significantly in recent years. The rising productivity of blending in the formation of new words is in line with previous quantitative investigations which regard blends as increasingly productive formations typical of newspapers and language use on the web (Ayto 2003). The final thesis by Andrea Birkmüller (see 1.5) has shown that blending has increased in productivity over the past 20 years, both in terms of innovations as such and in terms of diffusion. The fact that blending is often assumed to produce more short-lived formations (Algeo 1998) might partly account for the higher numbers of blends in our data of incipient diffusion compared with lower numbers of blends entered as fairly conventional words in the OED.
Results

Word-formation processes

It should be noted that our data are not collected by means of a systematic sampling method, but are based on the Discoverer's capacity to detect new words on the web and on Twitter. To check whether the composition of our sample covers the spectrum of lexical innovations as investigated by previous lexicographic work, a systematic investigation of new words recorded in the OED was conducted by the PhD candidate working on the project. A quantitative analysis of all neologisms which have entered the OED since 1800 has been found to be largely in line with the composition of word classes and word-formation processes in our sample.

1.4.2 Discussion in the light of initial hypotheses regarding factors affecting diffusion

1.4.2.1 Productivity

The distribution of our data in terms of word class and word-formation processes are an indicator of the productivity of the different word-formation patterns and word classes on a macroscopic scale. They reflect speakers' tendencies to coin new words and recruit word-formation patterns for lexical innovations.

The distribution of word classes in our sample (see Figure 2.1) is dominated by a high percentage of nouns (79%), followed by lower percentages of adjectives (15%) and verbs (12%), adverbs (1%) and phrases (1%). This is in line with the expectation that new nouns are particularly useful for naming innovative products, concepts and practices which are salient in public discourse. Our quantitative study of OED data has shown that the distribution of word classes among neologisms that have entered the lexicon since 1800 has remained very stable over time. In the period between 1950 and 2010, a total of 14,796 new nouns (69%), adjectives (22%), verbs (8%) and adverbs (1%) have been entered. In comparison with these data, our sample features a slightly higher proportion of nouns. While our sample cannot claim full representativeness, the numbers indicate that our database of neologisms at least reflects the distribution of new words in terms of word classes mirrored in the OED (with all due reservations regarding the OED's text sampling, lemma inclusion policy, etc.).

Regarding word-formation processes, compounding (37%), blending (31%) and derivation (24%) have given rise to the great majority of new words we detected (see Figure 2.2). The dominance in productivity of these three patterns is in accordance with previous studies (e.g. Bauer 1983). As with the evaluation of word classes, a quantitative comparison with OED data was drawn. While compounds (38%) and derivations (19%) account for a similarly large proportion of new words in the OED in the period between 1990 and 2010, blends (6%) are much rarer than in our data sample, even though this number has increased significantly in recent years. The rising productivity of blending in the formation of new words is in line with previous quantitative investigations which regard blends as increasingly productive formations typical of newspapers and language use on the web (Ayto 2003). The final thesis by Andrea Birkmüller (see 1.5) has shown that blending has increased in productivity over the past 20 years, both in terms of innovations as such and in terms of diffusion. The fact that blending is often assumed to produce more short-lived formations (Algeo 1998) might partly account for the higher numbers of blends in our data of incipient diffusion compared with lower numbers of blends entered as fairly conventional words in the OED. With regard to the relation between...
Results

Diffusion of all candidates
Results

Top 25 items

- trumpism
- alt-right
- blockchain
- Islamophobia
- Brexit
- dumpster fire
- brexit
- fake news
- hijab
- chatbot
- post-truth
- liveblog
- glamping
- Trumpian
- Gen Z
- trippy
- upskill
- momentous
- adulting
- virtue signalling
- DACA
- upcycling
- matcha
- circular economy

To give an idea of the words, we have selected the 25 top-ranking items (Figure 1.2), the 25 items in the middle of the frequency distribution, i.e. 12 above and below the median (Figure 1.3) and the 25 items at the end of the tail (Figure 1.4), omitting those which are not attested in the period. Note that these are cumulative counts which do reflect how long the words have existed. For example, the top-ranking item Trumpism has been in use over the whole period, while the compound fake news was coined later and has only been monitored for 61 weeks. Nevertheless, we regard cumulative counts as a suitable approximative indicator of diffusion in general, because they reflect the number of uses on the web and the number of occasions for the average Internet user to have come across these words on the web.

On the basis of this rationale, it seems legitimate to argue that the 25 top-ranking items listed in Figure 1.2 are the neologisms which have caught on best, while those in the middle part of the frequency line are less strongly conventionalized and those at the end of the tail are not at all conventionalized so far. This conclusion is confirmed by intuitive assessments of the items listed and by our questionnaire results, which suggest that frequency is a strong indicator of which words are generally more familiar to individual speakers. It should be noted that the data gloss over differences in the patterns of diffusion already identified by Kerremans (2015). These patterns are confirmed by the larger dataset we now have at our disposal. The main patterns are:

- fast and sustained diffusion, illustrated by most of the top-ranking items in Figure 1.2, although some, e.g. liveblog, show a less steep increase in the early stage after coinage;
- no diffusion, illustrated by the long tail of Figure 1.1 and 1.4;
- topical diffusion after noteworthy events, with subsequent reduction of usage intensity or sporadic topical peaks, e.g. Grexit, Catalexit, creepy clown;
- cyclical changes in usage frequency, depending on seasons, repeated events like elections or sports events, e.g. veganuary.

1.4.1.2 Distribution in terms of word class and underlying word-formation patterns

Figure 2 provides a summary of the data with regard to their distribution across word classes (Figure 2.1) and underlying word-formation patterns (Figure 2.2), counting the topmost or final word-formation process in the word-internal hierarchy if several apply.
Results

Items around median

- breadcrumb
- farmsler
- Trumple Dumple
- bed-blocking
- twatter
- ringxiety
- sologamy
- momster
- oniochalasia
- decyling
- bennifer
- catcaller
- oblication
- bloggergate
- milkshake duck
- metamour
- bloglet
- sexomniac
- germaphobe
- device mesh
tarology
- catfishing
- building
- epicanciacy
- loadly
Results
Bottom 25 items
How do new words spread on Twitter?
Methodology

▶ advantages
  ▶ going back in time
  ▶ high temporal resolution
  ▶ user metadata (social, geographic)
  ▶ social network data

▶ tools
  ▶ ongoing Twitter mining: TAGS
  ▶ web scraping: Twitter Scraper
A Case study of *alt-right* and *alt-left*

Background of *alt-right*

clipped form of earlier term *Alternative Right*, coined by White Supremacist Richard Spencer
A Case study of *alt-right* and *alt-left*

Background of *alt-left*

formed in analogy (and opposition) to pre-existing *alt-right*
Corpus examples

use of *alt-left* in 2016

Paul Joseph Watson
@PrisonPlanet

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
Corpus examples

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
Zooming in on diffusion
August 25, 2016: Hillary Clinton’s speech against alt-right
November 22, 2016: Trump publicly defends Steven Bannon
August 12, 2017: Charlottesville Rally
August 16, 2017: Trump attacking ‘alt-left’
Zooming in on diffusion

the ‘social’ network
Social network analysis

<table>
<thead>
<tr>
<th></th>
<th>alt-left</th>
<th>alt-right</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of tweets</td>
<td>295,968</td>
<td>1,760,777</td>
</tr>
<tr>
<td>number of individual speakers</td>
<td>117,607</td>
<td>550,798</td>
</tr>
<tr>
<td>avg. weighted degree</td>
<td>0.855</td>
<td>1.044</td>
</tr>
<tr>
<td>modularity</td>
<td>0.937</td>
<td>0.877</td>
</tr>
</tbody>
</table>

→ *alt-right* shows a high degree of diffusion over an extended time window

→ *alt-left* shows some diffusion, but remains to be used by smaller pockets of the speech community
Implications

- S-curves not to be expected due to effects of topicality\textsuperscript{2}
- differentiated view on diffusion: sub-communities
- 'influencers' drive innovation
- social network characteristics influence diffusion

\textsuperscript{2}and for other reasons that we could discuss ...
Thanks!

 strdup (OED Word of the Year 2015)