Using the Web and Social Media as Corpora for Monitoring the Spread of Neologisms.

The case of *rapefugee*, *rapeugee*, and *rapugee*.

Quirin Würschinger,
Mohammad Fazleh Elahi,
Desislava Zhekova,
Hans-Jörg Schmid

LMU Munich

ACL 2016, WAC-X
12 August, 2016

Table of Contents

1. Motivation
2. Related work
3. The case of *rapefugee*, *rapeugee* and *rapugee*
4. Data collection
5. Results
6. Discussion
7. Conclusion and future work

Project context

- Title: *Incipient diffusion of lexical innovations* (funded by DFG, grant SCHM 1232/5-1)
  - longitudinal study of English neologisms
  - development of a web crawler: *NeoCrawler* (http://www.neocrawler.de)
- Chair for Modern English Linguistics, LMU Munich, principal investigator: Prof Dr Hans-Jörg Schmid
- in cooperation with the Center of Information and Language Processing Munich (CIS):
  - Prof Dr Hinrich Schütze
  - Dr Desislava Zhekova

Motivation

Goal: systematic investigation of the spread of English neologisms

- detect neologisms as close to their coining as possible
- observe their conventionalization process
- analyse factors responsible for their (successful) establishment
    - prestige of coiner
    - nameworthiness of concept
    - transparency
    - formal appeal
    - ...

Related work: Publications

Investigating neologisms . . .

- based on traditional corpora:
  - usage contexts for 5,000 neologisms in a newspaper corpus (Bauer & Renouf, 2000)
  - productivity of prefixes such as *techno-* and *cyber-* plus development of four neologisms in newspaper articles (Renouf, 2007)
  - methods for retrieving and extracting neologisms from a 45-million-word corpus based on *Nature* (Paryzek, 2008)
- based on web data:
  - tracing *bouncebackability* in a web corpus (Hohenhaus, 2006)
  - harvesting neologisms from a *Wikipedia* corpus (Veale & Butnariu, 2010)
  - investigating succcess predicting factors (Grieve, Nini & Guo, 2016)

Related work: Websites

- *New Words* by Merriam Webster
- *About words* by Cambridge University Press
- *Urban Dictionary*
- *WordSpy: Dictionary of New Words*
- *Wortwarte* (Lemnitzer, 2011)

The case of *rapefugee*, *rapeugee* and *rapugee*

- socio-political background:
  - refugee crisis
  - New Year's Eve 2016: sexual assaults by refugees in Cologne
  - disclaimer: We strongly oppose any xenophobic motivations!
- linguistic background:
  - blends of *rape* and *refugee*
  - common meaning: {'rape' / 'refugee'}
  - onomasiological competition of different forms for occupying this semantic space
    - *rapefugee*
    - *rapeugee*
    - *rapugee*

Measuring conventionalization

- occurrences: single vs. multiple within one text
- absolute frequencies: overall usage intensity
- relative frequencies: relative dominance of each variant in the corpus
- special uses
    - web
        - token position: title of websites
        - metalinguistic uses: operationalized as uses in inverted commas (e.g. *“rapefugee”*, *‘rapefugee’*)
    - Twitter
        - hashtags: tokens preceded by #
        - retweets: tweets marked by the tag *RT*

Web corpus composition

- We updated and extended a previous version of the NeoCrawler (Kerremans, Stegmayr & Schmid, 2012), using Google Web Search
- intervals: weekly crawls
- timespan: from October 19th, 2015 until March 16th, 2016

	single	multiple	title	metaling.	total # words
rapefugee	169	849	125	59	273,961
rapeugee	122	281	24	3	627,077
rapugee	21	41	6	1	51,590

Table: Descriptive summary of data from the web corpus

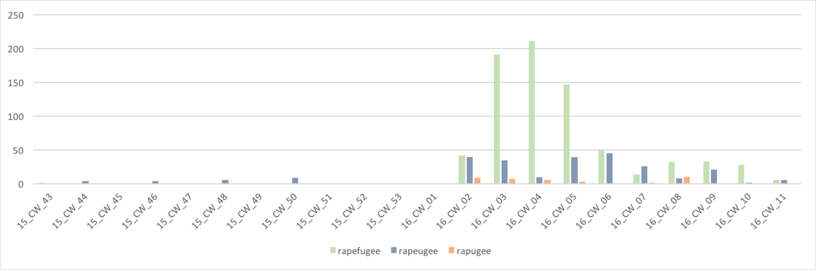
Twitter corpus composition

- using the *REFUGEE* corpus (Zhekova, 2016)
- timespan: October 19th, 2015 until March 16th, 2016
- collected using the Twitter Streaming API
- tracking the keyword *refugee*

	single	multiple	hashtag	direct	tweet	retweet	total # words
rapefugee	3,777	3,786	3,303	451	1,024	2,753	77,369
rapeugee	272	277	220	52	87	185	5,909
rapugee	92	92	88	4	22	70	1,740

Table: Descriptive summary of data from the Twitter corpus

Web corpus: usage intensity

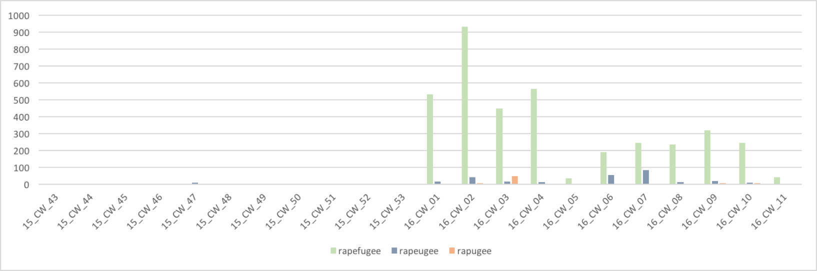


- before New Year:
  - little usage intensity for all three types
  - relative dominance of the variant *rapeugee*
  - discontinuous spikes → language-external triggers
- New Year turn:
  - increase in use for all three variants
  - trigger: sexual assaults in Cologne

Web corpus results

- language-external trigger: introduction of sexual education in courses for refugees in Denmark:
    - *Denmark has a rapeugee problem: They want to give the new 'migrants' classes so they don't rape the locals and the livestock. Sorry but classes aren't going to help with these savages.* (29 October 2015)
- usage types:
    - tokens in titles: 16 %
    - metalinguistic uses: often in non-disparaging function
        - New York Post (10 January 2016): *German clash over 'rapefugees' who carried out mass sex attack*

Twitter corpus: usage intensity



- Before New Year:
  - little usage intensity for all three variants
  - total dominance of *rapeugee*, no instances of the other two variants
- New Year turn:
  - one week earlier than on the web
  - immediate, strong dominance of *rapefugee*

Twitter corpus results

- language-external trigger: Cologne assaults
    - *Refugee = rapist. Flüchtling = Vergewaltiger. #Cologne #rapefugees* (6 January 2016)
- usage types:
    - 87 % of tokens used as hashtags
    - retweets/tweet ratio: 2.7

Competition across both corpora

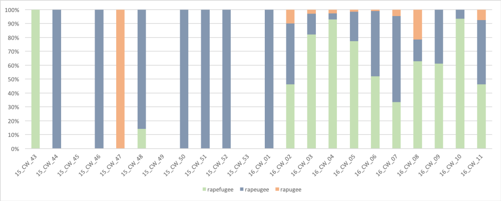


Figure: Relative frequencies in the web corpus

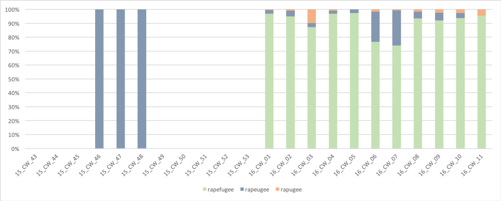


Figure: Relative frequencies in the Twitter corpus

Discussion

- The conventionalization of neologisms is not continous, but happens in spurts which are triggered by language-external events.
- Twitter reacts more quickly to these developments than the web.
- Twitter-specific features foster the spread of new words
- Twitter's most popular variant – *rapefugee* – spreads to also become most intensely used on the web.
- The results from Twitter and the language-external events can be regarded as cross-validation for the web corpus' results.

Conclusion and future work

Conclusion:

- Web and social media data can be effectively used to study the conventionalization of neologisms.
- Language use on social media platforms like Twitter shows community-specific characteristics.
- The use of new words on social media significantly affects the use across the whole web.
- Using the web as a corpus can provide a more balanced and representative data sample.

Future work:

- discover and observe large-scale set of neologisms
- analyse diffusion of neologisms into different domains-of-discourse

Thanks!

References

Bauer, L. & Renouf, A. (2000). Contextual clues to word-meaning. *International Journal of Corpus Linguistics*, *5*, 231–258.

Grieve, J., Nini, A. & Guo, D. (2016). Analyzing lexical emergence in Modern American English online. *English Language and Linguistics*.

Hohenhaus, P. (2006). Bouncebackability. A web-as-corpus-based study of a new formation, its interpretation, generalization/spread and subsequent decline. *SKASE Journal of Theoretical Linguistics*, *3*, 17–27.

Kerremans, D., Stegmayr, S. & Schmid, H.-J. (2012). The neocrawler: identifying and retrieving neologisms from the internet and monitoring ongoing change. *Current Methods in Historical Semantics*, 59–73.

Lemnitzer, L. (2011). Making sense of nonce words. In M. H. Andersen & J. N. Jensen (Eds.), *Sprognaevets konferenceserie 1* (pp. 7–18). Nye Ord. Kopenhagen.

Paryzek, P. (2008). Comparison of selected methods for the retrieval of neologisms. *Investigationes Linguisicae*, *16*, 163–181.

Renouf, A. (2007). Tracing lexical productivity and creativity in the British Media: 'The Chavs and the Chav-Nots'. *Lexical Creativity, Texts and Contexts*, 61–92.

Veale, T. & Butnariu, C. (2010). Harvesting and understanding on-line neologisms. *Cognitive perspectives on word formation*, 399–418.

Zhekova, D. (2016). Using Contemporary Media for the Humanities: The REFUGEE Twitter Corpus. *Digital Scholarship in the Humanities*. (submitted).