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 and Hans-Jörg Schmid
LMU Munich
80539 Munich, Germany
q.wuerschinger@lmu.de, fazleh.elahi@anglistik.uni-muenchen.de,
desi@cis.uni-muenchen.de, hans-joerg.schmid@anglistik.uni-muenchen.de

Abstract

This paper employs both a web-as-corpus and a Twitter-as-corpus approach to present a longitudinal case study of the establishment of three recently coined, synonymous neologisms: *rapefugee*, *rapeugee* and *rapugee*. We describe the retrieval and processing of the web and Twitter data and discuss the dynamics of the competition between the three forms within and across both datasets based on quantitative summaries of the results. The results show that various language-external events boost the usage of the terms both on the web and on Twitter, with the latter typically ahead of the former by some days. Beside absolute frequencies, we distinguish between several special usages of the target words and their effects on the establishment process. For the web corpus, we examine target words appearing in the title of websites and metalinguistic usages; for the Twitter corpus, we examine hashtag uses and retweets. We find that the use of hashtags and retweets significantly affects the spread of the neologisms both on Twitter and on the web.

1 Introduction

Electronic mass communication offers unique opportunities for the study of new words and the early phases of their establishment. Using the web and social media like Twitter as corpora offers an economical way of investigating whether newly coined words are taken up by language users and begin to spread and diffuse into other domains of discourse. Such investigations require longitudinal studies which keep track of new occurrences of neologisms on the web and/or in posts on Twitter and other social media.

This paper presents a web-as-corpus and Twitter-as-corpus study of the spread of three recently coined words which emerged in 2015 and compete for encoding the same meaning: *rapefugee*, *rapeugee*, and *rapugee*. All three target words are formed by blending the source words *rape* and *refugee*, and all three are mainly used as derogatory propaganda terms by opponents of policies that welcome asylum-seekers. We would like to note that our work does not support, but only explores and analyses the use of these terms, equally applicable to any other neologism.

The approach chosen in this paper complements an earlier study by Kerremans et al. (2012), who investigated the competition between the meanings of one polysemous neologism, viz. the verb *to de-tweet*. Analyzing material collected by means of a tailor-made webcrawler, the so-called *NeoCrawler*, the authors show how language users gradually begin to converge on one meaning, 'to sign off (from Twitter)', following a period where different users associate different meanings with the form and even explicitly promote them.

The current project addresses the mirror-image situation where several synonymous forms compete for encoding the same meaning. Investigations of this type are important for understanding how new words spread, because competition between forms is one of the factors that influence this process. Extending the methodology used in (Kerremans et al., 2012) in a second direction, we compare the data from the web with a second dataset collected for the same period from Twitter. We aim to provide a dense-data longitudinal analysis of the rivalry between these three recent neologisms, both separately within the web and the Twitter data and in comparison between these two

Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task, pages 35–43,
Berlin, Germany, August 7-12, 2016.

data sources. In the course of this, we discuss the specific advantages and challenges involved in retrieving, processing and analyzing data from the web and from Twitter respectively.

2 Related work

Efforts to investigate neologisms with the help of web-based data have been stepped up considerably over the past years. There are numerous websites, run by dictionary publishers or based on crowdsourced user-content, which list and define new words and provide selected quotations, often including the first known attestation. Prominent examples are *New Words* by Merriam-Webster[1], *About words* by Cambridge University Press[2], *UrbanDictionary*[3], and *WordSpy*: *Dictionary of New Words*[4]. A comparable project for German is *Wortwarte*[5], which documents German neologisms based on newspaper data (Lemnitzer, 2011).

As far as research projects on neologisms which apply the web-as-corpus method are concerned, Bauer and Renouf (2000) investigate the contexts of use for 5000 neologisms in a newspaper corpus. Combining data from a newspaper corpus and the web, Renouf (2007) analyzes the recent productivity of prefixes such as *techno-* and *cyber-* and traces the frequency development of four neologisms in newspaper articles. Hohenhaus (2006) investigates the word *bouncebackability* by means of the web-as-corpus method. Paryzek (2008) reviews different methods of retrieving neologisms and extracts neologisms from a 45-million-word corpus based on Nature. Veale and Butnariu (2010) harvest neologisms from a corpus which is derived from the English version of Wikipedia. Like the study by Kerremans et al. (2012) mentioned above, Grieve et al. (2016) aim to unveil the factors behind the emergence and success of neologisms. This is also the question that motivates the work presented in this paper.

3 Operationalizing the research question

As pointed out above, we aim at a comparative longitudinal analysis of attestations of three synonymous words on the web and on Twitter in order to investigate the dynamics of the competition between them. To operationalize this research question, the following types of data and data analyses must be provided by computational means:

- Absolute frequency counts of occurrences of the three words on the web and on Twitter over a defined period of time in a high temporal solution (i.e. weekly/daily counts of newly added occurrences). These counts are required to obtain a measure of *usage intensity as such* (cf. Stefanowitsch and Flach (forthcoming)).
- Relative frequency counts of the three words per time interval (days of weeks), i.e. the frequency of each word relative to the frequencies of the other two for the same time interval. For example, we detected a total number of 233 tokens across all three formal variants in the web corpus in the third week of January 2016. The variant *rapefugee* amounts to 191 occurrences, which corresponds to a relative frequency of about 0.82. These relative frequency counts are required to measure the *current relative success* of the three forms to occupy the onomasiological target space.
- A longitudinal analysis of the changes in absolute and relative frequencies over time: this is required to measure *the dynamics of the temporal development of relative success*. Examples can be found in Figure 1 and Figure 3.
- Classificatory analyses of different usage types of the three words which are suspected to have *differential effects on their chances* of being taken up again and thus being spread. Specifically, what we are interested in are:
  - *single* object-linguistic uses as opposed to
  - *metalinguistic* uses of talking about the word rather than actually using it (e.g. *Whenever people hear "refugee" they need to think #rapefugee.* (Tweet from 7 January 2016))
  - *multiple* uses within one web page / tweet as well as repetitions via *retweets*
  - uses as *hashtags* on Twitter or as parts of *titles* of web pages.

4 Data acquisition

4.1 Web as a corpus

We used the NeoCrawler (Kerremans et al., 2012) to collect timestamped web pages containing

[1] http://nws.merriam-webster.com/opendictionary

[2] https://dictionaryblog.cambridge.org/category/new-words/

[3] www.urbandictionary.com

[4] http://www.wordspy.com/

[5] www.wortwarte.de

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

Table 1: Descriptive summary of data from the web corpus

tokens of the three neologisms on the web. In order to have a comparable sample, we restricted the search to the timespan in which the Twitter data has been collected (see Section 4.2), namely from October 19th, 2015 until March 16th, 2016. The NeoCrawler uses Google searches for collecting web pages, as this has several benefits for neologism research (Lewandowski, 2008; Kerremans et al., 2012): Google provides the largest number of indexed pages, its index is updated fastest in comparison to other search engines, and it provides the web pages which are most relevant for a given search string.

The NeoCrawler searches by means of an automated version of the processes carried out in manual Google searches. The system builds a search string[6] defining values for a number of parameters (such as language, date, token etc.). There are several advantages of this approach over other Google search APIs[7], such as *Custom Search Engine* or *Google Site Search*. While the main functionality provided by *Custom Search Engine* is to search across a set of sites specified, it can also be configured to search the whole web. However, in that case, it provides a smaller number and less relevant search results than a manual Google search, which is not desirable if the project requires maximum recall. *Google Site Search* is an edition of *Google Custom Search* that provides additional functionality, but does not solve the problem either. Therefore, neither of these APIs is suitable for our goal, as we need to search the whole web in order to get as many relevant search results as possible. The automated version of the Google manual search implemented in the NeoCrawler is an optimal fit for our purpose. However, a large number of potential hits returned by Google searches turn out to be either false positives (i.e. pages that do not contain the search token), duplicate copies or otherwise useless pages. Therefore, we extracted only the pages containing the search token excluding duplicates and empty pages.

Following the operationalization procedure outlined in Section 3 above, we distinguished between single (each page is counted as a single occurrence independently of how often a neologism has been used on it) and multiple occurrences per page (each token on the page is counted separately), and between special usage types (i.e. usage in the title of a document) and metalinguistic usage (operationalized as uses in inverted commas). Table 1 shows a summary of the web data.

A key requirement for the longitudinal analysis of the temporal dynamics is to identify the correct timestamp of the web content that contains a given token. However, due to the decentralized nature of timestamps and the lack of standard meta-data for time and date, reliable timestamps are frequently not available for web documents. In its previous version, the NeoCrawler extracted the remote timestamp of the retrieved document using the CURL module for PHP, which is a library for getting files from various Internet protocols including HTTP/HTTPS. However, since CURL relies on the *Last-Modified* header value of the HTML page to extract the timestamp, which is often missing, it was impossible to extract a timestamp from a large proportion of the documents. Therefore, we have extended the NeoCrawler to extract the timestamp from the Google search page directly, where Google provides the timestamp of the content containing the token instead of that of the last update of the web page. Moreover, the NeoCrawler extracts both the absolute (i.e. 12/01/2016) and the relative (i.e. *a week ago*) timestamp found on the web page. It must be conceded, however, that Google's timestamps are not always correct either, among other things because the location of the content and its respective timestamp on the page is ambiguous, or because there are several tokens added at different dates to

[6] `https://encrypted.google.com/search?num=100&hl=en&lr=lang_en&start=0&tbs=lr%3Alang_1en%2Ccdr%3A1%2Ccd_min%3A10%2F01%2F2015%2Ccd_max%3A03%2F16%2F2016&q=%22rapefugee%22`

[7] `https://developers.google.com/custom-search/json-api/v1/overview`

	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 2: Descriptive summary of data from the Twitter corpus

the same page. In the latter case, only a single timestamp is provided by Google. Results related to the temporal development will be given in Section 5 below.

4.2 Twitter as a corpus

Unlike the web, Twitter cannot be queried for past events in an unlimited manner. Only the Firehose Twitter API[8], which is of highly limited access, can be used to collect all public statuses. An open access equivalent for part of this functionality is the Twitter Streaming API[9] which provides low latency access to Twitter's current global stream of data (i.e. a sample of the current stream fulfilling the query). However, the current Twitter stream cannot aid us in our attempt to observe how the three neologisms *rapefugee*, *rapeugee* and *rapugee* have been used since the time of their coining. The Twitter Search API, searches only against a sampling of recent Tweets published in the past seven days. Yet, the tokens have been in use a lot longer than seven days.

The only way to query Twitter for older posts is via using previously collected Twitter corpora. Based on the fact that the neologisms of interest are different blends of *rape* and *refugee*, we made use of an extended version of the REFUGEE corpus (Zhekova, 2016), which consists of tweets that were collected from October 19th, 2015 until March 16th, 2016 via the Twitter Streaming API by tracking the token *refugee*. We assume that the linguistic relation between the three neologisms and *refugee* will result in a representative sample of Twitter data containing these new words.

Another difference between Twitter and web data is that the meta-information is readily available in Twitter. Unlike in the web data, all relevant tweets are precisely timestamped. With respect to token identification and classification (single, multiple, metalinguistic use), we followed the same approach as for the web data. Additionally, for the Twitter corpus, we observed the difference between direct vs. hashtag usage (i.e. *No rapefugees!* vs. *No #rapefugees!*) and normal tweets vs. retweets (i.e. *No #rapefugees!* vs. *RT No #rapefugees!*). Table 2 provides a basic summary of the occurrences of the three neologisms in the Twitter data.

[8] https://dev.twitter.com/streaming/firehose

[9] https://dev.twitter.com/streaming/overview

5 Results

5.1 Web corpus

Usage intensity. In order to measure usage intensity (Stefanowitsch and Flach, forthcoming), we conduct absolute frequency counts of tokens for all three types (*rapefugee*, *rapeugee* and *rapugee*) in both datasets. We count multiple tokens per type within one website or one tweet separately. The counts are accumulated in weekly intervals corresponding to each calendar week in the timespan between October 19th, 2015 (i.e. 15_CW_43 – to be read as the 43rd calendar week of 2015) and March 16th, 2016. Figure 1 presents the absolute usage frequencies in the web corpus.

The graph shows a very small number of uses of the three types before 16_CW_02, with a maximum of 9 tokens of the form *rapeugee* in 15_CW_50. The period after New Year marks a turning point, after which numbers rapidly increase, with a maximum of 233 tokens in 16_CW_03.

The first attestation of any of the three target forms on the web is a single occurrence of *rapefugee* on January 19, 2015 (15_CW_43 in Figure 1).

Only a few days later, however, the type *rapeugee* appears and initially supersedes the other two types in popularity, representing an accumulated 79 % of all tokens of all three types in the period before the New Year turn. In 16_CW_02, the numbers for all three types rise significantly, indicating an increasing communicative need for expressing the underlying concept 'rape / refugee'. The use of *rapeugee* rises considerably and re-

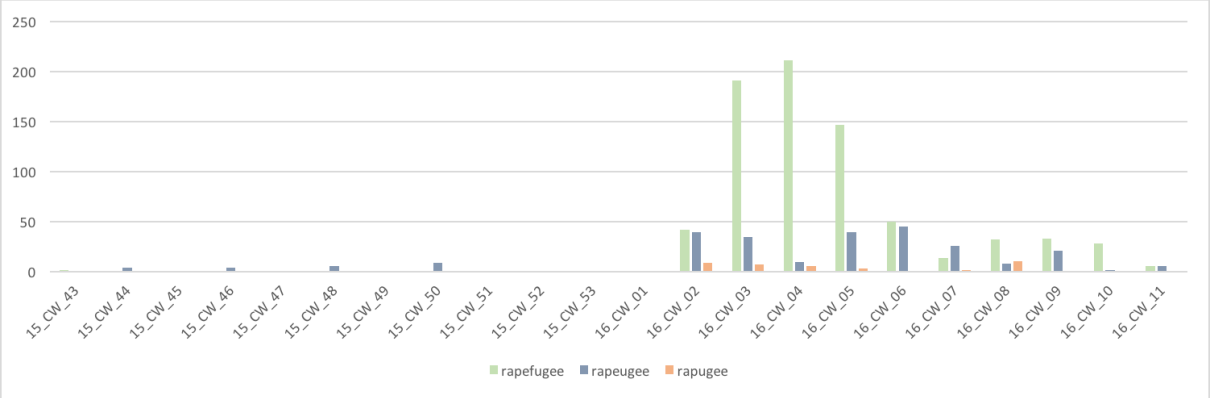

Figure 1: Absolute frequencies in the web corpus

mains fairly stable over the next few weeks. The form *rapugee*, which had up to this point been used only once, is used with moderate frequency until it vanishes again in 16_CW_09. Lastly, the form *rapefugee* shows the most radical increase by far. It reaches a maximum of 211 new tokens on 30 different websites in 16_CW_04. After New Year's Eve it represents an accumulated 73 % of tokens across all three types, making it the most dominant form in this period.

Figure 1 indicates that the spread of words expressing the concept 'rape / refugee' seems to happen in several spurts which do not follow a linear trend. Manual sample checks of the corpus data reveal that these spikes are closely related to real-life events in which refugees play an important role. Most often these events were various sexual harrassments, as we will exemplify further.

The first attestation of *rapeugee* we found is from a forum of an extremist propaganda website called *Shitskin Plantation*. On 29 October 2015, the user *canuckfmj* used the title *Denmark has a rapeugee problem* to publish the following post: *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.* The post contains a hyperlink to another extremist website which strongly criticises the introduction of sexual education in courses for refugees in Denmark. The use of the word *rapeugee* is clearly related to this particular recent political decision which serves as a trigger for coining the new term. The author expresses their critical attitude by questioning the adequacy of the neutral term *migrants* by using it in metalinguistic quotes. Instead, the author chooses the new term *rapeugee* to emphasize the propagated association between 'refugees' and 'rape'. In the following week, the new word seems to have already vanished again with the decreasing relevance of the real-life context, however, as we have not been able to find a single attestion of *rapeugee*. Similar patterns and connections to real-life events can be observed for the other spikes of *rapeugee* before New Year's Eve.

The turning point in the web corpus data is marked by the steep increase in the use of all three tokens after New Year's Eve and can be explained in the same manner. However this time, the variant *rapefugee* is preferred by most speakers. Its first attestation in 2016 is another blog post on a right-wing extremist blog named *Neoreactive*. A reader of the blog named Matt Bracken created a post entitled: *A Reader Says That The Cologne #Rapefugee Attacks Are Just A Pep Rally For The Coming Intifada In Europe*. Again, the author explicitly refers to the events in Cologne on New Year's Eve, when German media reported sexual assaults by refugees, and also instrumentalizes the blend of *rape* and *refugee* for anti-refugee propaganda.

The scale of the Cologne events and their presence in public media and in the Internet explain the explosive increase and the longer-lasting effect displayed in Figure 1. The numbers of new occurrences remain very high for a period of three weeks before the popularity of the three terms seems to run out of steam again after 16_CW_05.

The combination between such real-life triggers and the specific, quite uniform propaganda motivation of associating refugees with rape can be seen as the driving force behind the characteristic spurts in the usage intensity of the terms illustrated in Figure 1. These patterns are in line

with previous research by Kerremans (2015) who classified comparable cases as 'recurrent semi-conventionalization'.

Usage types. As pointed out in Section 3, besides measuring usage intensity as such, we examined different usage types of these words and their effects on the establishment process more closely.

Firstly, we investigated the tokens' position on the websites by counting tokens contained in titles separately. Across all three types, a high proportion of about 16 % of the tokens were used in the titles of websites. This fits the presumed motivation behind using the tokens as provocative propaganda terms in order to attract the readers' attention. We did not detect significant differences in usage frequencies regarding token position between the three types.

Secondly, we examined whether tokens were used in metalinguistic contexts. In these cases, speakers reflect/talk *about* the terms rather than just regularly using them. To identify these uses, we extracted quoted instances of all formal variants (i.e. *"rapefugee"*, *'rapugee'*). In total, about 7 % of the tokens were metalinguistic usages. On the one hand, we found that in most cases authors used inverted commas to distance themselves from the right-wing ideology behind the terms. For example, the website of the New York Post, an established conservative newspaper, published an article entitled *German clash over 'rapefugees' who carried out mass sex attack* (10 January 2016) in which they used the term *rapefugee* several times with a metalinguistic function. The article does not attack refugees, but the alarming growth of right-wing German extremists using the term for propaganda purposes. On the other hand, albeit in a much smaller number of cases, the terms are also sometimes used metalinguistically by anti-refugee activists who consciously try to spread them as propaganda terms. The results concerning metalinguistic uses indicate that they strongly differ from objectlinguistic uses and that they provide valuable information about the coinage and spread of neologisms.

5.2 Twitter corpus

Usage intensity. Figure 2 provides an overview of the Twitter data. In terms of usage intensity, the overall pattern is similar to that of the web corpus. The frequency of all three types remains relatively low before New Year, shows a steep increase in the first weeks of the new year and then declines to a lower level after that. However, there are also some differences.

First of all, there are no instances of *rapefugee* or *rapugee* before the New Year turn. This means that the dominance of *rapeugee* before New Year is even stronger in the Twitter data. There are only three weeks (15_CW_46 until 15_CW_48) that contain any tokens at all, and they only amount to a total of 15 tokens. Compared with the much higher usage intensity after the turn to 2016, this means an even steeper increase of use at the start of January than in the web corpus.

Secondly, the NY increase starts off earlier than in the web corpus. As a comparison of Figure 1 and Figure 2 shows, the turning point of usage intensity for all types on Twitter precedes that on the web by one week. This offset indicates that Twitter is the medium in which this change can be first observed. Being more flexible, social media are apparently faster in reacting to noteworthy events than web domains like blogs and forums.

The first tweet for *rapefugee* in 16_CW_01 in our dataset is *Refugee = rapist. Flüchtling = Vergewaltiger. #Cologne #rapefugees*, posted on Wednesday, 6 January 2016, and directly followed by its retweet. This tweet connects the neologism to the 2016 New Year's Eve sexual assaults in Cologne. Supposedly, these events were the trigger for the highly rapid boost in usage intensity for all three neologisms on Twitter. This is supported by the analysis of further tweets: The most frequent tweet for *rapefugee* in 16_CW_01 is *RT @DavidJo52951945: RT pictures from protest in Germany against immigrant/refugee abuse gangs #rapefugees https://t.co/USHsiXOtKZ*, which occurs 190 times during this week and also connects it to the sexual assaults in Germany.

The tweet *Where were the police water cannons when the Muslim rapeugees were terrorizing Cologne on NYE?!? https://t.co/dRcTMY9UJm*, retweeted twice, is the most frequent tweet for *rapeugee* during 16_CW_01 – also connected to the events in Cologne.

For *rapugee*, the two tweets during 16_CW_01: *@BBCBreaking @BBCWorld gangs of men??? Refugee men – say it: #rapugee https://t.co/AZK4fYLZLo* and a modified version of it, also relate it to these events.

The connection of the neologisms with the New

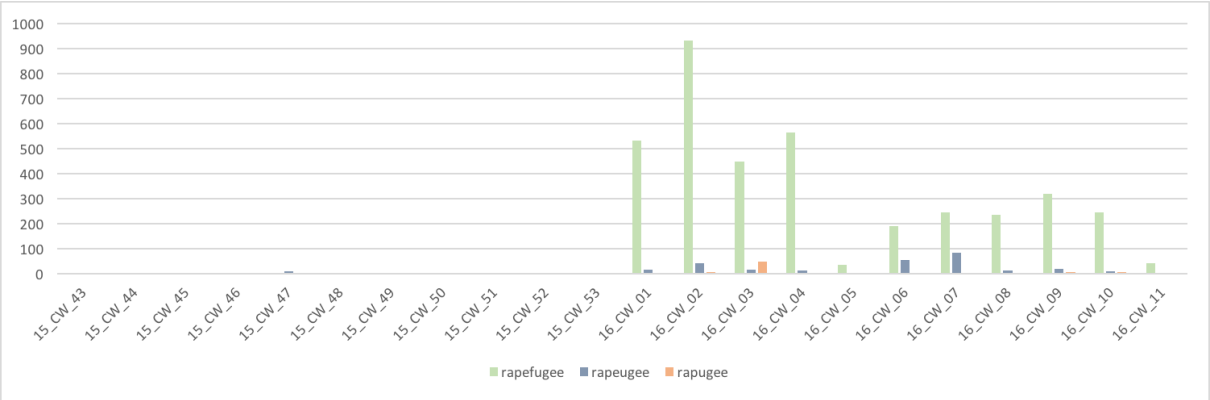

Figure 2: Absolute frequencies in the Twitter corpus

Year events and their respective usage intensity and relative success indicate that important real-life events play a significant role for the coining, rivalry and establishment of neologisms competing for occupying the same onomasiological space.

Usage types. With respect to usage types, a first distinction can be drawn between tweets and retweets. Retweets provide users with a very economical and efficient way of sharing tweets by other users with their own followers. As the original content is preserved and retweets are marked with the prefix *RT*, this can essentially be considered a quoting mechanism. The average number of retweets per tweets for all three forms is 2.7. This affects the establishment of words in at least two ways. On the one hand, it significantly increases the number of people reading the target words, which raises the chances that they will retweet or actively use it too. On the other hand, retweets are exact copies. So if the original author chooses the variant *rapefugee*, this choice is being replicated for all retweets. It is quite likely that these factors have contributed to the success of the form *rapefugee* on Twitter in the wake of New Year's Eve.

A second distinction can be drawn between hashtags and direct, i.e. normal uses of words. Hashtags are a second key feature of Twitter which has the potential to cause new effects on the pathways of the establishment of new words. Users can prefix words with # in order to turn them into labels. These labels build a fluctuating system tweeters use to refer to certain events or entities. Across all three types, we observed that 87 % of the tokens were used as hashtags. The very high proportion of tokens used as hashtags can be explained by their presumed communicative purpose. As was pointed out above, these terms mainly serve propaganda functions as they are used to label refugees as (potential) rapists. The establishment of a label like *#rapefugee* contributes to fixing the choice of the dominant variant.

5.3 Competition across both corpora

The composition and the sizes of the web corpus (about 950,000 words) and the Twitter corpus (about 85,000 words) differ greatly, which makes it hard to compare competition effects across both corpora. In order to measure the relative success of the three forms, we therefore normalized each type's frequency measures by the total frequency of all types within that dataset. The rationale behind this procedure is that the three forms lend themselves to encoding the same portion of semantic space and are thus in onomasiological competition. Even though the choice of individual language users may be determined by various factors such as whether they are familiar with all three terms, what they have heard or read just before (a priming effect possibly leading to the large numbers of retweets), or what they have become accustomed to (an entrenchment effect), this proportional measure is a good indicator of the relative success and spread of the three forms.

Figure 3 shows the relative counts for the web data where *rapeugee* appears to be the predominant type of choice between 15_CW_43 and 16_CW_02. 16_CW_02 marks the turning-point of the success of *rapefugee*. While *rapugee* still occurs following this period, there is a clear preference for the other two forms in the timespan from

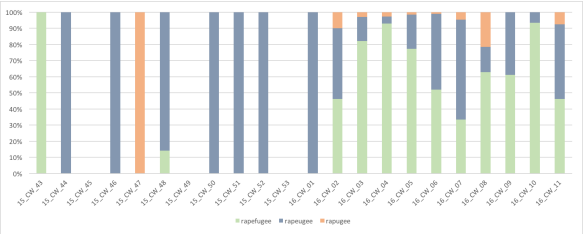

Figure 3: Relative frequencies in the web corpus

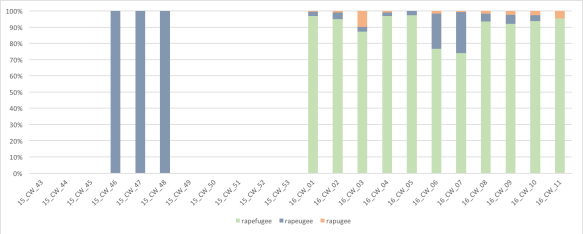

Figure 4: Relative frequencies in the Twitter corpus

16_CW_02 onwards, with an ongoing competition between them whose outcome does not seem to be determined at present.

In the Twitter data, which is visualized in Figure 4, the situation is considerably different. As mentioned above, the turning point in the relative success of the three types is one week before the one on the web, namely 16_CW_01. From this point onwards *rapefugee* is clearly the predominant choice although the other two types are also occasionally made use of.

Comparing the development in the web corpus to the Twitter data suggests that Twitter might have influenced the competition between the three competing forms in both domains decisively. Firstly, tweeters react to the events in Cologne on New Year's Eve more quickly than authors on the web. Secondly, the early establishment of the hashtag *#rapefugee* might have fuelled the increasing dominance of this formal variant. This is also supported by the fact that the type *rapefugee* often appears with the Twitter prefix # on the web in the early weeks of 2016, even though the hashtag does not serve any technical labelling function on the corresponding web pages. Thirdly, the high number of retweets seems to have supported the increasing dominance of the variant *rapefugee*. This is a particularly interesting finding, because it indicates that social media provide new ways of promoting the spread of new words.

What should be taken into consideration, however, is that all three of our target words are propaganda terms, whose users aim to spread their ideas and concepts. The people using these terms seem to belong to a like-minded community sharing the same communicative goals. This promotes the uniform use of the terms and the high number of retweets. Further research into less 'loaded' words will have to show whether the establishment process we observed is a special mechanism in the present case.

6 Conclusion

We have investigated the competition between three synonymous neologisms – *rapefugee*, *rapeugee* and *rapugee* – in a web and a Twitter corpus over a period of 22 weeks and found that the spread of the terms is closely related to preceding real-life events. Most importantly, the sexual assaults on New Year's Eve in Cologne lead to a steep increase in the use of these terms, mainly by right-wing extremists. Overall, the form *rapefugee* turned out to be the most likely candidate for establishment, although the final outcome remains uncertain at the present stage.

Analyzing data from the Twitter corpus allowed us to evaluate the web corpus' results more closely. We observed the same general development of the three neologisms in both datasets. Together with the language-external evidence of real-life events, this can be regarded as a cross-validation of both approaches. However, we also found that certain communicative practices within the Twitter domain, such as retweeting and hashtags, significantly influence the establishment of new words. Firstly, these mechanisms affected the competition between the three formal variants within the Twitter domain. It was presumably due to its high prominence in retweets and as a hashtag, that the variant *rapefugee* took the lead after New Year. Secondly, the Twitter domain seems to have influenced the use of the terms on the web. While the observed one-week offset could simply be due to the speed of social media, the use of hashtags on the web clearly suggests a causal explanation.

The results show that social media can be an important driving force in the coining of new words, and that social media corpora are thus an important data source for their detection and observation. Yet, the comparison of results between both datasets also shows that particular rules or conven-

tions on social media platforms like Twitter significantly shape the linguistic behaviour of users on that platform. Therefore, platform-specific features and mechanisms like retweeting and hashtags need to be taken into account to arrive at an adequate interpretation of results. A big advantage of using the web as a data source is its heterogeneity. It provides a much broader set of linguistic varieties, text types, authors and readers which makes it a much more representative sample. Platforms like Twitter might certainly often spark or react more quickly to the establishment of new words, yet their use on the heterogeneous and pervasive World Wide Web provides a more balanced indication for their eventual conventionalization.

7 Future work

As we have shown, differences between the linguistic behaviour of speakers on Twitter and on the web significantly influence the spread of neologisms in both domains. Given the heterogeneity of the Word Wide Web, it would be desirable to further classify different domains-of-discourse within the web corpus in order to observe how these sub-domains differ regarding the use of neologisms. For example, our case study indicates that the use of terms like *rapefugee* differs strongly between private domains like personal blogs and professional domains like newspaper websites. While the former seem to function as a driving force in the early spread of the term, the latter tend to use the term less frequently and more critically, which is also reflected in the increased proportion of metalinguistic uses.

For future work, automatic classifications of domains-of-discourse for the web should thus be implemented. When investigating a large set of neologisms, this would allow to monitor in which domains they first appear and whether and how their use extends to other domains-of-discourse. This promises very valuable information, as the diffusion of neologisms across several domains plays an important role in their conventionalization process.

References

Laurie Bauer and Antoinette Renouf. 2000. Contextual clues to word-meaning. *International Journal of Corpus Linguistics*, 5:231–258.

Jack Grieve, Andrea Nini, and Diansheng Guo. 2016. Analyzing lexical emergence in Modern American English online. *English Language and Linguistics*.

Peter Hohenhaus. 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.

Daphné Kerremans, Susanne Stegmayr, and Hans-Jörg Schmid. 2012. The Neocrawler: identifying and retrieving neologisms from the internet and monitoring ongoing change. In Kathryn Allan and Justyna A. Robinson, editors, *Current Methods in Historical Semantics*, pages 59–96. Berlin: de Gruyter Mouton.

Daphné Kerremans. 2015. *A Web of New Words: A Corpus-based Study of the Conventionalization Process of English Neologisms*. Frankfurt am Main: Peter Lang.

Lothar Lemnitzer. 2011. Making sense of nonce words. In Margrethe Heidemann Andersen and Jörgen Nörby Jensen, editors, *Sprognaevets Konferenceserie 1*, pages 7–18. Nye Ord. Kopenhagen.

Dirk Lewandowski. 2008. A three-year study on the freshness of Web search engine databases. *Journal of Information Science*, 34(6):817–831.

Piotr Paryzek. 2008. Comparison of selected methods for the retrieval of neologisms. *Investigationes Linguisicae*, 16:163–181.

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

Anatol Stefanowitsch and Susanne Flach. (forthcoming). The corpus-based perspective on entrenchment. In Hans-Jörg Schmid, editor, *Entrenchment and the Psychology of Language Learning: How We Reorganize and Adapt Linguistic Knowledge*. American Psychology Association and de Gruyter Mouton, Boston, USA.

Tony Veale and Cristina Butnariu. 2010. Harvesting and understanding on-line neologisms. *Cognitive perspectives on word formation*, pages 399–418.

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