



Lexical innovation on the web and social media.

Emergence, diffusion, and social variation in the use of English neologisms

INAUGURALDISSERTATION

zur Erlangung des Doktorgrades der Philosophie
an der Fakultät für Sprach- und Literaturwissenschaften
der Ludwig-Maximilians-Universität München

vorgelegt von

Quirin Würschinger

aus München

2023

Erstgutachter	Prof. Dr. Hans-Jörg Schmid
Zweitgutachterin	Prof. Dr. Stephanie Hackert
Drittgutachter	PD Dr. Wolfgang Schindler
Tag der mündlichen Prüfung	09.11.2022

für
Luisa

Acknowledgements

Herzlichen Dank an meinen Betreuer Hans-Jörg Schmid, der mir ermöglicht hat, dieses Dissertationsprojekt mit großem Freiraum durchzuführen und mir dabei immer mit kompetentem und effektivem Rat zur Seite gestanden hat, und auf dessen großes Verständnis, Geduld und Gelassenheit ich mich auch in schwierigen Zeiten immer verlassen konnte. Danke auch an Stephanie Hackert für die Übernahme meiner Zweitbetreuung sowie an Wolfgang Schindler für die Drittbegutachtung. Vielen Dank an meine KorrekturleserInnen Océane, Caro und Ricky; insbesondere an Océane, die mir beim Endspurt des Marathons über die Ziellinie geholfen hat. Größter Dank an meine Familie, Mama, Papa, Lisi und Kathi, und an meine Freunde, die mich in diesen herausfordernden letzten Jahren so sehr unterstützt haben. Und an Luisa, die für mich Motivation und Sinnbild war: Es gibt auch ein Leben neben und nach der Diss.

Contents

1	Introduction	9
1.1	Defining lexical innovations	11
1.2	Emergence	13
1.3	Diffusion	16
1.4	Outline	21
2	Emergence and diffusion on the web – The NeoCrawler	23
2.1	Research context	23
2.2	Using Data-Mining to Identify and Study Patterns in Lexical Innovation on the Web: The NeoCrawler	24
2.3	Conclusions	48
3	Diffusion on the web and social media	49
3.1	Research context	49
3.2	Using the Web and Social Media as Corpora for Monitoring the Spread of Neologisms	50
3.3	Conclusions	60
4	Social networks of diffusion	61
4.1	Research context	61
4.2	Social Networks of Lexical Innovation. Investigating the Social Dy- namics of Diffusion of Neologisms on Twitter	62
4.3	Conclusions	83
5	Semantic innovation of an old word – The case of <i>Anglo-Saxon</i>	85
5.1	Research context	85
5.2	Conclusions	86
6	Semantic innovation and social variation	89
6.1	Research context	89
6.2	Semantic Change and Socio-Semantic Variation. The Case of Covid- related Neologisms on Reddit	90
6.3	Conclusions	107
7	Conclusion	109
7.1	Research objectives	109
7.2	Emergence	110
7.3	Diffusion	112

0. Contents

7.4 Socio-semantic variation	113
Bibliography	117
Zusammenfassung	125
Appendix	131

which refers to the practice of ‘increase[ing] or enhanc[ing] the skills required in (a job)’². In addition, the list contains neologisms that have emerged specifically on the web such as *Internet of Things* or on Twitter, such as *twitterverse*.

From a formal perspective, the sample contains a large, representative selection of word classes and word-formation processes. Some neologisms result from morphemic word-formation processes like compounding (*deepfake*³), other terms such as *glamping*⁴ are the result of non-morphemic processes like blending, which merges the forms and concepts of *glamour* and *camping*.

More unusual examples include the term *covfefe*, coined by Donald Trump on May 31, 2017, in a tweet reading ‘Despite the constant negative press covfefe’. Even though it is most likely simply a misspelling of *coverage*, its meaning has never been officially clarified, sparking controversy over the term’s usage. Despite its unclear meaning, or perhaps because of it, the term went viral, with Trump’s followers using it extensively and primarily as a hashtag on Twitter to promote his policies. According to Trump himself, it is not the result of a misspelling of *coverage*; thus it would be a rare morphological case of an ex-nihilo formation (Bauer 1983: 239; Algeo 1998: 66).

Moreover, the neologisms in the sample show significant differences in terms of how successfully they have spread and managed to catch on. Some neologisms such as *hyperlocal*⁵ have enjoyed a relatively long lifespan and seem to be here to stay. Other terms like *covidiot*⁶ were coined only very recently, and their future is far more uncertain. During their lifetime, some terms such as *upcycling*⁷ exhibit stable use over time, while others show strong fluctuations in the form of sporadic, topical spikes. This is the case, for example, with the term *Brexit*, which was used most often before and after the withdrawal of the United Kingdom from the European Union, but has since experienced multiple surges whenever the topic has been more prominent in public discourse. In some cases, these spikes occur at regular, re-current intervals. This applies to *poppygate*, for example, which refers to scandals surrounding the wearing of poppy symbols for Remembrance Day in the United Kingdom and is consequently almost exclusively used in November.

Lastly, the neologisms in my sample vary greatly in how conventional they have

²upskill, v. OED Online. June 2022. Oxford University Press. (accessed June 10, 2022).

³an image or recording that has been convincingly altered and manipulated to misrepresent someone as doing or saying something that was not actually done or said’ (*deepfake*, n. Merriam-Webster.com. Merriam Webster. (accessed June 10, 2022)).

⁴A form of camping that involves accommodation and facilities more luxurious than those associated with traditional camping.’ (*glamping*, n. OED Online. Oxford University Press. (accessed June 10, 2022)).

⁵Extremely or excessively local. In later use spec.: designating (the output of) a television channel, website, or other media outlet focusing on matters of interest within a small geographical area.’ (*hyperlocal*, adj. OED Online. Oxford University Press. (accessed June 10, 2022)).

⁶someone who behaves in a stupid way that risks spreading the infectious disease Covid-19’ (*covidiot*, n. Cambridge Dictionary. Cambridge University Press. (accessed June 10, 2022)).

⁷To reuse (waste material) to create a product of higher quality or value than the original, and to reduce the need for new raw materials in production’ (*upcycle*, v. OED Online. Oxford University Press. (accessed June 10, 2022)).

become in the speech community at large. On the one hand, many terms failed to gain traction and fell into disuse soon after their coinage (e.g. *microflat*). On the other hand, some terms like *coworking* have attained a high degree of conventionality; they are used in a broad range of usage contexts and are familiar to large portions of the speech community. Other neologisms have attained a certain degree of conventionality, but their use remains limited to specific usage contexts, as in the example of *detweet*⁸, whose use is restricted to the social media platform Twitter – the ‘twitterverse’. Others, such as *alt-left*, are confined to particular communities of speakers, in this case on the basis of their political views. The term *alt-left* was coined in opposition to the term *alt-right* in 2015. It was introduced as a counterpart to the shortened form of *Alternative Right*, which serves as an umbrella term for far-right, white nationalist groups in the United States. Due to its background, it is ideologically loaded, and its use remains limited to a small community of like-minded individuals on the extreme right of the political spectrum.

The preceding examples provide an initial overview of the data covered in this dissertation. They present prototypical cases that illustrate the variety of lexical innovations. Because of this diversity, previous empirical studies of neologisms have tended to focus on a single type of neologisms or on their usage in specific contexts. In this dissertation, I aim to provide a more comprehensive account of lexical innovation by studying the emergence and diffusion of a diverse set of neologisms. To address this diversity, I collect a large sample of both formal and semantic neologisms, investigate their use in a variety of data sources obtained from the web, Twitter, and Reddit, and employ various methods to analyse their diffusion, such as frequency-based analyses, social network analysis, and distributional semantics.

Following the preceding introductory tour of my sample, the subsequent sections will disentangle the various aspects of lexical innovation represented in my sample and present the theoretical and methodological basis of my dissertation.

1.1 Defining lexical innovations

All of the lexical innovations in my sample are ‘lexical units that have been manifested in use and thus are no longer nonce-formations, but have not yet occurred frequently and are not widespread enough in a given period to have become part and parcel of the lexicon of the speech community and the majority of its members’ (Kerremans 2015: 31). I adhere to this definition of neologisms by Kerremans, which underlines two crucial characteristics of lexical innovation that are central to this dissertation.

Firstly, I regard neologisms as innovative ‘lexical units’, which includes both formal and semantic neologisms. Previous work has offered several terms and definitions for neologisms. Among others, the terms ‘novel lexical items’ (Leech 1981: 30; Lipka, Handl & Falkner 2004: 10) and ‘new lexeme’ (Bauer 1983: 63) have been used most widely. Theoretical accounts of lexical innovation draw a clear distinction between

⁸The term *detweet* has been used with several competing meanings; previous work has found that its most widespread meaning is ‘to delete a tweet (after posting)’ (Kerremans 2015: 81–83)

formal neologisms, which are characterised by the emergence of a new word (often paired with a new meaning), and semantic neologisms, which are distinguished by the emergence of a novel meaning for an existing word (Leech 1981: 30; Bauer 1983: 55). This distinction is reflected in previous empirical investigations, which have studied these two types of lexical innovation separately, with most studies focusing exclusively on formal neologisms.

Defining neologisms as novel ‘lexical units’ emphasizes the fact that lexical innovation can pertain to both sides of the linguistic sign: depending on changing communicative needs, speakers either add new forms to the lexicon to refer to name-worthy concepts (Schmid 2008) or use existing words in novel senses and meanings. Formal and semantic neologisms share the trait of being novel lexical form-meaning pairings (Tournier 1985; Geeraerts 2010) that are unknown to the majority of the speech community and are competing to become a part of the conventional lexical inventory. In this dissertation, Chapters 2 to 4 examine formal neologisms, whereas Chapters 5 and 6 focus on semantic neology.

Secondly, I approach the study of lexical innovation from a usage-based perspective. As indicated in the preceding definition by Kerremans (2015: 31), I consider neologisms to be lexical units whose usage indicates that they have been ‘manifested in use’, but that they have not yet fully diffused and become conventional parts ‘of the lexicon of the speech community and the majority of its members’. While many previous accounts have primarily focused on structural aspects (Leech 1981: 30; Bauer 1983: 48; Plag 2003: 52; Lipka, Handl & Falkner 2004: 10), Schmid adopts a usage-based approach to lexical innovation and differentiates between structural, socio-pragmatic, and cognitive perspectives on the establishment of neologisms (Schmid 2008: Ch. 2, 2016: Ch. 4). The structural view focuses on the ‘internal structure of the word itself’ (Schmid 2016: 71), its formal and semantic properties, as well as its syntagmatic and paradigmatic relations in the language system. The cognitive perspective explores how new words are processed and stored in the minds of individual speakers. This includes questions about the extent to which neologisms are entrenched in the mental lexicon and the conceptual status of lexical innovations. This dissertation focuses mostly on Schmid’s second, socio-pragmatic perspective. This perspective concentrates on the use and conventionality of neologisms in social interaction and examines new words in terms of their emergence and ‘the extent of [their] spread and diffusion, i.e. [their] degree of use and familiarity among the members of the speech community.’ (Schmid 2016: 71).

Of course, cognitive and socio-pragmatic processes interact with structural aspects of language use. A more detailed account of these interactions is provided by the Entrenchment and Conventionalization Model (Schmid 2015, 2020), whose implications for the diffusion of lexical innovations will be discussed below.

However, from all three perspectives, neologisms are characterised as new lexical units which have been used beyond their first emergence, but have not (yet) become established parts of the lexical inventory. In this dissertation, I study the pathways of neologisms from their initial emergence to their diffusion in the speech community.

1.2 Emergence

1.2.1 Defining emergence

Lexical innovations emerge when speakers coin new words to refer to new products (e.g. *blockchain*) or public concerns (e.g. *Covid*), or when established words acquire new meanings, due to cultural changes in society (*(social) distancing*), for example.

I use the term *emergence* for the initial appearance of new lexical units in the form of formal or semantic innovations since the related terms ‘creation’ (Schmid 2016: 82) and ‘coining’ (Ayto 2006: 1; Schmid 2020: 330) are typically reserved for formal neologisms. While formal neologisms are often consciously formed and introduced by the linguistic action of individual speakers, novel senses and meanings of words are typically the result of indeterminate, gradual shifts in the use of existing words, which appears to be more accurately captured by the term *emergence*. Notably, the term *emergence* is also used by Grieve, Nini & Guo (2016) and Grieve (2018a), but the authors do not draw a distinction between the appearance and subsequent diffusion of neologisms, which will be treated separately here.

In the previous section, neologisms were defined as lexical units that are new and have been ‘manifested in use’ (Kerremans 2015: 31) but have surpassed the state of ad-hoc formations, which are coined and used in a specific communicative act and fail to spread beyond this original usage context (Hohenhaus 1996: 38, 2005: 17). In order to study the emergence of neologisms, it is thus necessary to identify neologisms as close to their coining and first uses as possible. Methodologically, this presents fundamentally different challenges for the study of formal and semantic neology, as will be discussed in the following sections.

1.2.2 New data for new words

The emergence of formal neologisms can be tracked by identifying the appearance of new orthographic or phonetic forms in language use. Lacking the data sources and computational methods required for automated detection, earlier attempts had to rely on the expertise of lexicographers and lexicologists in identifying new words. Similarly, earlier linguistic studies on neology were restricted to studying small, manually selected samples of neologisms (e.g. Hohenhaus 2006), or to relying on domain experts and specialised dictionaries to investigate the emergence of neologisms in specific text types or semantic fields (Elsen 2004; Foubert & Lemmens 2018).

In recent years, the advent of new data sources and computational methods has opened up new possibilities for large-scale, data-driven approaches to detecting lexical innovations. Earlier linguistic corpora such as the British National Corpus (BNC Consortium 2007) and the Corpus of Contemporary American English (Davies 2008) were too small in size to capture recent neologisms since neologisms are, by definition, infrequent. The increasing availability of big web corpora like the News on the Web Corpus (Davies 2016) and the TenTen Corpus family (Jakubíček et al. 2013) has significantly extended the opportunities for large-scale corpus analyses of the emergence of

lexical innovations.

In addition to general-purpose web corpora, several research groups have used the web as a corpus to build dedicated tools and specialised corpora for detecting the emergence and monitoring the spread of formal neologisms such as WebCorp (Renouf, Kehoe & Banerjee 2007), Wortwarte (Lemnitzer 2010), NeoCrawler (Kerremans, Stegmayr & Schmid 2012), Logoscope (Gérard et al. 2017), and Neoveille (Cartier 2017). I will provide more details about tools for studying neologisms on the web in Section 1.3. For this dissertation, I have used an extended version of the NeoCrawler (Kerremans et al. 2018) to investigate the emergence of formal neologisms on the web; Chapter 2 will provide a more detailed account of its architecture and the results gathered from this approach.

More recently, social media data have emerged as an increasingly important alternative to web corpora. The use of language in social media is informal and inventive. It most rapidly reflects cultural and linguistic trends and changes. These characteristics make it a hotbed for lexical innovation. Social media datasets from platforms like Twitter or Reddit surpass the size of existing web corpora, and they provide metadata about usage contexts (e.g. timestamps, like counts), about speakers (e.g. follower counts), and about interactional patterns (e.g. replies, retweets), which enable more reliable and fine-grained analyses of language use in communicative interaction.

These advantages have sparked a growing body of research investigating the emergence of formal neologisms on Twitter (e.g. Eisenstein et al. 2014; Grieve, Nini & Guo 2016; Grieve 2018b; Nini et al. 2017) and Reddit (e.g. Stewart & Eisenstein 2018; Del Tredici & Fernández 2018).

1.2.3 Formal neologisms

Based on web and social media data, previous work has used two main methods for identifying formal neologisms. Firstly, emerging neologisms can be identified using dictionary matching. In short, this approach compiles a dictionary of all word forms contained in the corpus that may include neologisms and compares the resulting list to a reference dictionary containing a comprehensive list of established lexemes. All terms matched by the reference dictionary are excluded, leaving a list of potential neologisms. This approach, as implemented in an extended version of the NeoCrawler (Kerremans et al. 2018), was used to compile the majority of the neologism sample studied in this dissertation. A more detailed account of this methodology and its results will be presented in Chapter 2.

Secondly, emerging neologisms can be identified by determining all words whose usage frequency has increased considerably over a certain time period. This method requires a diachronic corpus with relative frequency counts for each word in the corpus. Several previous studies have successfully utilised frequency-based approaches for detecting formal neologisms, albeit with slight variations (Lapata & Lascarides 2003; Cabré Castellví & Nazar 2012; Grieve, Nini & Guo 2016). Most recently, Grieve, Nini & Guo (2016) identified a set of 131 potential neologisms in a large Twitter dataset. After grouping the data into temporal bins (e.g. using monthly intervals), they used

Spearman's rank correlation coefficient (Spearman 1961) to identify all words in the dataset that exhibit statistically significant increases in usage frequency and were therefore presumed to have emerged as neologisms during the specified time period.

In this dissertation, I study neologisms determined by both methods to arrive at a broad, data-driven sample of recent English neologisms. The majority of formal neologisms were identified using the dictionary matching method implemented in the NeoCrawler. In contrast to the statistical method, which provides a retrospective analysis of the diffusion of words with higher usage frequency, this approach allows for investigating the incipient diffusion of recent neologisms and includes cases of incomplete and unsuccessful diffusion, thus providing a more complete picture of diffusion.

Additionally, I have included neologisms determined by the statistical method presented in Grieve, Nini & Guo (2016). This complements the sample with additional cases of neologisms that have spread successfully. The statistical method is able to provide full coverage for the diffusion pathways of these candidates going back to their first use, which is impossible for the dictionary-based detection of neologisms on the web.

1.2.4 Semantic neologisms

Unlike formal neologisms, semantic neologisms have received little attention from previous work. This can largely be attributed to the increased methodological complexity for corpus-based studies of semantic neology. Semantic neologisms are, by definition, characterised by conventional forms and thus cannot be identified based on the presence or absence of word tokens in the corpus alone. Instead, the detection of words with new senses and meanings requires computational semantic representations of words before one can assess whether lexical meanings have changed. While previous corpus-linguistic methods such as collocational analysis have allowed some insights into co-occurrence patterns and semantic profiles of words (e.g. Hilpert & Correia Saavedra 2017; Schmid et al. 2020), these methods lack the reliability and robustness required for detecting semantic neologisms.

Recent advances in Natural Language Processing offer new methods for the data-driven generation of semantic representations that can be adopted for large-scale studies of the emergence of semantic neologisms. Research in Computational Linguistics has demonstrated that word embeddings (Mikolov et al. 2013) can successfully model the meaning of lexemes based on their distributional properties (Firth 1957). Going beyond collocation analysis and methods based on co-occurrence counts (Hilpert & Correia Saavedra 2017; Baroni, Dinu & Kruszewski 2014), word embedding algorithms such as word2vec (Mikolov et al. 2013) leverage machine learning to acquire semantic representations by predicting context words.

With notable exceptions, most applications of word embedding methods have remained within the field of Natural Language Processing. In the domain of semantic change detection, word embeddings have been successfully utilised for studying processes of long-term, gradual meaning change in corpora covering decades and centuries of historical English, such as the Corpus of Historical American English (e.g. Kim et al.

2014; Hamilton, Leskovec & Jurafsky 2016; Kutuzov et al. 2018). More recent advances have enabled increasingly fine-grained investigations of short-term lexical semantic change on the scale of years rather than decades (Del Tredici, Fernández & Boleda 2019; Shoemark et al. 2019; Tsakalidis et al. 2019). However, most previous work in Computational Linguistics has focused primarily on methodological advances, with a rapidly growing interest in optimising algorithms for semantic change detection in Natural Language Processing competitions like the annual SemEval tasks (e.g. Schlechtweg et al. 2020). There has been little focus on semantic neologisms, on the linguistic features observed by computational models of meaning, and on the characteristics of the underlying process of lexical emergence.

Within the linguistics community, word embedding approaches have only very recently started to be adopted for studying linguistic research questions. In a recent review article, Stevenson & Merlo (2022) presents a linguistic perspective on the theoretical basis, potential, and current limitations of employing word embeddings for linguistic research questions. Previous studies have shown promising results in domains such as syntax (e.g. Hilpert & Flach 2021), morphology (e.g. Shafaei-Bajestan et al. 2022), and semantic change (e.g. Fonteyn & Manjavacas Arevalo 2021). However, previous work has not focused on using word embeddings for the linguistic study of the emergence and diffusion of semantic neologisms.

In Chapter 5, I conduct a case study of semantic innovation based on the term *Anglo-Saxon* by using collocation analysis to analyse changes in its meaning. Chapter 6 provides a more comprehensive study of semantic neologisms employing word embeddings to obtain a more accurate picture of semantic change and socio-semantic variation. This chapter also gives a more detailed account of the methodological aspects of word embeddings.

1.3 Diffusion

1.3.1 Modelling and measuring diffusion

Neologisms have, by definition, survived their period of initial uses and have managed to spread across the speech community at least to some degree. By contrast, ad-hoc formations are coined by speakers in specific usage contexts and their use remains restricted to these initial contexts (Hohenhaus 1996: 38; Fischer 1998: 3; Hohenhaus 2005: 17; Kerremans 2015: 30). Neologisms can thus be viewed as ‘transient phenomena’ (Schmid 2008) on the continuum between ad-hoc formations and established words, and diffusion can be seen as the process that facilitates their spread to new usage contexts and new parts of the speech community.

The diffusion of neologisms shares features with the diffusion of cultural innovations, since the spread of new words is often closely associated with the spread of cultural products and practices. Models of cultural innovation (Rogers 1962) propose an S-shaped trajectory of diffusion, whereby the adoption of an innovation in the population is slow during initial stages, increases to higher rates during intermediate stages, and

slows down again before full adoption is achieved. In linguistics, the S-curve model has been successfully applied in studies of the diffusion of linguistic innovation and change in multiple domains of language (Milroy 1992; Labov 2007). There has been substantial evidence supporting the empirical adequacy of the S-curve model; however, previous work has mostly focused on diffusion in the areas of phonology and syntax (Blythe & Croft 2012; Nevalainen 2015)

The Entrenchment-and-Conventionalization Model (Schmid 2015, 2020) conceptualises the conventionalization of linguistic innovations as involving two processes: usualization and diffusion. Diffusion is defined as the 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 conformity regarding the *types of cotexts and contexts* in which they use it’ (Schmid 2020: 178–179, emphasis mine). In the case of lexical innovation, neologisms start out as new form-meaning pairings. When a neologism is used for the first time by an individual speaker, interlocutors need to ‘make sense’ of the new word in the given usage context for the utterance act to succeed (‘co-semiosis’ (Schmid 2020: Ch. 3.1)). The same word can then proliferate beyond this initial usage context, and can be adopted and used by the interlocutors (‘co-adaptation’ (Schmid 2020: Ch. 3.2)). At this stage, the novel formation has surpassed the status of an ad-hoc formation (Hohenhaus 1996) and successful diffusion leads to its use in new usage contexts and by new speakers.

The process of ‘usualization’ (Schmid 2020: Ch. 9) serves to establish, sustain, and adapt the conventional use of new words. A new word has to be used repeatedly in the same types of usage contexts and by the same speakers and communities until individual speakers store it as a lexical unit in their mental lexicon (‘entrenchment’ (Schmid 2020: Pt. 3)), and until it becomes a conventional lexical unit in a community of speakers.

As described in Section 1.2.2, the growing availability of large-scale web and social media corpora in recent years have enabled a growing number of large-scale empirical investigations on the diffusion of neologisms. Besides a modest number of studies based on dictionaries (e.g. Elsen 2004) and questionnaires (e.g. Link 2021), most related work in recent years has been grounded on data obtained from the World Wide Web. The approach of using the web as a corpus (Gatto 2014) has led to several research projects developing specialized tools for studying the diffusion of neologisms on the web. Among the earliest efforts in this field, the tool WebCorp (Renouf, Kehoe & Banerjee 2007) attempted to provide a search engine for the web designed for linguists, with a special focus on monitoring new words. Later approaches providing more involved technical features include the tools Logoscope (Gérard et al. 2017) and Neoveille (Cartier 2018), which are specifically focused on neologisms and offer elaborate interfaces with visualisations and descriptive statistics about the use of the neologisms in the corpus. Both of these methods compile their corpora based on newspaper articles, which enables them to provide large corpora, along with precise metadata in the form of timestamps and annotated sources. However, this restriction to newspaper articles comes at the expense of a more limited perspective on diffusion, since this approach

fails to capture neologisms emerging in other domains of less formal language use such as discussion forums or blogs. Moreover, the use of neologisms in such specialised corpora provides only a limited indication of the overall degree of diffusion outside the domain of newspapers.

Coming back to the above definition of diffusion by Schmid (Schmid 2020: 178–179), I will disentangle two dimensions of diffusion to provide a more fine-grained view on the degrees of diffusion of lexical innovations.

1. To what degree have neologisms spread across multiple *types of usage contexts*?
2. To what degree have neologisms spread across multiple *speakers and communities*?

1.3.2 Diffusion across usage contexts

The NeoCrawler project (Kerremans, Stegmayr & Schmid 2012) was initiated to gain a more comprehensive view of the diffusion of neologisms and to investigate factors involved in this process. Using a web-as-corpus approach, it aimed to monitor the spread of neologisms in a large corpus of authentic language use, encompassing a broad range of text types and authors. Consequently, the NeoCrawler compiles its corpus by searching the entire open web, based on Google's search index, which represents most speakers' view of the web.

Utilising this corpus, Kerremans (2015) studied the diffusion of 44 neologisms on the web. Using qualitative and quantitative indicators of diffusion, Kerremans established a continuum of four stages of diffusion based on her sample of neologisms (Kerremans 2015: Ch. 4.1). While many new words remain nonce-formations or quickly fall into disuse ('non-conventionalization'), others manage to gain traction to some degree and are used with sporadic ('topicality') or recurring ('recurrent semi-conventionalization') spikes in usage frequency. A small number of neologisms manage to spread with considerable success, showing 'advanced conventionalization'. Kerremans' results indicate that in addition to usage frequency as an indicator of diffusion, the use of neologisms across a variety of text types plays an important role for the assessment of degrees of diffusion, with words that exhibit advanced conventionalization typically being used in a more diverse set of usage contexts.

Expanding upon the work of Kerremans (2015), Chapter 2 presents an extension of the NeoCrawler approach. It enhances the NeoCrawler's ability to detect new words by incorporating social media data sourced from Twitter and employing partial string matching to detect neologisms. Through periodic Discoverer searches, these advancements enabled me to identify a significantly larger sample of 958 neologisms, a substantial increase compared to the previous study by Kerremans (2015), which was based on 40 neologisms. This larger sample allowed me to examine the diffusion of a more representative set of lexical innovations over a longer time period.

To study diffusion beyond usage contexts on the web, Chapter 3 extends the study of neologisms based on the NeoCrawler using additional data from the social media

platform Twitter. This serves to evaluate the results obtained from the NeoCrawler and investigate the extent to which the use of neologisms is tied to usage contexts on the web or social media.

1.3.3 Diffusion across speakers and communities

The second dimension of diffusion entails the spread of new words across speakers and communities. The study of social networks has a long history of research in sociolinguistics, and the social dynamics involved in the adoption of linguistic innovations are central to well-established models of diffusion like the S-curve model (Milroy & Milroy 1985; Labov 2007; Nevalainen 2015). It is commonly assumed, for example, that successful spread depends on the social status of early adopters, dense networks that promote diffusion in the early stages, and the presence of weak ties to disseminate innovations to larger portions of the speech community (Granovetter 1973).

However, it has long been impossible to study the spread of innovations across speakers and communities directly based on corpus data. While the availability of web corpora has facilitated large-scale, longitudinal studies, web data lacks precise information about how many speakers have used and adopted a specific neologism or whether its use remains limited to specific communities or extends to larger proportions of the speech community. Consequently, previous studies had to rely on usage frequency as an indicator for the degrees of social diffusion of neologisms. The underlying assumption is that neologisms that have been used many times in the corpus are likely to be familiar to a large group of speakers who have actively produced the observed uses ('corpus-as-output') or have been passively exposed to these neologisms ('corpus-as-input') (Stefanowitsch & Flach 2017). Aggregating all instances of usage to total frequency counts is taken to represent the total amount of exposure or active usage, indicating the degree of diffusion of a given neologism in the speech community at large.

The increasing availability of social media datasets have enabled more differentiated studies on the social diffusion of neologisms. Social media data from platforms like Twitter or Reddit provides data about how frequently a specific neologism has been used on the platform, along with metadata about usage contexts such as timestamps and like and retweet counts. Additionally, social media data include information about which individual speakers have used the term, as well as metadata about the location of users and the communities to which they belong. These additional data facilitate more detailed studies of social diffusion than earlier approaches, which had to rely on frequency measures alone.

Recent work has begun to exploit these new opportunities and has used social media data for studying the social dynamics of the diffusion of neologisms. By collecting a large dataset from Twitter, for example, Grieve, Nini & Guo (2016) were able to identify a set of 131 emerging formal neologisms and to track their use over a period of 14 months. In a follow-up study, Nini et al. (2017) analysed the temporal dynamics of diffusion based on the same dataset, confirming an S-curve trajectory in the case of successful spread of neologisms. Attempting to measure factors influencing these

trajectories, Grieve (2018b) found that new words spread more successfully when they are shorter, when they are formed using ‘standard word-formation processes’ (rather than acronyms and spelling variations), and when they mark a new meaning. In Chapter 3, I collect and analyse data from Twitter to extend the study of diffusion on the web based on the NeoCrawler with an investigation of neologisms on social media. I find that Twitter data regarding the spread of selected neologisms is largely consistent with web data on their usage. However, the results indicate that the emergence and diffusion of neologisms on Twitter precede and influence their spread on the web, highlighting the importance of social media for the dissemination of new words on the web.

Recently, an increasing number of studies have utilised social media data for social network analyses to get an even more detailed view of the social dynamics of diffusion. Data from platforms like Twitter offer metadata about speakers which can be used to capture all speakers and their interactions in the dataset in the form of a network graph. Based on these social graphs, one can extract additional information about the speakers (e.g. their social influence) and their relationships (e.g. whether they show strong or weak ties), and about the communities (e.g. size, density) involved in the observed language use.

So far, social network analyses have mainly been applied outside of linguistics. Network science approaches to social media data have been successfully employed in diverse fields, for example, to study the spread of diseases (Lu et al. 2018), public opinions (West et al. 2014), and political attitudes (Pew Research Center 2019). Given its ability to analyse social dynamics in communicative interaction, social network analysis also shows great potential for studying the social diffusion of neologisms. Despite this potential, social network analyses have only rarely been applied to lexical innovation.

One of few examples in previous literature, a recent study by Goel et al. (2016) uses network analyses for studying the spread of selected abbreviations, phonetic spellings, and lexical words on Twitter. They find that language change can be viewed as a form of social influence and that the diffusion of innovations generally follows patterns of complex contagion, i.e. adoption is more likely when speakers are exposed to an innovation multiple times. However, their results suggest that the transmission of new words follows simple contagion and is promoted by the existence of strong ties between speakers. Their approach complements earlier studies based on synthetic data obtained from simulations (e.g. Blythe & Croft 2012) and provides a very detailed view of the social dynamics of diffusion based on authentic language use. However, the results only allow for limited generalizations since they are based on a relatively small sample of only 19 lexical innovations. Moreover, the study covers a time window of only 12 months, which restricts the analysis to short-term trends. Aside from these limitations, Goel et al. (2016) convincingly show the potential of network analysis based on social media data for a detailed view of the social diffusion of neologisms.

In Chapter 4, I extend this approach by conducting social network analyses of the use of neologisms on Twitter. I study the spread of a bigger sample of neologisms identified

by the NeoCrawler over an extended period and employ network metrics in addition to frequency data to get a more detailed view of social diffusion across speakers and communities. The results suggest that the use of neologisms shows considerable social variation. While certain neologisms exhibit high overall usage frequency, their use differs strongly between communities.

Building on this observation, Chapters 5 and 6 delve into socio-semantic differences in the use of neologisms in social networks. Previous research on semantic neology has been mostly limited to studying semantic change over longer timeframes, without access to data about the social dynamics involved in the observed changes. Utilising social media data from Twitter and Reddit, I explore the extent to which the meaning of words differs between communities and whether variation and change in the use and meaning of neologisms correlate with the social dynamics in the social network of speakers.

1.4 Outline

This dissertation explores the emergence and diffusion of neologisms on the web and on social media, with a special focus on the social dynamics of diffusion and variation between communities of speakers. To this end, I collect a large sample of formal and semantic neologisms, and investigate their use on the web, on Twitter, and on Reddit using analyses based on usage frequency, social networks, and word embeddings.

Chapters 2 to 6 each present one paper along with an introductory and concluding section that situate these studies within the context of my dissertation project.

Chapter 2 studies the emergence and diffusion of formal neologisms on the web. It presents an extended version of the NeoCrawler, employed to identify neologisms in a data-driven fashion, and to compile a large monitor corpus that captures their spread over time. This approach yields a large sample of recent neologisms, which serves as the basis for the studies in the following chapters. A frequency-based analysis of the data indicates that the detected neologisms cover a broad spectrum of diffusion on the web.

Chapter 3 continues to use the NeoCrawler, but also studies the spread of selected neologisms on Twitter. This serves to evaluate whether the results obtained from the NeoCrawler extend beyond its web corpus and whether the selected neologisms show differences in use across web and social media contexts. The results indicate a high degree of convergence across these two usage contexts, but imply that the target words' diffusion gains traction on Twitter first, and that their use on social media, for instance in the form of hashtags, influences their spread on the web. Moreover, the findings indicate that usage frequency on the web presents an inaccurate picture of the degrees of diffusion of the selected cases, as their use seems to be largely limited to a small community of speakers.

Chapter 4 expands on these findings to provide a more detailed view of the social dynamics of diffusion on Twitter. Utilising a bigger sample of neologisms, it studies their spread from the launch of Twitter in 2006, a substantially longer time period than

was previously possible with the NeoCrawler. This study complements the frequency-based approach used in the preceding chapters with measures that capture the temporal dynamics of spread. Additionally, it employs social network analysis to assess the extent to which neologisms have spread across speakers and communities, and the role of social network features in diffusion. The results demonstrate considerable overlap between the frequency-based and network-based analysis of diffusion. However, the network analysis offers a more nuanced picture of the underlying social dynamics and reveals considerable social variation in the use of the observed neologisms.

Chapter 5 extends the study of variation in the use of formal neologisms by investigating differences in the use and meaning of a semantic neologism across communities. It examines the case of the term *Anglo-Saxon*, whose meaning and use have been highly contested recently in academia and on social media. The established, neutral uses of this term have been challenged by its use in an ‘ethno-racial’ sense, which has increasingly strengthened associations between the word *Anglo-Saxon* and the notion of white supremacy promoted by far-right communities. This study applies a social network approach to a longitudinal dataset obtained from Twitter, but goes beyond differences in usage intensity to study semantic variation based on collocation analysis. The results indicate that there is considerable semantic variation between communities depending on their socio-cultural background (e.g. US vs UK) and political views (e.g. far-right vs liberal).

Chapter 6 examines the emergence and socio-semantic variation of a larger set of semantic neologisms on Reddit. In the context of the Covid pandemic, it employs word embeddings for a data-driven identification of semantic neologisms and investigates meaning differences in the use of the detected terms between communities. This study demonstrates that word embeddings can be used effectively to detect recent semantic neologisms. Moreover, it reveals considerable socio-semantic variation between communities, which can be traced back, again, to diverging attitudes in polarised groups. A detailed investigation of the semantic variation reveals systematic differences along two dimensions. Communities with critical views about mainstream positions and public policies regarding the pandemic have more negative and more subjective associations with Covid-related terms compared to neutral communities. These results indicate the effects of polarisation on social media and the prevalence of social variation in the use of neologisms.

Chapter 7 encapsulates the dissertation’s main findings, discusses their theoretical and methodological implications, and proposes avenues for future research.

2 Emergence and diffusion on the web – The NeoCrawler

2.1 Research context

This first chapter investigates the emergence and diffusion of formal neologisms on the web. It presents an extended version of the NeoCrawler (Kerremans, Stegmayr & Schmid 2012), which is used to detect neologisms in a data-driven way and to compile a large monitor corpus that captures their spread over time.

The NeoCrawler is also the background and starting point of my dissertation. This research started with the DFG project ‘Incipient diffusion of lexical innovations’, which funded my work as a PhD student between 2016 and 2018.¹ The project was situated at the Chair of Modern English Linguistics at LMU Munich, and included Hans-Jörg Schmid (Principal Investigator), Daphné Kerremans (Postdoc Researcher), Fazleh Elahi (Computational Linguist), and Jelena Prokić (Computational Linguist), who all contributed to the project and its results, presented in the following section (2.2).

The paper *Using Data-Mining to Identify and Study Patterns in Lexical Innovation on the Web: The NeoCrawler*, which I co-authored with Daphné Kerremans (DK), Jelena Prokić (JP), and Hans-Jörg Schmid (HJS), presents the extended version of the NeoCrawler that was developed during this research project.² This paper was published in the journal *Pragmatics and Cognition* in 2018.

The extended version builds on an earlier version of the NeoCrawler developed by Kerremans, Stegmayr & Schmid (2012), which was used for a previous investigation of lexical innovation on the web by Kerremans (2015). The NeoCrawler’s objective has not changed: it attempts to discover formal neologisms and monitors their spread on the web. However, the paper presented in the following section (2.2) contributes several extensions.

Crucially, substantial improvements were made in neologism identification. As detailed in the paper, the enhanced version of the NeoCrawler enables partial string matching based on Levenshtein Distance (Levenshtein 1965), which was implemented by JK with assistance from the Center for Information and Language Processing (LMU). Partial matching made it possible to increase the proportion of high-quality candidates of formal neologisms by permitting the adjustment of the precision and recall of the dictionary matching method, which determines the number of false positive and false

¹Grant number: SCHM 1232/5-1

²For better readability, author contributions will be indicated with author initials, and references to sections, figures, and tables in the included papers will be marked with an asterisk (e.g. Section 1.1*).

negative candidates. I tested and used this method during weekly scans for new neologisms. In addition, I extended the detection of neologisms on the web with additional social media data obtained from Twitter. To this end, throughout the project, I collected random samples of tweets using TAGS (Hawksey 2020) in weekly intervals, and used the Discoverer module of the NeoCrawler to add the identified neologisms from Twitter to the database.

Using periodic Discoverer searches, these advances allowed me to identify a significantly larger sample of neologisms than was previously possible. At the time of writing the paper, the NeoCrawler database contained 958 neologisms, a substantial increase compared to the preceding study by Kerremans (2015), which was based on 40 neologisms.

Together with DK, I annotated this extended sample for word classes and word-formation processes. I showed that the resulting distributions observed for both formal categories are consistent with earlier research and data obtained from the OED (Section 3.2*). Monitoring the diffusion of this larger sample over an extended period of time using the Observer module of the NeoCrawler, I found that it captures a broad spectrum of diffusion, as measured by cumulative usage frequency (Figure 5*).

The following section (2.2) presents the architecture and results of the NeoCrawler, which was used to compile the database of recent neologisms that serves as the basis for the investigations in the following chapters. DK led the writing process and wrote the draft of this paper; together with JP and HJS, I contributed to the final version of the manuscript by providing additions, revisions, and comments.

Due to copyright restrictions, the following section contains the final version of the submitted manuscript of the published article:

Kerremans, Daphné, Jelena Prokić, Quirin Würschinger & Hans-Jörg Schmid. 2018. Using data-mining to identify and study patterns in lexical innovation on the web: The NeoCrawler. *Pragmatics and Cognition* 25(1). 174–200. <https://www.jbe-platform.com/content/journals/10.1075/pc.00006.ker>.

2.2 Using Data-Mining to Identify and Study Patterns in Lexical Innovation on the Web: The NeoCrawler

Using data-mining to identify and study patterns in lexical innovation on the web: The NeoCrawler

Daphné Kerremans, Jelena Prokić, Quirin Würschinger, and Hans-Jörg Schmid
Ludwig-Maximilians-Universität München

Abstract

This paper presents the NeoCrawler – a tailor-made webcrawler, which identifies and retrieves neologisms from the Internet and systematically monitors the use of detected neologisms on the web by means of weekly searches. It enables researchers to use the web as a corpus in order to investigate the dynamics of lexical innovation on a large-scale and systematic basis. The NeoCrawler represents an innovative web-mining tool which opens up new opportunities for linguists to tackle a number of unresolved and under-researched issues in the field of lexical innovation. This paper presents the design as well as the most important characteristics of two modules, the Discoverer and the Observer, with regard to the usage-based study of lexical innovation and diffusion.

Keywords: lexical innovation, neologisms, webcrawler, data-mining, string matching, innovation identification

1. Introduction

Lexical innovation is a pervasive phenomenon inherent to language and society. Unlike other kinds of linguistic innovation, e.g. affecting the phoneme inventory, innovation in the lexicon is very often motivated by extralinguistic concerns as a result of the predominant naming function of words. New products, experiences and phenomena require new names or semantic extensions of established ones. Lexical innovation therefore mostly involves a pairing of a novel form and a novel meaning or the pairing of an existing form and a novel sense, which derives from already conventionalized meanings. Less frequent are new formations in which an existing meaning is paired with a novel form; their creation is often motivated by pragmatic or social concerns in the speech community (humour, slang, jargon, etc.). Novel formations, whether semantic, morphological or morphosemantic (cf. Tournier 1985), which have been evidenced in language to some degree but have not become conventionalized yet are known as *neologisms* (cf. Fischer 1998, Schmid 2016).

Research into lexical innovation so far has primarily focused on the formal level of word-formation processes underlying neologisms and their productivity (cf. Plag 1999, Baayen & Neijt 1997). Much less empirical work has been done on the pathways on which lexical innovations enter the language and the speech community and spread through these, i.e. their diffusion

process (cf. Nevalainen 2000, Nevalainen & Raumolin-Brunberg 2003, Säily et al. this volume). This apparent lack of interest in the dynamics of diffusion of lexical innovation is particularly striking in view of the extensive, ground-breaking research into the diffusion of phonological innovation (cf. Labov 1980, 2001, Milroy & Milroy 1985, Tagliamonte & Denis 2014). A possible explanation may be methodological in nature. Whereas the use of certain novel phones and phonemes can be easily elicited by means of word lists, reading passages and scripted interviews (see the famous department store study by Labov 1966), triggering the use of a particular neologism is almost impossible. In addition, lexical innovation is less systematic than phonological innovation. Phonological innovations happen in a situation where speakers can select a variant from a restricted pool of possibilities. In contrast, in lexical innovation, language users seem to have almost infinite linguistic means at their disposal when it comes to labelling innovations or discussing a new topic.

As a consequence, the only feasible way to study the diffusion of lexical innovation is by means of corpora which represent a balanced cross-section of a given language for a given period. Since neologisms are infrequent linguistic units by definition and often confined to certain speaker groups or genres, such corpora should fulfil at least three requirements. They should be large, both in size and diversity of the varieties contained, to enable researchers to find as many occurrences of a neologism in as many linguistic, situational and social contexts as possible. The corpus should be recent, in order to study the incipient diffusion of lexical innovations before a given neologism is already fully conventionalized or has disappeared between its first occurrence in the corpus and the public availability of the corpus data. This condition is particularly important for observing the diffusion of fashion words that tend to become almost instantaneously established in the lexicon or quickly disappear into oblivion once their relevance and salience has diminished. Lastly, such a corpus should be dynamic to enable researchers to keep up with the pace of diffusion and gain fine-grained longitudinal data on the behaviour of neologisms, which is required to study the mechanisms and potential patterns of lexical innovations and their diffusion in particular.

This paper describes a corpus-based approach to investigating the diffusion of lexical innovations. When we started to form an interest in the study of neologisms and their diffusion almost a decade ago, a corpus fulfilling the three requirements listed above – size, recency, and dynamicity – in a satisfactory way was not in sight. The only option available was to use the Internet as such as corpus. Therefore, pursuing the web-as-corpus method, we developed a data-mining tool capable of identifying and closely monitoring the diffusion of lexical innovations on the web, the NeoCrawler. Our choice for the Web was motivated by the sheer dimensions of the data in terms of size and diversity on the one hand, and by its increasing importance as a communication mode in the speech community, also proving to be a very fertile breeding ground for creativity in the lexicon.

The web-based investigation requires three methodological steps: first, the identification of neologisms; second, the acquisition of data on the diffusion of neologisms; and third, the interpretation of this data regarding the factors influencing diffusion. In order to meet these requirements, the NeoCrawler is equipped with two modules. The first is an identification module, the Discoverer, which is able to semi-automatically detect new English forms as neologism candidates in online data (see Section 2). These candidates are stored in a database, currently including 958 recent neologisms. The second module, the Observer, performs weekly searches to extract any new occurrences of these 958 neologisms (see Section 3). It stores the

web pages these new words occur on together with contextual information, which can be used to investigate the factors influencing diffusion. With the help of the NeoCrawler important advances have been made regarding the application of state-of-the-art digital methods in lexicology, which in turn allow researchers to overcome previously existing hurdles and investigate the dynamics of lexical innovation on a large-scale and systematic basis.

2. The Discoverer

The Discoverer represents the innovation detection module of the NeoCrawler. It follows a form-based procedure, which operates with a string matching process to identify potential instances of lexical innovation (see Kerremans & Prokić 2018 for a detailed discussion). Novel instances of semantic change cannot be identified because they do not involve innovation of the formal pole of the linguistic sign but affect the semantic pole only. Recent advances in word embedding approaches currently seem to be the most promising technology to trace patterns of semantic change (Hamilton et al. 2016, Liao & Cheng 2016, Jatowt & Duh 2014, Cartier 2019). The remaining three possible sources of lexical innovation discussed above are unproblematic in form-based string matching since they all make use of morphological material, regardless of whether this consists of morphemes, non-morphemic splinters or foreign language forms in the case of borrowings.

String matching is a technique by means of which a string of entities in a source text, in our case graphemes, is compared against another string of entities in a reference text in order to gauge the degree of similarity. If the two strings are identical, a match will take place; if they are different, no match will take place. In the present case, all the words found in a selection of very recent websites are string-matched against a database listing existing words (see Section 2.1 and 2.3 below). Therefore, a positive match means that the source grapheme string is a known English word, whereas a non-match identifies the source string as a potential neologism. String matching in the Discoverer therefore involves four main components: (1) one or more source texts to be searched for neologisms, (2) a string matching algorithm, (3) a reference text or corpus, and (4) a final evaluation step in which the output is manually assessed in terms of neologism status. These four elements will now be described in detail. Due to the current operationalization of *word* as a gapless sequence of graphemes, only concatenated and hyphenated compounds can be identified. Options to include non-concatenated compounds are currently being explored and tested.

2.1. Source material and pre-processing

Our central goal is to investigate lexical innovation in the speech community at large. Therefore, when we look for new words on the Web, we aim at querying a diverse range of sources from different genres and topics rather than focussing on one specific genre such as newspapers (see Cartier 2017, Falk et al. 2018), blogs (see Megerdooian & Hadjarian 2010 for Persian blogs), specific web domains (see Liu et al. 2013 for the Taiwanese PTT forum) or academic papers (see Torres-del-Rey & Nava 2014). Moreover, the Discoverer uses the actual online webpages in their entirety rather than RSS feeds as in the case of Néoveille and the Logoscope or traditional corpora (see Cabré & de Yzaguirre 1995 for Catalan and the multilingual NeoTrack by Jansen 2005). Since

lexical innovation can occur in all kinds of linguistic materials, the NeoCrawler's general scope surpasses any genre- or topic-specific patterns and produces a more varied and representative output. The main difference to these comparable automatic identification tools and approaches therefore concerns the Discoverer's search scope and the nature of its textual input.

The Discoverer is equipped with three different input modalities. The first is to select specific target URLs to be queried. Since the selection of the input determines the nature of the output, the URL collection needs to be carefully balanced for genres and topics or fields. The URL selection currently consists of three categories:

- (1) British and American newspapers and news websites, e.g. The Guardian, The Daily Telegraph, The Sun, The Washington Post, The New York Times, Huffington Post
- (2) Popular blogs from lifestyle, technology, science, entertainment, business and politics, e.g. *Lifehacker*, *Wired*, *Mashable*, *Gizmodo*, *ScienceBlogs*, *TechCrunch*, *Buzzfeed*
- (3) British and American online dictionaries and lexicographic resources, e.g. *Urban Dictionary*, *Oxford Dictionaries Blog*, *Merriam Webster*

Secondly, local files from the user's local disk can be searched. This is particularly helpful if the texts stem from offline sources that have been digitalized or have been provided as offline collections in digital form. We also employ this option to scan large batches of Twitter data so as to have an exhaustive social media source in the selection. Thirdly, the user can also choose to access texts from the server in case this may prove necessary, e.g. to search particular pages, online domains or social media platforms in the NeoCrawler database.

The Discoverer does not use clean, text-only samples such as standard corpora or RSS feeds in the approaches mentioned above, but webpages in their entirety, i.e. complete with links, embedded videos, pictures, HTML code and other unsearchable or (linguistically) irrelevant material. The pages thus need to be converted to a clean text-only format that can be used in the following string matching process. The Discoverer performs several pre-processing tasks as listed in the process report in Figure 1. After checking the UTF-8 encoding of the page – in case the page is not in UTF-8 a warning will appear that errors might occur during the cleaning and identification procedure – the page will be tokenized, i.e. split into single words, which are also called *unigrams*. Unigrams are the standard token format in English; Chinese on the other hand rather uses bigrams because Chinese characters typically consist of two characters (see Liu et al. 2013). All contained tokens will be extracted and listed together with their frequency. The user can opt to see both the stripped text and the frequency list by ticking the box.


```

Preprocessing Text #4
Encoding is UTF-8, attempting full punctuation stripping method
Filtering out stopwords and strings with length < 3 and length > 64 Done.
Filtering out known names... Done.
Filtering out known tokens from Dictionary... Done.
Filtering out known tokens from previous program runs... Done.
Constructing Candidate objects... Done.

 Show stripped Text of URL "telegraph.co.uk", text number 4

 Show Frequency list of URL "telegraph.co.uk", text number 4

Adding unknown types to List...

Preprocessing Text #5
Encoding is UTF-8, attempting full punctuation stripping method
Filtering out stopwords and strings with length < 3 and length > 64 Done.
Filtering out known names... Done.
Filtering out known tokens from Dictionary... Done.
Filtering out known tokens from previous program runs... Done.
Constructing Candidate objects... Done.

 Show stripped Text of URL "nytimes.com", text number 5

 Show Frequency list of URL "nytimes.com", text number 5

No tokens remained after filtering.

```

Figure 1: String matching process report in the Discoverer.

Other semi-automatic identification methods include further pre-processing steps such as POS-tagging (see Iakovleva 2017, Megerdumian & Hadianian 2010) and/or spelling checking (see Iakovleva 2017). In order to keep the procedure fast and efficient we have opted not to include further pre-processing measures at this stage. As our data shows, the steps implemented so far produce output of a quality high enough not to require further cleaning and filtering. Once the texts have been selected, stripped and tokenized the Discoverer immediately performs the string matching procedure, which will be explained in detail in the next section.

2.2. String matching procedure

String matching is a form-based procedure operating with the degree of identity or comparability of two strings or n-grams. Since we are interested in detecting lexical innovations, these strings are lexical unigrams, i.e. uninterrupted English grapheme sequences corresponding to single tokens. By means of a matching algorithm all tokens of the source text, which consists of a collection of URLs and Twitter data in our case (see Section 2.1), are compared against the unigrams in a reference text, which consists of one or more dictionaries that are compiled into a reference corpus. In the present context string matching is therefore often called *dictionary matching* (see also “Exclusion Dictionary Architecture” as labelled by Cartier 2017). If the source token is found in the reference dictionary, a match will take place and the token will count as an existing English word. If no match arises between the source token and a token from the dictionary, the source token will be identified as an unknown English grapheme sequence and will represent a potential neologism. The degree of ‘neologism-ness’ thus derives from an

unsuccessful formal match between tokens from different texts. In more technical terms, the Discoverer uses a string library from the *Center for Information and Language Processing* (CIS) at LMU Munich, which compiles the reference texts into a dictionary in the required form. It represents a finite state automaton capable of very fast string matching in huge files such as the Wikipedia dump we use (see section 2.3).

Prior to running the string matching procedure, all tokens starting with a capital are removed because they are in most cases proper names and are of no further concern for the present purpose. Although this will exclude acronyms from the list of potential neologisms, this simplification was introduced to minimize the noise from non-standard spellings, nicknames and other typographic creativity frequently encountered in computer-mediated communication. Moreover, tokens consisting of a combination of digits and graphemes such as *18-year-old* or *400-ounce* are also removed. Previously the Discoverer did include tokens containing the digits 2, 4 and 8 because of their frequent substitutions for *to/too* (*2morrow*), *for* (*4real*) and *-ate* (*18*), but due to the low improvement of the output recall at the cost of reduced precision all digits are currently excluded.

A further complication that needs to be addressed concerns the effect of non-standard, creative spelling and spelling errors. Since both types can be intended by the coiner when launching a neologism, the Discoverer presently contains the option to make strict string matching more flexible by varying the degree of comparability. In default matching, a match is typically successful, and no neologism candidate is found, if one grapheme is different between the source and reference token. For instance, a search of fourhourworkweek.com/blog on August 1, 2018 produced the candidates *uncommit*, *tequila-fueled*, *safety-certified*, *drunk-dialing*, *reidhoffman*, *jasonfried*, *spressfield* and *bchesky*. Often, this type of exact matching produces numerous candidates, but it is also prone to the rampant inclusion of false positives, precisely because online language is by definition messy when orthography is concerned and creative with regard to the lexicon. In order to reduce the number of false positives and restrict the output to genuine candidates as much as possible in an automated approach, we currently rely on elastic searches, which involves the manipulation of the so-called *string distance*, which determines the precision of the match. The string distance can be measured with the Levenshtein distance (Levenshtein 1965). The Levenshtein distance represents the smallest number of operations (including insertions, deletions and substitutions), which is needed to transform two non-matching strings into matching ones. When comparing *Brexit* to *Grexit* the Levenshtein distance is one because one operation is required to change *Brexit* into *Grexit*, i.e. *B* to *G*. If the Levenshtein distance is specified as zero in the interface, exact string matching will take place and *Brexit* will not be matched against *Grexit* in the reference dictionary, because there is a difference of one: *Brexit* will be suggested as a potential candidate. If the Levenshtein distance is increased to one, all tokens that have a string distance of 1 or less when compared to *Grexit* (e.g. *Grixit*, *Grexis*, *Brexit*) will be identified as identical and will not appear as candidates. As a consequence, simple misspellings and other products of formal creativity will be considered to be existing English words and the output will not be cluttered with such genuine or potential false positives. In the example from fourhourworkweek.com/blog the last two user names, *spressfield* and *bchesky*, were not identified as potential neologisms and did not appear in the final output list once the string distance was increased to 1.

The extent to which varying the Levenshtein distance affects the output is further illustrated in Figures 2a and 2b. Both figures present the results of a daily neologism identification run of

The Sun conducted on May 23, 2018. Figure 2a contains the candidate list produced after exact matching procedure, i.e. with a Levenshtein distance of 0, whereas Figure 2b shows the results for the same text with a Levenshtein distance of 1. As the two figures show, recall is lower for an increased string distance. Only two strings, *gin-flavoured* and *unfollowing*, are suggested as candidates in Figure 2b. In contrast, the exact matching produces five more lexemes: *coolbox*, *non-VIPs*, *chef-level*, *li-ttle* and *mega-mansion*. Save for *li-ttle*, none of these can be reasonably classified as misspellings or typos and to be directly excluded as candidates. *Coolbox* represents the concatenated spelling variant of *cool box*, which is recorded as a special usage in the OED entry for *cool* (cf. OED online, s.v. *cool*, adj, adv., int.). Similarly, *unfollowing* does not present a genuine neologism either, since it merely is the present participle of *unfollow*, which is already included in the NeoCrawler database. Rather than classifying them as candidates in the database they will be marked as ‘missing in the dictionary’ and added to the reference corpus so as to exclude them from potential further matches. As a consequence, of the seven terms listed in Fig. 2 a *gin-flavoured*, *non-VIPs*, *chef-level* and *mega-mansion* are evaluated as candidates and stored in the NeoCrawler database for future crawling. Of these, only *gin-flavoured* would have been discovered in a less constrained search with a Levenshtein distance of 1.

Show unknown types in Text #7:

1. : <i>coolbox</i>	<input type="radio"/> Candidate	<input type="radio"/> Garbage	<input type="radio"/> Ignore	Type Quality: 0	Avg. Trigram Freq 0%
2. : <i>gin-flavoured</i>	<input type="radio"/> Candidate	<input type="radio"/> Garbage	<input type="radio"/> Ignore	Type Quality: 0	Avg. Trigram Freq 0%
3. : <i>non-VIPs</i>	<input type="radio"/> Candidate	<input type="radio"/> Garbage	<input type="radio"/> Ignore	Type Quality: 0	Avg. Trigram Freq 0%
4. : <i>chef-level</i>	<input type="radio"/> Candidate	<input type="radio"/> Garbage	<input type="radio"/> Ignore	Type Quality: 0	Avg. Trigram Freq 0%
5. : <i>li-ttle</i>	<input type="radio"/> Candidate	<input type="radio"/> Garbage	<input type="radio"/> Ignore	Type Quality: 0	Avg. Trigram Freq 0%
6. : <i>mega-mansion</i>	<input type="radio"/> Candidate	<input type="radio"/> Garbage	<input type="radio"/> Ignore	Type Quality: 0	Avg. Trigram Freq 0%
7. : <i>unfollowing</i>	<input type="radio"/> Candidate	<input type="radio"/> Garbage	<input type="radio"/> Ignore	Type Quality: 0	Avg. Trigram Freq 0%

Figure 2a: Discoverer output of *The Sun* search 23 May 2018 with Levenshtein distance of 0.

<input checked="" type="checkbox"/> Show unknown types in Text #7:					
1. : gin-flavoured	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Type Quality: 0	Avg. Trigram Freq 0%
	Candidate	Garbage	Ignore		
2. : unfollowing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Type Quality: 0	Avg. Trigram Freq 0%
	Candidate	Garbage	Ignore		

Figure 2b: Discoverer output of The Sun search 23 May 2018 with Levenshtein distance of 1.

These examples aptly show that a higher Levenshtein distance reduces the output and number of potential candidates because more matches between source text and reference corpus have taken place, perhaps also unwarrantedly. Although the number of possible false positives deriving from typos, misspellings and other causes can be decreased, it may also go at the cost of not detecting genuine neologisms. In many cases, the setting of the string distance depends on the users' preferences and the nature of the source texts: there is no universal measure that scores 'best' across the board.

2.3. Reference dictionary

The quality of the output, or the efficiency of the algorithm-driven identification procedure, not only depends on the cleanliness of the input or the use of different string distance measures, but also on the quality of the reference dictionary.¹ In the current framework dedicated to lexical innovation, such a dictionary is required to be large-sized, diverse in composition and up-to-date. These three conditions will guarantee an effortless and successful automatic identification procedure. The overall size will reduce the number of false positives in general. Diversity is required to go beyond the limits of standard English and include socio-pragmatic variation such as dialectal instances or technical language. Finally, keeping the reference dictionary up-to-date will minimize the number of false positives with regard to conventionalization in the speech community and exclude older words that have remained rare.

The Discoverer's dictionary is compiled from three sources. The core is formed by a 13 gigabyte English Wikipedia dump from November 2016. This dump contains all words from the English Wikipedia version up until November 2016 and meets the three conditions mentioned in the previous paragraph. It ensures through its sheer size and up-to-dateness that the reference sample to be matched against is recent and exhaustive both regarding the number and topic-related diversity of lexemes. As a consequence, the likeliness that the system suggests established words as candidates for neology is kept to a minimum. In order to guarantee recency of the reference material such a dump needs to be updated at regular intervals. In addition, the matching procedure also relies on user feedback. During the manual evaluation step (see below 2.4) the user can mark false positives, upon which they will be fed into the reference dictionary and excluded from future searches. The Discoverer architecture consequently enables users to

¹ We use a full forms dictionary. This makes lemmatization redundant, which is an important advantage since lemmatization is occasionally prone to mistakes and inconsistencies.

interact with the dictionary to continuously expand it so as to ensure its optimal composition and performance.

In order to maximize the speed and efficiency of the identification procedure, several special types of tokens are filtered out during the actual matching and contribute to the dictionary currently used. These types are also illustrated in Figure 1 above. First, stop words² are removed. Stop words are highly frequent (function) words in a given language, which rarely contribute relevant information to the NLP task at hand but considerably slow down automatic processing. By default, strings shorter than three and longer than 64 graphemes are also deleted. Shorter string can however be included by selecting such an option in the search interface. Next, named entities, or known proper names, are filtered out by means of a list which has been provided by the CIS at LMU Munich.³ These named entities are also efficiently covered by the choice of Wikipedia as our reference dictionary, which contains an exhaustive and up-to-date list of named entities. Because of the noisiness of web data with regard to our current research purposes such a multi-level and extensible reference dictionary is required to conduct a high-quality matching procedure which requires as little manual post-processing as possible. This final step will be briefly discussed now.

2.4. Manual evaluation

As shown so far, most of the neologism identification procedure in the Discoverer consists of an automatized sequence of steps with minimal user input. The final assessment of the output, i.e. the decision whether a candidate is to be fed into the NeoCrawler database for further observation, however, is entirely manual. After the texts have been processed, the interface provides the user with an overview of the matches found, as presented in Figure 3. The first column contains the candidate in bold, *non-OpenStack*. The next three columns are decision-making markers: the lexeme can be identified as a candidate, as garbage, or can be ignored. In the latter case, a final decision will be postponed and the candidate will re-appear in future searches of the same page/domain as long as the lexeme is included on the input page. Lexemes marked as 'garbage' will be added to the reference dictionary. Types of garbage that can be specified are 'orthographic error', 'garbled string' (nonsensical grapheme sequences), or as known lexemes (including proper names) that are missing in the dictionary. Since such a decision is rather difficult without the context, a further column lists the candidate in its immediate cotext with the option to expand it.



Figure 3: Discoverer output interface from techcrunch.com, May 22, 2018.

² <https://code.google.com/archive/p/stop-words/>.

³ This list was provided by our colleague Michaela Geierhos from the Center for Information and Language Processing (CIS) at the LMU Munich (2006).

In a prior version of the Discoverer,⁴ two additional measures were included to inform the decision: the type quality and the average trigram frequency, as also shown in the figure. The average trigram frequency provides an estimate of the probability that the identified form is indeed a valid English word in use. The trigram frequency is calculated on the basis of a corpus which is composed of the *King James Bible*, the *Brown Corpus* and a three-month period of *The New York Times*. First, the potential candidate is split into trigrams, e.g. *bib*, *ibl*, *ble* for *bible*, which are subsequently checked against the trigrams in the corpus. The resulting score is the average of all trigram frequencies in this corpus. The type quality, on the other hand, offers a measure of neologism fitness, or quality of neologismness. By default, all candidates are considered to be equally good candidates and receive a score of 10. By means of a penalty system based on formal linguistic properties that the candidate may exhibit, points are detracted. For instance, *Q&As* will be scored lower than *uncommit* from the search above because it contains non-grapheme characters (&) and a capital letter in non-initial position. The same would apply to *non-OpenStack* in Figure 3.

Once a potential candidate is marked accordingly, a first rudimentary word-formation classification can be performed via a drop-down menu. This menu contains the standard morphemic and non-morphemic labels 'compound', 'derivation', 'blend', 'conversion', 'clipping', 'back-formation', 'abbreviation', 'epo-/toponym' 'ex-nihilo' and 'phrase'. Unclear cases are identified as 'unclassified', ambiguous ones involving several word-formation processes from a synchronic perspective as 'multiple'. The final result of such a manual evaluation for the output from guardian.co.uk is given in Figure 4.⁵

⁴ Due to technical difficulties these options are not operational at the moment but will be re-included in the future.

⁵ For reasons of space the cotext is not included in the figure.

List of unknown types

Display Explanation:

List of all encountered unknown types:

Show unknown types in Text #3:

1. : <i>overtourism</i>	<input checked="" type="radio"/> Candidate	<input type="radio"/> Garbage	<input type="radio"/> Ignore	Type Quality: 0	Avg. Trigram Freq 0%
	of type: Compound				
2. : <i>the-lives-of-grenfell-tower</i>	<input type="radio"/> Candidate	<input checked="" type="radio"/> Garbage	<input type="radio"/> Ignore	Type Quality: 0	Avg. Trigram Freq 0%
		of type: Proper name			
3. : <i>pre-Weimar</i>	<input type="radio"/> Candidate	<input type="radio"/> Garbage	<input checked="" type="radio"/> Ignore	Type Quality: 0	Avg. Trigram Freq 0%
4. : <i>live-plucking</i>	<input checked="" type="radio"/> Candidate	<input type="radio"/> Garbage	<input type="radio"/> Ignore	Type Quality: 0	Avg. Trigram Freq 0%
	of type: Compound				

Figure 4: Manual evaluation results for a search output of guardian.co.uk on May 22, 2018

Among the candidates produced by the Discoverer, *overtourism* and *live-plucking* were assessed as potential neologisms to be further observed. Both were classified as compounds, although it should be noted that *overtourism* straddles the boundary to prefixation. *Overtourism* refers to the phenomenon that many places in the world attract an ever-growing number of tourists and experience inconveniences as a result. *Live-plucking* arose in the present case in the context of fashion. In order to meet the demands of fashion designers using feathers in their designs, apparently animals were abused to the extent that their feathers were plucked rather than collected after having naturally fallen off. In contrast, the phrase *the-lives-of-grenfell-tower* was assigned to the garbage category as a technical error since the original page does not contain the phrase in its hyphenated form as shown by the Discoverer but as a regular title with spaces between the words. Finally, *pre-Weimar* was moved into the 'ignore' list awaiting final judgement. In its extracted cotext no unambiguous clues for its neologism status were found.

After having been marked as candidates, *overtourism* and *live-plucking* were added to the NeoCrawler database, from which they are retrieved by the second module, the Observer, to be queried on the Internet during its weekly crawling round. The Observer module will be presented in more detail in the next chapter.

3. The Observer

The Observer is a custom-tailored web-mining tool designed (1) to search the web for new occurrences of selected neologisms in regular intervals and (2) store these hits in the NeoCrawler database. We use a relational database, which will be presented in 3.2. (see Kerremans et al. 2012

for technical details) and which can be accessed via the interface as described in 3.2. First, the basic architecture will be discussed.

3.1. Architecture of the Observer

The Observer first accesses the NeoCrawler database to retrieve the stored neologisms and searches the entire surface web, i.e. those parts that are commercially accessible as indexed by standard search engines with the Google Custom Search API, for new occurrences of each neologisms. Thus, only those occurrences which newly appear on the Internet between two search rounds are downloaded by the Observer. Currently, these searches are performed weekly. This short time interval not only allows us to observe small increments in the potential incipient diffusion process, but also to find as many new occurrences as possible. Since we rely on Google's Custom Search API to access the web (see below), the number of returned pages per query is capped at 100. Thus, stretching the crawling interval to one month or even longer would increase the risk of obtaining biased quantitative and qualitative results: a neologism may be much more conventionalized than its frequency in the database reveals since no pages exceeding 100 hits will be returned, regardless of the time depth. The first search for each newly added item is performed retroactively as much as possible by not only including data from the past week but from January 1, 1970 up until present. The choice for this data follows logically from our use of Unix time, which handles all time-related information in the database and which is part of the Unix Epoch starting at the said date.

As noted above, for lack of a non-commercial linguistic search engine the Observer uses the Google's Custom Search API to systematically obtain large-scale data from the web. The present choice for Google is motivated by the quality of its index: both in terms of size and freshness, both properties being crucial for studying incipient diffusion of neologisms, the Google index is unrivalled (cf. Lewandowski 2008, Wilson 2017). Besides, it provides a consistent search index which enables comparative longitudinal analyses. However, in the near future other search engines will be included so as to ease existing restrictions imposed by Google algorithms and commerce-driven policies, and increase our autonomy. For the time being, queries are restricted to the English language as identified by Google's language identification tools in the Google API, but principally the Observer's architecture can be extended to other languages. Manual evaluation has revealed a substantial degree of inaccuracy in Google's language identification. Therefore, we will incorporate an independent identification tool double-checking Google's return and filtering out any non-English pages.

After the results of the queries are returned, the pages found and their URLs are stored in the database. Together with their original html version, a txt file is also saved. However, as web data are very messy, extensive post-processing is required to display the data in the web interface in an accessible and user-friendly way and prepare it for subsequent linguistic analysis. For now, the first step consists of removing all duplicates, i.e. pages (near-)identical in content, and false positives (cf. Kerremans et al. 2012). Next, boilerplate removal takes place: all linguistically irrelevant material (for the present purpose) is deleted, which includes photos, videos, lists, html tags and script code. The original page nevertheless remains available in the database. The title of the document is automatically extracted too. Finally, the pages are tokenized: the entire text is split into words (and sentences) and the neologism tokens are identified and counted. The tokens are extracted together with their extended cotext of 500 characters so as to obtain immediate

information about their meaning and usage and produce searchable concordance lines for text analysis tools. The next step will be to implement POS tagging and incorporate a topic and text classifier, which automatically performs semantic and sociopragmatic linguistic analysis. The result of these post-processing features is a human- and machine-readable, relatively clean sample of new occurrences of neologisms in their linguistic context, which can be used for linguistic research. Information concerning the social context is available in the form of the coiner's profile, if known, and the online types of (social) media in which the neologism occurs. The entire processing of all nearly 1000 neologisms presently takes about seven hours.

The data can be downloaded as html or txt at this point and fed into different analysis tools or concordancers, but users also have the option to access the data from the database in an online interface. Before this interface will be discussed in 3.3, the structure and contents of the current database will be briefly presented in the next section.

3.2. The NeoCrawler database

At the time of writing in July 2018 the NeoCrawler database contained 958 neologisms and over 2,600,000 html pages. These 958 neologisms are currently queried by the Observer on a weekly basis in order to closely monitor their potential diffusion in the English language and society as represented on the web. All lemmas have been annotated with regard to meanings, part-of-speech and underlying word-formation processes. Unsurprisingly the vast majority of the 958 neologisms are nouns (79%), followed by adjectives (15%) and verbs (12%). Adverbs and phrases account for 1%. In terms of word-formation, compounds and blends contribute almost equal proportions of lexemes, viz. 37% and 31% respectively. Derivations are found in 24% of the neologisms. The remaining 8% are distributed across clippings, borrowings, acronyms, conversions and ex-nihilo formations. Both distributions, in terms of word classes and in terms of word-formation, strikingly mirror general tendencies as found in the OED additions between 1950 and 2010 and observations by Bauer (1983), Ayto (2003) and Algeo (1998).

The present structure and composition of the database is illustrated in Figures 5a–d. Only neologisms with new tokens between April 2016 and April 2018 are reported, totalling 716 items.⁶ The first figure 5a displays the cumulative number of tokens for all 716 types. Not surprisingly, the peak of highly frequent neologisms is rather thin compared to the long tail with a wide range of medium frequency lexemes petering out to 28 neologisms which have hardly been attested at all in use during the observed 2-year period. The 25 most frequent neologisms in the database are shown in Figure 5b. From *Trumpism* to *circular economy*, a diversity of topic domains is covered, ranging from politics and economy (e.g. *Brexit*, *post-truth*) to lifestyle (e.g. *liveblog*, *glamping*) and technology (e.g. *internet of things*, *blockchain*). The 25 items around the median in Figure 5c and the 25 least frequent items in Figure 5d display a similar distribution across topics.

⁶ It should be noted that the entire monitoring period differs for each lexeme as a result of the continuous addition of new items.

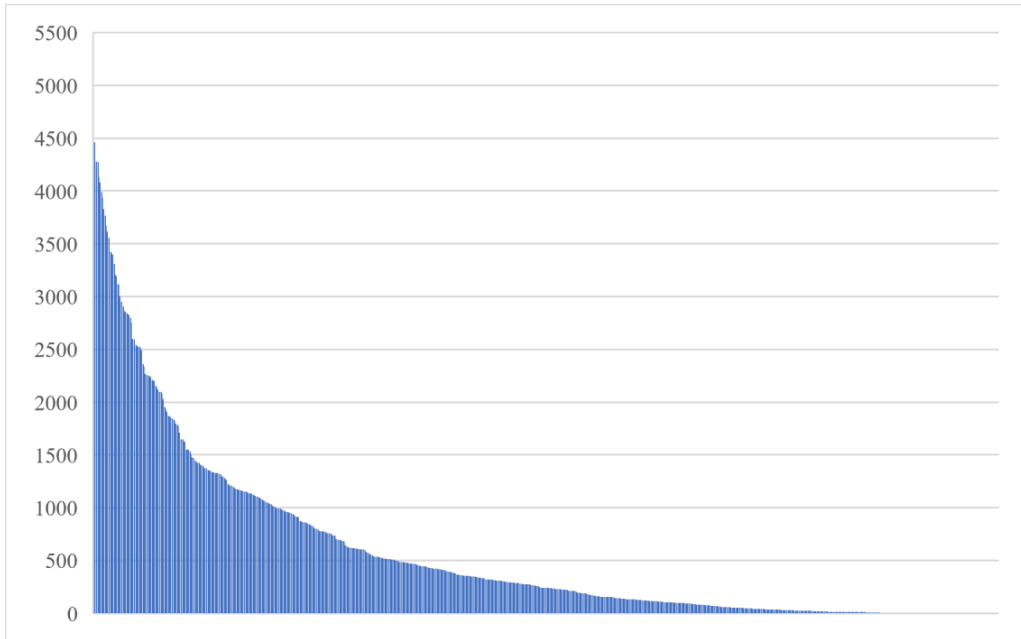


Figure 5a: Frequency for each item in the NeoCrawler database.

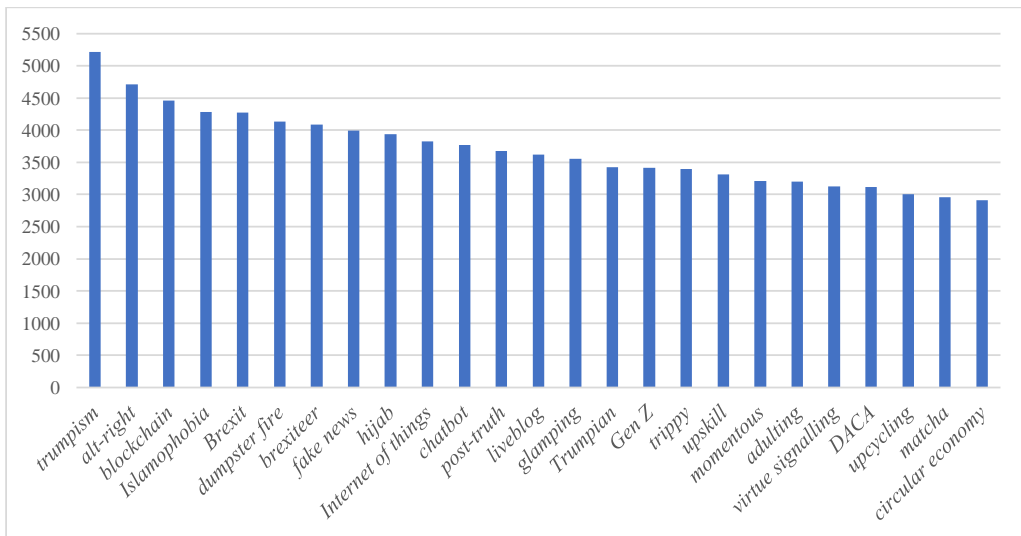


Figure 5b: 25 most frequent items in the NeoCrawler database.

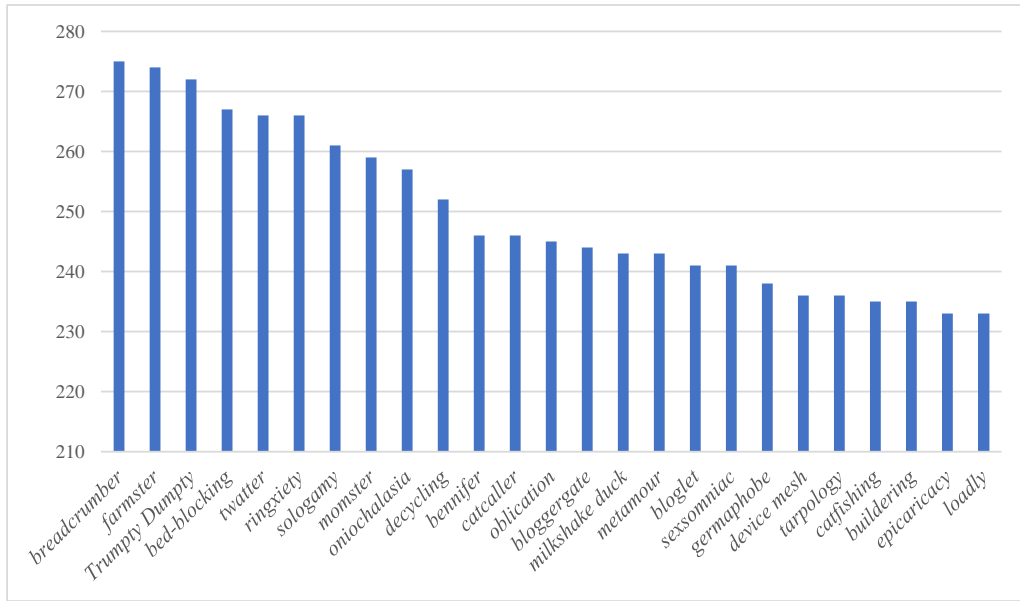


Figure 5c: 25 items around the median frequency in the NeoCrawler database.

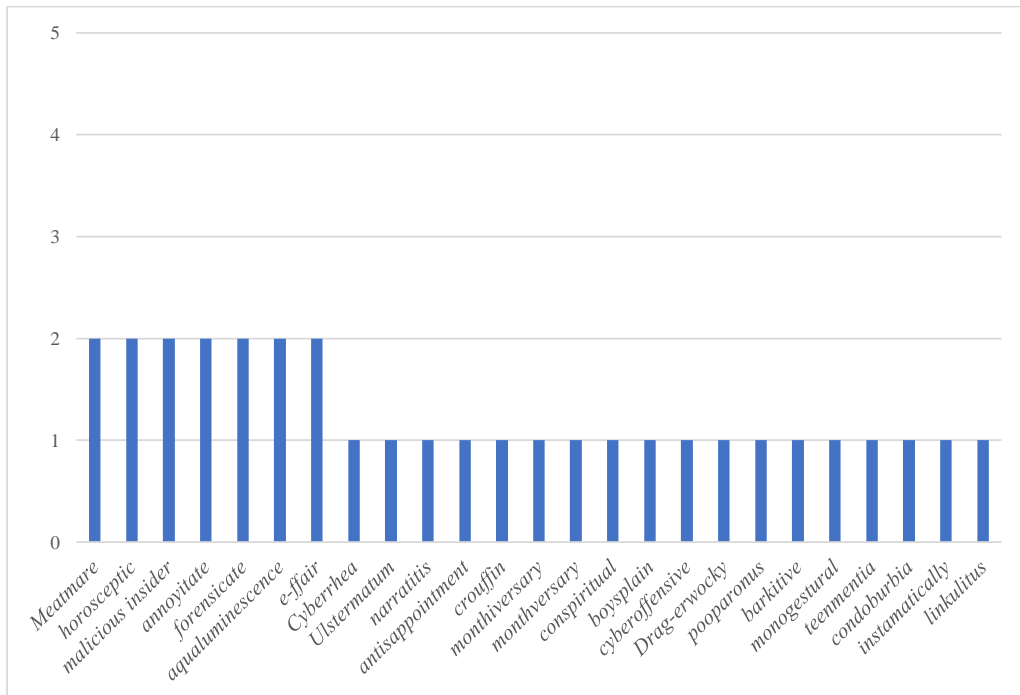


Figure 5d: 25 items with the lowest frequency higher than 0 in the NeoCrawler database.

3.3. The Observer interface

The online NeoCrawler interface has two sections: one for the Discoverer module, as shown in section 2 and one for the Observer module, which will be presented in brief in the following section. The front page in the Observer part consists of an alphabetical list of all neologisms under

current observation with their part-of-speech, word-formation process, meaning and the date at which the lemma was added to the database. Clicking on a specific entry activates a drop-down menu from which the user can select information to be displayed on a second, neologism-specific level, as shown for *lifhack* in Figure 6. *Lifhack* is classified as a nominal compound (*comp, n*) added in August 2016, which denotes ‘actions and activities that make life easier or better’.

▶ libtards	2012-02-13	bien	n	(n.) a derogatory term for a person who holds a liberal political view
▼ lifhack	2016-08-25	comp	n	
→ progress of lifhack				
→ edit lemma				
→ show statistics				an action you can take to make your life easier or better

Figure 6: First-level front page interface for *lifhack*.

Via the link ‘edit lemma’ the user can add or change semantic, morphological and morphosyntactic information provided for each lemma or insert notes on relevant characteristics (coiner, medium, extralinguistic topicality, etc.). Moreover, the user can decide to keep on monitoring the lexeme or exclude it from future searches. This level is therefore concerned with (meta)data management. Additional metadata pertaining to quantitative properties is accessible via the ‘show statistics’ button in the drop-down menu. As shown in Figure 7 this page contains a visual graph of the frequency development of the neologism measured as new tokens per week. In addition, basic facts on the duration of the observation so far, the total number of tokens in the database and, in view of future classification, the number and percentage of classified tokens are provided. At the time of writing in July 2018, *lifhack* had been observed for 111 weeks, producing a total number of 8801 tokens awaiting classification. The maximum number of tokens during one week is 283. These numbers taken together with the shape of the frequency curve provide a cursory glance into the dynamics of the diffusion process (cf. Kerremans 2015).

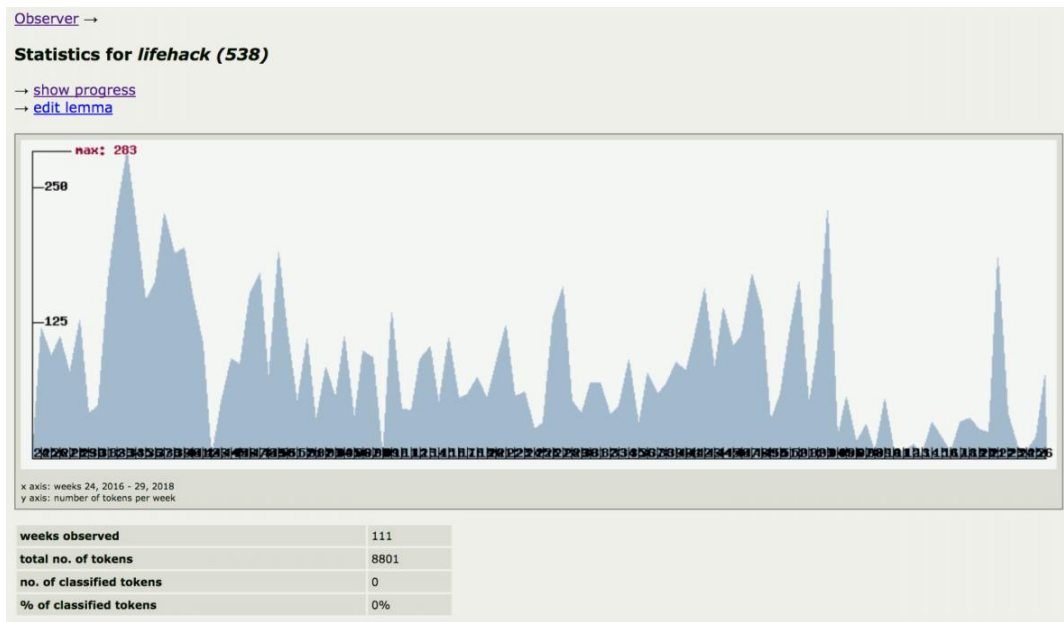


Figure 7: Second-level statistics page for *lifhack*.

Turning to the final option, clicking on ‘progress of lifehack’ leads the user to a second level, which contains a finer quantitative overview of the weekly search results together with the possibility to proceed to an even more detailed level for each of these weeks. Figure 8 shows the second-level general overview of the progress of *lifehack*. For the sake of clarity, only the first three months are displayed. Here, each week is included with the precise dates and its search ID from the NeoCrawler database. For each week the number of pages found and the total number of tokens contained on these pages is listed. Finally, three buttons lead to further options: the green arrow starts the download, which can be restricted to a specific time frame and file format (html or txt), the ‘summary’ button and the ‘details’ button.










ID	Date	Restr.	No. of Pages	No. of Tokens	Details
33865	2016-06-27 - 2016-07-03 (week 26-2016)	w	13	120	summary details 
33497	2016-07-04 - 2016-07-10 (week 27-2016)	w	8	93	summary details 
33129	2016-07-11 - 2016-07-17 (week 28-2016)	w	11	112	summary details 
32761	2016-07-18 - 2016-07-24 (week 29-2016)	w	10	77	summary details 
32393	2016-07-25 - 2016-07-31 (week 30-2016)	w	9	127	summary details 
35951	2016-08-01 - 2016-08-07 (week 31-2016)	w	5	41	summary details 
36034	2016-08-08 - 2016-08-14 (week 32-2016)	w	5	49	summary details 
31224	2016-08-15 - 2016-08-21 (week 33-2016)	w	6	164	summary details 
31433	2016-08-22 - 2016-08-28 (week 34-2016)	w	16	227	summary details 
32019	2016-08-29 - 2016-09-04 (week 35-2016)	w	20	283	summary details 
34655	2016-09-05 - 2016-09-11 (week 36-2016)	w	17	220	summary details 
35090	2016-09-12 - 2016-09-18 (week 37-2016)	w	17	144	summary details 

Figure 8: Second-level general overview of the progress of *lifehack*.

The latter two buttons display more detailed information about the page (via ‘summary’) and each token on this particular page (via ‘details’). These options, once more using the example of *lifehack*, are presented in Figures 9 and 10. First, Figure 9 shows the summary level of all pages retrieved in the second crawling week listed Figure 8. The overview in Figure 9 contains several kinds of data regarding the page level, for reasons of readability again restricted to the first four pages of a total of eight retrieved in that particular week.

Lemma: lifehack

Date: 2016-07-04 - 2016-07-10
 Restriction: w
 Pages: 8
 Tokens: 93
 Deleted tokens: 0
 (5.7289 sec.)









#	Tokens	Title	Co-text	Classify Site	Links
1	9	Parents Of Successful Kids Do These 10 Things In Common ... (2198414)	Sign up Please enter a valid email address I've already subscribed newsletter It's never too late to start over. If you weren't happy with yesterday, try something different today. Don't stay stuck. Do better. I agree I disagree x We think so, too! Join Lifehack newsletter and we will ipire you to puue a happier existence. Sign up for free! Please enter a valid email address We think so, too! x Like us on facebook and we will ipire you to puure a happier existence. Like us 9K Follow us on Twitter and we	--field of discourse-- --type of source-- --authorship--	 
2	9	How to change Google Maps navigation voice to its original settings ... (2198415)	How to change Google Maps navigation voice to its original settings [Lifehack] RocketNews24 RocketNews24 RocketNews24 Japanese RocketNews24 Bringing you yesterday's news from Japan and Asia, today.	--field of discourse-- --type of source-- --authorship--	 
3	9	Drink Honey Lemon Water Every Morning - Amazing Benefits For A ... (2198416)	stay stuck. Do better. I Agree I Disagree x We think so, too! Join Lifehack newsletter and we will ipire you to puue a happier existence. Sign up for free! Please enter a valid email address x Good to see you here and we hope you'll enjoy reading on Lifehack! Don't miss our confirmation email for you! x We think so, too! Like us on facebook and we will ipire you to puure a happier existence. x We think so, too! Follow us on pinterest and we will ipire you to pu ure a happier existence. Follow us on Pinterest	--field of discourse-- --type of source-- --authorship--	 
4	4	Lifehack: How To Make A Disposable Vegetable Peeler From A ... (2198417)	Home Music Lifehack: How To Make A Disposable Vegetable Peeler From A Soda Can Lifehack: How To Make A Disposable Vegetable Peeler From A Soda Can Music Jul 10, 2016 If you ever go camping and you find that you did not take a vegetable peeler with you, no worries! In this video, you'll learn a way to make a vegetable peeler out of a beverage can.	--field of discourse-- --type of source-- --authorship--	 

Figure 9: Page-level interface of lifehack.

At the top of the page the summary of the search is repeated: it lists the dates, the weekly search interval, the total number of pages (8) and tokens (93). If manual evaluation on the deeper level should require the deletion of tokens, information about these steps will be displayed here too. For each page the number of tokens is given as well as data on the content: the title of the page on the one hand and the first token (in bold) on this page in its extended 500-character cotext. In the present example the first page, which contains nine tokens, is called “Parents of successful children do these 10 things in common”. If necessary, the globe icon will lead the user to the original website online, the briefcase to its stored counterpart in the NeoCrawler database. We store websites in our database so as to ensure replicability and validity in case they disappear from the web which also allows us to perform more sophisticated post-processing and classification techniques on our whole corpus in the future. At this point the user can proceed with a manual sociopragmatic classification on the page level (see Kerremans 2015 for a systematic manual application of these classifications). First, the field of discourse, or topic, can be chosen from a drop-down menu. These categories are based on the labels used in the BNC. Furthermore, the type of source, which is essentially the genre or text type, can be identified. Here, no ready-made categories existed at the time of the Observer’s development, which necessitated an idiosyncratic approach suitable for the present purpose (see Kerremans 2015 for a description): examples of categories are ‘blog’, ‘newspaper’ and ‘social media’. Finally, if applicable, the author of the page can be characterized as ‘private’ (the proverbial man on the Clapham omnibus) or ‘professional’, i.e. experts in the field concerned such as politicians, scientists or journalists. As mentioned above, we are currently working on the implementation of an automatic web genre and topic classifiers, which will replace the manual assessment regarding field of discourse and genre.

The final level in the Observer’s online interface pertains to the individual token level, which can be accessed through the ‘details’ button at the top of the summary page or on the general overview page (not shown in Figures 9 and 10). Figure 10 presents the token-level information for two pages from a query of *lifehack* conducted in May 2018.

#	ID token	Co-text	Classify Tokens
10	12065260	Dysfunctional Family Become Functional? Scroll down to continue reading article Lifehack &quo;s CEO has written a definitive guide on how to focus, learn the tips:	<input type="checkbox"/> ext. Co-text
11	12065261	come an Author Contact Us Terms and Conditio Privacy Policy © 2005 - 2018 Lifehack - All Rights Reserved.	<input type="checkbox"/> ext. Co-text

#	ID token	Co-text	Classify Tokens
12	12065262	to main content Open site search Search Search Close site search Shortlist logo Lifehack Apps And Web Services That'll Make You Better At Life ShortList skip to mai	<input type="checkbox"/> ext. Co-text
13	12065263	ditio Cookie Declaration Contact Copyright © 2010-2018 ShortList Gadgets 10 Lifehack Apps And Web Services That'll Make You Better At Life Share Tweet	<input type="checkbox"/> ext. Co-text
14	12065264	illed this list. Each of these apps and online services acts as something of a " lifehack " - a tedious modern term that boils down to "completing chores to give you time	<input type="checkbox"/> ext. Co-text

Figure 10: Token-level interface of *lifehack*.

The tokens for each page are listed under their process ID in the database together with the globe and briefcase icons allowing the user to access the original page on this level too. For reasons of clarity the token is by default displayed in a limited cotext, which can be expanded to the 500-character window used on the page level by checking the ‘ext. co-text’ button. Further token-level classification of linguistic properties such as part-of-speech, usage, textual position, style and meaning is planned in order to gain detailed information on the possibly changing linguistic behaviour of neologisms during their diffusion process. For now, this is a time-consuming manual procedure. As mentioned above, pages can be deleted if they prove to be undetected false positives, irrelevant advertisement or other less useful material from a linguistic point of view.

In sum, the Observer interface offers users an elaborate linguistic analysis and annotation package for each neologism under observation, providing information required for systematic investigations of their diffusion and the key factors influencing it.

4. Summary and future work

The NeoCrawler presents an ambitious state-of-the-art project extending the use of digital methods in empirical linguistics. It represents an innovative web-mining tool to identify and closely monitor lexical innovation and diffusion in computer-mediated discourse, which opens up new opportunities for linguists to tackle a number of unresolved and under-researched issues in the field of lexical innovation systematically on a large scale. This paper has presented the design as well as the most important characteristics of the two modules, the Discoverer and the Observer, with regard to the usage-based study of lexical innovation and diffusion. It has discussed the form-based neologism identification procedure of the Discoverer and its four components in detail, illustrating the semi-automatic detection by means of recent results. In particular, we have shown that by including different measures of the string matching distance, the effectivity of the Discoverer can be flexibly adjusted so as to achieve a healthy balance between recall and precision, which is important given the messiness of web data resulting from

typos, deliberate misspellings and creative language instances. As a result, our neologism database has exponentially grown in size, which is the prerequisite for a systematic study of the dynamics of lexical innovation and diffusion.

The Observer conducts the actual data mining in the form of weekly searches for new occurrences for all 958 neologisms present in the database at the time of writing. Accessing the web via Google's Custom Search API, the Observer extracts all pages containing new tokens and stores them after preliminary cleaning (boilerplate, duplicates) in the NeoCrawler database. By means of a multi-level content-organization display system, users receive quantitative and qualitative information regarding the linguistic behaviour of these neologisms during their diffusion process through language and society. The interface offers not only general statistics on each neologism but contains an embedded manual sociopragmatic classification system which will be largely automatized in the future to ensure better coverage of all relevant linguistic and extralinguistic facets of diffusion.

We do not, however, want to gloss over the challenges and drawbacks involved in our current data-mining approach. Presently, three main areas for improvement guide our work programme. Firstly, the current focus on Google as an access point is assumed to produce the most accurate results pertaining to the index of web pages, but also increases the risk of a quantitative and qualitative bias in the weekly output, caused by Google's commercial policy. Since a specialized linguistic search engine is not operational yet, the near future will make it necessary to include other search engines. As a consequence, we hope to be able to reduce our dependence on Google for the amount and content of the search results.

Secondly, the noise inherent in web data requires advanced cleaning of the material. Although duplicate detection and boilerplate removal so far produce a fairly clean sample, the primary focus will concern optimising the language identification. As mentioned in section 3.1 the Observer uses Google's language identification component as part of the API. Manual evaluation, however, has revealed a rather high error rate, which is problematic for our qualitative and quantitative analysis. The Observer will therefore need to be equipped with a separate language identification module in its post-processing pipeline, which will remove all non-English pages extracted from Google.

Finally, web genre and topic classifiers need to be implemented in order to reduce the amount of time-consuming manual classification and enable us to process larger batches of data than presently possible. Regarding all three challenges the various options for extensions and improvements are currently being looked into. In the meantime, the NeoCrawler nevertheless is a unique tool for the formal identification of English neologisms on the web and continuous tracking of the diffusion of lexical innovation in language and society.

References

- Algeo, John. 1998. Vocabulary. In Suzanne Romaine (ed.), *The Cambridge history of the English Language*, vol. 3, Cambridge: Cambridge University Press. 57–91.
- Ayto, John. 2003. Newspapers and neologisms. In Jean Aitchison & Diana M. Lewis (eds.), *New media language*, 182–187. Routledge: New York.
- Baayen, Harald R. & Anneke Neijt. 1997. Productivity in context: A case study of a Dutch suffix. *Linguistics* 35. 565–587.
- Bauer, Laurie. 1983. *English word-formation*. Cambridge: Cambridge University Press.
- Cabré, Maria Teresa & Lluís de Yzaguirre. 1995. Stratégie pour la détection semiautomatique des néologismes de presse. *TTR: Traduction, Terminologie, Redaction* 8. 89–100.
- Cartier, Emmanuel. 2017. Néoveille, a web platform for neologism tracking. Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 95–98.
- Cartier Emmanuel. 2019. (to appear). Néoveille, plateforme de détection, de description et de suivi des néologismes en onze langues. *Néologica*.
- Falk, Ingrid, Delphine Bernhard & Christophe Gérard. 2017. The Logoscope: A semi-automatic tool for detecting and documenting French new words from the linguistic project to the web interface. Research Report, Université Strasbourg. <https://hal.archives-ouvertes.fr/hal-01896796>. [accessed 1 August 2018].
- Fischer, Roswitha. 1998. Lexical change in present-day English. A corpus-based study of the motivation, institutionalization, and productivity of creative neologisms. Tübingen: Narr.
- Gérard, Christophe, Lauren Bruneau, Ingrid Falk, Delphine Bernhard & Ann-Lise Rosio. 2017. Le Logoscope : Observatoire des innovations lexicales en français contemporain. In Joaquín García Palacios, Goedele de Sterck, Daniel Linder, Jesús Torre del Rey, Miguel Sánchez Ibanez & Nava Maroto García (eds.), *La neología en las lenguas Románicas: Recursos, estrategias y nuevas orientaciones*. Frankfurt : Peter Lang. 339-356.
- Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016. Cultural Shift or Linguistic Drift? Comparing Two Computational Models of Semantic Change. *Proceedings of Conference on Empirical Methods on Natural Language Processing, Austin, Texas, USA, 1-5 November 2016*. <http://aclweb.org/anthology/D/D16/D16-1229.pdf> [accessed 1 March 2018].
- Iakovleva, Tatiana. 2017. Automatic detection of neologisms in Russian newspaper corpora with Néoveille. *Proceedings of the International Conference CORPUS LINGUISTICS–2017, St Petersburg, 27-30 June 2017*. 43–47. <https://hal-univ-diderot.archives-ouvertes.fr/hal-01540995/document> [accessed 1 May 2018].
- Janssen, Maarten. 2005. NeoTrack: Semiautomatic neologism detection. *APL Conference 2005, Lisboa, Portugal*. <http://maarten.janssenweb.net/index.php?action=publications> [accessed 15 March 2018].
- Jatowt, Adam & Kevin Duh. 2014. A framework for analysing semantic change of words across time. *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*. 229–238.
- Kerremans, Daphné, Susanne Stegmayr & Hans-Jörg Schmid. 2012. The NeoCrawler: Identifying and retrieving neologisms from the internet and monitoring on-going change. In Kathryn Allan & Justyna Robinson (eds.), *Current methods in historical semantics*, 59–96. Berlin: Mouton de Gruyter.

- Kerremans, Daphné. 2015. A web of new words. A corpus-based study of the conventionalization process of English neologisms. Frankfurt am Main: Peter Lang.
- Kerremans, Daphné & Jelena Prokić. 2018. Mining the web for new words: Semi-automatic neologism identification with the NeoCrawler. *Anglia* 136(2). 239–268.
- Labov, W. 1966. *The social stratification of English in New York City*. Washington: Center for Applied Linguistics.
- Labov, William. 1980. The social origins of sound change. In William Labov (ed.), *Locating language in time and space*, 251–266. New York: Academic Press.
- Labov, William. 2001. Principles of Linguistic change. Volume II: Social Factors. Oxford: Blackwell.
- Levenshtein, Vladimir I. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10. 707–710.
- Lewandowski, Dirk. 2008. A three-year study on the freshness of web search engine databases. *Journal of Information Science* 34(6). 817–831.
- Liao, Xuanyi & Guang Cheng. 2016. Analysing the semantic change based on word embedding. In Natural language understanding and intelligent applications. *Proceedings of the 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2–6, 2016*. Cham: Springer. 213–223.
- Liu, Tsun-Jui, Shu-Kai Hsieh & Laurent Prevot. 2013. Observing features of PTT neologisms: A corpus-driven study with N-gram model. *Proceedings of the Twenty-Fifth Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*. 250–259.
- Megerdooimian, Karine & Ali Hadjarian. 2010. Mining and classification of neologisms in Persian blogs. *Proceedings of the 2nd Workshop on Computational Approaches to Linguistic Creativity (HLT 2010)*. 6–13.
- Milroy, James & Lesley Milroy. 1985. Linguistic change, social network and speaker innovation. *Journal of Linguistics* 21. 339–384.
- Nevalainen, Terttu. 2000. Mobility, social networks and language change in Early Modern England. *European Journal of English Studies* 4(3). 253–264.
- Nevalainen, Terttu & Helena Raumolin-Brunberg. 2003. Historical sociolinguistics: Language change in Tudor and Stuart England. London: Longman.
- Plag, Ingo. 1999. Morphological productivity: Structural constraints in English derivation. Berlin/New York: Mouton de Gruyter.
- Säily, Tanja. 2018. Explorations into the social contexts of neologism use in early English correspondence. To appear in *Pragmatics & Cognition*.
- Schmid, Hans-Jörg. 2016. *English morphology and word-formation: An introduction*. 3rd revised and extended edition. Berlin: Erich Schmidt.
- Tagliamonte, Sali A. & Derek Denis. 2014. Expanding the transmission/diffusion dichotomy: Evidence from Canada. *Language* 90(1). 90–136.
- Torres-del-Rey, Jesús & Nava Maroto. 2014. Building the interface between experts and linguists in the detection and characterisation of neology in the field of neurosciences. *Proceedings of the 4th International Workshop on Computational Terminology, Dublin, Ireland, August 2014*. 64–67. <https://aclanthology.info/papers/W14-4808/w14-4808> [accessed 25 March 2018].
- Tournier, Jean. 1985. Introduction Descriptive à la Lexicogénétique de l'Anglais Contemporain. Paris: Champion-Slatkine.

Wilson, Lee. 2017. Google Freshness Algorithm: Everything you need to know. *Search Engine Journal*. <https://www.searchenginejournal.com/google-algorithm-history/freshness-update/>. Last accessed August 1, 2018.

2.3 Conclusions

The NeoCrawler successfully retrieved a large sample of neologisms on the web. As indicated above, this sample appears to encompass a broad spectrum of lexical innovation, both in terms of its formal composition (word classes, word-formation processes; Section 3.2*) and its continuous coverage of the frequency spectrum (Figure 5*). However, the NeoCrawler approach also revealed a number of limitations for studying the diffusion of neologisms.

Firstly, utilising frequency counts as indicators for diffusion turned out to be less reliable than expected due to the dependence on Google’s search index, which remains a commercial black box for research. Google’s Custom Search API³ is officially limited to a maximum of 100 hits per search query. However, even highly frequent lexemes such as *Brexit* never reached this ceiling. Manual searches using Google and other search engines reveal that frequent terms like *Brexit* occur on far more pages per week than indicated by Google Custom Search, raising doubt on the reliability of the numbers returned by the API. Furthermore, the NeoCrawler is unable to provide relative frequency counts (e.g. per million words), since Google does not disclose the total size of their search index. This poses a challenge for time-series studies of frequency data since changes in the total size of the web corpus, which is most certainly growing with time, cannot be accounted for by statistical means.

Secondly, due to technical challenges, the NeoCrawler’s objective of enabling research into the diffusion of neologisms across usage contexts could not be reached. Due to the sample and corpus size involved, investigating the use of words in different contexts requires the automatic classification of websites by text type (e.g. newspaper, blog) or domain of discourse (e.g. sports, politics), for example. However, despite several attempts (Schlegel 2014; Maier 2016), the NeoCrawler could not be extended to provide these types of information.

The aim of the following chapter is to assess the problems discussed in this section and to address the inherent limitations of the current approach by supplementing the NeoCrawler data with data obtained from Twitter.

³<https://developers.google.com/custom-search>

3 Diffusion on the web and social media

3.1 Research context

This chapter explores the diffusion of neologisms on the web and on the social media platform Twitter. It examines a set of three selected recent neologisms to provide a more detailed view of their diffusion and the factors influencing their spread.

Supplementing the web data with additional data from Twitter also serves to address the problems associated with the NeoCrawler discussed above (Section 2.3), as this approach allows for the evaluation and cross-validation of the reliability of the NeoCrawler results. Additionally, it enables the investigation and comparison of the use of neologisms across two usage contexts: the web and social media.

Social media platforms like Twitter have become crucial for spreading new ideas and new words. Concluding her web-based study, Kerremans (2015) emphasises the importance of social media and its potential for the study of the emergence and diffusion of neologisms:

I argue that the traditional role of newspapers as channels of diffusion is complemented, perhaps even substituted by the rapid mode of exchange of information within social networks many Internet users form part of with increasing creativity and flexibility. (Kerremans 2015: 232)

While the web still remains an essential marketplace for the exchange of information and ideas, the importance of communication on social media has seen an even more significant increase in recent years. The potential of social media for the emergence and diffusion of new words was shown by the NeoCrawler's Discoverer module, which returned more neologisms for searches based on Twitter versus web data on average.

To leverage this potential, this chapter examines the use of three selected neologisms on the web and on Twitter: *rapefugee*, *rapeugee*, and *rapugee*. All three terms are blends of *rape* and *refugee* and were coined in 2015 as derogatory propaganda terms by far-right opponents of policies that welcome asylum-seekers.

The three neologisms are formal variants encoding the same meaning: 'A refugee who rapes people. Usually referred to the Muslim refugees pouring into Europe.' (Urban Dictionary¹) Thus, they constitute a case of onomasiological competition, which is particularly suitable for studying diffusion dynamics since established models of diffusion like the S-curve Model are based on the assumption that linguistic innovations

¹'rapefugee', <https://www.urbandictionary.com/define.php?term=rapefugee>, accessed 23 May 2022.

are involved in competition processes that require ‘replicator selection’ among a set of formal variants (Blythe & Croft 2012).

Moreover, the fact that all three neologisms are identical in meaning controls for the effect of semantics on the diffusion process. Previous studies assessing the factors influencing the spread of neologisms revealed that the ‘semantic carrying capacity’ (Nini et al. 2017), i.e. the semantic potential of words as influenced by topicality (Fischer 1998: 16), for example, ranks among the most influential factors for the success of neologisms (Karjus 2020). The present study of formal variants excludes semantic effects and thus allows for a more reliable assessment of usage-related factors such as salience (e.g. use in titles or hashtags) and metalinguistic use, which were the primary focus of this study.

The following section presents this study, titled *Using the Web and Social Media as Corpora for Monitoring the Spread of Neologisms. The case of ‘rapefugee’, ‘rapeugee’, and ‘rapugee.’*, which I co-authored with Hans-Jörg Schmid (HJS), Desislava Zhekova (DZ), and Fazleh Elahi (FE). The paper was published in *Proceedings of the 10th Web as Corpus Workshop* by the Association for Computational Linguistics in 2016. The study is based on web data, which were retrieved and described by FE in Section 4.1*, and on a Twitter corpus, which was compiled and described by DZ in Section 4.2*. I conducted all quantitative and qualitative analyses of the data and produced the figures and tables presented in the paper. I wrote the majority of the paper, with HJS contributing to the introduction and operationalization sections (1* and 3*). JP, FE, and HJS contributed to the final version of the manuscript by providing revisions and comments.

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

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 *Neo-Crawler*, 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

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-Webster¹, *About words* by Cambridge University Press², *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 or-

der 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 resolution (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/>

³www.urbandictionary.com

⁴<http://www.wordspy.com/>

⁵www.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 string⁶ defining values for a number of parameters (such as language, date, token etc.). There are several advantages of this approach over other Google search APIs⁷, 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

⁶https://encrypted.google.com/search?num=100&hl=en&lr=lang_en&start=0&tbs=lr%3Alang_len%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 API⁸, 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 API⁹ 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

⁸<https://dev.twitter.com/streaming/firehose>

⁹<https://dev.twitter.com/streaming/overview>

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.

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-

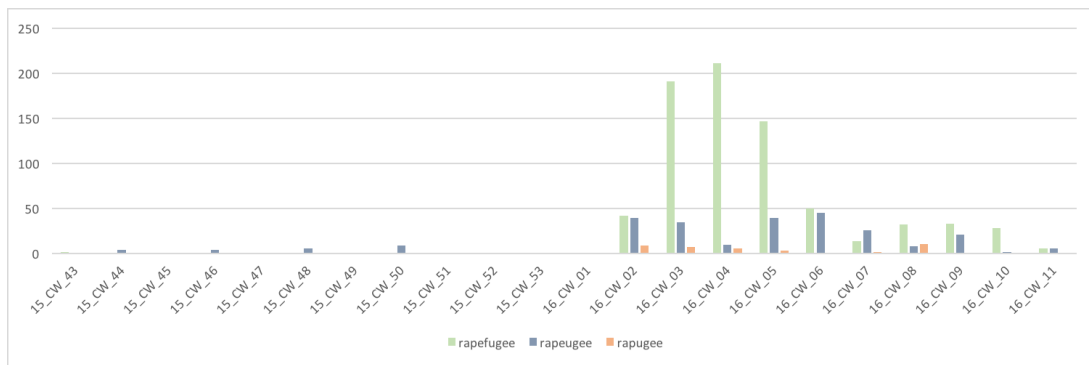


Figure 1: Absolute frequencies in the web corpus

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 attestation 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

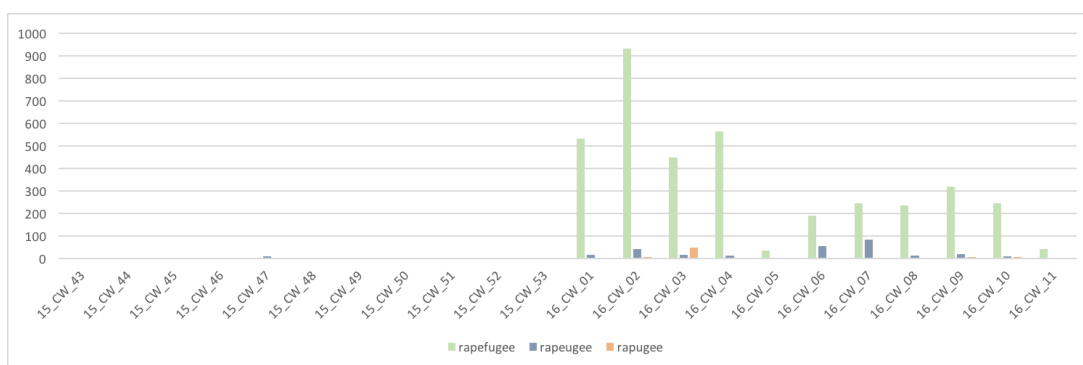


Figure 2: Absolute frequencies in the Twitter corpus

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

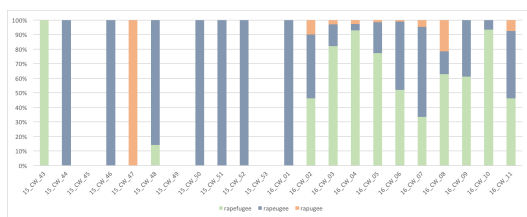


Figure 3: Relative frequencies in the web corpus

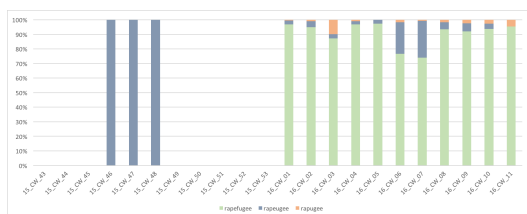


Figure 4: Relative frequencies in the Twitter corpus

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, how-

ever, 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 *rapegee* – 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 Nørby Jensen, editors, *Sprognaevets Konferencserie 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 Linguisticae*, 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).

3.3 Conclusions

Despite the issues regarding the reliability of the NeoCrawler's frequency-based measures for diffusion discussed above (Section 2.3), the current paper demonstrates broad agreement in the spread of the selected cases between the web data obtained from the NeoCrawler, the Twitter data, and extralinguistic events associated with the selected neologisms. This cross-validates the web and Twitter approaches, implying that both datasets can be used to analyse the diffusion of neologisms with a reasonable degree of reliability.

These findings should be interpreted with care, however, since they are based on a sample of just three neologisms, and their adoption has been heavily influenced by their topicality as well as their controversial meaning and pragmatic function. Moreover, since all three neologisms have low absolute frequency counts and both datasets are limited to a short time period, the quantitative results can only be interpreted tentatively.

The paper highlights the importance of social media as a driving force in the emergence and diffusion of neologisms. The selected neologisms were aggressively promoted on Twitter in the early stages of their diffusion, and their rising popularity on social media boosted their use on the web. The Twitter data clearly demonstrate the influence of certain communities in promoting higher usage intensity (Stefanowitsch & Flach 2017) of the selected neologisms. The qualitative analysis revealed that despite this increased usage intensity, the use of these terms remains limited to a small number of like-minded individuals and communities on the far-right of the political spectrum.

4 Social networks of diffusion

4.1 Research context

This chapter aims to provide a more detailed view of the social dynamics of diffusion on Twitter. The previous chapter suggested that communities promoting the spread of neologisms can significantly influence their spread. Moreover, it suggested that when a neologism has disproportionately high usage intensity in some portions of the speech community, total usage frequency may represent an inaccurate indicator and overestimate degrees of social diffusion. Based on these findings, this chapter presents an in-depth study of the spread of neologisms across speakers and communities on Twitter.

To improve generalizability, this study was based on a substantially larger sample of neologisms than the preceding study. It investigates the emergence and diffusion of 99 neologisms, the majority of which were detected using the NeoCrawler's Discoverer module, plus selected examples from a previous study on the emergence of neologisms on Twitter by Grieve, Nini & Guo (2016). Moreover, it is based on a large-scale Twitter dataset encompassing around 30 million tweets covering the time from Twitter's inception in 2006 to the end of 2018. The neologisms in the sample were selected to have emerged within this time frame, enabling this study to capture their incipient diffusion and trace their spread over substantially longer time periods than the previously used NeoCrawler approach.

To address the limitations of depending exclusively on frequency counts, as used in earlier chapters, it introduces more fine-grained measures such as the coefficient of variance (CV) to capture temporal dynamics of diffusion such as 'topicality' (Fischer 1998: 16) and 're-current semi-conventionalization' (Kerremans 2015: 129–136). Additionally, it employs a social network analysis approach to give direct insights into the dynamics of social diffusion. Several network-related metrics (e.g. centralization, in-degree) and network graph visualisations are used to determine whether neologisms successfully diffuse across speakers and communities. Finally, the paper evaluates and compares frequency-based and network-based indicators of diffusion for the entire sample of neologisms.

When applied to the previous case of the term *rapefugee*, this approach allows for a more comprehensive, quantitative picture of its diffusion. The findings confirm the preceding qualitative observation that its relatively high usage frequency overestimates its degree of social diffusion because its use remains largely limited to small portions of the speech community, which use the term with disproportionately high usage intensity.

The following section presents the paper *Social Networks of Lexical Innovation. Investigating the Social Dynamics of Diffusion of Neologisms on Twitter*, published in the Research Topic Computational Sociolinguistics in *Frontiers in Artificial Intelligence*. Being the single author of this paper, I am exclusively responsible for the data, code, and methods used.

High-resolution, zoomable versions of the network graphs can be viewed by downloading and opening the files from the following links:

Figure 3* | <https://osf.io/4ta6u/>

Figure 5* | <https://osf.io/9dpj8/>

4.2 Social Networks of Lexical Innovation. Investigating the Social Dynamics of Diffusion of Neologisms on Twitter



Social Networks of Lexical Innovation. Investigating the Social Dynamics of Diffusion of Neologisms on Twitter

Quirin Würschinger*

Department of English and American Studies, LMU, Munich, Germany

OPEN ACCESS

Edited by:

Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:

Alina Maria Cristea,
University of Bucharest, Romania
Alekssei Ioulevitch Nazarov,
Utrecht University, Netherlands

*Correspondence:

Quirin Würschinger
q.wuerschinger@lmu.de

Specialty section:

This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

Received: 31 December 2020

Accepted: 13 July 2021

Published: 01 November 2021

Citation:

Würschinger Q (2021) Social Networks
of Lexical Innovation. Investigating the
Social Dynamics of Diffusion of
Neologisms on Twitter.
Front. Artif. Intell. 4:648583.
doi: 10.3389/frai.2021.648583

Societies continually evolve and speakers use new words to talk about innovative products and practices. While most lexical innovations soon fall into disuse, others spread successfully and become part of the lexicon. In this paper, I conduct a longitudinal study of the spread of 99 English neologisms on Twitter to study their degrees and pathways of diffusion. Previous work on lexical innovation has almost exclusively relied on usage frequency for investigating the spread of new words. To get a more differentiated picture of diffusion, I use frequency-based measures to study temporal aspects of diffusion and I use network analyses for a more detailed and accurate investigation of the sociolinguistic dynamics of diffusion. The results show that frequency measures manage to capture diffusion with varying success. Frequency counts can serve as an approximate indicator for overall degrees of diffusion, yet they miss important information about the temporal usage profiles of lexical innovations. The results indicate that neologisms with similar total frequency can exhibit significantly different degrees of diffusion. Analysing differences in their temporal dynamics of use with regard to their age, trends in usage intensity, and volatility contributes to a more accurate account of their diffusion. The results obtained from the social network analysis reveal substantial differences in the social pathways of diffusion. Social diffusion significantly correlates with the frequency and temporal usage profiles of neologisms. However, the network visualisations and metrics identify neologisms whose degrees of social diffusion are more limited than suggested by their overall frequency of use. These include, among others, highly volatile neologisms (e.g., *poppygate*) and political terms (e.g., *alt-left*), whose use almost exclusively goes back to single communities of closely-connected, like-minded individuals. I argue that the inclusion of temporal and social information is of particular importance for the study of lexical innovation since neologisms exhibit high degrees of temporal volatility and social indexicality. More generally, the present approach demonstrates the potential of social network analysis for sociolinguistic research on linguistic innovation, variation, and change.

Keywords: lexicology, lexical innovation, sociolinguistics, diffusion, social media, Twitter, time-series analysis, social network analysis

1 INTRODUCTION

Societies continually evolve, new products and practices emerge, and speakers coin and adopt new words when they interact and share information. How do these new words spread in social networks of communicative interaction?

In a recent paper analysing contagion patterns of diseases in *Nature Physics*, Hébert-Dufresne et al. (2020) suggest that the spread of viruses like SARS-CoV-2 follows principles of complex contagion through social reinforcement, and that it matches the dynamics of diffusion of cultural and linguistic innovations such as new words and internet memes. Does this confirm the widespread perception that new words 'go viral'? Influential sociolinguistic models of the spread of linguistic innovations like the S-curve model (Milroy 1992) share fundamental features with earlier economic models of diffusion (Rogers 1962). It is often assumed that diffusion in social networks follows universal trajectories and that rates of spread depend on social dynamics such as network density and the presence or absence of weak ties (Granovetter 1973). Unlike research on biological and cultural diffusion processes, however, sociolinguistic research has only recently been provided with data sources that are equally suitable for large-scale, data-based approaches which can rely on network analyses to study these phenomena empirically.

Social media platforms like Twitter have changed the way we communicate and how information spreads, and they offer valuable data for empirical research. For linguists, social media provides large amounts of data of authentic language use which opens up new opportunities for the empirical study of language variation and change. The size of these datasets as well as their informal nature allow for large-scale studies on the use and spread of new words, for example, to gain insights about general trajectories of diffusion (Nini et al., 2017) or about factors that influence whether new words spread successfully (Grieve, 2018). Moreover, metadata about speakers facilitate the study of aspects of diffusion that go beyond what can be captured by usage frequency alone. Recent work has used Twitter data to investigate the geographical spread of lexical innovations (Eisenstein et al., 2014; Grieve et al., 2016), for example.

Data about the communicative interaction of speakers additionally allows performing network analyses of the social dynamics of diffusion processes. Network science approaches to social media data have been successfully employed in diverse fields, for example, to study the spread of diseases (Lu et al., 2018), opinions (West and Hristo, 2014) and political attitudes (Pew Research Center 2019). While the study of social networks has a long research tradition in sociolinguistics and has shaped influential models of diffusion (e.g., Milroy and Milroy 1985), large-scale network analyses of sociolinguistic phenomena have only recently become more widespread. These new data sources and methodological advances put computational sociolinguistics in an excellent position to gain new insights and to test long-standing theoretical models empirically.

In the area of lexical innovation, this can serve to evaluate important theoretical concepts like the role of early adopters, network density and weak ties in the diffusion of new words. For

example, previous approaches have used computational modelling to test the validity of the S-curve model (Blythe and Croft 2012), and to model processes of simple and complex contagion of linguistic innovations in social networks (Goel et al., 2016). Applying social network analysis to bigger samples of neologisms and tracking their use and spread on social media datasets promises to provide a more detailed picture of social diffusion. Social network information has the potential to more accurately assess the degrees to which the adoption of new words remains limited to closely connected sub-communities or whether they reach larger parts of the speech community.

This paper aims to explore the role of network information and temporal dynamics in assessing the diffusion of lexical innovations on Twitter. I use several quantitative and qualitative methods to study diffusion. I conduct a longitudinal study monitoring the use of a broad sample of neologisms to analyse their usage frequency and the temporal dynamics underlying their use. Next, I use social network analyses to get a better picture of the sociolinguistic dynamics at play, to assess different pathways and overall degrees of diffusion. Lastly, I combine both approaches to get a more detailed picture of the diffusion of the neologisms in the sample, and to assess the results of both approaches to diffusion.

The paper is structured as follows. **Section 2** introduces the theoretical framework for modelling and measuring the diffusion of lexical innovations which forms the basis for the empirical study. **Section 3** presents information about the sample of neologisms and the Twitter dataset this study is based on. **Section 4** describes the methods used for analysing diffusion. **Section 5** presents the results of the empirical study. I analyse diffusion on the basis of frequency and social networks and integrate the results obtained from both approaches. **Section 6** summarises and discusses the results from the empirical study and draws implications about the role of frequency and network-based measures for the study of diffusion.

2 MODELLING AND MEASURING THE DIFFUSION OF LEXICAL INNOVATIONS

2.1 Modelling Diffusion

Neologisms are on a continuum from entirely novel word-formations to fully established lexemes which are familiar to the majority of the speech community. Neologisms have spread to some extent, but are still perceived as new or unknown by many speakers (Schmid 2016). On one end of the continuum, 'ad-hoc formations' are new words that have been coined in a concrete communicative situation, but are not adopted by interlocutors and do not diffuse beyond their original usage contexts (Hohenhaus 1996). On the other end, fully established words are known and used by the majority of the speech community. Neologisms occupy an intermediate position between both poles and can be defined as '(...) lexical units, that have been manifested in use and thus are no longer non- formations, but have not yet occurred frequently and are not widespread enough in a given period to have become part and parcel of the lexicon of the speech community and the majority of its members' (Kerremans 2015, 31).

Diffusion can be seen as the process that transports successful neologisms along this continuum while they are becoming increasingly conventional in the speech community. The S-curve model (Milroy 1992; Nevalainen 2015; Labov 2007) expects an S-shaped trajectory for the spread of linguistic innovations and makes specific assumptions about the sociolinguistic characteristics of speakers involved in the diffusion process. In a first stage of slow diffusion, only a small number of early adopters take up the innovative words. These individuals typically form dense networks which are connected by strong ties. In the case of successful diffusion, the initial stages are followed by an acceleration in spread when new words increasingly reach speakers outside the initial communities. Weak ties (Granovetter 1973) play an important role in allowing the innovations to reach a bigger parts of the speech community. During later stages, rates of diffusion slow down again as the majority of the speech community has already adopted the new words, while a minority of speakers remains resistant to take up the new words.

The Entrenchment-and-Conventionalization Model (Schmid 2020) conceptualises the conventionalization of linguistic innovations as involving two processes: usualization and diffusion. Diffusion is defined as the 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 conformity regarding the types of cotexts and contexts in which they use it.' (Schmid 2020, 178–179, emphasis mine) In the case of a given new word, it is coined by an individual speaker and first reaches a community of speakers who might be closely-connected to the coiner and/or share interests related to the given neologism. With more advanced diffusion, the word spreads to larger numbers of speakers and increasingly also becomes conventional in other communities of speakers. The process of usualization, by contrast, leads to the increasing establishment of a given neologism by repeated use within one community of speakers. Neologisms thus show high degrees of conventionality, when they exhibit high usage intensity across a large number of speakers and communities.

2.2 Measuring Diffusion

Earlier empirical work on lexical innovation had to rely on smaller, general-purpose linguistic corpora. The low-frequency nature of neologisms limited earlier studies to conducting case studies on selected neologisms (Hohenhaus 1996) or on specific domains of neology (Elsen 2004). In recent years, research on lexical innovations has seen an upsurge in large-scale empirical investigations on the diffusion of neologisms, thanks to the availability of new data sources and computational methods.

The increasing availability of web corpora significantly extended the opportunities for large-scale corpus analyses. Modern corpora like the NOW corpus (Davies 2013) allow to study more comprehensive samples of neologisms and enable researchers to monitor their use over time, which is essential for investigating diffusion processes. In addition to general-purpose web corpora, several research groups built dedicated tools and specialized corpora for the monitoring and analysis of neologisms (Renouf et al., 2007; Kerremans et al., 2012; Lemnitzer, 2010; Gérard et al., 2017; Cartier 2017).

More recently, social media data have become an increasingly important alternative to web corpora. Language use on social media is informal and creative, which makes it a hotbed for lexical innovation. Recent work using Twitter data has focused, for example, on the identification of neologisms (Grieve et al., 2018), on their geographical diffusion (Eisenstein et al., 2014), and on trajectories of diffusion (Nini et al., 2017). Empirical investigations on the basis of Reddit data include studies of the linguistic dissemination of neologisms (Stewart and Jacob, 2018) and the role of innovators and adopters (Del Tredici et al., 2018).

The present study is based on Twitter data and goes beyond previous work in its focus on the sociolinguistic dynamics of diffusion, which are at the core of theoretical models of diffusion. Most previous empirical investigations of the spread of new words have been limited to using frequency measures as an indicator of diffusion. While frequency counts have proven useful in previous work, they can only provide limited insight into the sociolinguistic dynamics of diffusion (Stefanowitsch and Flach 2017). In addition to usage frequency, I will therefore use network information to assess the social pathways of diffusion in the present dataset.

3 DATA

3.1 Neologism Sample

The present study is based on a selection of 99 neologisms and investigates their use on Twitter from its launch in 2006 to the end of 2018. The lexemes were selected to cover a broad spectrum of lexical innovation. Previous work by Kerremans (2015, 115–147) has identified four main clusters of neologisms on the conventionalization continuum: 'non-conventionalization', 'topicality or transitional conventionalization', 'recurrent semi-conventionalization' and 'advanced conventionalization'. The present sample was designed to cover these categories and largely contains neologisms taken from the NeoCrawler (Kerremans et al., 2012), which uses dictionary-matching to retrieve a semi-automatic, bottom-up selection of recent neologisms on the web and on Twitter (Kerremans et al., 2019). I have additionally included several lexemes that were statistically identified to have been increasing in frequency on Twitter in recent years by Grieve et al. (2016). I limit my selection to neologisms whose diffusion started after 2006 to have full coverage of the incipient stages of their spread on Twitter.

3.2 Twitter Corpus

Twitter is a popular micro-blogging platform that was started in 2006 and has become one of the most popular social media platforms today. Its broad user base and informal nature allow for a more representative picture of language use than domain-specific studies of, for example, newspaper corpora.¹ Twitter corpora have been successfully used to identify patterns of sociolinguistic variation in numerous previous studies. A

¹The present dataset was restricted to tweets in the English language. Due to the absence of the required metadata, the data cannot be further restricted to specific geographical regions, and it is not possible to identify native speakers of English.

recent study by Grieve et al. (2019), for example, has demonstrated the reliability of large-scale Twitter datasets for studying lexical variation.

Twitter is particularly well-suited for studying lexical innovation due to the scale and types of data it provides, and due to the nature of language use on Twitter. The large size of Twitter's search index facilitates the quantitative study of neologisms, which requires large-scale datasets due to their inherently low frequency of occurrence. Twitter is widely used to discuss trends in society and technology, which makes it a good environment for studying the emergence of linguistic innovations. The informal and interactional nature of communication on Twitter fosters the rapid adoption of linguistic innovations, and the use of neologisms on social media platforms like Twitter often precedes and drives the diffusion of new words in more formal sources or on the web (Würschinger et al., 2016).

The data for this study were collected using the Python library *twint*, which emulates Twitter's Advanced Search Function. For each word in the sample, I performed a search query to retrieve all tweets found in Twitter's search index. Due to the large volume of more frequent lexemes, I limited the sample to contain only candidates for which I could collect all entries found in Twitter's index. The combined dataset for all 99 lexemes in the sample contains 29,912,050 tweets. The first tweet dates from May 5, 2006 and involves the neologism *tweeter*, the last tweet in the collection is from December 31, 2018, and includes *dotard*.

4 METHODS

I processed the dataset to remove duplicates, tweets that do not contain tokens of the target neologism in the tweets' text body. This was mostly relevant in cases where Twitter returned tweets in which the target forms were only part of usernames or URLs.² Hashtag uses were included in the analysis. Retweets were excluded, since the data did not provide reliable information about retweeting activity for the social network analysis. The resulting dataset contains about 30 million tweets, and each tweet contains at least one instance of the 99 neologism under investigation.

To investigate the diffusion of these lexemes in terms of usage frequency, I use time-series of the neologisms' frequency of occurrence over time. I binned the number of tweets per lexeme in monthly intervals to weaken uninterpretable effects of daily fluctuations in use, and to achieve a reasonable resolution to compare the use of all lexemes, which differ according to their overall lifespan. I visualize the resulting time series as presented in **Figure 2**.

To capture different degrees of stability vs. volatility in the use of neologisms over time, I calculated the coefficient of variance for all time series. The coefficient of variance (c_v) is a measure of the ratio of the standard deviation to the mean: $c_v = \frac{\sigma}{\mu}$. Higher

values indicate higher degrees of variation in the use of a neologism, which is typical of topical use of words such as *burquini*; lower values indicate relatively stable use of words such as *twitterverse*.

To investigate the diffusion across social networks over time, I subset the time series into four time frames of equal size, relative to the total period of diffusion observed for each neologism. I set the starting point of diffusion to the first week in which there were more than two interactions which featured the target lexeme. This threshold was introduced to distinguish early, isolated ad-hoc uses of neologisms by single speakers from the start of accommodation processes during which new words increasingly spread in social networks of users on Twitter. This specific limit was determined and validated empirically by systematically testing different combinations of threshold values for the offset of number of users and interactions among early users. Setting a low minimum level of interactions per week proved to reduce distortions in the size of time windows, and enabled a more robust coverage of the relevant periods of diffusion. For each neologism, I divided the time window from the start of its diffusion to the end of the period covered by the dataset into four equal time slices that are relative to the varying starting points of diffusion for all words in the sample. The starting points of each time frame are marked by dashed vertical lines in the usage frequency plots presented below (**Figure 2**).

To investigate the social dynamics of diffusion over time, I generated social networks graphs for each of these subsets. Nodes in the network represent speakers who have actively used the term in a tweet and speakers who have been involved in usage events in the form of a reply or a mention in interaction with others. The resulting graphs represent networks of communicative interaction. Communities are formed based on the dynamic communicative behaviour observed, rather than on information about users' social relations as found in follower–followee networks. This methodology is supported by previous research, which suggests that interactional networks of this kind are better indicators of social structure, since the dynamic communicative behaviour observed is more reliable and socially meaningful than static network information (Goel et al., 2016; Huberman et al., 2008). While users often follow thousands of accounts, their number of interactions with others provides a better picture of their individual social networks, which are much more limited in size (Dunbar 1992).

To construct the networks, I extracted users and interactions from the dataset to build a directed graph.³ Nodes in the graph correspond to individual Twitter users, edges represent interactions between users. I captured multiple interactions between speakers by using edge weights, and I accounted for active vs. passive roles in interaction by using directed edges. I assessed the social diffusion of all neologisms quantitatively by generating and comparing several network metrics, and I

²The post-processing and all quantitative analyses were performed in R Core Team (2018), and the source code is available on GitHub: <https://github.com/wuqui/sna>.

³I used several R packages (R Core Team 2018) from the *tidyverse* library collection (Wickham et al., 2019) for the network pre-processing; *igraph* and *tidygraph* were used for constructing the networks and for calculating network metrics.

produced network visualisations for all subsets for more detailed, qualitative analyses.

On the graph level, I rely on the measures of *degree centralization* and *modularity* to quantify the degree of diffusion for each subset. Degree centralization (Freeman 1978) is a graph-level measure for the distribution of node centralities in a graph. Nodes have high centrality scores when they are involved in many interactions in the network and thus play a 'central' role in the social graph of users. The degree centrality of a graph indicates the extent of the variation of degree centralities of nodes in the graph. A graph is highly centralized when the connections of nodes in the network are skewed, so that they center around one or few individual nodes. In the context of diffusion, the graph of a neologism tends to have high centralization in early stages when its use is largely confined to one or few centralized clusters of speakers. Diffusion leads to decreasing centralization when use of the term extends to new speakers and communities and the distribution of interactions in the speech community shows greater dispersion.

The normalized degree centralization of a graph is calculated by dividing its centrality score by the maximum theoretical score for a graph with the same number of nodes. This enables the comparison of graphs of different sizes, which is essential for drawing comparisons across lexemes in the present context. The neologisms under investigation differ with regard to their lifespan and usage intensity, resulting in substantial quantitative differences in network size. This needs to be controlled for to allow for an investigation of structural differences of the communities involved in their use.

Modularity (Blondel et al., 2008) is a popular measure for detecting the community structure of graphs. It is commonly used to identify clusters in a network and provides an overall measure for the strength of division of a network into modules. In the social context, this corresponds to the extent to which the social network of a community is fragmented into sub-communities. Networks with high modularity are characterized by dense connections within sub-communities, but sparse connections across sub-communities. In the context of the spread of new words on Twitter, diffusion leads from use limited to one or few densely connected communities to use in more and more independent communities. This is reflected by higher degrees of modularity of the full graph representing the speech community as a whole. Modularity complements degree centralization since it provides additional information about the number and size of sub-communities who use the target words. I rely on the modularity algorithm to perform community detection, and I visualize the eight biggest communities in each graph by colour.

Since modularity is sensitive to the number of edges and nodes in a graph and thus cannot provide reliable results for comparing graphs of different size, I use degree centralization to analyse diffusion over time, and to assess differences in degrees of diffusion between lexemes on the macro-level. Its conceptual clarity and reliable normalization allow for more robust comparisons on the macro-level.

For visualizing network graphs, I rely on the Force Atlas 2 algorithm (Jacomy et al., 2014) as implemented in *Gephi* (Bastian

et al., 2009). Force Atlas 2 is a force-directed algorithm that attempts to position the graph's nodes on a two-dimensional space such that edges should be of similar length and there should be as little overlap between edges as possible. In the present social network graphs, the algorithm places nodes (speakers) closer to each other if they have one or more edges connecting them (communicative interactions in the form of replies and mentions). Attempts to evaluate and compare these visualisations with results obtained from different algorithms such as Multi-Dimensional Scaling and Kamada Kawai showed similar results across methods for parts of the dataset, but could not be used for the full dataset due to the computational complexity involved in the generation of large-size graphs of high-frequency neologisms. Force Atlas 2 is particularly well-suited for handling social networks in big data contexts and has been widely applied in network science approaches to Twitter data (Bruns 2012; Bliss et al., 2012; Gerlitz and Rieder 2013).

To assess and visualize the influence of individual users in the social network, I use the PageRank algorithm (Brin and Page 1998). PageRank assesses the importance of nodes in a network based on how many incoming connections they have. It was initially used to analyse the importance of websites on the World Wide Web, but it is also frequently applied to determine the influence of agents in social networks (e.g., Halu et al., 2013; Pedroche et al., 2013; Wang et al., 2013). In the present context, PageRank assigns higher scores to speakers who receive more incoming replies and mentions, which I visualise by bigger node sizes in the network graphs. To account for varying degrees of strength in the connection between users, I use edge weights for repeated interactions, visualised by the edges' width in the graphs.

5 RESULTS

5.1 Frequency-Based Measures of Diffusion

5.1.1 Overall Usage Frequency

As described in Section 2.1, successful diffusion involves an increase in the number of speakers and communities who know and use a new word. The degree of diffusion of new words is often approximated by usage frequency, i.e., by how many times speakers have used a given word in the corpus. The most fundamental way of using this information is to aggregate usage counts and to rely on the total number of uses observed. The underlying assumption is that neologisms that have been used very frequently in the corpus are likely to be familiar to a large group of speakers who have actively produced the observed uses ('corpus-as-output') or have been passively exposed to these neologisms ('corpus-as-input') (Stefanowitsch and Flach 2017). Aggregating all instances of usage to total counts is taken to represent the total amount of exposure or active usage, indicating the degree of conventionality in the speech community. In the following, I will use this most basic measure of diffusion as a baseline before I zoom in to get a more differentiated picture of the temporal and social dynamics of diffusion.

The present sample of neologisms covers a broad spectrum of usage frequency. Tables 1–4 presents the candidates under investigation in four groups: six examples around the

TABLE 1 | Total usage frequency (FREQ) in the corpus. Most frequent lexemes.

Lexeme	FREQ
tweeter	7,367,174
fleek	3,412,807
bromance	2,662,767
twitterverse	1,486,873
blockchain	1,444,300
smartwatch	1,106,906

TABLE 2 | Total usage frequency (FREQ) in the corpus. Examples around the median.

Lexeme	FREQ
white fragility	26,688
monthiversary	23,607
helicopter parenting	26,393
deepfake	20,101
newsjacking	20,930
twittosphere	20,035

TABLE 3 | Total usage frequency (FREQ) in the corpus. Least frequent lexemes.

Lexeme	FREQ
microflat	426
dogfishing	399
begpacker	283
halfalogue	245
rapugee	182
bediquette	164

TABLE 4 | Total usage frequency (FREQ) in the corpus. Case study selection.

Lexeme	FREQ
alt-right	1,012,150
solopreneur	282,026
hyperlocal	209,937
alt-left	167,124
upskill	57,941
poppygate	3,807

minimum, around the median, and around the maximum total usage frequency observed in the corpus, as well as six words that will serve as case studies in the following sections. These cases reflect a set of prototypical examples of different pathways of diffusion, and I will use these cases to illustrate more detailed characteristics of diffusion before I present the general patterns found for the full sample of neologisms.

The grouping of neologisms on the basis of their total usage frequency presented in **Tables 1–4** largely seems to fit intuitions about diverging degrees of conventionality between the frequency-based groups listed in **Tables 1–3**. Neologisms such as *blockchain* and *smartwatch*, which are probably familiar to most readers, can be assumed to be more conventional than neologisms from the low end of the frequency continuum such as *dogfishing* ('using a dog to get

a date') or *begpacker* ('backpackers funding their holidays by begging').

However, total frequency counts only provide a limited picture of diffusion since they are insensitive to temporal dynamics of usage. Neglecting temporal information about the lifespan and the period of active use of a new word can distort the quantitative assessment of its degree of conventionality in two directions. Firstly, it carries the danger of overestimating the status of words such as *millennium bug*⁴, whose total usage frequency largely goes back to a short period of highly intensive usage, after which they fall into disuse, become unfamiliar to following generations of speakers, eventually becoming obsolete. Secondly, total counts can underestimate the conventionality of words such as *coronavirus*, which have already become familiar to the vast majority of speakers, but show comparatively moderate total frequency counts, since they have started to diffuse only fairly recently.

Among the most frequent neologisms presented in **Table 1**, words such as *twitterverse* and *blockchain*, for example, have similar total frequency counts, but differ significantly with regard to their temporal usage profiles. The neologism *twitterverse* has been in use ever since the start of Twitter, while the diffusion of the much younger *blockchain* only started in 2012. Despite its shorter lifespan, *blockchain* accumulated roughly the same number of uses, but shows significantly higher usage intensity in the more recent past, and can be assumed to be familiar to bigger parts of the speech community.

Similar effects are even more pronounced in the remaining groups of neologisms, since words from the lower ranges of the frequency spectrum are typically affected more strongly by temporal variation in their use. In the following sections, I will include temporal information to get a more fine-grained picture of diffusion.

5.1.2 Cumulative Frequency

Visualising the cumulative increase in usage frequency of new words complements total counts by taking into account the temporal dynamics of their usage intensity over time. **Figure 1** presents this information for the case study selection.

While the end points of the trajectories in **Figure 1** mark the target words' total frequency counts as shown in **Table 4**, the offsets and slopes of the trajectories of usage frequency reveal additional characteristics about differences in their diffusion patterns. The selected neologisms differ regarding their total lifespan observed, which is indicated by diverging starting points of diffusion. The term *hyperlocal*, for example, is the oldest new word among the selected neologisms, and it is commonly used to refer to information that has a strong focus on local facts and events. While it was hardly used in the first years of Twitter, it started to increase in its use in 2009 and was added to the OED's Third Edition in 2015. Around this time, the neologism *solopreneur* only started to significantly increase in its use. A blend of *solo* and *entrepreneur*, it keeps a low, flat trajectory

⁴The neologism *millennium bug* was used to refer to anticipated technical problems caused by inconsistent formatting of timestamps at the turn of the century.

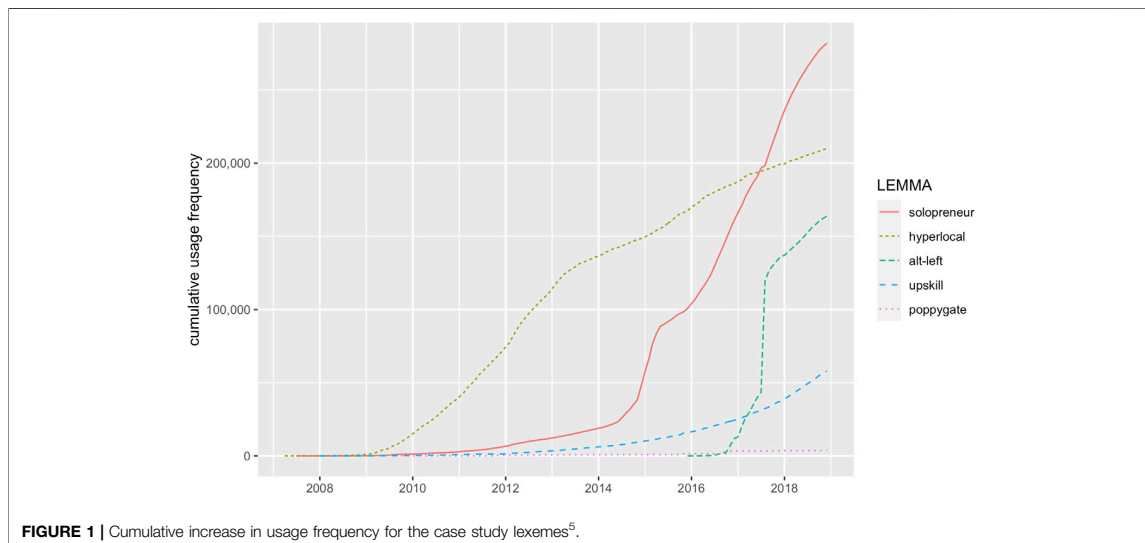


FIGURE 1 | Cumulative increase in usage frequency for the case study lexemes⁵.

of sporadic use for about 7 years after its first appearance in the corpus. The first two attestations in the corpus indicate the sense of novelty and scepticism towards the term in its early phases:

- 1) I'm trying to figure out if I like the term 'solopreneur' I just read (July 27, 2007).
- 2) hmmm new word added to my vocab = 'solopreneur' !! (January 6, 2008).

Most speakers increasingly 'like the term' and 'add them to their vocabulary' only much later, after 2014, when the phenomenon of individual entrepreneurship attracts increasing conceptual salience in the community, which seems to be both reflected and propagated by the publication of several self-help books for entrepreneurs in this year, which all explicitly use this new term in their titles (e.g., the popular guide *Free Tools for Writers, Bloggers and Solopreneurs* by Banes (2014)). The following short, but intense period of use results in a higher overall number of uses for *solopreneur* as compared with *hyperlocal*, even though the use of the latter term shows a longer lifespan of continual use⁵.

In addition to differences in age, the slopes of the cumulative trajectories in **Figure 1** indicate differences regarding the dynamics of diffusion underlying the aggregated total number of uses over time.

Neologisms such as *hyperlocal* and *upskill* ('to learn new skills') show a steady, gradual increase in usage frequency over longer periods of time. By contrast, the use of other candidates such as *solopreneur* and *alt-left* is much less stable and less evenly distributed over time.

⁵*alt-right* was omitted from this plot because its high usage frequency would have inhibited the interpretability of the other lexemes; its frequency over time is presented in **Figure 3D**.

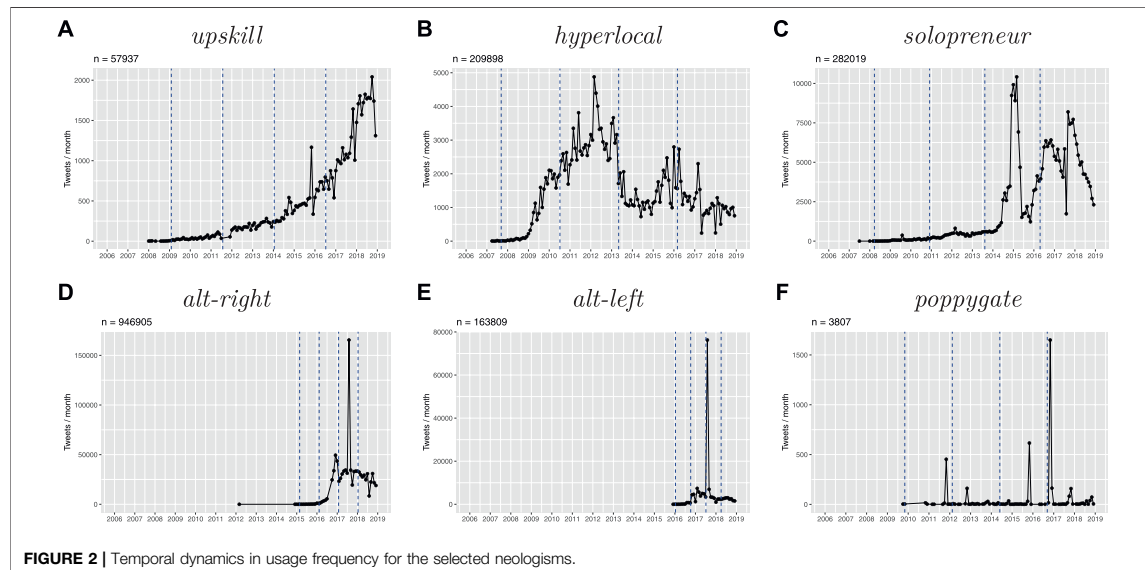
In the case of *solopreneur*, we observe a big spike in frequency following its increased popularity in the entrepreneurial community in 2014. While it shows the highest total frequency count in **Figure 1**, the majority of its uses fall into the second part of its observed lifespan.

An even shorter and steeper increase can be seen in the use of *alt-left*, which is the youngest neologism to enter the scene at the end of 2015. *alt-left* was coined as a counterpart to the term *alt-right*. The latter neologism is a shortening of *Alternative Right*, introduced by the white-supremacist Richard Spencer in 2010 as a new umbrella term for far-right, white nationalist groups in the United States. Facing substantial criticism for racist attitudes and actions, proponents of this far-right political camp coined and attempted to propagate the derogatory term *alt-left* to disparage political opponents. Despite its late appearance in the corpus, *alt-left* occurs in a total of 163,809 tweets, which places it in the medium range of the sample in terms of total frequency counts. However, its trajectory in **Figure 2** shows that the majority of its uses go back to a single period of highly intensive use in the second half of 2017, soon after which it slows down considerably.

The cumulative increase in usage intensity of the selected neologisms illustrates that similar total frequency counts of neologisms can be the product of highly different trajectories of diffusion. These data complement total counts in that they show differences in the total lifespan and in the intensity with which a given neologism was used over time – types of information that are highly relevant for assessing the degree to which they have spread in the speech community.

5.1.3 Usage Intensity

Going beyond cumulative counts, absolute usage frequency counts provide a more fine-grained view of the temporal dynamics of diffusion. Most importantly, analysing usage



intensity highlights to what degree new words are being used consistently over time. **Figure 2** presents this information for the selected neologisms. In the following section, I will illustrate prototypical differences by referring to the selected cases, before I discuss the results for the full sample⁶.

The absolute frequency plots confirm differences regarding the lifespan and dynamics of usage intensity among the neologisms discussed above. In terms of lifespan, **Figure 2** shows that *upskill* and *hyperlocal* are much older than *alt-right* and *alt-left*. The absolute counts also highlight the fact that while there is a low level of use of *solopreneur* since 2007, its main period of diffusion starts much later, in 2014, with a subsequent spike in usage intensity.

5.1.4 Volatility

Besides, the absolute frequency counts over time provide a more detailed picture of the temporal dynamics of use. While the cumulative counts in **Figure 1** suggest more gradual trajectories, the plots in **Figure 2** indicate that the selected neologisms differ significantly in terms of the volatility with which they are used in the corpus.

The neologism *upskill* shows the smoothest trajectory of diffusion among the candidate neologisms in **Figure 2**. Aside from two smaller spikes, at the end of 2016 and 2018, it has gradually increased in its use since its first attestation in the corpus at the end of 2007. Neither its frequency counts, nor the corpus data suggest that its spread was triggered or propagated by specific topical events or by the determining influence of individual users or user groups. After a long period of very slow, but consistent increase in frequency, its diffusion has

accelerated in recent years. While its future remains uncertain, its previous trajectory resembles most closely the earlier phases of spread as predicted by S-curve models.

While *hyperlocal* also exhibits a marked increase in usage frequency during its earlier stages, its peak in popularity is followed by a decline in use, after which it settles at a relatively stable level of about 1,000 tweets per month. This coincides with the OED's decision to take up *hyperlocal* in its 2015 edition. Despite fluctuations, *hyperlocal* has been used relatively consistently in the recent past.

The neologism *solopreneur* has been in use since 2007 and shows an overall increase in usage frequency, but its use fluctuates more strongly than that of *hyperlocal*. After its initial peak around 2015, which coincides with the release of several self-help books featuring the term, its frequency plummets, becomes less stable, and shows an overall downward trend.

As was mentioned above, *alt-right* and *alt-left* are closely related. Both terms show high levels of volatility in their usage frequency. The former, older term shows significant diffusion in 2016, particularly in the period leading up to Donald Trump's election, after which *alt-right* remains in consistent use to a relatively high degree, at about 25,000 tweets per month. Its counterpart, *alt-left*, enters the scene much later, during the infamous Charlottesville Rally in 2017, whose topical effect causes a huge spike in the use of both terms. However, unlike *alt-right*, which reverts to its previous usage intensity, the use of *alt-left* seems to largely disappear from Twitter in the aftermath of the event.

The final example among the selected candidates, *poppygate*, also exhibits high degrees of volatility, and it features the most distinctive pattern of spikes in its usage intensity. Unlike the single topical spike for *alt-right* and *alt-left*, its use follows a recurrent, regular pattern: speakers use it almost exclusively

⁶Neologisms with a lifespan shorter than 1 year and/or less than 2,000 tweets ($n = 5$) were excluded since the coefficient of variation does not provide robust measures for these infrequent, short-lived outliers.

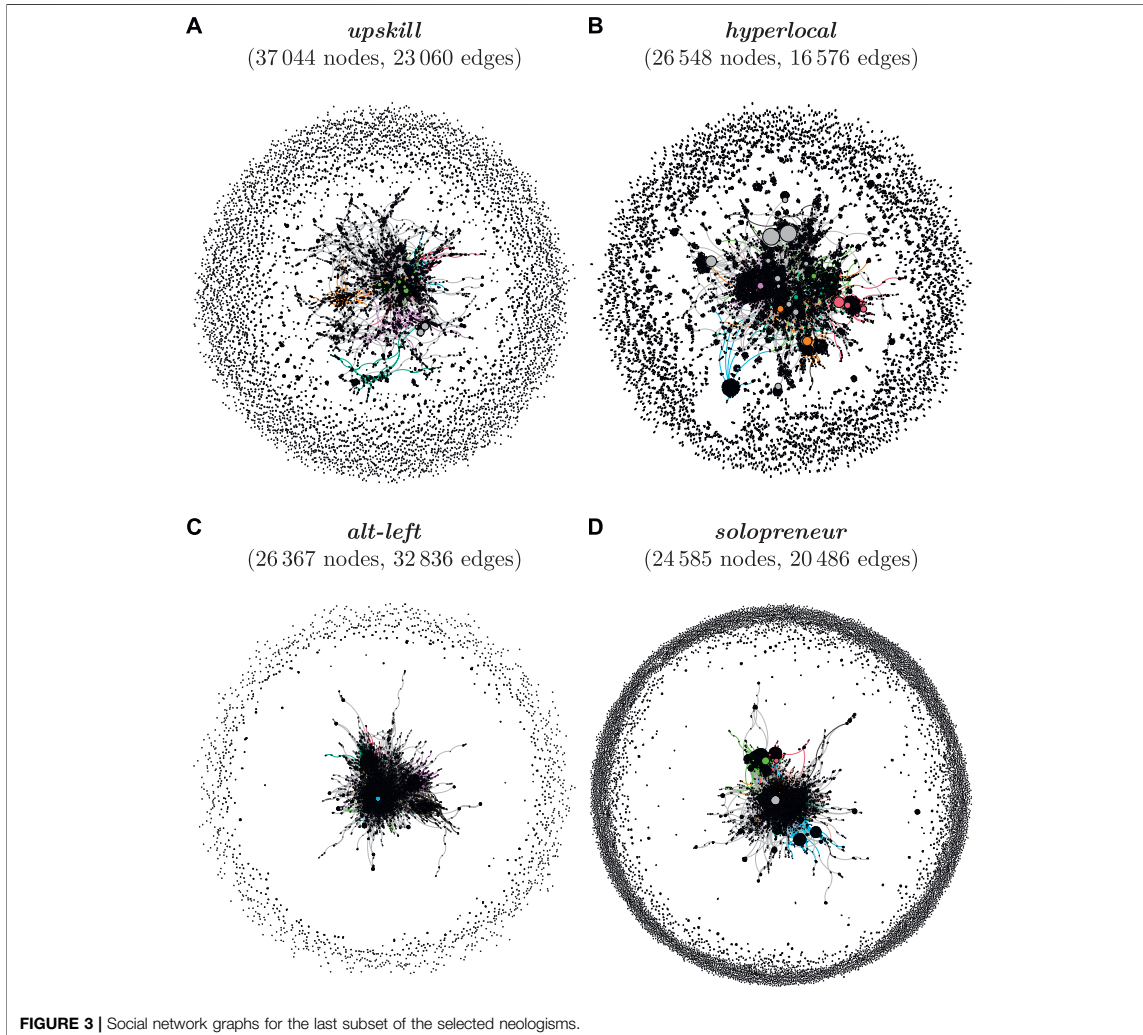


FIGURE 3 | Social network graphs for the last subset of the selected neologisms.

around Remembrance Day, which takes place in November. The term *poppygate* represents a last category of neologisms in the sample, which show strong fluctuations in usage intensity, but for which these patterns follow a regular temporal pattern.

To quantify the degree to which neologisms are used with consistent frequency over time, I calculate and compare the coefficients of variation for each neologism in the sample. This metric captures the overall volatility in usage frequency of words over their lifespan relative to their average frequency of occurrence in the corpus. Tables 5–7 presents the coefficients of variation for the selected neologisms, as well as for the top and bottom six neologisms that show the highest and lowest degrees of variation in the sample.

The results in Tables 5–7 show that the sample covers a broad spectrum of volatility in usage frequency. Among the neologisms that were used the most consistently, i.e., exhibit the lowest

degrees of variation, we find words whose frequency-based measures suggested high degrees of conventionality. For example, *twitterverse* is listed among the most frequent neologisms in Table 1 and is also one of the oldest neologisms, with its first attestation in the corpus dating back to December 19, 2006.

TABLE 5 | Coefficients of variation (VAR) for the selected neologisms.

Lexeme	VAR
hyperlocal	0.98
upskill	1.14
solopreneur	1.20
alt-right	1.81
poppygate	4.75
alt-left	5.31

TABLE 6 | Coefficients of variation (VAR) for the six neologisms with the lowest scores in the sample⁵.

Lexeme	VAR
followership	0.71
lituation	0.72
twitterverse	0.72
detweet	0.74
remoaners	0.76
twittersphere	0.77

TABLE 7 | Coefficients of variation (VAR) for the six neologisms with the highest scores in the sample.

Lexeme	VAR
upskirting	9.39
youthquake	6.32
alt-left	5.31
birther	5.00
poppygate	4.75
cherpumple	4.69

By contrast, the group of lexemes that show the highest degree of volatility in usage frequency is comprised of neologisms with lower degrees of conventionality, which are generally less frequent and were coined more recently. Notably, topical spikes play a crucial role in the diffusion processes of all examples in this category: the diffusion of *alt-left* and *birther*⁷ was promoted by extralinguistic political events, *upskirting*⁸ and *youthquake*⁹ were advanced through increased metalinguistic salience after they were added to the OED and awarded Word of the Year 2017 by Oxford University Press. Both *poppygate* and *cherpumple*¹⁰ exhibit recurrent topicality, and are typically only used in the contexts of their seasonal relevance in autumn and winter.

The selected neologisms cover the spectrum of volatility in usage frequency found in the full sample of neologisms, and the coefficients of variation represent quantitative measures which reflect the differences in volatility between the selected neologisms visualised in **Figure 2** and discussed above. The frequency-based analysis of the three neologisms discussed above demonstrates that usage frequency counts, particularly when combined with an analysis of their underlying temporal dynamics, can help to approximate the spread and success of neologisms to a certain degree. However, the results also point to

⁷Proponent of the 'birther movement', a conspiracy theory which claims that President Obama's birth certificate was forged and that he was not born in the United States.

⁸'The habit or practice of taking upskirt photographs or videos' (OED).

⁹'A significant cultural, political, or social change arising from the actions or influence of young people' (<https://languages.oup.com/word-of-the-year/2017/>).

¹⁰Cherpumple is short for cherry, pumpkin and apple pie. The apple pie is baked in spice cake, the pumpkin in yellow and the cherry in white (<https://en.wikipedia.org/wiki/Cherpumple>); typically consumed during the holiday season in the US.

TABLE 8 | Degree centrality scores (CENT) for the selected neologisms; the scores are based on the most recent time slice for each neologism in the corpus.

Lexeme	CENT
upskill	0.0021
hyperlocal	0.0085
alt-right	0.0144
alt-left	0.0238
solopreneur	0.0523
poppygate	0.0566

TABLE 9 | Degree centrality scores (CENT) for the six lexemes with the lowest scores in the sample; the scores are based on the most recent time slice for each neologism in the corpus.

Lexeme	CENT
baecation	0.0005
fleek	0.0009
ghosting	0.0013
man bun	0.0016
big dick energy	0.0018
twittersphere	0.0020

TABLE 10 | Degree centrality scores (CENT) for the six lexemes with the highest scores in the sample; the scores are based on the most recent time slice for each neologisms in the corpus.

Lexeme	CENT
rapugee	0.2580
leivdrome	0.2373
kushnergate	0.2309
dronography	0.1530
dotard	0.0979
ecocide	0.0922

substantial limitations of frequency-based approaches to studying diffusion.

The present data demonstrate considerable variation in the degrees of diffusion of neologisms with similar frequency of occurrence in the corpus. Total frequency counts alone would predict high degrees of diffusion for neologisms such as *alt-left*, for example. However, its usage history reveals that its use largely goes back to a short period of high usage intensity linked to a specific topical event. The term's background suggests that it might not have spread far beyond one particular community of speakers. Such potential distortions of frequency-based measures could partly be resolved by in-depth analyses of temporal usage profiles combined with insights from corpus data and extralinguistic events. However, these in-depth analyses of diffusion are not possible through a systematic frequency-based analysis alone, and they cannot be extended to the large-scale study of larger samples of neologisms. Hence it remains unknown to what degree frequency-based metrics adequately capture social pathways of diffusion. In the following section, I will complement the frequency-based approach by social network analyses to get a more differentiated view of the sociolinguistic aspects of diffusion.

5.2 Social Networks of Diffusion

As described in **Section 4**, the social network analysis is based on the interactions between all speakers who have used the neologisms in the sample. Speakers are represented as nodes in the network graph, and interactions between users in the form of replies or mentions are represented as edges. The network structure of the resulting graphs allows analysing the degree to which the target neologisms have diffused in these networks. To monitor diffusion over time, I split the observed lifespan of each neologism into four equally-sized time slices. These time windows are marked by dashed vertical lines in **Figure 2**. I then generated network graphs for each time window for each neologism in the sample to analyse the individual pathways of diffusion over time and to compare degrees of diffusion between all neologisms in the sample.

5.2.1 Degrees of Diffusion

As discussed in **Section 4**, I mainly rely on degree centralization as a quantitative measure of diffusion. I consider increasing diffusion to be reflected by decreasing degree centralization of the graph, thus lower values of centrality indicate higher degrees of diffusion across social networks.

For example, the social graph users of a new word shows high centralization in early stages when its use is largely confined to one or few centralized clusters of speakers. When increasing diffusion extends the use of the term to new speakers and communities, the distribution of interactions in the speech community shows greater dispersion, which should be reflected by lower centrality scores for the social network of speakers.

Tables 8–10 report the degree centrality scores for the selected neologisms and for six lexemes with the highest and lowest scores in the sample

The neologisms with the lowest scores for degree centrality are also among the most frequent lexemes in the sample. Overall, frequency and centrality generally tend to produce similar results when used to assess degrees of diffusion. This shows usage frequency and social diffusion correlate, as one might expect. Notable deviations exist, however, and will be further discussed in **Section 5.3**.

Correspondingly, the neologisms with the highest centrality scores rank among the least frequent candidates in the sample. Notable trends among lexemes with high centrality scores are that they tend to be more recent (e.g., *dronography*¹¹) and/or to exhibit high degrees of volatility (e.g., *ecocide*¹²). Moreover, this group includes political terms such as *Kushnergate*¹³ and *rapugee* which are controversially discussed on the left and right ends of the political spectrum. For example, *rapugee* is a derogatory term which was coined after sexual assaults by refugees during New Year's Eve 2015/16 in Cologne, Germany. Previous work has shown that this term was

consciously coined and propagated by a closely connected community of far-right activists to disparage refugees, and that its use on Twitter and on the Web has remained largely limited to these communities (Würschinger et al., 2016). This low degree of diffusion is reflected by the low centrality score for *rapugee*.

The following sections use network visualisations to provide a detailed, partly qualitative analysis of the diffusion for the selected cases to illustrate the social dynamics captured by the quantitative measure of centralization as an indicator of diffusion. The examples represent prototypical pathways based on centralization scores. The in-depth analysis of the social dynamics at play is guided by the detection of communities using modularity clustering (**Section 4**). The algorithm identifies the eight largest communities in each graph, visualised by colour. Moreover, I rely on the PageRank algorithm (**Section 4**) to assess the importance of users in the network, visualised by node colour. I use manual inspection of user accounts to validate and further investigate the role of these communities and influential users in the selected diffusion processes.

The centrality scores for the selected neologisms cover a broad spectrum of degrees of diffusion, as can be seen in **Table 8**. **Figure 3** presents the full network graphs for four of the selected cases to illustrate differences in the social networks of speakers which are captured by centrality scores.¹⁴ The network graphs in **Figure 3** are sorted according to their degrees of social diffusion—as measured by centrality scores—from (a) to (d). Note that the number of nodes in each graph is very similar, differences between the visualized structure of network graphs are thus due to differences in the underlying social structure of communities rather than a mere function of differences in network size.

The neologism *upskill* exhibits the highest degree of diffusion, which is reflected by the highest degree of dispersion of nodes across the graph in **Figure 3A**. At the center of the graph, we find a relatively large cluster of speakers who are only loosely connected. Many of these speakers are connected via their affiliations to the world of business, where the term *upskill* is most commonly used. However, on the whole, the use of *upskill* is not limited to a coherent, closely-connected community. The majority of nodes appear towards the fringes and have no connections to the rest of the graph. Speakers use the term independently from each other, without being unified in their motivations to use the term by a common affiliation with a certain community of practice. The social network of *upskill* thus shows an advanced degree of diffusion.

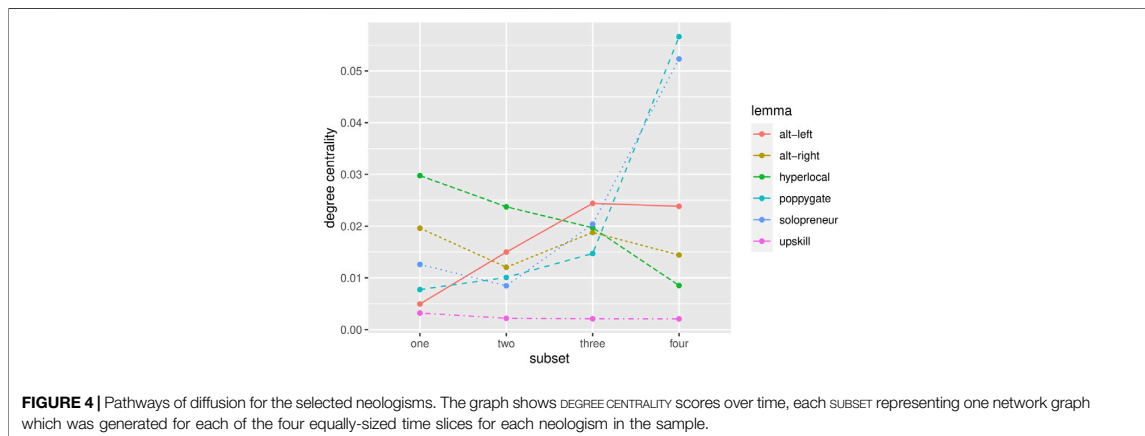
The graph for *hyperlocal* in **Figure 3B** also shows a high degree of social diffusion, but its use depends more strongly on a central community of users. This core sub-network of speakers forms several smaller clusters which can be linked to certain domains of interest such as journalism, business, and startups, in which the term is most popular. Notably, we observe a stronger role of

¹¹Dronography is the science, art and practice of creating durable images or video by recording light or other electromagnetic radiation by means of a drone flying around or above a certain scene (Urban Dictionary).

¹²the destruction of large areas of the natural environment as a consequence of human activity (Merriam Webster Online Dictionary).

¹³Referring to a political scandal involving Trump's senior adviser Jared Kushner allegedly meeting Russian officials.

¹⁴The network graphs for *alt-right* and *poppygate* were omitted as their difference in network size does not allow for comparative analyses (*alt-right*: 2,74,686 nodes, *poppygate*: 2473 nodes).



individual user accounts such as influencers and marketing agencies, which is illustrated by bigger node sizes (representing high PageRank scores). Yet, as in the graph for *upskill*, the majority of occurrences of *hyperlocal* can be traced back to a large number of speakers from a diverse set of sub-communities, which can be interpreted as a sign of advanced diffusion.

The social graph for *alt-left* shows very limited diffusion of the term. Almost all of its use can be traced back to one closely-connected community of users. This core community of users demonstrates typical characteristics of an echo chamber in that it is dense and features strong ties within the community, but has few weak ties connecting it to the rest of the social graph. This observation is in line with the socio-political background of the term, which was coined and propagated by far-right activists in an attempt to unify political efforts ('*Unite the Right Rally*') and to distance themselves from and protest against the political left. Inspection of the network reveals that the most influential node in the network is Donald Trump. His use of the term was followed by a sharp increase in usage intensity in the course of the Charlottesville Rally in August 2017. The high degree of social compartmentalization in the use of *alt-left* is also reflected in the ratio between the number of nodes and edges in its graph, which confirms that its community of speakers is much more closely connected than that of the remaining neologisms¹⁵. Notably, the same applies to the community of *alt-right*, which occupies the opposite pole of the political spectrum. The results for these two terms are in line with previous work reporting effects of political polarization in online social networks for these political communities (Sunstein, 2018). Overall, *alt-left* thus shows a low degree of diffusion. It has received significant popularity in certain parts of the speech community, but its use remains strongly limited to these communities.

¹⁵The numbers of edges per node for all selected cases in descending order: *alt-right*: 1.49, *alt-left*: 1.24, *solopreneur*: 0.83, *hyperlocal*: 0.62, *upskill*: 0.62, *poppygate*: 0.53.

Lastly, the social network of speakers using the term *solopreneur* also shows limited diffusion. A significant proportion of its use comes from a diverse set of individual speakers and micro-communities, which are placed at the fringes of the graph. However, similar to the social graph for *alt-left*, a relatively well-connected, large core of speakers is responsible for the majority of its use in the corpus. Moreover, unlike the example of *alt-left*, this central community of users is in turn dominated by the high centrality of a small number of individual accounts. Inspecting the network of users reveals that these 'influencers' are all either proud, self-proclaimed solopreneurs, or coaches and agencies that are using the term to promote their services to aspiring entrepreneurs. Overall, *solopreneur* has achieved significant popularity within certain communities, but its use in these communities is unevenly distributed and depends strongly on a small number of individual users. The term does not show signs of advanced diffusion since its use is largely limited to certain individual speakers and communities of practice.

In summary, the social networks of speakers reveal significant differences in the degrees of social diffusion for the neologisms in the present dataset, as observed in the period leading up to the cutoff point at the end of 2018.

While the centrality measures generally concur with the frequency-based analysis of the neologisms discussed in Section 5.1, the network metrics and visualisation add information by providing a more detailed picture of degrees of social diffusion and highlight cases for which the social dynamics of diffusion diverge from what could be observed by relying on usage frequency alone.

5.2.2 Pathways of Diffusion

To investigate the pathways of social diffusion, Figure 4 presents the degree centrality scores for the selected neologisms over time. The scores for Subset 4 represent the final degrees of diffusion as presented in Table 8. The corresponding network graphs for this stage were presented in Figure 3. The centrality scores for the preceding subsets now add information about the diffusion history of these neologisms. The diverging trajectories of centralization over time indicate significant changes over time as well as differences in the pathways of diffusion between neologisms.

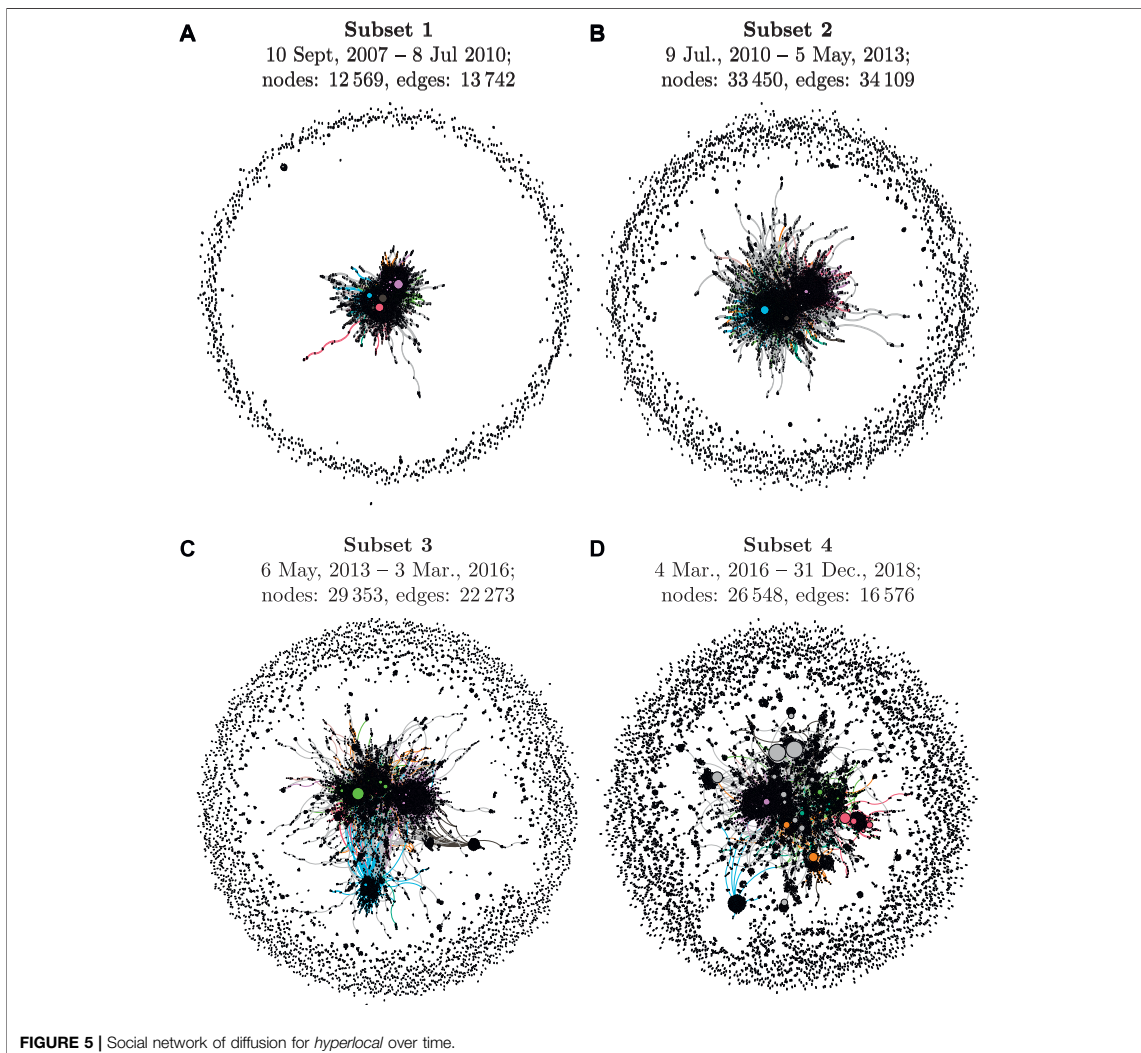


Figure 5 presents the full network graphs for all stages of diffusion for the term *hyperlocal* to illustrate the social dynamics underlying the quantitative measures.

Both the quantitative measure in **Figure 4** and the network visualizations in **Figure 5** indicate that *hyperlocal* shows increasing, successful diffusion over time. Its use is relatively centralized in its earlier stages, which can be seen from the fact that most speakers who have used the term are closely connected in the social graph in the first quarter of its observed lifespan. Inspecting the most influential speakers and sub-communities in the network (based on PageRank and Modularity scores) reveals that *hyperlocal* is mainly used by a relatively small community of individual journalists in the first subset, who are early adopters in trying to target news to

local audiences and use the term very frequently to label this new approach.

In Subset 2, the community of journalists grows and starts to include also bigger news outlets such as *The Guardian*. Additionally, a new community of practice adopts the term: several marketing agencies start promoting their services using the term *hyperlocal*. At this point, the usage intensity of the term peaks, as was demonstrated in **Figure 3B**. However, the social network data indicate that at this point its use is still mainly the product of high popularity and usage intensity within a small number of dense sub-communities rather than a sign of advanced diffusion across bigger parts of the speech community.

The network graphs show that the social diffusion of *hyperlocal* is only significantly advanced in the last two stages.

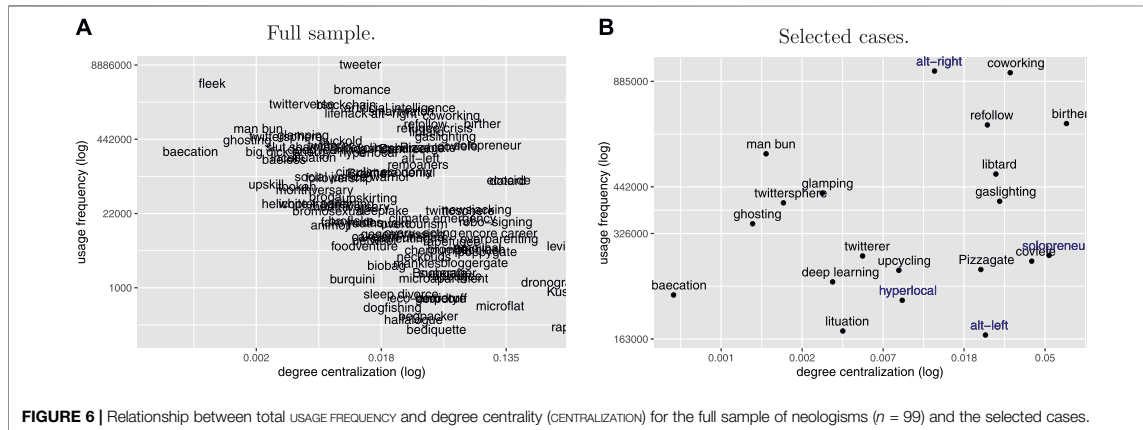


TABLE 11 | Correlations of 'degree centralization' (CENTRALITY) with the variables total usage frequency (FREQUENCY), coefficient of variation (VOLATILITY), and observed lifespan in the corpus (AGE) for the full sample of neologisms ($n = 99$) using Spearman's correlation coefficient (Spearman 1961)¹⁷.

	ρ	p
Frequency	-0.44	<0.001
Age	-0.29	0.004
Volatility	0.28	<0.001

While we see only few weak ties during the earlier stages of its use, the term now increasingly diffuses beyond its early adopters. Inspecting the network reveals that the use of the term becomes increasingly popular in the world of business and startups as well as the general public on Twitter. The network metrics indicate that individual agents and sub-communities now play a far smaller role in its overall use. While *hyperlocal* shows less usage intensity during these later stages, the network metrics indicate a high degree of diffusion for the second half of its observed lifespan. The timing of its addition to the OED in 2015 supports these observations. The term *hyperlocal* has successfully spread beyond its subcommunities of early adopters, and it seems to be used by a diverse community of speakers from different backgrounds, which renders it a case of advanced diffusion. This process of increasing diffusion for *hyperlocal* is also reflected in its decreasing measures for graph centrality in **Figure 4**.

The remaining cases in **Figure 4** show different pathways of diffusion, both in terms of their overall degree of diffusion and diachronic trajectory. Due to space limitations, I can only provide an overview of their development over time.

Besides *hyperlocal*, the second neologism which exhibits advanced diffusion is *upskill*. In this case, however, we observe little change over time, its degree centrality has been very low since its early attestations in the corpus. This indicates a gradual spread across speakers which is not significantly affected by a small group of influential speakers. The term *upskill* has been used by a wide variety of speakers throughout its observed lifespan and shows the highest degree of diffusion among the selected cases.

By contrast, *solopreneur* and *poppygate* show a negative trend in terms of diffusion. The term *solopreneur* features low degrees of diffusion in its earlier stages, but its use becomes more centralized over time. This is in contrast with its usage intensity over time (**Figure 2**): while its earlier period of moderate use goes back to a decentralized cluster of users, its increase in usage frequency coincides with a narrowing of its user base. As the network analysis in **Figure 3D** demonstrates, it becomes increasingly limited to a relatively small community which shares interest in a small professional niche.

The case of *poppygate* exhibits a similar trend towards increasing centralization. Its temporal dynamics show a pattern of recurrent topical usage (**Figure 2**). The social networks of *poppygate* suggest that while the term was used by a broader audience in its earlier stages, its use in the more recent past goes back to certain communities of speakers for which a specific topical event emerges as a salient occasion to use the term. For example, its most recent spike in usage intensity in November 2016 was caused by a controversy about whether Fifa was right to take disciplinary action against the national teams of England and Scotland after their players wore poppy armbands during a football match between the two nations on 11 November. Protests by the football community caused a spike in usage intensity for *poppygate*, but did not trigger its diffusion beyond this community¹⁷.

Lastly, *alt-right* and *alt-left* show limited degrees of diffusion over their lifespan. While the centrality of *alt-right* remains fairly stable over time, *alt-left* shows increasing centralization. Both terms are strongly tied to the political discourse surrounding the Unite the Right Rally in the United States and consequently exhibit a sharp increase in usage intensity in the course of the event in August 2017 (**Figure 2**). This increase in use is, however, reflected by increased centrality

¹⁷All variables entering the correlation analysis were log-transformed and centred. I report Spearman's correlation coefficients to avoid assumptions about the linearity of the variables involved. I additionally calculated Pearson's correlation coefficients, for which the correlation coefficients are slightly higher: FREQUENCY: $\rho = -0.45, p < 0.001$; AGE: $\rho = -0.38, p < 0.001$; VOLATILITY: $\rho = 0.23, p < 0.001$.

scores for both lexemes in **Figure 4**. This period of highly intense use is thus characterised by relatively smaller rather than larger degrees of diffusion for both lexemes. While the use of *alt-right* reverts to more decentralized use afterwards, the use of *alt-left* remains at this high level of centrality. This seems to confirm the echo chamber effect for *alt-left* discussed in **Section 5.2.1**: the term has become conventional and popular among a community of like-minded individuals, but its use remains limited to this community. Given the extreme, far-right attitudes and political orientations prevalent in this group, the majority of Twitter users do not want to be associated with this community of users. Since the term *alt-left* has become highly indexical of support and membership of this political camp, very few speakers are willing to adopt and use the term.

In summary, studying the temporal dynamics of social networks highlights changes in the use of neologisms over time and reveals different pathways of diffusion in the sample.

5.3 Combining Frequency and Network Information

Having applied the frequency-based and the social network approach to assess the diffusion of the present sample of neologism, this section will combine the results obtained from both approaches and show how they complement each other¹⁶.

5.3.1 Correlations

A first evaluation of the social network approach to diffusion relies on the correlations of degree centrality with the total usage frequency of neologisms, with their volatility, and with their age as observed in the corpus. **Table 11** reports the correlation coefficients for these variables.

Firstly, centrality shows a significant negative correlation with FREQUENCY. This confirms earlier observations in **Section 5.2** which indicated an inverse trend between total usage frequency and centrality. More frequent neologisms show on average higher degrees of diffusion, i.e. increase in frequency correlates with wider spread across the speech community. The fact these two central measures for diffusion correlate can be seen as a cross-validation of both approaches. While external data sources would be needed for a more rigorous evaluation, this overall convergence in results suggests that both metrics capture important aspects of diffusion.

Secondly, the AGE of neologisms in the sample shows a significant negative correlation with centrality. As expected, the use of more recent neologisms tends to still go back to more centralized communities, while neologisms with a longer history of use tend to show more advanced diffusion. Unlike frequency counts, which are directly influenced by the temporal usage history of neologisms, the centrality measure is blind to this

information. The fact that these age effects are captured by degree centrality supports the usefulness of the social network approach.

Lastly, VOLATILITY shows a significant positive correlation with centrality. Again, this result is in line with expectations. Neologisms such as *poppygate*, whose use exhibits substantial temporal variation tend to show lower degrees of diffusion than neologisms such as *hyperlocal*, whose use is more consistent and less dependent on the topical salience of extralinguistic events.

5.3.2 Deviations Between Centrality and Frequency

For a closer analysis of the interactions between these variables beyond correlation coefficients, **Figure 6** presents all neologisms according to their usage frequency and centrality scores. While **Figure 6A** covers the full sample, **Figure 6B** is based on the same data, but zooms in on the frequency range which covers four of the selected cases to provide a clearer view of this section of the sample.

The general trend in the plot confirms the inverse relation captured by the negative correlation coefficient between centrality and frequency. Neologisms with high frequency such as *fleek* have low centrality scores and would thus be assigned a high degree of diffusion by both approaches. The inverse applies to candidates from the lower end of the frequency spectrum such as *microflat*.

However, **Figure 6A** also shows substantial variation between frequency and centrality scores. Notably, the observed deviations are almost exclusively found towards the right of the diagonal trend, i.e., for cases where centrality assumes lower degrees of diffusion than frequency. For example, while *fleek* and *bromance* are assigned similar scores in terms of their usage frequency, their centrality scores suggest a much lower degree of diffusion for the latter neologism. Similar to cases like *solopreneur* and *alt-left*, which were discussed in detail in **Section 5.2.1**, centrality thus provides additional information for cases in which the social network structure indicates that the observed usage intensity overestimates the degree of diffusion of a target neologism. This can arise if its observed uses go back to a disproportionately smaller number of speakers and subcommunities.

Analysing these deviations highlights two main groups among the selected neologisms, for which total usage frequency and social network structure seem to diverge in systematic ways¹⁸. A first group contains neologisms marked by high degrees of volatility in their frequency of use. As shown above, centrality is significantly correlated with volatility. In addition to *poppygate* and *solopreneur*, which were already discussed above, *refollow*, *gaslighting*, *solopreneur*, and *coworking* also show little consistency in their usage. For all of these terms, social diffusion is out of sync with the increase in usage intensity in **Figure 6A**. It thus seems that the social network approach adds an extra layer of information which comes to the fore especially where frequency-based measures overestimate degrees of diffusion due to the strong impact of short periods of highly intensive use of neologisms in certain parts of the speech community.

¹⁶It should be noted that a strict evaluation of both approaches is in principle impossible without external data about the degrees of diffusion for the neologisms under investigation. While such a gold standard for evaluation is inconceivable in the present context, it would be desirable to use additional data sources such as questionnaires, dictionaries or web corpus data for a more rigorous validation of the present approach. This will have to be left for future work.

¹⁸The present dataset does not allow to assess whether the deviations of the two groups that emerge in this analysis are generalisable.

A second, converse group with diverging scores contains neologisms whose use is tied to political communities. The neologisms *alt-right*, *alt-left*, *birther*, *covfefe*, *Pizzagate*, and *Kushnergate* are politically controversial and differ strongly in popularity between political camps. It should be noted that these terms also exhibit considerable volatility in their use. **Figure 6A** shows comparatively lower centrality than frequency scores for these lexemes. Similarly to the cases of high volatility, centrality thus suggests that usage frequency overestimates degrees of diffusion for these cases. While neologisms such as *alt-right* show high frequency counts, the social network analysis reveals that these terms have not spread successfully across communities, and that their use remains limited to certain subcommunities.

5.3.3 Predicting the Success of Lexical Innovations

The results from the network approach show that community structure can be used to assess degrees of diffusion. The social structure of communities during the early stages of diffusion is commonly assumed to be an important factor for the successful spread of linguistic innovations. While a detailed analysis is beyond the scope of the present paper, the present approach yields initial results of the predictive power of social network information.

The dataset shows a significant correlation between the network structure in the first period of diffusion and the overall success of neologisms. Correlating CENTRALITY scores for all neologisms in Subset 1 with their total usage FREQUENCY observed across their full observed lifespan in the corpus yields Spearman correlation coefficient of -0.43 ($p < 0.001$). This means that neologisms are overall more likely to spread successfully if their use is not limited to a centralized network of speakers in their early stages. Among the selected cases presented above, *upskill* fits this pattern: it shows a consistent, successful trajectory of diffusion and its use has been the product of a decentralized bunch of users since its early attestations. Of course, the diverging pathways of diffusion for other words such as *hyperlocal* and *solopreneur* presented in **Figure 4** represent exceptions to this general trend. While this trend fits theoretical expectations and the empirical observations in the present dataset, these results remain preliminary. Since centrality correlates with frequency scores, future work based on larger samples, external data for evaluation, and more robust statistical tests is needed to test whether the predictive power of social network features can be confirmed.

6 DISCUSSION

In this paper, I have studied the spread of neologisms on Twitter to provide a multi-layered picture of the diffusion of lexical innovations in terms of 1) overall usage frequency, 2) changes in usage frequency over time (volatility), and 3) pathways of social diffusion across members and networks in a larger speech community. The process of diffusion entails social processes which lead to the spread of innovations in social networks (Rogers 1962). Theoretical models characterise the spread of linguistic innovations to new speakers and communities as the key feature of the process of diffusion (Weinreich et al., 1968; Schmid 2020). Despite a broad consensus over the fact that diffusion entails spread in networks of speakers, most previous

empirical investigations of lexical innovation have not been based on social network information, but have relied on frequency measures as an indicator for the diffusion of neologisms (Stefanowitsch and Flach 2017). The present study used a large Twitter dataset to investigate the sociolinguistic dynamics of diffusion of neologisms in online social networks. Aside from an in-depth analysis of the spread of neologisms in the present sample, the aim of this paper was to assess the usefulness of using usage frequency and social network data as indicators of diffusion.

6.1 Temporal Dynamics of Diffusion

The frequency-based approach revealed that frequency measures can be used to assess degrees of diffusion of lexical innovations with varying success. Total frequency counts (**Tables 1–4**) proved successful for a coarse-grained distinction between cases of high (e.g., *tweeter*, *smartwatch*), medium (e.g., *monthiversary*, *helicopter parenting*), and low degrees of diffusion (e.g., *begpacker*, *bediquette*). However, differences in the temporal dynamics of use have proved to be necessary for a more accurate assessment of the degrees and pathways of diffusion of neologisms.

Considering the nature of the process and products of *lexical* innovation, this temporal sensitivity is not surprising. Models of linguistic diffusion such as the S-curve model assume competition processes in which several formal variants compete to become the conventional linguistic means to express a certain meaning/function in the speech community. In cases of grammatical innovation, which is at the core of most models and most previous empirical investigations of diffusion, the communicative need for expressing the target concept/function remains stable over time. While grammatical means are, of course, also subject to language change (e.g., *going to*, *will* future), the salience of the target semasiological space (e.g., ‘expressing future intention’), remains stable over time for all speakers in the speech community. Both the direct competition between linguistic variants and the social and temporal invariance of the conceptual space over time are tacit assumptions of S-curve models of diffusion (Blythe and Croft 2012).

Earlier work by Nini et al. (2017) suggests that the diffusion of lexical innovations also follows S-curve trajectories, and the authors use the term ‘semantic carrying capacity’ to refer to the semantic potential of neologisms during diffusion. It seems plausible that the semantic carrying capacity of new words exhibits significant volatility over time and across communities of speakers. While the present study cannot measure or control for changes in semantic potential over time, it tries to account for the temporal sensitivity of neologisms by going beyond cumulated frequency counts and studying their temporal usage profiles.

The present study focused on three main aspects of the temporal dynamics of diffusion: trends in usage intensity, age and volatility. Firstly, trends in usage frequency add information about changes in the degrees of diffusion of neologisms over time. Going beyond total frequency counts, visualising the cumulative increases in usage frequency over time in **Figure 1** revealed significant differences in the pathways of diffusion of neologisms with similar total frequency counts. The neologism *hyperlocal* showed the most linear trajectory indicating fairly consistent use, the convex curve of *upskill* indicated a positive

trend in its use, and the concave trajectories of *solopreneur* and *alt-left* suggested negative trends in the recent past.

Cumulated frequency counts, which are, in their pure form as total counts, agnostic to temporal trends, have successfully been used as an approximation of the 'potential exposure' (Stefanowitsch and Flach 2017) of speakers to linguistic constructions in previous usage-based corpus-linguistic studies. The present results emphasize, however, that temporal trends and changes in usage frequency cannot be neglected when assessing the social diffusion of neologisms, since innovation in the lexicon is subject to high degrees of temporal variation. Notably, trends in usage frequency in the present sample can almost always be traced back to changes in the neologisms' semantic carrying capacity and are not merely the product of onomasiological competition between formal variants¹⁹. Typical examples of the influence of topical salience on the use of neologisms are re-current topical neologisms like *poppygate* discussed in **Section 5.1.1**.

Secondly, it was shown that the age of neologisms provides important information about their diffusion processes. Neologisms such as *hyperlocal* and *alt-left*, which are comparable in total use frequency, but differ strongly with regard to their observed lifespan in the corpus, show different pathways and degrees of diffusion. Older neologisms whose use is distributed more evenly across longer periods of consistent usage (*hyperlocal*) typically show higher degrees of social diffusion than younger neologisms whose use almost exclusively goes back to a short period of highly intensive use (*alt-left*). The positive relationship between the age of neologisms and their degrees of diffusion was supported by the significant correlation with centrality in the network analysis. While a longitudinal, predictive approach to the fate of lexical innovations is beyond the scope of the present paper, it seems possible that neologisms follow Lindy's Law: the longer new words have been in use in the speech community, the less likely they are to become obsolete in the (near) future (Eliazar 2017). The fate of new words ultimately depends on the conceptual salience of the objects and practices they denote, however: whether *smartwatch* and *blockchain* outlive previous neologisms such as *Walkman* and *Discman* ultimately depends on the future success of these products in our society.

Lastly, the results showed that volatility in use is an important factor in the diffusion of neologisms. While some candidates show fairly consistent usage frequency over time (e.g., *hyperlocal*, *upskill*), most exhibit considerable fluctuations. For some words in the sample, recurrent spikes in usage intensity are an inherent part of their usage profile. The neologism *youthquake* is characterised by spikes in usage intensity when relevant to current public affairs, but shows low frequency of use in the intermediate intervals. Due to the nature of this behaviour, this pattern has been termed 'topical' by Fischer (1998). Cases such as *poppygate*, for which these topical spikes occur in fairly regular, periodic intervals, have been classified as 'recurrent semi-conventionalization' by Kerremans (2015). For both groups of neologisms total frequency counts cannot provide an accurate estimation of degrees of diffusion since they lack information about these patterns of volatility which are central

to these cases of lexical innovation. The network approach to diffusion in **Section 5.2** revealed a negative correlation between volatility and degrees of diffusion. It seems that neologisms that are used less consistently over time are less likely to reach advanced degrees of diffusion. Moreover, comparing frequency counts and degree centrality indicated that frequency tends to overestimate the degree of diffusion of topical neologisms. This is in accordance with the observation that isolated spikes in usage intensity tend to go back to disproportionately smaller parts of the speech community.

6.2 Social Dynamics of Diffusion

To get a more differentiated view of the social dynamics of diffusion, I conducted a social network analysis of the present dataset. Successful diffusion was defined in **Section 2** as spread to new speakers and new communities. Unlike measures such as frequency and volatility which are solely based on the occurrence of neologisms in the corpus, the network approach is based on the social structure of the networks of speakers who have used the target neologisms and thus provides a more direct operationalisation of social pathways of diffusion.

The present results show considerable overlap between frequency and network measures of diffusion. Network centrality significantly correlates with usage frequency, and visualising the relationship between both metrics (**Figure 6A**) confirms this trend. Both metrics assign high scores for diffusion to established neologisms such as *man bun*, and low scores to less established candidates such as *microflat*. Moreover, centrality shows significant correlations with age and volatility, thus confirming the intuition and general finding that higher usage intensity correlates with wider social diffusion.

The more detailed evaluation of both approaches in **Section 5.3.2** also revealed that usage frequency is an imperfect predictor of social diffusion. Centrality generally tends to assign lower degrees of diffusion than frequency for some of the cases in the sample. The main groups affected consist of neologisms whose use goes largely back to specific communities of practice (e.g., *solopreneur*), political communities (e.g., *alt-left*), and/or highly volatile neologisms (e.g., *poppygate*). A closer analysis of these cases in **Section 5.2** showed that in these cases the observed number of uses of these neologisms stems from a comparatively smaller number of speakers and communities. It thus seems that the social network information contained in the measure of centrality manages to account for cases in which total usage frequency overestimates degrees of diffusion.

These discrepancies in results reflect two perspective on the process diffusion. Successful diffusion of neologisms was defined as spread to new speakers and new communities. Using the frequency of occurrence of a neologism in a corpus to approximate to what degree it is familiar to bigger parts of the speech community thus has to rely on several assumptions which are only accurate to a certain extent.

Firstly, the number of uses observed might diverge from the number of speakers who are familiar with the term. Frequency can overestimate the latter, for example, if the observed use is the product of high usage intensity by a smaller number of speakers (e.g., *solopreneur*) rather than moderate use by a higher number of speakers (e.g., *hyperlocal*).

¹⁹As an exception, the sample contains two sets of formal variants: *monthversary* & *monthiversary* and *rapefugee*, *rapeugee* & *rapugee*.

Secondly, usage frequency only captures active uses of the term and is blind to the number of speakers who are familiar with the term, but have not used it in the corpus. By contrast, social network metrics also include speakers who have only been passively exposed to the term, and thus covers a broader, and arguably more relevant definition of ‘familiarity’. Network metrics are free from the assumption that the observed output of speakers in the corpus is representative of the input to speakers in the speech community (Stefanowitsch and Flach 2017).

Lastly, the number of uses observed might not be indicative of whether a neologism has spread beyond certain sub-communities and has reached a broader spectrum of the speech community. Many of the neologisms for which centrality indicates significantly lower degrees of diffusion than frequency are socio-politically loaded and known to be used by fragmented and polarized communities, mainly from the far-right end of the political spectrum (Sunstein, 2018). **Figure 6B** features terms such as *alt-right*, *alt-left*, *birther*, *covfefe*, *Pizzagate*, and *Kushnergate*. Among the selected cases, *alt-left* and *hyperlocal* show a similar total number of uses. Moreover, the numbers of users involved in its use in the last temporal subset are almost identical: 26,367 vs. 26,548. Yet, their social network structure in **Figure 3** and their centrality scores indicate far lower degrees of diffusion for *alt-left*. While this political term has become popular among a closely connected community of users, its conventionality remains limited to this social niche and does not extend to bigger parts of the speech community. Its isolated use is in accordance with the socio-linguistic background of the term which was consciously coined by far-right activists as a disparaging out-group term in an attempt to ‘Unite the Right’.

The potential distortions that may arise when assessing the degrees of conventionality of linguistic constructions on the basis of usage frequency alone apply in principle to all linguistic domains. However, the underlying assumptions are particularly problematic in the case of lexical innovation.

Firstly, linguistic *innovations* are by definition new and not (yet) conventional among the speech community. It is therefore to be expected that their use is unevenly distributed across communities of speakers. Since frequency counts alone do not provide information about this distribution, sociolinguistic data are needed to assess the degrees of social diffusion of linguistic innovations.

Secondly, unlike linguistic innovations in other domains such as morphology or syntax, *lexical* innovations are often consciously coined and have a very specific communicative function. Their usefulness is closely tied to the conceptual salience of the entity they denote. The semantic carrying capacity of new words is thus much more likely to exhibit social and temporal variation than the functional potential of grammatical constructions. While speakers of English from all walks of life have felt the urge to talk about the future, the urge to talk about the future of ‘blockchain’ has only come up very recently, is (still) limited to specific parts of the speech community, and might not persist in the future. In other words, the use of lexical innovations exhibits greater social and temporal variation than innovations in other linguistic domains. The interpretation of aggregated frequency counts, which suggest a uniform distribution of use across time and across the speech community, is thus particularly problematic for assessing the diffusion of new words.

Moreover, neologisms typically arise in specific communities of practice and often show, at least initially, high degrees of social indexicality with regard to these communities. The present dataset includes several neologisms which are associated with youth language (*fleek*, *lituation*) and political discourse (*birther*, *alt-left*), for example. A term like *alt-left*, which could in principle be used neutrally to designate the political far-left, is highly socially indexical of the far-right community it emerged from. Therefore it is less likely to be used by speakers outside this community, unless they are willing to be associated with this community. Neologisms which are socially indexical are thus more community-specific. Even when speakers outside this community are familiar with these terms, they are less likely to use them. Usage frequency counts miss such effects, since they only capture active uses of neologisms.

7 CONCLUSION

In summary, the present study has shown that frequency and network-based approaches capture different kinds of information about the use and spread of new words. As we have seen, both approaches show considerable overlap in their overall assessment of degrees of diffusion. On the one hand, measures which are based on the occurrence of neologisms in the corpus such as frequency, age, and volatility capture important aspects about the temporal usage profiles of neologisms. On the other hand, social networks provide a more differentiated view of the social dynamics of diffusion. They allow to visualise and quantify different pathways and degrees of diffusion, which enables a more detailed analysis of the spread of new words to new speakers and communities. While the approaches differ in their strengths and weaknesses, combining information from both approaches provides the most complete picture of diffusion, of course. In corpus-linguistic practice, total frequency counts are the most readily available and most widely used measure for the conventionality of linguistic constructions. The present results suggest that the additional consideration of temporal dynamics of use and social network information can contribute substantially towards a more detailed and accurate picture of diffusion.

As I have argued, the use of network information is of particular importance for the study of neologisms, due to the nature of the process of lexical innovation. However, social network analysis also has great potential for sociolinguistic research in other domains. One of its biggest advantages is that it is usage-based and captures the communicative behaviour of speakers in interaction. It thus enables very fine-grained analyses of the sociolinguistic dynamics of communities, which can be visualised and qualitatively inspected on the basis of network graphs. Additionally, network science offers powerful algorithms to quantify and model the social characteristics of communities on a macro level.

The interactional dynamics discovered by network analyses can be a valuable addition to more traditional, static sociolinguistic information such as metadata about groups of speakers. Moreover, network analyses can be used in cases where metadata about speakers are unavailable, as in the present study. Since the

importance of online social networks like Twitter and Reddit is only going to grow in the future, both in terms of their role in society and in academic research, network analyses have great potential for future sociolinguistic research.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Twitter's Terms of Service. Requests to access these datasets should be directed to QW, q.wuerschinger@lmu.de.

REFERENCES

- Banes, K. (2014). *Free Tools for Writers, Bloggers and Solopreneurs*. Seattle, Washington: Amazon. Available at: <https://www.amazon.com/-/de/gp/product/B001OW2Q10>.
- Bastian, M., Heymann, S., and Jacomy, M. (2009). "Gephi: An Open Source Software for Exploring and Manipulating Networks," in Third International AAAI Conference on Weblogs and Social Media, March, 2009, San Jose, CA. Available at: <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>.
- Bliss, C. A., Kloumann, I. M., Harris, K. D., Danforth, C. M., and Dodds, P. S. (2012). Twitter Reciprocal Reply Networks Exhibit Assortativity with Respect to Happiness. *J. Comput. Sci.* 3 (5), 388–397. Advanced Computing Solutions for Health Care and Medicine. Available at: <http://www.sciencedirect.com/science/article/pii/S187775031200049X>. doi:10.1016/j.jocs.2012.05.001
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast Unfolding of Communities in Large Networks. *J. Stat. Mech.* 2008, P10008. doi:10.1088/1742-5468/2008/10/p10008
- Blythe, R. A., and Croft, W. (2012). S-curves and the Mechanisms of Propagation in Language Change. *Language* 88 (2), 269–304. doi:10.1353/lan.2012.0027
- Brin, S., and Page, L. (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine," in Seventh International World-Wide Web Conference (WWW 1998), April, 1998, Brisbane, Australia. Available at: <http://ilpubs.stanford.edu:8090/361/>. doi:10.1016/s0169-7552(98)00110-x
- Bruns, A. (2012). How Long Is a Tweet? Mapping Dynamic Conversation Networks Ontwitterusing GawK and Gephi. *Inf. Commun. Soc.* 15 (9), 1323–1351. doi:10.1080/1369118X.2011.635214
- Camenisch, J., Lambrinoukakis, C., Jódar, L., Cortés, J. C., and Acedo, L. (2011). Public Key Services and EUROPKI-2010-Mathematical Modelling in Engineering & Human Behaviour. *Math. Comp. Model.* 57 (7), 1577–2028. Available at: <https://www.sciencedirect.com/science/article/pii/S0895717711007898> (Accessed 02 06, 2021).
- Cartier, E. (2017). "Neovelle, a Web Platform for Neologism Tracking," in Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics, March, 2007, (Valencia, Spain: Association for Computational Linguistics), 95–98. Available at: <https://aclweb.org/anthology/E17-3024>. doi:10.18653/v1/e17-3024
- Davies, M. (2013). Corpus of News on the Web (NOW) - 3+ Billion Words from 20 Countries. Available at: <https://www.english-corpora.org/now/> Accessed August 6, 2021.
- Del Tredici, M., and Fernández, Rl. (2018). "The Road to Success: Assessing the Fate of Linguistic Innovations in Online Communities," in Proceedings of the 27th International Conference on Computational Linguistics. Available at: <https://arxiv.org/abs/1806.05838>.
- Dunbar, R. I. M. (1992). Neocortex Size as a Constraint on Group Size in Primates. *J. Hum. Evol.* 22 (6), 469–493. doi:10.1016/0047-2484(92)90081-j
- Eisenstein, J., O'Connor, B., Smith, N. A. P., and Xing, E. P. (2014). Diffusion of Lexical Change in Social Media. *PLOS ONE* 9 (11), e113114–13. doi:10.1371/journal.pone.0113114
- Eliazar, I. (2017). 'Lindy's Law'. *Physica A: Stat. Mech. its Appl.* 486, 797–805. doi:10.1016/j.physa.2017.05.077
- Elsen, H. (2004). *Neologismen. Formen Und Funktionen Neuer Wörter in Verschiedenen Varietäten Des Deutschen*. Tübingen: Narr.
- Fischer, R. (1998). *Lexical Change in Present Day English. A Corpus Based Study of the Motivation, Institutionalization, and Productivity of Creative Neologisms*. Tübingen: Narr.
- Freeman, L. C. (1978). Centrality in Social Networks Conceptual Clarification. *Social Networks* 1 (3), 215–239. Available at: <http://www.sciencedirect.com/science/article/pii/0378873378900217> (visited on 02/06/2020. doi:10.1016/0378-8733(78)90021-7
- Gérard, C., Bruneau, L., Falk, I., Bernhard, D., and Rosio, A.-L. (2017). 'Le Logoscope : Observatoire Des Innovations Lexicales En Français Contemporain'. In: *La Neología En Laslenguas Románicas: Recursos, Estrategias Y Nuevas Orientaciones*, Editor J. Palacios, G. de Sterck, D. Linder, J. del Rey, M. S. Ibanez, and N. M. García. Frankfurt a. M., Germany: Peter Lang. Available at: <https://hal.archives-ouvertes.fr/hal-01388255>.
- Gerlitz, C., and Rieder, B. (2013). Mining One Percent of Twitter: Collections, Baselines, Sampling. *M/C J.* 16. 2013 Available at: <http://www.journal.mediaculture.org.au/index.php/mcjournal/article/view/620>. doi:10.5204/mcj.620
- Goel, R., Soni, S., Goyal, N., Paparrizo, J., Wallach, H., Diaz, F., et al. (2016). 'The Social Dynamics of Language Change in Online Networks'. In: *Social Informatics*. Editors E. Spiro and Y.-Y. Ahn. Cham: Springer International Publishing, 41–57. doi:10.1007/978-3-319-47880-7_3
- Granovetter, M. S. (1973). 'The Strength of Weak Ties. *Am. J. Sociol.* 78 (6), 1360–1380. doi:10.1086/225469
- Grieve, J. (2018). "Natural Selection in the Modern English Lexicon," in Proceedings of EVOLANG XII (Poland: Torun). doi:10.12775/3991-1.037
- Grieve, J., Montgomery, C., Nini, A., Murakami, A., and Guo, D. (2019). Mapping Lexical Dialect Variation in British English Using Twitter. *Front. Artif. Intell.* 2, 11, 2019. Available at: <https://www.frontiersin.org/article/10.3389/frai.2019.00011>. doi:10.3389/frai.2019.00011
- Grieve, J., Nini, A., and Guo, D. (2016). Analyzing Lexical Emergence in Modern American English Online. *English Lang. Linguistics.* 21, 99–127. doi:10.1017/S1360674316000526
- Grieve, J., Nini, A., and Guo, D. (2018). Mapping Lexical Innovation on American Social Media. *J. English Linguistics.* 46 (4), 293–319. doi:10.1017/s1360674316000113
- Halu, A., Mondragón, R. J., Panzarasa, P., and Bianconi, G. (2013). 'Multiplex PageRank'. *PLOS ONE* 8 (10), e78293, 2013. Available at: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0078293> (Accessed 02 06, 2021). doi:10.1371/journal.pone.0078293
- Hébert-Dufresne, L., Scarpino, S. V., and Young, J.-G. (2020). Macroscopic Patterns of Interacting Contagions Are Indistinguishable from Social Reinforcement. *Nat. Phys.* doi:10.1038/s41567-020-0791-20791-210.1038/s41567-020-0791-2
- Hohenhaus, P. (1996). *Ad-Hoc-Wortbildung. Terminologie, Typologie Und Theorie Kreativer Wortbildung Im Englischen* Frankfurt a. M.: Lang.
- Hohenhaus, P. (2006). 'Bouncebackability. A Web-As-Corpus-Based Study of a New Formation, its Interpretation, Generalization/Spread and Subsequent Decline'. *SKASE J. Theor. Linguistics* 3, 17–27.
- Huberman, B. A., Romero, D. M., and Wu, F. (2008). *Social Networks that Matter: Twitter under the Microscope*. Available at: <http://arxiv.org/abs/0812.1045>.
- Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLOS ONE* 9 (6), e98679. doi:10.1371/journal.pone.0098679
- Kerremans, D. (2015). *A Web of New Words*. Bern, Schweiz: Peter Lang. doi:10.3726/978-3-653-04788-2

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

I would like to thank Stefano Coretta, Cornelius Fritz, Hans-Jörg Schmid, Maximilian Weigert, and the two reviewers for their comments on this paper.

- Kerremans, D., Stegmayr, S., and Schmid, H. J. (2012). "The NeoCrawler: Identifying and Retrieving Neologisms from the Internet and Monitoring Ongoing Change," in *Current Methods in Historical Semantics* (Berlin: Mouton de Gruyter), 59–96.
- Kerremans, D., Prokić, J., Würschinger, Q., and Schmid, H.-J. (2019). Using Data-Mining to Identify and Study Patterns in Lexical Innovation on the Web. *Pragmatics Cogn.* 25 (1), 174–200. doi:10.1075/pc.00006.ker
- Labov, W. (2007). Transmission and Diffusion. *Language* 83 (2), 344–387. doi:10.1353/lan.2007.0082
- Lemnitzer, L. (2010). Wortwarte. Available at: <http://www.wortwarte.de/> Accessed 6 August, 2021.
- Lu, F. S., Hou, S., Baltrusaitis, K., Shah, M., Leskovec, J., Sosis, R., et al. (2018). Accurate Influenza Monitoring and Forecasting Using Novel Internet Data Streams: A Case Study in the Boston Metropolis. *JMIR Public Health Surveill.* 4, e4. Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5780615/>. doi:10.2196/publichealth.8950
- Milroy, J. (1992). *Linguistic Variation and Change: On the Historical Sociolinguistics of English*. Oxford: Blackwell.
- Milroy, J., and Milroy, L. (1985). Linguistic Change, Social Network and Speaker Innovation. *J. Ling.* 21 (2), 339–384. Available at: <https://www.cambridge.org/core/article/linguistic-change-social-network-and-speaker-innovation/1/EB30A7117C09F6EDA5255BF9D788D5A>. doi:10.1017/s0022226700010306
- Nevalainen, T. (2015). Descriptive Adequacy of the S-Curve Model in Diachronic Studies of Language Change. *Studies in Variation, Contacts and Change in English* 16. Available at: <https://varieng.helsinki.fi/series/volumes/16/nevalainen/> Accessed August 6, 2021.
- Nini, A., Corradini, C., Guo, D., and Grieve, J. (2017). The Application of Growth Curve Modeling for the Analysis of Diachronic Corpora. *Lang. Dyn. Change.* 7 (1), 102–125. doi:10.1163/22105832-00701001
- Pedroche, F., Moreno, F., González, A., and Valencia, A. (2013). Leadership Groups on Social Network Sites Based on Personalized PageRank. *Math. Comp. Model.* 57 (7pp), 1891–1896. doi:10.1016/j.mcm.2011.12.026
- Pew Research Center (2019). National Politics on Twitter: Small Share of U.S. Adults Produce Majority of Tweets. Available at: <https://www.people-press.org/2019/10/23/national-politics-on-twitter-small-share-of-u-s-adults-produce-majority-of-tweets/> Accessed August 6, 2021.
- R Core Team (2018). *R: A Language and Environment for Statistical Computing. Manual*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>.
- Renouf, A., Kehoe, A., and Banerjee, J. (2007). WebCorp: An Integrated System for Web Text Search. Editors M. Hundt, N. Nesselhauf, and C. Biewer. *Corpus Linguistics and the Web*. Amsterdam, New York: Rodopi, 59–47.
- Rogers, E. M. (1962). *Diffusion of Innovations*. New York: Free Press of Glencoe.
- Schmid, H.-J. (2016). *English Morphology and Word-Formation - an Introduction*. 2nd ed. Berlin: Erich Schmidt Verlag.
- Schmid, H.-J. (2020). *The Dynamics of the Linguistic System. - Usage, Conventionalization, and Entrenchment*. Oxford: Oxford University Press.
- Spearman, C. (1961). "The Proof and Measurement of Association between Two Things," in *Studies in Individual Differences: The Search for Intelligence* (East Norwalk, CT, US: Appleton-Century-Crofts), 45–58. doi:10.1037/11491-005
- Stefanowitsch, A., and Flach, A. (2017). 'The Corpus-Based Perspective on Entrenchment'. In: *Entrenchment and the Psychology of Language Learning: How We Reorganize and Adapt Linguistic Knowledge*. Editors H. J. Schmid. Boston, USA: American Psychology Association and de Gruyter Mouton, 101–127. doi:10.1037/15969-006
- Stewart, I., and Jacob, E. (2018). Making 'Fetch' Happen: The Influence of Social and Linguistic Context on Nonstandard Word Growth and Decline. Available at: <http://arxiv.org/abs/1709.00345> Accessed August 6, 2021.
- Sunstein, C. R. (2018). *#Republic: Divided Democracy in the Age of Social Media*. Princeton and Oxford: University Press.
- Wang, R., Zhang, W., Deng, H., Wang, N., Miao, Q., and Zhao, X. (2013). 'Discover Community Leader in Social Network with PageRank'. In: *Advances in Swarm Intelligence*. Editors Y. Tan, Y. Shi, and H. Mo. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 154–162. doi:10.1007/978-3-642-38715-9_19
- Weinreich, Uriel, Labov, W., and Herzog, M. (1968). 'Empirical Foundations for a Theory of Language Change'. In: *Directions for Historical Linguistics*. Editors W. P. Lehmann and Y. Malkiel. Austin: University of Texas Press Austin, 95–188.
- West, Robert, Paskov, H. S., Leskovec, J., and Potts, C. (2014). Exploiting Social Network Structure for Person-To-Person Sentiment Analysis. Available at: <http://arxiv.org/abs/1409.2450> Accessed August 6, 2021. doi:10.1162/tacl_a_00184
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., et al. (2019). Welcome to the Tidyverse. *Joss* 4 (43), 1686. doi:10.21105/joss.01686
- Würschinger, Q., Elahi, M. F., Zhekova, D., and Schmid, H.-J. (2016). "Using the Web and Social Media as Corpora for Monitoring the Spread of Neologisms. The Case of 'rapefugee', 'rapeugee', and 'rapugee'." in Proceedings of the 10th Web as Corpus Workshop, August, 2016, Berlin, Germany (Berlin: Association for Computational Linguistics), 35–43. Available at: <http://aclweb.org/anthology/W16-2605>. doi:10.18653/v1/W16-2605

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Würschinger. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

4.3 Conclusions

This paper advanced the previous investigations of diffusion based on the NeoCrawler and the Twitter study of *rapefugee* in several ways.

To tackle the concerns regarding the reliability of usage frequency as an indicator of diffusion, which was raised in Sections 2 and 3, the study applied additional measures for the temporal dynamics of diffusion. It examined trends in usage intensity as captured by cumulative and absolute frequency counts, the volatility of neologisms as measured by the coefficient of variation, and the age of neologisms as indicated by their observed lifespan since their emergence. I argue that the additional consideration of temporal dynamics of use as captured by these measures may contribute towards a more detailed and accurate picture of diffusion. In addition, I suggest that these temporal dynamics are of particular importance for the study of *lexical* innovation, since neologisms are strongly influenced by changes in their ‘semantic carrying capacity’ (Nini et al. 2017).

The proposed methods were intended to provide a better view of cases of ‘topicality’ (Fischer 1998: 16) and ‘re-current semi-conventionalization’ (Kerremans 2015: 129–136), as observed in Kerremans (2015) and in Chapter 2. Regarding the case of the term *rapefugee*, which was found to show high topicality in Chapter 3, the current paper confirmed this impression. The new, extended Twitter dataset validates its initial emergence at the end of 2015 and demonstrates that its use until 2018 largely remains limited to the first spike in usage intensity covered in Chapter 3. It has limited semantic carrying capacity since the Cologne incident that led to its emergence remained unique. Its coefficient of variation of 1.37 reflects the fact that it exhibits high volatility.

In addition, the paper augmented frequency-based measures of diffusion by employing social network analysis on the Twitter dataset to provide direct insights into social diffusion. Several network-related metrics (e.g. centralization, in-degree) and network graph visualisations were used to determine whether neologisms successfully diffuse across speakers and communities. Applied to the term *rapefugee*, the network analysis supported earlier qualitative findings indicating that its use remained confined to a small community of densely-connected, like-minded individuals. With a centralization score of 0.26, it is the term in the sample that displays the lowest degree of social diffusion.

Comparing the frequency-based and network-based approaches unveiled a substantial degree of congruence in their overall assessments of the degrees of diffusion of the neologisms in the sample. However, it also highlighted several cases where usage frequency seems to overestimate degrees of social diffusion. The analysis revealed a pattern of cases that resemble the case of *rapefugee*. The data suggests that *rapefugee* shares questionable company with several other politically charged neologisms such as *alt-left* or *birther*, which also exhibit high usage intensity and centralization. These terms show high degrees of social variation and are predominantly used in small, politically polarized areas of the social network that resemble ideological echo chambers.

5 Semantic innovation of an old word – The case of *Anglo-Saxon*

5.1 Research context

After examining the emergence and diffusion of formal neologisms in the preceding three chapters, this chapter will pivot towards semantic innovation, which has been outside the focus of most previous work on lexical innovation. It studies recent changes in the use and meaning of the term *Anglo-Saxon* on Twitter.

The term *Anglo-Saxon* seemingly deviates from a typical neologism. In fact, it is very old, since it is tied to the roots of the English language. Yet, the term's meaning has lately undergone semantic innovation. It is commonly used with three primary senses, which previous work has categorised as 'historical/pre-Conquest', 'ethno-racial', and 'politico-cultural' (Wilton 2020). Recent controversy has surrounded the term due to its ethno-racial sense, which is strongly associated with the concept of white supremacy. This controversial sense has become more dominant and has transferred its contentious associations to the term itself, leaving little room for its use in the other two, less controversial senses. This battle between competing senses and associations of *Anglo-Saxon* has resulted in changes in its use and overall meaning. The competition between senses may be seen as the semasiological counterpart to the onomasiological competition between formal variants around the 'rapefugee' terms presented in Chapter 3.

Moreover, as in the case of *rapefugee* and the sample of formal neologisms studied in the previous chapter, the use and meaning of *Anglo-Saxon* exhibit considerable social variation and are influenced by the interests of particular communities. The term has been employed by far-right groups in its ethno-racial sense with positive connotations, while center-to-left and academic circles have used it mostly in its neutral, historical sense. The previous chapters have examined social variation in terms of the extent to which communities differ in the frequency with which they use certain terms. This chapter adds a semasiological lens to investigate the change and social variation in the meaning of the term *Anglo-Saxon*.

The following section covers the paper *Battling for Semantic Territory across Social Networks. The Case of Anglo-Saxon on Twitter*, which I co-authored with Hans-Jörg Schmid (HJS), Melanie Keller (MK), and Ursula Lenker (UL). This paper was published in the *Yearbook of the German Cognitive Linguistics Association* in 2020.

The approach taken in this study resembles that of the previous chapter in some regards. Firstly, I compiled a large, longitudinal corpus of Twitter data containing all instances of the term *Anglo-Saxon* since the inception of Twitter in 2006. I then extracted all occurrences and analysed its frequency of use over time (Figure 8*). Based

on these data, HJS conducted a collocation analysis inspecting the changes in the semantic profile of *Anglo-Saxon*. Next, I split the data into four temporal bins, and constructed social network graphs for each bin. This enabled me to analyse the degree of diffusion vs centralization of *Anglo-Saxon* over time, as well as the social structure of the speakers and communities involved in its usage. Lastly, together with MK, I manually labelled the most influential actors in the social network to examine their role in the battle for the meaning of *Anglo-Saxon*.

HJS was the principal author of the paper; UL wrote the introduction (Section 1*); I wrote the section on the data collection and processing (first part of Section 2*) and the section on diffusion (Section 5.2*). All authors contributed to the final version via revisions and comments.

Due to copyright restrictions, this paper is not reproduced in this dissertation. For reference, please consult the published version of the paper:

Schmid, Hans-Jörg, Quirin Würschinger, Melanie Keller & Ursula Lenker. 2020. Battling for Semantic Territory across Social Networks. The Case of Anglo-Saxon on Twitter. *Yearbook of the German Cognitive Linguistics Association* 8 (1): 3–26. <https://doi.org/10.1515/gcla-2020-0002>.

High-resolution, zoomable versions of the network graphs in Figure 10* can be viewed by downloading and opening the files from this address: <https://osf.io/vp43t/>.

The subsequent section presents the conclusions derived from this study, contextualized within the scope of my dissertation.

5.2 Conclusions

This paper has found notable socio-semantic variation in the use of the term *Anglo-Saxon*, which is subject to an intense battle over its meaning. The study of social variation in this paper expands on the differences in social diffusion found in the previous chapters. As in previous cases such as *rapefugee* (Chapter 3) or *alt-left* (Chapter 4), the use of *Anglo-Saxon* differs strongly between communities. Unlike the approaches in the preceding chapters, however, this study went beyond analysing differences in usage intensity and found that communities exhibit considerable social variation regarding the meaning of the term *Anglo-Saxon*.

The social network analysis in this paper suggests, like the preceding chapters, that the discourse on Twitter is highly polarised. Certain communities (e.g. historians or far-right activists in the US) seem to form echo chambers in which there is strong agreement in socio-political views. Within these communities, individuals mutually reinforce each other, leading to growing usualization of group conventions regarding whether and how to use the term *Anglo-Saxon*. Previous research in social psychology has shown that such social fragmentation into distinct groups with continual in-group affirmations leads to increasing divergence between groups and increasing polarisation in attitudes and conventions within the involved groups (Lukianoff & Haidt 2018; Sunstein 2018, 2019).

The present results indicate this increasing polarisation surrounding the term *Anglo-Saxon*. The network analysis revealed the growing centralization in its use over time, which the manual network analysis corroborated. Some communities continue to use the term in its previous meaning, while others have ceased using it since they have come to associate it with the political far-right. While the future of its use and meaning is currently undetermined, the results indicate that there is a significant chance that the current social dynamics will result in sustained changes to its meaning and use. Thus, the term *Anglo-Saxon* is an interesting case of semantic change and socio-semantic variation. It illustrates how an established word may exhibit significant semantic variation between communities and may be susceptible to meaning change over a brief period of time.

However, the observed results cannot be readily generalised to gain insights into the general dynamics of semantic change and socio-semantic variation. Due to the specifics of the current context, these findings are restricted to the present case study. Given the current controversy surrounding the term *Anglo-Saxon*, its use is likely to be strongly polarised, especially considering that the present investigation is confined to Twitter, which is known to be politically polarised (Yardi & Boyd 2010; Conover et al. 2011; Himelboim, McCreery & Smith 2013; Lotan & Minkov 2021).

Furthermore, the results depend on the community detection conducted by the social network analysis. While the quantitative analysis based on centralization and the qualitative analysis based on a manual inspection of the social network are in agreement, the reliability of this community analysis across the entire dataset is unclear since the modularity algorithm used for the network analysis does not provide readily interpretable results.

In addition, despite the fact that the initial results of the collocation analysis were convincing, they cannot be interpreted in terms of meaning change without manual inspection, nor can they be quantified; therefore, they cannot be easily applied to community-based comparisons and the investigation of larger, more representative samples of candidates for semantic change.

The purpose of the next chapter is to address the limitations mentioned in this section by using more refined measures of socio-semantic variation based on word embeddings derived from a broader set of semantic neologisms on Reddit.

6 Semantic innovation and social variation

6.1 Research context

This chapter continues to investigate semantic innovation and socio-semantic variation. It offers both a broader and more detailed account of semantic variation between communities and addresses the limitations of the previous study discussed above (5.2).

I investigate semantic innovation in the context of the Covid pandemic, since the large societal impact of Covid has spawned a considerable amount of linguistic innovation, both in terms of formal neologisms (Thorne 2020; Roig–Marín 2020; Scott 2020) and semantic neologisms (Dong, Buckingham & Wu 2021; Irshad, Arshad & Saba 2021; Ullah Shaheen, Qadeer & Rehman Khan 2021).

Firstly, I provide a broader view of socio-semantic variation by studying a large sample of semantic neologisms, as opposed to the case study of *Anglo-Saxon* in the previous chapter. To this end, I determine semantic neologisms in a data-driven way by using word embeddings, which enable generating distinct semantic representations for the years 2019 and 2020 and identifying the words exhibiting the greatest degree of semantic change during this time frame. This approach yields a large set of candidates for semantic change, from which I have selected the 20 words with the highest degrees of semantic change to analyse whether these words show semantic variation between communities.

Secondly, to get a more detailed view of socio-semantic variation, I generate community-based semantic representations for all semantic neologisms through training separate word embedding models. This enables the large-scale, quantitative analysis of differences in meaning between communities, which was not possible in the preceding study of *Anglo-Saxon*, since manual, qualitative analyses were required. I use these community-specific semantic representations to determine whether terms such as (*social*) *distancing*, whose meanings were found to have changed, are used with different meanings between groups. In addition, I study common dimensions of variation – for example whether groups differ in the extent to which they have positive vs negative connotations with the target neologisms – by analysing their meanings in specific semantic subspaces.

For this study, I collected data from the social media platform Reddit since it provides several benefits for analysing social variation. Reddit is organised into communities termed ‘subreddits’, which facilitates the study of social variation since groups do not need to be inferred using methods such as Modularity-based clustering as in Chapter 4. These groups are communities of practice (Leuckert 2020) that typically center around a

shared interest or (political) view and users explicitly specify the groups' characteristics through community descriptions. These descriptions enable more reliable and objective interpretations of groups and group differences than comparable investigations on Twitter, which could only be examined manually based on a subset of community members, as in the cases of *rapefugee* (Chapter 3) and *Anglo-Saxon* (Chapter 5).

The subsequent section presents an earlier manuscript version for the paper *Semantic Change and Socio-Semantic Variation. The Case of Covid-related Neologisms on Reddit*, co-authored with Barbara McGillivray (BMG). After the submission of the dissertation, this manuscript has been substantially revised and has been accepted for publication in the journal *Linguistics Vanguard*. I wrote the code, collected the data, implemented the methods, analysed the results, and wrote the paper. BMG contributed to the final version via revisions and comments.

6.2 Semantic Change and Socio-Semantic Variation. The Case of Covid-related Neologisms on Reddit

Semantic change and socio-semantic variation

The case of Covid-related neologisms on Reddit

Quirin Würschinger¹ and Barbara McGillivray²

¹LMU Munich

²King's College London

Covid-19 has triggered rapid innovations in science and society around the world. These innovations have led to the diffusion of numerous formal neologisms such as *infodemic* or *working from home (WFH)*. While previous work on Covid-related lexical innovation has primarily focused on such formal neologisms (Roig–Marín 2020; Mahlberg and Brookes 2021), this paper uses data from Reddit to study semantic neologisms like *lockdown* or *mask*, which have changed in meaning due to the pandemic.

In a first step, we identify words that have changed in meaning after the start of the pandemic. Our approach based on word embeddings (Mikolov et al. 2013) manages to detect a variety Covid-related terms, which dominate the resulting list of semantic neologisms.

Next, we generate community-specific semantic representations for the communities *r/Coronavirus* and *r/conspiracy*, which are highly engaged in Covid-related discourse. We analyse socio-semantic variation along two semantic dimensions and we find that the detected semantic neologisms consistently show more negative and subjective associations in the subreddit *r/conspiracy*, which is more critical towards Covid-related sociopolitical measures. Mapping the community-specific representations for the term *vaccines* on a shared semantic space confirms these differences and reveals more fine-grained denotational and connotational differences between the two communities.

Keywords: lexical innovation, semantic change, social variation, word embeddings, Reddit

1 Introduction

Covid-19¹ has imposed extensive changes to the social practices of people around the world, and it has triggered rapid innovations in science and society. Cultural changes like the Covid pandemic primarily affect language at the level of lexis. New concepts and practices enter the linguistic system as new words or result in changes in the meaning of existing entries in the lexicon. Previous work on Covid-related lexical innovation has primarily focused on formal neologisms like *infodemic* or *pancession* (Roig–Marín 2020; Mahlberg and Brookes 2021). We aim to add to this work by studying the interplay of semantic change and socio-semantic variation of Covid-related semantic neologisms like *lockdown* or *mask* on Reddit.

In a first step, we detect Covid-related semantic neologisms. We use word embeddings to identify lexemes that have changed their meaning after the onset of the pandemic (Section 5.1). Next, we study socio-semantic differences in the meanings of these words between two Reddit communities (Section 5.2). We investigate dimensions of variation in meaning by projecting embeddings onto two semantic axes. In addition, we provide a more differentiated picture of the variation found by visualising differences in the community-specific semantic representations in the semantic space of the term *vaccines*.

2 Theoretical background and previous work

2.1 Semantic neology

Lexical innovations can be divided into formal neologisms such as *Zoom fatigue* and semantic innovations as in *booster* (Tournier 1985; Geeraerts 2010). After the start of the pandemic, scholars have used data from the web, social media, and large-scale linguistic corpora such as the new Coronavirus Corpus (Davies 2019–) to identify Covid-related formal neologisms. These studies have demonstrated the creative linguistic potential of speakers and compiled and analysed extensive lists of lexical innovations (Thorne 2020; Roig–Marín 2020; Scott 2020; Roig–Marín 2020). Covid-related semantic neologisms have received little attention from previous work. This can largely be attributed to the increased methodological complexity of studying the emergence and diffusion of semantic change. Previous linguistic investigations on Covid-related meaning variation and change have been mostly limited to qualitative studies of selected cases (Dong et al. 2021; Irshad et al. 2021; Ullah Shaheen et al. 2021).

Lexical semantic change and innovation can pertain to denotational or connotational aspects of meaning (Leech 1981; Lipka 1992; Geeraerts 2010; Koch 2016). In denotational change, words are increasingly used to refer to new concepts and practices, while connotational change involves changing associations and attitudes of speakers. Different types of meaning change can further be distinguished according to the semantic relations between the old and new meanings of lexemes (Koch 2016). Processes typically involved in denotational change include ‘generalization vs specialization’ (e.g. the narrowing of *lockdown* to refer to Covid-related sociopolitical shutdowns), and ‘metonymic change’ (e.g. the word *jab* being used to refer to

¹In the rest of this article, we will use the terms “Covid” and “Covid-19” interchangeably to refer to the disease that was declared a global pandemic by the World Health Organization on 11 March 2020 and is caused by the SARS-CoV-2 virus.

injecting vaccines instead of its earlier sense of ‘stabbing’). Such semantic changes are often driven by processes of ‘subjectification’ and ‘intersubjectification’ (Koch 2016, pp. 45–46). Both processes represent special cases of metonymic change. Subjectification is a mechanism by which ‘meanings are recruited by the speaker to encode and recruit attitudes and beliefs’, whereas intersubjectification refers to the mechanism by which speakers ‘encode meanings centred on the addressee’ (Traugott 2010, p. 35). In the case of connotational change, semantic shifts often involve changes along an evaluative dimension (Koch 2016). Words are typically associated with increasingly negative (pejorization) or positive meanings (amelioration) over time.

2.2 Semantic change detection

Recent advances in Natural Language Processing (NLP) have enabled large-scale, quantitative studies of meaning change. The large majority of NLP approaches use word embeddings (Mikolov et al. 2013) to generate computational representations of meaning of lexemes based on their distributional properties (Firth 1957). Popular word embedding algorithms such as word2vec (Mikolov et al. 2013) or fasttext (Bojanowski et al. 2017) use neural networks to learn semantic representations by predicting context words.

Within the growing computational research on lexical semantic change detection, word embeddings have been successfully employed in large-scale empirical studies of long-term meaning change spanning decades and centuries (Kim et al. 2014; Hamilton et al. 2016; Kutuzov et al. 2018). More recent advances have enabled increasingly fine-grained investigations of short-term lexical semantic change on the scale of years rather than decades (Del Tredici, Fernández and Boleda 2019; Robertson et al. 2021; Shoemark et al. 2019; Tsakalidis et al. 2019). However, previous work in NLP has put little focus on the linguistic processes underlying semantic innovation and their effects on the nature of the meaning changes observed by computational models.

2.3 Socio-semantic variation in neologisms

Besides changes in meaning over time, word embeddings can also be used to study semantic variation between communities. Earlier theoretical work has pointed out the role of socio-semantic variation for the study of meaning and meaning change (Hasan 1989; Clark 1996; Geeraerts 2015). Except for some earlier attempts (e.g. Peirsman et al. 2010), there have been few large-scale empirical approaches on inter-community variation (Del Tredici and Fernández 2017; Gonen et al. 2020; Schmid et al. 2020; Hofmann et al. 2021). We aim to add to this work by focusing on community-specific differences in the use of semantic neologisms. New words show, by definition, low degrees of conventionality in the speech community, and so do new meanings associated with existing words. Neologisms are therefore more likely to exhibit high degrees of socio-semantic variation. Semantic neologisms related to Covid are particularly prone to exhibit socio-semantic variation since social polarization seems to have significantly increased since the start of the pandemic (Lang et al. 2021). This has been empirically studied in in news reports (Hart et al. 2020) and on social media (Green et al. 2020; Jiang et al. 2020; Jing and Ahn 2021).

From a sociolinguistic perspective, fragmentation and polarization of the speech community into echo chambers can drive socio-semantic variation and contribute to semantic change. Echo chambers prevents the diffusion of linguistic conventions across communities, which is essential for establishing and levelling shared norms of language use across the speech community (Schmid 2020).

3 Data

3.1 Reddit

We draw on data from the social media platform Reddit, which provides a large sample of authentic language use spanning the period before and after the pandemic. The platform features a large community of about 52 million daily active users, who take part in about 130,000 active communities, referred to as ‘subreddits’. Communities are referred to using the prefix *r/* and typically form around types of content (*r/pics*), specific topics (*r/UkrainianConflict*), shared interests (*r/politics*) or attitudes (*r/Conservative*). Users submit content as forum posts (‘submissions’) to subreddits and other users can comment on these posts in a hierarchically organised structure.

3.2 Data retrieval

For the present study, we collect Reddit data (Baumgartner et al. 2020) using the Python library `psaw`².

Firstly, to identify Covid-related semantic neologisms that have changed in their meaning before and after the start of the pandemic, we retrieve a random sample of comments for the years 2019 and 2020. We use these two datasets to train diachronic word embedding models.

Secondly, to study socio-semantic variation between communities, we use our 2020 dataset to identify those communities on Reddit that are most actively involved in the Covid discourse. We then select two communities that are representative of two main stances in Covid-related discourse: neutral, mainstream positions (*r/Coronavirus*), and more sceptical and critical stances towards the predominant public and sociopolitical attitudes and measures (*r/conspiracy*). For both subreddits, we obtain all comments for the year 2020. We then train individual word embeddings models for these datasets to study socio-semantic variation between these communities.

Table 1 contains an overview of the datasets used in our study.

4 Method

We perform a set of preprocessing steps to clean our Reddit datasets before we generate and analyse semantic representations.³ Firstly, we remove duplicate comments, comments in languages other than English, and comments that contain fewer than 10 tokens, which do not

²<https://github.com/dmarx/psaw>

³The code for this paper is available at <https://github.com/wuqui/neocov>.

Table 1: Datasets for semantic change detection (2019 and 2020) and socio-semantic variation; total number of comments (COMM.) and tokens (TOKS) in Millions.

Dataset	Comm.	Toks
2019	5.3	178
2020	5.4	185
r/Coronavirus	4.1	110
r/conspiracy	4.0	109

provide enough context for training our models. We then lowercase and tokenise all texts and remove punctuation, numeric and alphanumeric tokens. We further remove tokens with three or fewer characters due to the high prevalence of ambiguous and non-standard variants in this category. Additionally we use a blacklist to remove material involving bots and usernames, subreddit titles (e.g. *AskReddit*), and Reddit-specific jargon (e.g. *submission*).

For each of our datasets (Table 1), we then follow the same procedure to generate semantic representations for all words contained in the dataset. We train word embeddings using the word2vec algorithm (Mikolov et al. 2013) as implemented in Gensim (Rehurek and Sojka 2011)⁴. We then use Orthogonal Procrustes Alignment (Hamilton et al. 2016) to align vector spaces between models.

To measure change over time and variation across the Reddit communities, we calculate the cosine distance between the embeddings of the same words in the different models. To analyse dimensions of semantic variation, we project the embeddings into semantic subspaces. This allows us to study whether the representations of our models differ along semantic axes defined by antonyms such as *good vs bad*. We implement this embeddings projection following the approach proposed in An et al. (2018)⁵. Lastly, we visualise semantic spaces by performing dimensionality reduction via t-SNE (Maaten and Hinton 2008) on word2vec’s vector representations.

5 Results

5.1 Meaning change and semantic neology

In a first step, we aim to detect semantic neologisms that have changed in meaning before and after the start of the pandemic. To this end, we use our diachronic datasets (Table 1) to train embedding models for the years 2019 and 2020. The resulting models yield semantic representations for a vocabulary of 252,564 (2019) and 277,707 (2020) word types, respectively. After aligning the models using Orthogonal Procrustes alignment (Hamilton et al. 2016), we

⁴We follow a similar training procedure as in related approaches (Shoemark et al. 2019), but modify some hyperparameters slightly to adjust to the smaller corpus size of our individual models. We use the following hyperparameters for training with word2vec: *min_count* = 5 (minimum token frequency), *vector_size* = 300 (number of dimensions), *window* = 5 (context window size), and *epochs* = 20 (number of epochs).

⁵As recommended by An et al. (2018), we use the 10 nearest semantic neighbours in addition to each pole word for constructing the semantic axes.

Table 2: Words that show the highest degree of semantic change between 2019 and 2020. Semantic distance (SemDist) is based on the cosine distance between vector representations between the 2019 and the 2020 embedding spaces. All Covid-related words are in bold.

Word	SemDist
lockdowns	1.02
maskless	1.00
sunsetting	1.00
childe	0.98
megalodon	0.98
newf	0.96
corona	0.93
filtrate	0.92
chaz	0.90
klee	0.89
rona	0.89
cerb	0.87
rittenhouse	0.87
vacuo	0.86
moderna	0.84
pandemic	0.84
spreader	0.84
distancing	0.83
sars	0.83
quarantines	0.82

retain a shared vocabulary of 190,756 types. As described in Section 4, we then identify those words that show the greatest distance between their semantic representation in the 2019 vs. 2020 model. We calculate pairwise cosine distances between each word and all other words in the vocabulary, and we use a minimum frequency threshold of 100 to mitigate the effects of increases in frequency on our results. Table 2 presents a list of 20 candidates that are estimated to have undergone the most pronounced meaning shifts according to this method.

Overall, this list suggests our approach manages to detect semantic neologisms with considerable success, since most lexemes are Covid-related and can be plausibly assumed to have changed in meaning due to the socio-cultural impact of the pandemic. Among many well-established Covid-related words such as *lockdowns* or *distancing*, our model also captures less widespread terms such as *cerb* (‘Canada Emergency Response Benefit for Covid’) or *vacuo* (medical term for *vacuum*).

The semantic changes for most of the detected Covid-related neologisms are denotational in nature: the terms *spreader*, *distancing*, *maskless*, *pandemic* and *quarantines* are used in more diverse contexts before the start of the pandemic, and are used to refer a more narrow set of Covid-related concepts in 2020.

A second set of words in Table 2 show connotational differences that can be related to the

social and stylistic dimensions of meaning (Leech 1981). Our models seem to capture stylistic semantic variation for the terms *filtrate*, *sars*, and *vacuo*. While words such as *filtrate* or *PCR test* are still used to refer to the same concepts and entities in our 2020 corpus, their stylistic signature has changed considerably. The use of these terms was largely limited to formal, academic discourse before the start of the pandemic. Since then, they have spread into public discourse and have partly lost their connotations of jargon and formality. The term *sunsetting* represents an example of socio-semantic variation. It is not related to Covid and is mainly used to refer to the termination of programmes and services, often in a legal or business context⁶. Starting in 2020, however, *sunsetting* has increasingly been used by gamers to refer to the disappearance of specific items in the virtual world. The diachronic change captured by our models can thus be traced back to socio-semantic variation on the community-level.

Aside from these denotational and connotational changes, our models also capture distributional variation that is not the result of semantic change in the narrow, linguistic sense. We detect cases of homonymy for the Covid-related terms *moderna*, *corona*, *cerb*, and *rona*, and all remaining cases of non-Covid related terms in Table 2 can also be attributed to emerging homonymy.⁷

5.2 Socio-semantic variation

5.2.1 Covid-related communities

In this section, we aim to study socio-semantic variation in the use of the Covid-related neologisms identified by our semantic change detection approach.

First, we identify those communities that are most actively engaged in Covid-related discourse. To do so, we extract all Covid-related comments in our 2020 dataset based on the salience of the term *Covid*, following Hofmann et al. (2021) in using frequency of occurrence as a proxy for agenda-setting. The resulting dataset contains 3.8 million comments and 145 million word tokens. We then determine the communities with the highest number of Covid-related comments. Figure 1 presents the 15 most active communities in this dataset.

As described in Section 3.2, we select two communities that represent diverging viewpoints in the Covid discourse and provide sufficient data for generating community-specific semantic representations. The subreddit *r/Coronavirus* currently has about 2.4 million users and contains open discussions and mainstream positions about the pandemic⁸. The subreddit *r/conspiracy* has about 1.7 million users and represents a slightly smaller, more tightly-knit community of sceptics who are critical of Covid-related measures such as masks, lockdowns, and the general response by science, media, and politics. Its scepticism extends to sociopolitical issues outside the pandemic: ‘We hope to challenge issues which have captured the public’s imagination, from

⁶‘Termination or discontinuance of a programme, service, etc., after a fixed period of operation, under a sunset provision or sunset legislation. (“sunsetting, n.”. OED Online. December 2021. Oxford University Press. (accessed March 7, 2022))’

⁷We find two topical groups: *rittenhouse* and *chaz* are related to the Black Lives Matter movement in 2020; *childe*, *megalodon*, and *kleo* are used as gaming-related terms in 2020.

⁸Community description for *r/Coronavirus*: ‘This subreddit seeks to monitor the spread of the disease COVID-19, declared a pandemic by the WHO. This subreddit is for high-quality posts and discussion. Please be civil and empathetic.’

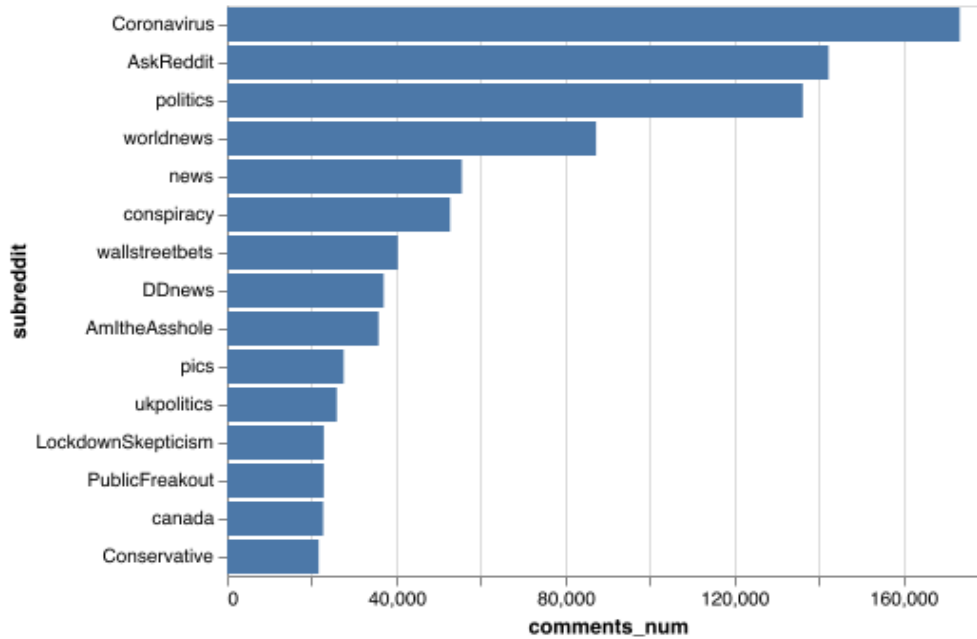


Figure 1: Most active Covid-related communities in our 2020 dataset.

JFK and UFOs to 9/11.’ For both subreddits, we retrieve all comments for the year 2020, resulting in the datasets described in Table 1.

5.2.2 Dimensions of socio-semantic variation

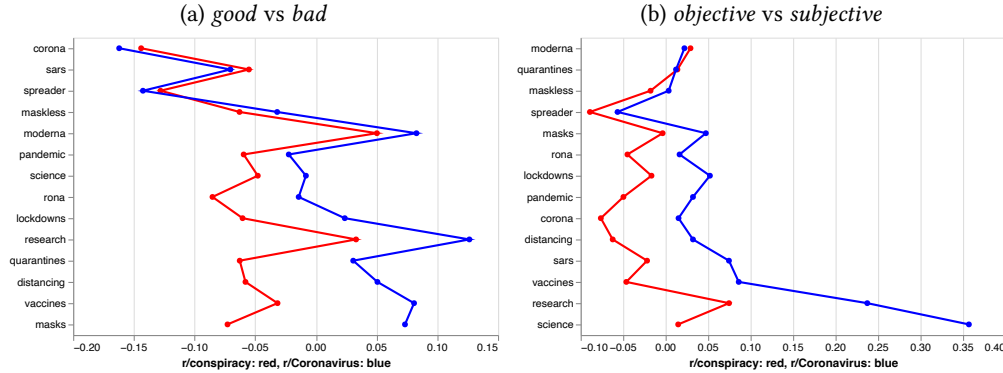
To study to which degree the neologisms show socio-semantic variation, we train community-specific word embeddings models for both communities. We follow the same procedure as for the detection of semantic change over time above, but measure semantic distances between two social models instead of two diachronic models. After aligning the models, we determine differences between the community-specific representations based on cosine distance. To generalise beyond single cases, we study social differences for the entire set of neologisms by analysing variation along two semantic dimensions.

As pointed out in Section 2.1, semantic variation often involves differences along an evaluative dimension (Koch 2016). We aim to detect such evaluative socio-semantic differences by projecting the semantic representations of the detected neologisms on an evaluative continuum. We follow the methodology proposed by An et al. (2018) and construct an evaluative semantic axis between the pole words *good* and *bad*.⁹ We then project the target words on this axis and use cosine similarity to determine whether the target word’s meaning is more positively or negative connotated.

On a second dimension, we aim to study whether the community-specific semantic representations show signs of (inter-)subjectification (Koch 2016). We follow the same procedure

⁹To make this semantic axis more robust, we include the ten nearest semantic neighbours for each pole word (e.g. *fantastic* or *terrible*) and use the average vector representation for each pole.

Figure 2: Projecting semantic representations for the communities *r/Coronavirus* and *r/conspiracy* on two semantic axes. Higher values in cosine similarity between the target words and the semantic axes indicate closer association with *good* and *objective*, respectively.



outlined above, and project the target neologisms on a second semantic axis defined by the pole words *subjective* and *objective*.

Figure 2 presents the results for the projections on both dimensions. It covers the set of semantic neologisms detected in the previous step, except for those unrelated to Covid¹⁰ and those that were not used in the selected communities¹¹. To compensate for this removal, we added five words that are known to be controversially discussed in Covid discourse (Ullah Shaheen et al. 2021): *masks*, *lockdowns*, *vaccines*, *science*, and *research*.

Figure 2a, which covers the evaluative dimension, shows that the Covid-related semantic neologisms are generally more negatively connotated in *r/conspiracy* than in *r/Coronavirus*. The cosine similarity between the target words’ semantic representations and the semantic axis is generally lower for this community. Notably, words that are more neutral or specific in their semantic scope such as *corona*, *sars*, and *moderna* show little difference between communities. However, terms that are associated with sociopolitical measures to fight the pandemic such as *masks*, *vaccines*, and *distancing* are evaluated much more negatively in *r/conspiracy*.

Next, Figure 2b presents our results for projecting the same set of words onto the axis of *subjective* vs *objective*. On the whole, the target words are more closely associated with the subjective pole in *r/conspiracy*. Similarly to the results on the evaluative dimension, the first set of more neutral terms such as *moderna*, *spreader*, and *maskless* exhibits little difference between communities. However, negatively connotated words related to government measures such as *vaccines*, *distancing*, and *lockdowns* are more closely aligned with subjectivity in *r/conspiracy*. Finally, the terms *science* and *research* show the greatest differences between the two communities. While *r/Coronavirus* associates both terms very strongly with objectivity, this is not the case for the conspiracy community.

Overall, we observe consistent patterns of socio-semantic variation along the evaluative and subjective dimensions, which indicate substantial semantic differences between the two communities. Covid-related terms such as *vaccines* or *distancing* are more closely associated

¹⁰ *sunsetting*, *childe*, *megalodon*, *newf*, *klea*, *rittenhouse chaz* were excluded because they are not related to Covid.

¹¹ *vacuo*, *filtrate*, and *cerb* had to be excluded because they were not used in the selected communities.

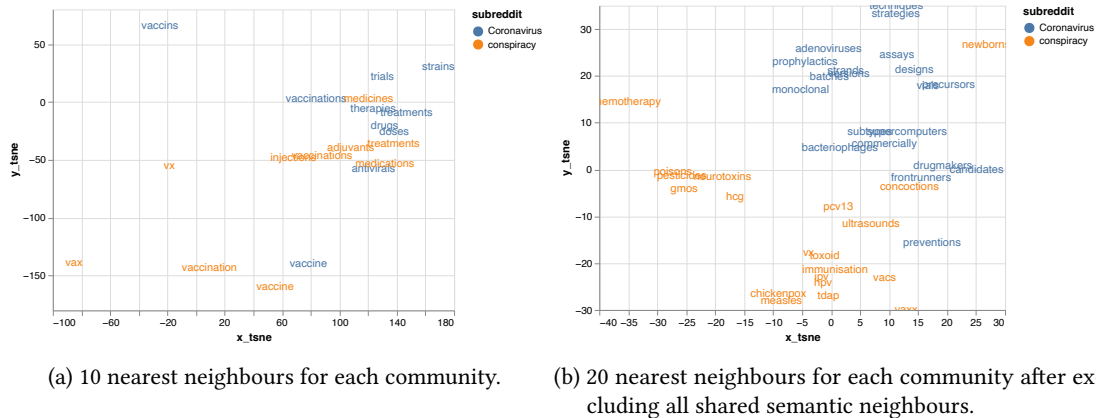


Figure 3: Semantic maps for the meaning of the term *vaccines* in the communities *r/Coronavirus* and *r/conspiracy*.

with negativity and subjectivity for *r/conspiracy*. This is also true for more general terms such as *research* and *science*, which are evaluated more critically in this community. These results are in keeping with the community’s general goal to ‘challenge issues which have captured the public’s imagination’, as stated in its community description. This scepticism seems to extend to Covid-related issues and to research and science as authoritative sources of objective truth.

5.2.3 Maps of socio-semantic variation

Lastly, we aim to get a more differentiated picture of the types of socio-semantic variation detected above. We visualise the semantic space of the word *vaccines*, which showed the highest degree of socio-semantic variation in Figure 2, and zoom in on the commonalities and differences in the community-specific semantic representations of *r/Coronavirus* and *r/conspiracy*. To this end, we align the embedding models of both communities and use t-SNE (Maaten and Hinton 2008) for reducing the dimensionality of the vector representations to visualise lexical meanings in a two-dimensional space.

Figure 3 presents two perspectives on the semantic space representing the meaning of *vaccines*, highlighting similarities and differences between the two communities. Figure 3a shows the ten nearest semantic neighbours for each community. Overall, the plot does not show strong differences in the semantic space of *vaccines* between *Coronavirus* and *conspiracy*. The semantic representations of its nearest neighbours show no clear separation by communities. Instead, they form semantic clusters. The bulk of the neighbours is located towards the top right and covers core aspects of the meaning of *vaccines*. There is high overlap between both communities, which is reflected in the close proximity of shared neighbours such as *vaccinations* and *treatments*.

The remaining neighbours form two groups. A first cluster at the centre bottom contains the singular form of the target word *vaccine* and its suffixation *vaccination*. While these terms are semantically most closely associated with *vaccines*, they are distinct from the bigger cluster in that they are singular forms. Our models capture these grammatical, more abstract differences, which is in accordance with previous studies that have shown the impact of word classes on the

semantic representations of word embedding models (e.g. Giulianelli et al. 2020). The remaining three terms are scattered across the semantic space. Notably, they are outliers in that they are all orthographic variants (*vax*, *vx*) or spelling mistakes (*vaccins*). Overall, the nearest neighbours in Figure 3a thus show a very high degree of overlap for the term *vaccines*, which indicates that both communities are in agreement about its core meaning.

To get a better view of the semantic differences between the two communities, in a last step, we now focus on the discrepancies in the semantic space of the term *vaccines*. Figure 3b presents the 20 nearest neighbours for each community after having filtered out those words that are shared between both communities. The resulting semantic space of *vaccines* now shows a clear separation between both communities. The semantic neighbours for *r/Coronavirus* cluster towards the top right of the plot. They cover a broad range of vaccine-related terms, including biological terms (*adenoviruses*) and terms related to vaccine development (*assays*, *supercomputers*) and the pharmaceutical industry (*drugmakers*).

The semantic neighbours for *r/conspiracy* are less diverse and show two main clusters. The first set towards the bottom centre covers vaccines that are unrelated to Covid: e.g. *chickenpox*, *measles*, and *hpv*. The appearance of these terms can be explained by the fact that speakers in this community are not only critical of the Covid vaccines, but show general scepticism towards vaccination. The second main group in the mid left part of the plot contains terms that are associated with conspiracy theories. These theories generally regard the vaccines as dangerous and claim that they cause a range of bio-medical side effects: causing brain damage due to *neurotoxins* and decreased fertility due to *hcg*¹², and turning people into genetically modified organisms (*gmos*).

These differences in the semantic space of the term *vaccines* in the two communities shed light on the socio-semantic variation on the evaluative and subjective dimensions identified by our embeddings projection approach in the previous section. The more negative evaluation in *r/conspiracy* seems to reflect the prevalent fears of side effects in this community. The stronger associations of subjectivity seem to go back to doubts about the objectivity and reliability of the established consensus in science and politics. The conspiracy theories shared in this community question the safety and efficacy of the vaccines.

6 Discussion

In this paper, we found that the coronavirus pandemic has caused considerable semantic innovation in the English lexicon.

Word embeddings models have allowed us to identify Covid-related semantic neologisms. While previous approaches to semantic change detection covered decades and centuries (e.g. Kim et al. 2014; Hamilton et al. 2016), and more recently also short-term meaning shifts over five to ten years (e.g. Del Tredici, Fernández and Boleda 2019; Shoemark et al. 2019), our approach managed to capture semantic changes in an even shorter period between 2019 and 2020. This indicates the potential of word embedding models for studying very recent semantic change and the strength of the impact of Covid on language and society.

¹²a hormone for the maternal recognition of pregnancy

Our results show that the large majority of words that show the greatest shifts between 2019 and 2020 are related to the pandemic. After a closer inspection, we find that the detected shifts represent different types of semantic variation over time. The first group contains homonyms like the proper noun *rittenhouse* that fall outside the scope of semantic neology. The detected neologisms mainly show denotational change, as in the semantic specialization of *distancing*. Besides, we find connotational changes that can be related to stylistic (e.g. *filtrate*) and social (e.g. *sunsetting*) dimensions of meaning.

While, on the whole, our semantic change detection method managed to identify neologisms with considerable success, our results also highlight the importance of distinguishing between these different types of detected changes. A closer inspection was necessary to exclude cases of homonymy and to distinguish between denotational and connotational change, which contributes to a more accurate picture of the semantic changes at play. More differentiated analyses of candidates for semantic changes have been outside the focus of most previous approaches in NLP, but recent approaches using token-based embeddings show promising results in that direction (Giulianelli et al. 2020).

In the second part of the paper, we focused on the socio-semantic variation of the Covid-related neologisms. Based on theoretical models of semantic change (Koch 2016), we analysed socio-semantic variation on the evaluative (*good vs bad*) and subjective (*subjective vs objective*) dimensions of meaning. Our results showed significant differences between communities, with *r/conspiracy* having more negative and subjective associations with our sample of Covid-related neologisms.

To get a more detailed view of socio-semantic variation, we then analysed the semantic space of the term *vaccines* by visualising its nearest semantic neighbours in both communities. We found considerable overlap between communities, which indicates that its core denotational meaning is shared between both communities. Yet, we also found denotational and connotational differences, which match the general pattern of differences found to its evaluative and subjective dimensions of meaning.

We argue that the considerable degree of socio-semantic variation found shows stronger effects in polarised echo chambers like *r/conspiracy*. Community-specific semantic representations are more likely to diverge from the norms in the speech community at large if communities are isolated from the ideas and linguistic influence from outside. Speakers with negative attitudes towards terms such as *vaccines* are more likely to reinforce and amplify their subjective associations in echo chambers of like-minded individuals.

Moreover, our results emphasise the importance of integrating socio-semantic variation in studies of semantic change. Studying socio-semantic variation can inform investigations of meaning change since community-specific variation can drive semantic change through processes of subjectification, ameliorization, and pejoration (Koch 2016). Besides, studies of semantic change that fail to account for community-specific effects might mistake the variation found for diachronic change. This is of particular relevance to studies using social media data, in which strong deviations of polarised communities might distort aggregate measures of semantic change. In such cases, considering socio-semantic variation can contribute to a more differentiated picture of semantic change and can simultaneously provide a lens into the underlying social differences between communities.

7 References

- An, Jisun, Haewoon Kwak and Yong-Yeol Ahn (2018). ‘SemAxis: A Lightweight Framework to Characterize Domain-Specific Word Semantics Beyond Sentiment’. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2018. Melbourne, Australia: Association for Computational Linguistics, pp. 2450–2461. DOI: 10.18653/v1/P18-1228.
- Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire and Jeremy Blackburn (2020). ‘The Pushshift Reddit Dataset’. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14, pp. 830–839.
- Bojanowski, Piotr, Edouard Grave, Armand Joulin and Tomas Mikolov (2017). ‘Enriching Word Vectors with Subword Information’. URL: <http://arxiv.org/abs/1607.04606> (visited on 04/02/2022).
- Clark, Herbert H. (1996). *Using Language*. ‘Using’ Linguistic Books. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511620539.
- Davies, Mark (2019–). *The Coronavirus Corpus*. URL: <https://www.english-corpora.org/corona/> (visited on 23/06/2021).
- Del Tredici, Marco and Raquel Fernández (2017). ‘Semantic Variation in Online Communities of Practice’. In: *IWCS 2017 - 12th International Conference on Computational Semantics - Long Papers*. URL: <https://aclanthology.org/W17-6804> (visited on 03/03/2022).
- Del Tredici, Marco, Raquel Fernández and Gemma Boleda (2019). ‘Short-Term Meaning Shift: A Distributional Exploration’. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. NAACL-HLT 2019. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 2069–2075. DOI: 10.18653/v1/N19-1210.
- Dong, Jihua, Louisa Buckingham and Hao Wu (2021). ‘A Discourse Dynamics Exploration of Attitudinal Responses towards COVID-19 in Academia and Media’. In: *International Journal of Corpus Linguistics* 26.4, pp. 532–556. DOI: 10.1075/ijcl.21103.don.
- Firth, John R. (1957). *A Synopsis of Linguistic Theory, 1930-1955*. Studies in Linguistic Analysis. Special Volume of the Philological Society. Oxford: Basil Blackwell.
- Geeraerts, Dirk (2010). *Theories of Lexical Semantics*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780198700302.001.0001.
- (2015). *How Words and Vocabularies Change*. The Oxford Handbook of the Word. DOI: 10.1093/oxfordhb/9780199641604.013.026.
- Giulianelli, Mario, Marco Del Tredici and Raquel Fernández (2020). ‘Analysing Lexical Semantic Change with Contextualised Word Representations’. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Online: Association for Computational Linguistics, pp. 3960–3973. DOI: 10.18653/v1/2020.acl-main.365.
- Gonen, Hila, Ganesh Jawahar, Djamé Seddah and Yoav Goldberg (2020). ‘Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora’. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Online: Association for Computational Linguistics, pp. 538–555. DOI: 10.18653/v1/2020.acl-main.51.

- Green, Jon, Jared Edgerton, Daniel Naftel, Kelsey Shoub and S. Cranmer (2020). 'Elusive Consensus: Polarization in Elite Communication on the COVID-19 Pandemic'. In: *Science Advances*. DOI: 10.1126/sciadv.abc2717.
- Hamilton, William L., Jure Leskovec and Dan Jurafsky (2016). 'Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change'. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL). Berlin, Germany: Association for Computational Linguistics, pp. 1489–1501. URL: <http://www.aclweb.org/anthology/P16-1141>.
- Hart, P. Sol, Sedona Chinn and Stuart Soroka (2020). 'Politicization and Polarization in COVID-19 News Coverage'. In: *Science Communication* 42.5, pp. 679–697. DOI: 10.1177/1075547020950735.
- Hasan, Ruqaiya (1989). 'Semantic Variation and Sociolinguistics'. In: *Australian Journal of Linguistics* 9.2, pp. 221–275. DOI: 10.1080/07268608908599422.
- Hofmann, Valentin, Janet B. Pierrehumbert and Hinrich Schütze (2021). 'Modeling Ideological Agenda Setting and Framing in Polarized Online Groups with Graph Neural Networks and Structured Sparsity'. URL: <http://arxiv.org/abs/2104.08829> (visited on 25/04/2021).
- Irshad, Sadia, Sadia Arshad and Kaukab Saba (2021). 'Lexicogrammatical Features of Covid-19: A Syntagmatic and Paradigmatic Corpus Based Analysis'. In: *CORPORUM: Journal of Corpus Linguistics* 4.2, pp. 76–94. URL: <https://journals.au.edu.pk/ojs/src/index.php/crc/article/view/167>.
- Jiang, Julie, Emily Chen, Shen Yan, Kristina Lerman and Emilio Ferrara (2020). 'Political Polarization Drives Online Conversations about COVID-19 in the United States'. In: *Human Behavior and Emerging Technologies* 2.3, pp. 200–211. DOI: 10.1002/hbe2.202.
- Jing, Elise and Yong-Yeol Ahn (2021). 'Characterizing Partisan Political Narrative Frameworks about COVID-19 on Twitter'. In: *EPJ Data Science* 10.1 (1), pp. 1–18. DOI: 10.1140/epjds/s13688-021-00308-4.
- Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde and Slav Petrov (2014). 'Temporal Analysis of Language through Neural Language Models'. In: *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Baltimore, MD, USA: Association for Computational Linguistics, pp. 61–65. DOI: 10.3115/v1/W14-2517.
- Koch, Peter (2016). 'Meaning Change and Semantic Shifts'. In: *The Lexical Typology of Semantic Shifts* 58, p. 21.
- Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski and Erik Velldal (2018). 'Diachronic Word Embeddings and Semantic Shifts: A Survey'. In: *Proceedings of the 27th International Conference on Computational Linguistics*. COLING 2018. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1384–1397. URL: <https://www.aclweb.org/anthology/C18-1117> (visited on 03/08/2020).
- Lang, Jun, Wesley W. Erickson and Zhuo Jing-Schmidt (2021). '#MaskOn! #MaskOff! Digital Polarization of Mask-Wearing in the United States during COVID-19'. In: *PLOS ONE* 16.4, e0250817. DOI: 10.1371/journal.pone.0250817.
- Leech, Geoffrey N. (1981). *Semantics*. 2nd ed. Harmondsworth: Penguin Books.
- Lipka, Leonhard (1992). *An Outline of English Lexicology*. Forschung Und Studium Anglistik. Tübingen: Niemeyer.

- Maaten, Laurens van der and Geoffrey Hinton (2008). 'Visualizing Data Using T-SNE'. In: *Journal of Machine Learning Research* 9.86, pp. 2579–2605. URL: <http://jmlr.org/papers/v9/vandermaaten08a.html> (visited on 07/03/2022).
- Mahlberg, Michaela and Gavin Brookes (2021). 'Language and Covid-19: Corpus Linguistics and the Social Reality of the Pandemic'. In: *International Journal of Corpus Linguistics* 26.4, pp. 441–443. DOI: 10.1075/ijcl.00043.mah.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeff Dean (2013). 'Distributed Representations of Words and Phrases and Their Compositionality'. In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc. URL: <https://papers.nips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html> (visited on 22/08/2021).
- Peirsman, Yves, Kris Heylen and Dirk Geeraerts (2010). 'Applying Word Space Models to Sociolinguistics. Religion Names before and after 9/11'. In: *Advances in Cognitive Sociolinguistics*. De Gruyter Mouton, pp. 111–138. DOI: 10.1515/9783110226461.111.
- Rehurek, Radim and Petr Sojka (2011). 'Gensim–Python Framework for Vector Space Modelling'. In: *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3.2.
- Robertson, Alexander, Farhana Ferdousi Liza, Dong Nguyen, Barbara McGillivray and Scott A. Hale (2021). 'Semantic Journeys: Quantifying Change in Emoji Meaning from 2012-2018'. URL: <http://arxiv.org/abs/2105.00846> (visited on 06/05/2021).
- Roig-Marín, Amanda (2020). 'English-Based Coroneologisms: A Short Survey of Our Covid-19-Related Vocabulary'. In: *English Today*, pp. 1–3.
- Schmid, Hans-Jörg (2020). *The Dynamics of the Linguistic System. - Usage, Conventionalization, and Entrenchment*. Oxford: Oxford University Press.
- Schmid, Hans-Jörg, Quirin Würschinger, Melanie Keller and Ursula Lenker (2020). 'Battling for Semantic Territory across Social Networks. The Case of Anglo-Saxon on Twitter'. In: *Yearbook of the German Cognitive Linguistics Association* 8.1, pp. 3–26. DOI: 10.1515/gcla-2020-0002.
- Scott, Ben (2020). 'Know Your Covidiot From Your Cove-Dwellers'. In: *Bloomberg.com*. URL: <https://www.bloomberg.com/opinion/articles/2020-04-03/coronavirus-know-your-covidiot-from-your-cove-dwellers> (visited on 22/08/2021).
- Shoemark, Philippa, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale and Barbara McGillivray (2019). 'Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings'. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 66–76.
- Thorne, Tony (2020). #CORONASPEAK – the Language of Covid-19 Goes Viral – 2. Language and innovation. URL: <https://language-and-innovation.com/2020/04/15/coronaspeak-part-2-the-language-of-covid-19-goes-viral/> (visited on 12/02/2021).
- Tournier, Jean (1985). *Introduction Descriptive à La Lexicogénétique de l'anglais Contemporain*. Paris: Champion-Slatkine.
- Traugott, Elizabeth Closs (2010). '(Inter)Subjectivity and (Inter)Subjectification: A Reassessment'. In: *Subjectification, Intersubjectification and Grammaticalization*. Ed. by Kristin Davidse, Lieven

- Vandelanotte and Hubert Cuyckens. De Gruyter Mouton, pp. 29–74. doi: doi : 10 . 1515/9783110226102 . 1 . 29.
- Tsakalidis, Adam, Marya Bazzi, Mihai Cucuringu, Pierpaolo Basile and Barbara McGillivray (2019). ‘Mining the UK Web Archive for Semantic Change Detection’. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. RANLP 2019. Varna, Bulgaria: INCOMA Ltd., pp. 1212–1221. doi: 10 . 26615/978-954-452-056-4_139.
- Ullah Shaheen, Zafar, Ayyaz Qadeer and Fouzia Rehman Khan (2021). ‘Conspiracy Theories (CT) vs Truth Based Reporting: A Corpus Driven Analysis of Covid-19 Online Newspaper(s) Discourse’. In: *CORPORUM: Journal of Corpus Linguistics* 4.2, pp. 112–135. url: <https://journals.au.edu.pk/ojsrcr/index.php/crc/article/view/169>.

6.3 Conclusions

This paper identified Covid-related semantic neologisms on Reddit and found that these terms exhibit considerable socio-semantic variation between communities.

In a first step, the study found a high proportion of Covid-related terms among the candidates for semantic change detected by the diachronic word embeddings approach. This validates the effectiveness of the approach since the high prevalence of Covid-related terms is in accordance with expectations. The pandemic has caused substantial socio-cultural changes and has introduced many new concepts, which is likely to be reflected in lexical innovation. Due to the prominence of Covid in public discourse, pandemic-related neologisms have experienced an extended period of high topicality (Fischer 1998) and high semantic carrying capacity (Nini et al. 2017), which has been shown to enhance the likelihood of successful diffusion in Chapter 4. The results of the semantic change detection approach are also consistent with previous studies, which have identified and studied a diverse set of Covid-related semantic neologisms (Dong, Buckingham & Wu 2021; Irshad, Arshad & Saba 2021; Ullah Shaheen, Qadeer & Rehman Khan 2021).

The fact that the present approach was able to detect recent semantic neologisms is promising for future studies on semantic innovation. Most previous approaches using word embeddings were unsuitable for studying semantic neologisms because they were limited to studying semantic change over decades and centuries (e.g. Kim et al. 2014; Hamilton, Leskovec & Jurafsky 2016). In line with more recent approaches detecting short-term meaning shifts over five to ten years (e.g. Del Tredici, Fernández & Boleda 2019; Shoemark et al. 2019), our approach managed to detect semantic changes in a period of only two years.

This demonstrates the potential of word embeddings for studying recent semantic innovation. As described in Section 1.2.4, previous work on lexical innovation lacked the necessary methods to conduct large-scale studies of semantic neologisms. In future work, the present approach could be applied to detect a larger, more representative sample of semantic neologisms, and thus enable studies on the diffusion of semantic neologisms comparable to the investigation of formal neologisms presented in Chapter 4.

In addition, the paper found considerable socio-semantic variation between communities. The subreddit *r/conspiracy*, which is more critical of sociopolitical policies regarding the pandemic, showed more negative and subjective associations with Covid-related neologisms than the neutral subreddit *r/Coronavirus*.

This analysis expands on the findings presented in the preceding chapter (5). The term *Anglo-Saxon* was also found to be used with different meanings by different communities on Twitter: generally, far-right groups use it more in its ethno-racial sense with positive connotations, while center-to-left and academic communities oppose this sense of the term and use it primarily in its neutral, historical sense or have stopped using the term altogether. The approach employed in this chapter enhances the earlier study by using data obtained from Reddit, which enables more robust interpretations of group

differences due to the fact that Reddit, unlike Twitter, is organised into communities that are centered around explicit shared interests of their members.

Moreover, the present study provides a more comprehensive, quantitative investigation by measuring socio-semantic variation based on community-specific word embedding models, as opposed to the manual, qualitative case study conducted for *Anglo-Saxon*. The data-driven identification of candidates for semantic change enabled the investigation of a larger sample than was previously possible. Owing to the quantitative design of the approach, communities-based differences could be analysed based on all words in this extended sample, allowing for a more comprehensive view of community-based differences than the study of a single lexeme.

In addition, the paper sought to expand on the previous studies by presenting a more detailed picture of the nature of semantic differences between communities. Motivated by previous research on semantic change (e.g. Geeraerts 2010; Koch 2016), which has underscored the significance of meliorization, pejorization, and subjectification in semantic change, the study analysed semantic differences along the dimensions *good vs bad*, and *objective vs subjective*. The results demonstrated consistent patterns of socio-semantic variation along these dimensions; both Covid-related terms such as *vaccines* and more general terms such as *research* are more closely associated with negativity and subjectivity for the sceptic community *r/conspiracy* than for the neutral community *r/Coronavirus*.

These findings relate back to the qualitative analyses in the previous chapters. Chapter 3 found that *rapefugee* had significantly more positive connotations among far-right activists, whereas the term was used much more critically by the majority of other users on Twitter. Chapter 5 discovered analogous evaluative differences across the political spectrum for the use of *Anglo-Saxon*. The present study offered a more principled and comprehensive account of these effects by leveraging a broader quantitative framework and using word embeddings to analyse socio-semantic variation.

This approach could be expanded in a number of ways through future work. Broadening the scope beyond the context of Covid and analysing variation across a larger number of communities would provide a more comprehensive and representative picture of socio-semantic variation. Moreover, this methodology has great potential for studies of semantic change. Due to the adaptability of the presented method, this could, for instance, be utilised to investigate the role of semantic dimensions such as *abstract vs concrete* in grammaticalization processes. Lastly, the current approach could be extended to examine the interaction between socio-semantic variation and semantic change. The study of *Anglo-Saxon* in the preceding chapter demonstrated that social variation and divergent semantic preferences between communities can induce semantic change. Extending the current method to study a bigger sample of communities over a longer period of time could help investigate the interplay between social variation and change on a large scale.

7 Conclusion

7.1 Research objectives

The aim of this dissertation was to investigate the emergence and diffusion of English neologisms on the web and social media, with a particular focus on the social dynamics of diffusion and social variation between communities of speakers.

As a first step toward achieving this overarching goal, I sought to collect a diverse sample of neologisms that is representative of the variety of lexical innovation. This includes formal neologisms emerging on the web such as *Internet of things* and Twitter such as *twitterverse*, as well as semantic neologisms such as *distancing*.

Based on this sample, and following Schmid's definition of diffusion (Schmid 2020: 178–179), I aimed to add to previous work on lexical innovation by providing a more differentiated view of the diffusion of neologisms by examining to what extent they spread across different usage contexts and speakers and communities. Therefore, it was essential to cover the entire conventionalization continuum (Kerremans 2015) and to be able to distinguish between cases like *microflat*, which failed to catch on at all, cases like *man bun*, which demonstrate advanced conventionalization, and neologisms that have diffused to some extent, yet remain confined to certain usage contexts (e.g. *twitterverse*) or specific communities (e.g. *rapefugee*). Due to the inclusion of semantic neologisms, this also entailed determining whether new meanings of existing words such as *Anglo-Saxon* have spread and become widely conventional, or whether these words have different meanings in different communities.

The aim of providing a thorough examination of a diverse set of lexical innovations, as well as a detailed analysis of their diffusion across usage contexts and communities presented a number of methodological challenges. Pursuing and achieving this objective required a series of adjustments and extensions, including utilizing a variety of data sources from the web, Twitter, and Reddit, as well as the application of several methods such as frequency-based analyses, social network analysis, and word embeddings.

Chapters 2 to 6 presented my steps towards this goal, with each chapter attempting to build upon the prior research in order to get closer to the overarching objective of studying the emergence, diffusion and social variation of lexical innovations. The following sections provide a conclusion to this dissertation by discussing my main findings, their theoretical and methodological implications, and propose directions for future work.

7.2 Emergence

The first objective in this dissertation was to collect a large sample of neologisms that is representative of the variety of lexical innovation on the web and social media, including both formal and semantic neologisms. As mentioned in Section 1.2.2, previous research has not provided an exhaustive account of the variety of emerging lexical innovations.

7.2.1 Formal neologisms

Earlier studies have predominantly focused on formal neologisms. In addition, they have focused primarily on particular domains of formal neologisms. Previous investigations have provided case studies of manually selected examples (e.g. Hohenhaus 2006), and specific text types (e.g. Elsen 2004) or semantic fields (e.g. Foubert & Lemmens 2018). Web corpora and specialised tools have facilitated the examination of larger samples, but most of the research has been confined to specific parts of the web, such as news, which feature more formal language (Gérard et al. 2017; Cartier 2017). The NeoCrawler (Kerremans, Stegmayr & Schmid 2012) sought to surmount these limitations by identifying neologisms on the open web, which enabled a broader view of the use of neologisms across various types of sources, such as blogs or discussion forums (Kerremans 2015).

In this dissertation, I used an extended version of the NeoCrawler to detect a large and diverse sample of formal neologisms. As described in Chapter 2, the NeoCrawler's ability to identify lexical innovations could be substantially improved by two main extensions. Firstly, partial string matching based on Levenshtein Distance (Levenshtein 1965) significantly increased the number of high-quality candidates of formal neologisms. Secondly, I supplemented the detection of neologisms on the web with additional social media data from Twitter, which I gathered as random samples of tweets using the tool TAGS (Hawksey 2020).

Through periodic Discoverer searches, managed to identify a significantly larger sample of neologisms than previously possible: the expanded database contained 958 neologisms, a substantial increase from the previous study using the NeoCrawler by Kerremans (2015), which was based on 40 neologisms. A formal analysis of the sample revealed that its distribution of word-formation processes aligns with prior research and OED data, indicating that the detected set of neologisms represents a wide range of lexical innovation. Furthermore, monitoring the diffusion of this larger sample over an extended period of time with the NeoCrawler's Observer module revealed that it captures a broad spectrum of diffusion, as gauged by cumulative usage frequency. The diffusion of the neologisms in this sample was analysed in further detail in the Chapters 3 and 4, which will be discussed in greater detail below.

7.2.2 Semantic neologisms

As described in Section 1.2.4, semantic neologisms have received little attention in prior research on lexical innovation. This is largely attributed to the heightened methodological challenge in discovering semantic neologisms. Word embedding approaches (Mikolov et al. 2013), which can be used to examine the emergence of semantic neologisms, have only very recently become accessible as a result of advances in Natural Language Processing.

Due to the intricacy of using word embedding models, however, the majority of word embedding applications have stayed within the field of Natural Language Processing. Previous work using word embeddings in Natural Language Processing has mostly focused on gradual, long-term meaning change (e.g. Kim et al. 2014; Hamilton, Leskovec & Jurafsky 2016; Kutuzov et al. 2018). Recent research has begun to focus on short-term lexical semantic change on the scale of years rather than decades (Shoemark et al. 2019; Del Tredici, Fernández & Boleda 2019; Tsakalidis et al. 2019), but semantic neologisms and the underlying characteristics of lexical innovation have received little attention.

To include semantic lexical innovations in my study of emergence and diffusion, as described in Chapter 6, I attempted to identify semantic neologisms by employing word embeddings. I investigated semantic innovation between the years 2019 and 2020, as previous research based on manually selected cases suggested that the Covid pandemic has resulted in semantic changes for a number of Covid-related terms (Dong, Buckingham & Wu 2021; Irshad, Arshad & Saba 2021; Ullah Shaheen, Qadeer & Rehman Khan 2021).

To identify semantic neologisms in a data-driven manner I utilised two extensive corpora for the years 2019 and 2020 obtained from Reddit, generated semantic representations for each word in the corpora using word2vec (Mikolov et al. 2013), and measured their diachronic semantic distance. Words demonstrating the largest distances were deemed to have undergone semantic innovation.

As described in further detail in Section 6.3, this method proved adept at discovering semantic neologisms with a high degree of precision. According to expectations, a large proportion of the detected candidates for semantic change between 2019 and 2020 were Covid-related terms. This affirms the effectiveness of the approach, aligning with the anticipated high prevalence of Covid-related terms.

The pandemic has resulted in considerable socio-cultural shifts and the introduction of numerous new concepts, which are likely to be reflected in lexical innovation. These findings are also compatible with earlier research on Covid-related semantic innovation based on manually selected cases (Dong, Buckingham & Wu 2021; Irshad, Arshad & Saba 2021; Ullah Shaheen, Qadeer & Rehman Khan 2021).

In addition, from a methodological standpoint, these results are encouraging because our approach was able to detect semantic changes in a period of only two years, which is even shorter than previous research on short-term meaning shift, which examined periods of five to ten years (e.g. Del Tredici, Fernández & Boleda 2019; Shoemark et al. 2019),

More generally, this approach has demonstrated the potential of word embeddings for studying the emergence of recent semantic innovations. As described in Section 1.2.4, prior research on lexical innovation lacked the methods required to conduct large-scale studies of semantic neologisms. In future research, the current method could be leveraged to identify larger, more representative samples of semantic neologisms, facilitating studies on the diffusion of semantic neologisms comparable to the investigation of formal neologisms presented in Chapter 4.

7.3 Diffusion

The second goal of this dissertation was to investigate the extent to which the neologisms in my sample show diffusion across different usage contexts and across speakers and communities. The NeoCrawler project (Kerremans, Stegmayr & Schmid 2012) aimed to gain a more differentiated view of the diffusion of neologisms. In addition to frequency of occurrence, Kerremans (2015) examined the use of neologisms in a variety of text types on the web, such as news and blogs, as an indicator of diffusion. However, due to the technical challenges described in Section 2.3, the NeoCrawler was unable to achieve its intended purpose of facilitating the study into the diffusion of neologisms across usage contexts. As a consequence, this dissertation has mainly focused on investigating the social dynamics of diffusion, i.e. to what extent neologisms spread across speakers and communities.

As described in Section 1.3.3, the social dynamics involved in the adoption of linguistic innovations are fundamental to well-established models of diffusion, such as the S-curve model (Milroy & Milroy 1985; Labov 2007; Nevalainen 2015). However, it has long been impossible to study the spread of innovations across speakers and communities directly based on corpus data.

Consequently, most previous studies had to rely on usage frequency to measure the degrees of social diffusion of neologisms. This approach is based on the underlying assumption that neologisms that have been used many times in the corpus are likely to be familiar to a large group of speakers (Stefanowitsch & Flach 2017). The reliance on usage frequency has been a common limitation of previous research studying the diffusion of neologisms on the web (e.g. Renouf, Kehoe & Banerjee 2007; Gérard et al. 2017; Cartier 2017).

This limitation also applies to the NeoCrawler, which served as the starting point of the investigation of diffusion in this dissertation in Chapter 2. The Observer module of the NeoCrawler allowed monitoring the spread of the neologisms in its database over an extended period of time. Assessing their degrees of diffusion as measured by cumulative usage frequency indicated that this sample encompasses a wide spectrum of diffusion. However, using frequency counts as indicators for diffusion turned out to be less reliable than expected due to the NeoCrawler's dependence on Google's search index.

The frequency-based comparative analysis of the NeoCrawler and Twitter data for the *rapefugee* terms in Chapter 3 demonstrated broad agreement between both datasets.

However, the analysis also revealed more general limitations associated with assessing social diffusion based solely on usage frequency. The results demonstrated the influence of certain communities in promoting higher usage frequency of the selected neologisms. Despite their high usage intensity, the qualitative analysis revealed that these terms exhibit relatively low degrees of social diffusion. They have not successfully spread across communities; their use is restricted to a small number of like-minded individuals and communities on the far-right of the political spectrum.

Chapter 4 addressed the issue of reliability of usage frequency to measure diffusion raised in Chapters 2 and 3, and extended the previous investigations of diffusion. Firstly, this study applied additional measures for the temporal dynamics of diffusion. The proposed methods aimed to provide a more comprehensive understanding of 'topicality' (Fischer 1998: 16) and 're-current semi-conventionalization' (Kerremans 2015: 129–136), as observed in Chapter 2. I examined trends in usage intensity as captured by cumulative and absolute frequency counts, the volatility of neologisms as measured by the coefficient of variation, and the age of neologisms as indicated by their observed lifespan since their emergence. I argued that the additional consideration of temporal dynamics captured by these measures contributes to a more precise picture of diffusion. In addition, I suggested that these temporal dynamics are of particular importance for the study of *lexical* innovation, since neologisms are strongly influenced by changes in their 'semantic carrying capacity' (Nini et al. 2017).

The paper also used social network analysis on Twitter to provide direct insights into social diffusion complementing the frequency-based measures of diffusion. To establish if neologisms successfully diffuse across speakers and communities, several network-related metrics (such as centralization and in-degree) and network graph visualisations were employed.

While the comparison of the degrees of diffusion of the neologisms between the frequency-based and network-based approaches revealed a substantial degree of congruence in their overall assessments, it also highlighted several cases where usage frequency seems to overestimate degrees of social diffusion. The analysis revealed a pattern of cases that resemble the case of *rapefugee*. The data indicate that *rapefugee* is in dubious company with a number of other politically charged neologisms such as *alt-left* or *birther*, which also exhibit high usage intensity and centralization. The small, politically polarized areas of the social network, which resemble ideological echo chambers, are where these terms are predominantly used and exhibit high degrees of social variation.

7.4 Socio-semantic variation

The last part of this dissertation covered socio-semantic variation. After examining diffusion and social variation of formal neologisms in the preceding three chapters, the last two addressed social variation among semantic neologisms, which has been outside the focus of earlier research on lexical innovation.

The study of social variation in Chapter 5 expands on the differences in social diffusion

found in the previous chapters. As in previous cases such as *rapefugee* (Chapter 3) or *alt-left* (Chapter 4), the use of *Anglo-Saxon* differs strongly between communities. Unlike the approaches in the preceding chapters, however, this study went beyond analysing differences in usage intensity and found that communities exhibit considerable social variation regarding the meaning of the term *Anglo-Saxon*.

Chapter 5 studied socio-semantic variation in the use of the term *Anglo-Saxon*, which is subject to an intense battle over its meaning. The present results indicate this increasing polarisation surrounding the term *Anglo-Saxon*. The network analysis revealed the growing centralization in its use over time, which the manual network analysis corroborated. While some communities continue to use the term in its previous meaning, others have stopped using it because they have come to associate it with the political far-right. The results show that there is a significant chance that the current social dynamics will cause sustained changes to its meaning and use, even though the future of its use and meaning is currently undetermined.

Chapter 6 looked at a wider range of semantic neologisms and found considerable socio-semantic variation in the use of Covid-related semantic neologisms on Reddit. Compared to the more neutral community *r/Coronavirus*, the community *r/conspiracy*, which is more critical of sociopolitical responses to the pandemic, displayed higher negative and subjective associations with Covid-related neologisms.

The goal of the paper was to provide a more thorough understanding of the nature of semantic differences between communities. Compared to the manual, qualitative case study of *Anglo-Saxon*, this chapter offered a more comprehensive, quantitative assessment by measuring socio-semantic variation based on community-specific word embedding models. Informed by prior research on semantic change (e.g. Geeraerts 2010; Koch 2016), which emphasise the significance of meliorization, pejorization, and subjectification in semantic change, this study examined semantic differences along the dimensions *good vs bad*, and *objective vs subjective*. The results showed consistent patterns of socio-semantic variation along these dimensions. For the sceptic community *r/conspiracy*, as opposed to the neutral community *r/Coronavirus*, Covid-related terms such as *vaccines* and more general terms such as *research* are more strongly associated with negativity and subjectivity.

This approach could be expanded in a number of ways through future work on semantic change and socio-semantic variation. This methodology has great potential for studies of semantic change. Due to the adaptability of the presented method, it could be applied, for instance, to investigate the role of semantic dimensions such as *abstract vs concrete* in grammaticalization processes. Moreover, a natural extension of this work is to extend the scope beyond the context of Covid and analyse variation across a larger number of communities to get a more comprehensive and representative picture of socio-semantic variation. Lastly, the current approach should be expanded to evaluate how socio-semantic variation and semantic change interact. The study of *Anglo-Saxon* in the preceding chapter demonstrated that social variation and divergent semantic preferences between communities can induce semantic change. Extending the present method to study a larger sample of communities over a longer period of

time would help investigate the relationship between social variation and change on a broad scale.

In this dissertation, I have investigated the emergence and diffusion of lexical innovations, with a particular focus on the social dynamics of diffusion. To this end, I have collected a large sample of formal and semantic neologisms and analysed their diffusion across the web, on Twitter, and on Reddit using analyses based on usage frequency, social networks, and word embeddings. By doing so, I hope to have contributed to a better view of the diverse phenomenon of lexical innovation.

Bibliography

- Algeo, John. 1998. Vocabulary. In Suzanne Romaine (ed.), *The Cambridge history of the English language*, vol. 4, IV 1776–1997 vols., 57–91. Cambridge: Cambridge University Press.
- Ayto, John. 2006. *Movers and shakers: a chronology of words that shaped our age*. Oxford University Press.
- Baroni, Marco, Georjina Dinu & Germán Kruszewski. 2014. Don't count, predict! - A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd annual meeting of the association for computational linguistics*, 238–247.
- Bauer, Laurie. 1983. *English word-formation*. Cambridge: Cambridge University Press.
- Blythe, Richard A. & William Croft. 2012. S-curves and the mechanisms of propagation in language change. *Language* 88(2). 269–304.
- BNC Consortium. 2007. *British National Corpus, XML Edition*. <http://hdl.handle.net/20.500.12024/2554>.
- Cabré Castellví, Maria Theresa & Rogelio Nazar. 2012. Towards a new approach to the study of neology. *Neologica*. <https://doi.org/10.15122/isbn.978-2-8124-4232-2.p.0063>.
- Cartier, Emmanuel. 2017. Neoveille, a web platform for neologism tracking. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 95–98. Valencia, Spain: Association for Computational Linguistics. <http://aclweb.org/anthology/E17-3024>.
- Cartier, Emmanuel. 2018. Néoveille, an automatic system for lexical units life-cycle tracking. In Jaka Cibej, Vojko Gorjanc, Iztok Kosem & Simon Krek (eds.), *The XVIII EURALEX International Congress. Lexicography in global contexts*. 34–35. Ljubljana, Slovenia.
- Conover, Michael, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer & Alessandro Flammini. 2011. Political Polarization on Twitter. In *ICWSM*, vol. 5, 89–96.
- Davies, Mark. 2008. *The Corpus of Contemporary American English (COCA): 560 Million Words, 1990-Present*. <https://www.english-corpora.org/coca/>.
- Davies, Mark. 2016. *Corpus of News on the Web (NOW) - 3+ Billion Words from 20 Countries, Updated Every Day*. <https://www.english-corpora.org/now/>.
- Del Tredici, Marco & Raquel Fernández. 2018. The Road to Success: Assessing the Fate of Linguistic Innovations in Online Communities.
- Del Tredici, Marco, Raquel Fernández & Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

7. Bibliography

- Technologies, Volume 1 (Long and Short Papers)*, 2069–2075. Minneapolis, Minnesota: Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1210>.
- Dong, Jihua, Louisa Buckingham & Hao Wu. 2021. A discourse dynamics exploration of attitudinal responses towards COVID-19 in academia and media. *International Journal of Corpus Linguistics* 26(4). 532–556. <https://doi.org/10.1075/ijcl.21103.don>.
- Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith & Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLOS ONE* 9(11). 1–13. <https://doi.org/10.1371/journal.pone.0113114>.
- Elsen, Hilke. 2004. *Neologismen. Formen und Funktionen neuer Wörter in verschiedenen Varietäten des Deutschen*. Tübingen: Narr.
- Firth, John R. 1957. *A synopsis of linguistic theory, 1930-1955* (Studies in Linguistic Analysis. Special Volume of the Philological Society). Oxford: Basil Blackwell.
- Fischer, Roswitha. 1998. *Lexical change in present day English. A corpus based study of the motivation, institutionalization, and productivity of creative neologisms*. Tübingen: Narr.
- Fonteyn, Lauren & E. Manjavacas Arevalo. 2021. Adjusting scope: a computational approach to case-driven research on semantic change. *Computational humanities research CEUR-WS* 2989. 280–298. <http://hdl.handle.net/1887/3280007>.
- Foubert, Océane & Maarten Lemmens. 2018. Gender-biased neologisms: the case of man-X. *Lexis. Journal in English Lexicology* (12). <https://doi.org/10.4000/lexis.2453>.
- Gatto, Maristella. 2014. *The web as corpus* (Studies in Corpus and Discourse). London: Bloomsbury.
- Geeraerts, Dirk. 2010. *Theories of Lexical Semantics*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198700302.001.0001>.
- Gérard, C, L Bruneau, Ingrid Falk, Delphine Bernhard & A.-L Rosio. 2017. Le Logoscope : observatoire des innovations lexicales en français contemporain. In Joaquín García Palacios, Goedele de Sterck, Daniel Linder, Jesús Torre del Rey, Miguel Sánchez Ibanez García & Nava Maroto (eds.), *La neología en las lenguas románicas: recursos, estrategias y nuevas orientaciones*. Frankfurt am Main: Peter Lang. <https://hal.archives-ouvertes.fr/hal-01388255>.
- Goel, Rahul, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz & Jacob Eisenstein. 2016. The social dynamics of language change in online networks. In Emma Spiro & Yong-Yeol Ahn (eds.), *Social informatics*, 41–57. Cham: Springer International Publishing.
- Granovetter, Mark S. 1973. The Strength of Weak Ties. *American Journal of Sociology* 78(6). 1360–1380. <https://doi.org/10.1086/225469>.
- Grieve, Jack. 2018a. Mapping emerging words in New York City. In *New Ways of Analyzing Variation (NWAY)*. New York, USA. <https://osf.io/9nu7v/>.

- Grieve, Jack. 2018b. Natural selection in the modern English lexicon. In *Proceedings of the International Conference on Language Evolution (EvoLang)*, 153–157. Torun, Poland. <https://doi.org/10.12775/3991-1.037>.
- Grieve, Jack, Andrea Nini & Diansheng Guo. 2016. Analyzing lexical emergence in Modern American English online. *English Language and Linguistics* (21). 99–127.
- Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In (Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)), 1489–1501. Berlin, Germany: Association for Computational Linguistics. <http://www.aclweb.org/anthology/P16-1141>.
- Hawkey, Matthew. 2020. TAGS. TAGS. <https://tags.hawkey.info/> (14 June, 2020).
- Hilpert, Martin & David Correia Saavedra. 2017. Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Linguistics and Linguistic Theory*. <https://www.degruyter.com/view/j/c11t.ahead-of-print/c11t-2017-0009/c11t-2017-0009.xml>.
- Hilpert, Martin & Susanne Flach. 2021. Disentangling modal meanings with distributional semantics. *Digital Scholarship in the Humanities* 36(2). 307–321. <https://doi.org/10.1093/l1c/fqaa014>.
- Himmelboim, Itai, S. McCreery & Marc A. Smith. 2013. Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter. *J. Comput. Mediat. Commun.* <https://doi.org/10.1111/jcc4.12001>.
- Hohenhaus, Peter. 1996. *Ad-hoc-Wortbildung. Terminologie, Typologie und Theorie kreativer Wortbildung im Englischen*. Frankfurt am Main: Lang.
- Hohenhaus, Peter. 2005. Lexicalization and institutionalization. In Pavol Stekauer & Rochelle Lieber (eds.), *Handbook of Word-Formation*, 353–373. Dordrecht: Springer.
- Hohenhaus, Peter. 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.
- Irshad, Sadia, Sadia Arshad & Kaukab Saba. 2021. Lexicogrammatical features of covid-19: A syntagmatic and paradigmatic corpus based analysis. *CORPORUM: Journal of Corpus Linguistics* 4(2). 76–94. <https://journals.au.edu.pk/ojs/crc/index.php/crc/article/view/167>.
- Jakubiček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý & Vít Suchomel. 2013. The TenTen corpus family. In *7th international corpus linguistics conference CL*, 125–127.
- Karjus, Andres. 2020. *Competition, selection and communicative need in language change: an investigation using corpora, computational modelling and experimentation*. University of Edinburgh dissertation.
- Kerremans, Daphné. 2015. *A Web of New Words*. Bern: Peter Lang. <https://doi.org/10.3726/978-3-653-04788-2>.
- Kerremans, Daphné, Jelena Prokić, Quirin Würschinger & Hans-Jörg Schmid. 2018. Using data-mining to identify and study patterns in lexical innovation on the web:

7. Bibliography

- The NeoCrawler. *Pragmatics and Cognition* 25(1). 174–200. <https://www.jbe-platform.com/content/journals/10.1075/pc.00006.ker>.
- Kerremans, Daphné, Susanne Stegmayr & Hans-Jörg Schmid. 2012. The NeoCrawler: Identifying and retrieving neologisms from the internet and monitoring ongoing change. In *Current Methods in Historical Semantics*, 59–96. Berlin: Mouton de Gruyter.
- Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde & Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 61–65. Baltimore, MD, USA: Association for Computational Linguistics. <https://doi.org/10.3115/v1/W14-2517>.
- Koch, Peter. 2016. Meaning change and semantic shifts. *The Lexical Typology of Semantic Shifts* 58. 21–66.
- Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski & Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, 1384–1397. Santa Fe, New Mexico, USA: Association for Computational Linguistics. <https://www.aclweb.org/anthology/C18-1117> (3 August, 2020).
- Labov, William. 2007. Transmission and diffusion. *Language* 83(2). 344–387.
- Lapata, Mirella & Alex Lascarides. 2003. Detecting novel compounds: The role of distributional evidence. In *10th Conference of the European Chapter of the Association for Computational Linguistics*. Budapest, Hungary: Association for Computational Linguistics. <https://aclanthology.org/E03-1073>.
- Leech, Geoffrey N. 1981. *Semantics*. 2nd edn. Harmondsworth: Penguin Books.
- Lemnitzer, Lothar. 2010. *Wortwarte*. <http://www.wortwarte.de/>.
- Leuckert, Sven. 2020. Towards a digital sociolinguistics. Communities of Practice on Reddit. In Sofia Rüdiger & Dasha Dayter (eds.), *Corpus Approaches to Social Media*, 15–40. John Benjamins Publishing Company. <https://benjamins.com/catalog/sc1.98.011eu>.
- Levenshtein, Vladimir. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10. 707–710.
- Link, Sabrina. 2021. *What makes a neologism a success story? An empirical Study of the diffusion of recent English Blends and German compounds*. Munich: LMU Munich. <http://nbn-resolving.de/urn:nbn:de:bvb:19-287605>.
- Lipka, Leonhard, Susanne Handl & Wolfgang Falkner. 2004. Lexicalization and institutionalization. The state of the art in 2004. *SKASE Journal of Theoretical Linguistics* (1). 2–19.
- Lotan, Nir & Einat Minkov. 2021. SocialVec: Social entity embeddings. *ArXiv* abs/2111.03514.
- Lu, Fred Sun, Suqin Hou, Kristin Baltrusaitis, Manan Shah, Jure Leskovec, Rok Susic, Jared Hawkins, John Brownstein, Giuseppe Conidi, Julia Gunn, Josh Gray, Anna Zink & Mauricio Santillana. 2018. Accurate influenza monitoring and forecasting using novel internet data streams: A case study in the boston metropolis. *JMIR Public Health and Surveillance* 4(1). <https://doi.org/10.2196/publichealth.8950>.

- Lukianoff, Greg & Jonathan Haidt. 2018. *The coddling of the American mind: How good intentions and bad ideas are setting up a generation for failure*. Penguin.
- Maier, Anna. 2016. *WebClass: Verfeinerte dokumentklassifizierung auf englischen webseiten*. Munich: LMU Munich BA thesis.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, vol. 26. Curran Associates, Inc. <https://papers.nips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html>.
- Milroy, James. 1992. *Linguistic variation and change: On the historical sociolinguistics of English*. Oxford: Blackwell.
- Milroy, James & Lesley Milroy. 1985. Linguistic change, social network and speaker innovation. *Journal of Linguistics* 21(2). 339–384.
- Nevalainen, Terttu. 2015. Descriptive adequacy of the S-curve model in diachronic studies of language change. *Studies in Variation, Contacts and Change in English* 16. <https://varieng.helsinki.fi/series/volumes/16/nevalainen/>.
- Nini, Andrea, Carlo Corradini, Diansheng Guo & Jack Grieve. 2017. The application of growth curve modeling for the analysis of diachronic corpora. *Language Dynamics and Change* 7(1). 102–125.
- Pew Research Center. 2019. *National Politics on Twitter: Small Share of U.S. Adults Produce Majority of Tweets*. <https://www.people-press.org/2019/10/23/national-politics-on-twitter-small-share-of-u-s-adults-produce-majority-of-tweets/>.
- Plag, Ingo. 2003. *Word-formation in English*. Cambridge: Cambridge University Press.
- Renouf, Antoinette, Andrew Kehoe & Jayeeta Banerjee. 2007. WebCorp: an integrated system for web text search. In Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (eds.), *Corpus Linguistics and the Web*, vol. 59 (Language and Computers), 47–67. Amsterdam, New York: Rodopi.
- Rogers, Everett M. 1962. *Diffusion of innovations*. New York: Free Press of Glencoe.
- Roig-Marín, Amanda. 2020. English-based coroneologisms: A short survey of our Covid-19-related vocabulary. *English Today*. 1–3.
- Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky & Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the fourteenth workshop on semantic evaluation*, 1–23. Barcelona, Spain: International Committee for Computational Linguistics. <https://aclanthology.org/2020.semeval-1.1>.
- Schlegel, Arnold. 2014. *Dokumentklassifizierung auf englischen Webseiten*. LMU Munich.
- Schmid, Hans-Jörg. 2008. New words in the mind: Concept-formation and entrenchment of neologisms. *Anglia – Zeitschrift für englische Philologie* 126. 1–36. <http://www.degruyter.com/view/j/angl.2008.126.issue-1/angl.2008.002/angl.2008.002.xml>.

7. Bibliography

- Schmid, Hans-Jörg. 2015. A blueprint of the entrenchment-and-conventionalization model. In *Yearbook of the German Cognitive Linguistics Association*, vol. 3 (Yearbook of the German Cognitive Linguistics Association), 1–27.
- Schmid, Hans-Jörg. 2016. *English morphology and word-formation - An introduction*. 2nd edn. Berlin: Erich Schmidt Verlag.
- Schmid, Hans-Jörg. 2020. *The dynamics of the linguistic system. - Usage, Conventionalization, and Entrenchment*. Oxford: Oxford University Press.
- Schmid, Hans-Jörg, Quirin Würschinger, Melanie Keller & Ursula Lenker. 2020. Battling for semantic territory across social networks. The case of Anglo-Saxon on Twitter. *Yearbook of the German Cognitive Linguistics Association* 8(1). 3–26. <https://doi.org/10.1515/gcla-2020-0002>.
- Scott, Ben. 2020. Know your covidiot from your cove-dwellers. *Bloomberg.com*. <https://www.bloomberg.com/opinion/articles/2020-04-03/coronavirus-know-your-covidiot-from-your-cove-dwellers> (22 August, 2021).
- Shafaei-Bajestan, Elnaz, Masoumeh Moradipour-Tari, Peter Uhrig & R. Harald Baayen. 2022. Semantic properties of English nominal pluralization: Insights from word embeddings. *ArXiv*.
- Shoemark, Philippa, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale & Barbara McGillivray. 2019. Room to glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 66–76.
- Spearman, Charles. 1961. *The proof and measurement of association between two things* (Studies in Individual Differences: The Search for Intelligence.). East Norwalk, CT, US: Appleton-Century-Crofts. 58. <https://doi.org/10.1037/11491-005>.
- Stefanowitsch, Anatol & Susanne Flach. 2017. The corpus-based perspective on entrenchment. In Hans-Jörg Schmid (ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge*, 101–128. Boston, USA: American Psychology Association and de Gruyter Mouton.
- Stevenson, Suzanne & Paola Merlo. 2022. Beyond the benchmarks: toward human-Like lexical representations. *Frontiers in Artificial Intelligence*. <https://doi.org/10.3389/frai.2022.796741>.
- Stewart, Ian & Jacob Eisenstein. 2018. Making “fetch” happen: The influence of social and linguistic context on nonstandard word growth and decline. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4360–4370. Brussels, Belgium: Association for Computational Linguistics. <https://doi.org/10.18653/v1/D18-1467>.
- Sunstein, Cass R. 2018. *#Republic: divided democracy in the age of social media*. Princeton & Oxford: Princeton University Press.
- Sunstein, Cass R. 2019. *Conformity: the power of social influences*. New York: NYU Press.
- Thorne, Tony. 2020. *#CORONASPEAK – the Language of Covid-19 Goes Viral – 2*. Language and innovation. <https://language-and-innovation.com/2020/04/>

- 15 / coronaspeak - part - 2 - the - language - of - covid - 19 - goes - viral / (12 February, 2021).
- Tournier, Jean. 1985. *Introduction descriptive à la lexicogénétique de l'anglais contemporain*. Paris: Champion-Slatkine.
- Tsakalidis, Adam, Marya Bazzi, Mihai Cucuringu, Pierpaolo Basile & Barbara McGillivray. 2019. Mining the uK web archive for semantic change detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 1212–1221. Varna, Bulgaria: INCOMA Ltd. https://doi.org/10.26615/978-954-452-056-4_139.
- Ullah Shaheen, Zafar, Ayyaz Qadeer & Fouzia Rehman Khan. 2021. Conspiracy theories (CT) vs truth based reporting: a corpus driven analysis of covid-19 online newspaper(s) discourse. *CORPORUM: Journal of Corpus Linguistics* 4(2). 112–135. <https://journals.au.edu.pk/ojsrcr/index.php/crc/article/view/169>.
- West, Robert, Hristo S. Paskov, Jure Leskovec & Christopher Potts. 2014. Exploiting social network structure for person-to-person sentiment analysis. *CoRR* abs/1409.2450. <http://arxiv.org/abs/1409.2450>.
- Wilton, David. 2020. What do we mean by Anglo-Saxon? Pre-Conquest to the present. *JEGP*.
- Würschinger, Quirin. 2021. Social networks of lexical innovation. Investigating the social dynamics of diffusion of neologisms on Twitter. *Frontiers in Artificial Intelligence* 4. 106. <https://doi.org/10.3389/frai.2021.648583>.
- Würschinger, Quirin, Mohammad Fazleh Elahi, Desislava Zhekova & Hans-Jörg Schmid. 2016. Using the web and social media as corpora for monitoring the spread of neologisms. The case of 'rapefugee', 'rapeugee', and 'rapugee'. In *Proceedings of the 10th web as corpus workshop*, 35–43. Berlin, Germany: Association for Computational Linguistics. <https://doi.org/10.18653/v1/W16-2605>.
- Würschinger, Quirin & Barbara McGillivray. N.d. Semantic change and socio-semantic variation. The case of Covid-related neologisms on Reddit. *Linguistics Vanguard* (accepted).
- Yardi, S. & D. Boyd. 2010. Dynamic debates: an analysis of group polarization over time on Twitter. <https://doi.org/10.1177/0270467610380011>.

Zusammenfassung

Lexikalische Innovation trägt dazu bei, das linguistische Repertoire von Sprachen an die kommunikativen Bedürfnisse ihrer SprecherInnen anzupassen. Dabei können lexikalische Innovationen eine Vielzahl unterschiedlicher Formen annehmen. In der vorliegenden Dissertation untersuche ich das Auftreten und die Diffusion von englischen Neologismen im Web und auf Social Media anhand eines vielfältigen Samples von 851 Neologismen, die mit Hilfe eines datenbasierten Ansatzes identifiziert wurden.

Neologismen werden in der vorliegenden Arbeit wie folgt definiert:

„[Neologisms are] lexical units, that have been manifested in use and thus are no longer nonce-formations, but have not yet occurred frequently and are not widespread enough in a given period to have become part and parcel of the lexicon of the speech community and the majority of its members.“ (Kerremans 2015: 31)

Ich schließe mich dieser Definition von Kerremans an, da sie zwei entscheidende Merkmale lexikalischer Innovation hervorhebt die für diese Dissertation von zentraler Bedeutung sind. Erstens betrachte ich Neologismen als innovative „lexikalische Einheiten“, was sowohl formale als auch semantische Neologismen einschließt. Zweitens nähere ich mich der Untersuchung lexikalischer Innovationen aus einer gebrauchsbasierten Perspektive. Gemäß der vorangegangenen Definition von Kerremans (2015: 31) betrachte ich Neologismen als lexikalische Einheiten, deren Gebrauch darauf hinweist, dass sie „im Gebrauch manifestiert wurden“, aber noch nicht vollständig verbreitet und zu konventionellen Bestandteilen „des Lexikons der Sprachgemeinschaft und der Mehrheit ihrer MitgliederInnen“ geworden sind. Aus gebrauchsbasierter Perspektive werden Neologismen weiterhin als neue lexikalische Einheiten angesehen, die über ihr erstes Auftreten hinaus verwendet wurden, aber (noch) nicht zu festen Bestandteilen des lexikalischen Inventars geworden sind. In dieser Dissertation untersuche ich die Verbreitung von Neologismen von ihrem erstmaligen Auftreten bis zu ihrer erfolgreichen Diffusion in der Sprachgemeinschaft.

Lexikalische Innovation tritt beispielsweise auf, wenn Sprecher neue Wörter einführen, um über neue Produkte (z. B. *blockchain*) oder neue gesellschaftliche Phänomene (z. B. *Covid*) zu sprechen, oder wenn etablierte Wörter aufgrund von kulturellen Veränderungen in der Gesellschaft neue Bedeutungen erhalten (z. B. (*social*) *distancing*). Neologismen haben per Definition die Zeit ihrer ersten Verwendung überlebt und haben es geschafft, sich zumindest bis zu einem gewissen Grad in der Sprachgemeinschaft zu verbreiten. Neologismen können daher als „transitional phenomena“ (Schmid 2008) auf dem Kontinuum zwischen Ad-hoc-Bildungen und etablierten Wörtern betrachtet

werden. Diffusion ist der Prozess, der ihre Verbreitung in neue Verwendungskontexte und neue Teile der Sprachgemeinschaft befördert (Schmid 2015, 2020).

In dieser Dissertation untersuche ich das Auftreten und die Diffusion von Neologismen im Internet und in sozialen Medien, mit einem besonderen Augenmerk auf den sozialen Dynamiken von Diffusion und auf sozialer Variation zwischen verschiedenen Gruppen von SprecherInnen. Zu diesem Zweck erhebe ich eine große Auswahl an formalen und semantischen Neologismen und untersuche ihre Verbreitung im Web, auf Twitter und auf Reddit, unter Verwendung von Methoden, die auf Gebrauchshäufigkeiten, Analysen von sozialen Netzwerken und *word embeddings* basieren.

Auftreten und Diffusion im Web – der NeoCrawler

Kapitel 2 untersucht die Entstehung und Diffusion von formalen Neologismen im Internet. Es präsentiert eine erweiterte Version des Tools NeoCrawler (Kerremans, Stegmayr & Schmid 2012), das dazu dient, Neologismen datenbasiert zu identifizieren und ein großes Monitor-Korpus zu erstellen, welches deren Diffusion beobachtet.

Ich verwende das Discoverer-Modul des NeoCrawlers, um ein breites Sample neuerer Neologismen zu identifizieren, das auch als Grundlage für die Untersuchungen in den folgenden Kapiteln dient. Die resultierende Neologismen-Datenbank enthält 958 Neologismen, was einen erheblichen Zuwachs im Vergleich zur vorangegangenen Studie von Kerremans (2015) darstellt, die auf 40 Neologismen begrenzt war. Dieses erweiterte Sample wurde nach Wortarten und Wortbildungsprozessen annotiert. Eine Auswertung des annotierten Samples ergibt, dass die Verteilungen für beide formalen Kategorien mit den Ergebnissen vorhergehender Untersuchungen und Daten aus dem Oxford English Dictionary übereinstimmen. Dies bestätigt, dass das Sample hinsichtlich Wortklassen und Wortbildungsprozessen ein breites Spektrum lexikalischer Innovation abdeckt.

Die Analyse der Diffusion dieser Neologismen über einen längeren Zeitraum mittels des Observer-Moduls des NeoCrawlers zeigt zudem, dass die untersuchten Neologismen auch ein breites Spektrum an Diffusionsgraden abdecken, was durch Auswertungen von kumulativen Gebrauchshäufigkeiten untersucht wird.

Diffusion im Web und in den sozialen Medien

Kapitel 3 untersucht die Diffusion von Neologismen im Web und auf der Social-Media-Plattform Twitter. Der Fokus liegt dabei auf drei ausgewählten Neologismen, um einen genaueren Blick auf ihre Verbreitung sowie auf die Faktoren, die ihre Diffusion beeinflussen, zu erlangen. Die Ergänzung der Webdaten durch zusätzliche Daten von Twitter dient dazu, die Reliabilität der Ergebnisse des NeoCrawlers zu evaluieren. Zudem ermöglicht dieser Ansatz, die Verwendung von Neologismen in zwei Gebrauchskontexten zu untersuchen, um festzustellen, ob deren Verwendung auf Gebrauchskontexte im Web oder in sozialen Medien beschränkt bleibt.

Das Kapitel untersucht den Gebrauch folgender drei Neologismen: *rapefugee*, *rapeugee*, und *rapugee*. Alle drei Begriffe sind Kontaminationsbildungen von *rape* und

refugee und traten im Jahr 2015 erstmalig auf. Es handelt sich bei diesen Begriffen um abwertende Propaganda-Begriffe von rechtsextremen Gegnern von Maßnahmen, die Asylsuchende willkommen heißen. Bei den drei Neologismen handelt es sich um formale Varianten, die in einem onomasiologischen Wettbewerb stehen, um dieselbe Bedeutung zu kodieren: „A refugee who rapes people. Usually referred to the Muslim refugees pouring into Europe.“ (Urban Dictionary¹)

Die Ergebnisse dieser Studie zeigen eine breite Übereinstimmung im Gebrauch der untersuchten Neologismen zwischen den mit dem NeoCrawler gewonnenen Webdaten und den Twitter-Daten. Zudem korreliert der Gebrauch der Begriffe in beiden Datensätzen mit außersprachlichen Ereignissen, die im Zusammenhang mit den untersuchten Neologismen stehen. Dies kann als Kreuzvalidierung der Web- und Twitter-Ansätze angesehen werden und legt nahe, dass beide Datenquellen dazu dienen können, die Diffusion von Neologismen mit angemessener Zuverlässigkeit zu analysieren. Das Kapitel hebt zudem die Bedeutung von sozialen Medien als treibender Faktor bei der Entstehung und Diffusion von Neologismen hervor. Die ausgewählten Neologismen wurden in den frühen Phasen ihrer Diffusion aggressiv auf Twitter verbreitet und ihre steigende Popularität in den sozialen Medien förderte ihren Gebrauch im Internet. Die Twitter-Daten zeigen darüber hinaus deutlich den Einfluss bestimmter Communities bei der Förderung einer höheren *usage intensity* (Stefanowitsch & Flach 2017) der ausgewählten Neologismen. Die qualitative Analyse zeigt jedoch, dass der Gebrauch dieser Lexeme trotz ihrer erhöhten Gebrauchintensität weiterhin auf eine vergleichsweise geringe Anzahl von gleichgesinnten Personen und Communities am politisch rechten Rand der Gesellschaft beschränkt bleibt.

Diffusion in sozialen Netzwerken

Nach einer Einführung in den Forschungskontext und den theoretischen Hintergrund der vorliegenden Arbeit (Kapitel 1) zielt Kapitel 4 darauf ab, ein detaillierteres Bild von den sozialen Dynamiken der Diffusion auf Twitter zu gewinnen. Um eine bessere Generalisierbarkeit zu erreichen, basiert diese Studie auf einer wesentlich größeren Stichprobe von Neologismen als die vorangegangene Studie. Sie untersucht das Aufkommen und die Diffusion von 99 Neologismen, von denen die meisten mittels des Discoverer-Moduls des NeoCrawlers identifiziert wurden sowie zusätzlich ausgewählte Neologismen aus einer früheren Studie zur Diffusion von Neologismen auf Twitter von Grieve, Nini & Guo (2016). Die Studie basiert auf einem großen Twitter-Datensatz von ca. 30 Millionen Tweets, der den Zeitraum von der Gründung Twitters im Jahr 2006 bis zum Ende des Jahres 2018 abdeckt. Die Neologismen in diesem Sample wurden so ausgewählt, dass sie innerhalb dieses Zeitrahmens erstmalig auftraten, wodurch die frühen Stadien ihrer Diffusion erfasst und ihre Verbreitung über einen wesentlich längeren Zeitraum verfolgt werden kann als dies mit dem zuvor verwendeten NeoCrawler-Ansatz möglich war.

¹„rapefugee“, <https://www.urbandictionary.com/define.php?term=rapefugee>, Zugriff am 23. Mai 2022.

Um die Reliabilität von Gebrauchshäufigkeit als Indikator für Diffusion zu untersuchen, wie in den Kapiteln 2 und 3 verwendet, wendet dieses Kapitel zusätzliche Maße an, um die zeitliche Dynamik von Diffusion genauer zu untersuchen. Dafür untersucht die Studie Trends in der Nutzungsintensität von Neologismen, die durch kumulative und absolute Häufigkeitszahlen erfasst werden, ihre Volatilität, die durch den Variationskoeffizienten gemessen wird, und ihr Alter, das nach der beobachteten Lebensdauer der Neologismen seit ihrem Auftauchen im verwendeten Korpus bemessen wird. Die Ergebnisse zeigen, dass die Berücksichtigung der zeitlichen Dynamik des Gebrauchs durch die vorgeschlagenen zusätzlichen Maße zu einem detaillierteren Bild von Diffusion beitragen kann. Zudem wird argumentiert, dass die zeitliche Dynamik von besonderer Bedeutung für die Untersuchung von Neologismen ist, da diese stark von Veränderungen in ihrer *semantic carrying capacity* (Nini u. a. 2017) abhängig sind. Die vorgeschlagenen Methoden gewährleisten eine genauere Betrachtung von Fällen von „Topikalität“ (Fischer 1998: 16) und „re-current semi-conventionalization“ (Kerremans 2015: 129–136).

Darüber hinaus ergänzt das Kapitel frequenzbasierte Maße von Diffusion durch die Anwendung sozialer Netzwerkanalysen auf Basis des Twitter-Datensatzes, um direkte Einblicke in die soziale Diffusion der untersuchten Neologismen zu erhalten. Mehrere netzwerkbasierende Maße (z. B. *centralization*, *in-degree*) und Visualisierungen von Netzwerkgraphen werden verwendet, um festzustellen, ob Neologismen erfolgreich diffundieren und sich über eine größere Anzahl von SprecherInnen und Communities verbreiten.

Ein Vergleich des frequenzbasierten und netzwerkbasierten Ansatzes ergibt ein hohes Maß an Übereinstimmung in der Gesamtbewertung des Grades der Diffusion der Neologismen in der Stichprobe. Es zeigen sich jedoch auch einige Fälle, in denen die Häufigkeit der Verwendung den Grad an sozialer Diffusion überbewerten zu scheint. Die Analyse zeigt ein Muster von Fällen die dem in Kapitel 3 behandelten Neologismus *rapefugee* ähneln. Die Daten deuten darauf hin, dass *rapefugee* in zweifelhafter Gesellschaft mit einer Reihe von anderen politisch aufgeladenen Neologismen wie *alt-left* oder *birther* ist, die ebenfalls eine hohe Nutzungsintensität und Zentralisierung aufweisen. Diese Begriffe zeigen ein hohes Maß an sozialer Variation und werden vorwiegend in kleinen, politisch polarisierten Bereichen des sozialen Netzwerks verwendet, die ideologischen Echokammern entsprechen.

Semantische Innovation eines alten Wortes – der Fall *Anglo-Saxon*

Nachdem in den vorangegangenen drei Kapiteln die Entstehung und Diffusion von formalen Neologismen untersucht wurde, befasst sich dieses Kapitel mit semantischen Neologismen, die in den meisten früheren Arbeiten zu lexikalischer Innovation nicht im Mittelpunkt standen. Das Kapitel untersucht die jüngsten Veränderungen im Gebrauch und in der Bedeutung des Begriffs *Anglo-Saxon* auf Twitter.

Der Begriff *Anglo-Saxon* wirkt zunächst nicht wie ein typischer Neologismus. Das Wort ist sehr alt, da es auf die Ursprünge der englischen Sprache zurückgeht. Dennoch war die Bedeutung des Begriffs in letzter Zeit Gegenstand semantischer Innovation. Der

Begriff wird im Allgemeinen in drei Bedeutungen verwendet, die in früheren Arbeiten wie folgt kategorisiert wurden: „historical/pre-Conquest“, „ethno-racial“, und „politico-cultural“ (Wilton 2020). In jüngster Zeit wurde der Begriff aufgrund seiner ethnisch-rassistischen Bedeutung kontrovers diskutiert, da diese stark mit dem Konzept der *white supremacy* verbunden ist. Diese kontroverse Bedeutung hat sich durchgesetzt, und ihre problematischen Assoziationen auf den Begriff selbst übertragen und lässt somit wenig Raum für seinen Gebrauch in den beiden anderen, weniger umstrittenen Bedeutungen. Dieser Konflikt zwischen konkurrierenden Bedeutungen und Assoziationen von *Anglo-Saxon* hat zu Veränderungen im Gebrauch und der Gesamtbedeutung des Begriffs geführt.

Das Kapitel zeigt erhebliche sozio-semantische Variation im Gebrauch des Begriffs *Anglo-Saxon*, der Gegenstand intensiven Wettbewerbs um seine Bedeutung ist. Die Analyse des sozialen Netzwerks in diesem Kapitel legt nahe, dass der öffentliche Diskurs auf Twitter stark polarisiert ist. Bestimmte Communities (z. B. HistorikerInnen oder rechtsextreme AktivistInnen in den USA) scheinen Echokammern zu bilden, in denen eine starke Übereinstimmung in gesellschaftspolitischen Ansichten herrscht. Innerhalb dieser Communities bekräftigen sich die SprecherInnen gegenseitig, was zu einer zunehmenden Usualisierung von Gruppenkonventionen darüber führt, ob und wie der Begriff *Anglo-Saxon* zu verwenden ist.

Die Netzwerkanalyse zeigt die zunehmende Zentralisierung des Gebrauchs von *Anglo-Saxon* im Laufe der Zeit, was durch manuelle Netzwerkanalysen untermauert wird. In einigen Communities wird der Begriff weiterhin in seiner früheren Bedeutung verwendet, während andere aufgehört haben, diesen Begriff zu verwenden, da sie ihn mit Rechtsextremismus assoziieren. Während die Zukunft dieses Begriffs und seiner Bedeutung derzeit ungewiss ist, deuten die Ergebnisse darauf hin, dass die derzeitige gesellschaftliche Dynamik zu einer nachhaltigen Veränderung seiner Bedeutung führen wird. Der Begriff *Anglo-Saxon* ist daher ein interessanter Fall von semantischem Wandel und sozio-semantischer Variation: Er zeigt, wie ein etabliertes Wort erhebliche semantische Variationen zwischen verschiedenen Communities aufweisen kann und wie es dadurch zu Bedeutungswandel innerhalb einer kurzen Zeitspanne kommen kann.

Semantische Innovation und soziale Variation

Kapitel 6 untersucht semantische Innovation und sozio-semantische Variation. Es geht über die Studie im vorhergehenden Kapitel hinaus, indem es eine umfassendere und detailliertere Darstellung der semantischen Variation zwischen Communities vorlegt. Dieses Kapitel untersucht semantische Innovation im Zusammenhang mit der aktuellen Covid-Pandemie, da die großen gesellschaftlichen Auswirkungen von Covid in den vergangenen zwei Jahren ein beträchtliches Ausmaß an sprachlicher Innovation hervorgebracht haben.

Zunächst unternehme ich eine datenbasierte Identifikation semantischer Neologismen im Gegensatz zu der Fallstudie von *Anglo-Saxon* im vorherigen Kapitel. Zu diesem Zweck bestimme ich semantische Neologismen auf datengesteuerte Weise durch den

Einsatz von *word embeddings*, die es ermöglichen, für alle Wörter im Korpus semantische Repräsentationen für die Jahre 2019 und 2020 zu erzeugen und zu bestimmen, welche dieser Wörter in diesem Zeitraum den höchsten Grad an semantischem Wandel aufweisen. Dieser Ansatz liefert eine große Menge an semantischen Neologismen, aus der ich die 20 Wörter mit dem höchsten Grad an semantischem Wandel auswähle, um zu analysieren, ob diese Wörter semantische Variation zwischen verschiedenen Communities zeigen.

Zweitens unternimmt dieses Kapitel eine groß angelegte quantitative Analyse von Unterschieden in der Bedeutung zwischen Communities. Die Ergebnisse zeigen erhebliche sozio-semantische Unterschiede zwischen verschiedenen Communities. Das Subreddit *r/conspiracy*, das der gesellschaftspolitischen Reaktion in Bezug auf die Pandemie kritisch gegenübersteht, zeigt mehr negative und subjektive Assoziationen mit Covid-bezogenen Neologismen als das neutrale Subreddit *r/Coronavirus*. Der in diesem Kapitel verwendete Ansatz erweitert die vorhergehende Studie durch den Einsatz von Reddit-Daten, was eine robustere Interpretation von Gruppenunterschieden ermöglicht, da Reddit, im Gegensatz zu Twitter, in Communities organisiert ist, die auf expliziten gemeinsamen Interessen ihrer Mitglieder basieren. Darüber hinaus bietet die vorliegende Studie eine umfassendere quantitative Untersuchung indem sie die sozio-semantische Variation auf der Grundlage von Community-spezifischen *word embeddings*-Modellen misst, im Gegensatz zu der manuellen, qualitativen Fallstudie, die für *Anglo-Saxon* durchgeführt wurde.

Das Kapitel versucht darüber hinaus ein detaillierteres Bild von der Art der semantischen Unterschiede zwischen Communities zu zeichnen. Die Ergebnisse zeigen konsistente Muster sozio-semantischer Variation entlang einer evaluativen und einer subjektiven Dimension: Sowohl Covid-bezogene Begriffe wie *vaccines* als auch allgemeinere Begriffe wie *research* weisen in der Corona-skeptischen Community *r/conspiracy* stärkere Assoziationen mit Negativität und Subjektivität als in der neutralen Community *r/Coronavirus* auf.

Kapitel 7 schließt diese Dissertation ab, indem es die Ergebnisse zusammenfasst und diskutiert. Das Kapitel ordnet die gewonnenen Erkenntnisse in den Forschungskontext ein und zeigt Implikationen und Desiderate für weitere Forschung auf.

Appendix: List of neologisms

accordionize	assault weapon	bird's nest parenting
acedia	ate-up	birther
adblocking	attitone	bitchcraft
administrativia	autocowrong	bitshaming
adulging	avatard	Bixby
AFOL	awesomesauce	blackgrass
AfPak	azurophil	bleisure
Afrofuturism	bae	blockchain
agism	baecation	bloggergate
al-amira	baeless	bloglet
alabastard	balayage	blue belt
alphology	balsy	bobu
alt-left	bandity	body shaming
alt-right	bankster	body-hating
alternative fact	barkini	bonespiration
alternative facts	barkitive	Boobgate
ambient snacking	beardruff	BookTuber
Anglo-Saxon	bed-blocking	bool out
animoji	bediquette	bootiful
annoyitate	begpacker	born day
anti-blackness	begpacking	boy mode
anti-diversity	bejumbled	boysplain
anti-Facebook	belfie	Brangelina
antifa	bennifer	breadatarian
antisappointment	Berkeley goggles	breadcrumb
antistalking	Berniesplain	breadcrumbr
apitourism	betrump	Brexit
aquacrunk	bezel	Brexiteer
aquafaba	bezel-less	brexiteer
aqualuminescence	BF	Brexiteers
architectophile	big dick energy	Brexiter
argumentarian	bing	brexiter
arthropodology	bingeable	bring-your-own-booze
askhole	biobag	broette
ass-to-heels	biometric border	broflake

7. Bibliography

Broga	chronocide	crizy
broga	ciggie	cronught
brogrammer	ciggy	cronut
brojob	circular economy	crouffin
bromance	clashion	cruffin
bromosexual	clashionista	crypto-mad
bronde	Clemsoning	crypto-mining
brongerie	climagate	cryptojacking
buildering	climate denial	cuck
bulletize	climate emergency	cuckold
bum-flashing	co-work	cultural Catholics
bumspiration	cobra effect	cunch
burgergate	coddiwomple	cupcakery
burkini	cold peace	cyber bullying
burqa	collaborative economy	cyber spying
burquini	Colognoisseur	cyberchondria
butt-dial	comfortability	cyberchondriac
buzz marketing	commjacking	cyberdisinhibition
cakeism	condoburbia	cybergang
Cancerversary	condominiumization	cyberloafing
cankles	conflictious	cyberoffensive
car keying	conspiritual	Cyberrhea
cashierless	constitutive resistance	cybersoldier
cat café	constracted	DACA
cat call	conurbation	dadchelor party
Catalexit	cook processor	daddymoon
catcaller	copist	dark kitchen
catfish	copy-Kate	dashcam
catfishing	Corbynista	datakinesis
catio	corona	dead tree book
cerb	corporatization	deathist
chador	Cortana	deathiversary
chankles	counter-radicalism	decepticon
chatbot	covfefe	dechristmas
chaz	Covfefegate	decrementally
checkout-free	cowork	decycling
cheffy	coworking	deep learning
chemicalism	crackberry	deepfake
cherpuple	craftivism	deja poo
chestical	creepy clown	democide
chiblings	cricketing	deomorphism
childe	crimmigration	deprofitizing
chronocidal	cringy	desk-share

desk-sharing	empty forest syndrome	flexting
destigmatise	encore career	flightseeing
detweet	Engelina	flipster
device mesh	epicaricacy	floordrobe
diabesity	equallyoked	followership
digerati	equivalate	FOLO
dimpsy	escape room	fomo
distancing	ethnomics	food digger
dockless	euneirophrenia	food swamp
dog manor	Eurofascists	foodventure footcial
Doga	evacation	forensicate
dogfishing	ex-Trump	Frankenmissile
dogtor	exabyte	freakshake
dotard	exahertz	friendsourcing
double double	exfixiation	froghurt
down-voting	exhaustipated	frosé
Drag-erwocky	exponentialize	frugalista
drink-fly	extreme phone pinching	frunk
droneboarding	eyebleach	fuel poverty
dronfie	fab lab	funformative
dronie	facedesk	gaslighting
dronograph	fake news	gegenpressing
dronographer	fake nudes	Gen Z
dronography	fake-apologise	gendercide
druggle	fake-apologize	generacast
drunkle	famfie	generation mute
dualie	fan-girled	generationist
dumpster fire	fangirling	genosucide
e-beg	farmster	germaphobe
e-ffair	fat shaming	gerontification
e-tailer	fauxmance	ghost driver
e-tivity	feminocracy	ghost species
e-waste	femivore	ghost surgery
earworm	fightmare	ghosting
eco-anxiety	filtrate	girther
eco-sexual	finishability	glamping
ecocide	firecrotch	glampsite
economicky	fitspiration	glanceable
egg coffee	flash-spread	globalists
elationship	flashion	globesity
election-hacking	flashpacking	globster
emojictionary	fleek	gloving
emojinal	flexi schooling	gorpcore

7. Bibliography

Gorpcore	inconvenience fee	lol
goth latte	infobesity	lolification
grawlix	informations	lolified
greenwashing	infotainment	luner
Grexit	ingenuine	lupper
grip-lit	inking	lyricality
gym-spiration	insta-mum	m-commerce
gymspiration	instamatically	malazy
half bricked	instasham	malhearted
half-false	Instastory	malicious insider
halfalogue	internet of me	malignant neglect
hamdog	Internet of things	man bun
hand-coder	Internet worms	manbabies
hangry	intexticated	manfant
hashtagger	iPadable	mankles
hashtivism	Islamophobia	mansplain
hatriot	Juggalo	manspreading
hawkish	kindergarchy	mantrum
hegan	kittenfishing	maranoia
helicopter parenting	klee	maskless
hepeating	kleptopredation	masks
hermiting	kompromat	matcha
hermitous	kosmemophobia	Maychine
hijab	Kushnergate	Meatmare
Honeycrisp	labradorable	mega-transfer
honeyteer	late-great	megadonor
hoptimist	latte levy	megalodon
horosceptic	laurel	melon-aid
horrideous	lecturous	melon-choly
hotumn	levidrome	memable
hunkvertising	lgbtp	memenials
hyperandrogenism	libtard	menu hacking
hyperlocal	libtards	mesofact
hyperthymesia	lifehack	metamour
hyphy	line-free	metarchon
ideation	linguaphile	micro-adventure
ideocentrism	linkulitus	micro-cheating
idiodysey	linner	microapartment
in-sprog-nito	litigarchy	microblading
incarceratory	lituation	microflat
incel	liveblog	midult
incentivation	loadly	milkshake duck
incestrial	lockdowns	minutarily

mirf	noodie	Pixelbook
misgender	nose-deaf	Pizzagate
misophonia	noseworm	plebgate
MLG	nugzilla	pocket-dial
moderna	nutricosmetics	Pokémoning
mom-of-three	obli-cation	policism
momcologist	oblication	policist
momentous	observationalist	polyhierarchy
mommymoon	old girls' network	polyreligious
momster	ominent	pooparonus
mondaze	oncofertility	poppygate
monkey dumpling	oniochhalasia	pornsexual
monogestural	open education	post-Brexit
monthiversary	opposite	post-eclipse
monthversary	ortho-bionomy	post-factual
moobs	orthosomniac	post-Trump
multiational	osmotic	post-truth
mum-of-one	otherize	powerdressing
mum-of-three	Oumuamua	pre-Brexit
mum-of-two	ousside	pre-juicing
mummymoon	out-Trumping	pre-reply
mumspiration	ovary-acting	pre-Trump
must-wear	overparenting	prebuckle
Mx	overshare	predatory lending
mysticious	oversharing	prequaintance
n00b	overtourism	presstitute
nanotecture	pabebe	previvor
narratitis	paggro	previvors
NBD	paleosphere	prisonic
Neckbuds	pandemic	pro-whiteness
neckbuds	papyrophile	procaffeinating
negromancy	paracosm	profit recession
neomascularity	parennial	promissory notes
neophile	peak stuff	pronunciate
newb	pedamantra	pugly
newf	pedosexual	pupper
newsjacking	peezing	quadraxexual
night czar	peoplekind	quarantines
niqab	perceptoid	radiculous
no-deal	peshmerga	rage quit
noice	peticure	rage-quit
non-rebellion	phonely	ragequit
noob	Phubbing	rap-head

7. Bibliography

rapefugee	sexposed	soliphilia
rapeugee	sexsomnia	sologamist
rapugee	shadow flipping	sologamy
Rapugee	shareable	solopreneur
reborn doll	sharent	sovereign debt
recency illusion	sharenting	spiralizer
recombobulation area	shideous	sploshing
reddiquette	shitcoin	snoop
Reddit	shitholegate	spreader
refollow	shopaholic	starballing
reimpeach	shopaholism	stealth health
relationsheep	short-termism	STEAM
remoaners	sibkid	strawberry-gate
research	siblet	subling
retrocise	sillious	subtweet
rewilding	singlism	sunsetting
ringxiety	sip slip	super-crops
rittenhouse	Siri	superager
robo-sign	situationship	superbness
robo-signer	skin-credible	superphone
robo-signing	skinny-shaming	survey knowledge
rona	slacktivism	suya
roofing	sleep divorce	tablet-phone
rooftopper	sleepcation	tabnabbing
rooftopping	sliver building	tarpology
root-to-stem	slut shaming	tea-scription
round pound	slutshaming	teabonics
royalmania	smartwatch	tech bro
runch	smirting	tech-giants
runchies	smize	techlash
runger	smober	teenmentia
sad rap	Smonday	telepresence
sanger	snaccident	tenebrous
sapiosexual	snackable	teraproject
sargasm	snaughling	testosteroom
sars	sneakerhead	text-walk
scale-up	Snowbergines	Textalyzer
science	snowicane	textament
scientry	snownado	texto
seenager	social eating	textpression
self-kind	social justice warrior	thinspiration
selfiegenic	sodcasting	third-wave coffee
set-jetting	solastalgia	thoughtscape

threequel	twittersphere	vote-shaming
thruple	twitterverse	wackaging
till-free	twittosphere	wackazoid
Tinder	typeractivity	wankle
tinder swipe	Ulstermatum	warrantless
tinderella	ultimullet	weaselflood
tip creep	ultra-processed	wefie
TOFI	un-Facebook	wellderly
tonow	un-reality	wext
tookah	unbe-leaf-able	white fragility
toxinologist	uncharacteristical	white walling
tragesty	uncomplicate	widhe
transload	unconvictable	winterscape
trippy	unfeesible	wisdomous
triquel	unfollow	witricity
trumpanzee	uninstructable	wob
Trumpian	unschooling	wokeness
trumpidation	unsend	womankini
Trumpish	unshare	womanspreading
trumpism	Unshare	work stoppage
trumpkin	upcycling	wrist-top
Trumpology	upskill	yanny
Trumpty Dumpty	upskirting	yardist
truth decay	urbexing	YIMBY
twatter	vaccines	youthquake
tweep	vacuo	Zexit
tweetathon	veganuary	zika virus
tweeter	vegducken	zoledronic acid
tweetoric	veggan	zoopharmacognosy
tweetstorm	vermicompost	ZumaExit
tweetup	vidcon	Zumexit
twerp	viral marketing	Zumxit
twitterer	virginality	
twitterrhea	virtue signalling	