Social networks of lexical innovation

Investigating the diffusion of neologisms on Twitter

Quirin Würschinger
LMU Munich

13th December 2020

Abstract

Societies continually evolve and speakers coin and use new words to talk about innovative products and practices. While most lexical innovations fail to catch on, others spread successfully and become part of the lexicon. This paper investigates the diffusion of English neologisms on Twitter. Previous work on lexical innovation has almost exclusively relied on usage frequency counts for measuring diffusion. Taking frequency as a baseline, I use social network analysis to zoom in on the sociolinguistic dynamics of diffusion.

My results show that frequency counts lend themselves to approximate overall degrees of diffusion with varying success. While absolute counts can be misleading, incorporating temporal dynamics of use provides a better picture of diffusion. However, frequency-based information alone fail to capture important sociolinguistic characteristics. Social network information are shown to add valuable information about whether new words are known and used by an increasing number of individuals and communities of speakers. Firstly, I distinguish different pathways of diffusion depending on whether and to which degree new words show increasing vs. decreasing centralized use over time. Secondly, I show that social network information allow for a more fine-grained assessment of degrees of diffusion, for example when new words are used with increasing frequency when use their remains limited to certain parts of the speech community. Lastly, I compare the results based on usage frequency and on social network analysis. Besides notable discrepancies, we find a significant correlation between both types of information, which serves to cross-validate both approaches.

My results suggest that social network information can complement frequency counts and that using information from both sources provides a more reliable and differentiated view of the sociolinguistic dynamics of diffusion. I argue that this is particularly important for investigating the diffusion of lexical innovations, as new words are often marked by high social indexicality and show substantial differences in use between communities of speakers. More generally, however, social network analysis shows great potential to study sociolinguistic dynamics of language variation and change beyong the domain of lexical innovation.

Keywords: lexicology, lexical innovation, sociolinguistics, diffusion, social media, Twitter, big data, 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?

Covid-19 has recently spread through social contagion with shocking speed and has tragically affected the lives of people around the world. Its fatal consequences have demonstrated the devastating power of exponential diffusion in social networks. In a recent paper analysing contagion patterns of diseases in *Nature Physics*, Hébert-Dufresne, Scarpino & Young (2020) suggested that the spread of viruses 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), and such models show commonalities between the spread of cultural and linguistic innovations. 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 1977). 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 using social network analysis to study these phenomena empirically.

Social media platforms like Twitter have changed the way we communicate and how information spreads, and they offer large amounts of data for empirical research. Sociological research has been concerned with pressing issues regarding the impact of online social networks for the spread of hate speech, fake news and the power of 'influencers', bots and institutions on public opinions and elections, which increasingly strain the social fabric. For (socio-)linguists, social media provide large amounts of data of authentic language use which opens up new possibilities 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 (Nini et al. 2017) or about factors that influence whether new words spread successfully (Grieve 2018). Moreover, metadata about speakers allows studying aspects of diffusion that go beyond what can be captured by usage frequency alone. Recent work, for example, has used Twitter data to investigate the geographical spread of lexical innovations. (Eisenstein et al. 2014, Grieve 2017, Grieve, Nini & Guo 2018)

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 et al. 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 & 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, earlier approaches have used computational modelling to test the validity of the S-curve model (Blythe & 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 diffusion on social media datasets promises to shed light on whether the adoption

references

of new words remains limited to closely connected sub-communities or whether they reach larger parts of the speech community and whether individuals or groups drive this process.

This paper makes use of social media data and social network analysis to study the diffusion of lexical innovations on Twitter. Taking usage frequency as a baseline, we conduct a longitudinal study monitoring the use of a broad sample of neologisms to analyse their cumulated usage frequency as well as the temporal dynamics underlying their spread. We additionally use social network analyses of our neologism sample to get a better picture of the sociolinguistic dynamics at play, in order to assess different pathways and overall degrees of diffusion. Lastly, we compare both approaches to assess their validity, and we combine information from both sources to draw a more differentiated picture of diffusion.

The paper is structured as follows. Section 2 presents an overview of previous attempts to modelling and measuring the diffusion of lexical innovations in order to contextualise and define the present theoretical framework and its operationalisation for the empirical study. Section 3 provides information regarding the present sample of neologisms and the collection and composition of the Twitter dataset this study is based on. Section 4 describes the methodological procedure for analysing diffusion in this dataset, focusing on the construction and analysis of social networks. Section 5 presents the empirical results based on usage frequency and social network analysis, and from comparing both approaches to studying diffusion. Section 6 summarises and discusses these results and suggests theoretical implications and directions for future work.

2 Modeling the diffusion of lexical innovations

Speakers continually coin new words, yet most fail to spread successfully and fall into oblivion. How do new words diffuse to be known and used by more and more speakers and to become conventional lexemes in a language system? And how can diffusion be modelled theoretically and measured empirically?

Neologisms are on a continuum from entirely novel word-formations to established lexemes that 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. On one end of the continuum, 'nonce-formations' are new words that have been coined in a concrete communicative situation, but are not adopted by interlocutors to be used in future usage contexts and do not enter a process of continuous diffusion. (Hohenhaus 1996)

Fully established words form the other end of the continuum. They are known and used by the majority of the speech community, and they are codified in dictionaries. This latter, lexicographic feature reflects speakers' agreement on how these words are to be used and marks their status as conventional lexical units in the language system. 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 nonce-formations, but have not yet occurred frequently and are not wide-

> spread 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 thus represents the process that transports successful neologisms along this continuum, becoming increasingly conventional in the speech community.

A more precise definition is provided by Schmid: 'I define diffusion as a process that brings about a change in the number of speakers and communities who conform to a regularity of co-semiotic behaviour and a change in the types of cotexts and contexts in which they conform to it.' (Schmid 2020)

page?

this definition includes three dimensions of diffusion

cotexts contexts speakers: number and groups

i will focus on sociolinguistic dimension in this paper spread across speakers communities

context: largely fixed as it is limited to use on Twitter comparison of contexts outside the scope of this paper; initial work: (Würschinger et al. 2016)

previous work has taken at structural and cognitive perspectives

i will focus on the sociolinguistic perspective

most relevant and important model: S-curve model

2.1 Research perspectives

A substantial body of linguistic research has tackled this question from different **perspectives**. (Schmid 2016: 16)

From a **structural** perspective, main areas of interest include which word-formation processes are involved in forming new words, whether they are formally and semantically transparent, whether they change in the process of lexicalization and which status the resulting neologisms have in the language system (institutionalization). (e.g. Bauer 1983, Lipka 2005)

Cognitive perspectives focus on how individuals process and store lexical innovations. Speakers generally use new words when they experience a communicative need to talk about entities or practices that cannot be readily expressed by their language's inventory of conventional words yet. In order for neologisms to successfully diffuse, speakers need to successfully negotiate their meaning (co-semiosis) in discourse, others need to adopt the behaviour of using these words (co-adaption). Continued exposure and use of new words can then lead to the entrenchment of new words in the mental lexicon of speakers. (Schmid 2008)

Sociolinguistic perspectives transcend the level of the individual to study the diffusion of new words across speakers. The diffusion of lexical innovations is commonly seen as successful when the majority of the speech community has accepted a new word as a conventional lexical unit which is commonly used in communicative practice.

2.2 The S-curve model

S-curve models of linguistic change (Milroy 1992, Nevalainen 2015, Labov 2007) assume universal sociolinguistic dynamics for the diffusion of linguistic innovations.

- The **trajectory** of spread is expected to follow an S-curve shape, with low rates of diffusion in early stages, followed by a period of accelerating spread with a tipping point at the mid point in the diffusion curve after which diffusion slows down and the curve flattens towards the end of the diffusion process.
- These temporal trajectories are assumed to correspond to the **sociolinguistic dynamics** of which individuals and groups interact with each other and adopt the target innovation.
  - In the **first stage** of slow diffusion only few early adopters take up the innovative words. Individuals who use the new word in this stage typically form dense networks which are connected by strong ties. The structure of tight-knit communities of closely-associated individuals facilitates the successful negotiation of meaning (co-semiosis) of new words. High rates of interactions in these communities lead to high rates of exposure for individuals, which fosters co-adaption, entrenchment and the increasing usualization of new words in these communities.
  - In cases of successful diffusion the initial stages are followed by an **acceleration in spread** when new words increasingly reach speakers outside these tight-knit communities via weak ties (Granovetter 1977). Rates of diffusion increase substantially when speakers who are not part of the initial group of early adopters start to accommodate the new words, allowing the innovations to reach a broader spectrum of the speech community.
  - In **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 reluctant to take up the new words.
- S-curve models have mainly been applied to the **linguistic domains** of phonology and syntax. Fundamental differences between lexemes and linguistic items on other levels such as phonemes and grammatical constructions might affect the validity and reliability of such models for *lexical* innovation.
  - For example, grammatical constructions such as the *going to* future used to express a speaker's future intention serve to fulfil relatively abstract communicative needs that remain stable over time. By contrast, on the lexical level, linguistic innovations are typically tied to concrete cultural referents such as products and practices whose conceptual relevance is much more volatile over time. For example, many lexical innovations such as *millennium bug*, which denotes the fear of a computer crash at the beginning of the new millennium, might show high rates of diffusion, and become entrenched and conventional among the majority of the speech community. Without continual conceptual relevance in public discourse, however, these words fail pass on to the next

generation of speakers. S-curves are commonly expected when linguistic innovations compete for **'semantic carrying capacity'** (Nini et al. 2017). In most cases of *lexical* innovation, however, the conceptual carrying capacity is not stable over time, but shows high degrees of variation as the use of new words is dependent on the salience of novel concepts in public discourse. This represents a critical deviation from the traditional assumptions behind S-curves in language change.

– Nevertheless, the strong theoretical and empirical basis of the S-curve model for the diffusion of linguistic and cultural innovations, and its explicit assumptions about the sociolinguistic dynamics underlying the process of diffusion make it an attractive **blueprint** for the empirical study of the sociolinguistic diffusion of lexical innovations.

2.3 Current framework: the EC-Model (Schmid 2020)

I use the Entrenchment-and-Conventionalization-Model (Schmid 2020) as a framework for modelling the diffusion of lexical innovations.

The EC-Model provides an approach integrating both structural, cognitive and sociolinguistic perspectives on the diffusion of lexical innovations.

The model also differentiates between the level of the individual ('entrenchment') and the community ('conventionalization').

Here I will only briefly outline the most important concept relevant for studying the sociolinguistic aspects of diffusion here.

Conventionalization:

definition: 'Conventionalization is the continual process of establishing and re-adapting regularities of communicative behaviour among the members of a speech community, which is achieved by repeated usage activities in usage events and subject to the exigencies of the entrenchment processes taking place in the minds of speakers.' (Schmid 2020)

Usualization 'Usualization can therefore be defined as a process that establishes, sustains, and changes regularities of behaviour with regard to co-semiotic mappings between forms and meanings or functions and communicative goals and linguistic forms. It affects the semasiological, onomasiological, syntagmatic, cotextual, and contextual dimensions of conformity behind conventionality and is relative to communities.' (Schmid 2020)

Diffusion 'Linking the three aspects of speakers, cotexts, and contexts, I define diffusion as a process that brings about a change in the number of speakers and communities who conform to a regularity of co-semiotic behaviour and a change in the types of cotexts and contexts in which they conform to it.' (Schmid 2020)

less relevant for sociolinguistic aspects and for this study cotexts contexts

According to the EC-Model, for studying the sociolinguistic aspects of diffusion, investigating 'changes in the number of speakers and communities' is thus essential.

3 Measuring the diffusion of lexical innovations

3.1 Previous approaches

- **before frequency** Empirical approaches studying the diffusion of lexical innovations have only recently become feasible with the advent of new data sources and computational methods.
- Earlier work had to rely on **traditional linguistic corpora**. Due to the low-frequency nature of neologisms, general linguistic corpora do not allow to study representative sets of neologisms, which poses limits to drawing strong generalizations about the nature lexical innovation. Despite these limitations, case studies on selected neologisms (Hohenhaus 2006) and studies on specific domains of neology (Elsen 2004) have managed to shed light on the spread of new words in more specific domains.
- The advent of **web corpora** in the last two decades has provided researchers with bigger and less formal data to study lexical innovation.
  - The sheer size of big corpora → bigger samples
  - monitoring corpora (Davies 2013): tracking dynamics of diffusion, closer to coining
  - In particular, a range of tools enabled the creation of specialized corpora for the investigation of neologisms. (Renouf, Kehoe & Banerjee 2006, Kerremans, Stegmayr & Schmid 2012, Lemnitzer 2018, Gérard 2017, Cartier 2017)
  - the nature of web corpus data is particularly suitable for investigating lexical innovations as
    * language on the web is very creative,
    * more informal sources, bigger spectrum of language use
    * new words often first occur on the web
    * and use on the web significantly influences whether these new formations catch on or not.
  - Web corpora thus promise insights into diffusion across
    * contexts: e.g. whether new words such as *blockchain* are increasingly used in less formal
    * cotexts: e.g. whether new words such as *XXX* are increasingly used in more diverse cotexts

3.2 Going beyond frequency

The conventionality of linguistic units is commonly assessed by counting how often they are found to be used in linguistic corpora, with high frequencies of occurrence seen as indicators of high levels of conventionality. Diffusion as a process that drives increasing conventionalization is thus usually assumed to be reflected by increases in the usage frequency of linguistic innovations. Previous research on lexical innovation has largely been limited to this approach and has evaluated the spread and the overall success of new words on the basis of the number of tokens found in linguistic corpora. This paper takes usage frequency as a baseline and uses social network analysis to go beyond frequency to discover sociolinguistic dynamics of diffusion and conventionality that might have eluded previous frequency-based approaches.

Frequency measures are widely used to study linguistic phenomena on all levels, from investigating phonological preferences between communities, to studying the increasing establishment of grammatical constructions like the *going to*-future over time, to assessing the degree to which words are conventional lexical units of a language.

Usage frequency is thus commonly used by a diverse set of linguistic sub-disciplines. From a structural perspective, for example, co-occurrence frequencies of multi-word units such as *handsome man* are taken as an indicator for whether these are free combinations or more or less fixed collocations in a language system. Historical linguistics investigates phenomena like language change and grammaticalization, by analysing changes in usage frequency of certain constructions like the *going to*-future over time. Cognitive and psycholinguistic research usually relies on frequency measures to approximate the degree to which speakers are familiar with words that are presented as linguistic stimuli in experiments to control for effects on experimental results.

This paragraph is unnecessary.

The reliance on usage frequency as a measure for different phenomena in these diverse research contexts has faced substantial criticism. Stefanowitsch & Flach 2017 provide a good overview of the theoretical assumptions and problems that underlie frequency-based approaches in corpus linguistics.

freq. esp. insufficient for lex. inn.

(1) highly socially indexical and thus especially prone to be used only by certain sub-communities, (2) topical which makes freq. less reliable bc. it fails to capture 'dormant' passive knowledge of the words

When assessing the suitability of usage frequency as a measure for the diffusion and conventionality of neologisms a set of assumptions underlying the frequency-based approach need to be disentangled. While these theoretical and methodological considerations generally apply to all corpus-linguistic work, the focus will be on the current issue of lexical innovation.

I adopt Schmid's EC-Model (Schmid 2020) as a framework for defining and delimitating the concepts of 'conventionalization' and 'diffusion'.

> [...] I define diffusion as a process that brings about a change in the number of speakers and communities who conform to a regularity of co-semiotic behaviour and a change in the types of cotexts and contexts in which they conform to it.

Aside from the sociolinguistic perspective on diffusion ('number of speakers and communities'), Schmid conceptualizes diffusion as a multi-dimensional process that also takes into account changes on the syntagmatic ('cotexts') and pragmatic ('contexts') level. This paper will focus on the sociolinguistic dimension of diffusion and will leave an integrative approach including all three perspectives for further research.

Applying this definition to the context of lexical innovation thus implies that the successful diffusion of a new word is marked by it being known and used by an increasing 'number of speakers and communities'.

By contrast, in a strict sense, usage frequency counts of a lexeme represent the total number of tokens produced by *all* speakers who have *contributed* to the *target text corpus*. The discrepancy between the theoretical definition of diffusion adopted here and the exact information contained by frequency show that this operationalization relies on a number of assumptions that let it only approximate the construct to be measured.

Firstly, usage frequency does not provide direct information as to how many *individual speakers* have used a new word. Especially in the case of neology, there are certain new words that are disproportionally used and propagated by a relatively small, but more active and dedicated users of the new term. This leads to high overall frequency counts which falsely suggest that larger parts of the speech community have adopted the term.

Secondly, usage frequency only captures active uses of a term and fails to include how many speakers have been passively exposed to neologisms. In the context of entrenchment, Stefanowitsch & Flach (2017) refer to this problem as the 'corpus-as-input' and 'corpus-as-output' hypotheses. The underlying assumption is that the output of the speakers who have contributed to a corpus can serve as an approximation for the potential linguistic input of a comparable speaker group. Frequency thus reflects the 'usage intensity' of neologisms in the speech community, which in turn indicates the degree of entrenchment in individual speakers as well as an approximation of the conventionality of the neologisms in the speech community. In the case of lexical innovation this can be problematic as questionnaires studies on the use of neologisms (Kerremans 2015) show that many speakers report that they have come across target neologisms, but have not actively used them in discourse. Relying on frequency counts only can thus often lead to under- or overestimating the degree of diffusion of neologisms.

Thirdly, usage frequency fails to capture where new words diffuse across 'communities of speakers', as suggested by Schmid's definition. This is, of course, a consequence of the fact that frequency counts cannot provide direct information about the number of speakers involved in the diffusion of neologisms, as was pointed out in the first two points above. New words often stem and quickly spread within tight-knit communities of practice that share common attitudes or interests. Frequency measures alone cannot detect whether neologisms only show increasing usualization within these groups or whether they diffuse and become conventional in other parts of the speech community, which represents an essential feature of the sociolinguistic dimension of diffusion.

how much the words might have diffused outside the *target text corpus*

temporal dynamics (e.g. *millennium bug*)

4 Data

4.1 Neologism sample

I base my empirical study on a selection of 100 neologisms and study their use on Twitter from the start of the platform 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‘. My sample was designed to cover these categories and largely contains neologisms taken from the NeoCrawler, 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, Nini & Guo (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.

4.2 Twitter data

Twitter is a popular micro-blogging platform that was started in 2006 and has become one of the most popular social media platforms today.

The Twitter community is not a perfect reflection of society and the speech community as a whole, of course, since certain social groups are over- or underrepresented according to social variables such as regional background and age. Nevertheless, 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. Twitter corpora have been successfully used to identify patterns of sociolinguistic variation in numerous previous studies. A recent study by Grieve et al. (2019) has, for example, shown the reliability of large-scale Twitter datasets for studying lexical variation.

reference

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).

retrospective, longitudinal, social network information, scope of the paper

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 retrospectively 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 100 lexemes in the sample contains 29,912,050 tweets. The first tweet dates from 5 May, 2006 and involves the neologism *tweeter*, the last tweet in the collection is from 31 December, 2018, and includes *dotard*.

5 Method

I processed the dataset to remove duplicate tweets, tweets that do not contain tokens of the target neologism in the tweet text, and all instances where tokens only occurred as parts of usernames.[1] Hashtag uses were included in the analysis. Retweets were excluded, since *twint* does not consistently provide metadata which would allow to include retweeting activity in the social network analysis. The resulting dataset contains about 30 Mio. tweets which each contain at least one instance of the 100 neologism under investigation.

To investigate the diffusion of these lexemes in terms of usage intensity (Stefanowitsch & Flach 2017), I compared time-series data based on 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 regarding their overall lifespan. I visualize the resulting time series as presented in Figure XXX, adding the *loess* function to indicate the smoothed trajectory of usage frequency over time.

I calculated the coefficient of variance for all time series to capture different degrees of stability vs. volatility in the use of neologisms over time. 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, e.g. 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 slices 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 accomodation processes during which new words increasingly spread in social networks of users on Twitter. This limit was validated empirically by testing different combinations of threshold values for the offset of number of users and interactions among early uses. 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 slice are marked by dashed vertical lines in the usage frequency plots presented below (e.g. Figure XXX).

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

[1] The post-processing as well as all quantitative analyses were performed in R (R Core Team 2018), and the source code is available on GitHub: `https://github.com/wuqui/sna`

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 suggest 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, Romero & Wu 2008) While users often follow thousands of accounts, their number of interactions with others provides a better picture of their individual social networks, which is much more limited in size (Dunbar 1992).

To construct the networks, I extracted users and interactions from the dataset to build a directed graph.[2] Nodes in the graph correspond to individual Twitter users, edges represent interactions between users. I capture multiple interactions between speakers by using edge weights, and I account 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 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 centalities 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 would have high centralization in early stages, for example, 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. insert formula This enables the comparison of graphs of different sizes, which is essential for drawing comparisons across lexems in the present context. The neologisms under investigation differ with regard to their lifespan and usage intensity, which results 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

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

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 8 biggest subcommunties in each graph by color.

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 analyze 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, Heymann & Jacomy 2009). 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 neologism. 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, Gerlitz & Rieder 2013, Bliss et al. 2012). To assess and visualize the influence of individual users in the social network, I use the PageRank algorithm (Brin & Page 1998) (visualized by node size), and I account for varying degrees of strength in the connection between users by using edge weigths for repeated interactions (visualized by edge thickness).

6 Results

6.1 Frequency-based measures of diffusion

6.1.1 Total usage frequency

As described in Section 2, 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 these words in the corpus. The most fundamental way of using this information is to cumulate 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 & 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

Table 1: Total usage frequency in the corpus.

(a) 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

(b) 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

(c) Least frequent lexemes.

lexeme	freq
microflat	426
dogfishing	399
begpacker	283
halfalogue	245
rapugee	182
bediquette	164

(d) 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

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. Table 1 presents the candidates under investigation in four groups: six examples around the 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 diverse set of prototypical examples of diffusion patterns, 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 clustering of neologisms on the basis of their total usage frequency presented in Table 1 largely seems to fit intuitions about diverging degrees of conventionality between the frequency-based groups 1a, 1b, and 1c. Neologisms such as *blockchain* and *smartwatch*, which are probably familiar to most readers, can be assumed to be more conventional than neologisms from the other 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 quantitave assessment of its degree of conventionality in two directions. Firstly, it bears the danger of overestimating the status of words such as *millenium bug*[3], whose total usage frequency largely goes back to a short period of highly intensive usage, after which they fall into oblivion and are unfamiliar to following generations of speakers, and eventually become 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 1a, 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. Due to its shorter lifespan, *blockchain* accumulated roughly the same number of uses, but shows significantly higher usage intensity in the more recent past, and could 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 provide a more fine-grained picture of diffusion.

total usage frequency will be used to evaluate the social network metrics trend: convex vs. concave

6.1.2 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 1d, the offsets and slopes of the trajectories of usage frequency reveal additional characteristics about differences in their diffusion patterns.

The selected cases 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 cases, 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 it 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 *entrepeneur*, it keeps a low, flat trajectory 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. (27 July, 2007)

(2) hmmmmmmm new word added to my vocab = 'solopreneur' !! (6 January, 2008)

[3] The neologisms *millenium bug* was used to refer to ancipated technical problems caused by inconsistent formatting of timestamps at the turn of the century.

[4] *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 2d.

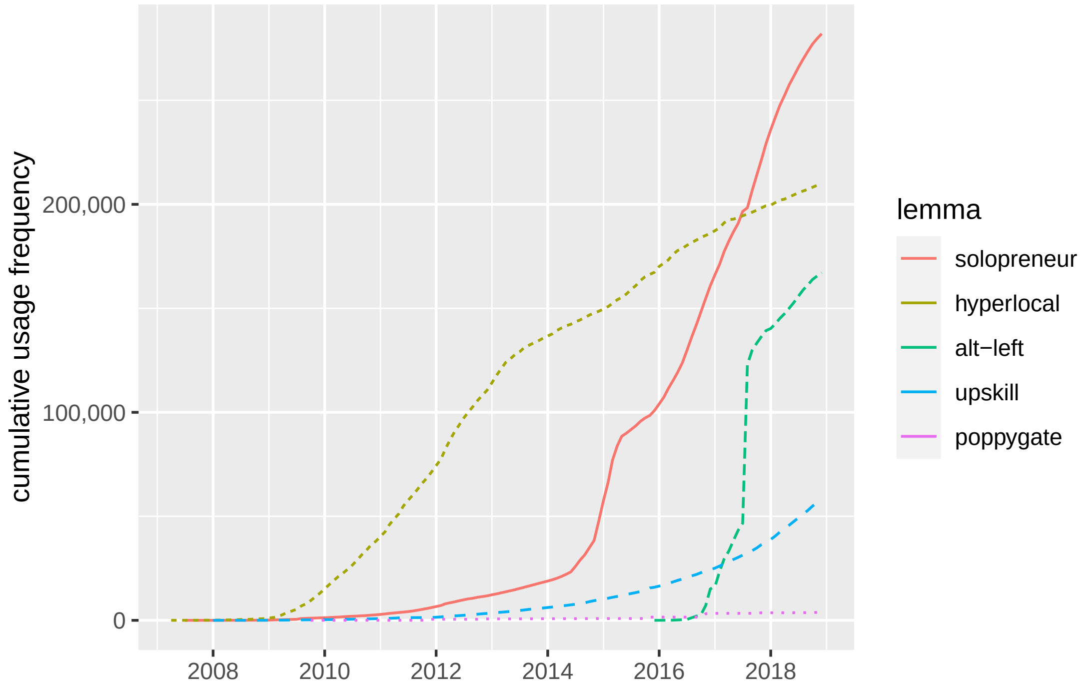


Figure 1: Cumulative increase in usage frequency for case studies.[4]

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 Karen Banes). 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 use.

In addition to lifespan differences, 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 evenly distributed over time.

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*, which was 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 1 shows that the majority of its use goes 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 cases illustrates that similar total frequency counts of neologisms can be the product of highly different pathways of diffusion.

These data complement total counts in that they show differences in the total lifespan and in the intensity and with which a neologism was used over time, which is relevant for assessing the degree to which it has spread in the speech community.

6.1.3 Absolute frequency

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

Most importantly, analysing usage intensity highlights to which degree new words are

LIFESPAN should be discussed here; with reference to diffusion cut-offs; for later evaluation of SNA metrics potentially also a table for

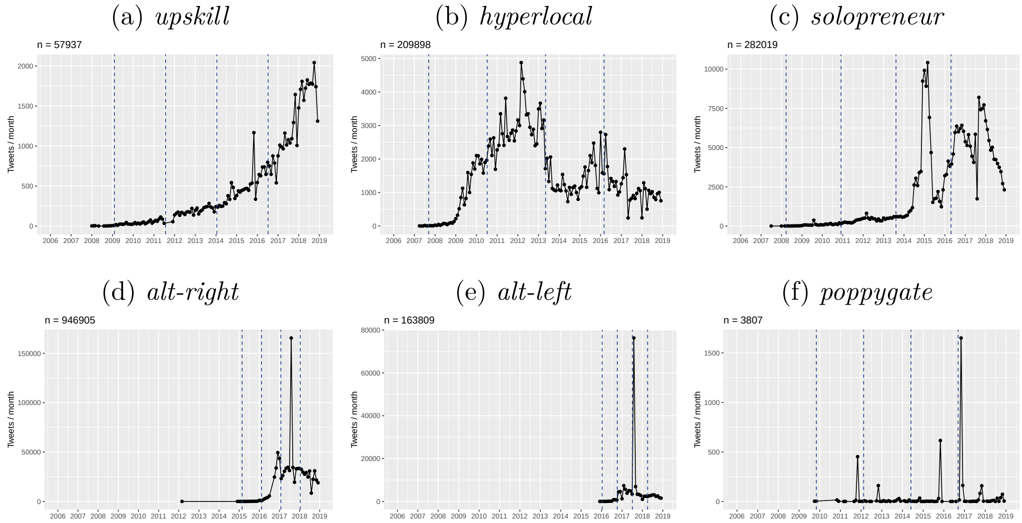


Figure 2: Temporal dynamics in usage frequency for the case study neologisms.

being used consistently over time. Figure 2 presents this information for the selected cases. In the following section, I will illustrate prototypical differences by referring to the selected cases, before I discuss the results for the full sample.

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.

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 smooth trajectories, the plots in Figure 2 indicate that the selected cases differ significantly in terms of the consistency with which they are used in the corpus.

gradual increase

The neologism *upskill* shows the smoothest trajectory of diffusion among the candidate neologisms. 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 the determining influence of individual influential 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.

consistent use

While *hyperlocal* also exhibits a marked increase in usage frequency during its earlier stages, its peak of popularity is followed by a decline in use, after which it settles at a relative 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, and seems to attract a stable community of users on Twitter.

inconsistent use

The neologisms *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.

topicality

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 showed significant diffusion in 2016, particularly in the period following up to Donald Trump's election, after which *alt-right* was consistently used 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 returns to its previous usage intensity, the use of *alt-left* seems to largely disappear from Twitter in the aftermath of the event.

recurrent topicality

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

'topicality' (Fischer 1998); 'recurrent semi-conventionalizat remans 2015))

6.1.4 Coefficient of variation

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 variation in usage frequency of words over their lifespan relative to their average frequency of occurrence in the corpus. Table 2 presents the coefficients of variation for the selected cases, 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 Table 2 show that the sample covers a wide spectrum of variability 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

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

Table 2: Coefficients of variations.(VAR)[5]

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

(a) Variation among the selected cases.

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

(b) Lowest degrees of variation.

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

(c) Highest degrees of variation.

neologisms in Table 1a and is also one of the oldest neologisms, with its first attestation in the corpus dating back to 19 December, 2006.

By contrast, the group of lexemes that show the highest degree of variation in usage frequency is comprised by 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*[6] was promoted by extralinguistic political events, *upskirting*[7] and *youthquake*[8] 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*[9] exhibit recurrent topicality, and are typically only used in the contexts of their seasonal relevance in autumn and winter.

The selected cases cover the spectrum of variability in usage frequency found in the full sample of neologisms, and the coefficients of variation are in line with the previous analysis of the frequency-based time-series visualisations presented in Figure 2.

6.1.5 Summary of frequency-based measures

So far, I have used frequency-based visualisations and metrics to assess the degrees and pathways of diffusion of the neologisms in the sample. In a first step, I used the most common measure for assessing the conventionality of new words: their total frequency of occurrence in the corpus. In the following steps, I extended the frequency-based approach by including temporal information in the analysis. Zooming in on the temporal dynamics of use surfaced different pathways of diffusion. Notably, it revealed substantial differences

[6] 'proponent 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 USA.'

[7] 'The habit or practice of taking upskirt photographs or videos.' (OED)

[8] '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/`)

[9] 'Cherpumple 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.

in the diachronic usage profiles of neologisms with comparable total frequency.

Within the group of selected cases, *hyperlocal*, *solopreneur*, and *alt-left*, for example, would all be placed in the medium range of the conventionality continuum if grouped by total usage frequency alone, as presented in Table 1d. Taking this most basic measure as an indicator of degrees of diffusion, it would seem that these words are roughly equally conventional among users on Twitter. However, adding the temporal dynamics of their use in the corpus to the picture revealed significant differences between their diachronic usage profiles, which seems important for assessing their pathways and degrees of diffusion in a more differentiated and accurate way.

Visualising the cumulative increase in uses over time (Table 1) for *hyperlocal*, for example, shows a stable linear trend, which indicates that its total frequency count has been the product of relatively consistent use over its relatively long lifespan. Its temporal usage profile in Figure 2b and confirms these observations and presents its initial period of accelerated diffusion followed by an extended stable level of relatively consistent use over the last five years of its observed lifespan. This consistency is further corroborated by its low coefficient of variation (Table 2a). In sum, the balanced nature of this frequency-based usage profile suggests a relatively organic trajectory of diffusion, culminating in a solid degree of conventionality in the recent past. The fact that *hyperlocal* was added to the OED in 2015 supports these observations.

By comparison, *solopreneur* has a slightly higher overall frequency of occurrence in the corpus, yet its use is less stable over time. While its overall lifespan is about as long as that of *hyperlocal*, its cumulative distribution shows that the majority of its use goes back to a relative short period of intensive use, after which it exhibits a slightly negative trend in later stages. Both the visualization of its temporal frequency profile as well as its coefficient of variation demonstrate a higher degree of fluctuation in its popularity. This temporal usage pattern suggests that its diffusion was influenced significantly by effects of topical salience. While *solopreneur* has been used in a high total number of tweets in the corpus, it thus seems less certain whether its use will become stable over time, and to which degree its use extends beyond the entrepreneurial community which triggered the main spurt of its diffusion in the second half of 2014.

Lastly, *alt-left* is in the same range of total usage frequency, but its use is much more unevenly dispersed across the corpus than that of the remaining selected cases. The term is much younger, and its cumulative increase in uses illustrates that diffusion is largely limited to a very short, highly intensive period of use, after which it shows a strong negativ trend in its usage frequency. Its diachronic frequency profile and its coefficient of variation correspondingly demonstrate very high levels of fluctuations in its use. Since the short period of intense use of *alt-left* can clearly be traced back to participants of the Unite the Right Rally in Charlottesville in August 2017, it seems plausible that its popularity has never extended beyond this topical event and beyond this particular community of like-minded individuals.

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 substantial limitations of frequency-based approaches to studying diffusion. The present data demonstrate high degrees of variation in the degrees of diffusion of neologims that are similar in terms of their frequency of occurrence in the corpus. Such discrepancies could partly be resolved by in-depth analyses of temporal usage profiles in combination with insights from corpus data and extralinguistic events.

These in-depth analyses of diffusion are not possible by a systematic frequency-based analysis alone, however, and they cannot be extended for large-scale analyses of bigger samples of neologisms. It hence remains unknown to which degree frequency-based metrics adequately capture pathways and degrees of diffusion. In the following section, I will address the limitations of the frequency-based approach by using Social Network Analysis to get a more differentiated view on the sociolinguistic aspects of diffusion.

6.2 Social Network Analysis

As discussed in Section 2.3, from a theoretical, sociolinguistic perspective, the degree of diffusion of lexical innovations crucially depends on the degree to which new words spread to become familiar among new speakers and communities of speakers.

Frequency-based analyses, as presented in the previous section, can by definition only provide information about the distribution of occurrences in the corpus, they cannot provide direct evidence about the size and composition of the community of speakers who produced these uses.

Social network analysis, by contrast, is based on the communicative behaviour of speakers in the corpus and can thus provide direct insights into the social characteristics of speakers and groups of speakers. This allows for a more direct operationalization of theoretical model of diffusion. The structural characteristics of the social network of speakers who have used a target neologism can be used to measure whether the term has been used by a broad section of the speech community, or whether its use remains limited to smaller communities.

going beyond frequency

results

total frequency $\neq$ number users/communities similar counts, different underlying temporal patterns (trend, lifespan, volatility) frequency time-series slopes (trend) gradual diffusion: *upskill* lifespan overestimate: obsolete: *millenium bug* underestimate: new, but established: *coronavirus* volatility one-hit wonders: *alt-left* consistent use: *hyperlocal*

particularly important for *lexical innovations* due to the nature of the process

bound to cultural conceptual salience (variable 'semantic carrying capacity' (Nini et al. 2017)) social indexicality open nature of the lexicon

full sample

stability: shows that freq. is problematic 'dormant' spikes distort representativity of frequency for degree of conventionality In this case, a less steady, more abrupt increase in usage over a relatively short period as with *solopreneur* could be a sign of high usage intensity within one community ('usualization'), rather than an indication of sustained spread to larger numbers of other speakers and communities ('diffusion'). underestimate:

poppygate not forgotten in troughs overestimate: cumulating hides the fact that words like *millenium* do get lost

going beyond frequency

In the following sections I will assess the value of usage frequency and compare and complement it with social network information about the diffusion of lexical innovations.

6.3 Social networks of diffusion

6.3.1 Centralization over time

(pathways of diffusion)

going beyond frequency

def. diffusion: numbers of users communities

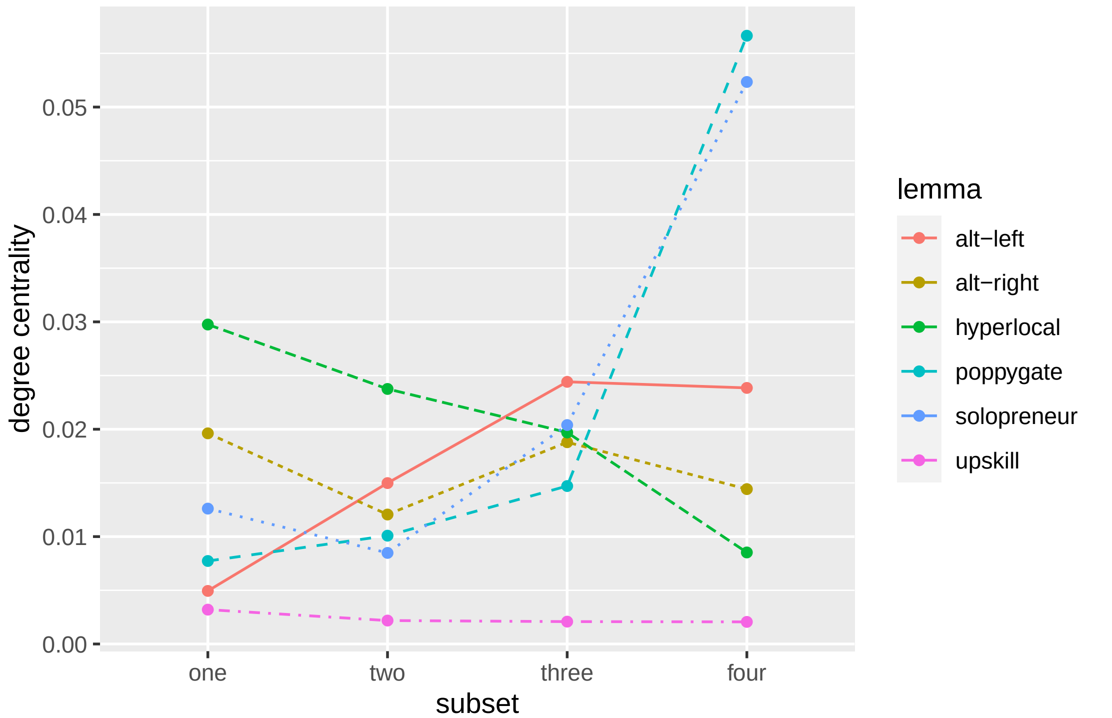
subsetting / time slices start of diffusion process 4 quarters

explain: degree centralization

case studies example where freq. meets nets example where nets add to freq.: *alt-left*

Overview of changes in centralization for case studies.

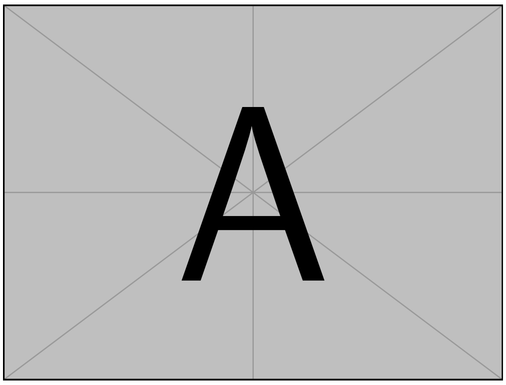
Figure 3: Degree centralization over time for case study words.



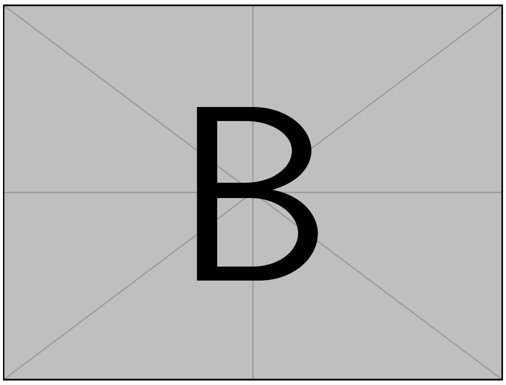
Social diffusion over time: *hyperlocal*

Figure 4: Social network of diffusion for *hyperlocal* over time.

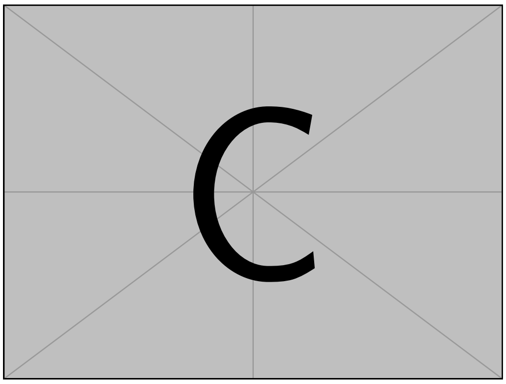
(a) First stage



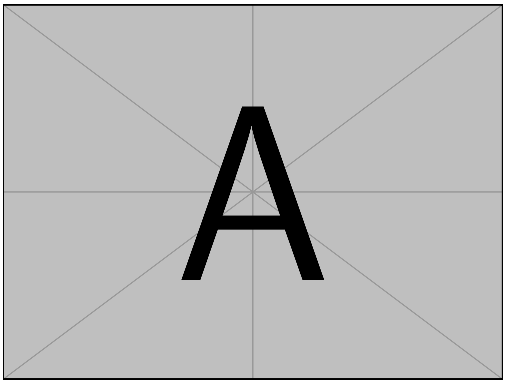
(b) Second stage



(c) Third stage



(d) Fourth stage



Full sample

density successful unsuccessful
biggest changes

6.3.2 Degrees of social diffusion

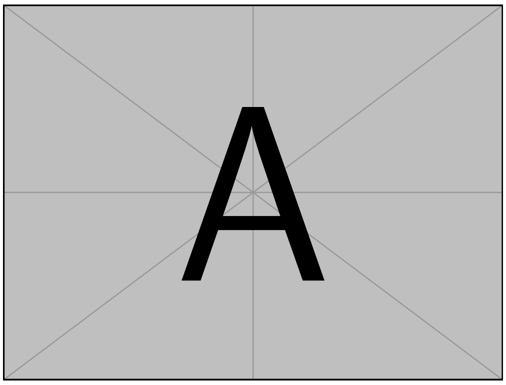
Metrics

most diffused least diffused

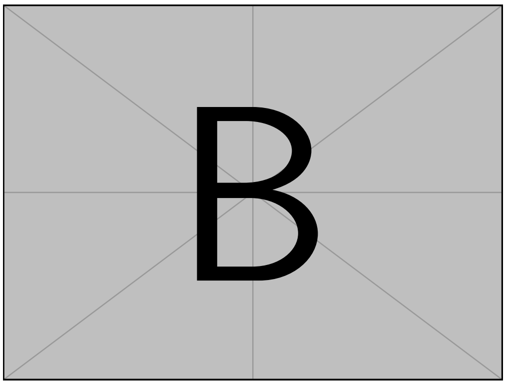
Network graphs

Figure 6: Social network of diffusion for *hyperlocal* over time.

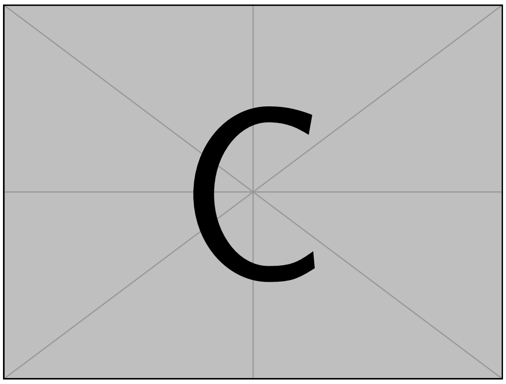
(a) *upskill*



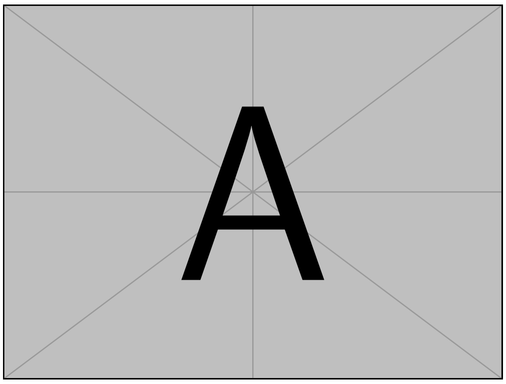
(b) *hyperlocal*



(c) *solopreneur*



(d) *alt-left*



6.4 Networks vs. frequency

6.4.1 Correlation

Table 3: Correlations of CENTRALITY ('degree centralization') with the variables total usage frequency (FREQUENCY), coefficient of variation (VARIATION IN FREQUENCY), and observed lifespan (LIFESPAN) across the full sample of neologisms ($n = 100$); reporting correlation coefficients and p-values for Spearman's ρ (Spearman 1961) and Kendall's τ (Kendall 1938).

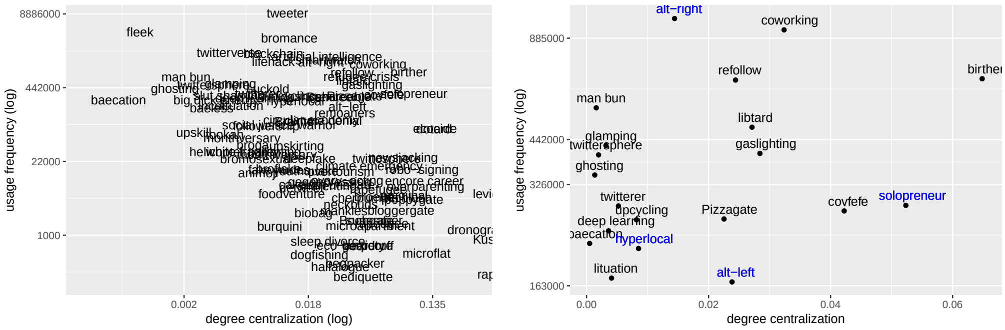
	Spearman		Kendall	
	ρ	p	τ	p
USAGE FREQUENCY	−0.40	0.000 05	−0.27	0.000 06
LIFESPAN	−0.29	0.003 67	−0.20	0.002 93
VARIATION IN FREQUENCY	0.29	0.003 45	0.20	0.004 12

6.4.2 Discrepancies

However, we also see discrepancies

plots

(a) Full sample. (b) Case studies.



cluster analysis

freq. overestimating topical propaganda: *alt-right*, *alt-left*, *covfefe*, *birther* Brexit terms: *Brexiteer*, *Brexiter*, *Brexit* nerds technical

freq. underestimating: XXX topical words

Social network metrics certainly add to freq.

7 Discussion

- I have investigated the diffusion of new words on Twitter.
- I have focused on diffusion as regards the spread to new speakers and communities.
- I have used frequency as a baseline and have gone beyond frequency to zoom in on the sociolinguistic dynamics of diffusion.
- I have shown that Social Network Analysis can complement frequency measures

frequency can be a reasonable baseline more differentiated view possible through SNA

case studies equally frequent words show different temporal patterns pathways of diffusion degrees of diffusion

sample: this holds for the full sample

import for lexical innovation

but also for other linguistic domains

should be used more in (computational) sociolinguistics

- freq. empirically proves to be a pretty good indicator
- drawbacks of cumulated, total counts: temporal dynamics important
- but rough approximation due to inferences

    - active vs. passive use
    - number of speakers
    - groups of speakers
- temporal dynamics important
    - time window: age, lifespan
        * starting point: late: very new words, would be under-represented by total counts
        * end point: words might have already grown out of use (e.g. *millenium bug*) > over-represented by total counts
        * length: quick vs. slow and steady increase: topicality
    - volatility: high topicality / highly fluctuating communicative need (cf. *going to*-future); thus fluctuating "semantic carrying capacity" (Grieve, Nini & Guo 2018)
- dynamics in usage intensity might reflect social dynamics
    - s. S-curve model: early adopters etc.
    - e.g.: fast rise due to rapid spread in certain communities
    - hard to infer from usage frequency > SNA needed
- social network dynamics important
    - esp. w.r.t. new words: often community-specific; are coined within tight-knit comminuties and (cf. grammatical change) and have socio-indexical function (e.g. youth language)
- cross-validation between
- cross-checking other data sources (NOW corpus) shows validity
- social network analysis can be an important tool for sociolinguistics
    - extend sociolinguistic research (on geographical variation; desideratum in Grieve et al. 2019)

- cross-validation of frequency and SNA information
    - systematic comparison with web data, e.g. NOW corpus (Davies 2013); early attempts: Würschinger et al. 2016
    - questionnaires; early work: Kerremans 2015
- investigation on diffusion across
- contexts: different web registers (Biber & Egbert 2016)
- cotexts: use word embeddings to study

- semantic innovation, meaning change
- and variation and change between communities (Tredici et al. 2019)

8 Conclusion

Summary

- I have investigated the diffusion of new words on Twitter.
- I have focused on diffusion as regards the spread to new speakers and communities.
- I have used frequency as a baseline and have gone beyond frequency to zoom in on the sociolinguistic dynamics of diffusion.
- I have shown that Social Network Analysis can complement frequency measures

  Importance of/for computational sociolinguistics

- new data like Twitter
- new methods like SNA, simulations, AI for embeddings
- progress: ideas, scholars and progress will "go viral"

References

Bastian, Mathieu, Sebastien Heymann & Mathieu Jacomy. 2009. *Gephi: An Open Source Software for Exploring and Manipulating Networks*. `http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154`.

Bauer, Laurie. 1983. *English word-formation*. Cambridge university press.

Biber, Douglas & Jesse Egbert. 2016. Register variation on the searchable web. *Journal of English Linguistics* 44(2). 95–137. `https://doi.org/10.1177/0075424216628955`.

Bliss, Catherine A., Isabel M. Kloumann, Kameron Decker Harris, Christopher M. Danforth & Peter Sheridan Dodds. 2012. Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Advanced Computing Solutions for Health Care and Medicine* 3(5). 388–397. `https://doi.org/10.1016/j.jocs.2012.05.001`.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte & Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10). P10008. `https://doi.org/10.1088/1742-5468/2008/10/p10008`.

Blythe, Richard A. & William Croft. 2012. S-curves and the mechanisms of propagation in language change. *Language* 88(2). 269–304.

Brin, Sergey & Lawrence Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Seventh International World-Wide Web Conference (WWW 1998)*. Brisbane, Australia. `http://ilpubs.stanford.edu:8090/361/` (4 June, 2020).

Bruns, Axel. 2012. How long is a tweet? Mapping dynamic conversation networks on Twitter using Gawk and Gephi. *Information, Communication & Society* 15(9). 1323–1351. https://doi.org/10.1080/1369118X.2011.635214.

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.

Davies, Mark. 2013. *Corpus of News on the Web (NOW) - 3+ Billion Words from 20 Countries, Updated Every Day.* https://www.english-corpora.org/now/.

Dunbar, Robin IM. 1992. Neocortex size as a constraint on group size in primates. *Journal of human evolution* 22(6). 469–493.

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.

Fischer, Roswitha. 1998. Lexical change in present day English. A corpus based study of the motivation, institutionalization, and productivity of creative neologisms.

Freeman, Linton C. 1978. Centrality in social networks conceptual clarification. *Social Networks* 1(3). 215–239. https://doi.org/10.1016/0378-8733(78)90021-7.

Gérard, Christophe. 2017. The logoscope: Semi-automatic tool for detecting and documenting the contexts of french new words.

Gerlitz, Carolin & Bernhard Rieder. 2013. Mining One Percent of Twitter: Collections, Baselines, Sampling. *M/C Journal* 16(2). http://www.journal.media-culture.org.au/index.php/mcjournal/article/view/620.

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. 1977. The strength of weak ties. In Samuel Leinhardt (ed.), *Social networks*, 347–367. Academic Press. https://doi.org/10.1016/B978-0-12-442450-0.50025-0.

Grieve, Jack. 2017. Geographical patterns of lexical innovation. Workshop 'Diffusion of Lexical Innovations', LMU Munich.

Grieve, Jack. 2018. Natural selection in the modern English lexicon. In *Proceedings of EVOLANG XII.* Torun, Poland.

Grieve, Jack, Chris Montgomery, Andrea Nini, Akira Murakami & Diansheng Guo. 2019. Mapping lexical dialect variation in British English using Twitter. *Frontiers in Artificial Intelligence* 2. 11. https://www.frontiersin.org/article/10.3389/frai.2019.00011.

Grieve, Jack, Andrea Nini & Diansheng Guo. 2016. Analyzing lexical emergence in Modern American English online. *English Language and Linguistics* (21). 99–127.

Grieve, Jack, Andrea Nini & Diansheng Guo. 2018. Mapping lexical innovation on American social media. *Journal of English Linguistics.*

Hébert-Dufresne, Laurent, Samuel V. Scarpino & Jean-Gabriel Young. 2020. Macroscopic patterns of interacting contagions are indistinguishable from social reinforcement. *Nature Physics*. https://doi.org/10.1038/s41567-020-0791-2.

Hohenhaus, Peter. 1996. Ad-hoc-Wortbildung. Terminologie, Typologie und Theorie kreativer Wortbildung im Englischen.

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.

Huberman, Bernardo A., Daniel M. Romero & Fang Wu. 2008. *Social Networks That Matter: Twitter under the Microscope*. http://arxiv.org/abs/0812.1045 (21 June, 2020).

Jacomy, Mathieu, Tommaso Venturini, Sebastien Heymann & Mathieu Bastian. 2014. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLOS ONE* 9(6). e98679. https://doi.org/10.1371/journal.pone.0098679.

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika* 30(1-2). 81–93. https://doi.org/10.1093/biomet/30.1-2.81.

Kerremans, Daphné. 2015. *A Web of New Words*. Bern, Schweiz: Peter Lang. https://doi.org/10.3726/978-3-653-04788-2.

Kerremans, Daphné, Jelena Prokić, Quirin Würschinger & Hans-Jörg Schmid. 2019. Using data-mining to identify and study patterns in lexical innovation on the web: The NeoCrawler. *Pragmatics and Cognition* 25(1). 174–200.

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.

Labov, William. 2007. Transmission and diffusion. *Language* 83(2). 344–387.

Lemnitzer, Lothar. 2018. *Wortwarte*. http://www.wortwarte.de/ (14 January, 2018).

Lipka, Leonhard. 2005. Lexicalization and institutionalization: Revisited and extended. *SKASE Journal of Theoretical Linguistics* 2(2). 40–42.

Lu, Fred Sun, Suqin Hou, Kristin Baltrusaitis, Manan Shah, Jure Leskovec, Rok Sosic, 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). e4. https://doi.org/10.2196/publichealth.8950.

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. https://www.cambridge.org/core/article/linguistic-change-social-network-and-speaker-innovation1/EB30A7117CC09F6EDA5255BF9D788D5A.

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. http://www.helsinki.fi/varieng/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/`.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing.* manual. R Foundation for Statistical Computing. Vienna, Austria. `https://www.R-project.org/`.

Renouf, Antoinette, Andrew Kehoe & Jayeeta Banerjee. 2006. WebCorp: an integrated system for web text search. *Language and Computers* 59(1). 47–67.

Rogers, Everett M. 1962. *Diffusion of innovations.* New York: Free Press of Glencoe.

Schmid, Hans-Jörg. 2008. New words in the mind: Concept-formation and entrenchment of neologisms. *Anglia – Zeitschrift für englische Philologie* 126. 1. `//www.degruyter.com/view/j/angl.2008.126.issue-1/angl.2008.002/angl.2008.002.xml`.

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.

Spearman, C. 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. 45 pp. `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.

Tredici, Marco Del, Diego Marcheggiani, Sabine Schulte im Walde & Raquel Fernández. 2019. You shall know a user by the company it keeps: Dynamic representations for social media users in NLP. `https://arxiv.org/pdf/1909.00412.pdf`.

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`.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo & Hiroaki Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4(43). 1686. `https://doi.org/10.21105/joss.01686`.

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: Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2605.

Acknowledgements

- UI: Max, Fabi
- SNA: Kauermann group
- Hans-Jörg