Social networks of lexical innovation

Investigating the diffusion of neologisms on Twitter

Quirin Würschinger
LMU Munich

List of Figures

List of Tables

Abstract

Societies continually evolve and speakers coin and use new words to talk about innovative products and practices. While most lexical innovations fall into oblivion, others spread successfully and become part of the lexicon. This paper investigates the diffusion of 99 English neologisms on Twitter, from the start of the platform in 2006 until the end of 2018. Previous work on lexical innovation has almost exclusively relied on usage frequency counts for measuring diffusion. After applying several frequency-based measures to the dataset, I use social network analysis to study the sociolinguistic dynamics of diffusion in more detail, and I cross-evaluate the results obtained from both approaches.

The results show that frequency counts lend themselves to approximate degrees of diffusion with varying success. While total usage frequency

While usage frequency counts can be misleading, incorporating temporal dynamics of use provides a better picture of diffusion.

However, frequency-based information alone fails to capture important sociolinguistic characteristics of diffusion. 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 what 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 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.

The 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 beyond the domain of lexical innovation.

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

under revision

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 alarming speed and has tragically affected the lives of people around the world. Its 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 and 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 (J. 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 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 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 provides 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 enables 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 2017; Grieve, Nini and Guo 2018)

should probably be removed

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. J. Milroy and L. 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 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 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, 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 of my 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, 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 Modeling and measuring the diffusion of lexical innovations

2 is under construction.

2.1 Modeling 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. 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 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 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)

Diffusion can be seen as the process that transports successful neologisms along this continuum while they are becoming increasingly conventional in the speech community.

One of the most influential frameworks for modelling language change and the social diffusion of linguistic innovations is the 'S-curve model' (J. Milroy 1992; Nevalainen 2015; Labov 2007). Since it focuses on the social dynamics and the roles of speakers in interactional networks during diffusion, the S-curve model provides an excellent framework for the present empirical analysis of spread in online social networks.

The model expects the trajectory of spread of innovations 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 trajectory, after which diffusion slows down and the curve plateaus towards the end of the process. These successive stages are assumed to correspond to sociolinguistic features of the community of speakers who are adopting the target innovation. In the 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 1977) play an important role in allowing the innovations to reach a broader spectrum 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. maybe shorten

EC Model

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').

conventionalization

'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) page number

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) page number

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)

page number

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

define 'community'

context: largely fixed as it is limited to use on Twitter comparison of cotexts outside the scope of this paper; initial work: (Schmid et al. 2020)

communities

'communities of practice' (Leuckert 2020)

definition

The resulting networks are interactional rather than static. (Goel et al. 2016) This makes them more similar to communities of practice than to traditional sociolinguistic networks based on static speaker characteristics such socio-economic status. In the case of lexical innovation, networks that are based on the interactions of speakers provide valuable information. In cases such as *alt-left*, for example, interactional networks show whether usage of the term remains centralized to a tight-knit community of speakers or whether it diffuses other sub-communities. Whether communities are distinct depends on whether users communicate with each other. While the reasons for theses communicative affiliations remain unknown (age, gender, socio-economic status), they are real in that users mutually engage in communicative interaction. (community = communication) It would be interesting to complement this information with static information (e.g. census data (Eisenstein et al. 2014)), however such data are currently not obtainable (geotags no longer provided, hard to infer; difficult to predict plus circular (e.g. gender)).

2.2 Measuring diffusion

2.2.1 Previous empirical approaches to the diffusion of lexical innovations

Large-scale empirical investigations of the diffusion of lexical innovations have only relatively recently become feasible with the advent of new data sources and new computational methods.

Earlier work on lexical innovation had to rely on smaller, general-purpose linguistic corpora. The low-frequency nature of neologisms thus prohibited studying larger, more representative samples of neologisms. Consequently, earlier studies conducted case studies on selected neologisms (Hohenhaus 2006) or on specific domains of neology (Elsen 2004).

The increasing availability of web corpora significantly extended the opportunities for large-scale corpus analyses. More recent 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, Kehoe and Banerjee 2006; Kerremans, Stegmayr and Schmid 2012; Lemnitzer 2018; Gérard 2017; Cartier 2017).

Aside from the technical advantages of size and temporal depth, the registers available on web corpora are better suited for the study of neologisms. Since the web covers informal and creative language, it constitutes an ideal environment for studying neo-

logisms. Moreover, new words are often coined on the web and propagated influences whether these new formations catch on or not.

social media corpora

characteristics language even more informal drive innovation even more influencers provide network information > number of speakers > SNA

previous studies general: Grieve, Nini and Guo (2016) geographical spread: Eisenstein et al. (2014), diffusion: Nini et al. (2017) communities: Stewart and Eisenstein (2018), Ryskina et al. (2020), Del Tredici and Fernández (2018)

2.2.2 Operationalising diffusion

diffusion spread to new speakers new communities

frequency counts used in most previous work several drawbacks (Stefanowitsch and Flach 2017) number of uses cannot directly capture key features number of users number of communities 'baseline'

temporal dynamics

social dynamics

3 Data

3.1 Neologism sample

I base my empirical study on a selection of 99 neologisms and study 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'. 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, Prokić 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 and 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.

3.2 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 reference

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, Montgomery et al. (2019) has shown, for example, the reliability of large-scale Twitter datasets for studying lexical variation.

Twitter is particularly well-suited to study 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 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 99 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*.

4 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 Million tweets which each contain at least one instance of the 99 neologism under investigation.

To investigate the diffusion of these lexemes in terms of usage intensity (Stefanowitsch and 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 according to 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.

To capture different degrees of stability vs. volatility in the use of neologisms over time,

[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`

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 limit was validated empirically by 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 (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 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, Romero and 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 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.[2] 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 produced network visualisations for all subsets for more detailed, qualitative analyses.

[2] I 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.

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 lexems 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.

insert formula

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 8 biggest subcommunities 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, Heymann and 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 and 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 and Page 1998) (visualized by node size), and I account for varying degrees of strength in the connection between users by using edge weights for repeated interactions (visualized by edge thickness).

5 Results

5.1 Usage frequency and diffusion

5.1.1 Total usage frequency

As described in Section NN, 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. 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 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 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 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 as-

Table 1: Total usage frequency (FREQ) 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

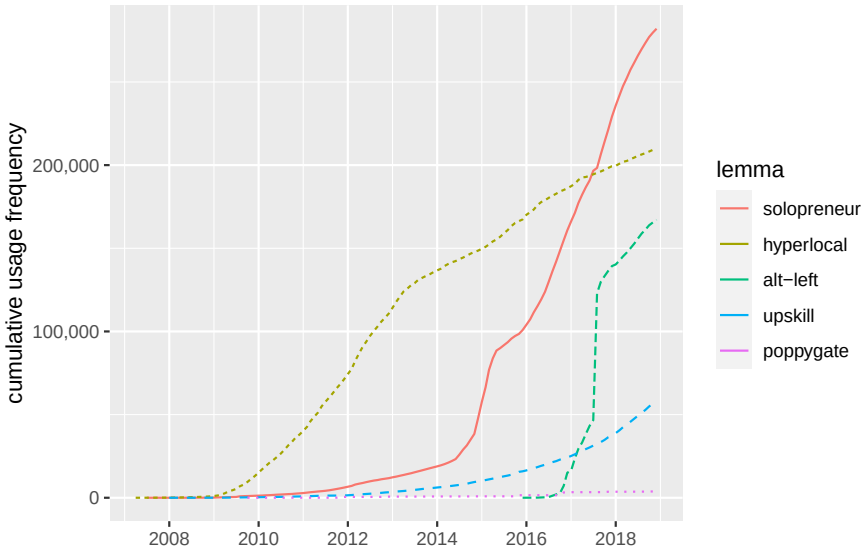
sessment of its degree of conventionality in two directions. Firstly, it carries the danger of overestimating the status of words such as *millennium bug*[3], whose total usage frequency largely goes back to a short period of highly intensive usage, after which they fall into oblivion, become unfamiliar 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 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. 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.

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

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



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

Shift the focus of this part to the TREND variable; discuss AGE primarly with reference to absolute counts.

TREND: convex trajectories » successful; concave » unsuccessful/obsolete

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 *entrepeneur*, it keeps a low, flat trajectory of sporadic use for about seven 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)

[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 3d.

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 less 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*, 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 uses go back to a single period of highly intensive use in the second half of 2017, soon after which it slows down considerably.

poppygate missing

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 clearly relevant for assessing the degree to which they have spread in the speech community.

5.1.3 Absolute frequency

change terminology

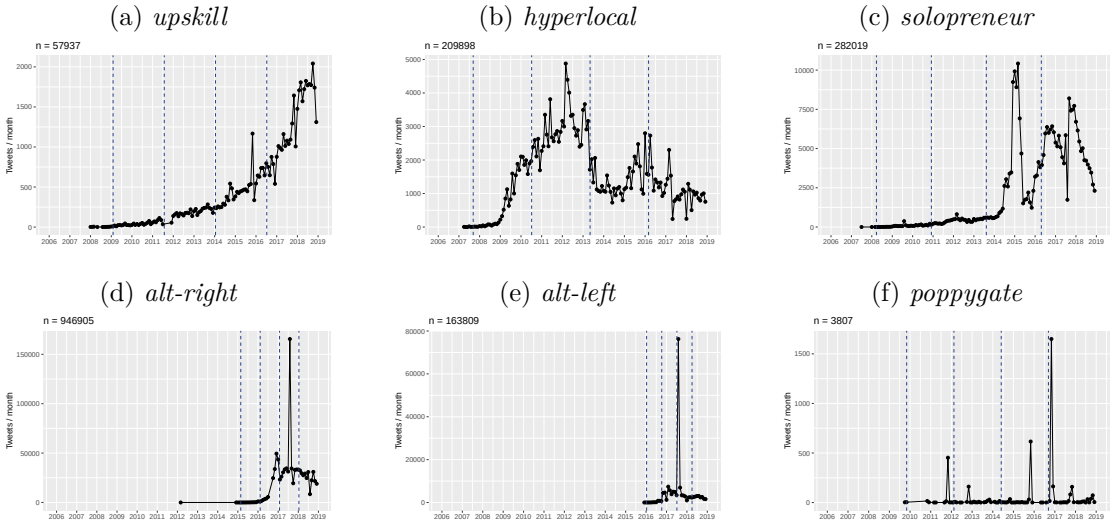
Going beyond cumulative counts, absolute usage frequency counts provide a more fine-grained view of the temporal dynamics of diffusion.

The following three paragraphs should probably be removed.

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

AGE should be discussed here; w. reference to diffusion cut-offs

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



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 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 neologisms differ significantly in terms of the consistency with which they are used in the corpus.

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 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.

add reference to Fig. 1

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 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.

definition

'topicality' (Fischer 1998); 'recurrent semi-conventionalization' (Remans 2015))

5.1.4 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 neologisms, as well as for the top and bottom six neologisms that show the highest and lowest degrees of variation in the sample.

reference to methods section

The results in Table 2 show that the sample covers a broad 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 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.

needs revision

[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 variation (VAR) for the selected neologisms, and for six neologism with the highest and lowest scores in the sample.[5]

(a) selected neologisms.

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

(b) Lowest scores.

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

(c) Highest scores.

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 variation 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*[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 neologisms 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.

5.1.5 Summary of frequency-based measures

I delete to save space.

So far, I have used frequency-based visualisations and metrics to assess different degrees of diffusion of the neologisms in the sample. In the 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 identified different pathways of diffusion. Notably, it revealed substantial differences in the diachronic usage profiles of neologisms with comparable total frequency.

[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.

Within the group of selected neologisms, *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 3b 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.

paragraph needs revision

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 similar to 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 profile 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 what 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 neologisms. 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 negative trend in its usage frequency. Its diachronic frequency profile and its coefficient of variation correspondingly demonstrate very high volatility 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.

mention *upskill*, potentially in place of *hyperlocal*

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 neologisms 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. However, these in-depth analyses of diffusion are not possible by means of a systematic frequency-based analysis alone, and they cannot be extended to the large-scale analysis of bigger samples of neologisms. Hence it remains unknown to what degree frequency-based metrics adequately capture pathways and degrees of diffusion. In the following section, I will complement the frequency-based approach by using Social Network Analysis to get a more differentiated view of the sociolinguistic aspects of diffusion.

needs revision; possibly remove

5.2 Social networks of diffusion

As discussed in Section NN, from a theoretical, sociolinguistic perspective, the degree of diffusion of lexical innovations depends on the degree to which new words become conventional 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 of neologisms in the corpus; they cannot provide direct evidence about the size and composition of the community of speakers who have produced the observed attestations. Social network analysis, by contrast, is based on data about the communicative behaviour of speakers in the corpus and can thus provide direct insights into the social characteristics of the speech community. This allows for a more direct operationalization of the theoretical model of diffusion. The structural characteristics of the social network of speakers who have used a target neologism can be used to assess whether the term has been used by large swaths of the speech community, or whether its use remains limited to certain pockets of the speech community.

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 a. 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

Table 3: Degree centrality scores (CENT) for the selected neologisms and six lexemes each for the highest and lowest scores in the sample; the scores are based on the most recent time slice for each neologisms in the corpus.

(a) Selected neologisms.

Lexeme	CENT
upskill	0.002 1
hyperlocal	0.008 5
alt-right	0.014 4
alt-left	0.023 8
solopreneur	0.052 3
poppygate	0.056 6

(b) Lowest scores.

Lexeme	CENT
baecation	0.000 5
fleek	0.000 9
ghosting	0.001 3
man bun	0.001 6
big dick energy	0.001 8
twittersphere	0.002 0

(c) Highest scores.

Lexeme	CENT
rapugee	0.258 0
levidrome	0.237 3
Kushnergate	0.230 9
dronography	0.153 0
dotard	0.097 9
ecocide	0.092 2

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.

removeable @HJS

Table 3 reports 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. Frequency and centrality generally tend to produce similar results when used to assess degrees of diffusion. This indicates that generally theres is a correlation between usage frequency and social diffusion, as one might expect. Notable deviations exist, however, and will be further discussed in Section 5.3.

to revise wrt. scores

Correspondingly, the neologisms with the highest centrality scores rank among the least frequent candidates in the sample. Notable trends among lexemes with low centrality scores are that they tend to be more recent (e.g. *dronography*[10]) and/or to exhibit high degrees of volatility (e.g. *ecocide*[11]). Moreover, this groups includes political terms such as *Kushnergate*[12] 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*.

[10] 'Dronography 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).

[11] 'the destruction of large areas of the natural environment as a consequence of human activity' (Merriam Webster Online Dictionary).

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

The centrality scores for the selected neologisms cover a broad spectrum of degrees of diffusion, as can be seen in Table 3a. Figure 4 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.[13]

The network graphs in Figure 4 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 network graphs are thus due to differences in the underlying social structure of communities rather than a mere function of differences in network size.

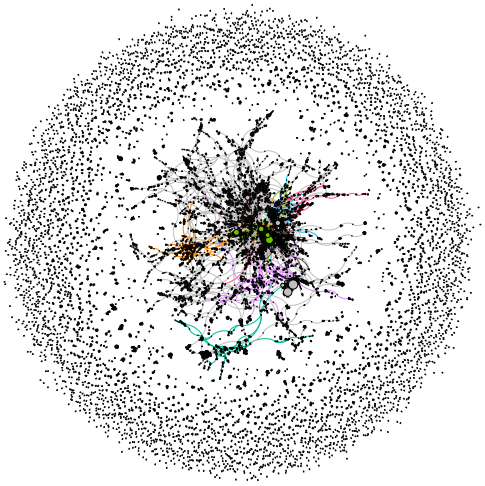
The neologism *upskill* exhibits the highest degree of diffusion, which is reflected by the highest degree of dispersion of nodes across the graph in Figure 4a. At the center of the graph, we find a relatively large cluster of speakers who are only loosely connected. Many of these speakers are connected 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 of its graph 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 4b 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 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.

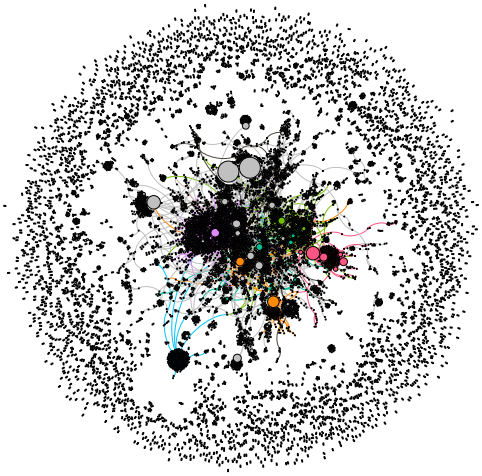
The social graph for *alt-left* shows very limited diffusion of the term. Almost all of its use can be traced back to one closely-connected community of users. This core community of users demonstrates typical characteristics of an echo chamber in that it is dense and features strong ties within the 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 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

[13]The network graphs for *alt-right* and *poppygate* were omitted as their difference in network size does not allow for comparative analyses (*alt-right*: 274 686 nodes, *poppygate*: 2 473 nodes).

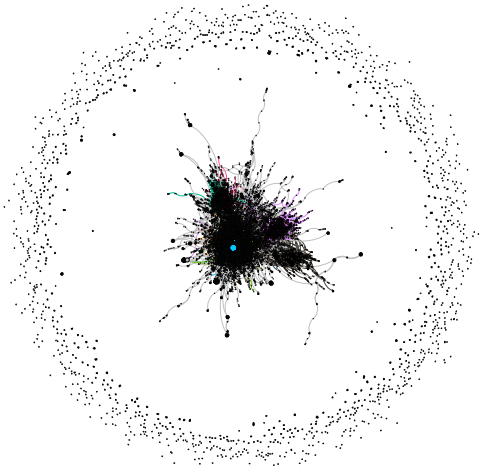
(a) ***upskill***
(37 044 nodes, 23 060 edges)



(b) ***hyperlocal***
(26 548 nodes, 16 576 edges)



(c) ***alt-left***
(26 367 nodes, 32 836 edges)



(d) ***solopreneur***
(24 585 nodes, 20 486 edges)

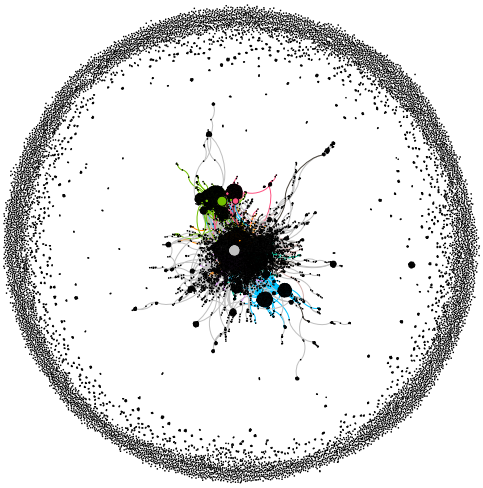


Figure 4: Social network graphs for the last subset of the selected neologisms.

neologisms[14]. 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 thus 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.

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 results obtained from the frequency-based analysis 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 development of social diffusion over time, Figure 5 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 3a. The corresponding network graphs for this stage were were presented in Figure 4. 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.

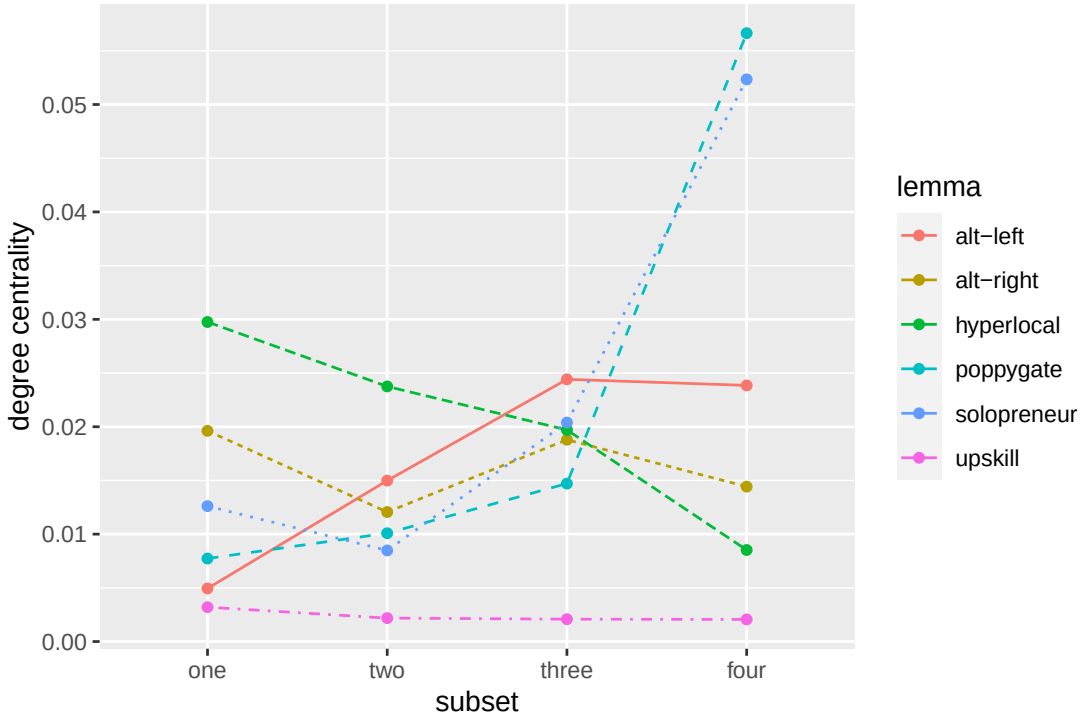
6 presents the full network graphs for all stages of diffusion for the term *hyperlocal* to illustrate the social dynamics underlying the quantitative measures.

Both the quantitative measure in Figure 5 and the network visualizations in Figure 6 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

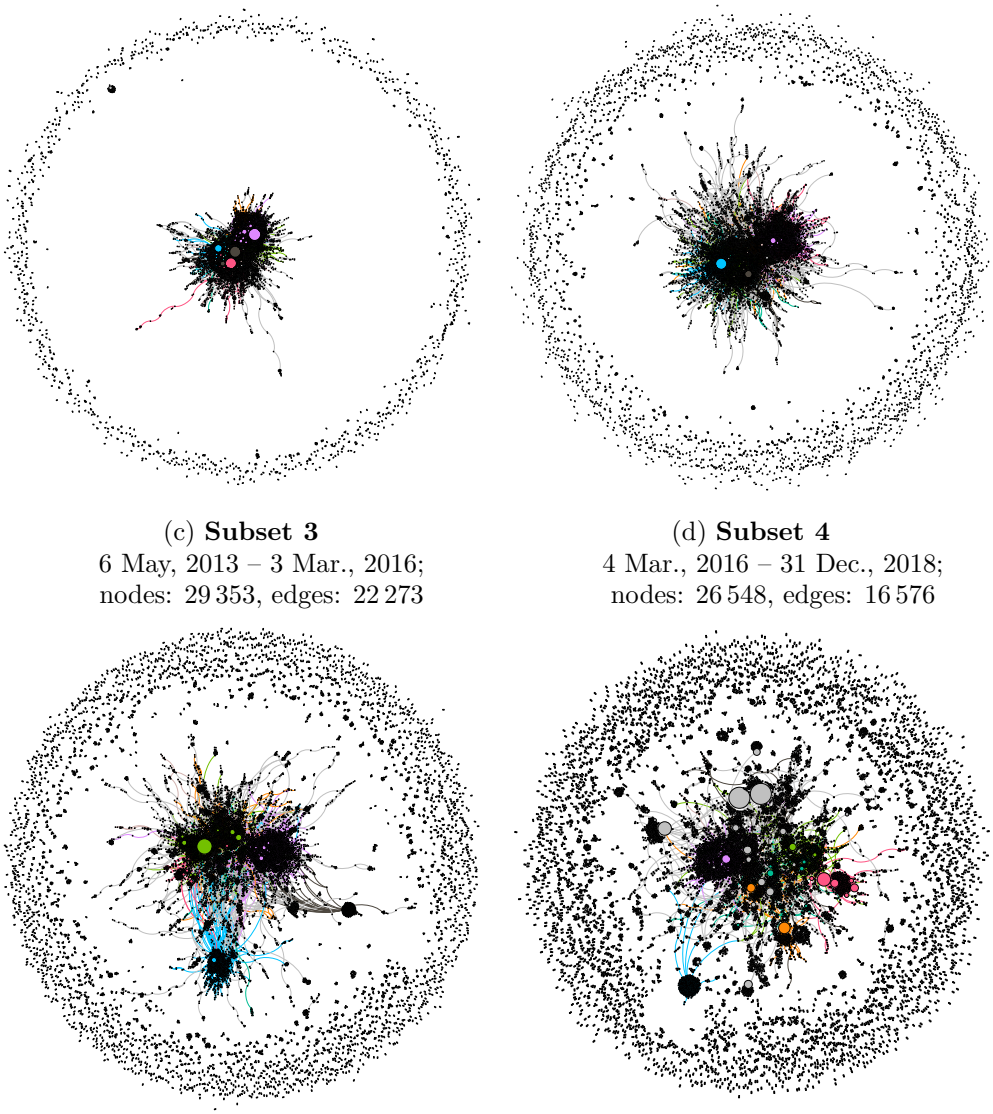
[14]The numbers of edges per node for all selected cases in descending order: *alt-right*: 1.49, *alt-left*: 1.24, *solopreneur*: 0.83, *hyperlocal*: 0.62, *upskill*: 0.62, *poppygate*: 0.53

Figure 5: Pathways of diffusion for the selected neologisms. The graph shows DEGREE CENTRALITY scores over time, each SUBSET representing one network graph which was generated for each of the four equally-sized time slices for each neologism in the sample.



(a) **Subset 1**
10 Sept, 2007 – 8 Jul 2010;
nodes: 12 569, edges: 13 742

(b) **Subset 2**
9 Jul., 2010 – 5 May, 2013;
nodes: 33 450, edges: 34 109



(c) **Subset 3**
6 May, 2013 – 3 Mar., 2016;
nodes: 29 353, edges: 22 273

(d) **Subset 4**
4 Mar., 2016 – 31 Dec., 2018;
nodes: 26 548, edges: 16 576

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

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. 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 use of 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 5.

The remaining cases in Figure 5 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 3c): 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 4d 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 or recurrent topical usage (Figure 3f). 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 community.

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 scores for both lexemes in Figure 5. 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 political orientations involved, it seems plausible that the majority of Twitter users do not want to be associated with this term nor its community of users.

In summary, studying the temporal dynamics of social networks highlights changes in the use of neologisms over time and reveals differenct 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 other. Due to the prevalence of using frequency as an indicator of diffusion in previous work, I will largely use frequency as a base of comparison for evaluating the social network approach.[15]

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

[15]It should be noted that a strict evaluation of both approaches is in principle impossible without external data about the true 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.

Table 4: 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).[16]

	ρ	P
FREQUENCY	-0.44	< 0.001
AGE	-0.29	0.004
VOLATILITY	0.28	< 0.001

with their age as observed in the corpus. Table 4 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 substantially 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 validity 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 centrality between and frequency

For a closer analysis of the interactions between these variables beyond correlation coefficients, Figure 7 presents all neologisms according to their usage frequency and centrality scores. While Figure 7a covers the full sample, Figure 7b 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 graph.

The general trend in the plot confirms the inverse relation captured by the negative

[16] All variables entering the correlation analysis were log-transformed and centred.

(a) Full sample. (b) Selected cases.

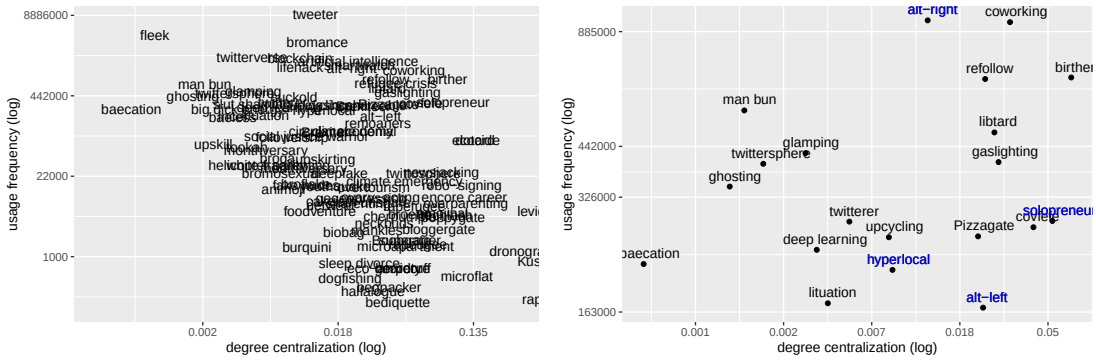


Figure 7: Relationship between total USAGE FREQUENCY and degree centrality (CENTRALIZATION) for the full sample of neologisms ($n = 99$) and the selected cases.

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 7a 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 examples, 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 seems to provide additional information for cases in which the social network structure suggests that the observed usage intensity overestimates the degree of diffusion of a target neologism since its observed uses go back to a disproportionately smaller number of speakers and subcommunities.

Analysing these deviations highlights two main groups of neologisms, for which total usage frequency and social network structure seem to diverge in systematic ways. 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 7a. 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.

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 7a 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 5 represent exceptions to this general trend. While this trend fits theoretical expectations and the empirical observations in the present dataset, these present results are only 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 (a) overall usage frequency, (b) changes in usage frequency over time (volatility), and (c) 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, Labov and Herzog 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 is to assess the validity 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 (Table 1) 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). Previous 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. However, previous work has not systematically analysed to what degree the semantic carrying capacity of neologisms is stable over time and across communities of speakers.

The present study focused on three main aspects of the temporal dynamics of diffusion: trend, age and volatility. Firstly, trends in usage frequency provide 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, total frequency counts, which are, in their pure form, 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 degrees of 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 the product of onomasiological competition between formal variants[17].

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 exhibit the Lindy Effect: the longer new words remain in use by the speech community, the less likely they are to become obsolete. 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

[17] As an exception, the sample contains the two formal variants *monthversary* and *monthiversary*.

isolated spikes in usage intensity tend to go back to disproportionally smaller parts of the speech community.

TREND + VOLATILITY ≠ S-curve

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 NN 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 7a) 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.

remove @HJS

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*).

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 out-

put 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 7b 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 4 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 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 it. 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.

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 across the full sample. On the one hand, measures which are based on the occurrence of neologisms in the corpus such as frequency, trend, 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 allows 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 will of course provide the most complete picture of diffusion. 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 suggests that the additional consideration of temporal dynamics of use and social network information can provide a much more detailed and accurate picture of diffusion.

6.3 Future work

While the present paper has attempted to assess the potential of frequency network-based approaches, a more extensive evaluation will have to be left for future work. A more rigorous attempt could include a systematic evaluation on the basis of external data sources like web corpora or online dictionaries. Moreover, questionnaires could be used to determine to what degree corpus-based measures generalise and converge with speaker judgements (s. Kerremans (2015) for an earlier attempt using a smaller sample of neologisms).

Furthermore, initial results in the present study suggested that social network information might be informative for predicting the success of lexical innovations. Using statistical models to test the predictive power of social network information, possibly in combination with temporal information, constitutes an attractive objective for future work, which might also provide insights about the determining factors of diffusion (Ryskina et al. 2020; Stewart and Eisenstein 2018).

Lastly, the present results demonstrated that neologisms show different degrees of social diffusion between speakers and communities. The broader definition by Schmid (2020) also includes diffusion across 'contexts' and 'cotexts', however. To study the former, it would seem desirable to investigate to what degrees neologisms diffuse across different usage contexts, e.g. different web registers (Biber and Egbert 2016). To study the latter, future work could continue to investigate to what degree the use of neologisms

differs between communities (Tredici et al. 2019; Schmid et al. 2020).

S-curve model: early adopters, influencers, density, weak ties

7 Conclusion

The present study has analysed the pathways and degrees of diffusion of neologisms on Twitter. Aside from an in-depth study of the spread of the neologisms in the sample, the aim of this paper was to apply and assess a range of frequency and network-based measures to study the diffusion of neologisms to new speakers and communities. It was shown that investigating the temporal and social dynamics underlying the use of neologisms is crucial for a more detailed and accurate assessment of their social diffusion in the speech community. 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, and 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 are promising to enable many new opportunities for sociolinguistic research in the near future.

References

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

Biber, Douglas and Jesse Egbert (2016). 'Register Variation on the Searchable Web'. In: *Journal of English Linguistics* 44.2, pp. 95–137. DOI: 10.1177/0075424216628955.

Bliss, Catherine A., Isabel M. Kloumann, Kameron Decker Harris, Christopher M. Danforth and Peter Sheridan Dodds (1st Sept. 2012). 'Twitter Reciprocal Reply Networks Exhibit Assortativity with Respect to Happiness'. In: *Advanced Computing Solutions for Health Care and Medicine* 3.5, pp. 388–397. DOI: 10.1016/j.jocs.2012.05.001.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre (9th Oct. 2008). 'Fast Unfolding of Communities in Large Networks'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008. DOI: 10.1088/1742-5468/2008/10/p10008.

Blythe, Richard A. and William Croft (2012). 'S-Curves and the Mechanisms of Propagation in Language Change.' In: *Language* 88.2, pp. 269–304.

Brin, Sergey and Lawrence Page (1998). 'The Anatomy of a Large-Scale Hypertextual Web Search Engine'. In: *Seventh International World-Wide Web Conference (WWW 1998)*. Seventh International World-Wide Web Conference (WWW 1998). Brisbane, Australia. URL: http://ilpubs.stanford.edu:8090/361/ (visited on 04/06/2020).

Bruns, Axel (1st Dec. 2012). 'How Long Is a Tweet? Mapping Dynamic Conversation Networks on Twitter Using Gawk and Gephi'. In: *Information, Communication & Society* 15.9, pp. 1323–1351. DOI: 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*. Valencia, Spain: Association for Computational Linguistics, pp. 95–98. URL: 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.* URL: https://www.english-corpora.org/now/.

Del Tredici, Marco and Raquel Fernández (15th June 2018). *The Road to Success: Assessing the Fate of Linguistic Innovations in Online Communities.* URL: http://arxiv.org/abs/1806.05838 (visited on 31/07/2020).

Dunbar, Robin IM (1992). 'Neocortex Size as a Constraint on Group Size in Primates'. In: *Journal of human evolution* 22.6, pp. 469–493.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith and Eric P. Xing (2014). 'Diffusion of Lexical Change in Social Media'. In: *PLOS ONE* 9.11, pp. 1–13. DOI: 10.1371/journal.pone.0113114.

Elsen, Hilke (2004). 'Neologismen'. In.

Fischer, Roswitha (1998). 'Lexical Change in Present Day English. A Corpus Based Study of the Motivation, Institutionalization, and Productivity of Creative Neologisms'. In.

Freeman, Linton C. (1st Jan. 1978). 'Centrality in Social Networks Conceptual Clarification'. In: *Social Networks* 1.3, pp. 215–239. DOI: 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'. In.

Gerlitz, Carolin and Bernhard Rieder (2013). 'Mining One Percent of Twitter: Collections, Baselines, Sampling'. In: *M/C Journal* 16.2. URL: 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 and Jacob Eisenstein (2016). 'The Social Dynamics of Language Change in Online Networks'. In: *Social Informatics*. Ed. by Emma Spiro and Yong-Yeol Ahn. Cham: Springer International Publishing, pp. 41–57.

Granovetter, Mark S. (1977). 'The Strength of Weak Ties'. In: *Social Networks*. Ed. by Samuel Leinhardt. Academic Press, pp. 347–367. DOI: 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 and Diansheng Guo (2019). 'Mapping Lexical Dialect Variation in British English Using Twitter'. In: *Frontiers in Artificial Intelligence* 2, p. 11. URL: https://www.frontiersin.org/article/10.3389/frai.2019.00011.

Grieve, Jack, Andrea Nini and Diansheng Guo (2016). 'Analyzing Lexical Emergence in Modern American English Online'. In: *English Language and Linguistics* 21, pp. 99–127.

– (2018). 'Mapping Lexical Innovation on American Social Media'. In: *Journal of English Linguistics*.

Hébert-Dufresne, Laurent, Samuel V. Scarpino and Jean-Gabriel Young (2020). 'Macroscopic Patterns of Interacting Contagions Are Indistinguishable from Social Reinforcement'. In: *Nature Physics*. DOI: 10.1038/s41567-020-0791-2.

Hohenhaus, Peter (1996). 'Ad-Hoc-Wortbildung. Terminologie, Typologie Und Theorie Kreativer Wortbildung Im Englischen'. In.

– (2006). 'Bouncebackability. A Web-as-Corpus-Based Study of a New Formation, Its Interpretation, Generalization/Spread and Subsequent Decline'. In: *SKASE Journal of Theoretical Linguistics* 3, pp. 17–27.

Huberman, Bernardo A., Daniel M. Romero and Fang Wu (4th Dec. 2008). *Social Networks That Matter: Twitter under the Microscope*. URL: http://arxiv.org/abs/0812.1045 (visited on 21/06/2020).

Jacomy, Mathieu, Tommaso Venturini, Sebastien Heymann and Mathieu Bastian (10th June 2014). 'ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software'. In: *PLOS ONE* 9.6, e98679. DOI: 10.1371/journal.pone.0098679.

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

Kerremans, Daphné, Jelena Prokić, Quirin Würschinger and Hans-Jörg Schmid (2019). 'Using Data-Mining to Identify and Study Patterns in Lexical Innovation on the Web: The NeoCrawler'. In: *Pragmatics and Cognition* 25.1, pp. 174–200.

Kerremans, Daphné, Susanne Stegmayr and Hans-Jörg Schmid (2012). 'The NeoCrawler: Identifying and Retrieving Neologisms from the Internet and Monitoring Ongoing Change'. In: *Current Methods in Historical Semantics*. Berlin: Mouton de Gruyter, pp. 59–96.

Labov, William (2007). 'Transmission and Diffusion'. In: *Language* 83.2, pp. 344–387.

Lemnitzer, Lothar (2018). *Wortwarte*. URL: http://www.wortwarte.de/ (visited on 14/01/2018).

Leuckert, Sven|Leuckert (2020). *Chapter 1. Towards a Digital Sociolinguistics*. scl.98.01leu. URL: https://benjamins.com/catalog/scl.98.01leu (visited on 08/12/2020).

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 and Mauricio Santillana (2018). 'Accurate Influenza Monitoring and Forecasting Using Novel Internet Data Streams: A Case Study in the Boston Metropolis'. In:

JMIR Public Health and Surveillance 4.1. Ed. by Gunther Eysenbach, e4. DOI: 10.2196/publichealth.8950.

Milroy, James (1992). *Linguistic Variation and Change: On the Historical Sociolinguistics of English*. Oxford: Blackwell.

Milroy, James and Lesley Milroy (1985). 'Linguistic Change, Social Network and Speaker Innovation'. In: *Journal of Linguistics* 21.2, pp. 339–384. URL: 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'. In: *Studies in Variation, Contacts and Change in English* 16. URL: http://www.helsinki.fi/varieng/series/volumes/16/nevalainen/.

Nini, Andrea, Carlo Corradini, Diansheng Guo and Jack Grieve (2017). 'The Application of Growth Curve Modeling for the Analysis of Diachronic Corpora'. In: *Language Dynamics and Change* 7.1, pp. 102–125.

Pew Research Center (23rd Oct. 2019). *National Politics on Twitter: Small Share of U.S. Adults Produce Majority of Tweets*. URL: 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. URL: https://www.R-project.org/.

Renouf, Antoinette, Andrew Kehoe and Jayeeta Banerjee (2006). 'WebCorp: An Integrated System for Web Text Search'. In: *Language and Computers* 59.1, pp. 47–67.

Rogers, Everett M. (1962). *Diffusion of Innovations*. New York: Free Press of Glencoe.

Ryskina, Maria, Ella Rabinovich, Taylor Berg-Kirkpatrick, David Mortensen and Yulia Tsvetkov (1st Jan. 2020). 'Where New Words Are Born: Distributional Semantic Analysis of Neologisms and Their Semantic Neighborhoods'. In: *Proceedings of the Society for Computation in Linguistics* 3.1, pp. 43–52. DOI: 10.7275/1jra-8m83.

Schmid, Hans-Jörg (2020). *The Dynamics of the Linguistic System. - Usage, Conventionalization, and Entrenchment*. Oxford: Oxford University Press.

Schmid, Hans-Jörg, Quirin Würschigner, 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 vols. De Gruyter Mouton.

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, p. 58. 45 pp. DOI: 10.1037/11491-005.

Stefanowitsch, Anatol and Susanne Flach (2017). 'The Corpus-Based Perspective on Entrenchment'. In: *Entrenchment and the Psychology of Language Learning: How We Reorganize and Adapt Linguistic Knowledge*. Ed. by Hans-Jörg Schmid. Boston, USA: American Psychology Association and de Gruyter Mouton, pp. 101–128.

Stewart, Ian and Jacob Eisenstein (31st Aug. 2018). *Making "Fetch" Happen: The Influence of Social and Linguistic Context on Nonstandard Word Growth and Decline*. URL: http://arxiv.org/abs/1709.00345 (visited on 02/08/2020).

Sunstein, Cass R. (3rd Apr. 2018). *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press. 333 pp.

Tredici, Marco Del, Diego Marcheggiani, Sabine Schulte im Walde and Raquel Fernández (Sept. 2019). 'You Shall Know a User by the Company It Keeps: Dynamic Representations for Social Media Users in NLP'. In: URL: https://arxiv.org/pdf/1909.00412.pdf.

Weinreich, Uriel, William Labov and Marvin Herzog (1968). 'Empirical Foundations for a Theory of Language Change'. In: *Directions for Historical Linguistics*. Ed. by Winfried P. Lehmann and Yakov Malkiel. Austin: University of Texas Press Austin, pp. 95–188.

West, Robert, Hristo S. Paskov, Jure Leskovec and Christopher Potts (2014). 'Exploiting Social Network Structure for Person-to-Person Sentiment Analysis'. In: *CoRR* abs/1409.2450. URL: 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 and Hiroaki Yutani (2019). 'Welcome to the Tidyverse'. In: *Journal of Open Source Software* 4.43, p. 1686. DOI: 10.21105/joss.01686.

Würschinger, Quirin, Mohammad Fazleh Elahi, Desislava Zhekova and 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*. Berlin: Association for Computational Linguistics, pp. 35–43. DOI: 10.18653/v1/W16-2605.