English Language and Linguistics 21.1: 99–127. © Cambridge University Press 2016
doi:10.1017/S1360674316000113

Analyzing lexical emergence in Modern American English online[1]

JACK GRIEVE and ANDREA NINI
Aston University

DIANSHENG GUO
University of South Carolina

(Received 28 October 2015; revised 10 March 2016)

This article introduces a quantitative method for identifying newly emerging word forms in large time-stamped corpora of natural language and then describes an analysis of lexical emergence in American social media using this method, based on a multi-billion-word corpus of Tweets collected between October 2013 and November 2014. In total 29 emerging word forms, which represent various semantic classes, grammatical parts-of-speech and word formation processes, were identified through this analysis. These 29 forms are then examined from various perspectives in order to begin to better understand the process of lexical emergence.

1 Introduction

A distinction has traditionally been made between two types of lexical variation (Geeraerts *et al.* 1994; Geeraerts 2010). *Semasiological variation* refers to variation in the meanings of words, such as variation in the meaning of the word *cell*, which denotes various concepts, including the basic structural unit of life, a small room and a portable telephone. *Onomasiological variation* refers to variation in the way that concepts are named, such as variation in the use of the words that refer to a portable telephone, with *cell* predominating in American English and *mobile* predominating in British English. Just as a distinction can be made between these two types of lexical variation, a distinction can be made between two types of lexical change (Grondelaers *et al.* 2007). *Semasiological change* involves change in the meanings of words, while *onomasiological change* involves change in the way that concepts are named, including the formation of new words. Although generally presented in opposition to each other, semasiological change can be seen as a type of onomasiological change – a process through which new pairings of forms and meanings are created by modifying the meanings of existing words (Geeraerts 2010).

Traditional research on semasiological change (e.g. Reisig 1839; Darmester 1887; Bréal 1897) catalogued processes that affect the meanings of existing words (see Geeraerts 2010), including *generalization* (e.g. *blog* expanding in meaning

[1] The research reported in this article was funded by AHRC, ESRC and Jisc (grant reference number 3154) as part of Round 3 of the Digging into Data Challenge. We would also like to thank Alice Kasakoff, Bernd Kortmann, Emily Waibel and two anonymous reviewers for their comments on this article.

from a website consisting of a collection of personal posts to include larger professional websites), *specialization* (e.g. *hacker* narrowing in meaning from a talented programmer to a talented programmer who accesses unauthorized computers) and *metaphor* (e.g. the word *virus* being used to refer to computational as opposed to biological infectious agents). Alternatively, traditional research on onomasiological change (e.g. Marchand 1969; Bauer 1983; Cannon 1987) catalogued word formation processes (i.e. *lexicogenesis*) (see Miller 2014), including *compounding* (e.g. *weblog* from *web log*), *truncation* (e.g. *blog* from *weblog*), *blending* (e.g. *vlog* from *video blog*), *morphological derivation* (e.g. *blogger* from *blog*) and *borrowing* (e.g. the use of *blog* in languages other than English).

Although new words can be introduced to name new concepts, new words are often introduced to name concepts for which other words already exist. Consequently, onomasiological change not only involves lexicogenesis but also the competition between synonyms over time. This type of change has recently been the subject of considerable research (e.g. Sweetser 1991; Kleparski 2000; Geeraerts *et al.* 2012; Zhang *et al.* 2015). For example, Geeraerts *et al.* (2012) investigated how the word *anger* became the dominant way to express its current meaning as opposed to *ire* and *wrath*, based on an analysis of a corpus of Middle and Early Modern English. Research in sociolinguistics has also analyzed onomasiological change, such as in studies of quotatives. For example, Tagliamonte & D'Arcy (2004) found a clear rise in the use of quotative *be like* over both real and apparent time (i.e. across age groups). Sociolinguistic research, however, generally focuses on variation in phonology and grammar as opposed to lexis. Onomasiological variation is more commonly analyzed in regional dialectology (e.g. Kurath 1949), but the diachronic dimension is usually absent in these studies.

Research has also considered how new words enter the standard vocabulary of a language (Bauer 1983; Lipka *et al.* 1994; Fischer 1998; Hohenhaus 2005). New words are constantly being formed during everyday language use, but only a small percentage of these *neologisms* will ever find their way into dictionaries and the written standard, usually after a considerable time has passed. This process is generally referred to as *institutionalization* (Bauer 1983; Brinton & Traugott 2005) and has also been the subject of empirical research (e.g. Aitchison & Lewis 1995; Fischer 1998). Most notably, Fischer (1998) tracked change in the usage of a variety of relatively new words (e.g. *GUI*, *sitcom*, *cyborg*) in a corpus of *Guardian* newspaper writing from 1990 to 1996, consisting of approximately 25 million words per year. Methods for identifying neologisms online have also been developed, including the *Neocrawler* program (Kerremans *et al.* 2011), which discovers new words by searching the web.

Lexical change has also been examined from the complementary perspectives of *lexicalization* (Brinton & Traugott 2005) and *grammaticalization* (Hopper & Traugott 2003). Lexicalization is the process through which words or multi-word units whose meanings cannot fully be derived from the meanings of their constituents gradually change in form and meaning over time as they become distinct lexical items (Brinton & Traugott 2005). Lexicalization therefore includes certain word formation processes

such as compounding (e.g. *hard drive*) as well as the codification of set phrases (e.g. *boot up*). For example, Méndez-Naya (2006) analyzed the lexicalization of *downright* based on the records of the *Oxford English Dictionary*. Research on lexicalization overlaps with research on onomasiological change, specifically regarding word formation processes, although lexicalization generally focuses on certain types of word formation processes and on longer-term patterns of lexical change. Lexicalization is also often contrasted with institutionalization. For example, Bauer (1983) saw lexicalization as a process that followed institutionalization. Alternatively, grammaticalizaton is the process through which lexical items lose referential meaning as they gradually develop into function words that express grammatical information (Hopper & Traugott 2003). For example, Krug (2000) tracked the transformation of various main verbs (e.g. *want*) into semi-modals (e.g. *wanna*) in a variety of historical corpora. Research on grammaticalization therefore overlaps with research on semasiological change, although it focuses specifically on the development of grammatical meaning.

Quantitative research on lexical change has also analyzed how the relative frequencies of words and multi-word units have risen and fallen over time (e.g. Krug 2000; Nevalainen & Raumolin-Brunberg 2003; Gries & Hilpert 2010; Geeraerts *et al.* 2012; Siemund 2014). In research on grammaticalization and semasiological change, analyzing the relative frequencies of words is primarily of interest because it can help identify, describe and explain changes in word meaning. For example, by plotting the frequency of various semi-modals since the seventeenth century, Krug (2000) showed that there was an exponential rise in their usage as they became grammaticalized. In research on institutionalization, lexicalization and onomasiological change, analyzing the relative frequencies of words is primarily of interest because it allows for the rise of new lexical items and competition between synonyms to be tracked over time. For example, Nevalainen & Raumolin-Brunberg (2003) charted the frequency of subject *you* relative to *ye* from the fifteenth to seventeenth century and found the rise of *you* followed a clear *s-shaped curve*, with its frequency rising gradually at first, then rapidly, then gradually once again, as the change neared completion. Because this type of research has produced graphs that allow for quantitative change in word usage to be visualized, comparison of different patterns of change is also possible. Most notably, s-shaped curves of language change have been repeatedly identified in linguistic research across linguistic levels, including both in the relative frequencies of individual linguistic forms (e.g. the occurrence of some form per million words) and in the frequencies of one form measured relative to the frequencies of other equivalent forms.

While there has been a long tradition of empirical research on lexical change from a variety of different perspectives, almost all of this research has focused either on word formation processes or on how the meanings and usages of words have changed over relatively long periods of time. Even research on institutionalization has focused on long-term patterns of change, as new words enter into standard usage. Very little is known about how the usage of new word forms changes following their introduction, as they spread across a population of speakers for the first time. This type of *lexical emergence* has been so difficult to analyze because linguists have not had access to

sufficient amounts of language data with the necessary temporal resolution to track the spread of emerging word forms. In general, lexical variation is difficult to study because most words are very rare. For example, out of the top 100,000 most frequent word forms in the 450 million-word *Corpus of Contemporary American English* (Davies 2010), the 50,000th form, *sympathizes*, occurs only 124 times – once every 3.6 million words. The majority of words in the English language therefore occur on average less than once per million words, requiring very large corpora for their analysis. However, the analysis of lexical emergence, which involves forms that are especially rare but that can also rise quickly in frequency over relatively short periods of time, requires access to incredibly large corpora that are very densely sampled over time.

Compiling corpora that meet these requirements has recently become possible by mining language data from the internet – an approach that is referred to as *web as corpus* (see Kilgarriff 2001). More specifically, over the last five years, linguists have begun to analyze very large corpora drawn from social media websites, especially Twitter, which makes its data easy to obtain for academic research using their internal API. This research has included numerous studies on language variation and change (e.g. Eisenstein *et al.* 2010, 2014; O'Connor *et al.* 2010; Hadican & Johnson 2012; Bamman *et al.* 2014; Huang *et al.* 2016). For example, based on a geo-coded corpus of Twitter data, Eisenstein *et al.* (2014) found that demographic similarities between cities is an especially important factor for explaining the spread of lexical change in addition to their geographical proximity. The multi-billion-word *Google Books Corpus*, which spans approximately 200 years of fiction writing, has also been used recently for research on lexical change (e.g. Petersen *et al.* 2012a, 2012b).

Building on this research, this article presents a quantitative corpus-based analysis of lexical emergence – the process through which new word forms spread across a population of speakers. The study has three primary goals. First, it introduces a method for identifying instances of lexical emergence in large time-stamped corpora. Second, it describes an application of this method to identify emerging word forms in Modern American English based on an analysis of an 8.9 billion-word corpus of American Twitter data collected between October 2013 and November 2014. Finally, it explores the set of emerging forms identified through this analysis from a variety of perspectives in order to better understand the process of lexical emergence.

2 Data

This study analyzes lexical emergence in Modern American English based on a multi-billion-word corpus of American Twitter data. Twitter (www.twitter.com) was primarily selected for analysis because it provides very large amounts of time-stamped data over a short period of time. Twitter is also often a very informal variety of natural language that millions of people from across the United States participate in, including younger speakers and speakers from lower socio-economic classes, who presumably are often responsible for the introduction of new words, especially slang. Analyzing Twitter data should therefore allow for new word forms to be identified at a relatively early stage of

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

their development and for their usage to be tracked over time. Of course, any patterns of lexical spread identified in Twitter are not necessarily representative of other registers of American English. This is true, however, of any register, and at this point in time Twitter is one of only sources of natural language data that is suitable for the analysis of lexical emergence.

The corpus analyzed in this study consists of 8.9 billion words of geocoded and time-stamped American Twitter data, totaling 980 million tweets written by 7 million unique users, collected between 11 October 2013 and 22 November 2014 at the University of South Carolina using the Twitter API (http://dev.twitter.com). The Twitter API allows for Tweets to be downloaded soon after they are posted, as well as a variety of metadata to be obtained, including the username of the poster, a time-stamp, language information about the Tweet and geocoding information, where available, in the form of the longitude and latitude of the user when posting that message. Geocoded tweets in particular are generated when users post on mobile devices (with the geotracking option activated). The corpus was compiled by extracting all geocoded English language Tweets (as identified by Twitter) from the Twitter API between October 2013 and November 2014. All Tweets were then sorted by county. Tweets were excluded from the corpus if they did not occur within a county in the contiguous United States. The corpus only contains geocoded Tweets because it was primarily compiled to analyze geolinguistic variation (e.g. see Huang *et al.* 2016); although no regional results are reported in this article, focusing on geocoded data guarantees that all the Tweets in the corpus have come from the United States, ensuring that the results of this study have a clear regional scope.

To analyze temporal patterns of lexical emergence, the corpus was divided into 397 daily subcorpora. On average, these daily subcorpora contain 22 million word tokens per day, but the size of the daily subcorpora ranges from 10 to 29 million word tokens. Although the period from 11 October 2013 to 22 November 2014 includes 409 days, the corpus only includes 397 daily subcorpora due to power failures and other technical difficulties, which interrupted the harvesting of Tweets. However, given that this is a relatively small percentage of missing data, these missing days are spread across the timeline, and robust statistics are used for analysis (see below), it is assumed that the missing data will have no substantial effect on the analysis of this corpus. Note that re-Tweets and quotations were not removed from the corpus, primarily because this should be seen as evidence of the spread of a new word form.

3 Analysis

To identify emerging word forms in the Twitter corpus, all 67,022 word forms (defined as a string of alphabetical characters plus hyphens, insensitive to case) that occur at least 1,000 times in the complete 8.9 billion-word corpus were extracted for analysis. No multi-word units were analyzed. In addition, no lemmatization was conducted (e.g. *computer* and *computers* were analyzed as distinct word forms) and alternative spellings were analyzed separately. This is because the goal of this analysis is to identify newly

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

emerging word forms. Related word forms can be lemmatized or combined at a later stage of the analysis if the analyst so chooses, but this is not a necessary step. Indeed, alternative forms, including variant spellings, can often have different meanings or social distributions, for example, which would be lost if the forms were automatically lemmatized.[2] Finally, no attempt at word sense disambiguation was made (e.g. *cell* as a small room and *cell* as a mobile phone were not analyzed separately).

The relative frequencies of each of these 67,022 forms were then measured over each of the 397 days in the corpus by dividing the frequency of that form in all posts from that day by the total number of word tokens in those posts and multiplying this value by 1 billion in order to obtain a normalized frequency count per billion words (PBW). Normalizing the frequency of each form by day makes it possible to compare the frequency of words across the daily subcorpora, even though the number of words per day is inconsistent. Frequencies were normalized PBW to allow for results to be expressed in whole numbers, as this analysis is focusing on very rare forms; normalizing by PBW has no effect on the results of the analysis. This procedure yielded a 67,022-word-form-by-397-day temporal data matrix, representing daily change over time in the relative frequency of each of these 67,022 forms from October 2013 to November 2014.

To identify emerging word forms based on this temporal data matrix two values were computed for each form: the relative frequency of that form at the start of the period of time represented by the corpus and the degree to which the relative frequency of that form had risen over the course of this period.

To measure the relative frequency of each of the forms at the start of the period under analysis, the average relative frequency per day of each form was measured from 11 October 2013 to 31 December 2013. The selection of 31 December 2013 as the end date of the initial period is of relatively little consequence, at least in this case, as the corpus only spans a little over a year. Had relative frequency at the start of the time period been measured using a somewhat different end date, the results would have been largely the same. However, given that the period covered by the corpus includes most of 2014 in addition to the last few months of 2013, the end of 2013 was deemed to be a natural cut off for this corpus: i.e. it allows for forms to be identified that were very uncommon in 2013 but that showed substantial increases over the course of 2014.[3]

To measure the degree to which each of the 67,022 forms shows a consistent rise in frequency over the course of the time period, a Spearman rank correlation coefficient was calculated for each form by correlating the (rank) relative frequency of that form per day (i.e. in each daily subcorpus) to the rank of that day in the time series (i.e. 11 October 2013 has a rank of 1 and 22 November 2014 has a rank of 397). A positive correlation between daily relative frequency and day of the period indicates

[2] Grouping related forms would, however, help to allow for lower-frequency emerging word forms to be identified, although this is a difficult task, especially when dealing with non-standard forms.

[3] If the method were used to identify emerging words in a corpus with more chronological depth, the analysis could be repeated using different initial periods and even different end dates to identify a larger number of emerging words.

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

that the usage of that particular form has increased over time. Alternatively, a negative correlation indicates that the usage of that form has decreased over time, whereas a correlation approaching zero indicates that the usage of that form does not exhibit an overall pattern of rise or decline. A Spearman correlation coefficient was used because it identifies monotonic patterns, where rise in value of one variable increases with the values of another variable, regardless of the shape of the rise. Alternatively, a Pearson correlation coefficient would only have allowed for linear patterns to be identified accurately, where the rise in the relative frequency of a form progresses at a stable rate. A Spearman correlation coefficient therefore allows for a much wider range of rising patterns to be identified.[4]

Two measurements were therefore computed for each of the 67,022 word forms in the data matrix: a relative frequency representing the commonness of the form at the start of the period of time and a Spearman correlation coefficient representing the degree to which that form has risen in frequency over the course of that period of time. By then identifying forms with low relative frequencies at the start of that period of time and relatively high positive correlations over the course of that period of time, a list of potential emerging word forms was generated. In particular, all 131 forms with an average relative frequency of less than 1,000 PBW (i.e. less than once per million words) at the end of 2013 and with a Spearman correlation coefficient of larger than .80 (i.e. a strong correlation coefficient) were extracted for further analysis. It should be noted that different settings could have been used, which would have resulted in a larger or smaller set of emerging words being extracted. Emerging word forms, however, is not a definitive category and as such it is impossible to set these values in a definitive way. For the purpose of this study, these settings were found to identify a sufficient number of forms to demonstrate the application of the method and to allow for the analysis of lexical emergence in the corpus.

The relationship between the complete set of 67,022 forms and the set of 131 potential emerging forms in terms of these two measurements is visualized in figure 1 using a series of graphs. In each of these graphs, the *x*-axis plots the Spearman correlation coefficients of each word, with rising forms on the right and falling forms on the left, while the *y*-axis plots the relative frequency PBW of each form at the end of 2013, with more frequent forms at the top and less frequent forms at the bottom. As discussed by Zipf (1935, 1949), the most frequent forms in this corpus account for most of the word tokens and therefore most forms occur very infrequently. Only the most frequent forms are therefore visible in the graph in the top-right corner of figure 1, which plots all 67,022 forms; the vast majority of these forms are clustered together at the bottom of the graph, including the emerging word forms, which are found in the bottom right-hand corner of the cloud (i.e. forms that are both infrequent at the start of the period and that rise substantially over the course of the period). To visualize the entire cloud as well as the small part of the cloud where emerging forms are found, figure 1 therefore

[4] Using Kendall's tau (see Hilpert & Gries 2009) yields very similar results.

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

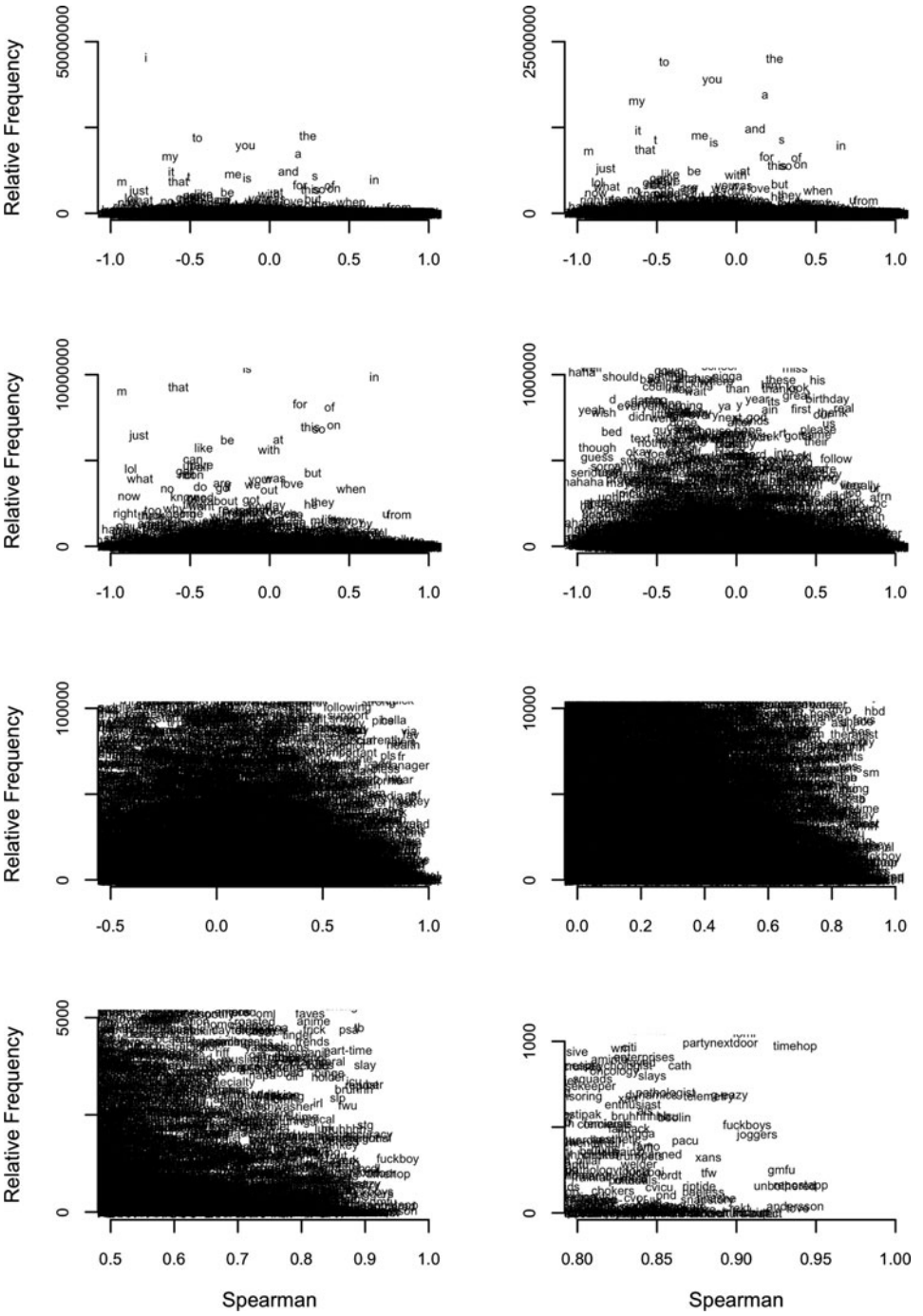

Figure 1. Spearman's coefficient vs. 2013 relative frequency (67,022 forms)

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

presents a series of graphs that gradually zoom in on the bottom right-hand corner of the cloud.

After generating this list of 131 potential emerging forms, concordance lines drawn from the corpus that exemplify the usage of each of these form were inspected by hand. A number of issues with these 131 forms were identified. First, a number of these forms were found to be proper nouns, including the names of people (e.g. *tove*, *partynextdoor*), products (e.g. *repostapp*, *timehop*) and companies (e.g. *aurstaff*, *marinemax*). Because the frequency of proper nouns over time primarily depends on the amount of discussion online about a specific person or thing, change in the relative frequency of proper nouns does not primarily reflect linguistic constraints on how new word forms emerge over time. All proper nouns were therefore manually excluded from the list. Second, the list contained an inordinately large number of forms relating to the healthcare sector (e.g. *telemetry*, *PACU*). Further analysis of the concordance lines revealed that the rise in the frequency of these forms was almost entirely due to a rise in the practice of posting geocoded healthcare job advertisements on Twitter over the course of 2014 by a relatively small but prolific group of commercial job agencies. In addition, increases in job advertisements were also found to be responsible for the rise of a small number of other words (e.g. *concierge*, *housekeeper*). Because the goal of this study is to analyze lexical emergence, rather than how Twitter advertisement strategies have changed in 2014, these forms were also excluded from further analysis. Finally, after removing these two sets of words, nine established words (i.e. words that are included in standard dictionaries) remained on the list. Because the goal of the analysis is to find recent word forms that are currently emerging in American English, these nine established word forms were also excluded from the list, although they are returned to at the end of this article.

Following this procedure, 29 emerging word forms were identified. These forms are plotted in figure 2, which corresponds to the final graph in figure 1 except that the 102 forms that fall within this range but that were removed are not plotted. These 29 forms are also listed in table 1, along with their 2013 relative frequency, their Spearman correlation coefficient and a working definition.

5 Discussion

In this section, the 29 emerging word forms identified through the analysis of the Twitter corpus are inspected from various perspectives, including their meanings and grammatical status, the word formation processes through which they were generated, their recentness, their patterns of growth over time, and their participation in both onomasiological competition and semasiological change. In addition, the status of the nine established forms that were identified during the analysis is discussed.

5.1 Word meanings and word classes

The 29 forms express a range of different meanings; however, all 29 can be characterized as *slang*, in the sense that they are highly informal forms that have more well-established

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

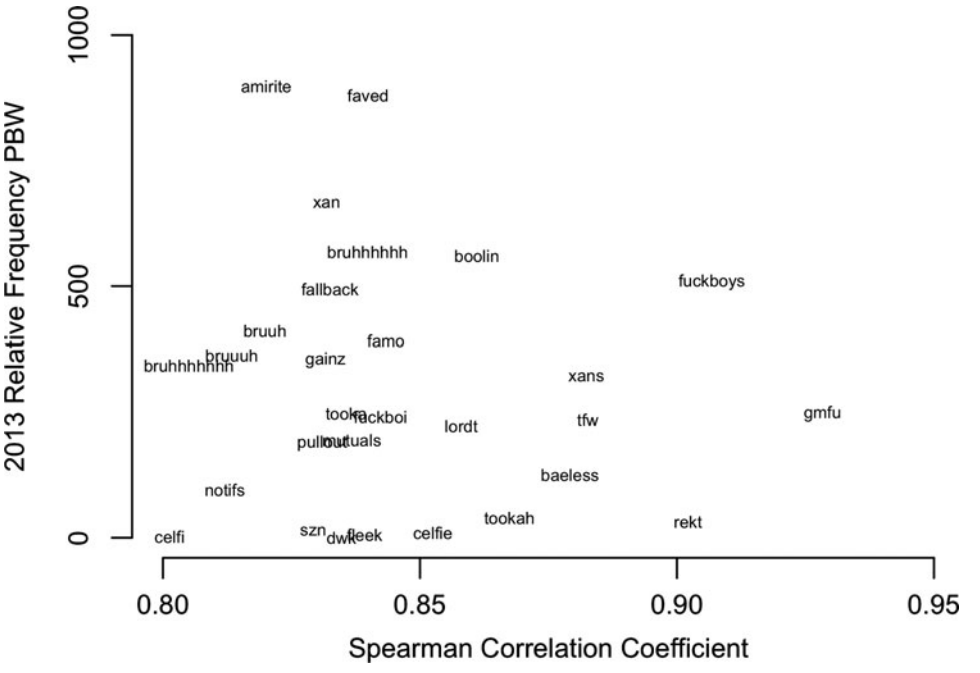

Figure 2. Spearman's coefficient vs. 2013 relative frequency (29 forms)

synonyms (Green 2011). Although these forms are not predominantly drawn from any one subject area, a number come from specific semantic domains, including profanity and insult (e.g. *gmfu*, *fuckboys*), recreational drug use (e.g. *xans*, *tookah*), social media (e.g. *faved*, *notifs*), and family and friends (e.g. *famo*, *boolin*).

A majority of these 29 forms are used primarily as nouns (e.g. *fuckboys*, *tooka*), although there are also forms that are used primarily as adjectives (e.g. *baeless*), verbs (e.g. *boolin*) and interjections (*lordt*). This distribution of word classes is not surprising. It is to be expected that emerging forms would be part of open word classes, which commonly accept new words, especially nouns, as opposed to closed word classes. The list, however, lacks adverbs, although they are an open class in the English language. In addition, several forms represent multi-word sequences, including acronyms (e.g. *gmfu*), blends of more than two words (e.g. *amirite*), and individual forms that occur as part of multi-word sequences (e.g. (*on*) *fleek*).

5.2 Word formation processes

Most of the 29 forms appear to have been created through standard word formation processes. Truncation is most common. Examples include *xan*(*s*) from *Xanax*, *famo* from *family*, *faved* from *favorited* and *notifs* from *notifications.* The formation of *mutuals* from *mutual friends* is also a form of truncation, although it also involves the

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

Table 1. *Emerging word forms*

Word form	Spearman correlation coefficient	2013 frequency PBW	Meaning
gmfu	0.928	247	get/got me fucked up
fuckboys[a]	0.907	508	insult for men (e.g. asshole)
rekt	0.902	32	wrecked
tfw	0.883	235	that feel when
xans[b]	0.882	320	benzodiazepine pills
baeless	0.879	125	single
tookah	0.867	39	marijuana
boolin	0.861	560	cooling (i.e. relaxing)
lordt	0.858	223	lord
celfie	0.853	10	photograph of oneself
famo	0.843	392	family and friends
fuckboi	0.842	241	insult for men (e.g. asshole)
faved	0.840	880	favorited
bruhhhhhh	0.840	568	bro
(on) fleek	0.839	6	good, on point
mutuals	0.837	194	mutual friends
tooka	0.836	247	marijuana
dwk	0.835	0	driving while kissing
fallback (game)	0.833	494	ability to escape difficult conversations
xan	0.832	665	benzodiazepine pill
gainz	0.832	354	weight gains from exercise
pullout (game)	0.831	189	*coitus interruptus*
szn	0.829	13	season
amirite	0.820	898	am I right?
bruuh	0.820	412	bro
bruuuh	0.813	363	bro
notifs	0.812	96	notifications
bruhhhhhhh	0.805	343	bro
celfi	0.801	3	photograph of oneself

[a]The singular form *fuckboy* was slightly more frequent than 1000 PMW in the 2013 data.
[b]Although *xan(s)* appears to be a proper noun, it is primarily used to refer to benzodiazepine pills in general.

grammatical conversion of *mutual* from an adjective to a noun. A number of emerging forms were also generated through compounding, including *fuckboys*, *fuckboi*, *fallback (game)*, *pullout (game)* and *amirite*. Particularly notable is the productive use of *game* in compounds, which refers to the ability to extricate oneself from difficult situations.

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

There are also several examples of acronymization, including *gmfu* and *tfw*, which along with compounding can be seen as a form of lexicalization and which represent multi-word units that are generally present in other varieties of language, including the spoken vernacular. In addition, *baeless* was formed through the derivation of the slightly older term *bae*, which appears to be a truncation of the word *babe*, and *boolin* appears to have been formed through blending (*blood* + *coolin*), introduced by members of the Bloods street gang as an alternative to *coolin* to avoid uttering words that contain the letter *c*, which is associated with the rival Crips.

In addition, 10 of the 29 forms represent spelling variation, either of established forms (e.g. *rekt*, *gainz*) or of other emerging forms (e.g. *tooka*, *fucboi*). Spelling variation is not generally considered a standard word formation process, as it is not an option in spoken language. From an orthographic perspective, however, these are new linguistic forms. Furthermore, most of these spelling variations appear to either mark specific pronunciations, including *lordt* and the various forms related to *bruuh*, or a specific usages of a word, including *rekt* for *wrecked* as a past participle, and *gainz* for *gains* as a noun.

The forms *tooka(h)* and *fleek* have more singular origins. *Tooka(h)* was apparently formed through the conversion of a proper noun, specifically the name of a dead Chicago gang member who was killed or *smoked* by a rival gang, which led to these words becoming associated with marijuana. The etymology of (*on*) *fleek* is less clear (see Whitman 2015) and may represent a true word creation, although it may also be related to the word *flick*. It is clear, however, that the rise of the term in 2014 is largely attributable to one video that went viral online where the term is used by Peaches Monroe to describe well-groomed eyebrows.

5.3 Recency

The 29 forms were not necessarily used for the first time in 2013 or 2014. Although usage of these 29 forms rose dramatically over the course of 2014, they may have been in existence for a considerably longer period of time. The goal of the analysis, however, was not to identify words that were first formed over this time period (i.e. neologism detection), but to identify rare forms not listed in dictionaries that were spreading rapidly on Twitter in 2014. Presumably most words are not first used online and therefore attempting to date the formation of words through the analysis of Twitter data or any other variety of computer-mediated communication is not generally reliable, aside perhaps for Twitter-related terms and some acronyms. Nevertheless, it is still informative to consider just how new these emerging forms are.

Given the lack of sufficiently large and dense diachronic corpora of informal spoken language, it is probably impossible to trace the exact time or place where these forms were introduced. It does appear, however, that most are relatively recent formations. In order to test this assumption, each of the form was searched for on *Google Trends* (www.google.com/trends), which allows for search term frequency on Google to be tracked over time, and *Urban Dictionary* (www.urbandictionary.com), which is a popular and

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

Table 2. *Recency of emerging words*

Word form	First search (*Google Trends*)	Earliest entry *Urban Dictionary*	*Urban Dictionary* related forms
gmfu	2014-11	2009	
fuckboys	2014-10	NA	*fuckboy* (2004)
rekt	2013-09	2011	
tfw	2006-08	2011	
xans	2012-12	NA	*xan* (2008)
baeless	NA	2014	*bae* (2008)
tookah	2013-09	2013	
boolin	2013-09	2005	
lordt	2015-02	NA	
celfie	2013-08	2014	
famo	2007-03	2005	
fuckboi	2014-10	2008	
faved	2013-01	NA	*fave* (2004)
bruhhhhhh	NA	NA	*bruh* (2003)
(on) fleek	2014-07	2014	
mutuals	2004-06	NA	*mutual* (2015)
tooka	2009-12	2014	
dwk	2008-08	NA	
fallback (game)	2006-10	2003	*fallback game* (2014)
xan	2004-11	2008	
gainz	2014-04	2013	
pullout (game)	2004-10	2003	*pullout game* (2014)
szn	2013-12	2015	
amirite	2009-04	2003	
bruuh	2014-08	2014	
bruuuh	NA	NA	
notifs	NA	2010	
bruhhhhhhh	NA	NA	
celfi	2014-07	NA	

immense online dictionary consisting of user-contributed definitions for slang words and phrases. The earliest occurrence of each of the 29 forms on both of these websites (i.e. earliest search, earliest dictionary entry) was then recorded. In addition, for *Urban Dictionary*, it was possible to read the definitions provided in order to verify that the definition matched the meaning used in the Twitter corpus. In most cases, this was the only meaning listed, but some of the acronyms, for example, had multiple definitions. In these cases, the date of the first definition of the form with the same general meaning attested in the Twitter corpus was recorded. Such semantic control, however, was not possible to implement for *Google Trends*. These data are presented in table 2, as well as additional information about the chronology obtained through these analyses.

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

Although this approach to dating the introduction of new forms no doubt underestimates the true dates of introduction, it does allow for a reliable upper limit to be established, which in most cases is well after 2000. *Urban Dictionary* appears to be particularly useful for making such estimates, as its first definitions often pre-date first searches on *Google Trends*. Indeed, in cases where the first searches identified by *Google Trends* pre-date the first *Urban Dictionary* entry (e.g. *tfw*, *mutuals*), it seems likely that a form with a different meaning was being searched for on Google. Overall, it therefore appears that most of the emerging forms on the list were introduced after or around 2000, with *baeless*, *tookah*, *lordt*, *celfie*, *(on) fleek*, *tooka*, *gainz*, *szn* and *celfi* in particular potentially having been introduced as late as 2013 or 2014. One exception is *amirite*, for which additional searches on Google revealed some webpages from before 2000 containing uses of the form.[5] Regardless of the exact dates of word formation, however, it is clear that a number of these forms have been in existence for many years. An important descriptive result of this study is therefore that new forms are often characterized by very infrequent use for years until they eventually emerge and see relatively widespread usage.

5.4 Change in relative frequency over time

To visualize change over time in the relative frequencies of the 29 forms, a time chart was generated for each, plotting the relative frequency per day of that form over the course of the period represented by the corpus (October 2013 to November 2014), which is a direct visualization of the data analyzed by the correlation analysis (i.e. the correlation analysis measures the degree to which the use of each from shows a steady rise over time). The time charts for each of the 29 forms are presented in figures 3 to 5. Because the relative frequencies of all these forms are associated with strong positive correlation coefficients, these time charts will necessarily exhibit relatively consistent increases over time. Nevertheless, the fact that forms with such high correlation coefficients were identified is remarkable and so too therefore is the fact that these 29 time charts show such smooth and regular patterns of change.

These time charts also generally show non-linear patterns of change, in the sense that the relative frequencies of most forms do not rise at a consistent rate over time. Perhaps the only form that shows what could reasonably be characterized as a linear pattern is *amirite*, which also interestingly appears to be one of the older forms under analysis. The majority of these time charts show relatively simple *super-linear* patterns of growth, where the rate of change speeds up steadily over time. In many cases these super-linear patterns of growth continue until the end of the corpus, especially for those forms that are associated with the highest correlation coefficients (e.g. *gmfu*, *fuckboys*, *rekt*, *tfw*, *xans*, *tookah*, *boolin*). In other cases, these time charts resolve themselves along a continuum between two extremes: the frequencies gradually stabilize, forming

[5] Each of the 29 forms was also dated by making date-restricted searches on Google, but in general it was difficult to judge the true age of the websites being returned.

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

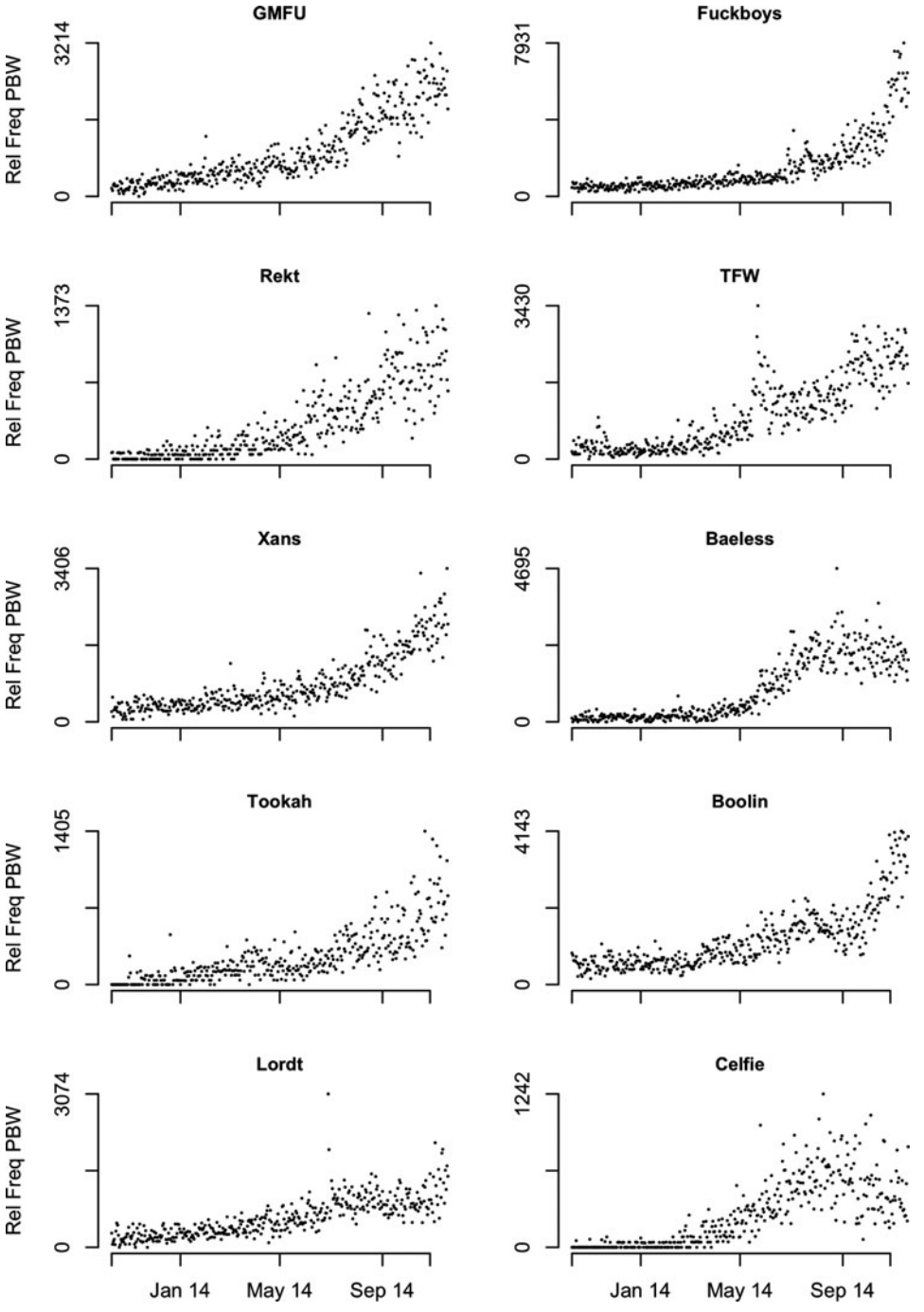

Figure 3. Emerging words time charts (part 1)

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

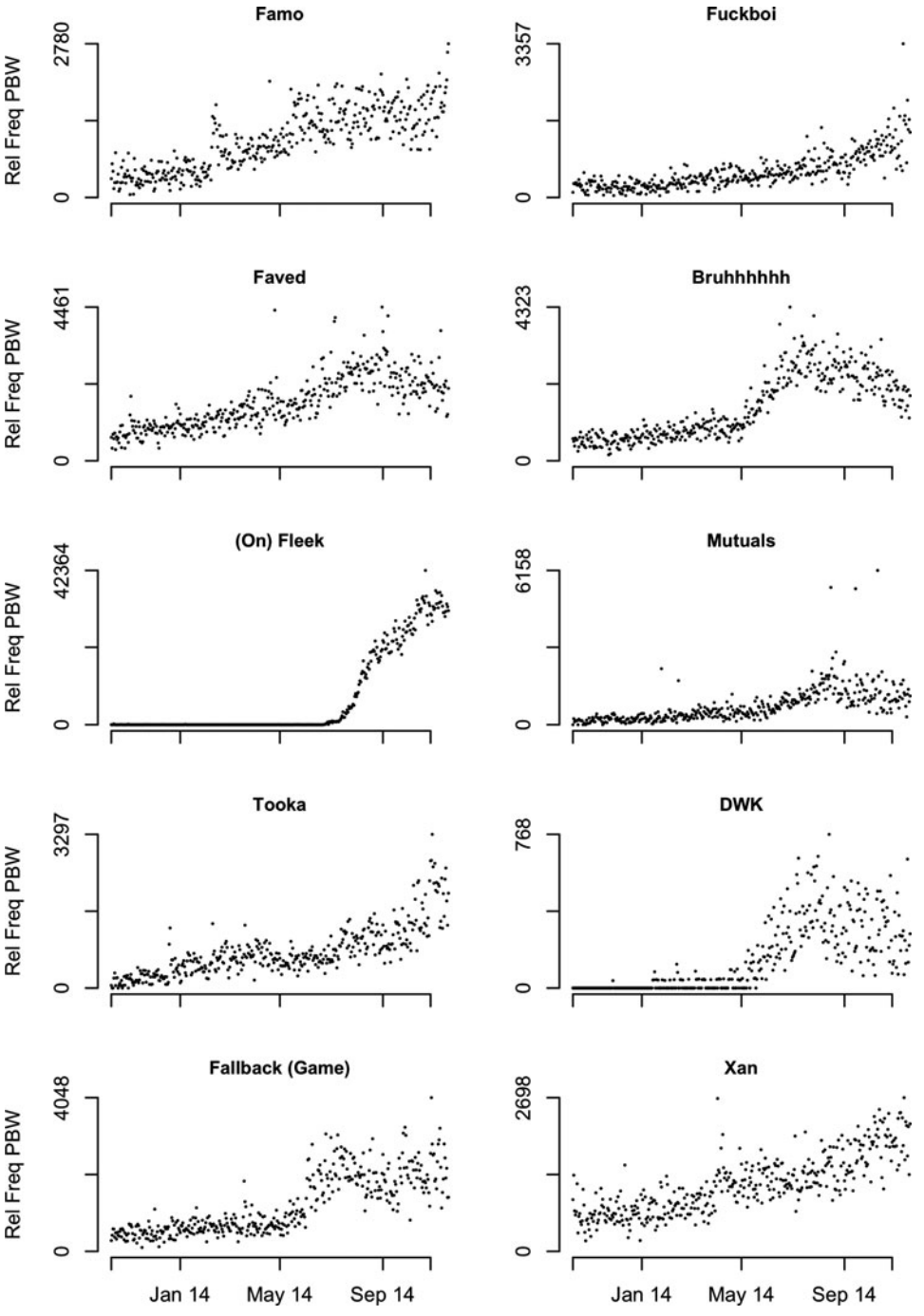

Figure 4. Emerging words time charts (part 2)

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

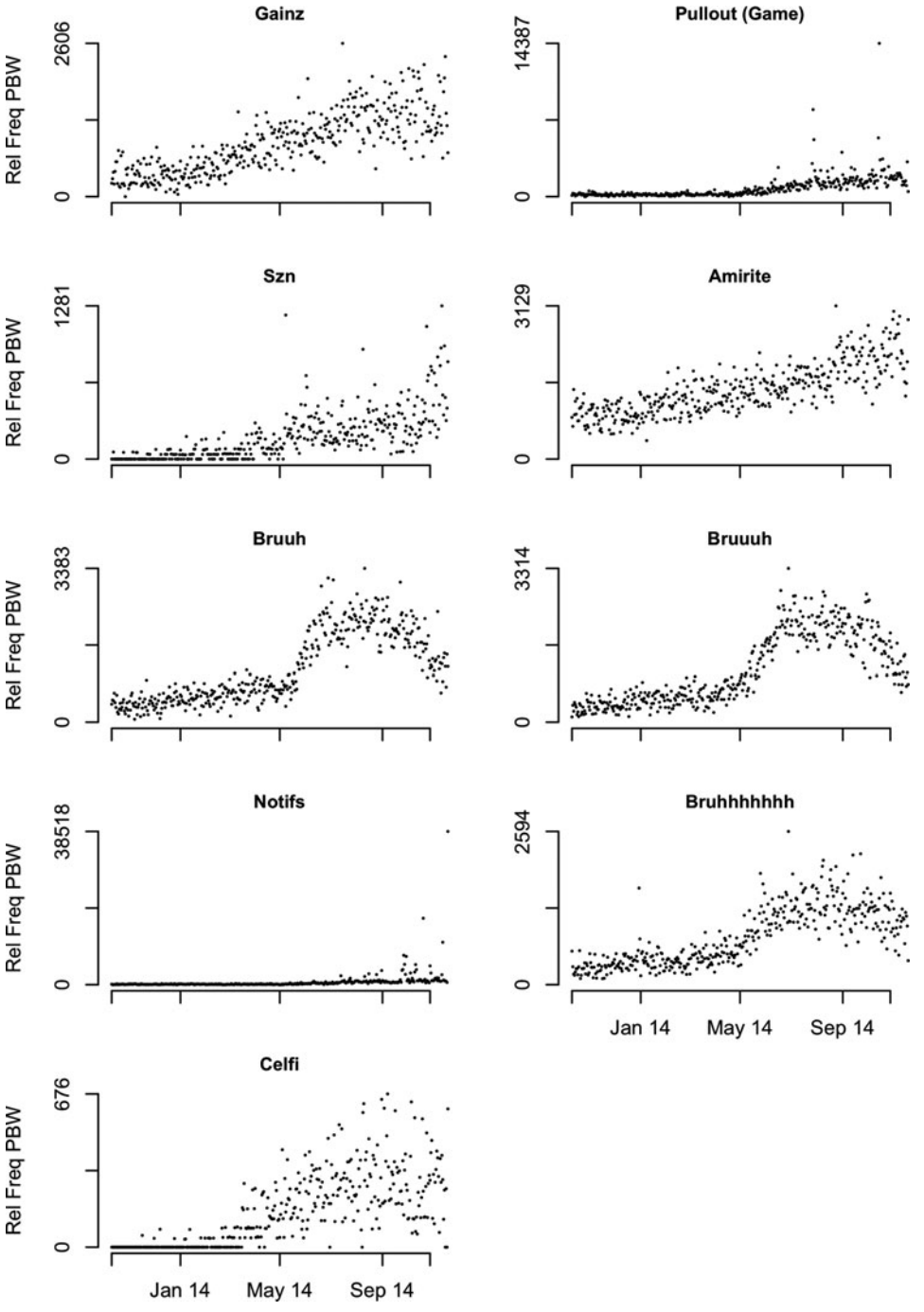

Figure 5. Emerging words time charts (part 3)

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

an s-shaped curve (e.g. *baeless*, *fleek*, *famo*, *lordt*, *gainz*), or the frequencies sharply fall (e.g. *celfie*, *DWK, bruuh*), often notably with high-frequency outliers right before the decline begins. There are also some more complex rising patterns that show multiple inflection points (e.g. *fallback*, *tooka*, *boolin*). Finally, the time charts for some forms show local spikes in relative frequency (e.g. *TFW*, *fuckboys*, *famo*) and outlier days with very high relative frequencies (e.g. *notifs*, *lordt*, *pullout*). Overall, the relative frequencies of these emerging forms tend to follow super-linear patterns of growth, sometimes gradually stabilizing and other times quickly falling off.

The results of this study are therefore largely consistent with an s-shaped curve theory of language change. There has been considerable discussion on why language change often follows an s-shaped curve (e.g. Labov 1972, 1994, 2001; Kroch 1989; Denison 2003; Nevalainen & Raumolin-Brunberg 2003; Aitchison 2001, Blythe & Croft 2012). Perhaps most notably, Labov (2001: 66) explains that the frequency of an incoming form will follow an s-shaped curve over time when measured relative to the frequency of the established form it is replacing, if one assumes that the 'probability of contact between the two governs the rate of change' (see also Kroch 1989; Denison 2003; Nevalainen & Raumolin-Brunberg 2003). In other words, an s-shaped curve should occur if the primary factor that determines the rate of adoption of an incoming form is the rate of contact between speakers who have and do not have that form, as opposed to the rate of contact between speakers who both have or both do not have that form. This is because in a population of speakers where some proportion use the incoming form and some proportion use the established form, the likelihood that these two types of speakers will communicate is equal to the product of their respective proportions. Consequently, the rate of change will be slowest at the beginning and end of the change, when the probability of matching divergent speakers is most unlikely, and fastest when the two forms are equally distributed across the population, when the probability of matching divergent speakers is most likely, giving rise to the s-shaped curve of language change.[6]

Although explanations for the s-shaped curve of language change have generally focused on accounting for the rise in frequency of an incoming form measured relative to the frequency of an established form, the same basic explanation applies to change in the relative frequencies of individual linguistic forms measured relative to the total number of words. In particular, the relative frequency of an emerging form should follow an s-shaped curve of change if the probability of interaction between people who know the form and people who do not know the form drives the rate of change in the relative frequency of that form. Arguably, this version of the s-shaped curve of language change is closer to the way the s-shaped curve is conceived in population dynamics, where it was discovered by Pierre François Verhulst in 1838, who argued that the basic rate of growth of a population is determined by the current size of the population and

[6] Alternatively, Blythe & Croft (2012) compare evolutionary-inspired mathematical models of linguistic diffusion, concluding that only models where incoming variants have different social value reliably produce an s-shaped curve.

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

the maximum possible size of the population given the available resources, which is referred to as the *carrying capacity* of an environment.[7] At first population growth is largely unconstrained and rises super-linearly; however, as the population nears carrying capacity, competition for resources causes population growth to slow until it eventually stabilizes at carrying capacity. Verhulst modeled how a population would grow over time based on these two quantities, which he formalized and called the *logistic growth model*.

The rise of an emerging form is similar in the sense that the frequency of a form is analogous to the size of the population and the maximum frequency of a form is analogous to the carrying capacity of the environment, although in this case the maximum frequency of a form reflects both the frequency of other equivalent forms and the frequency with which the meaning expressed by these forms is discussed in that variety – what might be called the *semantic carrying capacity* of that form. Furthermore, the late decline of some of the emerging forms identified in this study (e.g. *celfie*, *DWK*, *bruuh*) also appears to be analogous to exponential population growth, which is often contrasted with logistic population growth and is characterized by super-linear patterns of population growth that exceed carrying capacity and then abruptly crash.

5.5 Onomasiological competition

The 29 forms all appear to have various synonyms in the English language, as is generally the case with slang. These forms are therefore all necessarily involved in onomasiological competition with other lexical items. It is generally difficult, however, to analyze onomasiological competition directly, because it is often unclear what is the exact meaning (or range of meanings) denoted by a form and because it is challenging to identify contexts where different forms are synonymous. This is why most studies of onomasiological change have been based on the manual analysis of a relatively small number of carefully selected lexical alternations.

Despite the difficulties associated with directly analyzing onamasiological competition over time, clear synonyms can be identified for some of the 29 emerging forms identified in this study. As such, it is possible to measure the frequency of some of these forms relative to the frequency of their established synonyms (as opposed to the total number of words). This can be achieved by dividing the frequency of the emerging form by the combined frequency of the emerging form and the established forms in each daily subcorpus and multiplying this value by 100 to obtain a percentage. The graphs for four relatively straightforward examples are provided below in figure 6: *fuckboi/fuckboy, xan/xanax, celfie/selfie* and *faved/favorited.* It should be noted that in none of these cases were all synonymous forms taken into consideration. For example, there are countless other verbs that could be interchanged with *faved* on Twitter without the loss of basic

[7] Research on the diffusion of innovations has also identified s-shaped curves in how new technology and ideas are adopted by individuals in a society (Rogers 2010).

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

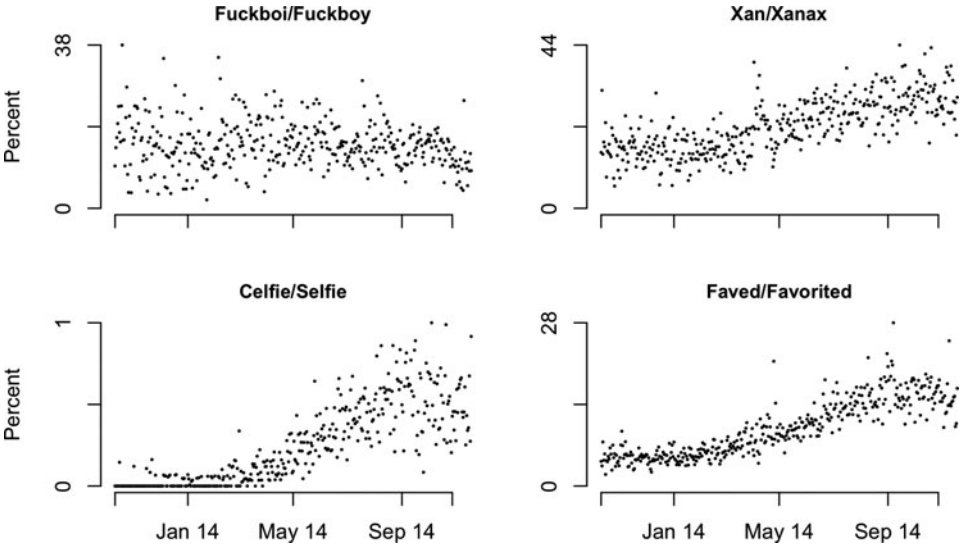

Figure 6. Onomasiological competition time charts

referential meaning, including *starred* or *liked*; however, as noted above, it is very difficult to establish this complete set of possibly synonymous forms or to distinguished usages of these forms that are actually synonymous (e.g. *he liked that Tweet* vs. *he liked that blog*). Nevertheless, the forms selected for comparison provide obvious, common, consistent and meaningful comparators for each of the emerging forms, thereby yielding interpretable measures and graphs of onomasiological competition.

It is particularly informative to compare these four time charts to the corresponding time charts presented in figures 3 to 5. For example, whereas the frequency of *fuckboi* shows a super-linear pattern of growth when measured relative to the total number of words in the corpus (figure 4), it shows no pattern of growth when measured relative to the synonymous form *fuckboy.* This discrepancy occurs because the use of *fuckboy* is increasing at a similar rate over this same period of time (Spearman correlation = 0.952). The time chart in figure 6, however, shows a decrease in the amount of variance in the alternation between these two forms over time, which reflects a stabilization of their proportional use. Alternatively, the alternation between *xan* and *xanax* shows a clear s-shaped curve. This same pattern is not visible, however, in the time chart for *xan* on its own, which simply shows a super-linear pattern of growth. This discrepancy occurs in part because *xanax* is rising at a much slower rate than *xan* (Spearman correlation = .287). Alternatively, the time chart for the alternation between *celfie* and *selfie* is almost identical to the time chart for *celfie*. This is because the term *selfie* is so much more common than *celfie*, even though *selfie* is also rising moderately over the same time period (Spearman correlation = .440). Finally, the time chart for the alternation between *faved* and *favorited* shows a clear s-shaped curve. This is similar to the time chart for the relative frequency of *faved* (figure 4), which also shows an

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

Table 3. Fleek *meaning generalization over time*

Date	Tweet
26 June 2014	*eyebrows on what? on fleek???*
	eyebrows on fleek, the fuck?
	Eyebrows on fleek
	Fleek?
	Eyebrows on "fleek" lmmfao.
	Eyebrows on fleek. da fuq
	EYEBROWS ON FLEEK
	what the hell is fleek
	Lmfao on fleek?
	eyebrows on fleek
22 November 2014	*"... so all ya bitches got eyebrows?" Yeahh and my shits on fleek*
	"I find my paradise when you look me in the eyes. Jobros on fleek..."
	Makeup was on fleek
	When your brows be on fleek but you ain't going no where
	Apparently my eyebrows are on fleek
	Today in autocorrect the shade queen: "on fleek" became "on fleet"
	Braids on fleek
	I fleek a leek a week
	I'm on fleek
	After winter break everything gone be on fleek

s-shaped curve, although the time chart for the individual word shows a moderate decline in usage over time. This decline is less pronounced when the frequency of *faved* is measured relative to the frequency of *favorited* because this measure better controls for the overall fall in references to *favoriting* in general, as indicated by the relatively strong negative correlation for *favorited* (Spearman correlation $= -.616$), which is still the more common form. The comparisons of these sets of time charts demonstrate just how complex the analysis of lexical change can be and how analyzing onomasiological competition can provide an important perspective on the process of lexical emergence.

5.6 Semasiological change

Despite their short history, it is also possible that the meanings of these 29 forms are changing over time. For example, consider the change in the meaning of (*on*) *fleek* as illustrated in table 3, which presents the first 10 Tweets from the corpus containing *fleek* on 26 June 2014, which is the first day in the corpus with at least 10 usages of the form, and the last 10 Tweets from the corpus on 22 November 2014. Seven out of the first 10 usages of *fleek* co-occur with a form of the word *eyebrow* and the other three usages question what *fleek* means. Alternatively, of the final 10 occurrences only three usages of *fleek* occur with a form of the word *eyebrow*, with the form now being used

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

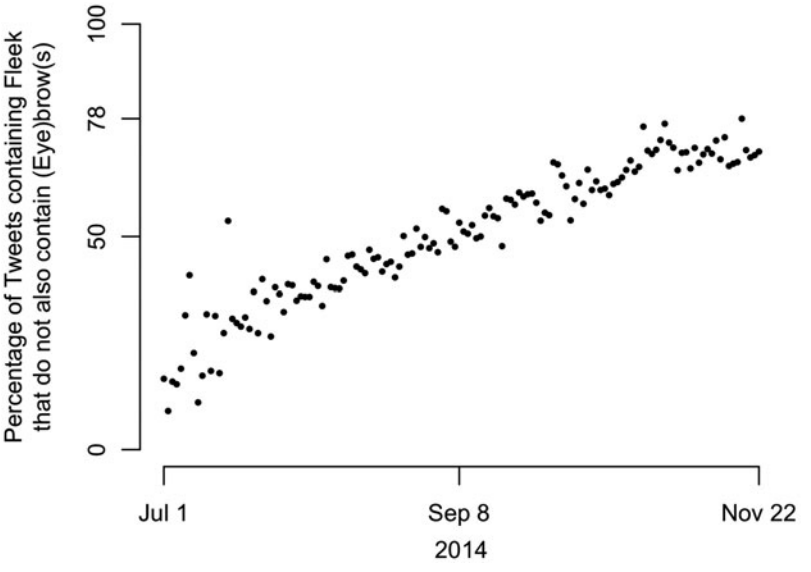

Figure 7. Semasiological change in (*on*) *fleek*

to refer to various other well-presented things, including makeup, braids and people. The form (*on*) *fleek* therefore appears to have quickly expanded in meaning, becoming more generalized. This semasiological change is visualized in figure 7, which plots the percentage by day of Tweets containing *fleek* that do not include contain a reference to eyebrows (i.e. *brow*, *brows*, *eyebrow*, *eyebrows*) from 1 July until the end of the corpus. This chart shows a clear sublinear rise in percentage over time, with usage of the form quickly broadening at first before gradually stabilizing. None of the other emerging forms shows such clear semantic shifts, but further manual analyses of these forms might identify patterns of semasiological change in their usage as well. Taking a broader diachronic perspective would also most likely allow for semasiological change to be observed in greater detail.

5.7 The status of established words

As noted above, in addition to the 29 emerging forms, nine established forms that were very uncommon at the end of 2013 were found to show substantial increase in usage over the course of 2014. These nine forms were excluded from the list of emerging forms because they are establish lexical items that are listed in standard dictionaries and do not appear to be the product of relatively recent word formations. These nine forms, however, are presented in table 4, along with their 2013 relative frequency, their Spearman correlation coefficient and a definition. Time charts are also presented in figure 8, which in general are very similar to the charts for emerging forms presented

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

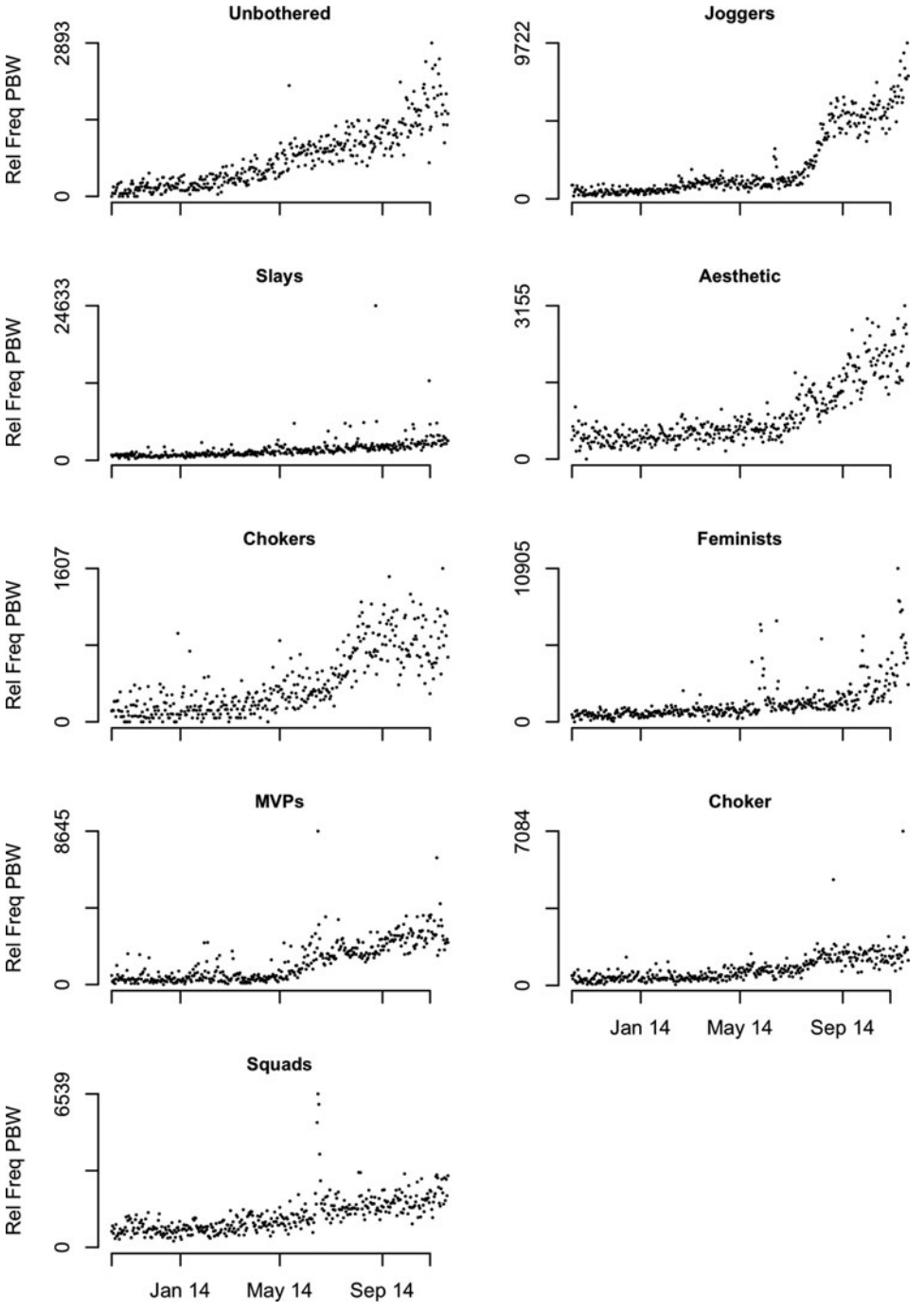

Figure 8. Established words time charts

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

Table 4. *Established words*

Word form	Spearman correlation coefficient	2013 frequency PBW	Meaning
unbothered	0.931	159	happily oblivious
joggers	0.913	453	jogging pants
slays	0.847	794	succeeds
aesthetic	0.825	432	personal style
chokers	0.822	133	choker necklaces
feminists	0.819	523	feminists
mvps	0.818	363	most valuable players
choker	0.815	341	choker necklace
squads	0.81	778	group of friends

in figures 3–5, although *joggers* in particular shows a notably singular pattern, which evokes Aitchison's (2001) theory of multiple s-shaped curves of language change.

Semasiological change appears to explain the rise in the use of 6 of these 9 forms, which are often used on Twitter with different meanings than their dictionary definitions. Specifically, *unbothered* is used not just to mean *not bothered* but *happily oblivious*, *joggers* is used primarily to refer to *jogging pants* rather than *people who jog*, *slays* is used primarily with a metaphorical meaning of defeating or conquering something (e.g. a guitar solo), *aesthetic* is used almost exclusively to mean *personal aesthetic* (i.e. *a personal style*), *squads* is used to refer mainly to gangs of young men and *mvps* is used to refer to valuable people in general as opposed to players in sports leagues. It is arguable that all these words should be included in the list of emerging words, especially if semasiological change is seen as being a type of word formation process (see Geeraerts 2010). Semasiological change, however, is probably best seen as being distinct from lexical emergence, as it does not necessarily involve the initial rise of a relatively new word form. Furthermore, the method introduced in this article identifies emerging forms by looking for forms that are rare at the start of the period under analysis and therefore cannot consistently identify semasiological change, which generally involves more frequent forms.

Regardless of the theoretical relationship between onomasiological and semasiological change, it is important to note that *unbothered* and *joggers* may actually be straightforward emerging forms generated through the recent application of standard word formation processes. Specifically, *unbothered* is arguably a formation derived by attaching the prefix *un-* to the relatively common and well-known word *bothered* by people who did not know that the relatively rare form *unbothered* already existed, while *joggers* to refer to *jogging pants* is arguably not a shift in the meaning of the existing word *joggers* (i.e. people who jog), but a truncation and derivation of the term *jogging pants*, most likely inspired by various other related forms (e.g. *trousers*, *slippers*, *sneakers*). These two forms should therefore perhaps be considered emerging forms

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

and be added to table 1, which would make *unbothered* in particular the most quickly rising emerging form on American Twitter in 2014. These two forms also demonstrate why it is important to inspect concordance lines for each potential emerging form identified by the method even if they appear to already exist.

Finally, the three remaining words in table 2 do not appear to be undergoing major semasiological shifts. Rather, the increase in the use of both the words *choker(s)* (i.e. choker necklaces) and *feminists* appears to reflect topics in which interest has increased over the course of 2014, although the relative frequency of both *choker* and *chokers* notably does show an s-shaped curve. Although culturally interesting, analyzing change in the usage of such forms is not central to the study of lexical emergence.

6 Conclusions

This article has introduced the concept of lexical emergence and has described a simple statistical method for identifying emerging word forms in large, time-stamped corpora. This method was then used to identify a set of 29 emerging word forms in Modern American English based on an 8.9 billion-word corpus of Twitter posts collected between 2013 and 2014. Finally, these forms were inspected from a variety of perspectives in order to better understand the process of lexical emergence.

At the most general level, this study has shown that patterns of lexical emergence can be identified in Modern American English through the quantitative analysis of large corpora of social media. Admittedly, the analysis presented in this article has not identified an especially large set of emerging word forms and there can be little doubt that this is not the complete set of emerging word forms attested on American Twitter in 2014, much less in Modern American English. The main reason why a larger number of emerging forms was not identified is that although the analysis presented in this study is based on a very large corpus by modern standards, it is still not large enough to allow for a comprehensive analysis of lexical variation and change. As more and more data become available, however, it will be possible to conduct more and more detailed studies of lexical emergence. In addition, two of the main parameters that must be set when applying this method (the minimum frequency of the forms under analysis and the Spearman correlation coefficient cut off) were assigned relatively conservative values; lowering these requirements would have allowed for additional potential emerging word forms to be identified.

Despite these limitations, the analysis identified a number of emerging forms from a variety of different topical domains and has led to numerous interesting observations about the nature of lexical emergence. Perhaps most notably, this analysis has found that new word forms are often introduced many years before they eventually emerge. This is not necessarily how one would assume that lexical emergence operates: another possibility is that out of the countless nonce words that are being formed every day in spontaneous language use, a small number of these forms would happen to spread across the population very soon after they are introduced, while most

would simply be forgotten. But this study presents evidence that this is not always the case: most of the emerging forms analyzed in this study appear to have lain dormant for years before they began to spread. Identifying what factors trigger the emergence of these forms is an important area for future research on lexical emergence.

In addition, this study has shown that the relative frequencies of emerging word forms generally follow super-linear patterns of growth, which resolve themselves either by stabilizing to create an s-shaped curve or quickly falling out of use. These results are largely consistent with previous research on language change, where it has been theorized that the frequency of incoming forms follow s-shaped patterns of change, when measured relative to the frequency of equivalent forms, due to the changing probability of interaction of speakers who have those forms and speakers who do not. This study has shown that the relative frequencies of individual incoming word forms also follow this type of s-shaped pattern, and it has been argued that the same basic explanation for the pattern applies. The mathematical modeling of the general process of lexical emergence, in much the same way that population growth has been modeled in biology, is another important area for future research, as it would allow the factors that affect the emergence of words and ultimately their success or failure to be studied in a systematic way. It is also important to note that the method introduced in this article does not necessarily identify forms that will go on to become full-fledged lexical items in the standard vocabulary of the English language. Some may, but many others may fall out of usage just as quickly as they have risen to prominence. This method, however, provides a basis for studying this type of phenomenon, once the amount of language that linguists have access to increases, both in terms of volume and chronological depth.

Finally, perhaps the most interesting area for future research opened up by this study is the analysis of the regional and social origins of emerging word forms. For example, many of the emerging words identified in this study appear to originate in the African American community, which would seem to be an especially influential segment of the American population. In addition, analyzing the regional spread of emerging forms is a particularly exciting area for future research, which is possible using geocoded corpora, including the corpus analyzed in this study.

As well as these methodological, descriptive and theoretical results, this study has also shown how adopting a big data, corpus-based approach to linguistics can open up new areas for research, especially related lexical variation and change, which requires massive amounts of language data. As more language data become available online, including eventually large amounts of spoken language data, research on lexical variation and change, including lexical emergence, will undoubtedly increasingly rely on analyzing these types of data. There are certainly inherent difficulties working with data that have been harvested online, but as this article has demonstrated they are outweighed by the advantages of analyzing very large corpora of natural language, which provide an unprecedented view of the complexity and dynamicity of language.

Authors' addresses:

Centre for Forensic Linguistics
Aston University
Birmingham B4 7ET
UK
j.grieve1@aston.ac.uk
a.nini1@aston.ac.uk

Department of Geography
University of South Carolina
Columbia South Carolina 29208
USA
guod@mailbox.sc.edu

References

Allan, Kathryn & Justyna Robinson (eds.). 2011. *Current methods in historical semantics*. Berlin: De Gruyter.
Aitchison, Jean & Diana Lewis. 1995. How to handle wimps: Incorporating new lexical items as an adult. *Folia Linguistica* 29, 7–20.
Aitchison, Jean. 2001. *Language change: Progress or decay?* Cambridge: Cambridge University Press.
Bamman, David, Jacob Eisenstein & Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18, 135–60.
Bauer, Laurie. 1983. *English word-formation*. Cambridge: Cambridge University Press.
Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.
Blythe, Richard A. & William Croft. 2012. S-curves and the mechanisms of propagation in language change. *Language* 88, 269–304.
Bréal, Michel. 1897. *Essai de sémantique: sciences des significations*. Paris: Hachette.
Brinton, Laurel J. & Elizabeth C. Traugott. 2005. *Lexicalization and language change*. Cambridge: Cambridge University Press.
Cannon, Garland. 1987. *Historical change and English word formation*. New York: Lang.
Davies, Mark. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25, 447–64.
Darmester, Arsène. 1887. *La vie des mots étudiée dan leur significations*. Paris: Delagrave.
Denison, David. 2003. Log(ist)ic and simplistic S-curves. In Hickey (ed.), 54–70.
Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith & Eric P. Xing. 2010. A latent variable model for geographic lexical variation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1277–87.
Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith & Eric P. Xing. 2014. Diffusion of lexical change in social media. *PloS one* 9, e113114.
Fischer, Roswitha. 1998. *Lexical change in Present-day English: A corpus-based study of the motivation, institutionalization, and productivity of creative neologisms*. Tübingen: Narr.
Geeraerts, Dirk. 2010. *Theories of lexical semantics*. Oxford: Oxford University Press.
Geeraerts, Dirk & Hubert Cuyckens (eds.). 2007. *The Oxford handbook of cognitive linguistics*. Oxford: Oxford University Press.

https://doi.org/10.1017/S1360674316000113 Published online by Cambridge University Press

Geeraerts, Dirk, Caroline Gevaert & Dirk Speelman. 2012. How anger rose: Hypothesis testing in diachronic semantics. In Allan & Robinson (eds.), 109–l32.
Geeraerts, Dirk, Stefan Grondelaers & Peter Bakema. 1994. *The structure of lexical variation: Meaning, naming, and context*. Berlin: De Gruyter.
Green, Jonathan. 2011. *Green's dictionary of slang*. Edinburgh: Chambers.
Gries, Stefan Th. & Martin Hilpert. 2010. Modeling diachronic change in the third person singular: A multifactorial, verb- and author-specific exploratory approach. *English Language and Linguistics* 14, 293–320.
Grondelaers, Stefan, Dirk Geeraerts & Dirk Speelman, D. 2007. Lexical variation and change. In Geeraerts & Cuyckens (eds.), 988–1011.
Haddican, Bill & Daniel E. Johnson. 2012. Effects on the particle verb alternation across English dialects. *Selected papers from NWAV 40*. Philadelphia: University of Pennsylvania Press.
Hickey, Raymond (ed.). 2003. *Motives for language change*. Cambridge: Cambridge University Press.
Hilpert, Martin & Stefan Th. Gries. 2009. Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing* 24, 385–401.
Hohenhaus, Peter. 2005. Lexicalization and institutionalization. In Pavol Štekauer & Rochelle Lieber (eds.), *Handbook of word-formation*, 353–73. Dordrecht: Springer.
Hohenhaus, Peter. 2006. Bouncebackability: A web-as-corpus-based study of a new formation. *SKASE Journal of Theoretical Linguistics* 3, 17–27.
Hopper, Paul. J. & Elizabeth C. Traugott. 2003. *Grammaticalization*. Cambridge: Cambridge University Press.
Huang, Yuan, Diansheng Guo, Alice Kasakoff & Jack Grieve. 2016. Understanding US regional linguistic variation with Twitter Data. Forthcoming in *Computers, Environment and Urban Systems*.
Kerremans, Daphne, Susanne Stegmayr & Hans-Joerg Schmid. 2011. The NeoCrawler: Identifying and retrieving neologisms from the internet and monitoring ongoing change. In Allan & Robinson (eds.), 59–96.
Kilgarriff, Adam. 2001. Web as corpus. *Proceedings of Corpus Linguistics 2001*, 342–4.
Kleparski, Grzegorz. 2000. Metonymy and the growth of lexical categories related to the conceptual category Female Human Being. *Studia Anglica Resoviensia* 1, 17–26.
Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1, 199–244.
Krug, Manfred G. 2000. *Emerging English modals: A corpus-based study of grammaticalization*. Berlin: De Gruyter.
Kurath, Hans. 1949. *A word geography of the Eastern United States*. Ann Arbor, MI: University of Michigan Press.
Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
Labov, William. 1994. *Principles of linguistic change*, vol. 1: *Internal factors*. Oxford: Blackwell.
Labov, William. 2001. *Principles of linguistic change*, vol. 2: *Social factors*. Oxford: Blackwell.
Lipka, L., Handl, S. & Falkner, W. 1994. Lexicalization and institutionalization. In Asher (ed.), *the encyclopedia of language and linguistics*, 2164–7.
Marchand, Hans. 1969. *The categories and types of present-day English word-formation: A synchronic-diachronic approach*. Munich: Beck.
Méndez-Naya, Belén. 2006. Adjunct, modifier, discourse marker: On the various functions of *right* in the history of English. *Folia Linguistica Historica* 40, 141–69.

Méndez-Naya, Belén. 2008. On the history of *downright*. *English Language and Linguistics* 12, 267–87.
Miller, D. Garry. 2014. *Lexicogenesis*. Oxford: Oxford University Press.
Nevalainen, Terttu & Helena Raumolin-Brunberg. 2003. *Historical sociolinguistics*. Harlow: Longman.
O'Connor, Brendan, Jacob Eisenstein, Eric P. Xing & Noah A. Smith. 2010. A mixture model of demographic lexical variation. *Proceedings of NIPS Workshop on Machine Learning for Social Computing* 14.
Petersen, Alexander M., Joel Tenenbaum, Shlomo Havlin & H. Eugene Stanley. 2012a. Statistical laws governing fluctuations in word use from word birth to word death. *Scientific Reports* 2, 313.
Petersen, Alexander M., Joel Tenenbaum, Shlomo Havlin & H. Eugene Stanley & Matjaz Perc. 2012b. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific Reports* 2, 943.
Reisig, Karl. 1839. *Vorlesungen über die lateinische Sprachwissenschaft*. Leipzig: Lehnhold.
Rogers, Everett. 2010. *Diffusion of innovations*. New York: Simon and Schuster.
Siemund, Peter. 2014. The emergence of English reflexive verbs: An analysis based on the *Oxford English Dictionary*. *English Language and Linguistics* 18, 49–73.
Sweetser, Eve. 1991. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*. Cambridge: Cambridge University Press.
Tagliamonte, Sali A. & Alexandra D'Arcy. 2004. *He's like, she's like*: The quotative system in Canadian youth. *Journal of Sociolinguistics* 8, 493–514.
Whitman, Neal. 2015. Geeking out on fleek. www.vocabulary.com/articles/dictionary/geeking-out-on-fleek/ (accessed 28 May 2015).
Zhang, Weiwei, Dirk Geeraerts & Dirk Speelman. 2015. Visualizing onomasiological change: Diachronic variation in metonymic patterns for Woman in Chinese. *Cognitive Linguistics* 26, 289–330.
Zipf, George. K. 1935. *The psycho-biology of language*. New York: Houghton Mifflin.
Zipf, George. K. 1949. *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.