NATURAL SELECTION IN THE MODERN ENGLISH LEXICON

JACK GRIEVE[*1]

[*]j.grieve@bham.ac.uk
[1]English Language and Linguistics, University of Birmingham, Birmingham, UK

This study tests the degree to which the form and function of 54 newly emerging words predicts their success over time in a multi-billion word corpus of American Twitter collected between 2013 and 2016. A linear model of the change in the relative frequency of each word is computed as a function of word length, part-of-speech, word formation process, and meaning. The analysis finds that the most important predictor of the success of these words is marking a new meaning. Shorter words and words created through standard word formation processes are also found to be more successful over time. These results are interpreted as supporting the theory that natural selection is the driving force behind lexical evolution.

1. Introduction

The evolution of the lexicon is difficult to study because most words, especially new words, are incredibly rare. Observing general patterns of lexical innovation therefore requires access to extremely large and densely sampled corpora of natural language. It has only recently become possible to compile such corpora with the rise in popularity of social media, which deposits huge amounts of informal written language online.

For example, based on an analysis of a 9-billion-word corpus of American Twitter, we identified 54 newly emerging words, which were very uncommon at the end of 2013 but whose usage increased substantially over the course of 2014 (Grieve et al. 2017; Grieve et al. 2018). By extracting a relatively large sample of emerging words, we have been able to make general claims about the form, function, and origin of new word formations in Modern American English. This research has also opened up other lines of research, including investigating the factors that predict whether or not emerging words survive over time.

The goals of this study are therefore to begin to understand the forces that drive the evolution of the modern English lexicon by (1) measuring the degree to which the frequencies of these 54 emerging words have changed on Twitter

DOI:10.12775/3991-1.037

between 2014 and 2016 and (2) testing the degree to which a range of factors predict the success of these words.

2. Analysis

In our previous research on lexical innovation, we identified 54 emerging words in a 9 billion word corpus of geocoded American Twitter collected between October 2013 and November 2014 (Grieve et al. 2017, 2018). These 54 emerging words, which are listed in Table 1, grew steadily in popularity on Twitter in 2014 and represent a wide range of different parts-of-speech, semantic domains, and word formation processes.

Table 1. The 54 emerging words used in the analysis

amirite	*cosplay*	*gainz*	*lordt*	*rekt*	*thotful*
baeless	*dwk*	*gmfu*	*lw*	*rq*	*thottin*
baeritto	*fallback*	*goalz*	*mce*	*scute*	*tookah*
balayage	*famo*	*idgt*	*mmmmmmmuah*	*senpai*	*traphouse*
boolin	*faved*	*lfie*	*mutuals*	*shordy*	*unbae*
brazy	*fhritp*	*lifestyleeee*	*nahfr*	*slayin*	*waifu*
bruuh	*figgity*	*litt*	*notifs*	*sqaud*	*wce*
candids	*fleek*	*litty*	*pcd*	*tbfh*	*xans*
celfie	*fuckboys*	*lituation*	*pullout*	*tfw*	*yaas*

To quantify the degree to which the popularity of these words changed between 2014 and 2016, I measured the relative frequency of each word in the November section of the 2014 Twitter corpus, the last month in that corpus, and in a new 9 billion word corpus of geocoded American Twitter from 2016. I then computed the factor by which the relative frequency of each word changed between 2014 and 2016 (i.e. the 2016 relative frequency divided by the 2014 relative frequency). Finally, I calculated the log of this factor so that rises and falls in frequency are measured on comparable scales.

This analysis found that the use of the 54 emerging words changed considerably over time. For example, *unbae* (i.e. 'to break up with') dropped from 148 occurrences per million words (pmw) in November 2014 to 4 occurrences pmw in 2016 (-2.0 logged factor of change), while *brazy* (i.e. 'crazy') rose from

1,745 occurrences pmw in November 2014 to 10,723 occurrences pmw in 2016 (+0.8). Overall, a majority (30/54) words fell in usage over this period.

I then constructed a linear model to predict change in the frequency of the 54 emerging words between 2014 and 2016 as a function of four independent variables: length (in characters), part-of-speech (nominal, verbal, adjectival, other), word formation process (acronym, creative spelling, standard), and whether or not the word marks a new meaning.These four variables were selected as predictors because they provide distinct and basic information about the form and the function of these words. The last three predictors require some explanation.

The part-of-speech variable includes 'Nominal', 'Verbal', and 'Adjectival' categories (rather than 'Noun', 'Verb', and 'Adjective') to allow for multiword units, including phrases represented by acronyms, to be classified. The 'Other' category is small and consists mainly of inserts.

The word formation process variable includes 'Acronyms' and 'Creative Spellings' because these are very common orthographic word formation processes on Twitter, even though they are uncommon in speech. Alternatively, the 'Standard' word formation process category includes all processes that are common in spoken language (e.g. compounds, blends, truncations, derivations, borrowings) (Bauer 1982).

The meaning variable was the most difficult to code. The basic distinction being drawn is between words that have meanings that are not already listed in a standard dictionary (e.g. *balayage,* which refers to a specific hair style) and words that have existing synonyms in Standard English (e.g. *baeless,* which means "to be single"). Creative spellings, which always represent existing words, were coded as marking new meanings only if they were associated with a specific non-standard meaning of that word (e.g. *gainz,* which specifically refers achieving 'weight gains' through exercise), as opposed creative spelling used for emphasis or other functions (e.g. *yaas*).

The linear model for frequency change as a function of these 4 independent variables was found to be significant ($F(8, 45) = 4.28$, $p < 0.001$) with an adjusted r-squared of .33. Most notably, meaning was found to be a relatively strong predictor of emerging word success, with words that mark new meanings being especially successful. In addition, shorter words and words formed using standard word formation processes were also found to be more likely to succeed. Alternatively, part-of-speech was found to have relatively little effect on change in the usage of these words. The complete analysis, including R code and data, is available online (Grieve 2018).

3. Conclusion

Overall, the analysis identified three factors that predict if emerging words will survive on Twitter. The most important of these predictors is whether or not the word marks a new meaning, with words that express new meanings being substantially more likely to survive. This finding sugegsts that the communicative utility of an emerging word is a strong predictor of its success, at least in this variety of language. If an emerging word fills a semantic gap in the standard lexicon, rather than simply providing a synonym for an existing word, it is more likely to be retained.

Similarly, the analysis found that shorter words are more likely to be succeed. This may be because Twitter places strict limits on text length, creating a communicative context that favours shorter words. Although this effect may therefore be restricted to this particular variety of language, given Zipf's (1936) observation that shorter words are generally more common than longer words, this result may also due to a more general principle of lexical change.

Finally, words generated through standard word formation processes were found to be more likely to succeed than words generated through specialised processes that are largely restricted to written language and computer-mediated communication. This finding suggests that emerging words that are suitable for use across varieties of language are more likely to succeed even on Twitter.

All three of these results support the claim that natural selection is a driving force behind lexical evolution, as Darwin himself first proposed in the *Descent of Man* (2003, 1871), where he wrote that "the survival or preservation of certain favoured words in the struggle for existence is natural selection." In particular, this analysis has found that words that are more useful for communication are more likely to succeed, including words that express unique meanings, that are more efficient, and that can be used across a wide range of communicative contexts.

Whether or not these results hold for other varieties of language is an open question. There are also numerous limitations with the present study, most notably the relatively small number of emerging words under analysis and the limited time frame. Given these issues, the main contributions of this study are (1) to illustrate how lexical evolution can be explored through the quantitative analysis of very large corpora of modern language, (2) to provide a preliminary exploratory analysis of the effect of a variety of factors on the success of emerging words on American Twitter, and (3) to present initial empirical support for a general theory of lexical evolution based on natural selection.

References

Bauer, L. (1983). *English Word-Formation*. Cambridge: Cambridge University Press.

Darwin, C. (2003). *The Descent of Man*. London: Gibson Square.

Grieve, J. (2018). R Analysis for Natural Selection in the Modern English Lexicon. Rpubs. http://rpubs.com/jwgrieve/evolangxii

Grieve, J., Nini, A. & Guo, D. (2017). Analyzing lexical emergence in American English online. *English Language and Linguistics* 21, pp. 99-127.

Grieve, J., Nini, A. & Guo, D. (2018). Mapping lexical innovation on American social media. Forthcoming in *Journal of English Linguistics*.

Zipf, G. (1936). The Psychobiology of Language. London: Routledge.