

Multivariate analysis of short-term lexical change

Jack Grieve (University of Birmingham, UK)

Most research on lexical change has analysed variation in the use of relatively few words over relatively long periods of time (Traugott and Dasher, 2001). In part this is because of limitations in the quantity of language data that has been available for analysis. As most words are very rare (Zipf, 1949), it is difficult to study change in word usage on a large scale based on small corpora. The situation, however, has recently changed with the rise of online communication, especially social media, which has allowed for extremely large and densely sampled corpora to be compiled. These new multi-billion word corpora have opened up several areas of research on lexical variation and change (e.g. Bamman et al., 2014; Eisenstein et al., 2014), including making it possible to analyse change in the relative frequencies of large numbers of words simultaneously over very short periods of time. To better understand the mechanisms of short-term lexical change in modern English, this paper presents a multivariate analysis of change over time in the relative frequencies of tens of thousands of word forms based on a multi-billion word time-stamped corpus of American social media data from 2013-2014. Rather than analyse change in the frequency of each word individually, this set of words was analysed together using a multivariate statistical analysis to identify the most common time series patterns in this variety, allowing for the dynamics of lexical change to be analysed from a new perspective and in unprecedented detail.

This analysis is based on an 8.9 billion word corpus of American Tweets collected between October 2013 and November 2014 using the Twitter API (see Huang et al., 2016; Grieve et al., 2017; Nini et al., 2017). In order to analyse general patterns of lexical frequency change in this corpus, the relative frequency of the 67,022 most common words in the corpus (all words forms that occur at least 1,000 times in the complete corpus) were measured across the 397 days in the corpus yielding 67,022 relative frequency time series. Each lexical time series can be visualised as a line graph showing how the relative frequency of that word form has changed over the days in the corpus. This complete dataset was then subjected to multivariate statistical analysis to identify common patterns of diachronic variation.

The time series for each word was first scaled to range between 0 and 1 to preserve the shape of the time series while allowing for the set complete set of words to be compared to each other despite occurring across vastly different relative frequency ranges. The 67,022 scaled time series were then subjected to a non-linear iterative partial least squares (NIPALS) principal component analysis (PCA). A PCA is common multivariate statistical method (Everitt & Hothorn, 2011) that takes a set of variables (in this case, word relative frequencies) measured over a set of observations (in this case, days of the corpus) and extracts a series of principal components that represent the most important patterns of variation in that dataset (in this case, time series). In other words, PCA is a dimension reduction technique: starting with a large set of variables that generally exhibit substantial inter-correlations, PCA identifies a smaller set of uncorrelated dimensions that maximally explain the variability exhibited in the original set of variables. A NIPALS PCA (Wold, 1975) was used in this study because it is a version of PCA that allows for very large

multivariate datasets to be processed efficiently, including datasets with far more variables than observations, as is the case here, which is not possible using a standard PCA. Based on the NIPALS PCA of the scaled time series, the most important patterns of lexical frequency change in the dataset were identified, allowing both for these patterns to be visualised (by plotting the component scores) and for the degree to which the time series for each of the individual words is characterized by these aggregate patterns to be assessed (by inspecting the component loadings).

The first 10 dimensions identified by the NIPALS PCA are plotted in Figure 1, ranked in order of importance (i.e. by amount of variance explained). Each of these graphs plots the component scores for one dimension, in essence allowing for the 10 most common patterns of relative frequency change over time in the 67,022 word forms in the dataset to be visualised. In each case, two time series are plotted on the graph, which are mirror images of each other. This is because each dimension extracted by the PCA identifies words that not only show highly positively correlated patterns but also words that show highly negatively correlated (i.e. complementary) patterns. For example, the first dimension identifies words whose relative frequencies show a relatively consistent rise *or* fall over time. A secondary weekly trend is also visible in Dimension 1, with rising words tending to be used especially frequently on Fridays and Saturdays. In addition to plotting the general shape of these trends, inspecting the component loadings allows for word forms that most clearly exhibit that pattern to be identified. In the case of Dimension 1, rising words include various proper nouns (e.g. *timehop*, *soliant*) and recent word formations (e.g. *fleek*, *fav*), whereas falling words include numerous interjections (e.g. *haha*, *uh*). Subsequent dimensions can be interpreted in similar ways, including identifying both smaller and larger cyclical trends. For example, Dimension 2 primarily identifies a weekly pattern, with certain words being more common on weekends (e.g. *hangout*, *fun*) or weekdays (e.g. *desk*, *lecture*), although a secondary seasonal pattern is also apparent, with weekend words being more common in the summer and weekday words in the winter. Similarly, Dimension 3 identifies a more focused distinction between summer (e.g. *vacation*, *popsicles*) and winter (e.g. *butternut*, *jacket*) words.

Identifying common temporal trends in the relative frequencies of a large numbers of words can inform our understanding of lexical change in numerous reasons. For example, this analysis finds that rising words generally follow a roughly s-shape curve of change on Twitter, which agrees with standard theories of linguistic innovation (e.g. Blythe & Croft, 2012). This analysis also finds that rising words tend to be used especially frequently on weekends, suggesting that certain topics and people may be more or less likely to drive lexical change, at least in this variety of language. More generally, most dimensions show both short-term (e.g. weekly) and long-term (e.g. monthly) cycles, demonstrating that word frequencies pattern at multiple temporal scales concurrently. The long-term cycles that characterise all but the first dimensions appear to primarily reflect cyclical topical trends, raising questions about how such trends might affect the rate of lexical change in different semantic domains.

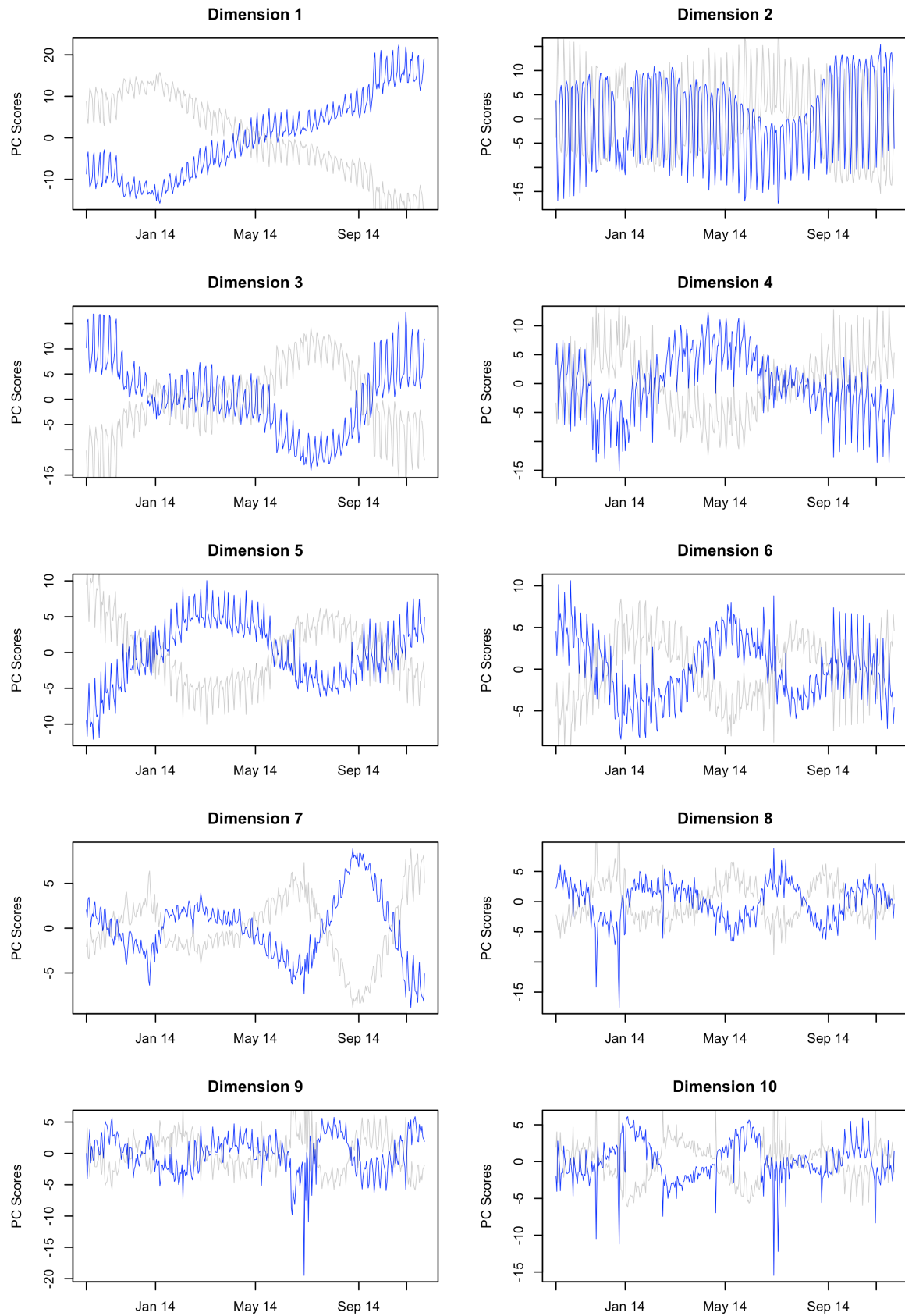


Figure 1 First 10 PCA Lexical Time Series Dimensions

References

- Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18, 135-160.
- Blythe, R. A., & Croft, W. (2012). S-curves and the mechanisms of propagation in language change. *Language*, 88, 269-304.
- Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. *PloS one*, 9, e113114.
- Everitt, B., & Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R*. Springer.
- Grieve, J., Nini, A., & Guo, D. (2017). Analyzing lexical emergence in Modern American English online. Forthcoming in *English Language and Linguistics*, 21, 99-127.
- Huang, Y., Guo, D., Kasakoff, A., & Grieve, J. (2015). Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59, 244-255.
- Traugott, E. C., & Dasher, R. B. (2001). *Regularity in Semantic Change*. Cambridge University Press.
- Nini, A., Corrandini, C., Guo, D., & Grieve, J. (2017). The application of growth curve modeling for the analysis of diachronic corpora. *Language Dynamics and Change*, 7, 102-125.
- Wold, H. (1975). Path models with latent variables: the NIPALS approach. In Balock, H. M., Agabegian, A., Borodkin, F. M., Boudon, R., & Cappecchi (Eds.), *Quantitative Sociology: International Perspectives on Mathematical and Statistical Modelling* (pp. 383-407). Academic Press.
- Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Welsey.