The application of growth curve modeling for the analysis of diachronic corpora

Nini, Andrea; Corradini, Carlo; Guo, Diansheng; Grieve, Jack

Document Version
Peer reviewed version

The application of growth curve modeling for the analysis of diachronic corpora

Andrea Nini, University of Manchester, andrea.nini@manchester.ac.uk
Carlo Corradini, Aston University, c.corradini@aston.ac.uk
Diansheng Guo, University of South Carolina, guod@mailbox.sc.edu
Jack Grieve, Aston University, j.grieve1@aston.ac.uk

Abstract

This paper introduces growth curve modeling for the analysis of language change in corpus linguistics. In addition to describing growth curve modeling, which is a regression-based method for studying the dynamics of a set of variables measured over time, the technique is demonstrated through an analysis of the relative frequencies of words that are increasing or decreasing over time in a multi-billion word diachronic corpus of Twitter. This analysis finds that increasing words tend to follow a trajectory similar to the s-curve of language change, whereas decreasing words tend to follow a decelerated trajectory, thereby showing how growth curve modeling can be used to uncover and describe underlying patterns of language change in diachronic corpora.

1

1. Introduction

The study of language change has always been one of the main concerns of corpus linguistics. Research in diachronic corpus linguistics has focused both on understanding the change in the relative frequency of a single feature (e.g. Biber 2004, Biber & Burges 2000, Hundt 2014, Siemund 2014) and on analyzing change in the relative frequencies of many features in order to describe change across entire varieties of language and to discover general principles of language change (Biber & Finegan 1989, Nevalainen & Raumolin-Brunberg 2003, Säily et al. 2011, Biber & Gray 2013). Corpus linguists have also analyzed the chronological change in the frequency of words to uncover changes in culture and society as opposed to language itself (e.g. Baker et al. 2013, Baker et al. 2008). Similarly, big data approaches have been used in *culturomics* to explore trends of cultural or historical change in the Google Books Corpus, such as the decrease in the frequency of the word *God* or the increase in the frequency of the word *feminism* (Michel et al. 2011). Analyses of the Google Books Corpus have also found that linguistic evolution slows down as the vocabulary of a language grows richer (Petersen et al. 2012a) and that diachronic fluctuation in the frequency of words is negatively related to a word's frequency rank (Petersen et al. 2012b). More recently, corpus-based studies on language change have begun to analyze very large data sets harvested from social media platforms to study language change over short

periods of time (Eisenstein et al. 2012, 2014, Grieve et al. 2015). For example Grieve et al. (2016) identified newly emerging words through a diachronic analysis of the relative frequency of lexical items in a multi-billion word corpus of tweets from 2014.

Several methodological advances have also recently been made in diachronic corpus linguistics. For example, Gries & Hilpert's (2008) use of neighbor clustering and Hilpert & Gries' (2009) iterative sequential interval estimation and regression with breakpoints are new methods for uncovering diachronic stages in time data. Similarly, Hilpert's (2011) motion charts offer a better approach for the visualization of time series data in corpus linguistics. Finally, Millar (2009) introduced an approach for fitting a curve to language time series data using regression, which is especially useful for testing for particular types of distribution over time in frequency data. Despite these methodological advances, an issue with many diachronic studies in corpus linguistics is that although numerous features are taken into consideration, eventually the values of these variables are summed together in one category and treated as a single feature in order to obtain an overall picture of change. For example, Säily et al. (2011) sum the frequencies of nouns and pronouns, while Biber (2004) sums the frequencies of stance-related features such as modal verbs and stance adverbials. Although informative, this operation can result in a substantial loss of information and thus risk obscuring important patterns in the distribution of these forms, especially when large numbers of forms with varying frequencies of different magnitude are analyzed together.

In addition to corpus linguistics, language change is also commonly analyzed in sociolinguistics. For example, understanding the general mechanism and principles of language

change is one of the main goals of variationist sociolinguistics (e.g. Labov 1978, Labov 1995, Chambers 2001). This type of research commonly analyzes the frequency of a linguistic form relative to the frequency of one or more equivalent forms, such as alternative pronunciations and grammatical constructions, in language data collected through sociolinguistic interviews. Variationist research also generally adopts an *apparent time* methodology, where change over time in the values of these alternation variables is estimated by sampling subjects within different age strata. Based on this approach to the analysis of language variation and change, sociolinguists have proposed that an *s-shaped curve model of language change* generally describes the replacement of one linguistic form with another (Denison 2003, Kroch 1989, Labov 1995, 2001). The model specifies that this replacement occurs following a logistic curve, that is, a slow-fast-slow change trajectory tracing an s-shaped pattern. This model has been validated by many studies, a comprehensive review of which is Blythe & Croft (2012). However, as Blythe & Croft (2012: 278) specify, no model of change has been so far proposed for *introduced changes*, that is, no model has been so far proposed for the trajectories taken by the frequencies of single variants changing in time—the types of feature commonly analyzed in corpus linguistics. Furthermore, most research in sociolinguistics has focused on analyzing temporal trends of individual variables, rather than on multivariate datasets, as is common in corpus linguistics.

Part of the reason for the lack of research on modeling diachronic frequency change in corpus linguistics is due to the absence of methods for analyzing the type of multivariate lexical and grammatical frequency data generally investigated. The object of this study is therefore to

introduce and explain the use of *growth curve modeling*, a statistical method primarily used in economics and behavioral sciences for modeling multilevel data that can be applied to the analysis of multivariate diachronic frequency data in corpus linguistics. Generally speaking, this regression-based technique allows the modeling of the latent change trajectory of a number of variables in longitudinal data while accounting for the information given by the trajectory of each variable. Within corpus linguistics, growth curve modeling allows the analysis of the mean growth trajectory of a frequency change for a bundle of linguistic variables while controlling for the variation of each individual trajectory. In the remainder of the paper growth curve modeling is first introduced and its application is then illustrated through the analysis of the trajectories of increasing and decreasing words in a diachronic corpus of Twitter.

2. Growth Curve Modeling

Growth curve modeling is a regression-based method for the analysis of longitudinal data (i.e. where the same subjects are observed repeatedly over time) commonly applied in social and behavioral sciences such as psychology and economics (Field 2009). The use of growth curve modeling allows for the estimation of a latent trajectory representing the single trajectories over time of a bundle of different observations while accounting for the clustered nature of the data (Hardy & Bryman 2014). The usefulness of such a method for corpus-based diachronic linguistic research lies in its power to test and predict the average pattern of growth or decay of the relative frequency of a bundle of words, grammatical structures, or other linguistic forms in a longitudinal corpus. The advantage of growth curve modeling over other methods is that this technique permits the testing of several types of trajectories until the one with the best fit to the

data is found. If the analyst's goal is to arrive at a simple model that summarizes, explains, and predicts the average trajectory of a bundle of variables, growth curve modeling is a technique that returns an output far more precise than other statistical means, such as plotting the mathematical average values of the bundle at each time point. This increased precision is achieved by controlling the variance across the several single trajectories while calculating the trajectory that best represents the whole bundle.

Growth curve modeling has not been applied in corpus linguistics research that analyzes changes in frequency of linguistic features and, more generally, it is still uncommon in linguistics. An area of linguistics in which growth models have seen some use is psycholinguistics. However, in this strand of research the variable being predicted is the speed of learning of a linguistic structure as opposed to its development and change in time (e.g. Mirman et al. 2008). In terms of the application of growth curve modeling to language change, recently Winter & Wieling (2016) have pioneered the use of growth curve modeling to study patterns of linguistic evolution in data sets of iterated learning experiments.

The application of growth curve modeling in diachronic corpus linguistics can be illustrated by considering a hypothetical example. Let us imagine that a linguist has measured the relative frequencies of numerous words or other linguistic forms longitudinally in a corpus—for example across several days, months, or years—in order to identify the overall trajectory of this bundle over time, such as whether the relative frequencies of this set of features has accelerated or whether their growth has slowed at some point or whether their increase has been constant. To tackle this problem, the linguist could plot the relative frequency of one of these linguistic

features over the time on a graph, producing a line that fluctuates and changes, perhaps identifying a relatively clear trend, if a diachronic pattern does indeed exist. The linguist could then add lines for more and more of the features under analysis to the same graph in an attempt to identify a general pattern of change in that set of features. Unfortunately, this process would likely result in a noisy graph characterized by a multitude of lines at different scales, each with its own idiosyncrasies, thus making the interpretation of a general trend impossible. The power of growth curve modeling lies in its ability to test and provide estimates for the coefficients of the overall latent trajectory that best fits the data, while also controlling for the differences among the trajectories of each single item of the bundle. In other words, growth curve modeling is a statistical method for fitting a line or curve in order to estimate growth trajectories of a set of outcome variables modeled as a function of time. As there are several different regression techniques that can be used to fit a line or curve, there thus exist several ways of performing growth curve modeling. In particular, this paper performs growth curve modeling through multi-level linear regression using maximum likelihood estimation.

A standard linear regression is a statistical method that derives the equation of the line that best fits the input dataset and that thus allows the prediction of a *response variable* using some set of *predictors*. A *multi-level* linear regression, sometimes also called *mixed-model* regression, is a linear regression that also accounts for the idiosyncratic behavior of different observations, so that a latent pattern can be extracted more accurately (Snijders & Bosker, 1999). This technique has been recently proposed as a more efficient statistical method for many linguistic applications. For example, multi-level linear regression has been applied to the study of

diachronic morphological change (Gries 2013) and in the creation of models that can predict native-speakers behavior (Gries & Deshors 2014). Multi-level regression has also been adopted to improve the application of variable rule analysis in sociolinguistics (Johnson 2009). The use of multi-level linear regression for growth curve modeling is another powerful application of this statistical technique in linguistics.

Growth curve models estimated using multi-level regressions are composed of two parts. The first part, usually referred to as *fixed effect*, represents the latent trajectory of the time trend underlying the entire bundle of observations. This may be used to represent the trajectory of frequency change across time for the bundle as a whole. The second part, defined as *random effect*, represents the variance for each individual observation around the value given by the fixed effect, which can be estimated for both the intercept and the slopes of the trajectories. Taken together, these two elements allow for the characteristics and patterns of change of the entire bundle of individual trajectories over time to be captured jointly.

A multi-level regression used to perform a growth curve modeling for a set of linguistic variables predicts *relative frequency* using *time* as predictor. Such a regression model can be represented formally using the following equation:

$$f = \quad Ct + z \qquad + ut + zu$$

$$\textit{fixed effect} \quad \textit{random effect}$$

where *f* is the relative frequency of any item of the bundle, *t* is time in whichever scale it was measured, and *z* is the value of the intercept, that is the value of frequency when time is equal to

0. The equation above is broken down into its fixed effect and random effect. The random effect involves the term *u,* which represents a parameter that accounts for the idiosyncratic variation between the single trajectories. Once the regression estimates the coefficients of such an equation, the fixed part of the equation can be used to draw the line that best fit the growth trajectory of the bundle of observations as a whole.

It is common practice in growth model analysis to test for three different linear and curvilinear time trends: linear growth, quadratic growth, and cubic growth. More complex polynomials including powers over four can also be fitted. However, the use of polynomials of degree higher than three often leads to models that are more complex but that do not add significant amount of new information (Field 2009). Using the same notation as above, this procedure can be formalized in the use of, respectively, these three equations:

$$f = \qquad\qquad\quad Ct + z \quad + ut + zu \qquad \textit{linear growth}$$

$$f = \qquad Bt^2 + Ct + z \quad + ut + zu^1 \quad \textit{quadratic growth}$$

$$f = At^3 + Bt^2 + Ct + z \quad + ut + zu \qquad \textit{cubic growth}$$

$$\textit{fixed effect} \qquad \textit{random effect}$$

If the first equation is found to be the best fit for the data , then this suggests that the bundle follows a linear growth, i.e. the average trajectory for the bundle is a straight line corresponding

---

1   In the mixed model framework, it is also possible to estimate random slopes for the quadratic and cubic effects but this comes with a significant increase in the complexity of the model and the within-variation in the data is often insufficient to calculate the covariance for all different combinations. However, it is generally possible to include random slopes only for the linear term without significant loss of efficiency in the model (Rabe-Hesketh and Skrondal, 2012).

to increasing or decreasing constant rate of change in time, depending on the sign of the coefficient *C*. Alternatively, if the second equation is found to be the best fit, then the bundle follows a quadratic growth, i.e. the average trajectory for the bundle contains a curve with variable degrees of steepness and corresponding to an acceleration or deceleration, depending on the sign and value of the coefficient *B*. Finally, if the third equation is found to be the best fit, then the bundle follows a cubic growth, i.e. the average trajectory for the bundle behaves quadratically until a further curve occurs, which can correspond to an acceleration or deceleration with variable degrees of steepness, depending on the value and sign of the coefficient *A*. The growth curve modeling used for exploring the latent trend of a dataset essentially consists in examining the significance, the sign, and the value of the coefficients estimated to understand the dynamics of change of the data in time. In the remainder of this paper, this methodology is further demonstrated through an application on lexical change in a dataset of American Twitter.

3. Case study

To demonstrate the application of growth curve modeling in diachronic corpus linguistics, an analysis of the dynamics of the growth and decay of word frequencies is presented below. As explained above, more research is needed on the modeling of the dynamics of frequency change for linguistic features, including on the dynamics of word frequency change. This case study makes a first step toward filling these gaps by using growth curve modeling to discover the general trajectories of change taken by words that are increasing and decreasing in frequency in a multi-billion word corpus of American Tweets. In addition to demonstrating the application of

growth curve modeling in diachronic corpus linguistics, this case study also uncovers models of language change that inform our understanding of the dynamics of word frequencies.

3.1 *Data*

The data used for this analysis consists of a corpus of American English Tweets produced from 10 October 2013 to 22 November 2014, which was collected at the University of South Carolina using the Twitter API (see Grieve et al. 2015, 2016, Huang et al. 2015, Wieling et al. 2016). The corpus includes approximately 1 billion tweets written by 7 million users totaling 8.9 billion words. Due to some technical problems, for 16 days data could not be collected. The final number of days studied was therefore 392, each one containing on average 22 million word tokens. In addition, this corpus contains only English language tweets that were geocoded with a longitude and latitude that falls within the contiguous United States, because this corpus was compiled with the intention of studying geolinguistic variation.

The corpus was analyzed by first extracting all word types from the corpus. For this study, a word type is defined as an orthographic string of characters divided from other word types by white space or punctuation marks. The set of word types included in the analysis thus consists of many strings of characters that might not be considered as standard English words, such as acronyms (e. g. *tbh* for *to be honest*), creative spellings (e. g. *b* for the verb *be*), abbreviations (e. g. *bro* for *brother*), or interjections (e. g. *ohh, ahah*). These strings, however, are distinct and often very common forms in this particular variety of language and are therefore included in this analysis, although a similar analysis of a corpus representing this variety of language or another variety of language could be repeated using a different set of features.

To focus only on the most common words with a reliable number of observations, a minimum frequency threshold of at least 1 million total occurrences in the whole corpus was chosen, which resulted in a sub-sample of 12,580 word types. For each of these words, the relative frequency per billion words was calculated for each day by dividing the number of occurrences of a word by the total number of word tokens for the day and multiplying by 10^9 in order to control for variation in sample size. In this way, each word type was represented by a time series consisting in word frequency per billion words evolving in time from day 1 to day 392.

To answer the research question regarding the description of the dynamics of words that are increasing or decreasing in the corpus, the sample of 12,580 word types was filtered to identify forms that showed a clear increase or decrease over the course of the year. To identify these words, a Spearman correlation test was performed by correlating the rank of the relative frequency of each of the 12,580 words of the sample against the rank of the 392 days, following the procedure outlined in Grieve et al. (2016). If the Spearman correlation coefficient for a word was greater than 0.7 then the word was classified as increasing. Likewise, if the coefficient for a word was smaller than -0.7 then the word was classified as decreasing. The value of ±0.7 was chosen as it is a common threshold in correlation analysis used to identify strong correlations. Other values could have been used, however using a more conservative cutoff such as 0.8 or 0.9 would have resulted in far fewer words to examine. This operation of filtering is by no means necessary for the growth curve modeling. However, the filtering of the increasing and decreasing

words is essential for the present study as its aim is to understand the dynamics of change of these two types of words.

After applying the Spearman correlation test in the way described, two groups of words were thus isolated: a set of 344 increasing words and a set of 345 decreasing words, the analysis of which is described in the section below.

3.2 *Analysis*

The first set of words examined is the 344 increasing words. When these words are plotted on the same graph the interpretation of the underlying trend is largely unintelligible, aside from the fact that the words are all clearly increasing in frequency, which is already known given the process through which they were selected. This problem is illustrated in Fig. 1, which shows the top 50 increasing words plotted on the same graph against time.
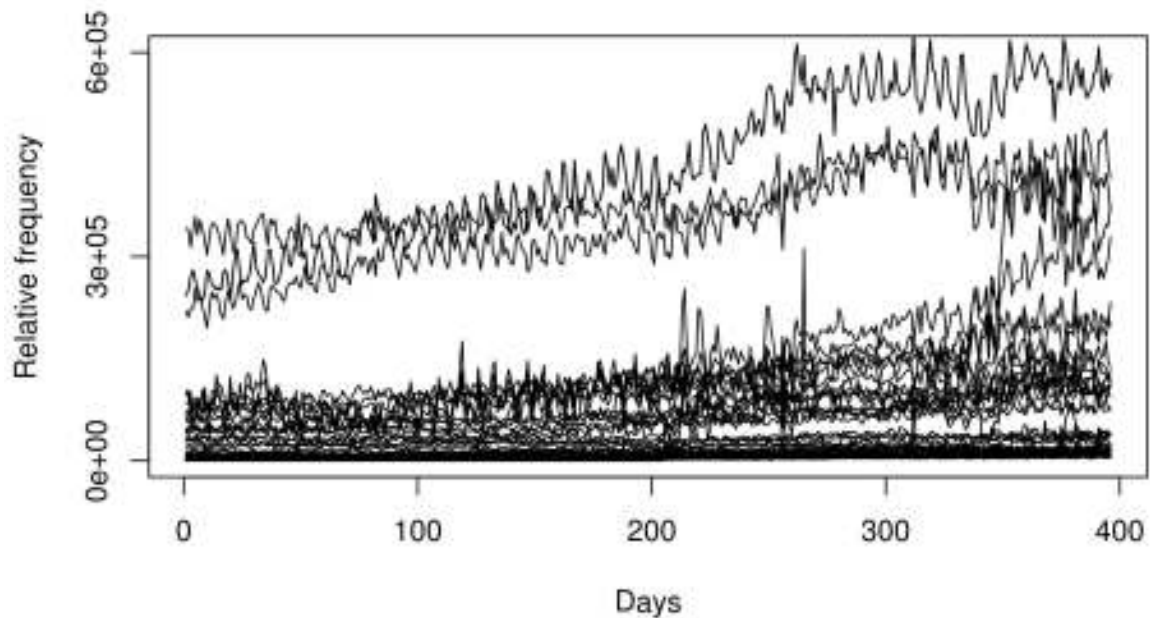
Figure 1: Frequency of the top 50 increasing words plotted on the same graph against time

Although it is possible to perceive a pattern, even by limiting the set to the top 50 increasing words out of 344, this is far from clear. Furthermore, the graph reveals how different words vary within different frequency bands. With the growth curve modeling analysis reported below, the latent trend underlying these trajectories can be estimated, also accounting for the fact that words are changing in time at different frequency bands.

The application of growth curve models to the 344 increasing words can be summarized by three tables and their relative models that are reproduced in Table 1 below. As explained above, the analysis of this set of words starts with fitting a multi-level regression model to predict *relative frequency* using *time*, which is reported in Table 1a. The χ^2 statistic of this model is

significant, indicating that the model as a whole can be considered accurate. However, because

the words analyzed were selected due to their positive correlation with time, it is not surprising

that time significantly predicts relative frequency or that the coefficient for *time* has a positive

sign, indicating that as time increases so does relative frequency. The values *SD cons* and *SD*

time represent the random effect parts of the equation and indicate the standard deviation of the

intercept and the standard deviation of the slope. The relatively high standard deviations obtained

indicate that there is considerable variation across the words, as seen in Fig. 1. As explained

above, the fixed part of the equation can be used to represent the latent trajectory in the data,

which is plotted in a scatter plot next to Table 1a. The plot shows time in days on the x axis and

the relative frequency on the y axis.

    After carrying out the regression with only time as predictor, the next step of the growth

model analysis is to add *time²* as a predictor to the equation in addition to *time* to verify whether

there is a significant quadratic term that contributes to the modeling of relative frequency

increasing in time. Adding a quadratic term amounts to testing for the presence of a curve in the

latent trajectory of the increasing words, either facing up or down and with varying degrees of

steepness depending on the value of the quadratic coefficient. The statistics reproduced in Table

1b show that a quadratic effect is present. The χ^2 statistic of the quadratic model is significant

and the significant *p*-value of the *time²* term suggests that the average trajectory of increasing

words curves at some point in time. Because the quadratic term is positive, this curve is facing

upwards. The random effects again indicate that there is considerable variation for intercepts and

slopes. Similarly to the result of the linear analysis, it is possible to produce the visualization of

the quadratic model in the graph next to Table 1b by using the coefficients and the constant resulting from the regression.

The final step of the growth curve modeling as applied in this work is the fitting of the cubic growth model, performed by adding *time³* to the analysis. The statistics reported in Table 1c indicate that a cubic effect is also present. The χ^2 statistic for the cubic model is again significant and the results reveal that the cubic term is a significant predictor of relative frequency increase. The cubic coefficient is negative, thus pointing to the presence of a point in which the curve changes direction and starts decreasing. The quadratic term of this model is, instead, positive, indicating a curve facing upwards. Finally, the linear coefficient for this model is negative, which suggests that the average line for the cubic model of increasing words starts with a decrease. In terms of the random effects, similar conclusions can be drawn as for the other two models. Overall, these results mean that if a cubic model is fitted, the increasing words are best modeled as growing slowly at the beginning, then having an acceleration in increase and then slowly increasing again until almost declining. The cubic model thus indicates that relative word frequency behaves in a slow-fast-slow pattern similar to the s-shaped curve of change that has been extensively found in sociolinguistics. Plotting the graph as before shows clearly how to the model translates visually.

Increasing words (N = 344, N obvs = 134796)		
Linear model (a)		

	Coefficient	P-value
time	71.88	< 0.01
constant	32190.25	< 0.01
SD cons	145062.6	
SD time	136.55	
χ^2	95	< 0.01
AIC	3010648	

Quadratic model (b)		

	Coefficient	P-value
time	47	< 0.01
$time^2$	0.06	< 0.01
constant	33886.33	< 0.01
SD cons	145047.6	
SD time	136.52	
χ^2	351.65	< 0.01
AIC	3010393	

Cubic model (c)		

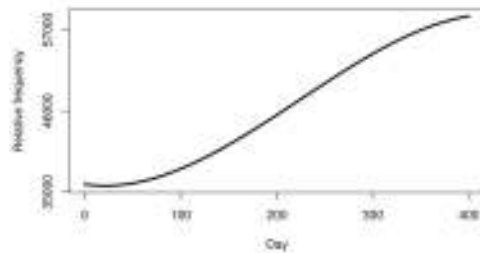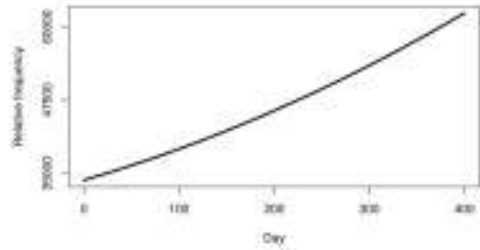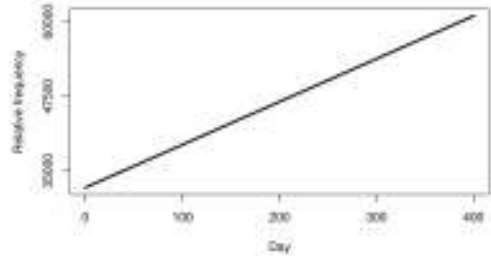|	Coefficient	P-value
time	-19.06	0.023
$time^2$	0.47	< 0.01
$time^3$	-0.0007	< 0.01
constant	36055.19	< 0.01
SD cons	145066.8	
SD time	136.55	
χ^2	678.85	< 0.01
AIC	**3010069**	

Table 1(a, b, c): Multi-level linear regression models for increasing words

If no other polynomials are fitted, the growth model analysis concludes with the interpretation of the results and the selection of the best model. A standard measure to compare goodness of fit among competing models to select the best fitting model is Akaike's Information Criterion (AIC), reported in each table in the AIC rows (Sakamoto et al. 1986). Using this model selection method, the model with the lowest AIC is the model that best fits the data. The best overall model for the increasing words considered for this analysis is therefore the cubic model. This result is also supported by likelihood ratio tests used to test the contributions of the each additional coefficient: the addition of the quadratic coefficient significantly improves the linear model (LR $\chi^2 = 256.36$, p < 0.001), as does the addition of the cubic coefficient to the quadratic model (LR $\chi^2 = 326.22$, p < 0.001).

In order to explore how the model relates to the data, Fig. 2 below shows the scatterplots of the top eight rising words in the corpus: *fuckboy, rn, timehop, fw, ft, sm, squad,* and *asf*.
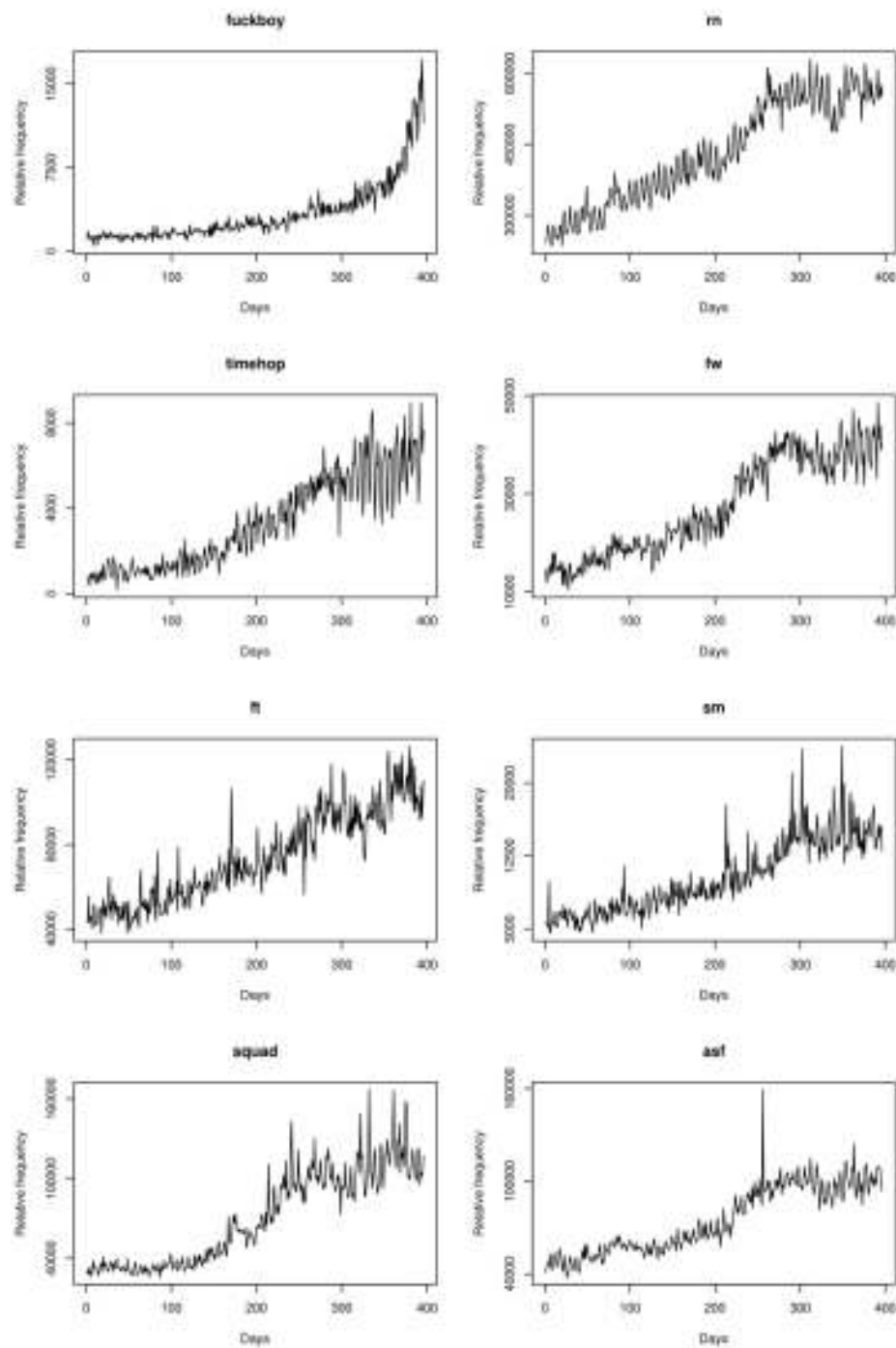
Figure 2: Scatterplots for top eight increasing words

19

Although a glance at these top rising words does suggest an s-shaped cubic pattern, it is evident that not all the words follow the same trajectory and that not all of them follow a cubic trend. The large variation that is present in the data is reflected statistically in the large standard deviations reported above. The usefulness of growth curve modeling however lies in the possibility of extracting a general latent pattern that would otherwise be unnoticeable by examining a small data sample such as in Fig. 2. The cubic pattern of increase shown in Table 1c is the common trajectory that best summarizes the behavior of all of the 344 increasing words.

Given the method explained above, the same growth curve analysis can now be applied to the set of decreasing words. The results of the analysis is in Table 2a,b,c below.

Decreasing words (N = 345, N obvs = 137050)		

Linear model (a)

	Coefficient	P-value
time	-130.42	< 0.01
constant	309847.6	0.03
SD cons	2599862	
SD time	840.36	
χ^2	8.29	< 0.01
AIC	3504926	

Quadratic model (b)

	Coefficient	P-value
time	-52.83	0.25
$time^2$	-0.19	< 0.01
constant	304675.3	0.03
SD cons	2599860	
SD time	840.36	
χ^2	110.22	< 0.01
AIC	3504826	

Cubic model (c)

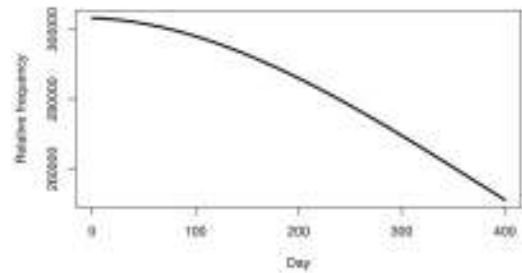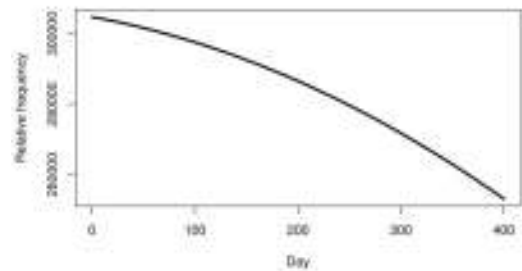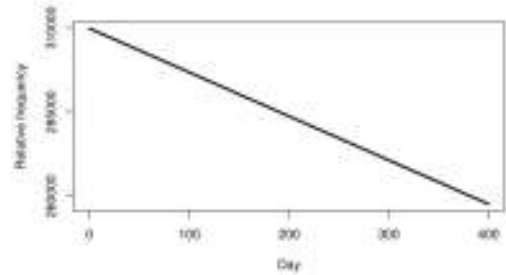|	Coefficient	P-value
time	-10.42	0.83
$time^2$	-0.46	< 0.01
$time^3$	0.0004	0.02
constant	303253.5	0.03
SD cons	2599861	
SD time	840.36	
χ^2	115.62	< 0.01
AIC	**3504822**	

Table 2 (a, b, c): Multi-level linear regression models for decreasing words.

For the decreasing words the linear model is again appropriate, and this is the result of the fact that the words were selected only if they had a negative relationship with time. In the case of the quadratic model, the analysis reveals a decreasing trend that is similar to the linear model. While the negative coefficient for the quadratic term seems to suggest a curve facing downwards, as visualized in the line plot in Table 2b, the *p*-value of the linear coefficient is, however, not significant, meaning that it is not possible to reject the hypothesis that the initial decline may in fact be null. Finally, the cubic model reveals a decreasing pattern characterized initially by a decelerated trend followed by an acceleration in the decline. The coefficient of the first term is however not significant. Furthermore, the magnitude of the cubic coefficient is quite limited, as visualized in Table 2c, indicating that this effect does not substantially change the decreasing trend from the one obtained from the quadratic model, at least within the time span from 0 to 400 words considered for this study, as can be seen from the figures of Table 2. The random effects for all the three models are similar to the random effects for increasing words, with both intercepts and slopes varying considerably across words. An examination of the AIC values suggest that the cubic model is once again the best fitting model, as do likelihood ratio tests that evaluate the contribution of the quadratic coefficient to the linear model (LR $\chi^2 =$ 101.89, p < 0.001) and of the cubic coefficient to the quadratic model (LR $\chi^2 = 5.39$, p = 0.02).

As for the analysis of increasing words, an exploration of how the model relates to the data can be performed using Fig. 3 below, which shows the top eight decreasing words in the corpus: *haha, fdb, uoeno, ooo, ratchet, ohh, yolo, cx*.
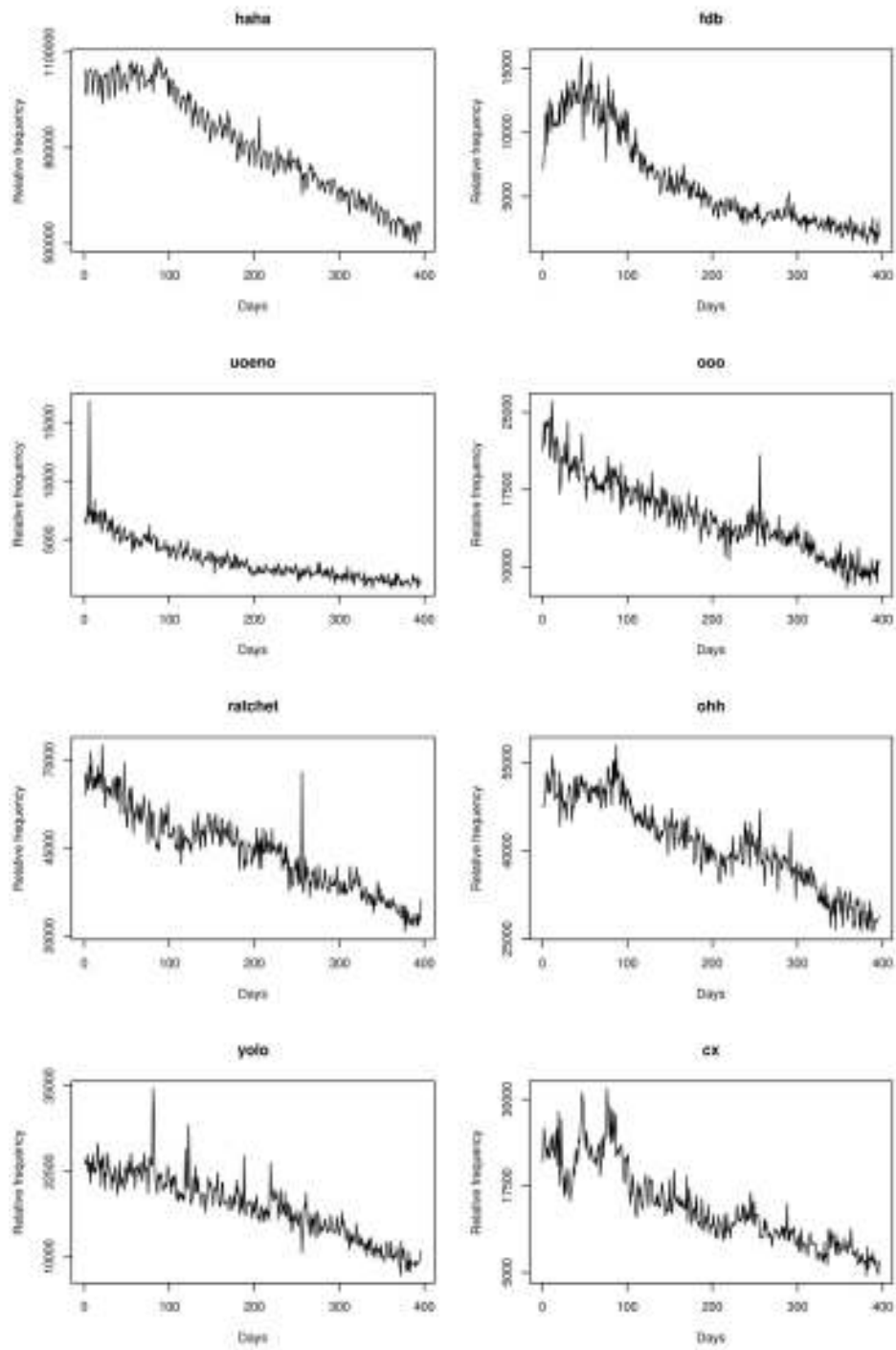
Figure 3: Scatterplots for top eight decreasing words

As for the increasing words, although most of the words in Fig. 3 above approximate the trajectory found by the growth curve model, not all of them follow exactly the same pattern, and this variety is again reflected statistically by the large standard deviations. Given the idiosyncrasies that each word exhibits, large variations are expected, as it is expected that commonalities are almost invisible to the naked eye unless a technique such as growth curve modeling is applied.

The two growth curve models thus provided describe the increase or decrease of words that were in a state of strong frequency change in American Twitter in 2014. These two models, represented by two equations that can be plotted onto a graph, are useful to visualize the latent tendency of the bundle of hundreds of words considered.

3.3 *Discussion*

In addition to demonstrating the application of growth curve modeling for diachronic corpora, the analysis of decreasing and increasing words are of relevance to theories of language change. The results have shown that words on the increase in American Twitter in 2014 follow a trajectory that is similar to the s-shaped pattern of language change already familiar to sociolinguists. Although this pattern has been repeatedly found in cases of alternations between two or more forms and occasionally in studies concerning frequencies over total number of words, this study identified a similar pattern across a large number of word frequencies measured in a corpus over real time. The model of growth for increasing words found in this study suggests that word frequency change tends to start slowly and then develop quickly until reaching a point of saturation while decelerating. In the research on sociolinguistic variables, several explanations

have been proposed for the emergence of such an s-shaped pattern for cases of competing changes. Labov (1995) suggests that it is the probability of contact between individuals who use the new variant with those who do not that explains the pattern, since this probability increases until the new variant becomes equally likely as the old variant and then it decreases, thus forming an s-shaped pattern. Alternatively, Blythe & Croft (2012) explain that the s-shape results from the fact that competing variants are unequally weighted by social factors.

The findings of the case study are in all probability accounted for by similar explanations, although the methodology here adopted consisted in the measurement of frequency relative to the total number of words as opposed to the frequency of related variants. The cubic model found demonstrates that the average trajectory of a word that is increasing in popularity firstly accelerates, and this phase corresponds to a state of diffusion in the community. However, there is a point in which this increase diminishes in power until leveling off, as if reaching saturation in the speech community. As recently proposed for emergent words, in accordance with the Verhulstian model of population dynamics, the maximum frequency of a word corresponding to the top of the s-curve is likely to correspond to the semantic carrying capacity of the word, or the maximum frequency of use of that form and of all the other forms that cover the same semantic space (Grieve et al. 2016). Since the words examined are not novel or emergent forms, the case study reported in the present paper provides evidence that a similar mechanism of diffusion also applies to words already established in a community.

Another important finding of the present study that concerns theoretical diachronic linguistics and language change is that the way relative word frequency increases is not exactly

symmetrical to the way relative word frequency decreases. On the one hand, the average increase of a word tends to be an s-shaped cubic pattern. On the other hand, the average decrease tends to approximate a decelerated trajectory. This asymmetry is not easy to explain in relation to previous research, as not enough research has been dedicated to the decrease in frequency of linguistic items. However, similar effects have been recently noted by Maslennikova et al. (2015), who found that 'the phase of decrease in the frequency is longer than the phase of increase in the frequency' for 500 billion words of the Google N-gram database. Similar asymmetry has been observed by Krawczyk et al. (2014) studying the dynamics of baby names adoption in the US, who noted that 'the popularity [of names] increases more quickly than it decreases' (Krawczyk et al. 2014: 387). The explanation for this effect is only hinted at by these authors, who mention that when a baby name is declining this name is already well spread and popular. The explanations for this asymmetry observed are rather speculative and require more attention in future research.

Overall, in terms of the general pattern of development of word frequency increase and decrease, it is possible to draw interesting parallelism between the dynamics of word frequency change derived by the growth curve models of the present study and the patterns of developments of products according to the marketing concept of the Product Life Cycle (Armstrong et al. 2015; Hunt 2010). According to this model, typically new products develop in time following four stages: *introduction, growth, maturity*, and *decline.* A product starts its life very slowly, then increases rapidly until a point of saturation after which it faces a slow decline. The growth models found in the present study for both increasing and decreasing words

combined reveal a similar type of behavior. This similarity could suggest possible theoretical connections between models of market change and language change requiring further investigation.

4. Conclusion

As well as making first steps towards a better understanding of the dynamics of lexical frequency change, the main goal of this paper is to introduce growth curve modeling for the analysis of large longitudinal diachronic linguistic datasets.

The present study has demonstrated how growth curve models can be used to capture the latent trajectory of a large bundle of linguistic features that are changing in time by estimating and testing the trend representing their tendencies that best fit the data. If an analyst is confronted with a longitudinal dataset of frequency data and their goal is to understand the dynamics of a category or bundle of features, growth curve modeling can help reveal such patterns.

Since the model returned by a growth curve modeling is in the form of an equation, the model can be easily plotted onto a scatter plot to be visualized. This method is therefore an excellent tool for diachronic research, providing a statistical tool to analyze the latent trend of a bundle of feature that is more powerful than summing up the frequencies in one category. Although the method was applied to relative word frequency change, the frequency of any kind of linguistic structure can be used instead. For example, bundles of relative frequencies of grammatical structures or of morphological items can also be used instead of lexical frequencies. The frequency of a variant over the frequency of another variant can be used instead of frequencies over the total number of words in a day.

Although growth curve models as described in this paper are powerful and flexible tools, they do present limitations. For example, polynomials of the kind explored in the present work are only an approximation of the dynamics of the relative frequencies considered, since the proposed models assume that the predicted variable is unbounded when, in fact, relative frequency is bounded (i. e. can only vary from 0 to 1) and the response data may approach an asymptote. Although for the sake of the present study this limitation is less important, as the focus of the study is the understanding of the shape of the latent trajectories, researchers dealing with relative frequencies who are are primarily interested in more flexible and precise curve fitting properties may adopt other techniques that explicitly account for bounded data. In addition, further improvements in modeling reliability, for example in case of variables that present many fluctuations, can also be adopted with more sophisticated techniques such as Generalized Additive Modeling (Winter & Wieling 2016). However, with its balance between power and simplicity of application as well as interpretation, growth curve modeling lends itself well to become a standard technique for the analysis of trends of change in diachronic corpora.

References

Armstrong, Gary, Philip Kotler, Michael Harker, and Ross Brennan. 2015. *Marketing: An Introduction*. Harlow, England: Pearson.

Baker, Paul, Costas Gabrielatos, and Tony McEnery. 2013. Sketching Muslims: A corpus driven analysis of representations around the word 'Muslim' in the british press 1998-2009. *Applied Linguistics* 34: 255–278.

Baker, Paul, Costas Gabrielatos, Majid Khosravinik, Michał Krzyżanowski, Tony McEnery, and Ruth Wodak. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society* 19: 273–306.

Biber, Douglas and Edward Finegan. 1989. Drift and the evolution of English style: A history of three genres. *Language* 65: 487–517.

Biber, Douglas. 2004. Historical patterns for the grammatical marking of stance: A cross-register comparison. *Journal of Historical Pragmatics* 5: 107–136.

Biber, Douglas, and Jená Burges. 2000. Historical change in the language use of women and men: Gender differences in dramatic dialogue. *Journal of English Linguistics* 28: 21–37.

Biber, Douglas, and Bethany Gray. 2013. Being specific about historical change: The influence of sub-register. *Journal of English Linguistics* 41: 104–134. doi:10.1177/0075424212472509.

Blythe, Richard, and William Croft. 2012. S-curves and the mechanisms of propagation in language change. *Language* 88: 269–304. doi:10.1353/lan.2012.0027.

Chambers, Jack K. 2001. Patterns of variation including change. In Jack K. Chambers and Natalie Schilling (eds), *The Handbook of Language Variation and Change*, 358–361. Malden, MA: Blackwell Publishers.

Denison, David. 2003. Log(ist)ic and simplistic S-curves. In Raymond Hickey (ed.), *Motives for Language Change*, 54–70. Cambridge, New York: Cambridge University Press

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2012. Mapping the geographical diffusion of new words. *arXiv:1210.5268 [cs.CL]*, 1–13. http://arxiv.org/abs/1210.5268.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLOS one* 9.

Field, Andy. 2009. *Discovering Statistics Using SPSS*. London: SAGE.

Gries, Stefan Th. 2013. Sources of variability relevant to the cognitive sociolinguist, and corpus - as well as psycholinguistic methods and notions to handle them. *Journal of Pragmatics* 52: 5–16.

Gries, Stefan Th., and Martin Hilpert. 2008. The identification of stages in diachronic data: Variability-based neighbour clustering. *Corpora* 3: 59–81. doi:10.3366/E1749503208000075.

Gries, Stefan Th., and Sandra C. Deshors. 2014. Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora* 9: 109–136. doi:10.3366/cor.2014.0053.

Grieve, Jack, Andrea Nini, Diansheng Guo, and Alice Kasakoff. 2015. Recent changes in word formation strategies in American social media. Presented at *Corpus Linguistics 2015*. Lancaster University: Lancaster.

Grieve, Jack, Andrea Nini, Diansheng Guo. 2016. Analyzing lexical emergence in American English online. Forthcoming in *English Language and Linguistics*.

Hardy, Melissa and Alan Bryman. 2014. *Handbook of Data Analysis*. London: Sage.

Hilpert, Martin. 2011. Dynamic viualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora. *International Journal of Corpus Linguistics* 16: 435–461. doi:10.1075/ijcl.16.4.01hil.

Hilpert, Martin, and 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. doi:10.1093/llc/fqn012.

Huang, Yuan, Diansheng Guo, Alice Kasakoff, Jack Grieve. 2015. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*.

Hundt, Marianne. 2014. *Late Modern English Syntax*. Cambridge: Cambridge University Press.

Hunt, Shelby D. 2010. *Marketing Theory: Foundations, Controversy, Strategy, Resource-Advantage Theory*. Abingdon: Routledge.

Johnson, Daniel E. 2009. Getting off the GoldVarb Standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3: 359–383.

Krawczyk, J. Małgorzata, Antoni Dydejczyk, and Krzysztof Kułakowski. 2014. The Simmel Effect and babies names. *Physica A: Statistical Mechanics and Its Applications* 395: 384–391. doi:10.1016/j.physa.2013.10.018.

Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1, 199-244.

Labov, William. 1978. *Sociolinguistic Patterns*. Oxford: Blackwell.

Labov, Williams. 1995. *Principles of Linguistic Change - Volume I: Internal Factors*. Oxford: Blackwell.

Labov, Williams. 2001. *Principles of Linguistic Change - Volume II: Social Factors*. Oxford: Blackwell.

Maslennikova, S. Yulia, Vladimir V. Bochkarev, and Inna A. Belashova. 2015. Cluster analysis of word frequency dynamics. *Journal of Physics: Conference Series* 574: 012120. doi:10.1088/1742-6596/574/1/012120.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331: 176–182. doi:10.1126/science.1199644.

Millar, Neil. 2009. Modal verbs in TIME: Frequency changes 1923–2006. *International Journal of Corpus Linguistics* 14: 191–220. doi:10.1075/ijcl.14.2.03mil.

Mirman, Daniel, James A. Dixon, and James S. Magnuson. 2008. Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language* 59: 475–494. doi:10.1016/j.jml.2007.11.006.

Nevalainen, Terttu and Helena Raumolin-Brunberg. 2003. *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. Edinburgh: Pearson.

Petersen, Alexander M., Joel N. Tenenbaum, Shlomo Havlin, H. Eugene Stanley, and Matjaž Perc. 2012a. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific Reports* 2: 1-10. doi:10.1038/srep00943.

Petersen, Alexander M., Joel Tenenbaum, Shlomo Havlin, and H. Eugene Stanley. 2012b. Statistical laws governing fluctuations in word use from word birth to word death. *Scientific Reports* 2: 1–9. doi:10.1038/srep00313.

Rabe-Hesketh, Sophia and Anders Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata*. Texas: Stata Press.

Säily, Tanya, Harri Siirtola, and Terttu Nevalainen. 2011. Variation in noun and pronoun frequencies in a sociohistorical corpus of English. *Literary and Linguistic Computing* 26 (2): 167-188.

Sakamoto, Yosiyuki, Makio Ishiguro, and Genshiro Kitagawa. 1986. *Akaike Information Criterion Statistics*. Tokyo Dordrecht; Boston Hingham, MA: KTK Scientific Publishers.

Siemund, Peter. 2014. The emergence of English reflexive verbs: An analysis based on the Oxford English Dictionary. *English Language and Linguistics* 18: 49–73. doi:10.1017/S1360674313000270.

Snijders, Tom A. B. and Roel J. Bosker. 1999. *An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.

Wieling, Martijn, Jack Grieve, Gosse Bouma, Josef Fruehwald, John Coleman and Mark

Liberman. 2016. Variation and change in the use of hesitation markers in Germanic

languages. Forthcoming in *Language Dynamics and Change.*

Winter, Bodo & Martijn Wieling. 2016. How to analyze linguistic change using mixed models,

Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language*

Evolution 1(1), 7-18.