Article:
Ruck, Damin, Bentley, R. Alexander, Acerbi, Alberto et al. (2017) ROLE OF NEUTRAL EVOLUTION IN WORD TURNOVER DURING CENTURIES OF ENGLISH WORD POPULARITY. Advances in Complex Systems. ISSN: 1793-6802

https://doi.org/10.1142/S0219525917500126

Role of neutral evolution in word turnover during centuries of English word popularity

Damian Ruck[1,3], R. Alexander Bentley[2,3,*], Alberto Acerbi[4], Philip Garnett[5], Daniel J. Hruschka[6],

1 School of Social and Community Medicine, University of Bristol, UK
2 Anthropology Department, University of Tennessee, USA
3 Hobby School of Public Affairs, University of Houston, USA
4 Eindhoven University of Technology, Netherlands
5 York Management School, University of York, UK
6 School of Human Evolution and Social Change, Arizona State University, USA

* rabentley@uh.edu

Here we test Neutral models against the evolution of English word frequency and vocabulary at the corpus scale, as recorded in annual word frequencies from three centuries of English language books. Against these data, we test both static and dynamic predictions of two neutral models, including the relation between corpus size and vocabulary size, frequency distributions, and turnover within those frequency distributions. Although a commonly used Neutral model fails to replicate all these emergent properties at once, we find that modified two-stage Neutral model does replicate the static and dynamic properties of the corpus data. This two-stage model is meant to represent a relatively small corpus of English books, analogous to a 'canon', sampled by an exponentially increasing corpus of books among the wider population of authors. More broadly, this model—a smaller neutral model within a larger neutral model—could represent more broadly those situations where mass attention is focused on a small subset of the cultural variants.

Introduction

English has evolved continually over the centuries, in the branching off from antecedent languages in Indo-European prehistory [35, 40], in the rates of regularisation of verbs [35] and in the waxing and waning in the popularity of individual words [3, 13, 38]. At a much finer scale of time and population, languages change through modifications and errors in the learning process [14, 28].

This continual change and diversity contrasts with the simplicity and consistency of Zipf's law, by which the frequency a word, f, is inversely proportional to its rank k, as $f \sim k^{-\gamma}$ and Heaps law, by which vocabulary size scales sub-linearly with total number of words, across diverse textual and spoken samples [33, 42, 48, 51, 15, 22, 50, 44].

The Google Ngram corpus [38] provides new support for these statistical regularities in word frequency dynamics at timescales from decades to centuries [23, 42, 44, 1, 29]. With annual counts of n-grams —an n-gram being n consecutive character strings, separated by spaces —derived from

millions of books over multiple centuries [36], the n-gram data now covers English books from the year 1500 to year 2008.

Further research on common words and phrases made possible by the n-gram data demonstrates the "Matthew effect" of stochastic proportional growth, which has been observed in a range of natural, biological and socio-cultural realms [43]. In English, the Zipf's law in the n-gram data [42] exhibits two regimes: one among words with frequencies above about 0.01% (Zipf's exponent $\gamma \approx 1$) and another ($\gamma \approx 1.4$) among words with frequency below 0.0001% [44]. The latter Zipf's law exponent γ of 1.4 is equivalent to a probability distribution function (PDF) exponent, α, of about 1.7 ($\alpha = 1 + 1/\gamma$).

While the well-known Zipf's law demonstrates a necessary but incomplete characterization of stochastic proportional growth, a more complete characterization requires analyzing change in time-resolved data [43]. In this respect, word frequency data have at least two other statistical properties. One, known as Heaps law, refers to the way that vocabulary size scales sub-linearly with corpus size (raw word count). The n-gram data show Heaps law in that, if N_t is corpus size and v_t is vocabulary size at time t, then $v_t \approx N_t^{\beta}$, with $\beta \approx 0.5$, for all English words in the corpus [44]. If the n-gram corpus is truncated by a minimum word count, then as that minimum is raised the Heaps scaling exponent increases from $\beta < 0.5$, approaching $\beta < 1$ [44].

The other statistical property is dynamic turnover in the ranked list of most commonly used words. This can be measured in terms of how many words are replaced through time on "Top y" ranked lists of different sizes y of most frequently-used words [12, 18, 20, 24]. We can define this turnover $z_y(t)$ as the number of new words to have entered the top y most common words in year t, which is equivalent to the the top y in that year. The plotting of turnover z_y for different list sizes y can therefore be useful in characterising turnover dynamics [2].

Many functional or network models readily yield the static Zipf distribution [22, 15, 43] and Heaps law [37], but not the dynamic aspects such as turnover. Here we focus on how Heaps law and Zipf's law can be modeled together with continual turnover of words within the rankings by frequency [4, 24]. We focus on the 1-grams in Google's English 2012 data set, which samples English language books published in any country [26].

Our overall finding is a model that can replicate observed the Google corpus, which we assume to be representative of overall language through time. Even if the Google sample is biased toward more recent texts [16], the model reveals its utility in replicating multiple dynamic properties, including growing corpus and vocabulary sizes, frequency distributions, and turnover within those frequency distributions.

Neutral models of vocabulary change

One promising, parsimonious approach incorporates the class of neutral evolutionary models [11, 12, 7, 25, 39] that are now proving insightful for language transmission [13, 10, 47]. The null hypothesis of a Neutral model is that copying is undirected, without biases or different 'fitnesses' of the words being replicated [2, 30].

A basic neutral model, which we will call the *full-sampling Neutral model* (FNM), would assume simply that authors choose to write words by copying those published in the past and occasionally inventing or introducing new words. As shown in Fig 1a, the FNM represents each word choice by an author as selecting at random among the N_t words that were published in the previous year [47, 10]. This copying occurs with probability $1 - \mu$, where $\mu \ll 1$ is the fixed, dimensionless probability that an author invents a new word (even if the word had originated somewhere 'outside' books, e.g. in spoken slang). Each newly-invented word enters with frequency one, regardless of N_t. In terms of the modeled corpus, a total of about μN_t unique new words are invented per time step. Note that N_t represents the total number of written words, or corpus size, for year t, which contrasts with the smaller "vocabulary" size, v_t, defined as the number of *different* words in each year t regardless of their frequency of usage (these terms, which we use for

generality, are equivalent to *token* and *type* in corpus linguistics, where token is the number of words in the corpus and type is the number of unique words).

As has been well demonstrated, the FNM readily yields Zipf's law [11, 9, 49], which can also be shown analytically (see Appendix 1). Also, simulations of the FNM show that the resulting Zipf distribution undergoes dynamic turnover [12]. Extensive simulations [20] show that when list size y is small compared to the corpus ($0.15y < N_t\mu$), this neutral turnover z_y per time step is more precisely approximated by:

$$z_y = 1.4 \cdot \mu^{0.55} \cdot y^{0.86} \cdot n^{0.13}, \tag{1}$$

where n is the number of words per time interval.

This prediction can be visualized by plotting the measured turnover z_y for different list sizes y. The FNM predicts the results to follow $z_y \propto y^{0.86}$, such that departures from this expected curve can be identified to indicate biases such as conformity or anti-conformity[2]. It would appear from eq. 1 that turnover should increase with corpus size. This is the nominal equilibrium for FNM with constant N_t. If corpus size N_t in the FNM is growing exponentially with time, however, then there may be no such nominal equilibrium. In this case we predict that the turnover z_y can actually decrease with time as N_t increases. This is because newly invented words start with frequency one, and under the neutral model they must essentially make a stochastic walk into the top 100, say. As N_t grows, so does the minimum frequency needed to break into the top 100. As the "bar" is raised, words are more likely to 'die' before they ever reach the bar by stochastic walk [45]. As a result, turnover in the Top y can slow down over time and growth of N_t.

The FNM does not, however, readily yield Heaps law ($v_t = N_t^\beta$, where $\beta < 1$), for which $\beta \approx 0.5$ among the 1-gram data for English [44]. In the FNM, the expected exponent β is 1.0, as the number of different variants (vocabulary) normally scales linearly with μN_t [11].

While the FNM has been a powerful null model, in the case of books, we can make a notable improvement to account for the fact that most published material goes unnoticed while a relatively small portion of the corpus is highly visible. To name a few examples across the centuries, literally billions of copies of the Bible and the works of Shakespeare have been read since the seventeenth century, as well as tens or hundreds of millions of copies of works by Voltaire, Swift, Austen, Dickens, Tolkien, Fleming, Rawling and so on. While these and hundreds more books become considered part of the "Western Canon," that canon is constantly evolving [29] and many books that were enormously popular in their time —*e.g., Arabian Nights* or the works of Fanny Burney—fall out of favour. As the published corpus has grown exponentially over the centuries, early authors were more able to sample the full range of historically published works, whereas contemporary authors sample from an increasingly small and more recent fraction of the corpus, simply due to its exponential expansion [29, 41].

As a simple way of capturing this, we propose a modified neutral model, called the *partial-sampling Neutral model* (PNM), of an evolving "canon" that is sampled by an exponentially-growing corpus of books. As shown in Fig 1b, the PNM represents an exponentially growing number of books that sample words from a fixed size canon over all previous years since 1700. Our PNM represents a world where there exists an evolving canonical literature as a relatively small subset of the world's books on which all writers are educated. As new contributions to the canon are contributed, authors sample from the recent generation of writers with occasional innovation. Because the canon is a high-visibility subset of all books, only a fixed, constant number words of text per year are allowed into a year's canon. The rest of the population learns from the cumulative canon since our chosen reference year of 1700.

Results

The average result from 100 runs in each of the FNM and PNM were used to match summary statistics with the 1-gram data. Several key statistical results emerge from analysis of the 1-gram

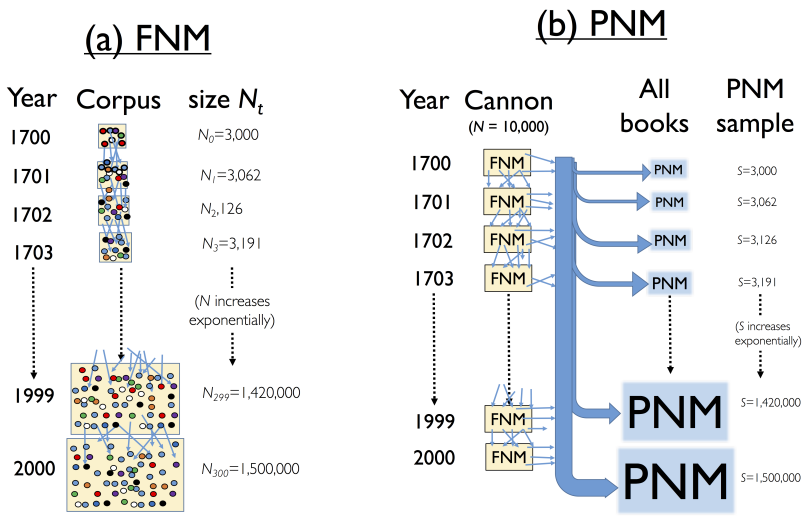

Figure 1: **Schematic representation of the Full-sampling Neutral model (FNM) and Partial sampling Neutral model (PNM). (a)** In the FNM, each of the N_t words in year t, represented by different colored circles in each box, is copied (arrows) from from the previous year $t-1$ with probability $1-\mu$, or newly-invented with probability μ. The FNM shown in (a) has a corpus size N_t that grows through time. In **(b)** the PNM samples from all previous results of the FNM since the initial time step representing year 1700. The PNM corpus grows exponentially ($N_0 e^{0.021t}$) through time, from 3000 to 1.5 million. As the PNM samples from all previous years of FNM corpus, the PNM samples from a corpus that increases linearly (by 10,000 words per year) from 10,000 words in year 1700 to 3 million words by year 2000. For the PNM, the big blue arrows represent how each generation can sample any year of the canon randomly, all the way back to 1700, the smaller arrows representing individual sampling events.

data which we compare the FNM to the PNM in terms of these results: (1) Heaps law, which is the sublinear scaling of vocabulary size with corpus size, (2) a Zipf's law frequency distribution for unique words, (3) a rate of turnover that decreases exponentially with time and a turnover vs popular list size that is approximately linear. Here we describe our results in terms of rank-frequency distributions, turnover and corpus and vocabulary size. We compare the partial-sample Neutral model (PNM) to the full 1-gram data for English.

First, we check that the model replicates the Zipf's law that characterizes the 1-gram frequencies in multiple languages [42]. Our own maximum likelihood determinations, applying available code [15] to the Google 1-gram data, confirm that the mean $\alpha = 1.75 \pm 0.12$ for the Zipf's law over all English words in the hundred years from 1700 to 1800 (beyond 1800, the corpus size becomes too large for our computation). Normalising by the word count [22], the form of the Zipf distribution is virtually identical for each year of the dataset, reaching eight orders of magnitude by the year 2000 (Fig 2a). The FNM replicates the Zipf (Fig 2b) but the PNM replicates it better and over more orders of magnitude (Fig 2c). It was not computationally possible with either the FNM or PNM to replicate the Zipf across all nine orders of magnitude, as the modeled corpus size N_t grows exponentially (Fig 2d).

Fig 3a illustrates the relationship between corpus size and vocabulary size in our partial-sampling Neutral model. Due to the exponentially increasing sample size, the ratio of vocabulary size over corpus size becomes increasingly small, thus the model gives us the sub-linear relationship described by $v_t = N_t^{\beta}$, where $\beta < 1$. On the double-logarithmic plot in Fig 3a, the

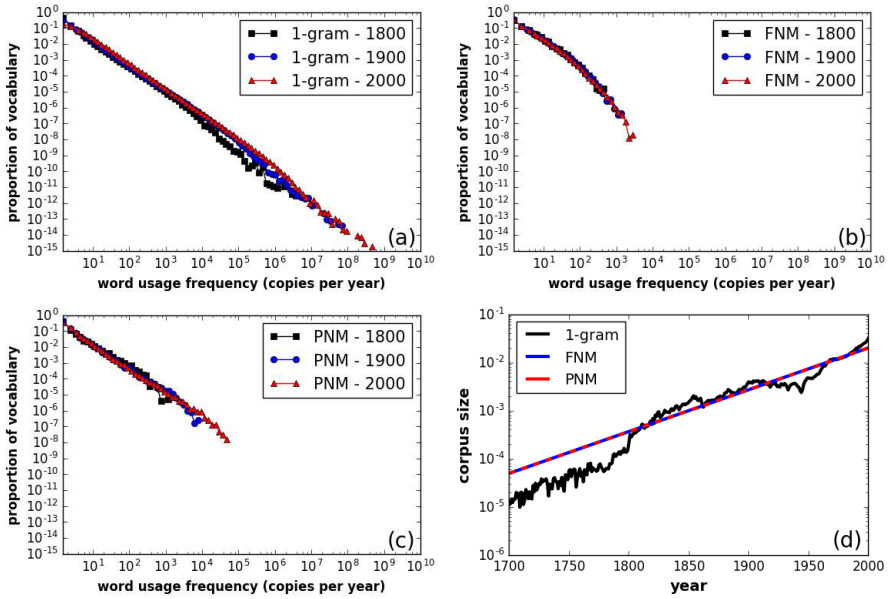

Figure 2: **Rank-frequency distributions among English words, (a)** In the 1-gram corpus. Black symbols show the distribution for the year 1800, blue shows year 1900 and red shows year 2000. The simulated results are shown for the FNM in **(b)** and the PNM in **(c)**. Panel **(d)** shows the actual number of English words, N_t in the 1-gram corpus versus the modeled corpus size $N_0 e^{0.021t}$, where t is number of years since1700.

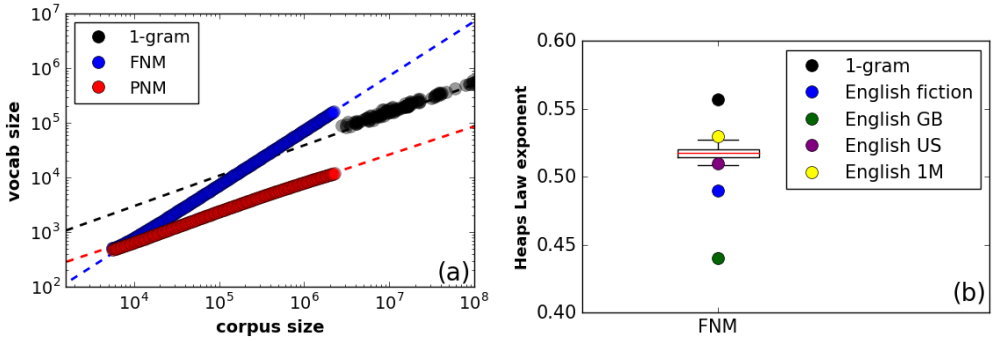

Figure 3: **Heaps law in simulated Neutral models versus 1-gram data. (a):** A double-logarithmic plot, showing corpus size versus vocabulary size, i.e. Heaps Law, for all 1-grams (black), the FNM (blue) and the PNM (red). **(b):** The Heaps law exponents, β, for the data series on the left, as well as additional data series, using Table 1 in [44]: all English 1-grams: 0.54 ± 0.01; English fiction: 0.49 ± 0.01; English GB: 0.44 ± 0.01; English US: 0.51 ± 0.01. The 100 independent runs of each neutral model, using parameters listed in the text, yielded $\beta = 0.52 \pm 0.07$ for the PNM, and $\beta = 1.00 \pm 0.002$ for the FNM (not shown).

Heaps law exponent is equivalent to the slope of the data series. The PNM matches the 1-gram data with Heaps exponent (slope) of about 0.5, whereas the FNM, with exponent about 1.0, does not. Fig 3b shows how 100 runs of the PNM yields a Heaps law exponent within the range derived by [44] for several different n-grams corpora (all English, English fiction, English GB, English US

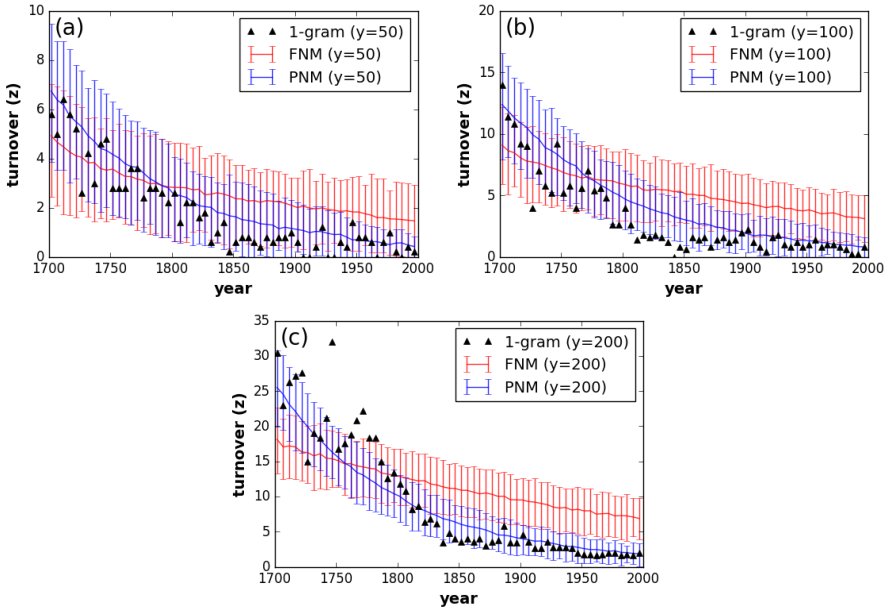

Figure 4: **Turnover decay in neutral model versus 1-gram data**, for different toplist sizes. Each panel shows the annual turnover among the ranked lists of the top y most frequently-used 1-grams, for list sizes of **(a)**$y = 50$, **(b)** $y = 100$ and **(c)** $y = 200$. The respective line and error bars in each color represent the range of FNM and PNM simulation results. Bands indicate 95% range of simulated values.

and English 1M). The PNM yields Heaps law exponent $\beta \approx 0.52 \pm 0.006$, within the range of English corpora, whereas the FNM yields a mismatch with the data of $\beta \approx 1 \pm 0.002$ (Fig 3b).

In Fig 3a, there is a constant offset on the y-axis between vocabulary size in the PNM ($\alpha = 0.02$, $N = 10000$) versus the 1-gram data. Both data series follow Heaps exponent $b \approx 0.5$, but the coefficient, A, is several times larger for the 1-gram data than for the PNM. We do not think this is due to our choice of canon size N in the PNM, because if we halve it to 5000, the resulting A does not significantly change. The difference could be resolved, however, with larger exponential growth in PNM corpus size, S_t, over the 300 time steps. Computationally, we could only model the PNM with growth exponent $\alpha = 0.02$—using $\alpha = 0.03$, as would fit the actual growth of the n-gram corpus over 300 years [8], makes the PNM too large to compute. Nevertheless, we can roughly estimate the effect; when we reduce α from 0.02 to 0.01, while keeping $N = 10000$, we find that A averaged over one hundred PNM runs is reduced from 6.3 ± 0.5 to 1.4 ± 0.3. Given an exponential relationship, increasing alpha to 0.03 would increase A to about 20, which is within the magnitude of offset we see in Fig 3a. Of course, this question can be resolved precisely when the much larger PNM can be simulated.

Regarding dynamic turnover, we consider turnover in ranked lists of size y, varying the list size y from the top 1000 most common words down to the top 10 (the top word has been "the" since before the year 1700). We measure turnover in the word-frequency rankings by determining the top y rankings independently for each year, and then counting the number of new words to appear on the list from one year to the next. Fig 4 shows the number of 1-grams to drop out of the top 1000, top 500 and top 200 per year in the 1-gram data. Annual turnover among the top 1000 and the top 500 decreased exponentially from the year 1700 to 2000, proportional to $e^{-0.012t}$ ($r^2 > 0.91$ for both), where t is years since 1700. This exponential decay equates to roughly a halving of turnover per century.

Since the corpus size was increasing with time, Fig 4 effectively also shows how turnover in top

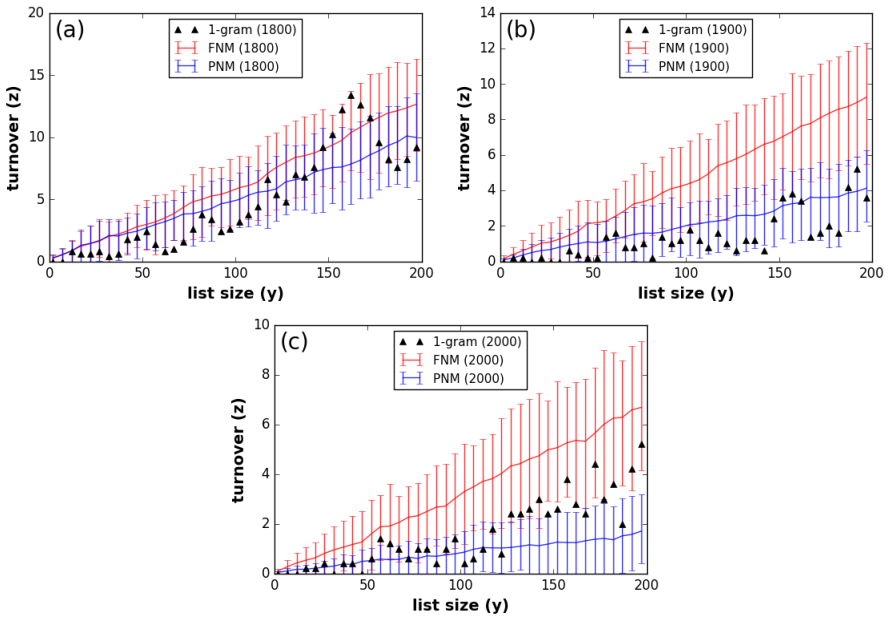

Figure 5: **Turnover profiles in 1-gram data and in simulated results**, for **(a)** the year 1800, **(b)** the year 1900 and **(c)** the year 2000. In each panel, the circles show turnover z in 1-grams versus list size y, averaged over the decade from five years before the new century to five years after. For the FNM, the corpus size, N_t, is 1.5 million by year 2000. For the PNM, the sample S_t grows exponentially as $S_0 e^{\alpha t}$ and the sampled *canon* size, N_t, grows linearly at 10,000 words per year, reaching 3 million by year 2000 ($t = 301$). For the FNM and the PNM, bands indicate 95% range of simulated values.

Table 1: Seven predictions of the Full Neutral Model (FNM) and Partial Sampling Neutral Model (PNM) and how they fare against 1-gram data.

Model	Zipf's Law	Heaps exponent	Heaps coefficient	Turnover $y = 50$	Turnover $y = 200$	z vs y yr 1800	z vs y yr 2000
FNM	Yes/No	No	No	Yes	No	Yes	No
PNM	Yes	Yes	No	Yes	Yes?	Yes	Yes

y list decreases as corpus size increases in the partial-sampling Neutral model. The exponential decay in turnover in the partial-sampling Neutral model is markedly different than the base Neutral model, in which turnover would be growing as corpus size grew, due to term $n_s^{0.013}$ in equation 1.

Finally, we also look at the "turnover profile", plotting list size y versus turnover z_y for different time slices (Fig 5). For all words, $z_y \propto y^{1.26}$ for different time periods (Fig 5). We can then compare the turnover profile for the 1-grams to the prediction from eq. 1 that turnover will be proportional to $y^{0.86}$, as shown in Fig 5b.

Table 1 lays out the specific predictions of each of the models and how they fare against empirical data. Bands indicate 95% range of simulated values. While the predictions for the FNM and PNM are similar for $y = 50$ and for the year 1800 (Fig 4a and Fig 5a), they do differ substantially in their predictions for Zipf's law and Heaps law under list size $y = 200$ and for the year 2000 (Fig 4c and Fig 5c). Although the FNM can fit Zipf's Law with the right parameters, it cannot also fit Heaps law or the turnover patterns at the same time as matching Zipf's Law. In contrast, the PNM can fit Zipf's law, Heaps law exponent (Fig 3a), and the 2000 series in Fig 4 (but starts to breakdown at $y > 150$). Neither the FNM nor the PNM does very well at $y = 200$.

Discussion

We have explored how 'neutral' models of word choice could replicate a series of static and dynamic observations from a historical 1-gram corpora: corpus size, frequency distributions, and turnover within those frequency distributions. Our goal was to capture two static and three dynamic properties of word frequency statistics in one model. The static properties are not only the well-known (a) Zipf's law, which a range of proportionate-advantage models can replicate, but also (b) Heaps law. The dynamic properties are (c) the continual turnover in words ranked by popularity, (d) the decline in that turnover rate through time, and (e) the relationship between list size and turnover, which we call the turnover profile.

We found that, although the full-sample Neutral model (FNM) predicts the Zipf's law in ranked word frequencies, the FNM does not replicate Heaps law between corpus and vocabulary size, or the concavity in the non-linear relationship between list size y and turnover z_y, or the slowing of this turnover through time among English words.

It is notable that we found it impossible to capture all five of these properties at once with the FNM. It was a bit like trying to juggle five balls, as soon as the FNM could replicate some of those properties, it dropped the others. Having explored the FNM under broad range of under a range of parameter combinations, we ultimately determined that it could never replicate all these properties at once. This is mainly because both vocabulary size in the FNM is proportional to corpus size (rather than roughly the square root of corpus size as in Heaps law) and also because turnover in FNM should increase slightly with growing corpus, not decrease as we see in the 1-gram data over 300 years. Other hypotheses to modify the FNM, such as introducing a conformity bias [2], can also be ruled out. In the case of conformity bias—where agents choose high-frequency words with even greater probability than just in proportion to frequency—both the Zipf law and turnover deteriorate under strong conformity in ways that mis-match with the data.

What did ultimately work very well was our partial-sampling Neutral model, or PNM (Fig 1b), which models a growing sample from a fixed-sized FNM. Our PNM, which takes exponentially increasing sample sizes from a neutrally evolved latent corpus, replicated the Zipf's law, Heaps law, and turnover patterns in the 1-gram data. Although it did not replicate exactly the particular 1-gram corpus we used here, the Heaps law exponent yielded by the PNM does fall within the range—from 0.44 to 0.54—observed in different English 1-gram corpora [44]. Among all features we attempted to replicate, the one mismatch between PNM and the 1-gram data is that the PNM yielded an order of magnitude fewer vocabulary words for a given corpus size, while increasing with corpus size according to the same Heaps law exponent. The reason for this mismatch appears to be a computational constraint: we could not run the PNM with exponential growth quite as large as that of the actual 300 years of exponential growth in the real English corpus.

As a heuristic device, we consider the fixed-size FNM to represent a canonical literature, while the growing sample represents the real world of exponentially growing numbers of books published ever year in English. Of course, the world is not as simple as our model; there is no official fixed canon, that canon does not strictly copy words from the previous year only and there are plenty of words being invented that occur outside this canon. Also, the Google dataset is an imperfect sample of the language for earlier years. At least some of the growth observed over time is due to greater availability and easier digitization of later texts, such that the Google corpus grows faster than the language itself over the years [16].

This does not change our overall result, however, in the PNM can replicate dynamic properties observed in an exponentially-growing corpus (even if that exponent were smaller) that the FNM cannot. In particular, our canonical model of the PNM differs from the explanation by [44], in which a "decreasing marginal need for additional words" as the corpus grows is underlain by the "dependency network between the common words ... and their more esoteric counterparts." In our PNM representation, there is no network structure between words at all, such as "inter-word statistical dependencies" [46] or grammar as a hierarchical network structure between words [21].

Conclusion

Since the PNM performed quite well in replicating multiple static and dynamic statistical properties of 1-grams simultaneously, which the FNM could not do, we find two insights. The first is that the FNM remains a powerful representation of word usage dynamics [13, 47, 27, 25, 9, 5], but it may need to be embedded in a larger sampling process in order to represent a very large data sample. Case studies where the PNM succeeds and the FNM fails could represent situations where mass attention is focused on a small subset of the cultural variants. The same idea seems appropriate for a digital world, where many cultural choices are pre-sorted in ranked lists [25]. In the present century, published books contain only a few percent of the verbiage recorded online, with the volume of digital data doubling about every three years. Centuries of prior evolution in published English word use provides valuable context for future study of this digital transition.

Models and data

Our aim is to compare key summary statistics from simulated data generated by the hypothetical FNM and PNM processes with summary statistics from Google 1-gram data. See Acknowledgements for data source address and the repository location for the Python code used to generate the FNM and PNM.

Neutral models

The FNM assumes words in a corpus at time t are selected at random from the corpus at time $t-1$. The corpus size N_t increases exponentially, $N_0 e^{0.021t}$, through time to simulate the exponentially increasing corpus size observed in the Google n-grams data [8]. We ran a genetic algorithm (described in the Appendix 2) to search the model state space to obtain parameter combinations—latent corpus size N_t, innovation fraction μ and initial corpus size N_0—that yielded similar summary statistics to the 1-gram data. With the corpus growth exponent fixed at 0.021, initial corpus size, N_0, was constrained by computational capacity.

Following the genetic algorithm search, the model was initialized with corpus size $N_0 = 3000$ and invention fraction $\mu = 0.003$. Once steady state was achieved, we permitted the corpus size in each successive generation to increase at an exponential growth rate comparable to the average annual growth rate of Google 1-gram data until it finally reached $N_{300} = 1.5$ million by time step $t = 301$.

At each time t in the FNM, a new set of N_t words enter the modeled corpus. Each word in the corpus, at time t, is either a copy of a word from the previous generation of books, with probability $1-\mu$, or else invented as a new word with probability μ. Each of the copied words is selected from v_{t-1} possible words (the vocabulary in the previous time step), which follow a discrete Zipf's law distribution with the probability a word is selected being proportional to the number of copies the word had in the previous corpus in time step $t-1$ [7].

The PNM, represented schematically in Fig 1, draws an exponentially increasing sample (with replacement) from a latent neutrally-evolving canon. We designate the number of words in the sample as S_t, and the cumulative number of words in the canon as N_t, which grows by a fixed number of words in each time step. This exponentially increasing sample, $S_0 e^{\alpha t}$, has an initial corpus size $S_0 = 3000$, growth exponent $\alpha = 0.021$, yielding a final sample size $S_{300} = 1.5$ million, matching the FNM. The latent corpus evolves by the rules of the FNM, but with a constant corpus size of 10000 for each year t (representing a canonical literature from which the main body of authors sample). The *cumulative* canon, N_t, thus grows by 10,000 words per year. The partial sample, S_t, at time t can copy words from all canonical literature, N_t, up to that time step. We set $\mu = 0.003$ and run for $t = 301$ time steps representing years between 1700 and 2000, which are the same parameters used in the FNM.

1-gram data

The 1-gram data are available as csv files directly from Google's Ngrams site [26]. As in a previous study [1], we removed 1-grams that are common symbols or numbers, and 1-grams containing the same consonant three or more times consecutively. As in our other studies [1, 8, 6], we normalized the count of 1-grams using the yearly occurrences of the most common English word, *the*. Although we track 1-grams from the year 1700, for turnover statistics we follow other studies [44] in being cautious about the n-grams record before the year 1800, due to misspelled words before 1800 that were surely digital scanning errors related to antique printing styles of that may conflate letters such as 's' and 'f' (*e.g., myfelf, yourfelf, provifions, increafe, afked* etc). The code used for modeling is available at: https://github.com/dr2g08/Neutral-evolution-and-turnover-over-centuries-of-English-word-popularity.

Acknowledgments

We thank William Brock for comments on an early draft. RAB thanks the Northwestern Institute on Complex Systems for support as a visiting scholar. DR is supported by a grant from the Hobby School of Public Affairs, University of Houston and also by EPSRC grant to the Bristol Centre for Complexity Sciences (EP/I013717/1). AA was supported by a Royal Society Newton Fellowship at Bristol University entitled "Cultural evolution online"; PG was supported by the Leverhulme Trust grant on "Tipping Points" (F/00128/BF) awarded to Durham University.

References

1. Acerbi, A, Lampos V, Garnett P, Bentley RA (2013). The expression of emotions in 20th century books. *PLoS ONE* 8(3): e59030.
2. Acerbi A, Bentley RA (2014). Biases in cultural transmission shape the turnover of popular traits. *Evolution & Human Behavior* 35: 228–236.
3. Altmann EG, Pierrehumbert JB, Motter AE (2011). Niche as a determinant of word fate in online groups. *PLoS ONE* 6(5): e19009.
4. Batty M (2006). Rank clocks. *Nature* 444: 592–596.
5. Barucca P, Rocchi J, Marinari E, Parisi G, Ricci-Tersenghi F (2015). Cross-correlations of American baby names. *PNAS* 112: 7943–7947.
6. Bentley RA, Acerbi A, Lampos V, Ormerod P (2014). Books average previous decade of economic misery. *PLoS ONE* 9(1): e83147.
7. Bentley RA, Caiado C, Ormerod P (2014). Effects of memory on spatial heterogeneity in neutrally transmitted culture. *Evolution & Human Behavior* 35: 257–263.
8. Bentley RA,Garnett P, O'Brien MJ, Brock WA (2012). Word diffusion and climate science. *PLoS ONE* 7(11): e47966.
9. Bentley RA, Ormerod P, Batty M (2011). Evolving social influence in large populations. *Behavioral Ecology & Sociobiology* 65: 537–546.
10. Bentley RA, Shennan SJ, Ormerod P (2011). Population-level neutral model already explains linguistic patterns. *Proceedings B* 278: 1770–1772.
11. Bentley RA, Hahn MW, Shennan SJ (2004). Random drift and culture change. *Proceedings B* 271: 1443–1450.
12. Bentley RA, Lipo CP, Herzog HA, Hahn MW (2007). Regular rates of popular culture change reflect random copying. *Evolution & Human Behavior* 28: 151–158.
13. Bentley RA (2008). Random drift versus selection in academic vocabulary. *PLoS ONE* 3(8): e3057.
14. Christiansen MH, Chater N (2008). Language as shaped by the brain. *Behavioral & Brain Sciences* 31: 489–509.
15. Clauset A, Shalizi CR, Newman MEJ (2007). Power-law distributions in empirical data. *SIAM Review* 51: 661–703.
16. Cuskley CF, Pugliese M, Castellano C, Colaiori F, Loreto V, Tria F (2014). Internal and external dynamics in language: Evidence from verb regularity in a historical corpus of English. *PLoS ONE* 9(8): e102882.
17. Dehaene S, Mehler J (1992). Cross-linguistic regularities in the frequency of number words. *Cognition* 43: 1–29.
18. Eriksson K, Jansson F, Sjöstrand, J (2010). Bentley's conjecture on popularity toplist turnover under random copying. *Ramanujan Journal* 23: 371–396.
19. Evans TS (2007). Exact solutions for network rewiring models. *European Physical Journal B* 56: 65–69.
20. Evans, TS, Giometto, A (2011). Turnover rate of popularity charts in neutral models. arXiv: 11054044v1.
21. Ferrer i Cancho R, Riordan O, Bollobás B (2005). The consequences of Zipf's law for syntax and symbolic reference. *Proceedings B* 2005; 272: 561–565.
22. Gabaix X (2009). Power laws in economics and finance. *Annual Review of Economics* 1: 255-293.
23. Gao J, Hu J, Mao X, Perc M (2012). Culturomics meets random fractal theory: Insights into long-range correlations of social and natural phenomena over the past two centuries. *J. R Soc Interface* 9: 1956–1964.
24. Ghoshal G, Barabási A-L (2011). Ranking stability and super-stable nodes in complex networks. *Nature Communications* 2: 394.

25. Gleeson JP, Cellai D, Onnela J-P, Porter MA, Reed-Tsochas F (2014). A simple generative model of collective online behavior. *PNAS* 111: 10411–10415.

26. Google Books. https://booksgooglecom/ngrams/info

27. Hahn MW, Bentley RA (2003). Drift as a mechanism for cultural change: an example from baby names. *Proceedings B* 270: S1–S4.

28. Hruschka DJ, Christiansen MH, Blythe RA, Croft W, Heggarty P, Mufwene SS, Pierrehumbert JB, Poplack S (2009). Building social cognitive models of language change. *Trends in Cognitive Sciences* 13: 464–469.

29. Hughes JM, Foti NJ, Krakauer DC, Rockmore DN (2012). Quantitative patterns of stylistic influence in the evolution of literature. *PNAS* 109: 7682–7686.

30. Kandler A, Shennan S (2013). A non-equilibrium neutral model for analysing cultural change. *J. Theoretical Biology* 330: 18–25.

31. Laherrère J, Sornette D (1998). Stretched exponential distributions in nature and economy: 'fat tails' with characteristic scales. *European Physical Journal B* 2: 525–539.

32. Lanfear R, Kokko H, Eyre-Walker A (2014). Population size and the rate of evolution. *Trends in Ecology & Evolution* 29: 33–41.

33. Li W (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Trans Inf Theory* 38: 1842–1845.

34. Lieberman E, Hauert C, Nowak MA (2005). Evolutionary dynamics on graphs. *Nature* 433: 312–316.

35. Lieberman E, Michel J-P, Jackson J, Tang T, Nowak MA (2007). Quantifying the evolutionary dynamics of language. *Nature* 449: 713–716.

36. Lin Y, Michel JB, Aiden EL, Orwant J, Brockman W, Petrov S (2012). Syntactic annotations for the google books ngram corpus. In: *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, pp.169–174.

37. Lü L, Zhang Z-K, Zhou T (2010). Zipf's Law leads to Heaps' Law: Analyzing their relation in finite-size systems. *PLoS ONE* 5(12): e14139.

38. Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J, Pinker S, Nowak MA, Aiden EL (2011). Quantitative analysis of culture using millions of digitized books. *Science* 331:176–182.

39. Neiman FD (1995). Stylistic variation in evolutionary perspective. *American Antiquity* 60: 7–36.

40. Pagel M, Atkinson QD, Meade A (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449: 717–721.

41. Pan RK, Petersen AM, Pammolli F, Fortunato S (2016). The memory of science: inflation, myopia, and the knowledge network. arXiv: 160705606v1.

42. Perc M (2012). Evolution of the most common English words and phrases over the centuries. *J R Soc Interface* 9: 3323–3328.

43. Perc M (2014). The Matthew effect in empirical data. *J R Soc Interface* 11: 20140378.

44. Petersen AM, Tenenbaum J, Havlin S, Stanley HE, Perc M (2012). Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific Reports* 2: 943.

45. Petersen AM, Tenenbaum J, Havlin S, Stanley HE (2012). Statistical laws governing fluctuations in word use from Word Birth to Word Death. *Scientific Reports* 2: 313.

46. Piantadosi ST, Tily H, Gibson E (2011). Word lengths are optimized for efficient communication. *PNAS* 108: 3526–3529.

47. Reali F, Griffiths TL (2010). Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. *Proceedings B* 277: 429–436.

48. Sigurd B, Eeg-Olofsson M, van de Weijer J (2004). Word length, sentence length and frequency–Zipf revisited. *Studia Linguistica* 58: 37–52.

49. Strimling P, Sjöstrand J, Eriksson K, Enquist M (2009). Accumulation of cultural traits. *Theoretical Population Biology* 76: 77–83.

50. Williams JR, Lessard PR, Desu S, Clark E, Bagrow JP, Danforth CM, Dodds PS (2015). Zipf's law holds for phrases, not words. *Scientific Reports* 5: 12209.

51. Zipf GK (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison Wesley.

Appendix 1 Neutral model yields Zipf's law. Recent analytical results [49] show that the expected number of variants of popularity rank k under the stationary distribution is

$$f_k = \mu N_t \frac{(1-\mu)^{k-1}}{k} \prod_{i=1}^{k-1} \frac{N_t - i}{N_t - i - 1 + i\mu}. \tag{2}$$

Note from this expression [49], we can find the ratio of f_{k+1}/f_k, which is

$$\frac{f_{k+1}}{f_k} = \frac{\mu N_t \frac{(1-\mu)^k}{k+1} \prod_{i=1}^{k} \frac{N_t - i}{N_t - i - 1 + i\mu}}{\mu N_t \frac{(1-\mu)^{k-1}}{k} \prod_{i=1}^{k-1} \frac{N_t - i}{N_t - i - 1 + i\mu}}. \tag{3}$$

which simplifies to

$$\frac{f_{k+1}}{f_k} = \frac{k(1-\mu)(N_t - k)}{(k+1)(N_t - k - 1 + k\mu)}. \tag{4}$$

If N_t is large compared to k and μ is small, then this simplifies to

$$\frac{f_{k+1}}{f_k} \approx \frac{k}{k+1}, \tag{5}$$

which is an expression for Zipf's law, because the ratio of the word frequencies is inversely proportional to the ratio of their ranks.

Appendix 2. Genetic algorithm. The PNM has five parameters N, μ, S_0, α and T. The number of time steps, T is fixed at 301 (representing calendar years). The exponential growth rate of the sampled corpus, α, is fixed at 0.02. The other three parameters - initial sampled corpus size (S_0), latent corpus size, N, and innovation rate (μ) - are free. We bound potential values of Nbetween 5000 and 30000 and S_0 between 1000 and 10000. In both cases the lower bound is chosen to ensure a minimum acceptable vocabulary size is reached and the upper bound is limited by computational constraints. The product $N\mu$ was limited between 5 and 90, as the region in which Neutral model yields a reasonable Zipf's law. For the genetic algorithm, the fitnesses were scored by the following equations and a variable values:

Summary statistic	Equation	Target variables
Heaps Law	$v = An^b$	A and b
Zipf's law	$f \sim k^{-\gamma}$	γ
Turnover decay ($y = 50$)	$z(50) = z_0 e^{-\beta_{50} t}$	β_{50} and z_0
Turnover decay ($y = 100$)	$z(100) = z_0 e^{-\beta_{100} t}$	β_{100} and z_0
Turnover decay ($y = 200$)	$z(200) = z_0 e^{-\beta_{200} t}$	β_{200} and z_0

The PNM parameter combination receives a point when each of the target statistics is approximately the same as the equivalent value from the n-grams data. The genetic algorithm starts with 100 random parameter combinations then the following steps are repeated until they converge on parameter combinations that maximize fitness scores:

1. The fittest 20% from the corpus is passed to the next generation.
2. The remaining 80% is populated by recombinations of two randomly selected parents from the fittest 20% from the previous generation.
3. 15% of the new agents are subject to random mutation of a single parameter to ensure diversity in the corpus.