2015

Lexical mechanics: Partitions, mixtures, and context

Jake Ryland Williams
University of Vermont

Lexical Mechanics:
Partitions, mixtures, and context

A Dissertation Presented

by

Jake Ryland Williams

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fullfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Mathematical Sciences

May, 2015

Defense Date: March 18, 2015
Dissertation Examination Committee:

Peter Sheridan Dodds, Ph.D., Advisor
Christopher M. Danforth, Ph.D., Advisor
Jacques A. Bailly, Ph.D., Chairperson
Richard M. Foote, Ph.D.
Cynthia J. Forehand, Ph.D., Dean of the Graduate College

Abstract

Highly structured for efficient communication, natural languages are complex systems. Unlike in their computational cousins, functions and meanings in natural languages are relative, frequently prescribed to symbols through unexpected social processes. Despite grammar and definition, the presence of metaphor can leave unwitting language users "in the dark," so to speak. This is not problematic, but rather an important operational feature of languages, since the lifting of meaning onto higher-order structures allows individuals to compress descriptions of regularly-conveyed information. This compressed terminology, often only appropriate when taken locally (in context), is beneficial in an enormous world of novel experience. However, what is natural for a human to process can be tremendously difficult for a computer.

When a sequence of words (a phrase) is to be taken as a unit, suppose the choice of words in the phrase is subordinate to the choice of the phrase, i.e., there exists an inter-word dependence owed to membership within a common phrase. This word selection process is *not* one of independent selection, and so *is* capable of generating word-frequency distributions that are not accessible via independent selection processes. We have shown in Ch. 2 through analysis of thousands of English texts that empirical word-frequency distributions possess these word-dependence anomalies, while phrase-frequency distributions do not. In doing so, this study has also led to the development of a novel, general, and mathematical framework for the generation of frequency data for phrases, opening up the field of mass-preserving mesoscopic lexical analyses.

A common oversight in many studies of the generation and interpretation of language is the assumption that separate discourses are independent. However, even when separate texts are each produced by means of independent word selection, it is possible for their composite distribution of words to exhibit dependence. Succinctly, different texts may use a common word or phrase for different meanings, and so exhibit disproportionate usages when juxtaposed. To support this theory, we have shown in Ch. 3 that the act of combining distinct texts to form large 'corpora' results in word-dependence irregularities. This not only settles a 15-year discussion, challenging the current major theory, but also highlights an important practice necessary for successful computational analysis—the retention of meaningful separations in language.

We must also consider how language speakers and listeners navigate such a combinatorially vast space for meaning. Dictionaries (or, the collective editorial communities behind them) are smart. They know all about the lexical objects they define, but we ask about the latent information they hold, or should hold, about related, undefined objects. Based solely on the text as data, in Ch. 4 we build on our result in Ch. 2 and develop a model of context defined by the structural similarities of phrases. We then apply this model to define measures of meaning in a corpus-guided experiment, computationally detecting entries missing from a massive, collaborative online dictionary known as the Wiktionary.

Citations

Material from this dissertation has been accepted for publication in *Physical Review E* on 03/17/2015 in the following form:

Williams, J. R. and and Bagrow, J. P. and Danforth, C. M. and Dodds, P. S.. (2015). Text mixing shapes the anatomy of rank-frequency distributions: A modern Zipfian mechanics for natural language. *Physical Review E*.

AND

Material from this dissertation has been submitted for publication in *Nature Scientific Reports* on 12/29/2014 in the following form:

Williams, J. R. and Lessard, P. R. and Desu, S. and Clark, E. M. and Bagrow, J. P. and Danforth, C. M. and Dodds, P.S.. (2014). Zipf's law holds for phrases, not words. *Nature Scientific Reports*.

AND

Material from this dissertation has been submitted for publication in *Physical Review E* on 03/05/2015 in the following form:

Williams, J. R. and Clark, E. M. and Bagrow, J. P. and Danforth, C. M. and Dodds, P.S.. (2015). Frequency-conserving context models detect missing dictionary entries. *Physical Review E*.

Dedicated to

my beloved wife and partner in all things

Sharon

Acknowledgements

I must take this opportunity to thank all of those friends and family who have cared for me—both mentally and physically—over the years. I could not have even dreamt of this accomplishment without you. I would also like to thank my advisors, Peter and Chris, for their guidance, thought, and direction; Richard Foote, for his wisdom and knowledge; Jacques Bailly, for first teaching me about language; Jim Bagrow, for his spirited realism and keen intellect; Andrea Elledge, for facilitating all things; Jim Lawson, for being the kindest system administrator I have ever known; David Van Horn, Josh Auerbach, Andy Reagan and Nick Allgaier, the LaTeXwizards who produced this document's template; Cathy Bliss, for all of her thoughts and guidance as a graduate student; my teachers and professors, for their patience and kindness; my fellow graduate students over the years, for all of our shared pain and growth; and my dear cats, Emmett and Thurston, for keeping me sane.

Table of Contents

List of Figures

[1]Data: The complete historical romances of Georg Ebers.

List of Tables

Chapter 1

Introduction and Literature Review

In this chapter we introduce the topic of study, the statistical mechanics of natural lexica. Though all of the studies in this dissertation fall within this greater physical framework, they naturally fall into three focuses, namely: (1) the definition of phrase-generalized lexical frequencies, (2) the dependence of lexical frequencies on mixed corpora, and (3) the definition of models of context. Here we review this work's predecessors, and discuss their context from earlier in the 20^{th} century.

1.1 Introduction

The basis for the studies contained in this dissertation is composed of several works from early in the 20^{th} century focusing on the topics of evolution (Yule, 1924), social preference (Zipf, 1935, 1949; Simon, 1955), and information theory (Shannon, 1948; Mandelbrot, 1953). When focusing on the topic of natural language, debates have sparked both early on with Zipf (1949) and Miller (1957), and Simon and Mandelbrot (Simon, 1955; Mandelbrot, 1959; Simon, 1960; Mandelbrot, 1961a; Simon, 1961b; Mandelbrot, 1961b; Simon, 1961a), and more recently with others (Piantadosi et al., 2011b; Reilly and Kean, 2011; Piantadosi et al., 2011a; Ferrer-i-Cancho and P., 2012; Piantadosi et al., 2013). The main body of this chapter will discuss highlights and concerns with the more current work.

In the first section we will focus in detail on the parsing of phrases as lexical objects and the production of frequency data (Becker, 1975; Michel et al., 2011; Ha et al., 2009;

Lin et al., 2012), which is straightforward for words, but requires nuance when considering phrases. In the second section we will focus on theories describing language formation and structure (Ferrer-i-Cancho and Solé, 2001; Kwapien et al., 2010; Gerlach and Altmann, 2013; Corominas-Murtra et al., 2014), and finally in the third, we will discuss collocation-based context models and their applications (Church and Hanks, 1990; Smadja, 1993; Piantadosi et al., 2011b; Garcia et al., 2012).

1.2 Higher-order lexical data

Becker, J. D., 1975. The phrasal lexicon. In: Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing. TINLAP '75. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 60–63, http://dx.doi.org/10.3115/980190.980212

In this work, in 1975, Joseph D. Becker asserts the existence and dominance of the phrasal lexicon of English, hypothesizing that most utterances are produced in common social situations, where the demand of communication is not for novelty, but instead for fomulaic language, such as idioms, cliches, and turns of phrase. He suggests further that the majority of our social language is formed by the repitition, modification, and concatenation of previously-known phrases consisting of more than one word. In this work Becker notes that while (at the time) no English dictionary comes close to encompassing the variety of English phrases he discusses, he has seen phraseological dictionaries of more than $25,000$ entries, encompassing rare phrases like "knee-high to a grasshopper." This early suggestion of the existence of an enormous and unexplored phrasal lexicon has served as an impetus, a base-theory for much of the work in this thesis, pointing to the determination of the lexicon's size as an independent way of gauging its importance (in addition to proposing phrasal formation mechanisms that describe important structural relations).

Ha, L. Q., Sicilia-Garcia, E. I., Ming, J., Smith, F. J., 2002. Extension of Zipf's law to words and phrases. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING). pp. 315–320

A first computational step to studying the phrasal lexicon as a whole came from work by Ha et al. in 2002, and focused on word-sequence, or, N-grams data. For them and others, N-grams were generally defined by the 'tokens' appearing between segments of whitespace. While earlier work had considered N-gram frequencies, plotting them individually (Smith and Devine, 1985), this study combined N-gram frequency distributions of varying N-lengths to find a better conformation to Zipf's law than by words or any N-gram length alone. However, there are major issues with this approach, as N-grams overlap and are not counted independently of one another. As such, the consideration of N-gram frequencies is lacking in physical meaning, since one has no way to derive the true mass of words appearing on "the page," and produce an appropriate N-gram normalization for probabilistic modeling . Furthermore, as more and more lengths are combined, the misrepresentation caused by overlap is exacerbated, leading us to ask if there is a better way. Nevertheless, this work serves as an important step in the acknowledgment and study of an integrated phrasal lexicon, which has set the stage for much of the work presented in this dissertation.

Michel, J., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., Aiden, E. L., 2011. Quantitative analysis of culture using millions of digitized books. Science 331 (6014), 176–182, http://www.sciencemag.org/content/331/6014/176.abstract

An early computational step to the large-scale excavation of the phrasal lexicon has been the production of N-gram frequency data on a massive scale by the Google machine translation team (Google, 2006; Lin et al., 2012). This data gained wide attention in 2011, when the

article considered here emerged, exploring the data in a temporal fashion, and making inferences about cultural behavior. While their observations were widely upheld, empirically confirming transformations of the English language (like verb regularization), these observations were made on data that we know now to be prone to issues of curation (Pechenick et al., 2015), unduly placing scientific texts as books in the dataset, muddying the inferences that researchers might wish to make from this data. Further, as with other N-gram analyses, this work is likewise subject to the issue of word-frequency misrepresentation, stemming from the construction of all N-gram data sets, discussed above and in the main body of the dissertation.

Lin, Y., Michel, J., Aiden, E. L., Orwant, J., Brockman, W., Petrov, S., 2012. Syntactic annotations for the google books ngram corpus. In: Proceedings of the ACL 2012 System Demonstrations. ACL '12. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 169–174, http://dl.acm.org/citation.cfm?id=2390470.2390499

In 2012, a second generation of N-gram data emerged from the Google team, addressing several issues present in the original data (Google, 2006; Michel et al., 2011). Most notably, the original N-gram parsing technique tokenized 'words' by whitespace, which is quite reasonable as a first pass, but unfortunately includes highly non-lexical objects, such as punctuation and all manners of markup. As such, when this algorithm was applied in an automated fashion to millions of books and web pages, the results contained massive amounts of junk text, of little interest to researchers. The authors here improved the methodology, not only adding syntactic annotations to the data set, but performing the tokenization within the bounds of punctuation, eliminating much of the junk text present in the previous versions. However, with the improvements came exacerbation of an old issue: under the old methodology the first and last words of a text would lack some 2-gram

membership and thereby be underrepresented in the 2-gram frequencies. Now, appearance in many short sentences precludes words from appearance in many 2-gram scenarios, an issue which gets magnitudinally worse for larger values of N. Despite this issue now exacerbated, this production made great strides toward integrating grammatical, punctuation and boundary information into text parsing techniques, which as we will see in the main body of the dissertation may be accomplished in a mathematically sound and physically principled manner.

1.3 Models of vocabulary formation

Ferrer-i-Cancho, R., Solé, R. V., 2001. Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. Journal of Quantitative Linguistics 8 (3), 165–173

This paper first proposed the hypothesis that word vocabularies naturally decompose into two subsets. The theory considered the existence of a kernel (core) lexicon of versatile words, and an unlimited (non-core) lexicon for specific communication. This hypothesis came about with the rise of computation: both the paper discussed and a concurrent article by Montemurro (2001) first noted the existence of multiple scaling regimes in the rank-frequency distributions of large corpora. However, while Ferrer-i-Cancho and Solé (2001) speculated as to the reasons for the two regimes, Montemurro (2001) cautioned against the study of this phenomena in the presence of large mixed corpora. Since then, the core/non-core vocabulary theory has prevailed in the community, and even led to work by Gerlach and Altmann (2013) (which we focus on more closely below) on a generative selection model that is capable of producing the observed multiple scaling regimes.

Kwapien, J., Drozdz, S., Orczyk, A., 2010. Linguistic complexity: English vs. polish, text vs. corpus. Acta Physica Polonica, A. 117, 716

In 2010, Kwapien et al. continued the work of Montemurro (2001), investigating the multiple scaling regimes present in rank-frequency word distributions. In this work the authors studied some effects of corpus composition on the second scaling regime and were even able to show its presence in corpora in languages other than English. There, it was found with two corpora of comparable size that one by a mixture of authors had a more abrupt and severe scaling break than one by a single author. While this work is brief on the side of analytics, it approaches the rank-frequency scaling break with subtlety, guiding us in our work to consider mixtures of texts of varying compositions and languages in our investigation of the core language hypothesis in the main body of the dissertation.

Gerlach, M., Altmann, E. G., 2013. Stochastic model for the vocabulary growth in natural languages. Phys. Rev. X 3, 021006

In this work, Gerlach et al. considered the early observations of Ferrer-i-Cancho and Solé (2001) and Montemurro (2001), and produced a stochastic model based off of that of Simon (1955) for the vocabulary growth of natural languages. The notable product of this work is its derivation of a means for producing rank-frequency distributions with severe scalings. While this was a huge advancement for the relevance of preferential selection as a mechanism for the production of social data, the execution of their development was limited in its focus on supporting the core/non-core vocabulary theory. The authors did a masterful job integrating this theory with the preferential selection mechanism, and were able to produce very realistic simulations, but failed to consider other physical processes (which we show in Ch. 3 are dominant in the creation of large corpora). This work has ultimately shown that preferential selection and decaying innovation are two very important social process, but the field is still open and there are other mechanisms that have been proposed for the

production of natural language vocabularies, such as the Dirichlet processes discussed by MacKay (2002) (which utilizes both preferential selection and decaying innovation), and the history-dependent dice process proposed by Corominas-Murtra et al. (2014), which we review at length.

Corominas-Murtra, B., Hanel, R., Thurner, S., 2014. Understanding zipf's law with playing dice: history-dependent stochastic processes with collapsing sample-space have power-law rank distributions. CoRR abs/1407.2775, http://arxiv.org/abs/1407.2775

In this work, the authors propose to investigate an alternative mechanism for the generation of Zipf's law, and hence natural language vocabularies. Here, the authors consider the effects of rolling dice with fewer and fewer faces in a history-dependent way, where the size of the next die is equal to one less than the value of the previous roll. This process biases heavily toward the lowest numbers (as every sequence terminates in a roll of 1), and so is capable of generating heavy-tailed distributions, converging to scalings in the limit. However despite this feature of the model and its extension by the authors, producing a wide range of scaling exponents, we note that it comes up short on a few accounts. First, the range of scalings produced by this model are still less severe than those observed in nature. Second, the model does not realistically represent the class of once-appearing words known as hapax legomena, which generally comprise half the words appearing in texts (approximately). Finally, what is perhaps the most important limitation of this model is the fact that the scalings they observe are *not* scalings of ranks. Since the authors do not rank their model output, but instead analyze its dice-face distributions, their results are actually incomparable to empirical rank-frequency distributions. Hence, such a process (while still quite interesting) unfortunately informs us of little (if anything) about Zipf's law, and the appearance of scaling through social processes.

1.4 Collocation context models

Church, K. W., Hanks, P., Mar. 1990. Word association norms, mutual information, and lexicography. Comput. Linguist. 16 (1), 22–29, http://dl.acm.org/citation.cfm?id=89086.89095

In 1990 Church and Hanks proposed an alternative measure for word association norms. In particular they produced an asymmetric version of mutual information, which has been widely applied and cited in the community (e.g., by Justeson and Katz 1991; Smadja 1993; Seretan 2008; Pecina 2010; Ramisch 2014 to name a few). Not only did these authors first propose entropic measures of frequency distributions for word association, but they also go in to detail, exploring (and correcting for) issues of frequency conservation in marginal probability distributions. It is good to see physical considerations like this early on, but as we will discuss in our review of the more recent works (both below and in Ch. 4), these consideration are often left out. Moreover, even when Church and Hanks accommodate for the over-counting in their measure, they do so by assuming texts are circular—an assumption whose approximation breaks down when small texts are considered, or when punctuation is observed (as is done with modern N-gram data (Lin et al., 2012)).

Smadja, F., Mar. 1993. Retrieving collocations from text: Xtract. Comput. Linguist. 19 (1), 143–177, http://dl.acm.org/citation.cfm?id=972450.972458

In this work the author applies the model of Church and Hanks (1990) for the purposes of extracting collocations, whose definition they take as arbitrary and recurrent word combinations. They go beyond the word-word associations measured by the base model, and use its output to extend to large collocations of more than two words. Using these rigid forms, the author then defines and exhibits an algorithm for the identification of insertive forms, and has notable success extracting phrasal templates, which add used to add syn-

tactic annotations. All together, the author referred to the algorithm/tool as "Xtract," and while this package is no longer updated or widely used, it has served as a basis for more current work by authors such as Seretan (2008), Pecina (2010), and Ramisch (2014), guiding the community toward the development of general extraction techniques for large lexical objects.

Piantadosi, S. T., Tily, H., Gibson, E., 2011b. Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences 108 (9), 3526, http://colala.bcs.rochester.edu/papers/PNAS-2011-Piantadosi-1012551108.pdf

Here, the authors consider another information-theoretic measure using a context model derived from word collocations, or N-grams. In particular, their model is an extension from the word-transition probabilities investigated by Shannon (1948), but for higher-order patterns of usage. As was the case with Church and Hanks (1990), the model presented by Piantadosi et al. (2011b) is asymmetric, and only considers N-grams of a fixed N/length at a time, which must be specified to define the model, discarding valuable information and making the model non-general. However, the asymmetry built into their model afforded an approximate frequency preservation when applied to the original, white space-tokenized N-gram distributions (Google, 2006)—an approximation that breaks down when modern N-gram data is considered (Lin et al., 2012) (which we note both here and in Ch. 4). In application, the authors exhibit the power of their model using an entropic measure on words, which they compare with orthographic lengths to find significant correlations. However, as is pointed out in Ch. 4, this result has been of concern to others (Reilly and Kean, 2011; Piantadosi et al., 2011a; Ferrer-i-Cancho and P., 2012; Piantadosi et al., 2013), and hence is taken as guiding work for us only for its manipulation of the context model.

Garcia, D., Garas, A., Schweitzer, F., 2012. Positive words carry less information than negative words. EPJ Data Science 1 (1), http://dx.doi.org/10.1140/epjds3

In this article the authors apply the model and information-theoretic measure (referred to point-wise as the Information Content (IC) of a word) produced by Piantadosi et al. (2011b) to the Google N-grams corpus (Google, 2006), and compare its output to existing word-sentiment norms. In their results they find a result, claimed succinctly in their title: "Positive words carry less information than negative words." However, despite this result we note two concerns with the work. First, the IC-measure is strongly (inversely) associated with word frequency, making their result an implication of the frequency-dependence of sentiment norms, observed in other recent work (Kloumann et al., 2012; Dodds et al., 2015). Most importantly, however, their application of the context model described by Piantadosi et al. (2011b) makes use of a special formula, which technically only applies to uncompressed, human readable text, and *not* the frequency-based N-grams. This second point casts their results into question, and calls attention to the care needed when handling these kinds of models.

Chapter 2

Zipf's law holds for phrases, not words

With Zipf's law being originally and most famously observed for word frequency, it is surprisingly limited in its applicability to human language, holding over no more than three to four orders of magnitude before hitting a clear break in scaling. Here, building on the simple observation that phrases of one or more words comprise the most coherent units of meaning in language, we show empirically that Zipf's law for phrases extends over as many as nine orders of rank magnitude. In doing so, we develop a principled and scalable statistical mechanical method of random text partitioning, which opens up a rich frontier of rigorous text analysis via a rank ordering of mixed length phrases.

2.1 Introduction

Over the last century, the elements of many disparate systems have been found to approximately follow Zipf's law—that element size is inversely proportional to element size rank (Zipf, 1935, 1949)—from city populations (Zipf, 1949; Simon, 1955; Batty, 2008), to firm sizes (Axtell, 2001), and family names (Zanette and Manrubia, 2001). Starting with Mandelbrot's optimality argument (1953), and the dynamically growing, rich-get-richer model of Simon (1955), strident debates over theoretical mechanisms leading to Zipf's law have continued until the present (Miller, 1957; Ferrer-i-Cancho and Elvevåg, 2010; D'Souza et al., 2007; Corominas-Murtra and Solé, 2010). Persistent claims of uninteresting randomness underlying Zipf's law (Miller, 1957) have been successfully challenged (Ferrer-i-

Cancho and Elvevåg, 2010), and in non-linguistic systems, good evidence supports Simon's model (Simon, 1955; Bornholdt and Ebel, 2001; Maillart et al., 2008) which has been found to be the basis of scale-free networks (de Solla Price, 1976; Barabási and Albert, 1999).

For language, the vast majority of arguments have focused on the frequency of an individual word which we suggest here is the wrong fundamental unit of analysis. Words are an evident building block of language, and we are naturally drawn to simple counting as a primary means of analysis (the earliest examples are Biblical corcordances, dating to the 13th Century). And while we have defined morphemes as the most basic meaningful 'atoms' of language, the meaningful 'molecules' of language are clearly a mixture of individual words and phrases. The identification of meaningful phrases, or multi-word expressions, in natural language poses one of the largest obstacles to accurate machine translation (Sag et al., 2002). In reading the phrases "New York City" or "Star Wars", we effortlessly take them as irreducible constructions, different from the transparent sum of their parts. Indeed, it is only with some difficulty that we actively parse highly common phrases and consider their individuals words.

While partitioning a text into words is straightforward computationally, partitioning into meaningful phrases would appear to require an additional level of sophistication requiring online human analysis. But in order to contend with the increasingly imposing sizes and rapid delivery rates of important text corpora—such as news and social media—we are obliged to find a simple, necessarily linguistically naive, yet effective method.

A natural possibility is to in some way capitalize on N-grams, which are a now common and fast approach for parsing a text. Large scale N-gram data sets have been made widely available for analysis, most notably through the Google Books project (Google, 2014). Unfortunately, all N-grams fail on a crucial front: in their counting they overlap, which obscures underlying word frequencies. Consequently, and crucially, we are unable to

properly assign rankable frequency of usage weights to N-grams combined across all values of N.

Here, we introduce 'random partitioning', a method that is fast, intelligible, scalable, and sensibly preserves word frequencies: i.e., the sum of sensibly-weighted partitioned phrases is equal to the total number of words present. As we show, our method immediately yields the profound basic science result that phrases of mixed lengths, as opposed to just individual words, obey Zipf's law, indicating the method can serve as a profitable approach to general text analysis. To explore a lower level of language, we also partition for sub-word units, or graphemes, by breaking words into letter sequences. In the remainder of the paper, we first describe random partitioning and then present results for a range of texts.

2.2 Text partitioning

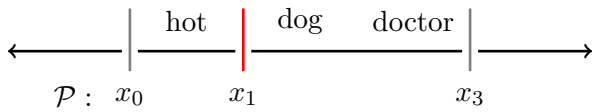
To begin our random partitioning process, we break a given text T into clauses, as demarcated by standard punctuation (other defensible schemes for obtaining clauses may also be used), and define the length norm, ℓ, of a given clause t (or phrase, $s \in S$) as its word count, written $\ell(t)$. We then define a partition, $\mathcal{P}$, of a clause t to be a sequence of the boundaries surrounding its words:

$$\mathcal{P}: \; x_0 < \cdots < x_{\ell(t)}, \tag{2.1}$$

and note that $x_0, x_{\ell(t)} \in \mathcal{P}$ for any $\mathcal{P}$, as we have (a priori) the demarcation knowledge of the clause. For example, consider the highly ambiguous text:

"Hot dog doctor!"

Forgoing punctuation and capitalization, we might attempt to break the clause down, and interpret through the partition:



i.e., $\mathcal{P} = \{x_0, x_1, x_3\}$, which breaks the text into phrases, "hot" and "dog doctor", and assume it as reference to an attractive veterinarian (as was meant in (Cougar Town, 2013)). However, depending on our choice, we might have found an alternative meaning:

hot dog; doctor:	A daring show-off doctor.
:	One offers a frankfurter to a doctor.
hot; dog doctor:	An attractive veterinarian (vet).
:	An overheated vet.
hot dog doctor:	A frank-improving condiment.
:	A frank-improving chef.
hot; dog; doctor:	An attractive vet of canines.
:	An overheated vet of canines.

Note in the above that we (as well as the speaker in (Cougar Town, 2013)) have allowed the phrase "dog doctor" to carry idiomatic meaning in its non-restriction to canines, despite the usage of the word "dog".

Now, in an ideal scenario we might have some knowledge of the likelihood for each boundary to be "cut" (which would produce an 'informed' partition method), but for now our goal is generality, and so we proceed, assuming a uniform boundary-cutting probability, q, across all $\ell(t) - 1$ word-word (clause-internal) boundaries of a clause, t. In general, there are $2^{\ell(t)-1}$ possible partitions of t involving $\frac{1}{2}\ell(t)(\ell(t)+1)$ potential phrases. For each integral pair i, j with $1 \leq i < j \leq \ell(t)$, we note that the probability for a randomly chosen partition of the clause t to include the (contiguous) phrase, $t_{i\cdots j}$, is determined by

successful cutting at the ends of $t_{i\cdots j}$ and failures within (e.g., x_2 must *not* be cut to produce "dog doctor"), accommodating for $t_{i\cdots j}$ reaching one or both ends of t, i.e.,

$$P_q(t_{i\cdots j} \,|\, t) = q^{2-b_{i\cdots j}}(1-q)^{\ell(s)-1} \tag{2.2}$$

where $b_{i\cdots j}$ is the number of the clause's boundaries shared by $t_{i\cdots j}$ and t. Allowing for a phrase $s \in S$ to have labeling equivalence to multiple contiguous regions (i.e., $s = t_{i\cdots j} = t_{i'\cdots j'}$, with $i, j \neq i', j'$) within a clause e.g., "ha ha" within "ha ha ha", we interpret the 'expected frequency' of s given the text by the double sum:

$$f_q(s \,|\, T) = \sum_{t\in T} f_q(s \,|\, t) = \sum_{t\in T} \sum_{s=t_{i\cdots j}} P_q(t_{i\cdots j} \,|\, t). \tag{2.3}$$

Departing from normal word counts, we may now have $f_q \ll 1$, except when one partitions for word ($q = 1$) or clause ($q = 0$) frequencies. When weighted by phrase length, the partition frequencies of phrases from a clause sum to the total number of words originally present in the clause:

$$\ell(t) = \sum_{1\leq i<j\leq \ell(t)} \ell(t_{i\cdots j}) P_q(t_{i\cdots j} \,|\, t), \tag{2.4}$$

which ensures that when the expected frequencies of phrases, s, are summed (with the length norm) over the whole text:

$$\sum_s \ell(s) f_q(s \,|\, T) = \sum_{t\in T} \ell(t) f(t), \tag{2.5}$$

the underlying mass of words in the text is conserved (see SI-2 for proofs of Eqs. 2.4 and 2.5). Said differently, phrase partition frequencies (random or otherwise) conserve word frequencies through the length norm ℓ, and so have a physically meaningful relationship to the words on "the page."

2.3 Statistical mechanical interpretation

Here, we focus on three natural kinds of partitions: $q = 0$: clauses are partitioned only as clauses themselves; $q = \frac{1}{2}$: what we call 'pure random partitioning'—all partitions of a clause are equally likely; $q = 1$: clauses are partitioned into words.

In carrying out pure random partitioning ($q = \frac{1}{2}$), which we will show has the many desirable properties we seek, we are assuming all partitions are equally likely, reminiscent of equipartitioning used in statistical mechanics (Goldenfeld, 1992). Extending the analogy, we can view $q = 0$ as a zero temperature limit, and $q = 1$ as an infinite temperature one. As an anchor for $f_{\frac{1}{2}}$, we note that words that appear once within a text—hapax legomena—will have $f_q \in \{\frac{1}{4}, \frac{1}{2}, 1\}$ (depending on clause boundaries), on the order of 1 as per standard word partitioning.

2.4 Experiments and Results

Before we apply the random partition theory to produce our generalization of word count, f_q, we will first examine the results of applying the random partition process in a 'one-off' manner. We process through the clauses of a text once, cutting word-word boundaries (and in a parallel experiment for graphemes, cutting letter-letter boundaries within words) uniformly at random with probability $q = \frac{1}{2}$.

In Fig. 2.1A, we present an example 'one-off' partition of the first few lines of Charles Dickens' "Tale of Two Cities" We give example partitions at the scales of clauses (red), pure random partition phrases (orange), words (yellow), pure random partition graphemes (green), and letters (blue). In Fig. 2.1B, we show Zipf distributions for all five partitioning scales. We see that clauses ($q = 0$) and pure random partitioning phrases ($q = \frac{1}{2}$) both adhere well to the pure form of $f \propto r^{-\theta}$ where r is rank. For clauses we find $\theta \simeq 0.78$ and for random partitioning, $\theta \simeq 0.98$ (see supplementary material for measurement details and

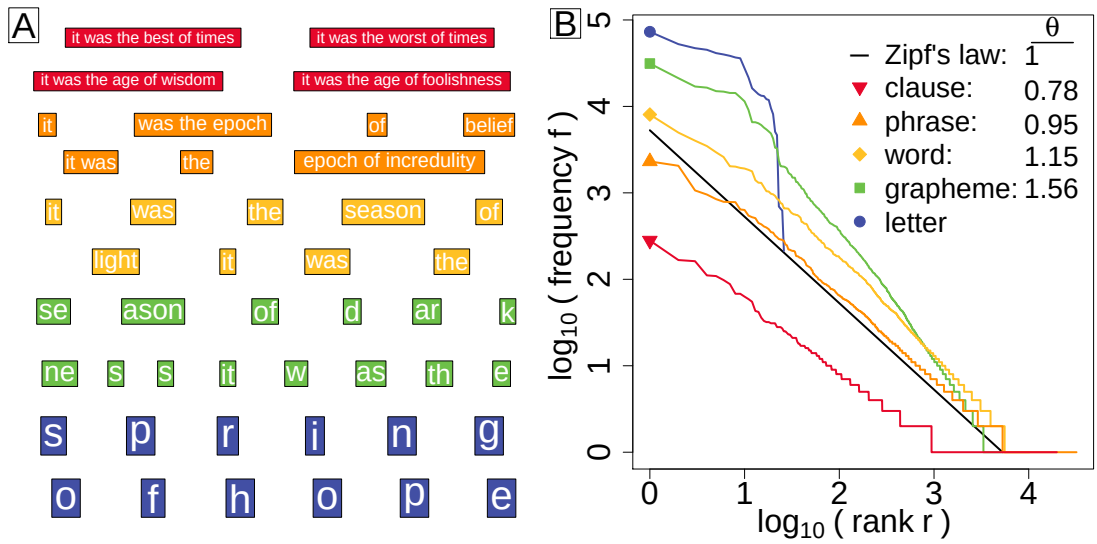


Figure 2.1: **A.** Partition examples for the start of Charles Dickens's "Tale of Two Cities" at five distinct levels: clauses (red), pure random partitioning phrases ($q = \frac{1}{2}$, orange), words (yellow), pure random partitioning graphemes ($q = \frac{1}{2}$, green), and letters (blue). The specific phrases and graphemes shown are for one realization of pure random partitioning. **B.** Zipf distributions for the five kinds of partitions along with estimates of the Zipf exponent θ when scaling is observed. No robust scaling is observed at the letter scale. The colors match those used in panel **A**, and the symbols at the start of each distribution are intended to strengthen the connection to the legend. See Ref. (Clauset et al., 2009) and supplementary material for measurement details.

for examples of other works of literature). The quality of scaling degrades as we move down to words and graphemes with the appearance of scaling breaks (Ferrer-i-Cancho and Solé, 2001; Gerlach and Altmann, 2013; Williams et al., 2014). Scaling vanishes entirely at the level of letters.

Moving beyond a single work, we next summarize findings for a large collection of texts (Project Gutenberg, 2010) in Fig. 2.2A, and compare the Zipf exponent θ for words and pure random $q = \frac{1}{2}$ 'one-off' partitioning for around 4000 works of literature. We plot the corresponding marginal distributions in Fig. 2.2B, and see that clearly $\theta \lesssim 1$ for $q = \frac{1}{2}$ phrases, while for words, there is a strong positive skew with the majority of values of $\theta > 1$. These steep scalings for words (and graphemes), $\theta > 1$, are not dynamically accessible for Simon's model (D'Souza et al., 2007).

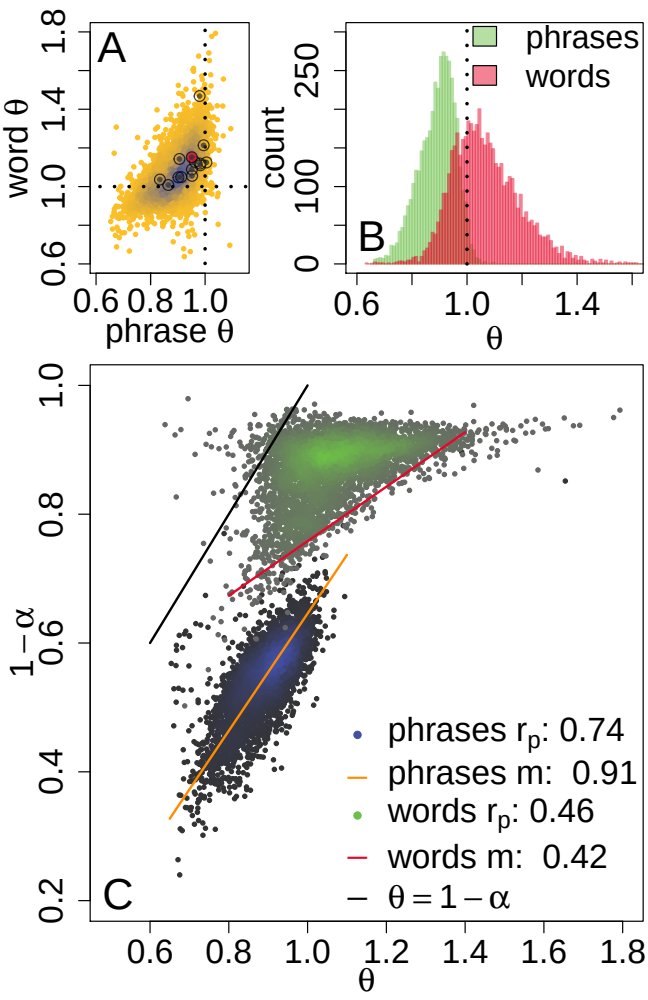


Figure 2.2: **A.** Density plot showing the Zipf exponent θ for 'one-off' randomly partitioned phrases and word Zipf distributions ($q = 1$ and $q = \frac{1}{2}$) for around 4000 works of literature. We indicate "Tale of Two Cities" by the red circle, and with black circles, we represent measurements for 14 other works of literature analyzed further in the supplementary material. **B.** Histograms of the Zipf exponent θ for the same set of books (marginal distributions for **A**). Phrases typically exhibit $\theta \leq 1$ whereas words produce unphysical $\theta > 1$, according to Simon's model **C.** Test of Simon's model's analytical connection $\theta = 1 - \alpha$, where θ is the Zipf exponent and α is the rate at which new terms (e.g., graphemes, words, phrases) are introduced throughout a text. We estimate α as the number of different words normalized by the total word volume. For both words and phrases, we compute linear fits using Reduced Major Axis (RMA) regression (Rayner, 1985) to obtain slope m, along with the Pearson correlation coefficient r_p. Words (green) do not exhibit a simple linear relationship whereas phrases do (blue), albeit clearly below the $\alpha = 1 - \theta$ line in black.

Leaving aside this non-physicality of Zipf distributions for words and concerns about breaks in scaling, we recall that Simon's model connects the rate, α, at which new terms are introduced, to θ in a simple way: $1 - \alpha = \theta$ (Simon, 1955). Given frequency data from a pure Simon model, the word/phrase introduction rate is determined easily to be $\alpha = N/M$, where N is the number of unique words/phrases, and M is the sum total of all word/phrase frequencies. We ask how well works of literature conform to this connection in Fig. 2.2C, and find that words (green dots) do not demonstrate any semblance of a linear relationship,

whereas phrases (blue dots) exhibit a clear, if approximate, linear connection between $1-\alpha$ and θ.

Despite this linearity, we see that a pure Simon model fails to accurately predict the phrase distribution exponent θ. This is not surprising, as when $\alpha \to 0$, an immediate adherence to the rich-get-richer mechanism produces a transient behavior in which the first few (largest-count) word varieties exist out of proportion to the eventual scaling. Because a pure Zipf/Simon distribution preserves $\theta = 1 - \alpha$, we expect that a true, non-transient power-law consistently makes the underestimate $1 - N/M < \theta$.

Inspired by our results for one-off partitions of texts, we now consider ensembles of pure random partitioning for larger texts. In Fig. 2.3, we show Zipf distributions of expected partition frequency, f_q, for $q = \frac{1}{2}$ phrases for four large-scale corpora: English Wikipedia, the New York Times (NYT), Twitter, and music lyrics (ML), coloring the main curves according to the length of a phrase for each rank. For comparison, we also include word-level Zipf distributions ($q = 1$) for each text in gray, along with the canonical Zipf distribution (exponent θ=1) for reference.

We observe scalings for the expected frequencies of phrases that hover around $\theta = 1$ for over a remarkable 7–9 orders of magnitude. We note that while others have observed similar results by simply combining frequency distributions of N-grams (Ha et al., 2002), these approaches were unprincipled as they over-counted words. For the randomly partitioned phrase distributions $f_{\frac{1}{2}}$, the scaling ranges we observe persist down to 10^{-2}, beyond the hapax legomena, which occur at frequencies greater than 10^{-1}. Such robust scaling is in stark contrast to the very limited scaling of word frequencies (gray curves). For pure word partitioning, $q = 1$, we see two highly-distinct scaling regimes exhibited by each corpus, with shallow upper (Zipf) scalings at best extending over four orders of magnitude, and typically only three. (In a separate work, we investigate this double scaling finding evidence that text-mixing is the cause (2014).)

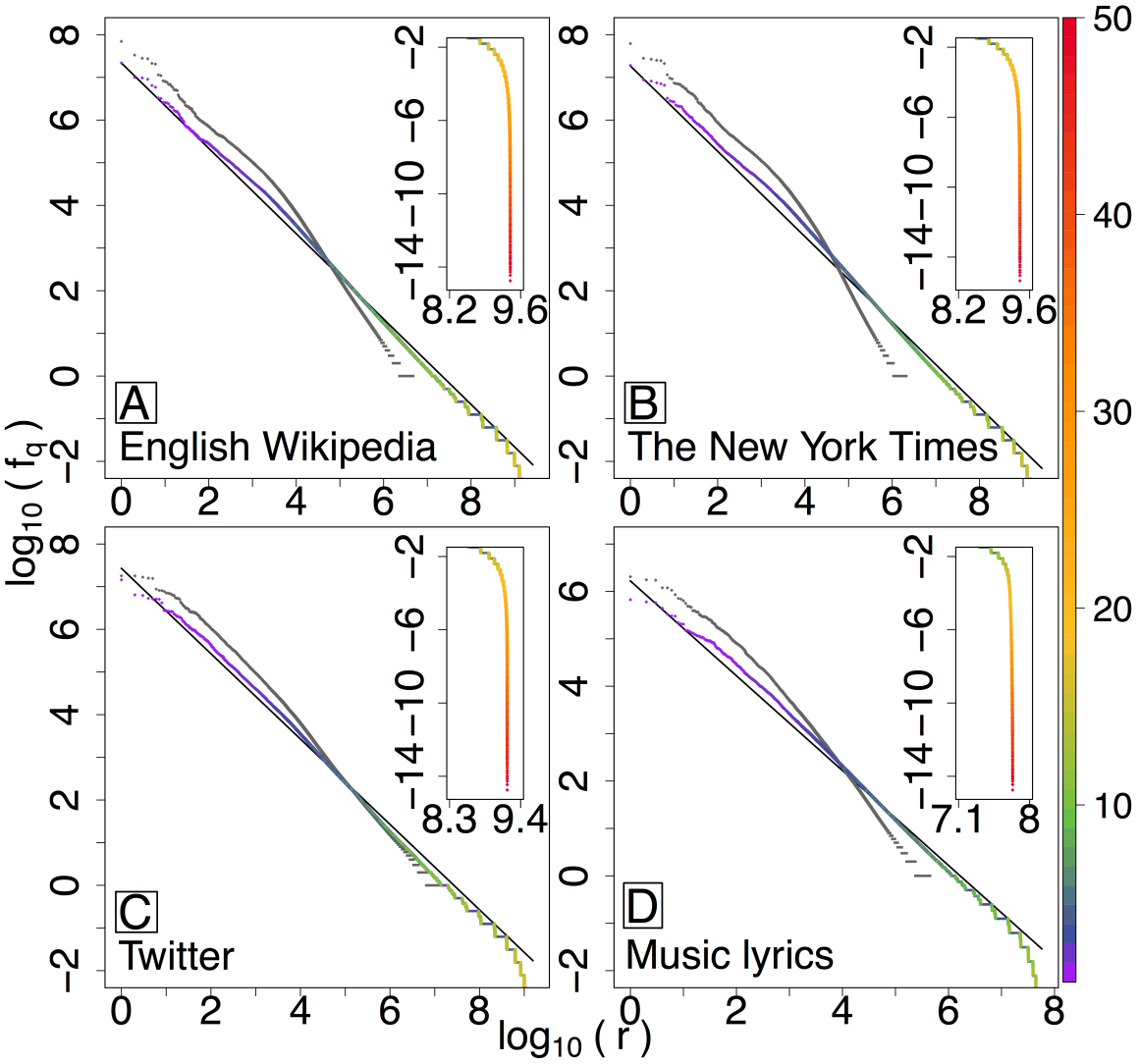


Figure 2.3: Random partitioning distributions ($q = \frac{1}{2}$) for the four large corpora: (A) Wikipedia (2010); (B) The New York Times (1987–2007); (C) Twitter (2009); and (D) Music Lyrics (1960–2007). Top right insets show the long tails of random partitioning distributions, and the colors represent phrase length as indicated by the color bar. The gray curves are standard Zipf distributions for words ($q = 1$), and exhibit limited scaling with clear scaling breaks. See main text and Tabs. A.1–A.4, for example phrases.

For all four corpora, random partitioning gives rise to a gradual interweaving of different length phrases when moving up through rank r. Single words remain the most frequent (purple), typically beginning to blend with two word phrases (blue) by rank $r = 100$. After the appearance of phrases of length around 10–20, depending on the corpus, we see the phrase rank distributions fall off sharply, due to long clauses that are highly unique in their construction (upper right insets).

In Appendix A, we provide structured tables of example phrases extracted by pure random partitioning for all four corpora (Tabs. A.1–A.4), along with complete phrase data sets. As with standard N-grams, the texture of each corpus is quickly revealed by examining phrases of length 3, 4, and 5. For example, the second most common phrases of length 5 for the four corpora are routinized phrases: "the average household size was" (EW), "because of an editing error" (NYT), "i uploaded a youtube video" (TW), and "na na na na na" (ML). By design, random partitioning allows us to quantitatively compare and sort phrases of different lengths. For music lyrics, "la la la la la" has an expected frequency similar to "i don't know why", "just want to", "we'll have", and "whatchu" (see Tab. A.4), while for the New York Times, "the new york stock exchange" is comparable to "believed to have" (see Tab. A.2).

2.5 Discussion

The phrases and their effective frequencies produced by our pure random partitioning method may serve as input to a range of higher order analyses. For example, information theoretic work may be readily carried out, context models may be built around phrase adjacency using insertion and deletion, and specific, sentence-level partitions may be realized from probabilistic partitions.

While we expect that other principled, more sophisticated approaches to partitioning texts into rankable mixed phrases should produce Zipf's law spanning similar or more orders

of magnitude in rank, we believe random partitioning—through its transparency, simplicity, and scalability—will prove to be a powerful method for exploring and understanding large-scale texts.

To conclude, our results reaffirm Zipf's law for language, uncovering its applicability to a vast lexicon of phrases. Furthermore, we demonstrate that the general semantic units of statistical linguistic analysis can and must be phrases—not words—calling for a reevaluation and reinterpretation of past and present word-based studies in this new light.

2.6 References

Axtell, R., 2001. Zipf distribution of U.S. firm sizes. Science 293 (5536), 1818–1820.

Barabási, A. L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–511.

Batty, M., 2008. The size, scale, and shape of cities. Science Magazine 319 (5864), 769–771.

Bornholdt, S., Ebel, H., 2001. World Wide Web scaling exponent from Simon's 1955 model. Phys. Rev. E 64, 035104(R).

Clauset, A., Shalizi, C. R., Newman, M. E. J., 2009. Power-law distributions in empirical data. SIAM Review 51, 661–703.

Coromina-Murtra, B., Solé, R., 2010. Universality of Zifp's law. Phsyical Revew E 82, 011102.

Cougar Town, 2013. I should have known it. Cougar Town, season 4, episode 4: http://www.imdb.com/title/tt2483134/.

de Solla Price, D. J., 1976. A general theory of bibliometric and other cumulative advantage processes. J. Amer. Soc. Inform. Sci. 27, 292–306.

D'Souza, R. M., Borgs, C., Chayes, J. T., Berger, N., Kleinberg, R. D., 2007. Emergence of tempered preferential attachment from optimization. Proc. Natl. Acad. Sci. 104, 6112–6117.

Ferrer-i-Cancho, R., Elvevåg, B., 03 2010. Random texts do not exhibit the real Zipf's law-like rank distribution. PLoS ONE 5, e9411.

Ferrer-i-Cancho, R., Solé, R. V., 2001. Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. Journal of Quantitative Linguistics 8 (3), 165–173.

Gerlach, M., Altmann, E. G., 2013. Stochastic model for the vocabulary growth in natural languages. Phys. Rev. X 3, 021006.

Goldenfeld, N., 1992. Lectures on Phase Transitions and the Renormalization Group. Vol. 85 of Frontiers in Physics. Addison-Wesley, Reading, Massachusetts.

Google, 2014. http://ngrams.googlelabs.com/.

Ha, L. Q., Sicilia-Garcia, E. I., Ming, J., Smith, F. J., 2002. Extension of Zipf's law to words and phrases. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING). pp. 315–320.

Maillart, T., Sornette, D., Spaeth, S., von Krogh, G., 2008. Empirical tests of Zipf's law mechanism in open source Linux distribution. Phys. Rev. Lett. 101 (21), 218701.

Mandelbrot, B. B., 1953. An informational theory of the statistical structure of languages. In: Jackson, W. (Ed.), Communication Theory. Butterworth, Woburn, MA, pp. 486–502.

Miller, G. A., 1957. Some effects of intermittent silence. American Journal of Psychology 70, 311–314.

Project Gutenberg, 2010. http://www.gutenberg.org.

Rayner, J. M. V., 1985. Linear relations in biomechanics: the statistics of scaling functions. J. Zool. Lond. (A) 206, 415–439.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A., Flickinger, D., 2002. Multiword expressions: A pain in the neck for NLP. In: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing. CICLing '02. Springer-Verlag, London, UK, pp. 1–15.

Simon, H. A., 1955. On a class of skew distribution functions. Biometrika 42, 425–440.

Williams, J. R., Bagrow, J. P., Danforth, C. M., Dodds, P. S., 2014. Text mixing shapes the anatomy of rank-frequency distributions: A modern zipfian mechanics for natural language. CoRRhttp://arxiv.org/abs/1409.3870.

Zanette, D. H., Manrubia, S. C., 2001. Vertical transmission of culture and the distribution of family names. Physica A 295, 1–8.

Zipf, G. K., 1935. The Psycho-Biology of Language. Houghton-Mifflin.

Zipf, G. K., 1949. Human Behaviour and the Principle of Least-Effort. Addison-Wesley.

Chapter 3

Text mixing shapes the anatomy of rank-frequency distributions: A modern Zipfian mechanics for natural language

Natural languages are full of rules and exceptions. One of the most famous quantitative rules is Zipf's law which states that the frequency of occurrence of a word is approximately inversely proportional to its rank. Though this 'law' of ranks has been found to hold across disparate texts and forms of data, analyses of increasingly large corpora over the last 15 years have revealed the existence of two scaling regimes. These regimes have thus far been explained by a hypothesis suggesting a separability of languages into core and non-core lexica. Here, we present and defend an alternative hypothesis, that the two scaling regimes result from the act of aggregating texts. We observe that text mixing leads to an effective decay of word introduction, which we show provides accurate predictions of the location and severity of breaks in scaling. Upon examining large corpora from 10 languages in the Project Gutenberg eBooks collection (eBooks), we find emphatic empirical support for the universality of our claim.

3.1 Zipf's law and (non) universality

Given some collection of distinct kinds of objects occurring with frequency f and associated rank r according to decreasing frequency, Zipf's law is said to be fulfilled when ranks and frequencies are approximately inversely proportional:

$$f(r) \sim r^{-\theta}, \tag{3.1}$$

typically with $\theta \simeq 1$. Though Zipf's functional form has been found to be a reasonable one for disparate forms of data, ranging from frequencies of words to sizes of cities in Zipf's original work (1935; 1949), its lack of *total* universality in application to natural languages is now widely acknowledged (Ferrer-i-Cancho and Solé, 2001; Montemurro, 2001; Gerlach and Altmann, 2013; Kwapien et al., 2010; Petersen et al., 2012; Williams et al., 2014).

Recently it was suggested (Ferrer-i-Cancho and Solé, 2001; Montemurro, 2001) that large corpora exhibit two scaling regimes (delineated by some $b > 0$):

$$f(r) \sim \begin{cases} r^{-\theta}, & : r \leq b \\ r^{-\gamma}, & : r > b \end{cases}, \tag{3.2}$$

the first being that of Zipf ($\theta = 1$) and the second distinctly more variable (Montemurro, 2001), (though generally $\gamma > 1$). Ferrer-i-Cancho and Solé hypothesized in (2001) that these two regimes reflected a division of natural languages into two lexical subsets—the kernel (core) and unlimited (non-core) lexica.

We observe that in all studies finding dual scalings that the texts analyzed are of mixed origin, that is, they are not derived from a single author, or even a single topic. Montemurro indicated in 2001 that combining heterogeneous texts could generate effects that shield investigators from the true underlying nature of this second scaling regime:

> To resolve the behavior of those [high rank] words we need a significant increase in volume of data, probably exceeding the length of any conceivable single text. Still, at the same time it is desirable to maintain as high a degree of homogeneity in the texts as possible, in the hope of revealing a more complex phenomenology than that simply originating from a bulk average of a wide range of disparate sources.

With this inspiration, we focus on understanding the effects of combining texts of varying heterogeneity—a process we refer to as "text mixing".

3.2 Stochastic models

In the years following Zipf's original work, various stochastic models have been proposed for the generation of natural language vocabularies. The first of these was that proposed by Simon (1955), and based on Yule's model of evolution (1924). This work is a powerful companion to understanding Zipf's empirical work, and can be seen as the natural antecedent of the rich-gets-richer models (Barabási and Albert, 1999; Krapivsky and Redner, 2001) for growing networks that have interested the complex systems community over recent years. Indeed, perhaps the most important piece we may draw from Simon's model is that a rich-gets-richer mechanism is a reasonable one for the growth of a vocabulary.

An important limitation of Simon's model is that it is only capable of producing a single scaling regime, which, as we know is an incomplete picture. Furthermore, the scalings accessible via the Simon model were strictly less severe than the 'universal' $\theta = 1$ exponent. So, if one assumes the Simon model as truth, with a fixed word introduction rate α_0, Zipf's exponent should be variable and necessarily less than 1, though empirically found indistinguishable from 1, that is $\theta = 1 - \alpha_0$, with $\alpha_0 \ll 1$ (Simon, 1955).

Recently, a modification to Simon's model was proposed in which two types of words could be produced—core and non-core words (Gerlach and Altmann, 2013). As a built-in feature of the core/non-core vocabulary (CNCV) model, the size of the core set of words was prescribed to be finite, while the non-core was allowed to expand indefinitely. Aside from introducing two classes of words, the most important distinction of this model from its predecessor was a rule for the decay in the rate of introduction of new words, α. Along with producing the CNCV model they showed that when α decays as a power-law with exponent $-\mu$, of the number of unique words, n, the relationship between μ and the lower rank-frequency exponent, γ, is a difference of θ, i.e.,

$$\alpha(n) = \alpha_0 \cdot n^{-\mu} \Rightarrow f(r) \sim r^{-(\theta+\mu)}, \tag{3.3}$$

with $\gamma = \theta + \mu$ (Gerlach and Altmann, 2013). The distinction between word types provided a means for postponing the point at which their power law decay would occur, thereby generating two scaling regimes. We note that the severity of the second scaling was *only* contingent upon the existence of a decay in the rate of introduction of new words, and that this decay was imposed, rather than the result of the existence of two word types. We are therefore led to find an explicit mechanism capable of producing power-law decaying word introduction rates, and hence multiple scaling regimes.

3.3 Text mixing

As we have described, the CNCV model offers a means by which one can obtain a second scaling. The model is, like Simon's, framed as a model of the generation of a vocabulary. However, we are led to question whether lower scalings are a product of vocabulary generation or an artifact of an interaction between disparate texts. Suppose a collection of texts, $\mathcal{C} = \{T_1, ..., T_k\}$, is read sequentially, and that each has rank-frequency distribution

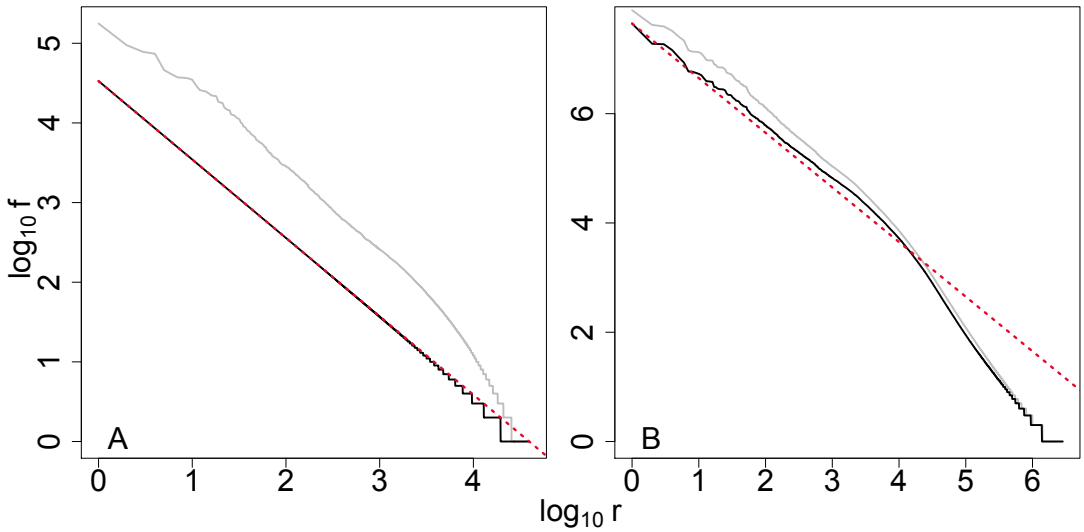


Figure 3.1: (**A**) An idealization (black points) of a rank-frequency distribution (gray points) for a single text[1] from the English eBooks collection. Idealization is defined by a pure power law of scaling $1 - N/M$ (red dashed line, see Materials and Methods). (**B**) The mixtures of all texts (gray points) and their idealizations (black points) from the English eBooks collection. Note that neither mixture results in a pure power law such as Zipf's ($\theta = 1$, red, dashed line).

of Zipf/Simon form. Upon constructing idealized rank-frequency distributions from empirical data (see Sec. 4.5), we find that their combined distribution possesses multiple scaling regimes (see Fig. 3.1). Though each individual vocabulary might have been created without a decay of word introduction, an overlap in the words they use has it *seem* as though the appearance of new words is rarer by the time the later texts are read. If one reads the texts repeatedly and in permuted orders, the resulting decay in the rate of word introduction likely does not evince itself until the mean text size (mean number of unique words per text) is reached, but certainly not before the minimum text size is reached.

Operating under this ansatz—that a text mixing-derived scaling break, b, covaries with the mean number of unique words per text, N_{avg}, in a corpus—we investigate thousands of corpora defined by samples from the English eBooks database (see Sec. 4.5 for more details on text sampling and a complete description of the eBooks database). Obtaining $1,000$ text-sample corpora from each of the 10 deciles of the text-size distribution, we regress

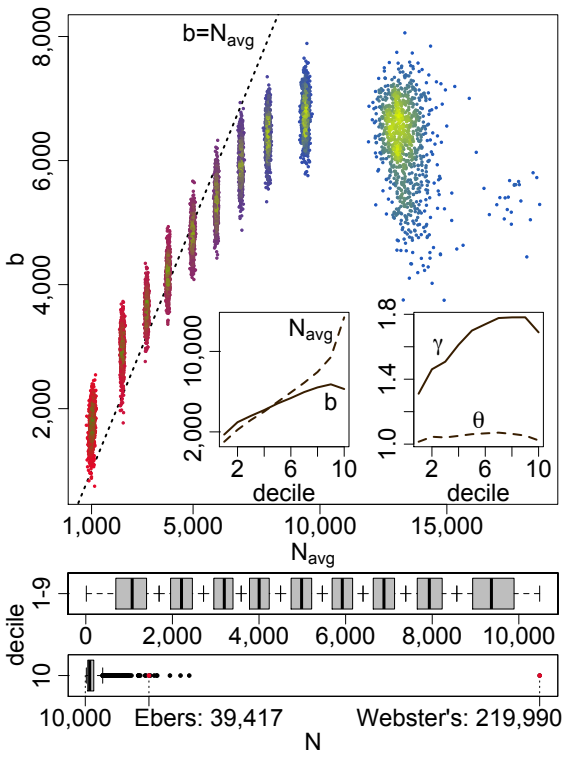


Figure 3.2: (**Top**) For each of the 10 deciles of the English distribution of text sizes, we measure the parameters b, γ, N_{avg}, and θ from 50-book sample corpora. Each cloud represents $1,000$ sample corpora from deciles 1–10 (low-to-high from left to right, where red to blue also indicates increasing decile and fade to green or yellow indicates increasing density). The line $b = N_{\text{avg}}$ is also presented (dashed line, main axis), and shows that b increases with decile for all but the most extreme (10^{th}) decile. Main axes insets show parameter variation across deciles for both b and N_{avg} (left); and γ and θ (right), where we note that Zipf's parameter, θ, is the only one that exhibits signs of stationarity. (**Bottom**) Box plots providing a more detailed look at the ten deciles of the distribution of text sizes. For clarity we have separated the plots for deciles 1–9 from the 10^{th}. This highlights the extreme nature of the later deciles (most notably the 10^{th}), where the presence of poorly refined texts throw off estimates of N_{avg}, which we also note corresponds to the roll over in the distributions off of the $b = N_{\text{avg}}$ axis above.

for b (see Sec. 4.5), and record N_{avg} to find that the two covary strongly along the line $b = N_{\text{avg}}$ for all but the most extreme deciles (see main axes Fig. 3.2, which we return to later in the discussion). We see that this relationship breaks down in the presence of large-N texts, which upon closer inspection appear ill formed in the sense of being of mixed origin themselves (e.g., posthumous/longitudinal compendia, dictionaries, encyclopedias, etc...; see Sec. 4.5 and Fig. 3.6 for more details on corpus formation and internally-mixed texts). Additionally, we see from these preliminary experiments that both of the quantities, b and γ, do not appear as universal for a given language (see Fig. 3.2), but rather depend quite severely on corpus composition. In fact, the only regressed parameter that presents any signs of universality for a language is Zipf's exponent, θ, which remains quite close to 1. These initial results indicate that hypotheses of the locations of scaling breaks, b,

Consider the two excerpts from Charles Dickens' "A Tale of Two Cities", taken as texts:

$$T_1 : (it, was, the, best, of, times, it, was, the, worst, of, times), \quad \text{and}$$
$$T_2 : (it, was, the, age, of, wisdom, it, was, the, age, of, foolishness)$$

Supposing we read T_1 first, the sequence of words is:

$$(T_1, T_2) : (it, was, the, best, of, times, it, was, the, worst, of, times,$$
$$it, was, the, age, of, wisdom, it, was, the, age, of, foolishness)$$

where we have highlighted initial (growing text) word appearances in red. The corresponding sequences of values, $m, n_m, N_m, \alpha_m, A_m$ and α_m/A_m, are then

$$m : (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24)$$
$$n_m : (1, 2, 3, 4, 5, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 8, 8, 9, 9, 9, 9, 9, 9, 10)$$
$$N_m : (1, 2, 3, 4, 5, 6, 6, 6, 6, 7, 7, 7, 8, 9, 10, 11, 12, 13, 13, 13, 13, 13, 13, 14)$$
$$\alpha_m : (1, 1, 1, 1, 1, 1, \tfrac{6}{7}, \tfrac{6}{8}, \tfrac{6}{9}, \tfrac{7}{10}, \tfrac{7}{11}, \tfrac{7}{12}, \tfrac{7}{13}, \tfrac{7}{14}, \tfrac{7}{15}, \tfrac{8}{16}, \tfrac{8}{17}, \tfrac{9}{18}, \tfrac{9}{19}, \tfrac{9}{20}, \tfrac{9}{21}, \tfrac{9}{22}, \tfrac{9}{23}, \tfrac{10}{24})$$
$$A_m : (1, 1, 1, 1, 1, 1, \tfrac{6}{7}, \tfrac{6}{8}, \tfrac{6}{9}, \tfrac{7}{10}, \tfrac{7}{11}, \tfrac{7}{12}, \tfrac{8}{13}, \tfrac{9}{14}, \tfrac{10}{15}, \tfrac{11}{16}, \tfrac{12}{17}, \tfrac{13}{18}, \tfrac{13}{19}, \tfrac{13}{20}, \tfrac{13}{21}, \tfrac{13}{22}, \tfrac{13}{23}, \tfrac{14}{24})$$
$$\frac{\alpha_m}{A_m} : (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, \tfrac{7}{8}, \tfrac{7}{9}, \tfrac{7}{10}, \tfrac{8}{11}, \tfrac{8}{12}, \tfrac{9}{13}, \tfrac{9}{13}, \tfrac{9}{13}, \tfrac{9}{13}, \tfrac{9}{13}, \tfrac{9}{13}, \tfrac{10}{14}).$$

Example 3.1: A concrete example of the text mixing effect, where we consider two passages (T_1 and T_2) as separate texts that are then mixed. The similarity of word use between these excerpts provides an excellent example for understanding the differences between the growing text, where we count new word appearances (n_m) with the awareness of previous texts, and the memoryless text, where we count word appearances (N_m) as new with each initial appearance in each text. Note that both α_m and A_m are simply the quotients of n_m and N_m with m (respectively), and that their quotient (α_m/A_m) is equivalent to n/N, and is not equal to 1 *only* when texts are mixed.

corresponding to language-universal lexical-core sizes are in strong need of reevaluation, or should be reformulated as corpus-relative.

In the following, we run text mixing experiments that measure decay in rates of word introduction directly attributable to mixing texts to predict lower scalings in composite distributions. As we read out texts (in some order) let m be the volume of words observed

at any point, and n_m be the number of distinct words in the volume m, which we will refer to as the vocabulary size of the *growing* text. To exhibit the effects of text mixing we contrast the vocabulary size of the growing text with the vocabulary size of the *memoryless* text, N_m, where we "forget" the words read in all previous texts and continuing counting appearances of words that were initial *in their text* (regardless of appearances in previous texts). From n_m and N_m we then have two proxies for the word introduction rate, one for the growing text $\alpha_m = n_m/m$ and one for the memoryless text $A_m = N_m/m$. We may consider α_m to be the word introduction rate of the composite (which includes mixing effects), and A_m to be the word introduction rate of the individual texts (excluding mixing effects).

There are many conceivable mechanisms that lead to a power-law decay in the rate of word introduction. To measure the severity of scaling breaks we do not need to know the true values of the word introduction rates, but instead just their scalings. So, to determine the extent to which text mixing generates word introduction decay, we isolate the portion of the scaling that results from mixing by measuring α_m/A_m, the portion of word introduction remaining after mixing texts. Note that since $n_m \leq N_m$, one has $\alpha_m \leq A_m$, and hence $\alpha_m/A_m \leq 1$ for all m. Hence, this normalized rate behaves as a non-constant only when mixing ensues, and so any decay measured via α_m/A_m implies the presence and is the direct consequence of text mixing (see Example 3.1 for an intuitive understanding of all text mixing quantities). Since α_m/A_m will be the only quantity used in the measurement of word introduction decay, we relax the notation, and simply write α for α_m/A_m and n for n_m in what follows.

To test the effects of text mixing, we not only observe the word introduction rate $\alpha(n)$, but consider its ability to predict the scalings of rank-frequency distributions. To do this, we note that by design, the data for $\alpha(n)$ are aligned with $f(r)$—both have domain $\{1, ..., N_{\text{corp}}\}$ (where N_{corp} is the vocabulary size of the corpus). Further, since the theory has $\gamma = \theta + \mu$,

we may also observe that $\alpha(n) \cdot n^{-\theta}$, need only be normalized

$$\hat{p}(n) = \frac{\alpha(n) \cdot n^{-\theta}}{C}, \text{ where } C = \sum_{1}^{N_{\text{corp}}} \alpha(n) \cdot n^{-\theta} \tag{3.4}$$

to produce a model for the normalized rank-frequency distribution $p(r) = f(r)/\sum_{1}^{N_{\text{corp}}} f(r)$. To determine a model's Zipf scaling, θ, we scan the range $\{0.75, 0.751, ..., 1.25\}$ and accept the θ for which $\hat{p}$ minimizes the sum of squares error

$$\sum_{1}^{N_{\text{corp}}} (\log_{10} p(r) - \log_{10} \hat{p}(r))^2 \tag{3.5}$$

over as many as $10,000$ log-spaced ranks.

3.4 Materials and methods

In our experiments we worked with a subset of the eBooks (2010) collection. We collected those texts which were annotated sufficiently well to allow for the removal of meta-data as well as for the parsing of authorship, title, and language. All together, this resulted in the inclusion of $23,309$ books from across ten languages (broken down in Tab. 3.1).

To idealize texts as discussed in Fig. 3.1 we note that a resultant rank-frequency distribution from a pure Simon model of constant word introduction rate, α_0, will scale with Zipf exponent $\theta = 1 - \alpha_0$, such that $N/M \to \alpha_0$ as the text grows. Therefore, for an observed text of size N and volume M, we define the idealized Zipf/Simon exponent as $\theta_0 = 1 - N/M$, and apply θ_0 to the collection of ranks, $r = 1, \cdots, N$, as

$$f_{\text{ideal}}(r) = \left\lfloor \left(\frac{r}{N}\right)^{-\theta_0} + \frac{1}{2} \right\rfloor, \tag{3.6}$$

while preserving their word-labels from the empirical data.

For all of the rank-frequency distributions analyzed, we regress over as many as $10,000$ log-spaced ranks (taken over the range $r = 1, ..., N$) to determine estimates for θ, b, and γ. This estimation is done by applying a two-line least-squares regression, constrained by intersection at the point of scaling break. Given data points (x, y), and a point of break, x_b, we solve for the model

$$\hat{y} = \begin{cases} \beta_1 + \beta_2 x, & : x \leq x_b \\ \beta_3 + \beta_4 x, & : x > x_b \end{cases}, \tag{3.7}$$

constrained by $\beta_1 + \beta_2 x_b = \beta_3 + \beta_4 x_b$, through standard minimization of the sum of squares error. We compute this regression for $1,000$ log-spaced points, x_b, across the middle 20–80% of the log r domain. For given distribution we then perform these $1,000$ regressions and accept the value b for which we have observed the smallest SSE.

To understand our text mixing results we must note that there is measurement error for both b and N_{avg}. As a regressed quantity, this may be expected for b, but for N_{avg}, the existence of measurement error is less obvious, and generally results from poor corpus composition. The main effect stems from the fact that many texts in the eBooks data set

	N_{books}	N_{char}	N_{min}	N_{ave}	b	N_{max}	N_{corp}
en	19,793	46	5	5,899.3	5,849	219,990	2,836,900
fr	1,360	44	395	8,300.7	17,715	26,171	528,314
fi	505	31	1,144	8,872.6	7,761	31,623	811,742
nl	434	48	133	6,747.1	6,098	82,246	443,816
pt	375	38	203	4,675.8	10,363	17,818	246,497
de	327	30	153	7,554.9	7,259	113,089	477,274
es	223	34	406	8,735.1	15,079	29,452	237,874
it	194	29	1,083	9,388.7	13,954	29,445	258,509
sv	56	34	1,389	7,499.8	5,315	18,726	123,806
el	42	35	2,047	6,414.7	7,613	17,774	110,940

Table 3.1: Table of information concerning the data used from the eBooks database. For each language we record the number of books (N_{books}); the number of characters (N_{char}), which we take to be the number of letters (Wikipedia Latin Alphabets, 2014; Wikipedia Greek Alphabet, 2014) (including diacritics and ligatures); the minimum text size (N_{min}); the maximum text size (N_{max}); and the total corpus size (N_{corp}). For reference, we additionally record the regressed point of scaling break, b.

are internally mixed. The longitudinal compendia of individual authors and genres are the most intuitive and abundant examples of internally mixed texts, and the most extreme cases are generally reference texts, e.g., dictionaries, encyclopedias, and textbooks (see Fig. 3.2). The major point is that when a compendium is not refined, but taken as an individual text in a corpus, the calculation of N_{avg} considers only a single book of large size (wrongly), instead of many books of smaller size (correctly). Within the English data set we have found that the large-N texts are generally of this variety and dominate the 10^{th} decile. Reading down the top ten N-ranking texts makes this abundantly clear:

1. Webster's Unabridged Dictionary
2. Diccionario Ingles-Español-Tagalog
3. The Complete Project Gutenberg Works of George Meredith
4. The Anatomy of Melancholy
5. A Concise Dictionary of Middle English
6. A Pocket Dictionary
7. The Nuttall Encyclopaedia
8. The Complete PG Works of Oliver Wendell Holmes, Sr.
9. The Complete Historical Romances of Georg Ebers
10. The Complete Project Gutenberg Works of Galsworthy

Note here that among these compendia and reference texts lies a two way (Spanish/English) dictionary whose placement in the top 10 likely results from dual word forms (English and Spanish translations) of the majority of words that it possesses. We have explored the impact of these under-refined and ill-formed texts in detail in Fig. 3.2, where we have found a clear association of b with N_{avg} along the line $b = N_{\text{avg}}$ that breaks down in the larger deciles, where these strange texts occur.

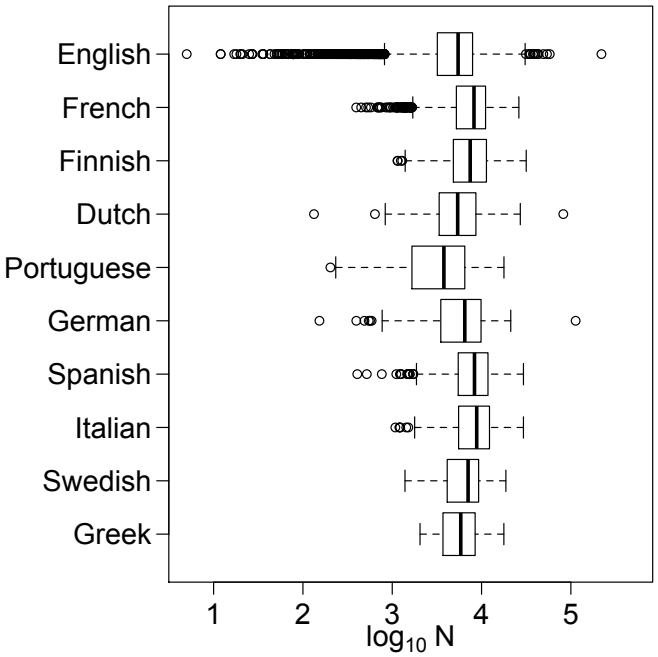


Figure 3.3: Box plots of the base ten logarithm vocabulary sizes of the texts contained in the 10 eBooks corpora studied. Center bars indicate means and whiskers extend to most extremal values up to 1.5 times the I.Q.R. length, whereupon more extremal values are plotted as points designated 'outliers'.

We also note that N_{avg} is subject to measurement error from overrefined texts as well, most notably in the Portuguese data set, which has the smallest average text size, while having the fifth largest number of books (see Tab. 3.1 and Fig. 3.3). There we note that Portuguese presents the most significant deviation between N_{avg} and b (b is notably more than 120% larger than N_{avg}), and moreover that this deviation is in the expected direction, i.e., $N_{\text{avg}} \ll b$. Note also that this observation is in agreement with those other languages that have $N_{\text{avg}} \ll b$ in Tab. 3.1 (specifically Italian, Spanish, and French), where in Fig. 3.3 we see that having many low-N outliers with no high-N outliers biases the corpus-wide measurement of N_{avg}.

To estimate μ we perform common least squares linear regression on the log-transformed data over the region $[N_{\text{avg}}, N_{\text{corp}}]$, since N_{avg} is generally the point at which mixing-derived decay becomes clear.

Computation of $\alpha(n)$ involves running many realizations of the text mixing procedure, randomizing the order in which the texts are read. To ensure that our measurements are

accurate, we adhere to a heuristic—that the number of text mixing runs be no less than $10 \cdot N_{\text{books}}$ for the given corpus. The final values used is in our experiments are computed as averages of the α_m/A_m from the more than $10 \cdot N_{\text{books}}$ runs. However, we note that $\alpha_m/A_m = n_m/N_m$, where n_m ranges with rank: $n_m = 1, 2, 3, \cdots, N_{\text{corp}}$. So, the only quantities that vary across runs that are necessary to compute $\alpha(n)$ are the N_m. Hence we take the average as $\alpha(n_m) = n_m/\langle N_m \rangle$ (where $\langle N_m \rangle$ indicates the average N_m of the memoryless text across runs), which is in fact the harmonic mean of the $\alpha(n_m)$ (the truest mean for rates).

In our investigation of the different divisions of the internally mixed corpus, "The complete historical romances of Georg Ebers," we have shown how important it is to have meaningfully defined texts to be able to produce an accurate text mixing model for a corpus. An important component of this exhibition presented the extremal refinement, where each word is treated individually as a separate text (a highly non-realistic scenario). To conduct a text mixing experiment for such a refinement can be quite computationally taxing, as this requires taking permutations of the word orders of the entire corpus. Since this process is entirely independent of the original word orderings from the corpus, it may be computed directly from the rank-frequency distribution via expected gap sizes. In particular, we wish to determine the average number of previously seen words appearing between the n^{th} and $n+1^{\text{st}}$ "new" words, given all permutations of the corpus words. Denoting this number by $\overline{M}_n$, we note that the average word introduction rate over this range is easily found as $\alpha_n = 1/\overline{M}_n$. We then define i_n as the total number of previously-observed words that were not yet counted by the time the n^{th} new word was observed, and define j_n to be the total number (out of all corpus words) that were not yet counted by the time the n^{th} new word was first observed (including those word types that were not yet observed). Then, if $P_n(M)$ is the probability that the n^{th} and $n+1^{\text{st}}$ "new" words were separated by

precisely M previously seen words,

$$\begin{aligned}\overline{M}_n &= \sum_{M=0}^{i_n} M \cdot P_n(M) \\ &= \sum_{M=0}^{i_n} M \cdot \frac{j_n - i_n}{j_n - M} \prod_{k=0}^{M-1} \frac{i_n - k}{j_n - k}\end{aligned} \tag{3.8}$$

where in the last expression, the product is the probability of seeing M consecutive previously-observed words, with the first factor being the probability that the "new" word is seen as the $M + 1^{\text{st}}$. These expressions for the $\overline{M}_n$ are iteratively computable, and in addition, since the sums appear (empirically) to converge quickly, we find that it suffices to take their first $1,000$ terms for added computational efficiency.

3.5 Results and discussion

To understand our results we define N_{min}, N_{avg} and N_{max} as the minimum, average, and maximum text sizes (by numbers of unique words) respectively (see Tab. 3.1). These three values obviate four text mixing regimes:

$$\begin{aligned} n < N_{\text{min}};&\ \text{Zipf/Simon (no mixing)} \\ N_{\text{min}} \leq n \leq N_{\text{avg}};&\ \text{initial (minimal mixing)} \\ N_{\text{avg}} \leq n \leq N_{\text{max}};&\ \text{crossover (partial mixing)} \\ n > N_{\text{max}};&\ \text{terminal (full mixing)} \end{aligned}$$

In the Zipf/Simon regime we expect the result of an unperturbed Simon model, but because mixing is also minimal over the initial regime, we expect that behavior over the first two regimes to more or less be consistent. Once in the crossover regime, words will on average have appeared under the effects of text mixing and so there is the expectation that N_{avg}

will mark the macroscopically observable change in behavior, or scaling break of the rank-frequency distribution, i.e., we expect $b \approx N_{\text{avg}}$. Plotting the two against one another, we have see this relationship holds across sample corpora from the well-behaved deciles of the English distribution of text sizes (see Fig. 3.2), and breaks down in the presence of ill-formed texts. Finally, over the terminal regime, all words will appear in the presence of mixing, and so this regime exhibits the stabilized second scaling, characterized by the decay parameter μ.

Our main results from text mixing, comparing the text mixing-derived model, $\hat{p}$, with the normalized empirical rank-frequency data, p, may be found for the English data set in Fig. 3.4, and for the nine other languages studied in Fig. 3.5. For all 10 languages we observe that the models defined by text mixing, $\hat{p}$, produce excellent predictions of the rank-frequency distributions (Main axes, Figs. 3.4 and 3.5), which is made quite clear by plotting point-wise squared error (lower-left insets, Figs. 3.4 and 3.5). For each corpus we see a broad range of ranks beginning not far before 10^2, and extending into the second scaling where the error is quite low (disregarding the effect of the finite-size plateaux).

We also perform text mixing analysis at different scales for a single, large, and internally mixed text from the English data set, "The complete historical romances of Georg Ebers." It is important to note before interpreting these results that the text itself is a compendium, combining series' that were each written by the author over the course of more than 30 years, writing and publishing volumes independently. With this in mind, the text offers an important example for text mixing that helps us to understand several important details. First, that not all texts are well formed—an individual text such as this may in and of itself present a scaling break that has resulted from text mixing. Second, that the scaling break of a single, large text may be understood through text mixing analysis. This second point is more difficult to observe, as it requires an appropriate refinement of the internally-mixed text, i.e., one must be able to break the mixed text into appropriately independent sub-

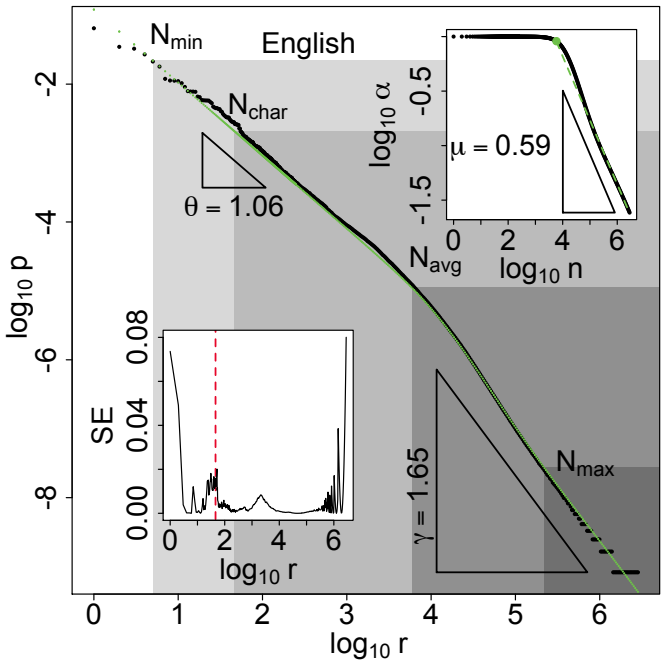


Figure 3.4: Results for the English corpus from the eBooks collection. The main axes show the empirical, normalized rank-frequency distribution (black), p, and the text mixing model (green points), $\hat{p}$. The measured lower and upper exponents, γ and θ, are depicted in the lower-right and upper-left respectively, with triangles indicating the measured slopes. We also present gray boxes in the main axes to highlight the different mixing regimes, marked by N_{char}, N_{min}, N_{avg}, and N_{max} (see Sec. 4.5 and Tab. 3.1 for complete descriptions). The lower left inset shows the squared errors $(p(r)-\hat{p}(r))^2$, whose sum is minimized in the production of $\hat{p}$ from the word introduction rate, α, depicted with black points in the upper right inset with the decay exponent μ (green dashed line's slope).

texts. From our example in Fig. 3.6, we can see that the division of the text into a corpus of 28 series' (left panel) renders a text mixing model for the empirical data with much higher error than a division into a corpus of 143 volumes (center panel, a refinement of the series' division). We also present text mixing results from the extremal refinement, where each individual word is treated as a text (right panel, see Sec. 4.5 for more information on the extremal refinement), which shows that a text can be over-refined to produce a poor text mixing model.

It is worth noting from our results that the parameter, θ, is frequently measured to lie outside the Simon-productive range, $(0, 1)$. Therefore, we are left to conclude that individually, many texts are subject to internally-derived decay in word introduction rates (as is exemplified by the Ebers text in Fig. 3.6), i.e., the underlying rank-frequency distributions are not of pure Zipf/Simon form (as we suggest in other work (2014)), but, instead, subject to internal mixing. Though we do not exhaustively investigate the occurrence of internally-derived decay in the rates of word introduction across the eBooks data set, it seems quite

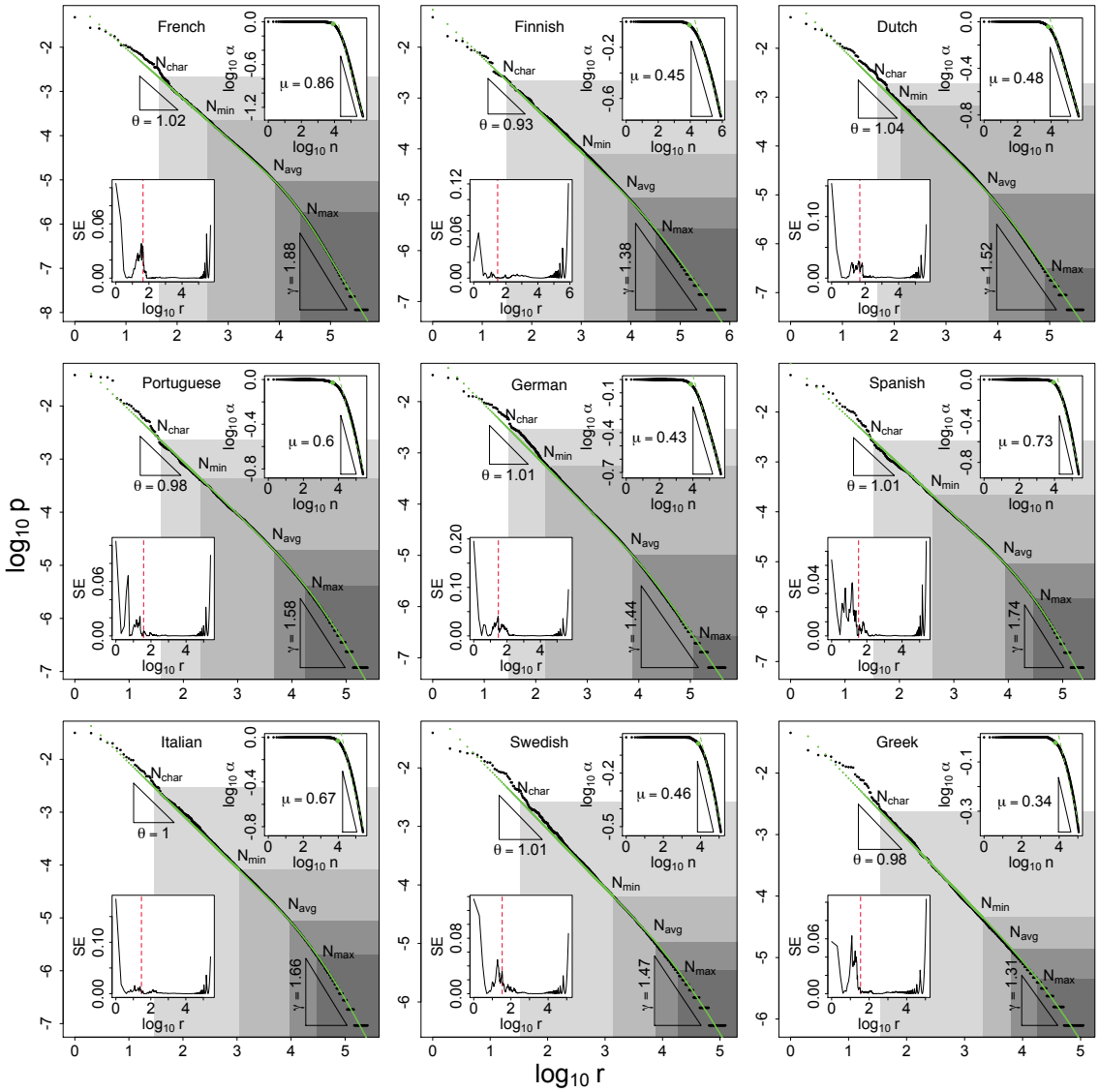


Figure 3.5: The results of text mixing experiments for the nine smaller corpora analyzed. All insets, color-coding, and labels are consistent with those from the larger, English presentation in Fig. 3.4, whose caption possesses full descriptions of all axes and plotted data.

possible that all of the texts parsed are subject to some internal mixing effects, whether from non-original annotation by the Project Gutenberg e-Text editors, or just the mixing of differing components (e.g., chapters, series', volumes, prologues, etc...). This of course

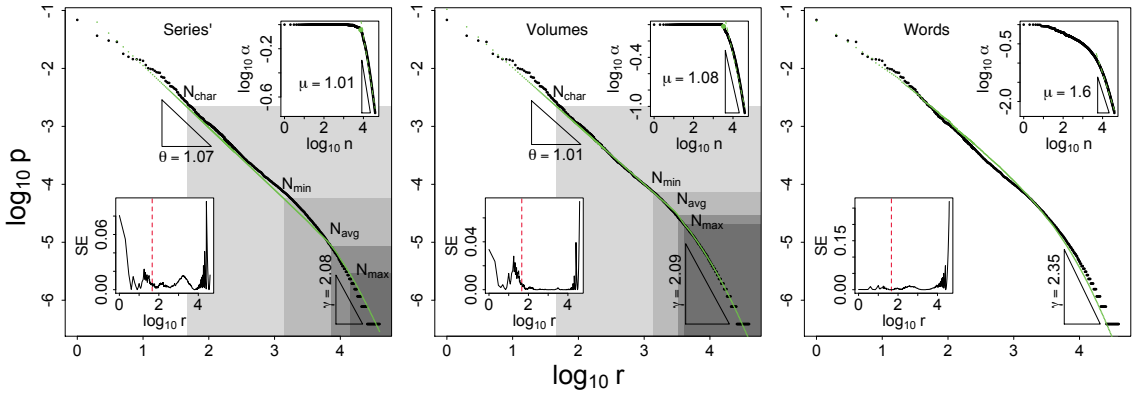


Figure 3.6: Text mixing results for a single-author corpus. Here, α was measured for differing refinements of the Egyptological fiction compendium/text "The complete historical romances of Georg Ebers" into sub-texts. All insets, color-coding, and labels are consistent with those from the English presentation in Fig. 3.4, whose caption possesses full descriptions of all axes and plotted data. **(Left)** Each series is considered a separate text. **(Middle)** Each volume of each series is considered a separate text. **(Right)** Each word (the extremal refinement, see Materials and Methods) in the compendium is considered a separate text. Note that in the upper right insets, α decreases overall with each refinement (as by definition it must), and that there appears to be an optimal refinement for producing a text mixing model, likely close to the scale of volumes.

would require that these mixing effects be of low-impact in the cases generally considered strong examples of Zipf's law.

We also note a strange behavior (which is captured by the text mixing model) in the English data set. There, we have found a relatively shallow lower scaling ($\gamma \approx 1.65$), but notice that it appears to be one of possibly two lower scalings. For English, the crossover regime exhibits a consistently steeper scaling that dies away in the terminal regime. Though we have no certain explanation for this behavior, part of what makes the English collection so different from the others is the sheer number of texts (see Tab. 3.1). However, upon looking closer at the distribution of English text sizes, we also notice that the collection possess some extremely large-N outliers. In the largest text (which has nearly an order of magnitude more words than any other text), approximately one tenth of all words are represented (out of nearly $20,000$ books), which must have a profound impact on the combined rank-frequency distribution, and hence lower scaling. Further, this large-N hypothesis is supported by

our preliminary investigation (see Fig. 3.2) where we observed that those (large) texts in the tenth decile not only generated scaling break points that went against the $b = N_{\text{avg}}$ correspondence, but also, generated relatively shallow lower scalings, against the trend of steepening with increasing decile. English is also well-known for its willingness to adopt foreign words, which may lead to an increased rate of appearance of low-count loan words. Regardless of the reasons for this difference with English, we find that text mixing captures the shape of both lower scaling regimes, and so both are well explained by the text mixing model.

We also take time to make note of and discuss another anomalous behavior of the rank-frequency distributions investigated. Upon viewing a rank-frequency distribution for Zipf's law, one generally finds a "wobble" of the frequency data around Zipf's scaling (regardless of the existence of a scaling break). We refer to the termination of this "wobble" as the point of stabilization of the Zipf/Simon regime. Looking at the empirical data from the ten languages, we see that this stabilization point generally appears early on the in Zipf/Simon regime, and generally not before the first 10^2 ranks. Though we have no definitive explanation for the existence of this anomaly, we note upon looking at the pointwise-squared errors that the stabilization point frequently occurs near each language's number of characters, N_{char} (depicted as a red dotted vertical line in each of the lower left insets of Figs. 3.4, 3.5, and the center panel of 3.6). Whether the numbers of characters spawned in the generation of primordial, character-based languages still influence the shapes of rank-frequency distributions of descendant languages today, we cannot say for sure. However this anomalous regime appears consistently across languages, and may potentially be of consistent shape across the corpora of a language. If so, we might view such anomalies as universal properties of languages, and so highlight them in the hopes of opening a broader discussion.

In light of the results presented, we take time to consider the validity of the core language hypothesis. We have seen significant variation in both the location and severity of scaling breaks both across and within languages. Upon sampling the English corpus by deciles, we have observed that the regressed point of scaling break, b, is not stationary (see Fig. 3.2). We take this as indication of the lack of validity of and language-universal core/non-core hypothesis, as a core should exhibit a strong consistency of size. Moreover, languages closely related via a common, recent ancestor should likewise exhibit this consistency, but notably two of the languages most closely related in the study, Spanish and Portuguese, present a large difference in b, ($10,363$ for Spanish, and $15,079$ for Portuguese—see Tab. 3.1). Both of these results seem to indicate that scaling breaks in rank-frequency distributions are likely consequences of text and corpus composition. Hence, it may then be more reasonable to consider a language core as a collection of words necessary for basic description, but not overlapping in use or meaning. However, such a core lexicon would need to be determined by native practitioners, and not necessarily to be an observable property of rank-frequency distributions. Alternatively, one could consider a corpus-core by its collection of words common to its texts. However, such a "common core" would be entirely dependent on the composition of the corpus, and hence not a universal property of a language proper.

3.6 References

Barabási, A. L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–511.

Ferrer-i-Cancho, R., Solé, R. V., 2001. Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. Journal of Quantitative Linguistics 8, 165–173.

Gerlach, M., Altmann, E. G., 2013. Stochastic model for the vocabulary growth in natural languages. Phys. Rev. X 3, 021006.

Krapivsky, P. L., Redner, S., 2001. Organization of growing random networks. Phys. Rev. E 63, 066123.

Kwapien, J., Drozdz, S., Orczyk, A., 2010. Linguistic complexity: English vs. polish, text vs. corpus. Acta Physica Polonica, A. 117, 716.

Montemurro, M. A., 2001. Beyond the Zipf-Mandelbrot law in quantitative linguistics. Physica A: Statistical Mechanics and Its Applications 300, 567–578.

Petersen, A. M., Tenenbaum, J., Havlin, S., Stanley, H. E., Perc, M., 2012. Languages cool as they expand: allometric scaling and the decreasing need for new words. Scientific Reports 2.

Project Gutenberg, 2010. http://www.gutenberg.org.

Simon, H. A., 1955. On a class of skew distribution functions. Biometrika 42, 425–440.

Wikipedia Greek Alphabet, 2014. https://en.wikipedia.org/wiki/Greek_alphabet.

Wikipedia Latin Alphabets, 2014. https://en.wikipedia.org/wiki/Latin_alphabets.

Williams, J. R., Lessard, P. R., Desu, S., Clark, E. M., Bagrow, J. P., Danforth, C. M., Dodds, P. S., 2014. Zipf's law holds for phrases, not words. CoRR abs/1406.5181, http://arxiv.org/abs/1406.5181.

Yule, G. U., 1924. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Phil. Trans. B 213, 21.

Zipf, G. K., 1935. The Psycho-Biology of Language. Houghton-Mifflin.

Zipf, G. K., 1949. Human Behaviour and the Principle of Least-Effort. Addison-Wesley.

Chapter 4

Identifying missing dictionary entries with frequency-conserving context models

In an effort to better understand meaning from natural language texts, we explore methods aimed at organizing lexical objects into contexts. A number of these methods for organization fall into a family defined by word ordering. Unlike demographic or spatial partitions of data, these collocation models are of special importance for their universal applicability in the presence of ordered symbolic data (e.g., text, speech, genes, etc...). Our approach focuses on the phrase (whether word or larger) as the primary meaning-bearing lexical unit and object of study. To do so, we employ our previously developed framework for generating word-conserving phrase-frequency data. Upon training our model with the Wiktionary—an extensive, online, collaborative, and open-source dictionary that contains over $100,000$ phrasal-definitions—we develop highly effective filters for the identification of meaningful, missing phrase-entries. With our predictions we then engage the editorial community of the Wiktionary and propose short lists of potential missing entries for definition, developing a breakthrough, lexical extraction technique, and expanding our knowledge of the defined English lexicon of phrases.

4.1 Background

Starting with the work of Shannon (1948), joint probability distributions between word-types (denoted $w \in W$), and their groupings by appearance-orderings, or, *contexts* (denoted $c \in C$), were first used for the prediction of upcoming symbols. For a word appearing in text, Shannon's model assigned context according to the word's immediate antecedent. In other words, the sequence

$$\cdots\ w_{i-1}\ w_i\ \cdots$$

places this occurrence of the word-type of w_i in the context of $w_{i-1}\ \star$ (uniquely defined by the word-type of w_{i-1}), where "$\star$" denotes "any word". This experiment was novel, and when these transition probabilities were observed, he found a method for the automated production of language that far better resembled true English text than simple adherence to relative word frequencies.

Later, though still early on in the history of modern computational linguistics and natural language processing, theory caught up with Shannon's work. In 1975, Becker wrote:

> My guess is that phrase-adaption and generative gap-filling are very roughly equally important in language production, as measured in processing time spent on each, or in constituents arising from each. One way of making such an intuitive estimate is simply to listen to what people actually say when they speak. An independent way of gauging the importance of the phrasal lexicon is to determine its size.

Since then, with the rise of computation and increasing availability of electronic text, there have been numerous extensions of Shannon's context model. These models have generally been information-theoretic applications as well, mainly used to predict word associations (Church and Hanks, 1990) and to extract multi-word expressions (MWEs) (Smadja,

1993). This latter topic has been one of extreme importance for the computational linguistics community (Ramisch, 2014) and has seen many approaches aside from the information-theoretic, including use of part-of-speech taggers (Justeson and Katz, 1995) and use of syntactic parsers (Seretan, 2008). However, almost all of these methods have the common issue of scalability (Pecina, 2010), making them difficult to use for the extraction of phrases of more than two words.

Information-theoretic extensions of Shannon's context model have also been used by Piantadosi et al. (2011b) to extend the work of Zipf (1935), using an entropic derivation called the Information Content (IC):

$$I(w) = -\sum_{c \in C} P(c \mid w) \log P(w \mid c) \tag{4.1}$$

and measuring its associations to word lengths. Though there have been concerns over some of the conclusions reached in this work (Reilly and Kean, 2011; Piantadosi et al., 2011a; Ferrer-i-Cancho and P., 2012; Piantadosi et al., 2013), Shannon's model was somewhat generalized, and applied to 3-gram, 4-gram and 5-gram context models to predict word lengths. This model was also used by Garcia et al. (2012) to assess the relationship between sentiment (valence) norms and IC measurements of words. However their application of the formula

$$I(w) = -\frac{1}{f(w)} \sum_{i=1}^{f(w)} \log P(w \mid c_i), \tag{4.2}$$

to N-grams data was wholly incorrect, as this special representation applies only to corpus-level data, i.e., uncompressed, human readable text, and *not* the frequency-based N-grams.

In addition to the above considerations, there is also the issue of word frequency conservation, which is exacerbated by the Piantadosi et al. extension of Shannon's model. To be precise, for a joint distribution of words and contexts that is *physically* related to the

appearance of words on "the page", there should be conservation in the marginal frequencies:

$$f(w) = \sum_{c \in C} f(w, c), \tag{4.3}$$

much like that discussed by Church and Hanks (1990). This property is not upheld using any true, sliding-window N-gram data (e.g., Google 2006; Michel et al. 2011; Lin et al. 2012). To see this, we recall that for both of Garcia et al. (2012) and Piantadosi et al. (2011b), a word's N-gram context was defined by its immediate $N-1$ antecedents. However, by this formulation we note that the first word of a page appears as *last* in no 2-gram, the second appears as *last* in no 3-gram, and so on.

These word frequency misrepresentations may seem to be of little importance at the text or page level, but since the methods for large-scale N-gram parsing have adopted the practice of stopping at sentence and clause boundaries (Lin et al., 2012), word frequency misrepresentations (like those discussed above) have become very significant. In the new format, 40% of the words in a sentence or clause of length five have no 3-gram context (the first two). As such, when these context models are applied to modern N-gram data, they are incapable of accurately representing the frequencies of words expressed. We also note that despite the advances in processing made in the construction of the current Google N-grams corpus (Lin et al., 2012), other issues have been found, namely regarding the source texts taken (Pechenick et al., 2015).

We also note that there exist many other methods for grouping occurrences of lexical units to produce informative context models. As early as 1992, Resnik showed class categorizations of words (e.g., verbs and nouns) could be used to produce informative joint probability distributions. In recent work, Montemurro and Zanette (2010) used joint distributions of words and arbitrary equal-length parts of texts to entropically quantify the semantic information encoded in written language. Texts tagged with metadata like genera (Dodds and Danforth, 2009), time (Dodds et al., 2011), location (Mitchell et al., 2013),

and language (Dodds et al., 2015), have rendered straightforward and clear examples of the power in a (word-frequency conserving) joint probability mass function, at shedding light on social phenomena by relating words to classes. Though metadata approaches to context are informative, with their power there is simultaneously a loss of applicability (metadata is frequently not present), as well as a loss of bio-communicative relevance (humans are capable of inferring social information from text in isolation).

4.2 Frequency-conserving context models

In previous work (2014) we developed a scalable and general framework for generating frequency data for N-grams, called random text partitioning. Since a phrase-frequency distribution, S, is balanced with regard to its underlying word-frequency distribution, W,

$$\sum_{w \in W} f(w) = \sum_{s \in S} \ell(s) f(s) \tag{4.4}$$

phrase	$\ell(s_{i \cdots j}) = 1$	$\ell(s_{i \cdots j}) = 2$	$\ell(s_{i \cdots j}) = 3$	$\ell(s_{i \cdots j}) = 4$	$\cdots$
w_1	$\star$	-	-	-	$\cdots$
$w_1\ w_2$	$\star\, w_2$ $w_1\ \star$	$\star\,\star$	-	-	$\cdots$ $\cdots$
$w_1\ w_2\ w_3$	$\star\, w_2\ w_3$ $w_1\ \star\ w_3$ $w_1\ w_2\ \star$	$\star\ \star\ w_3$ $w_1\ \star\ \star$	$\star\ \star\ \star$	-	$\cdots$ $\cdots$ $\cdots$
$w_1\ w_2\ w_3\ w_4$	$\star\, w_2\ w_3\ w_4$ $w_1\ \star\ w_3\ w_4$ $w_1\ w_2\ \star\ w_4$ $w_1\ w_2\ w_3\ \star$	$\star\ \star\ w_3\ w_4$ $w_1\ \star\ \star\, w_4$ $w_1\ w_2\ \star\ \star$	$\star\ \star\ \star\, w_4$ $w_1\ \star\ \star\star$	$\star\ \star\ \star\star$	$\cdots$ $\cdots$ $\cdots$ $\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Table 4.1: A table showing the expansion of context lists for longer and longer phrases. We define the internal contexts of phrases by the removal of individual sub-phrases. These contexts are represented as phrases with words replaced by $\star$'s. Any phrases whose word-types match after analogous sub-phrase removals share the matching context. Here, the columns are labeled 1–4 by sub-phrase length.

(where ℓ denotes the phrase-length norm) it is easy to produce a symmetric generalization of Shannon's model that integrates all phrase/N-gram lengths and all word placement/removal points. To do so, we define W and S to be the sets of words and (text-partitioned) phrases from a text respectively, and let C be the collection of all single word-removal patterns from the phrases of S. A joint frequency, $f(w, c)$, is then defined by the partition frequency of the phrase that is formed when c and w are composed. In particular, if w composed with c renders s, we then set $f(w, c) = f(s)$, which produces a context model on the words whose marginal frequencies preserve their original frequencies from "the page." In particular we refer to this, or such a model for phrases, as an 'external context model,' since the relations are produced by structure external to the semantic unit.

It is good to see the external word-context generalization emerge, but our interest actually lies in the development of a context model for the phrases themselves. To do so, we define the 'internal contexts' of a phrase by the patterns generated through the removal of sub-phrases. To be precise, for a phrase s, and a sub-phrase $s_{i\cdots j}$ ranging over words i through j, we define the context

$$c_{i\cdots j} = w_1 \cdots w_{i-1} \star \cdots \star w_{j+1} \cdots w_{\ell(s)} \tag{4.5}$$

to be the collection of same-length phrases whose analogous word removal (i through j) renders the same pattern (when word-types are considered). We present the contexts of generalized phrases of lengths 1–4 in Tab. 4.1, as described above. Looking at the table, it becomes clear that these contexts are actually a mathematical formalization of the generative gap filling proposed by Becker (1975), which was semi-formalized by the phrasal templates discussed at length by Smadja (1993). Between our formulation and that of Smadja, the main difference of definition lies in our restriction to contiguous word sequence (i.e., sub-phrase) removals, as is necessitated by the mechanics of the secondary partition process, which defines the context lists.

The weighting of the contexts for a phrase is accomplished simultaneously with their definition through a secondary partition process describing the inner-contextual modes of interpretation for the phrase. The process is as follows. In an effort to relate an observed phrase to other known phrases, the observer selectively ignores a sub-phrase of the original phrase. To retain generality, we do this by considering the *random* partitions of the original phrase, and then assume that a sub-phrase is ignored from a partition with probability proportional to its length, to preserve word (and hence phrase) frequencies. The conditional probabilities of inner context are then:

$$\begin{aligned} P(c_{i\cdots j} \mid s) &= P(\text{ignore } s_{i\cdots j} \text{ given a partition of } s) \\ &= P(\text{ignore } s_{i\cdots j} \text{ given } s_{i\cdots j} \text{ is partitioned from } s) P(s_{i\cdots j} \text{ is partitioned from } s). \end{aligned} \tag{4.6}$$

Utilizing the partition probability and our assumption, we note from our work in 2014 that

$$\ell(s) = \sum_{1 \le i < j \le \ell(s)} \ell(s_{i\cdots j}) P_q(s_{i\cdots j} \mid s), \tag{4.7}$$

which ensures through defining

$$P(c_{i\cdots j} \mid s) = \frac{\ell(s_{i\cdots j})}{\ell(s)} P_q(s_{i\cdots j} \mid s), \tag{4.8}$$

the production of a valid, phrase-frequency preserving context model:

$$\begin{aligned} \sum_{c \in C} f(c, s) &= \sum_{i < j \le \ell(s)} P(c_{i\cdots j} \mid s) f(s) \\ &= f(s) \sum_{1 \le i < j \le \ell(s)} \frac{\ell(s_{i\cdots j})}{\ell(s)} P_q(s_{i\cdots j} \mid s) = f(s), \end{aligned} \tag{4.9}$$

which preserves the underlying frequency distribution of phrases. Note here that beyond this point in the document we will used the normalized form,

$$P(c,s) = \frac{f(c,s)}{\sum\limits_{s\in S}\sum\limits_{c\in C} f(c,s)}, \tag{4.10}$$

for convenience in the derivation of expectations in the next section.

4.3 Likelihood of dictionary definition

In this section we exhibit the power of the internal context model through a lexicographic application, deriving a measure of meaning and definition for phrases with empirical phrase-definition data taken from a collaborative open-access dictionary (Wiktionary, 2014) (see Sec. 4.5 for more information on our data and the Wiktionary). With the rankings that this measure derives, we will go on to propose phrases for definition with the editorial community of the Wiktionary in an ongoing live experiment, discussed in Sec. 4.4.

To begin, we define the dictionary indicator, D, to be a binary norm on phrases, taking value 1 when a phrase appears in the dictionary, (i.e., has definition) and taking value 0 when a phrase is unreferenced. The dictionary indicator tells us when a phrase has reference in the dictionary, and in principle can be replaced with other indicator norms, for other purposes. Moving forward, we note an intuitive description of the distribution average:

$$\overline{D}(S) = \sum_{t\in S} D(t)P(t) = P(\text{randomly drawing a defined phrase from S}),$$

and go on to derive an alternative expansion through application of the context model:

$$\begin{aligned}
\overline{D}(S) = \sum_{t \in S} D(t)P(t) &= \sum_{t \in S} D(t)P(t) \sum_{c \in C} P(c \mid t) \sum_{s \in S} P(s \mid c) \\
&= \sum_{c \in C} P(c) \sum_{t \in S} D(t)P(t \mid c) \sum_{s \in S} P(s \mid c) \\
&= \sum_{c \in C} P(c) \sum_{s \in S} P(s \mid c) \sum_{t \in S} D(t)P(t \mid c) \\
&= \sum_{s \in S} P(s) \sum_{c \in C} P(c \mid s) \sum_{t \in S} D(t)P(t \mid c) \\
&= \sum_{s \in S} P(s) \sum_{c \in C} P(c \mid s)\overline{D}(c \mid S).
\end{aligned} \tag{4.11}$$

In the last line we then interpret:

$$\overline{D}(C \mid s) = \sum_{c \in C} P(c \mid s)\overline{D}(c \mid S), \tag{4.12}$$

to be the likelihood (analogous to the IC equation presented here as equation 4.1) that a phrase, which is randomly drawn from a context of s, to have definition in the dictionary. To be precise, we say $\overline{D}(C \mid s)$ is the likelihood of dictionary definition of the context model C, given the phrase s. When only one $c \in C$ is considered, we say $\overline{D}(c \mid S) = \sum_{t \in S} D(t)P(t \mid c)$ is the likelihood of dictionary definition of the context c, given S. Numerically, we note that the distribution-level values, $\overline{D}(C \mid s)$, "extend" the dictionary over all S, smoothing out the binary data to the full lexicon (uniquely for phrases of more than one word, which have no interesting space-defined internal structure) through the relations of the model. In other words, though $\overline{D}(C \mid s) \neq 0$ may now only indicate the *possibility* of a phrase having definition, it is still a strong indicator, and most importantly, may be applied to never-before-seen expressions.

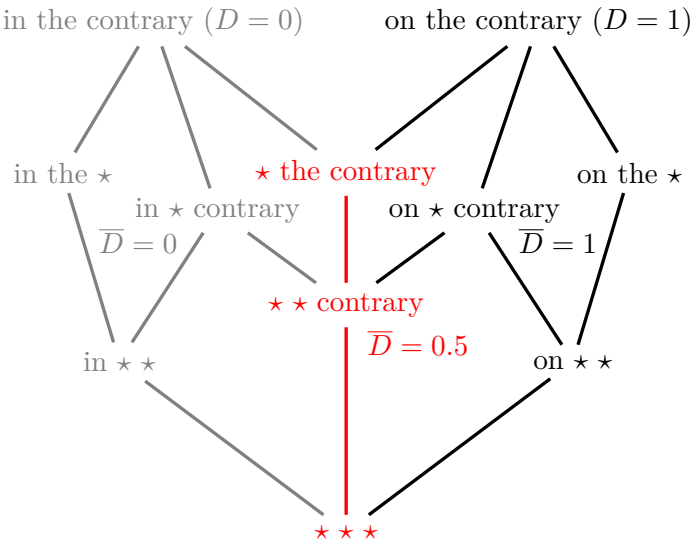


Figure 4.1: An example showing the sharing of contexts by similar phrases. Suppose our text consists of the two phrases, "in the contrary" and "on the contrary", and that each occurs once, and that the latter has definition ($D = 1$) while the former does not. In this event, we see that the three shared contexts: "⋆ ⋆ ⋆", "⋆ ⋆ contrary", and "⋆ the contrary", present elevated likelihood ($\overline{D}$) values, indicating that the phrase "in the contrary" may have meaning and be worthy of definition.

4.4 Predicting missing dictionary entries

Starting with the work of Sinclair et al. (1987) (though the idea was proposed more than 10 years earlier by Becker (1975)), lexicographers have been building dictionaries based on language as it is spoken and written, including idiomatic, slang-filled, and grammatical expressions (e.g., Collins English Cobuild Dictionary; Wiktionary; The Urban Dictionary; The Online Slang Dictionary). These dictionaries have proven highly-effective for non-primary language learners, who may not be privy to cultural metaphors. In this spirit, we utilize the context model derived above to discover phrases that are undefined, but which may be in need of definition for their similarity to other, defined phrases. We do this in a corpus-based way, using the definition likelihood $\overline{D}(C \mid s)$ as a secondary filter to frequency. The process is in general quite straightforward, and first requires a ranking of phrases by frequency of occurrence, $f(s)$. Upon taking the first $s_1, ..., s_N$ frequency-ranked phrases ($N = 100,000$, for our experiments), we reorder the list according to the values $\overline{D}(C \mid s)$ (descending). The top of such a double-sorted list then includes phrases that are both frequent and similar to defined phrases.

With our double-sorted lists we then record those phrases having no definition or dictionary reference, but which are at the top. These phrases are quite often meaningful (as we have found experimentally, see below) despite their lack of definition, and as such we propose this method for the automated generation of short lists for editorial investigation of definition.

4.5 Materials and methods

For its breadth, open-source nature, and large editorial community, we utilize dictionary data from the Wiktionary (2014) (a Wiki-based open content dictionary) to build the dictionary-indicator norm, setting $D(s) = 1$ if a phrase s has reference or redirect. We also note that the minimum information necessary for a phrase to be included in the Wiktionary, is a language, part of speech, and meaning.

We apply our filter for missing entry detection to several large corpora from a wide scope of content. These corpora are: twenty years (1987–2007) of New York Times (NYT) articles (Sandhaus, 2008), approximately 4% of a year's (2009) tweets from twitter, music lyrics from thousands of songs and authors (Lyrics, 1960–2007) (Dodds and Danforth, 2009), complete Wikipedia articles (2010), and a Project Gutenberg eBooks collection (eBooks) (2010) of more than $30,000$ public-domain texts. We note that these are all unsorted texts, and that Twitter, eBooks, Lyrics, and to an extent, Wikipedia are mixtures of many languages (though majority English). We only attempt missing entry prediction for phrase lengths (2–5), for their inclusion in other major collocation corpora (Lin et al., 2012), as well as their having the most data in the dictionary. We also note that all text processed is taken lower-case.

To understand our results, we perform a 10-fold cross-validation on the frequency and likelihood filters. This is executed by random splitting the Wiktionary's list of defined phrases into 10 equal-length pieces, and then performing 10 parallel experiments. In each

of these experiments we determine the likelihood values, $\overline{D}(C \mid s)$, by a distinct $\frac{9}{10}$'s of the data. We then order the union set of the $\frac{1}{10}$-withheld and the Wiktionary-undefined phrases by their likelihood (and frequency) values descending, and accept some top segment of the list, or, 'short list', coding them as positive by the experiment. For such a short list, we then record the true positive rates, i.e., portion of all $\frac{1}{10}$-withheld truly-defined phrases we coded positive, the false positive rates, i.e., portion of all truly-undefined phrases we coded positive, and the number of entries discovered. Upon performing these experiments, the average of the ten trials is taken for each of the three parameters, for a number of short list lengths (scanning $1,000$ log-spaced lengths), and plotted as a receiver operating characteristic (ROC) curve (see Figs. 4.2–B.4). We also note that each is also presented with its area under curve (AUC), which measures the accuracy of the expanding-list classifier as a whole.

4.6 Results and discussion

Before observing output from our model we take the time to perform a cross-validation (10-fold), and compare our context filter to a sort by frequency alone. From this we have found that our likelihood filter renders missing entries much more efficiently than by frequency (see Tab. 4.2, and Figs. 4.2–B.4), already discovering missing entries from short lists of as little as twenty (see the insets of Figs. 4.2–B.4 as well as Tabs. 4.2, 4.3, and B.1–B.4). As such we adhere to this standard, and only publish short lists of 20 predictions per corpus per phrase lengths 2–5. In parallel, we also present phrase frequency-generated short-lists for comparison.

In addition to listing them in the appendices, we have presented the results of our experiment from across the 5 large, disparate corpora on the Wiktionary in a pilot program,

where we are tracking the success of the filters [1]. Looking at the lexical tables, where defined phrases are highlighted in red, we can see that many of the predictions by the likelihood filter (especially those obtained from the Twitter corpus) have already been defined in the Wiktionary following our recommendation (as of February 19th 2015) since we accessed its data in September of 2014 Wiktionary (2014). We also summarize these results from the live experiment in Tab. 4.2.

Looking at the lexical tables more closely, we note that all corpora present highly idiomatic expressions under the likelihood filter, many of which are variants of existing id-

[1]Track the potential missing entries that we have proposed: https://en.wiktionary.org/wiki/User:Jakerylandwilliams/Potential_missing_entries

	Corpus	2-gram	3-gram	4-gram	5-gram
Cross-val	Twitter	4.22 (0.40)	1.11 (0.30)	0.90 (0.10)	1.49 (0)
	NYT	4.97 (0.30)	0.36 (0.50)	0.59 (0.10)	1.60 (0)
	Lyrics	3.52 (0.50)	1.76 (0.40)	0.78 (0)	0.48 (0)
	Wikipedia	5.06 (0.20)	0.46 (0.80)	1.94 (0.20)	1.54 (0)
	eBooks	3.64 (0.30)	1.86 (0.30)	0.59 (0.60)	0.90 (0.10)
	Corpus	**2-gram**	**3-gram**	**4-gram**	**5-gram**
Live exp.	Twitter	6(0)	4 (0)	5 (0)	5 (0)
	NYT	5 (0)	0 (0)	2 (0)	1 (0)
	Lyrics	3 (0)	1 (0)	3 (0)	1 (0)
	Wikipedia	0 (0)	1 (0)	1 (0)	2 (0)
	eBooks	2 (0)	1 (0)	3 (0)	6 (1)

Table 4.2: Summarizing our results from the cross-validation procedure **(Above)**, we present the mean numbers of missing entries discovered when 20 guesses were made for N-grams/phrases of lengths 2, 3, 4, and 5, each. For each of the 5 large corpora (see Materials and Methods) we make predictions according our likelihood filter, and according to frequency (in parentheses) as a baseline. When considering the 2-grams (for which the most definition information exists), short lists of 20 rendered up to 25% correct predictions on average by the definition likelihood, as opposed to the frequency ranking, by which no more than 2.5% could be expected. We also summarize the results to-date from the live experiment **(Below)** (updated February 19, 2015), and present the numbers of missing entries correctly discovered on the Wiktionary (i.e., reference added since July 1, 2014, when the dictionary's data was accessed) by the 20-phrase shortlists produced in our experiments for both the likelihood and frequency (in parentheses) filters. Here we see that all of the corpora analyzed were generative of phrases, with Twitter far and away being the most productive, and the reference corpus Wikipedia the least so.

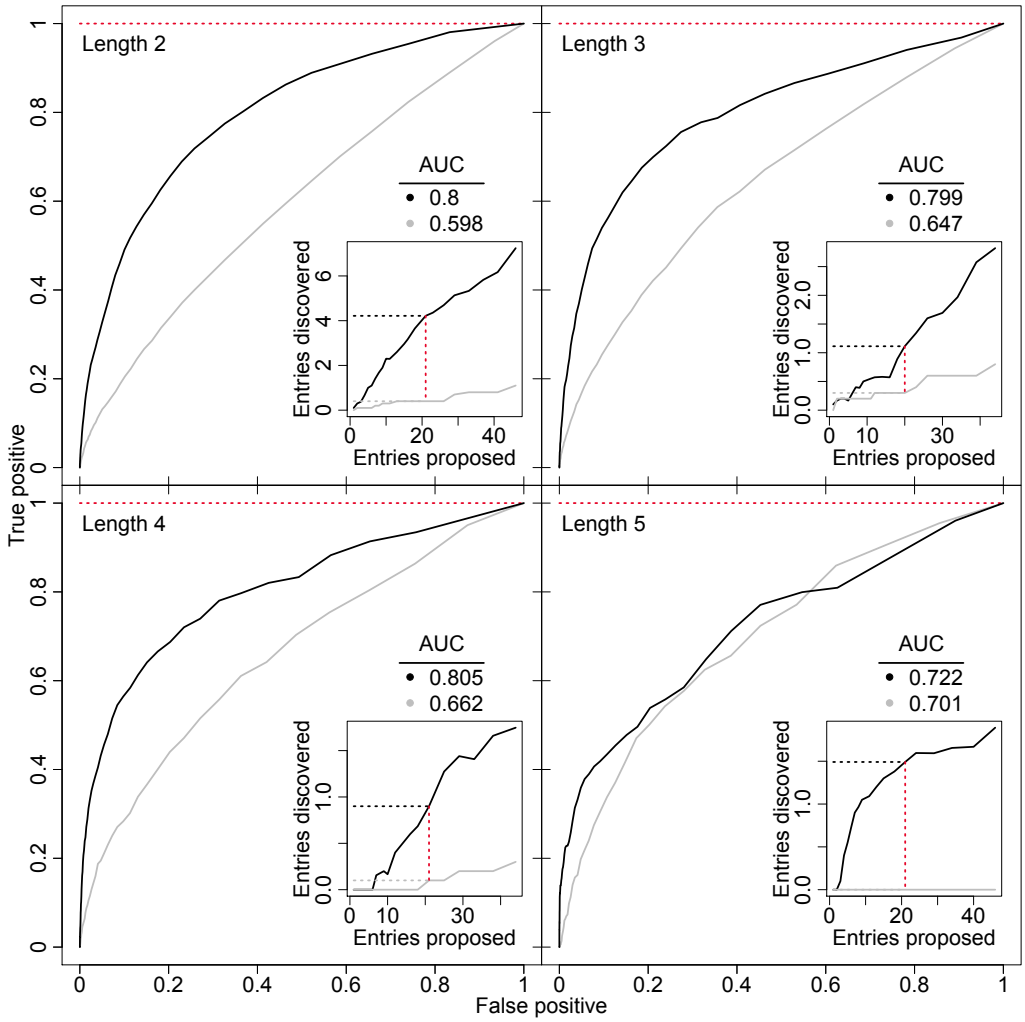


Figure 4.2: With data taken from the Twitter corpus, we present (10-fold) cross-validation results for the filtration procedures. For each of the lengths 2, 3, 4, and 5, we show the ROC curves (**Main Axes**), comparing true and false positive rates for both the likelihood filters (black), and for the frequency filters (gray). There, we see increased performance in the likelihood classifiers (except possibly for length 5), which is reflected in the AUCs (where an AUC of 1 indicates a perfect classifier). We also monitor the average number of missing entries discovered as a function of the number of entries proposed (**Insets**), for each length. There, the horizontal dotted lines indicate the average numbers of missing entries discovered for both the likelihood filters (black) and for the frequency filters (gray) when short lists of 20 phrases were taken (red dotted vertical lines). From this we see an indication that even the 5-gram likelihood filter is effective at detecting missing entries in short lists, while the frequency filter is not.

iomatic phrases that will likely be granted inclusion into the dictionary through redirects or alternative-forms listings. To name a few, the Twitter (Tab. 4.3), Times (Tab. B.1), and Lyrics (Tab. B.2) corpora consistently predict large families derived from phrases like "at

	rank	2-gram	3-gram	4-gram	5-gram
definition likelihood	1	buenos noches	knock it out	in the same time	actions speak louder then words
	2	north york	walk of fame	on the same boat	no sleep for the wicked
	3	last few	piece of mind	about the same time	every once and a while
	4	holy hell	seo-search engine optimization	around the same time	to the middle of nowhere
	5	good am	puta q pariu	at da same time	come to think about it
	6	going away	who the heck	wat are you doing	dont let the bedbugs bite
	7	right up	take it out	wtf are you doing	you get what i mean
	8	go sox	fim de mundo	why are you doing	you see what i mean
	9	going well	note to all	hell are you doing	you know who i mean
	10	due out	in the moment	better late then never	no rest for the weary
	11	last bit	note to myself	here i go again	as long as i know
	12	go far	check it here	every now and again	as soon as i know
	13	right out	check it at	what were you doing	going out on a limb
	14	fuck am	check it http	was it just me	give a person a fish
	15	holy god	check it now	here we are again	at a lost for words
	16	rainy morning	check it outhttp	keeping an eye out	de una vez por todas
	17	picked out	why the heck	what in the butt	onew kids on the block
	18	south coast	memo to self	de vez em qdo	twice in a blue moon
	19	every few	reminder to self	giving it a try	just what the dr ordered
	20	picking out	how the heck	pain in my ass	as far as we know
	rank	**2-gram**	**3-gram**	**4-gram**	**5-gram**
frequency	1	in the	new blog post	i just took the	i favorited a youtube video
	2	i just	i just took	e meu resultado foi	i uploaded a youtube video
	3	of the	live on http	other people at http	just joined a video chat
	4	on the	i want to	check this video out	fiddling with my blog post
	5	i love	i need to	just joined a video	joined a video chat with
	6	i have	i have a	a day using http	i rated a youtube video
	7	i think	quiz and got	on my way to	i just voted for http
	8	to be	thanks for the	favorited a youtube video	this site just gave me
	9	i was	what about you	i favorited a youtube	add a #twibbon to your
	10	if you	i think i	free online adult dating	the best way to get
	11	at the	i have to	a video chat with	just changed my twitter background
	12	have a	looking forward to	uploaded a youtube video	a video chat at http
	13	to get	acabo de completar	i uploaded a youtube	photos on facebook in the
	14	this is	i love it	video chat at http	check it out at http
	15	and i	a youtube video	what do you think	own video chat at http
	16	but i	to go to	i am going to	s channel on youtube http
	17	are you	of the day	if you want to	and won in #mobsterworld http
	18	it is	what'll you get	i wish i could	live stickam stream at http
	19	i need	my daily twittascope	just got back from	on facebook in the album
	20	it was	if you want	thanks for the rt	added myself to the http

Table 4.3: With data taken from the Twitter corpus, we present the top 20 unreferenced phrases considered for definition (in the live experiment) from each of the 2, 3, 4, and 5-gram likelihood filters **(Above)**, and frequency filters **(Below)**. From this corpus we note the juxtaposition of highly idiomatic expressions by the likelihood filter (like "holy hell"), with the domination of the frequency filters by semi-automated content. The phrase "holy hell" is an example of the model's success with this corpus, as it achieved definition (February 8$^{\text{th}}$, 2015) concurrently with the preparation of this manuscript (several months after the Wiktionary's data was accessed in July, 2014).

the same time", and "you know what i mean", while the eBooks and Wikipedia corpora predict families derived from phrases like "on the other hand", and "at the same time". In general we see no such structure or predictive power emerge from the frequency filter.

We also observe that from those corpora which are less pure of English context (namely, the eBooks, Lyrics, and Twitter corpora), extra-English expressions have crept in. This highlights an important feature of the likelihood filter—it does not intrinsically rely on the syntax or grammar of the language to which it is applied, beyond the extent to which syntax and grammar effect the shapes of collocations. For example, the eBooks predict (see Tab. B.4) the undefined French phrase "tu ne sais pas", or "you do not know", which is a syntactic variant of the English-Wiktionary defined French, "je ne sais pas", meaning "i do not know". Seeing this, we note that it would be straightforward to construct a likelihood filter with a language indicator norm to create an alternative framework for language identification.

There are also a fair number of phrases predicted by the likelihood filter which in fact are spelling errors, typos, and grammatical errors. In terms of the context model, these erroneous forms are quite near to those defined in the dictionary, and so rise in the short lists generated from the less-well edited corpora, e.g., "actions speak louder *then* words" in the Twitter corpus. This then seems to indicate the potential for the likelihood filter to be integrated into auto-correct algorithms, and further points to the possibility of constructing syntactic indicator norms of phrases, making estimations of tenses and parts of speech (whose data is also available from the Wiktionary) possible through application of the model in precisely the same manner presented in Sec. 4.3. Regardless of the future applications, we have developed and presented a novel, powerful, and scalable MWE extraction technique.

4.7 References

Becker, J. D., 1975. The phrasal lexicon. In: Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing. TINLAP '75. As-

sociation for Computational Linguistics, Stroudsburg, PA, USA, pp. 60–63, http://dx.doi.org/10.3115/980190.980212.

Church, K. W., Hanks, P., Mar. 1990. Word association norms, mutual information, and lexicography. Comput. Linguist. 16 (1), 22–29, http://dl.acm.org/citation.cfm?id=89086.89095.

Collins English Cobuild Dictionary, 2015. http://www.collinsdictionary.com/dictionary/english-cobuild-learners.

Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., Mitchell, L., Harris, K. D., Kloumann, I. M., Bagrow, J. P., Megerdoomian, K., McMahon, M. T., Tivnan, B. F., Danforth, C. M., 2015. Human language reveals a universal positivity bias. Proceedings of the National Academy of Scienceshttp://www.pnas.org/content/early/2015/02/04/1411678112.abstract.

Dodds, P. S., Danforth, C. M., 2009. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. Journal of Happiness StudiesDoi:10.1007/s10902-009-9150-9.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., Danforth, C. M., 12 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. PLoS ONE 6 (12), e26752, http://dx.doi.org/10.1371/journal.pone.0026752.

Ferrer-i-Cancho, R., P., M. F. M., 2012. Information content versus word length in random typing. CoRR abs/1209.1751, http://arxiv.org/abs/1209.1751.

Garcia, D., Garas, A., Schweitzer, F., 2012. Positive words carry less information than negative words. EPJ Data Science 1 (1), http://dx.doi.org/10.1140/epjds3.

Google, 2006. Official Google Research Blog: All Our N-gram are Belong to You. http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html.

Justeson, J., Katz, S., 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering, 9–27.

Lin, Y., Michel, J., Aiden, E. L., Orwant, J., Brockman, W., Petrov, S., 2012. Syntactic annotations for the google books ngram corpus. In: Proceedings of the ACL 2012 System Demonstrations. ACL '12. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 169–174, http://dl.acm.org/citation.cfm?id=2390470.2390499.

Michel, J., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., Aiden, E. L., 2011. Quantitative analysis of culture using millions of digitized books. Science 331 (6014), 176–182, http://www.sciencemag.org/content/331/6014/176.abstract.

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., Danforth, C. M., 05 2013. The geography of happiness: Connecting twitter sentiment and expression,

demographics, and objective characteristics of place. PLoS ONE 8 (5), e64417, http://dx.doi.org/10.1371/journal.pone.0064417.

Montemurro, M. A., Zanette, D. H., 2010. Towards the quantification of the semantic information encoded in written language. Advances in Complex Systems 13 (02), 135–153, http://www.worldscientific.com/doi/abs/10.1142/S0219525910002530.

Pechenick, E. A., Danforth, C. M., Dodds, P. S., 2015. Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. CoRR abs/1501.00960, http://arxiv.org/abs/1501.00960.

Pecina, P., 2010. Lexical association measures and collocation extraction. Language Resources and Evaluation 44 (1-2), 137–158, http://dx.doi.org/10.1007/s10579-009-9101-4.

Piantadosi, S., Tily, H., Gibson, E., 2011a. Reply to Reilly and Kean: Clarifications on word length and information content. Proceedings of the National Academy of Sciences 108 (20), E109, http://colala.bcs.rochester.edu/papers/PNAS-2011-Piantadosi-1103550108_reply.pdf.

Piantadosi, S. T., Tily, H., Gibson, E., 2011b. Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences 108 (9), 3526, http://colala.bcs.rochester.edu/papers/PNAS-2011-Piantadosi-1012551108.pdf.

Piantadosi, S. T., Tily, H., Gibson, E., Jul. 2013. Information content versus word length in natural language: A reply to Ferrer-i-Cancho and Moscoso del Prado Martin [arXiv:1209.1751]. ArXiv e-printshttp://adsabs.harvard.edu/abs/2013arXiv1307.6726P.

Project Gutenberg, 2010. http://www.gutenberg.org.

Ramisch, C., 2014. Multiword Expressions Acquisition: A Generic and Open Framework. Springer Publishing Company, Incorporated.

Reilly, J., Kean, J., 2011. Information content and word frequency in natural language: Word length matters. Proceedings of the National Academy of Sciences 108 (20), E108, http://www.pnas.org/content/108/20/E108.short.

Resnik, P., 1992. Wordnet and distributional analysis: A class-based approach to lexical discovery. AAAI Technical Report WS-92-01http://www.aaai.org/Papers/Workshops/1992/WS-92-01/WS92-01-006.pdf.

Sandhaus, E., 2008. The New York Times Annotated Corpus. Linguistic Data Consortium, Philadelphia.

Seretan, V., 2008. Collocation Extraction Based on Syntactic Parsing. http://books.google.com/books?id=nIrjSAAACAAJ.

Shannon, C. E., Jan. 1948. A mathematical theory of communication. SIGMOBILE Mob. Comput. Commun. Rev. 5 (1), 3–55, http://doi.acm.org/10.1145/584091.584093.

Sinclair, G., Hanks, P., Fox, G., Moon, R., et. al. Stock, P., 1987. The Collins Cobuild English Language Dictionary. Collins, Glasgow.

Smadja, F., Mar. 1993. Retrieving collocations from text: Xtract. Comput. Linguist. 19 (1), 143–177, http://dl.acm.org/citation.cfm?id=972450.972458.

The Online Slang Dictionary, 2015. http://onlineslangdictionary.com/.

The Urban Dictionary, 2015. http://www.urbandictionary.com/.

Twitter, 2009. Twitter API. http://dev.twitter.com/.

Wikipedia, 2010. http://dumps.wikimedia.org/enwiki/.

Wiktionary, 2014. http://dumps.wikimedia.org/enwiktionary/2014.

Wiktionary, 2015. https://www.wiktionary.org/.

Williams, J. R., Lessard, P. R., Desu, S., Clark, E. M., Bagrow, J. P., Danforth, C. M., Dodds, P. S., 2014. Zipf's law holds for phrases, not words. CoRR abs/1406.5181, http://arxiv.org/abs/1406.5181.

Zipf, G. K., 1935. The Psycho-Biology of Language. Houghton-Mifflin.

Chapter 5

Conclusion

Over the course of the work presented here we have accomplished several important tasks that will guide future research. In Ch. 2 our study resulted in the development of a general and scalable framework for producing frequency data for intermediate-sized lexical objects, which has already enabled us (in Ch. 4) to define a context model that conserves word frequencies, and use it effectively to detect missing entries from an online dictionary and extend our knowledge of the greater English lexicon of phrases. Beyond this, there is still much to be explored with random text partitioning—we have not even quantified the effects of temperature (q) on rank-frequency scalings. Additionally, the value apparent with random text partitioning leads us to consider how we might define other, informed methods for text partitioning.

In Ch. 3 we showed how large corpora are affected by their composition, and in doing so we clarified a discussion of 15 years regarding an empirical phenomenon of unknown origin. While our result has contended the core/non-core language hypothesis (Ferrer-i-Cancho and Solé, 2001), we connected the highly insightful analysis from some of its proponents (Gerlach and Altmann, 2013) to empirical data, confirming a mathematical connection between word dependences and rank-frequency scalings. Understanding this connection has large implications for future theory, as it directs us to look for and test other mechanisms that lead to the dependence of word appearance, like the subordinate selection process we have discussed in the abstract and approached lightly in Ch. 2. A clear next step then is to begin modeling subordinate selection as a stochastic process, and measure an empirical analog (much as we have done with text mixing) to determine

its relevance on language production. Beyond the implications of text mixing for future theory, applying its analysis (Eq. 3.8) to social data on twitter has already shown us that automatons have highly constrained vocabularies that are distinguishable, (a property we are leveraging in other work (Clark et al., 2015), separating automatons from human users on Twitter).

Finally, while the work in Ch. 4 was an application of our results from Ch. 2, the context model we have proposed and its application to the dictionary indicator norm have cleared a path toward applications that will be highly valuable in the natural language processing industry. In much the same manner as we have in Ch. 4, we can construct norms for tense, part of speech, and language, which could be applied to auto-correct and machine translation tasks. Furthermore, since the model is general with norms, we will be able to apply it in future work to non-binary norms such as valence (Bradley and Lang, 1999), with which we have already seen considerable success at detecting large events with social media (Dodds et al., 2011). We could then build a phrase-based 'story finder' with access to context-informed sentiment norms for an unlimited vocabulary of phrases, and create an early warning system for large-scale social events.

Bibliography

Axtell, R., 2001. Zipf distribution of U.S. firm sizes. Science 293 (5536), 1818–1820.

Barabási, A. L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–511.

Batty, M., 2008. The size, scale, and shape of cities. Science Magazine 319 (5864), 769–771.

Becker, J. D., 1975. The phrasal lexicon. In: Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing. TINLAP '75. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 60–63, http://dx.doi.org/10.3115/980190.980212.

Bornholdt, S., Ebel, H., 2001. World Wide Web scaling exponent from Simon's 1955 model. Phys. Rev. E 64, 035104(R).

Bradley, M. M., Lang, P. J., 1999. Affective norms for english words (anew): Stimuli, instruction manual and affective ratings. Technical report c-1, University of Florida, Gainesville, FL.

Church, K. W., Hanks, P., Mar. 1990. Word association norms, mutual information, and lexicography. Comput. Linguist. 16 (1), 22–29, http://dl.acm.org/citation.cfm?id=89086.89095.

Clark, E. M., Williams, J. R., Danforth, C. M., Dodds, P. S., Jones, C. A., 2015. Humans can't hide on the cyber linguistic frontier of the twittersphere.

Clauset, A., Shalizi, C. R., Newman, M. E. J., 2009. Power-law distributions in empirical data. SIAM Review 51, 661–703.

Collins English Cobuild Dictionary, 2015. http://www.collinsdictionary.com/dictionary/english-cobuild-learners.

Coromina-Murtra, B., Solé, R., 2010. Universality of Zifp's law. Phsyical Revew E 82, 011102.

Corominas-Murtra, B., Hanel, R., Thurner, S., 2014. Understanding zipf's law with playing dice: history-dependent stochastic processes with collapsing sample-space have power-law rank distributions. CoRR abs/1407.2775, http://arxiv.org/abs/1407.2775.

Cougar Town, 2013. I should have known it. Cougar Town, season 4, episode 4: http://www.imdb.com/title/tt2483134/.

de Solla Price, D. J., 1976. A general theory of bibliometric and other cumulative advantage processes. J. Amer. Soc. Inform. Sci. 27, 292–306.

Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., Mitchell, L., Harris, K. D., Kloumann, I. M., Bagrow, J. P., Megerdoomian, K., McMahon, M. T., Tivnan, B. F., Danforth, C. M., 2015. Human language reveals a universal positivity bias. Proceedings of the National Academy of Scienceshttp://www.pnas.org/content/early/2015/02/04/1411678112.abstract.

Dodds, P. S., Danforth, C. M., 2009. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. Journal of Happiness StudiesDoi:10.1007/s10902-009-9150-9.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., Danforth, C. M., 12 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. PLoS ONE 6 (12), e26752, http://dx.doi.org/10.1371/journal.pone.0026752.

D'Souza, R. M., Borgs, C., Chayes, J. T., Berger, N., Kleinberg, R. D., 2007. Emergence of tempered preferential attachment from optimization. Proc. Natl. Acad. Sci. 104, 6112–6117.

Ferrer-i-Cancho, R., Elvevåg, B., 03 2010. Random texts do not exhibit the real Zipf's law-like rank distribution. PLoS ONE 5, e9411.

Ferrer-i-Cancho, R., P., M. F. M., 2012. Information content versus word length in random typing. CoRR abs/1209.1751, http://arxiv.org/abs/1209.1751.

Ferrer-i-Cancho, R., Solé, R. V., 2001. Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. Journal of Quantitative Linguistics 8 (3), 165–173.

Ferrer-i-Cancho, R., Solé, R. V., 2001. Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. Journal of Quantitative Linguistics 8, 165–173.

Garcia, D., Garas, A., Schweitzer, F., 2012. Positive words carry less information than negative words. EPJ Data Science 1 (1), http://dx.doi.org/10.1140/epjds3.

Gerlach, M., Altmann, E. G., 2013. Stochastic model for the vocabulary growth in natural languages. Phys. Rev. X 3, 021006.

Goldenfeld, N., 1992. Lectures on Phase Transitions and the Renormalization Group. Vol. 85 of Frontiers in Physics. Addison-Wesley, Reading, Massachusetts.

Google, 2006. Official Google Research Blog: All Our N-gram are Belong to You. http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html.

Google, 2014. http://ngrams.googlelabs.com/.

Ha, L. Q., Hanna, P., Ming, J., Smith, F. J., 2009. Extending Zipf's law to n-grams for large corpora. Artif. Intell. Rev. 32, 101–113.

Ha, L. Q., Sicilia-Garcia, E. I., Ming, J., Smith, F. J., 2002. Extension of Zipf's law to words and phrases. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING). pp. 315–320.

Justeson, J., Katz, S., 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering, 9–27.

Justeson, J. S., Katz, S. M., March 1991. Co-occurrences of antonymous adjectives and their contexts. Comput. Linguist. 17 (1), 1–19, http://dl.acm.org/citation.cfm?id=971738.971739.

Kloumann, I. M., Danforth, C. M., Harris, K. D., Bliss, C. A., Dodds, P. S., 01 2012. Positivity of the english language. PLoS ONE 7 (1), e29484, http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0029484.

Krapivsky, P. L., Redner, S., 2001. Organization of growing random networks. Phys. Rev. E 63, 066123.

Kwapien, J., Drozdz, S., Orczyk, A., 2010. Linguistic complexity: English vs. polish, text vs. corpus. Acta Physica Polonica, A. 117, 716.

Lin, Y., Michel, J., Aiden, E. L., Orwant, J., Brockman, W., Petrov, S., 2012. Syntactic annotations for the google books ngram corpus. In: Proceedings of the ACL 2012 System Demonstrations. ACL '12. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 169–174, http://dl.acm.org/citation.cfm?id=2390470.2390499.

MacKay, D. J. C., 2002. Information Theory, Inference & Learning Algorithms. Cambridge University Press, New York, NY, USA.

Maillart, T., Sornette, D., Spaeth, S., von Krogh, G., 2008. Empirical tests of Zipf's law mechanism in open source Linux distribution. Phys. Rev. Lett. 101 (21), 218701.

Mandelbrot, B. B., 1953. An informational theory of the statistical structure of languages. In: Jackson, W. (Ed.), Communication Theory. Butterworth, Woburn, MA, pp. 486–502.

Mandelbrot, B. B., 1959. A note on a class of skew distribution function. Analysis and critique of a paper by H. A. Simon. Information and Control 2, 90–99.

Mandelbrot, B. B., 1961a. Final note on a class of skew distribution functions: analysis and critique of a model due to H. A. Simon. Information and Control 4, 198–216.

Mandelbrot, B. B., 1961b. Post scriptum to 'final note'. Information and Control 4, 300–304.

Michel, J., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., Aiden, E. L., 2011. Quantitative analysis of culture using millions of digitized books. Science 331 (6014), 176–182, http://www.sciencemag.org/content/331/6014/176.abstract.

Miller, G. A., 1957. Some effects of intermittent silence. American Journal of Psychology 70, 311–314.

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., Danforth, C. M., 05 2013. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. PLoS ONE 8 (5), e64417, http://dx.doi.org/10.1371/journal.pone.0064417.

Montemurro, M. A., 2001. Beyond the Zipf-Mandelbrot law in quantitative linguistics. Physica A: Statistical Mechanics and Its Applications 300, 567–578.

Montemurro, M. A., Zanette, D. H., 2010. Towards the quantification of the semantic information encoded in written language. Advances in Complex Systems 13 (02), 135–153, http://www.worldscientific.com/doi/abs/10.1142/S0219525910002530.

Pechenick, E. A., Danforth, C. M., Dodds, P. S., 2015. Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. CoRR abs/1501.00960, http://arxiv.org/abs/1501.00960.

Pecina, P., 2010. Lexical association measures and collocation extraction. Language Resources and Evaluation 44 (1-2), 137–158, http://dx.doi.org/10.1007/s10579-009-9101-4.

Petersen, A. M., Tenenbaum, J., Havlin, S., Stanley, H. E., Perc, M., 2012. Languages cool as they expand: allometric scaling and the decreasing need for new words. Scientific Reports 2.

Piantadosi, S., Tily, H., Gibson, E., 2011a. Reply to Reilly and Kean: Clarifications on word length and information content. Proceedings of the National Academy of Sciences 108 (20), E109, http://colala.bcs.rochester.edu/papers/PNAS-2011-Piantadosi-1103550108_reply.pdf.

Piantadosi, S. T., Tily, H., Gibson, E., 2011b. Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences 108 (9), 3526, http://colala.bcs.rochester.edu/papers/PNAS-2011-Piantadosi-1012551108.pdf.

Piantadosi, S. T., Tily, H., Gibson, E., Jul. 2013. Information content versus word length in natural language: A reply to Ferrer-i-Cancho and Moscoso del Prado Martin [arXiv:1209.1751]. ArXiv e-printshttp://adsabs.harvard.edu/abs/2013arXiv1307.6726P.

Project Gutenberg, 2010. http://www.gutenberg.org.

Ramisch, C., 2014. Multiword Expressions Acquisition: A Generic and Open Framework. Springer Publishing Company, Incorporated.

Rayner, J. M. V., 1985. Linear relations in biomechanics: the statistics of scaling functions. J. Zool. Lond. (A) 206, 415–439.

Reilly, J., Kean, J., 2011. Information content and word frequency in natural language: Word length matters. Proceedings of the National Academy of Sciences 108 (20), E108, http://www.pnas.org/content/108/20/E108.short.

Resnik, P., 1992. Wordnet and distributional analysis: A class-based approach to lexical discovery. AAAI Technical Report WS-92-01http://www.aaai.org/Papers/Workshops/1992/WS-92-01/WS92-01-006.pdf.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A., Flickinger, D., 2002. Multiword expressions: A pain in the neck for NLP. In: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing. CICLing '02. Springer-Verlag, London, UK, pp. 1–15.

Sandhaus, E., 2008. The New York Times Annotated Corpus. Linguistic Data Consortium, Philadelphia.

Seretan, V., 2008. Collocation Extraction Based on Syntactic Parsing. http://books.google.com/books?id=nIrjSAAACAAJ.

Shannon, C. E., Jan. 1948. A mathematical theory of communication. SIGMOBILE Mob. Comput. Commun. Rev. 5 (1), 3–55, http://doi.acm.org/10.1145/584091.584093.

Simon, H. A., 1955. On a class of skew distribution functions. Biometrika 42, 425–440.

Simon, H. A., 1960. Some further notes on a class of skew distribution functions. Information and Control 3, 80–88.

Simon, H. A., 1961a. Reply to Dr. Mandelbrot's post scriptum. Information and Control 4, 305–308.

Simon, H. A., 1961b. Reply to 'final note' by Benoît Mandelbrot. Information and Control 4, 217–223.

Sinclair, G., Hanks, P., Fox, G., Moon, R., et. al. Stock, P., 1987. The Collins Cobuild English Language Dictionary. Collins, Glasgow.

Smadja, F., Mar. 1993. Retrieving collocations from text: Xtract. Comput. Linguist. 19 (1), 143–177, http://dl.acm.org/citation.cfm?id=972450.972458.

Smith, F. J., Devine, K., 1985. Storing and retrieving word phrases. Information Processing & Management 21 (3), 215–224, http://www.sciencedirect.com/science/article/pii/0306457385901062.

The Online Slang Dictionary, 2015. http://onlineslangdictionary.com/.

The Urban Dictionary, 2015. http://www.urbandictionary.com/.

Twitter, 2009. Twitter API. http://dev.twitter.com/.

Wikipedia, 2010. http://dumps.wikimedia.org/enwiki/.

Wikipedia Greek Alphabet, 2014. https://en.wikipedia.org/wiki/Greek_alphabet.

Wikipedia Latin Alphabets, 2014. https://en.wikipedia.org/wiki/Latin_alphabets.

Wiktionary, 2014. http://dumps.wikimedia.org/enwiktionary/2014.

Wiktionary, 2015. https://www.wiktionary.org/.

Williams, J. R., Bagrow, J. P., Danforth, C. M., Dodds, P. S., 2014a. Text mixing shapes the anatomy of rank-frequency distributions: A modern zipfian mechanics for natural language. CoRRhttp://arxiv.org/abs/1409.3870.

Williams, J. R., Lessard, P. R., Desu, S., Clark, E. M., Bagrow, J. P., Danforth, C. M., Dodds, P. S., 2014b. Zipf's law holds for phrases, not words. CoRR abs/1406.5181, http://arxiv.org/abs/1406.5181.

Yule, G. U., 1924. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Phil. Trans. B 213, 21.

Zanette, D. H., Manrubia, S. C., 2001. Vertical transmission of culture and the distribution of family names. Physica A 295, 1–8.

Zipf, G. K., 1935. The Psycho-Biology of Language. Houghton-Mifflin.

Zipf, G. K., 1949. Human Behaviour and the Principle of Least-Effort. Addison-Wesley.

Appendix A

Random text partitions

A.1 Materials and methods

To obtain the results in Fig. 2.2, we utilize the maximum likelihood estimation (MLE) procedure developed by Clauset et al. (2009). In applying this procedure to clause and phrases distributions, several quantities are generally considered:

- $\hat{\theta}$: Zipf exponent estimate.
- $x_{\min}$: upper cutoff in rank r determined by MLE procedure.
- D: Kolmogorov-Smirnov (KS) statistic.
- p-value determined by the MLE procedure (note that higher is better in that the null hypothesis is more favored).
- $1-\alpha$: Estimate of Zipf exponent $\hat{\theta}$ based on the Simon (1955) model where α is the introduction rate of new terms. We estimate α as the number of unique terms (N) divided by the total number of terms (M).

which we report for 14 famous works of literature in SI-3.

In Fig. 2.2C we measure covariation between regressed values of $\hat{\theta}$ and the Simon model prediction $1-\alpha$. Since both are subject to measurement error ($\hat{\theta}$ is a regressed quantity and α is only coarsely approximated by N/M), we adhere to Reduced Major Axis regression (Rayner, 1985), which produces equivalent results upon interchanging x and y variables, and hence guarantees that no information is assumed or lost when we place $\hat{\theta}$ as the x-variable).

Random text partitions

To produce the rank-frequency distributions in Fig. 2.3 and words in tables S1–S4, we apply the random partition process to several large corpora from a wide scope of content. These corpora are: twenty years of New York Times articles (Sandhaus, 2008), approximately 4% of a year's tweets (Twitter, 2009), music lyrics from thousands of songs and authors (Dodds and Danforth, 2009), and a collection of complete Wikipedia articles (Wikipedia, 2010). In Fig. 2.2 we also use a subset of more than $4,000$ books from the Project Gutenberg eBooks collection (Project Gutenberg, 2010) of public-domain texts.

A.2 Proof of f_q word conservation

In the body of this document we claim that the random partition frequencies of the phrases within a text T conserve the text's underlying mass of words, M_T. This claim relies on the fact that the partition frequencies of phrase-segments, $t_{i\cdots j}$, emerging from a single clause, t, preserve its word mass, $\ell(t)$. We represented this by the summation presented (Eq. 2.4) in the body of this document, which is equivalent to, $f_q(S \mid t)E_S[\ell(s) \mid t]$, i.e., the total number of words represented by the frequency of appearance of all phrases generated by the q-partition:

$$\begin{aligned} f(S \mid t) \cdot E_S[\ell(s) \mid t] &= \sum_{s \in S} \ell(s) f_q(s \mid t) \\ &= \sum_{s \in S} \sum_{s = t_{i\cdots j}} \ell(t_{i\cdots j}) P_q(t_{i\cdots j} \mid t) \\ &= \sum_{1 \le i < j \le \ell(t)} \ell(t_{i\cdots j}) P_q(t_{i\cdots j} \mid t), \end{aligned} \tag{A.1}$$

which we now denote by $M(S \mid t)$ for brevity. For convenience, we now let $n = \ell(t)$ denote the clause's length and observe that for each phrase-length $k < n$ there are two single-boundary phrases having partition probability $q(1-q)^{k-1}$, and $n - k - 1$ no-boundary phrases having partition probability $q^2(1-q)^{k-1}$. The contribution to the above sum by all k-length phrases is then given by

$$2kq(1-q)^{k-1} + (n-k-1)kq^2(1-q)^{k-1}. \tag{A.2}$$

Upon noting the frequency of the single phrase (equal to the clause t) whose length is n, $(1-q)^{n-1}$, we consider the sum over all $k \leq n$,

$$\begin{aligned} M(S \mid t) \;&=\; (1-q)^{n-1} \\ &+\; [2q + nq^2] \sum_{k=1}^{n-1} k(1-q)^{k-1} \\ &-\; q^2 \sum_{k=1}^{n-1} k(k+1)(1-q)^{k-1}, \end{aligned} \tag{A.3}$$

which we will show equals n. We now define the quantity $x = 1 - q$ (the probability that a space remains intact), and in these terms find the sum to be:

$$\begin{aligned} M(S \mid t) \;&=\; nx^{n-1} \\ &+\; \left[2(1-x) + n(1-x)^2\right] \sum_{k=1}^{n-1} kx^{k-1} \\ &-\; (1-x)^2 \sum_{k=1}^{n-1} k(k+1)x^{k-1}. \end{aligned} \tag{A.4}$$

This framing through x affords a nice representation in terms of the generating function

$$f(x) = \frac{1 - x^{n+1}}{1 - x}, \tag{A.5}$$

which allows us to express the summations through derivatives of $f(x)$:

$$\begin{aligned} &\sum_{k=1}^{n-1} kx^{k-1} = f'(x) - nx^{n-1}, \text{ and} \\ &\sum_{k=1}^{n-1} k(k+1)x^{k-1} = f''(x), \end{aligned} \tag{A.6}$$

to find

$$\begin{aligned} M(S \mid t) &= nx^{n-1} \\ &+ \left[2(1-x) + n(1-x)^2\right](f'(x) - nx^{n-1}) \\ &- (1-x)^2 f''(x). \end{aligned} \tag{A.7}$$

Substitution of the second derivative term

$$f''(x)(1-x) = 2f'(x) - n(n+1)x^{n-1} \tag{A.8}$$

then produces the reduced form:

$$\begin{aligned} M(S \mid t) &= n[f'(x)(1-x)^2 \\ &- (nx^{n+1} - (n+1)x^n)], \end{aligned} \tag{A.9}$$

into which we substitute the first derivative term

$$f'(x)(1-x)^2 = 1 + nx^{n+1} - (n+1)x^n, \tag{A.10}$$

to render

$$\begin{aligned} M(S \mid t) &= n[1+nx^{n+1} - (n+1)x^n \\ &- (nx^{n+1} - (n+1)x^n)] = n, \end{aligned} \tag{A.11}$$

which proves Eq. 2.4. Putting this together into a sum over all clauses, we see proof of Eq. 2.5 naturally follows:

$$\begin{aligned} \sum_{s \in S} \ell(s) f_q(s \mid T) &= \sum_{t \in T} \sum_{s \in S} \ell(s) f_q(s \mid t) \\ &= \sum_{t \in T} M(S \mid t) = \sum_{t \in T} \ell(t). \end{aligned} \tag{A.12}$$

A.3 Parameters for well-known texts

Below are tables showing fits of Zipf's exponent, $\hat{\theta}$, for 14 famous works of literature, along with details of the maximum likelihood estimation (MLE) procedure developed by Clauset et al. (2009). The quantities used in these table are described in SI-1, Materials and Methods.

A.3.1 A Tale of Two Cities

level	$\hat{\theta}$	$x_{\min}$	D	p-value	$1-\alpha$
clause	0.783	3	0.0124	0.961	0.176
phrase	0.951	3	0.00742	0.772	0.603
word	1.15	4	0.0077	0.811	0.925
grapheme	1.56	4	0.0146	0.359	0.986

A.3.2 Moby Dick

level	$\hat{\theta}$	$x_{\min}$	D	p-value	$1-\alpha$
clause	0.296	1	0.0192	0	0.154
phrase	0.902	3	0.0132	0.0626	0.576
word	1.05	7	0.00986	0.61	0.912
grapheme	1.42	13	0.0109	0.953	0.986

A.3.3 Great Expectations

level	$\hat{\theta}$	$x_{\min}$	D	p-value	$1-\alpha$
clause	0.301	1	0.0199	0	0.186
phrase	0.995	5	0.0164	0.225	0.622
word	1.21	4	0.00943	0.526	0.938
grapheme	1.66	3	0.0147	0.181	0.988

A.3.4 Pride and Prejudice

level	$\hat{\theta}$	$x_{\min}$	D	p-value	$1-\alpha$
clause	1	3	0.0204	0.911	0.172
phrase	0.983	3	0.0148	0.149	0.617
word	1.11	18	0.0201	0.662	0.947
grapheme	1.43	24	0.0226	0.698	0.989

A.3.5 Adventures of Huckleberry Finn

level	$\hat{\theta}$	$x_{\min}$	D	p-value	$1-\alpha$
clause	0.881	4	0.0192	0.977	0.197
phrase	0.98	3	0.0119	0.385	0.625
word	1.47	1	0.0183	0.83	0.94
grapheme	1.66	6	0.0239	0.203	0.987

A.3.6 Alice's Adventures in Wonderland

level	$\hat{\theta}$	$x_{\min}$	D	p-value	$1-\alpha$
clause	0.707	2	0.0198	0.711	0.191
phrase	0.906	2	0.0108	0.687	0.555
word	1.14	6	0.0353	0.105	0.899
grapheme	1.19	49	0.0338	0.972	0.975

A.3.7 The Adventures of Tom Sawyer

level	$\hat{\theta}$	$x_{\min}$	D	p-value	$1-\alpha$
clause	0.321	1	0.0208	0	0.188
phrase	1.01	6	0.0173	0.826	0.555
word	1.12	3	0.0162	0.108	0.893
grapheme	1.51	4	0.0134	0.683	0.978

A.3.8 The Adventures of Sherlock Holmes

level	$\hat{\theta}$	$x_{\min}$	D	p-value	$1-\alpha$
clause	0.308	1	0.0231	0	0.191
phrase	0.952	4	0.0093	0.892	0.586
word	1.09	9	0.0144	0.733	0.921
grapheme	1.44	12	0.0191	0.663	0.983

A.3.9 Leaves of Grass

level	$\hat{\theta}$	$x_{\min}$	D	p-value	$1-\alpha$
clause	0.486	2	0.00768	0.783	0.0717
phrase	0.865	3	0.00971	0.463	0.543
word	1.01	6	0.0095	0.78	0.886
grapheme	1.39	7	0.0131	0.692	0.981

A.3.10 Ulysses

level	$\hat{\theta}$	$x_{\min}$	D	p-value	$1-\alpha$
clause	0.34	1	0.0192	0	0.193
phrase	0.912	4	0.0062	0.854	0.551
word	1.05	5	0.00773	0.515	0.887
grapheme	1.48	4	0.00874	0.61	0.983

A.3.11 Frankenstein; Or, The Modern Prometheus

level	$\hat{\theta}$	$x_{\min}$	D	p-value	$1-\alpha$
clause	0.257	1	0.0121	0	0.0741
phrase	0.834	2	0.0085	0.55	0.532
word	1.04	5	0.0215	0.057	0.906
grapheme	1.31	12	0.019	0.682	0.982

A.3.12 Wuthering Heights

level	$\hat{\theta}$	$x_{\min}$	D	p-value	$1-\alpha$
clause	0.927	3	0.0217	0.751	0.178
phrase	0.952	7	0.0104	0.978	0.581
word	1.06	10	0.0163	0.533	0.917
grapheme	1.54	5	0.0165	0.345	0.984

A.3.13 Sense and Sensibility

level	$\hat{\theta}$	$x_{\min}$	D	p-value	$1-\alpha$
clause	0.274	1	0.0176	0	0.142
phrase	0.982	3	0.00945	0.611	0.614
word	1.12	20	0.017	0.907	0.946
grapheme	1.41	28	0.0264	0.584	0.989

A.3.14 Oliver Twist

level	$\hat{\theta}$	$x_{\min}$	D	p-value	$1-\alpha$
clause	0.93	3	0.0152	0.808	0.242
phrase	0.962	3	0.00945	0.439	0.622
word	1.13	8	0.0118	0.695	0.931
grapheme	1.52	7	0.0153	0.521	0.987

A.4 Phrase frequency tables

The following tables contain selected phrases extracted by random partitioning for the four corpora examined in the main text. We provide complete phrase lists in csv format along with other material online at:
http://www.uvm.edu/storylab/share/papers/williams2014a/.

Random text partitions

rank	order=1	order=2	order=3	order=4	order=5
1	the (21763834.00)	of the (1332433.25)	one of the (42955.88)	in the united states (8425.91)	years of age or older (3363.23)
2	in (9935182.25)	in the (1095178.50)	as well as (41878.69)	at the age of (5873.75)	the average household size was (1669.62)
3	and (9708982.00)	to the (443282.25)	the united states (37460.25)	a member of the (5534.50)	were married couples living together (1662.12)
4	of (9015261.00)	and the (404687.00)	part of the (23948.69)	under the age of (5287.88)	from two or more races (1530.73)
5	a (6458405.25)	on the (335456.00)	at the time (17591.44)	the end of the (5013.12)	at the end of the (1512.25)
6	to (5890435.75)	at the (308288.50)	the age of (17212.81)	at the end of (4780.31)	the median income for a (1251.14)
7	was (3290575.00)	for the (282949.75)	the end of (16135.31)	as well as the (3805.84)	the result of the debate (1123.98)
8	is (3203926.00)	he was (276889.75)	according to the (16111.19)	at the same time (3609.44)	of it is land and (863.06)
9	he (2583977.75)	it is (246804.50)	may refer to (15914.88)	years of age or (3375.91)	the racial makeup of the (854.42)
10	on (2577531.25)	with the (233894.38)	member of the (15805.50)	of age or older (3364.88)	has a total area of (847.59)
11	as (2520721.50)	as a (230830.62)	the university of (15243.00)	the population density was (3354.00)	the per capita income for (841.80)
12	for (2409743.75)	it was (209433.25)	a number of (14994.00)	the median age was (3332.41)	and the average family size (838.66)
13	with (2107098.50)	from the (202985.38)	in the early (14390.50)	as of the census (3325.94)	and the median income for (832.59)
14	by (2010245.50)	the first (177129.12)	as a result (14356.69)	households out of which (3290.84)	the average family size was (831.62)
15	it (1960890.50)	as the (172026.62)	a member of (13692.75)	one of the most (2952.97)	had a median income of (831.28)
16	from (1688878.50)	was a (153285.75)	in the united (13589.25)	people per square mile (2875.00)	of all households were made (830.97)
17	that (1616682.00)	in a (152800.25)	he was a (13201.88)	at the university of (2866.38)	at an average density of (830.95)
18	s (1588172.00)	to be (142233.38)	of the population (13129.81)	was one of the (2728.66)	males had a median income (830.89)
19	at (1574302.50)	one of (128960.50)	in order to (12507.44)	for the first time (2684.28)	housing units at an average (829.80)
20	his (1461713.50)	during the (128190.62)	was born in (11809.50)	the result of the (2675.75)	made up of individuals and (829.12)
21	this (1187743.00)	of a (126613.62)	end of the (11779.88)	has a population of (2658.84)	had children under the age (828.27)
22	an (1121850.50)	with a (120564.38)	in the late (11641.56)	on the other hand (2654.81)	someone living alone who was (827.98)
23	are (965128.75)	and a (117848.38)	also known as (11477.12)	as part of the (2650.53)	income for a family was (825.89)
24	or (962634.50)	such as (116356.12)	in addition to (11229.75)	of those under age (2626.69)	had someone living alone who (825.33)
25	were (894722.00)	united states (107440.38)	it is a (11059.50)	during world war ii (2600.66)	householder with no husband present (823.45)
26	also (771224.25)	as well (105543.38)	world war ii (11018.56)	of the united states (2591.84)	had a female householder with (816.72)
27	be (736999.75)	th century (102688.62)	such as the (10948.00)	the median income for (2504.88)	population was spread out with (813.75)
28	has (711456.75)	was the (102566.25)	the result was (10670.12)	as a result of (2403.53)	this is a list of (784.14)
29	after (699095.75)	that the (98832.00)	most of the (10051.62)	he was born in (2381.19)	were below the poverty line (761.86)
30	however (689592.50)	and was (93389.38)	as part of (9636.19)	to the united states (2366.12)	the united states census bureau (743.83)
31	who (678548.00)	there were (88907.25)	he was the (9630.62)	in new york city (2292.53)	of the population were below (735.98)
32	they (674922.00)	after the (86291.12)	due to the (9612.06)	in the summer of (2204.88)	according to the united states (734.53)
33	one (657238.50)	new york (84445.25)	some of the (9501.06)	at the time of (2114.16)	was a member of the (711.28)
34	she (628094.25)	citation needed (83924.50)	at the end (9428.75)	the rest of the (2013.94)	result of the debate was (692.72)
35	had (625329.50)	he is (82930.25)	a population of (9329.19)	in the united kingdom (1997.19)	hispanic or latino of any (671.59)
36	first (581708.50)	there are (81538.25)	it is the (9142.31)	as well as a (1862.66)	at the time of the (643.62)
37	their (565228.50)	and in (79643.25)	there is a (9116.25)	of the population were (1848.16)	he was a member of (640.23)
38	there (563650.50)	part of (79108.00)	new york city (8808.00)	the result was delete (1845.00)	it has a population of (583.19)
39	when (554108.75)	for a (78926.88)	years of age (8519.56)	as one of the (1833.75)	national register of historic places (571.44)
40	new (542938.25)	to a (75288.88)	members of the (8365.31)	a total area of (1827.31)	it had a population of (562.61)
41	i (541554.00)	the united (73217.62)	of the year (8232.00)	was a member of (1718.38)	this list is intended to (554.30)
42	its (540743.50)	has been (72469.62)	the city of (8122.94)	below the poverty line (1715.28)	it does not imply that (553.88)
43	may (501345.50)	according to (71437.38)	this is a (8075.19)	the racial makeup of (1708.22)	interest in adding the link (553.47)
44	have (501019.00)	of his (71418.38)	was the first (7792.12)	the per capita income (1692.88)	below a full report on (552.31)
45	th (499317.50)	for example (67477.38)	was one of (7664.56)	the average household size (1691.25)	or that the involved accounts (550.88)
46	her (490852.25)	the new (66748.88)	in the world (7551.44)	married couples living together (1681.62)	accounts are spamming the link (550.88)
47	years (472193.25)	and is (65799.75)	in the first (7480.19)	a median income of (1677.53)	the external link gets used (549.97)
48	m (470841.75)	the same (65643.25)	in new york (7475.44)	and the average family (1677.31)	the village has a population (541.09)
49	all (448565.75)	the city (64700.38)	the united kingdom (7361.56)	average household size was (1670.56)	village has a population of (538.23)
50	other (443913.00)	this is (61094.62)	in the u (7321.44)	median income for a (1668.47)	of the debate was delete (470.31)
100	under (265543.75)	who was (39346.88)	at the same (5320.88)	in adding the link (1106.97)	does not imply that involved (275.44)
150	because (181436.25)	and he (30431.00)	in the town (4118.00)	for a household in (831.50)	household in the town was (172.14)
200	games (136824.00)	who had (26478.88)	the poverty line (3444.69)	at the start of (677.09)	like some other vfd subpages (128.88)
250	still (114752.25)	the american (22596.00)	the soviet union (3232.75)	is located on the (551.88)	the united states department of (103.78)
300	great (98223.75)	the uk (19042.25)	in the region (2778.00)	a municipality in the (491.28)	in the first round of (91.98)
350	court (84906.00)	in late (16349.38)	name of the (2500.62)	afds for this article (422.00)	an archive of the discussion (84.06)
400	further (75941.25)	and are (14785.12)	the government of (2289.19)	one of the oldest (383.66)	income for the village was (78.66)
450	does (68667.00)	size was (13455.25)	as early as (2140.56)	the origin of the (343.66)	township has a total area (71.88)
500	wrote (60868.25)	less than (12458.00)	according to a (1982.56)	in the development of (320.78)	can be seen in the (66.02)
600	hit (51657.00)	place in (10748.25)	also refer to (1758.62)	the new york city (280.81)	of the new york times (56.72)
700	ground (44973.75)	he took (9618.50)	of the show (1622.31)	he studied at the (252.50)	archive of an rfd nomination (49.69)
800	lower (39646.75)	the canadian (8583.50)	so that the (1498.31)	in the fourth quarter (226.56)	of the church of england (44.61)
900	fall (34990.25)	republic of (7820.88)	of the french (1378.50)	played college football at (210.66)	peerage of the united kingdom (41.45)
1000	ad (31756.75)	with other (7171.38)	united states navy (1246.62)	city has a total (197.91)	fis nordic world ski championships (39.11)
1500	garden (21653.75)	from which (5251.62)	throughout the world (916.56)	at the outbreak of (154.88)	it is found in europe (30.12)
2000	ben (16336.25)	the need (4203.12)	of the oldest (753.75)	also part of the (125.28)	as a result of their (24.70)
2500	band's (12613.25)	formed the (3456.62)	a small village (640.88)	it was performed in (107.12)	the first game of the (21.09)
3000	modified (10278.00)	of information (2951.50)	season in the (562.62)	was entered into the (93.94)	the second single from the (18.69)
3500	md (8490.25)	system is (2609.38)	the highest point (501.00)	as late as the (84.00)	he became chairman of the (16.84)
4000	mythology (7228.25)	water and (2340.88)	team for the (455.06)	he served with the (77.25)	the most important of these (15.38)
4500	joan (6291.25)	in singapore (2118.25)	to the french (416.31)	majority of the population (71.12)	in burgundy in north-central france (14.12)
5000	politically (5492.50)	cold war (1922.25)	has an area (386.50)	old at the time (65.78)	station went on the air (13.20)
6000	ignored (4359.25)	be given (1641.62)	national association of (334.69)	claimed to be the (57.66)	the population has grown at (11.66)
7000	lexington (3519.25)	and perhaps (1437.00)	for two weeks (297.25)	there may be a (51.59)	the university of new brunswick (10.50)
8000	blackburn (2945.50)	much larger (1281.50)	cities in the (268.31)	advanced to the final (46.97)	of the church of st (9.58)
9000	eighteenth (2495.75)	the ride (1156.88)	with the american (244.19)	used to determine the (42.91)	whom he had two children (8.86)
10000	validity (2140.00)	bit of (1053.75)	birth to a (224.81)	of the songs on (39.78)	the new york court of (8.23)
15000	topical (1181.00)	of alcohol (726.62)	the australian national (162.31)	continued to serve as (29.50)	over the next three seasons (6.17)
20000	timeslot (765.75)	history is (556.75)	queen elizabeth i (127.94)	he was also chairman (23.97)	unsuccessful candidate for election in (5.03)
25000	wheatley (545.50)	the peruvian (451.38)	and a friend (107.12)	and asks him to (20.25)	during the making of the (4.31)
30000	epithelial (411.50)	can change (376.50)	it difficult for (92.19)	from george washington university (17.69)	during the two world wars (3.78)
35000	awakes (324.25)	footage from (324.50)	the march of (81.12)	with this surname include (15.75)	was the eldest child of (3.38)
40000	ruck (262.75)	marion county (283.88)	over the area (72.56)	please give a reliable (14.25)	the immigration and nationality act (3.06)
45000	verbandsliga (218.50)	a sculpture (252.88)	his father worked (65.56)	so in order to (13.06)	in new york during the (2.83)
50000	imageshack (185.25)	break with (227.12)	fire at the (59.88)	departed new york on (12.06)	and his family lived in (2.61)
60000	partito (140.50)	injured his (188.62)	for economic development (51.19)	were only able to (10.50)	the permanent court of arbitration (2.30)
70000	akatsuki (110.50)	ships had (160.88)	on the sky (44.88)	as an actress in (9.34)	was first known as the (2.05)
80000	salley (89.75)	various characters (140.12)	stanford university and (39.94)	there is some overlap (8.41)	the northeast end of the (1.86)
90000	huila (74.75)	the sodium (123.88)	though in a (36.06)	during the following decade (7.69)	the australian national rugby league (1.70)
100000	leaven (63.75)	the condemned (110.88)	other two were (32.88)	closely linked with the (7.06)	already at the age of (1.58)

Table A.1: Example phrases for English Wikipedia extracted by random partitioning.

Random text partitions

rank	order=1	order=2	order=3	order=4	order=5
1	the (19034045.00)	of the (922676.50)	the united states (48226.25)	in the united states (7162.22)	at the end of the (599.03)
2	a (8722183.25)	in the (778571.88)	one of the (34160.31)	at the same time (5127.59)	because of an editing error (556.81)
3	and (8175499.25)	he said (506762.62)	in new york (32747.94)	for the first time (3893.78)	the new york stock exchange (514.61)
4	of (7463223.50)	to the (321805.25)	the new york (19706.31)	the new york times (3282.12)	for the first time in (481.62)
5	to (7094522.25)	and the (312622.62)	as well as (19019.81)	in new york city (3036.69)	he is survived by his (478.02)
6	in (6553996.25)	for the (275765.75)	new york city (17266.12)	at the end of (2664.31)	is survived by his wife (454.94)
7	that (3251408.00)	at the (266174.25)	a lot of (14997.94)	the end of the (2560.50)	an initial public offering of (400.08)
8	for (2849787.25)	new york (234356.50)	some of the (12923.62)	a spokesman for the (2556.88)	by the end of the (391.30)
9	he (2720690.75)	in a (228202.25)	part of the (12009.06)	at the university of (2224.84)	the end of the year (354.31)
10	is (2668672.00)	to be (182396.25)	of new york (11626.38)	one of the most (2167.66)	the securities and exchange commission (340.56)
11	it (2252598.00)	with the (180261.50)	president of the (10928.75)	of the united states (2105.25)	for the first time since (328.12)
12	but (2134976.50)	that the (179624.88)	the end of (10895.50)	a member of the (2028.19)	for students and the elderly (298.50)
13	on (2102270.50)	it is (171736.38)	there is a (10682.38)	the rest of the (1907.81)	beloved wife of the late (292.89)
14	with (2090580.50)	from the (165015.00)	director of the (10320.38)	at the age of (1877.81)	he said in an interview (287.44)
15	at (2042863.25)	of a (161459.62)	it was a (10318.81)	to the united states (1832.50)	the dow jones industrial average (276.14)
16	as (1808659.75)	she said (160297.25)	as a result (10075.00)	in lieu of flowers (1794.28)	the executive director of the (270.16)
17	i (1626505.00)	by the (159916.25)	according to the (10053.56)	executive director of the (1718.41)	tonight and tomorrow night at (253.62)
18	by (1573509.50)	it was (159603.00)	in the last (9828.88)	the united states and (1653.31)	in the last two years (243.44)
19	his (1418411.25)	as a (146938.88)	the white house (9593.25)	is one of the (1549.75)	in the new york times (240.67)
20	from (1397015.25)	he was (146862.00)	in the united (9578.31)	of the new york (1541.53)	in the last few years (235.52)
21	who (1317491.75)	is a (142374.75)	the university of (9083.88)	by the end of (1524.62)	in the united states and (229.91)
22	an (1253617.50)	with a (135244.50)	there is no (9027.81)	as well as the (1447.84)	in the middle of the (228.61)
23	are (1179629.75)	and a (126899.75)	it is a (8987.25)	the chairman of the (1339.56)	there are a lot of (222.73)
24	they (1177411.75)	but the (120749.75)	the first time (8735.56)	he is survived by (1330.34)	at the university of california (222.31)
25	not (1163949.50)	one of (118009.62)	in the first (8607.00)	the new york city (1322.84)	the federal bureau of investigation (221.33)
26	be (1140990.25)	for a (113570.88)	a spokesman for (8528.75)	in a telephone interview (1289.75)	the museum of modern art (220.48)
27	this (1017793.00)	the new (107764.88)	at the time (8300.88)	at a news conference (1162.12)	of the new york times (214.25)
28	which (985107.50)	the first (105144.75)	out of the (8246.56)	in the new york (1153.72)	graduated from the university of (210.23)
29	or (927178.00)	united states (103164.62)	in the past (8010.69)	for the most part (1147.06)	the food and drug administration (207.61)
30	new (892914.75)	as the (100548.38)	to be a (7877.38)	a son of mr (1144.06)	but at the same time (201.62)
31	had (865149.00)	is the (95388.62)	this is a (7856.44)	a spokeswoman for the (1103.06)	as a result of the (200.59)
32	one (826293.50)	will be (94356.50)	for the first (7789.44)	as a result of (1066.22)	the metropolitan museum of art (200.20)
33	about (820268.00)	to a (92111.75)	in an interview (7685.56)	a lot of people (1060.12)	the university of california at (193.88)
34	she (799892.00)	the united (91259.75)	he said he (7576.50)	a few years ago (1047.81)	years old and lived in (193.58)
35	s (796792.25)	there is (83281.62)	the number of (7551.12)	of new york city (1034.91)	for the new york times (189.27)
36	we (781654.50)	th street (81072.25)	of the new (7016.19)	new york stock exchange (1024.41)	received a master's degree in (179.23)
37	when (752716.25)	for example (74955.88)	the same time (6904.50)	at a time when (1023.19)	a memorial service will be (179.17)
38	will (704428.00)	according to (70748.12)	it was the (6859.56)	the director of the (1007.72)	new york and new jersey (176.58)
39	there (700976.25)	would be (70553.75)	it would be (6843.44)	survived by his wife (998.25)	president and chief executive of (175.53)
40	their (699595.50)	of his (70529.62)	in the world (6814.81)	as part of the (986.62)	president and chief operating officer (172.33)
41	p (687358.75)	this is (69945.38)	it is not (6789.88)	in the middle of (970.16)	at the time of the (167.44)
42	were (676437.25)	there are (69653.25)	in recent years (6653.56)	and the united states (956.59)	the rest of the world (165.12)
43	years (672249.00)	that he (69545.88)	in the early (6652.31)	from the university of (916.47)	th street and amsterdam avenue (164.03)
44	would (664100.25)	he is (69104.00)	in addition to (6584.25)	i don't want to (901.09)	the end of the day (156.91)
45	you (616708.00)	they are (68165.50)	the united nations (6541.31)	in addition to the (897.94)	the united states court of (155.62)
46	its (611930.00)	years ago (66357.25)	at the same (6344.44)	the first time in (897.12)	for more than a decade (151.02)
47	if (608648.75)	when the (65028.62)	but it is (6272.62)	in an effort to (888.00)	this film is rated r (149.75)
48	her (571742.75)	in his (62736.00)	at the end (6264.12)	as well as a (883.31)	spoke on condition of anonymity (148.44)
49	all (568749.50)	who is (62527.25)	i don't think (6247.25)	in the first half (883.22)	court of appeals for the (148.30)
50	been (552982.75)	and mr (61636.88)	i don't know (6171.06)	president and chief executive (882.94)	in the last five years (147.31)
100	here (259618.25)	to have (44901.62)	executive director of (4271.94)	in the middle east (614.88)	he graduated from the university (111.53)
150	st (168117.75)	trying to (32876.25)	and chief executive (3384.31)	tens of thousands of (498.16)	the virus that causes aids (92.28)
200	information (133141.25)	kind of (26368.75)	not going to (2943.56)	the heart of the (425.56)	secretary of state george p (79.97)
250	young (108081.25)	where he (21971.38)	a long time (2503.81)	the first half of (387.38)	came to the united states (69.20)
300	enough (93902.75)	he did (19303.00)	vice president for (2282.38)	for a total of (347.19)	salt and pepper to taste (63.61)
350	county (79788.75)	to pay (17182.75)	declined to comment (2116.62)	time on the market (315.03)	the new york city opera (58.09)
400	tax (72699.25)	the west (15687.75)	would like to (1993.69)	salt and freshly ground (288.91)	in state supreme court in (53.38)
450	became (65631.00)	to come (14304.75)	to more than (1884.75)	the vast majority of (272.25)	who is in charge of (49.69)
500	doing (59774.25)	the soviet (13439.88)	to build a (1787.25)	he said it was (257.59)	at the university of wisconsin (47.52)
600	quarter (51948.25)	a more (11832.00)	would be the (1569.44)	new york city police (231.50)	the good news is that (42.91)
700	someone (44616.75)	in november (10500.62)	a part of (1436.62)	state supreme court in (211.69)	he is also survived by (39.03)
800	weekend (39540.00)	get a (9667.62)	in a new (1314.62)	they don't want to (194.69)	in the next five years (36.31)
900	plays (35724.50)	given the (8871.62)	but for the (1227.50)	the last several years (184.69)	that he not be identified (33.95)
1000	ask (32280.00)	to show (8172.12)	they would be (1141.31)	those of us who (174.66)	upper east side of manhattan (31.92)
1500	reduce (21437.25)	in late (5853.50)	who heads the (872.69)	of the same name (135.16)	i don't know what to (24.73)
2000	seventh (15906.00)	and up (4597.88)	ought to be (721.38)	will continue to be (111.94)	the democratic congressional campaign committee (20.64)
2500	expansion (12556.75)	why the (3791.38)	of the biggest (621.31)	of the iraq war (97.66)	in the second half and (17.94)
3000	importance (10172.50)	and get (3287.00)	believed to have (545.69)	in front of his (86.75)	a good place to start (16.02)
3500	andy (8297.75)	idea that (2869.38)	he has made (492.75)	it is unclear how (78.78)	of the foreign relations committee (14.52)
4000	assessment (7023.75)	due to (2576.38)	of the report (453.06)	original moldings and detail (72.62)	there's no question about it (13.31)
4500	rye (6046.50)	which may (2336.75)	which he was (417.50)	the second and third (67.31)	the book review last year (12.30)
5000	officiated (5247.00)	ceremony at (2147.50)	affected by the (387.06)	to a multiyear contract (62.81)	the first day of school (11.50)
6000	distinctive (4090.00)	while others (1826.88)	economist at the (340.38)	to pay more than (55.56)	we are unable to acknowledge (10.25)
7000	racist (3296.75)	day for (1604.50)	the number to (305.81)	trinity college in hartford (50.03)	it is a question of (9.31)
8000	cracked (2726.75)	long term (1428.25)	and i hope (278.56)	the results have been (45.59)	filed in state supreme court (8.56)
9000	shrine (2306.25)	three and (1294.75)	throughout the state (256.19)	in the last seven (41.94)	the company went public in (7.92)
10000	handel's (1978.75)	new generation (1181.38)	of the home (236.62)	that the police had (39.03)	if there is such a (7.41)
15000	forgo (1063.50)	states supreme (818.12)	there are fewer (175.75)	its way through the (29.50)	of economics at the university (5.64)
20000	fujitsu (666.75)	come at (627.62)	a room with (140.88)	the history of american (24.12)	that donations be made to (4.62)
25000	refrained (456.50)	north fork (508.38)	to explain to (118.50)	and mayor david n (20.56)	that the soviet union would (3.97)
30000	tree' (335.25)	to disarm (427.50)	going for it (102.44)	the best they can (18.12)	a former republican senator from (3.52)
35000	afrikaans (256.00)	close and (367.12)	out of character (90.56)	end zone for a (16.22)	the east and the west (3.17)
40000	rushers (201.75)	by louis (321.62)	a maze of (81.19)	it also plans to (14.75)	and does not want to (2.89)
45000	andrews's (162.25)	after hitting (285.50)	sit in a (73.88)	the new law will (13.50)	he said the white house (2.67)
50000	hearne (133.00)	candidate is (256.88)	eastern european countries (67.81)	confirmed that he had (12.53)	it was the first victory (2.48)
60000	inxs (94.50)	accounting standards (213.50)	to use for (58.38)	of people in this (11.00)	and this was one of (2.19)
70000	airships (69.75)	compensation and (181.75)	doing enough to (51.31)	as if it could (9.81)	until the end of world (1.97)
80000	wei-sender (53.75)	dairy farmers (157.50)	he had missed (45.88)	new jersey attorney general (8.88)	game in the eighth inning (1.78)
90000	willan's (42.75)	table tennis (138.88)	and special events (41.50)	this is a town (8.12)	pleaded not guilty to all (1.64)
100000	prosecutable (35.00)	caught with (124.00)	you are ready (37.88)	i can't say enough (7.50)	the end of the new (1.53)

Table A.2: Example phrases for the New York Times extracted by random partitioning.

Random text partitions

rank	order=1	order=2	order=3	order=4	order=5
1	http (14482019.75)	in the (196458.75)	new blog post (34056.56)	i just took the (5910.19)	i favorited a youtube video (1839.47)
2	com (6428552.75)	i am (157031.25)	check it out (18386.69)	e meu resultado foi (5061.88)	i uploaded a youtube video (1453.28)
3	i (6227774.25)	i just (141596.00)	i love you (15578.25)	other people at http (3254.81)	just joined a video chat (1185.88)
4	ly (5320341.75)	of the (140377.62)	i just took (15341.56)	check this video out (3243.06)	fiddling with my blog post (917.62)
5	the (5180235.50)	on the (137894.88)	live on http (14544.62)	just joined a video (2371.72)	joined a video chat with (813.86)
6	bit (5140992.75)	i love (137768.62)	i want to (13955.88)	a day using http (2061.75)	i rated a youtube video (642.88)
7	a (5044536.50)	i have (136816.38)	i need to (12812.56)	on my way to (2006.66)	i just voted for http (582.91)
8	to (4183208.25)	going to (121491.12)	i have a (12131.88)	favorited a youtube video (1842.59)	this site just gave me (581.73)
9	o (2747181.50)	i think (120492.75)	quiz and got (11955.56)	i favorited a youtube (1839.59)	add a #twibbon to your (472.08)
10	rt (2735865.00)	to the (105588.38)	thanks for the (11796.62)	free online adult dating (1659.81)	the best way to get (454.42)
11	and (2671876.00)	to be (103771.38)	what about you (10897.31)	a video chat with (1628.81)	just changed my twitter background (444.41)
12	tinyurl (2630837.25)	i was (92118.25)	i think i (10602.31)	uploaded a youtube video (1461.91)	a video chat at http (375.84)
13	you (2594872.75)	if you (89098.88)	i have to (10443.56)	i uploaded a youtube (1453.50)	photos on facebook in the (356.89)
14	is (2589257.50)	at the (85136.38)	how are you (9339.94)	video chat at http (1435.12)	check it out at http (351.31)
15	in (2278977.00)	i know (81260.50)	looking forward to (9084.25)	what do you think (1435.09)	own video chat at http (341.72)
16	it (2209243.75)	have a (81252.50)	acabo de completar (9008.31)	i am going to (1398.34)	s channel on youtube http (304.75)
17	me (2049880.50)	to get (79410.75)	i love it (8357.44)	if you want to (1359.84)	and won in #mobsterworld http (293.28)
18	d (2017195.75)	this is (78757.50)	a youtube video (8342.50)	i wish i could (1356.75)	live stickam stream at http (289.97)
19	my (1967677.25)	and i (78420.62)	to go to (8035.69)	just got back from (1344.84)	on facebook in the album (289.03)
20	of (1925590.25)	but i (77363.75)	of the day (8032.19)	at the same time (1310.53)	added myself to the http (275.00)
21	on (1888067.25)	are you (76166.25)	what'll you get (7927.12)	thanks for the rt (1302.22)	just added myself to the (274.16)
22	bitly (1814009.75)	it is (73377.25)	my daily twittascope (7900.69)	channel on youtube http (1284.31)	alot of followers using http (251.34)
23	s (1644797.75)	i need (71952.50)	if you want (7526.38)	have a great day (1262.06)	has just done a job (235.92)
24	n (1528713.00)	it was (70856.62)	going to be (7514.31)	joined a video chat (1186.28)	of followers check out http (232.44)
25	lol (1514066.50)	is a (68456.88)	i don't know (7512.19)	is going to be (1167.62)	i love you so much (228.77)
26	p (1439044.75)	i want (67746.75)	i wish i (7496.06)	trying to figure out (1145.16)	if you want to get (225.47)
27	that (1325029.75)	i don't (67579.62)	is going to (7426.94)	thanks for the follow (1090.50)	hey i just got alot (202.48)
28	at (1297915.75)	i can (67569.12)	going to bed (7393.75)	to your avatar now (1077.56)	mb and the humidity is (202.31)
29	just (1260488.00)	to go (67275.00)	one of the (7351.12)	what are you doing (1069.03)	more followers go to http (201.09)
30	u (1232849.00)	just voted (66498.25)	a lot of (7075.44)	can't wait to see (1031.53)	make your own video chat (200.94)
31	de (1217523.25)	thank you (65912.62)	i feel like (7008.06)	com twitter directory under (1011.69)	you should check this site (199.98)
32	lt (1189756.75)	want to (64412.75)	i just got (6909.56)	check it out http (1004.12)	site out if you want (197.48)
33	this (1086481.00)	listening to (63914.62)	i need a (6752.81)	i have no idea (947.03)	where the wild things are (197.03)
34	e (1076997.00)	in a (63833.88)	in the morning (6713.38)	add a #twibbon to (944.44)	o luansantanaevc liga para voc (190.50)
35	so (1069598.50)	right now (63056.50)	on my way (6635.31)	with my blog post (933.28)	joined a video chat at (186.12)
36	www (1061481.75)	to do (61255.25)	let me know (6630.00)	i don't want to (929.91)	getting ready to go to (186.09)
37	no (1042255.25)	have to (59581.12)	just took the (6078.25)	fiddling with my blog (924.09)	keep up the good work (183.22)
38	gt (974932.50)	is the (58904.38)	meu resultado foi (6037.75)	i need to get (893.50)	gets you tons of followers (181.47)
39	t (973331.25)	on my (58570.00)	can't wait to (5670.00)	i want to go (865.00)	i just become a member (178.52)
40	with (964515.25)	you are (57962.88)	to be a (5662.25)	just got home from (861.56)	am i the only one (174.88)
41	but (962587.00)	do you (57464.62)	just woke up (5596.50)	thank you so much (855.88)	let me know if you (174.84)
42	im (961024.75)	at http (56957.25)	i just voted (5560.06)	the rest of the (851.44)	if you trying to get (173.91)
43	now (942129.00)	i got (56806.25)	what do you (5274.19)	going to be a (848.88)	on my way to the (173.20)
44	do (940366.75)	need to (56329.25)	just joined a (5162.19)	the best way to (831.09)	your own video chat at (170.86)
45	m (935591.50)	vote too (56158.62)	i am so (5077.88)	i wish i was (817.91)	at the end of the (163.81)
46	have (870991.75)	the best (55650.75)	e meu resultado (5062.62)	the end of the (789.47)	on my way to work (163.31)
47	be (868701.00)	try it (52316.88)	in the world (4978.25)	check out this site (787.66)	trying to figure out how (163.19)
48	twitpic (836558.50)	will be (51355.38)	happy new year (4964.62)	i can't wait to (784.03)	looking forward to your tweets (158.67)
49	up (816280.25)	i will (50980.00)	getting ready to (4894.19)	i am listening to (760.44)	this is going to be (154.58)
50	what (809227.00)	took the (50614.88)	getting ready for (4842.00)	to go to bed (757.81)	i want to go to (154.06)
100	know (434311.75)	on twitter (35430.38)	go to the (3525.19)	of followers using http (517.44)	sign up free and get (102.12)
150	come (265721.25)	good night (25500.38)	to see you (2833.94)	boa noite a todos (420.22)	don't know what to do (87.19)
200	watch (195026.50)	i mean (20696.62)	are you doing (2347.06)	more for gemini http (358.44)	the end of the world (65.86)
250	music (154807.00)	of it (18114.25)	at the end (2019.00)	a lot of people (310.47)	you know you want to (58.53)
300	soon (131072.75)	get the (15471.12)	out of my (1846.62)	all of a sudden (281.06)	us has given loan amount (52.56)
350	tell (113651.00)	i saw (13997.00)	to get the (1718.12)	i miss you too (259.38)	has been updated on nicedealz (48.59)
400	id (98939.25)	at a (12575.50)	to make it (1603.75)	a member of this (236.84)	now playing on smooth sounds (44.88)
450	gotta (85315.75)	i might (11760.38)	what happened to (1474.69)	how to make money (223.25)	i love him so much (42.11)
500	ne (76840.25)	able to (10855.50)	on the radio (1385.44)	calling it a night (209.53)	a day it work great (39.33)
600	care (63429.75)	my hair (9497.50)	to hear that (1256.88)	what do you want (191.09)	learn the trick discovered by (34.69)
700	once (52620.75)	as the (8335.25)	first day of (1133.31)	thanks for the #followfriday (177.00)	i just snapped a new (31.94)
800	final (45764.00)	me out (7537.88)	day at work (1047.62)	going to bed early (159.59)	only a matter of time (29.42)
900	search (40726.00)	what u (6958.75)	all is well (957.31)	that would be a (148.81)	need to go to the (27.66)
1000	jackson (37068.25)	next year (6506.38)	be on the (906.00)	know what that means (139.56)	you want more followers check (26.00)
1500	program (24182.75)	but then (4861.50)	when you are (689.25)	is supposed to be (106.97)	just woke up from my (20.38)
2000	jones (17577.75)	reminds me (3908.38)	not the only (576.38)	to see you in (88.22)	made me laugh out loud (16.84)
2500	pengen (13684.50)	playing with (3222.12)	not sure i (495.19)	i feel for you (76.06)	my friend made this great (14.50)
3000	host (11228.75)	a person (2784.50)	they are the (435.19)	long way to go (67.12)	get in the way of (12.94)
3500	ghost (9481.00)	when a (2445.00)	there was no (386.44)	hope it gets better (60.88)	so sorry for your loss (11.73)
4000	chi (8112.25)	this a (2158.62)	and if i (350.12)	this minute was presented (55.94)	home for rent in houston (10.75)
4500	attempt (7138.25)	no homo (1952.12)	to do what (322.62)	i sound like a (52.03)	want to take a nap (9.97)
5000	strength (6254.00)	wanna get (1783.00)	i can just (298.00)	u know what i (48.66)	or however you spell it (9.33)
6000	andr (4908.00)	off your (1514.00)	don't know where (259.31)	best of luck to (43.06)	im going to take a (8.25)
7000	jeremy (3970.00)	a world (1312.62)	of things to (231.00)	will there be a (38.56)	medical and nursing staff wanted (7.44)
8000	domestic (3314.75)	em out (1177.38)	just voted demi (208.12)	what not to wear (35.31)	get more great followers at (6.81)
9000	aje (2837.00)	so wrong (1060.62)	yes you do (190.75)	you have a link (32.59)	it just me or are (6.31)
10000	fase (2468.50)	now following (966.38)	will be available (175.81)	to be in bed (30.25)	longs to be romantically entangled (5.91)
15000	lagunya (1421.00)	head off (662.12)	know each other (127.38)	will do my best (22.84)	it was good seeing you (4.47)
20000	bts (946.00)	lovin the (502.75)	is in need (101.44)	we should meet up (18.56)	a little bit of both (3.70)
25000	grandino (690.00)	more photos (407.75)	in the light (84.62)	join our site for (15.72)	going to see where the (3.19)
30000	treysongz (536.50)	jus had (342.38)	in this week (73.31)	a break from work (13.81)	this is a great site (2.80)
35000	helluva (438.00)	some guys (295.38)	too hot for (64.69)	would like to share (12.34)	i am so sorry i (2.50)
40000	woots (368.00)	very scary (259.12)	don't get mad (57.94)	have to stay in (11.16)	how bad can it be (2.28)
45000	combi (315.00)	red dwarf (230.75)	hard is it (52.69)	see some of you (10.22)	we are working on the (2.09)
50000	rfd (274.25)	recession proof (208.00)	little help from (48.25)	giving it a try (9.44)	to your sign today can (1.95)
60000	wwwstickamjp (217.00)	video converter (173.25)	i woukla been (41.50)	want to say i (8.22)	a long time ago i (1.72)
70000	casbah (180.00)	wonderful evening (148.12)	holidays to all (36.31)	vu all over again (7.31)	don't have enough on my (1.55)
80000	coolman (151.00)	via de (129.62)	min to get (32.38)	i thought of u (6.59)	but i can't take it (1.41)
90000	caius (123.25)	a share (115.00)	that's just sad (29.31)	lady gaga or beyonce (6.03)	my body wants to be (1.28)
100000	aliena (102.50)	class right (103.25)	eating a bowl (26.75)	i can't imagine that (5.56)	never do much of anything (1.19)

Table A.3: Example phrases for Twitter extracted by random partitioning.

Random text partitions

rank	order=1	order=2	order=3	order=4	order=5
1	i (668838.75)	in the (28174.25)	i love you (2556.75)	la la la la (514.06)	la la la la la (184.89)
2	you (600813.50)	and i (25040.88)	i don't know (2094.00)	i don't want to (315.31)	na na na na na (93.98)
3	the (576318.50)	i know (17993.00)	i want to (1750.06)	na na na na (281.78)	on and on and on (48.28)
4	and (440698.25)	you know (16977.75)	la la la (1449.50)	in love with you (237.28)	i want you to know (47.70)
5	to (330196.75)	i don't (16237.12)	i want you (1229.00)	i want you to (227.75)	you know what i mean (45.64)
6	me (305085.75)	on the (14977.12)	you and me (1159.00)	i don't know what (201.38)	don't know what to do (45.22)
7	a (301126.50)	if you (13856.62)	i don't want (1105.88)	i don't know why (187.59)	oh oh oh oh oh (40.80)
8	it (219505.25)	to me (13048.50)	i know you (1086.00)	oh oh oh oh (181.59)	da da da da da (40.41)
9	my (205611.00)	to the (12940.75)	i need you (1065.12)	i want to be (172.69)	do do do do do (40.02)
10	in (203916.25)	to be (12614.00)	and i know (1051.62)	know what to do (144.06)	one more chance at love (35.66)
11	that (150464.50)	i can (12372.12)	i don't wanna (914.00)	what can i do (141.41)	i don't want to be (35.38)
12	of (149402.75)	and the (11679.88)	i got a (904.25)	yeah yeah yeah yeah (138.19)	in the middle of the (34.66)
13	on (143576.50)	but i (11512.50)	i know that (903.00)	you don't have to (137.38)	i don't give a fuck (33.81)
14	your (135024.00)	of the (11239.88)	you know i (902.69)	i close my eyes (130.31)	yeah yeah yeah yeah yeah (33.05)
15	but (132235.00)	i can't (10372.88)	i can see (872.62)	you want me to (129.19)	i don't know what to (32.39)
16	all (124985.50)	for you (10147.75)	and i don't (844.81)	you make me feel (128.31)	all i want is you (31.78)
17	so (121375.75)	when i (10046.38)	in your eyes (844.06)	i just want to (128.00)	you know i love you (26.88)
18	no (116877.00)	come on (9924.25)	i don't care (832.06)	da da da da (123.78)	the middle of the night (26.73)
19	we (113865.25)	you can (9686.00)	and if you (825.94)	if you want to (123.06)	the rest of my life (26.34)
20	is (113375.25)	i got (9577.88)	the way you (824.94)	come back to me (121.56)	no no no no no (26.11)
21	for (108828.50)	in my (9473.12)	all the time (817.62)	in the middle of (119.16)	at the end of the (25.30)
22	oh (107477.25)	all the (9467.25)	na na na (790.38)	and i don't know (118.72)	i wanna be with you (22.77)
23	be (107432.75)	i want (9396.50)	don't you know (766.62)	let me tell you (117.66)	all i wanna do is (22.44)
24	love (104438.50)	that i (9190.88)	this is the (766.25)	give it to me (111.97)	no matter what i do (22.41)
25	it's (99026.75)	i am (9141.88)	can't you see (761.19)	you are the one (111.94)	the way you love me (21.42)
26	now (95016.75)	and you (9048.75)	you love me (753.44)	do do do do (111.28)	no matter what you do (21.36)
27	don't (94956.00)	i was (9028.12)	oh oh oh (749.56)	i love you so (111.16)	what you do to me (20.83)
28	yeah (92807.00)	tell me (8783.50)	i wanna be (744.50)	all i want is (109.81)	when i close my eyes (20.31)
29	when (91600.75)	like a (8614.12)	you know that (714.38)	how does it feel (109.69)	and i don't know why (20.09)
30	with (90323.75)	the way (8512.38)	you want to (709.62)	know what i mean (109.12)	let me be the one (19.86)
31	what (90190.50)	to you (8289.50)	you don't know (707.62)	no no no no (104.03)	the end of the day (18.64)
32	this (90120.00)	when you (8157.62)	in my heart (693.69)	to be with you (100.81)	in the name of love (18.50)
33	know (89600.00)	if i (7941.50)	you and i (691.50)	i don't wanna be (97.50)	lemme see you drip sweat (18.00)
34	like (84259.00)	in a (7893.38)	you make me (675.19)	and on and on (96.47)	i like the way you (17.91)
35	just (83346.75)	my heart (7882.88)	if you want (663.81)	the end of the (94.66)	it's been a long time (17.89)
36	baby (83182.75)	for me (7880.50)	yeah yeah yeah (662.38)	i wish i could (93.09)	till the end of time (17.67)
37	do (81926.00)	this is (7754.62)	don't want to (654.62)	don't give a fuck (92.94)	i wish that i could (17.61)
38	up (81529.00)	for the (7570.88)	want to be (624.56)	can you feel it (91.88)	if you want me to (17.47)
39	if (74941.25)	let me (7539.25)	in my life (622.44)	the way i feel (91.00)	see it in your eyes (17.20)
40	chorus (72833.00)	with you (7482.62)	if i could (619.25)	i don't know how (90.47)	no matter what they say (16.78)
41	can (67057.50)	i need (7424.62)	you know what (615.06)	gon play with it (90.00)	and i don't know what (16.73)
42	down (66636.75)	with me (7386.00)	what you want (605.19)	you know that i (89.84)	let me hear you say (16.70)
43	get (63408.50)	you are (7208.25)	i used to (604.88)	at the end of (89.38)	i look into your eyes (16.70)
44	time (62579.50)	i wanna (7083.00)	on and on (595.94)	can you hear me (89.06)	i love the way you (16.64)
45	out (62562.50)	what you (6949.00)	i see you (592.88)	want you to know (88.38)	and i don't want to (16.45)
46	go (62101.75)	love you (6900.38)	in the sky (587.75)	out of my mind (86.62)	when i think of you (16.38)
47	quot (61793.50)	the world (6774.62)	in the air (584.06)	i need to know (86.56)	i look in your eyes (16.31)
48	got (60347.00)	do you (6733.50)	what to do (577.12)	all i wanna do (84.03)	the end of the world (16.16)
49	one (59306.50)	from the (6679.88)	all night long (558.19)	on the other side (83.88)	when the sun goes down (16.11)
50	see (58662.50)	want to (6649.88)	i know i (557.00)	do you love me (83.72)	still in love with you (16.02)
100	that's (28709.75)	in love (4324.38)	i just want (441.88)	that you love me (61.00)	you want me to do (11.83)
150	always (17981.50)	i won't (3225.00)	make me feel (369.31)	take a look at (51.09)	the end of the line (9.78)
200	en (13668.25)	without you (2692.50)	for you to (308.31)	you make me wanna (43.81)	that's the way it goes (8.77)
250	side (10606.00)	when i'm (2280.75)	who i am (277.81)	the rest of my (39.50)	the way that you do (7.88)
300	words (8896.00)	so long (2050.12)	on the wall (254.81)	open up your eyes (36.66)	i want to see you (7.23)
350	coming (7424.50)	have a (1815.25)	no one else (236.19)	get out of my (34.09)	makes the world go round (6.59)
400	ground (6669.25)	that's the (1645.12)	that's what i (218.94)	i don't want no (31.16)	tell me what you need (6.22)
450	death (5688.75)	then you (1506.12)	come back to (206.38)	don't mean a thing (29.25)	hey hey hey hey hey (5.81)
500	slow (5006.25)	i try (1382.25)	just want to (194.12)	goes on and on (27.84)	my my my my my (5.44)
600	cut (3808.00)	here i (1196.62)	i see your (172.44)	me like you do (25.44)	hey ladies drop it down (5.00)
700	grow (3091.25)	love with (1066.25)	in the game (158.62)	in front of you (23.47)	don't know if i can (4.61)
800	shut (2569.75)	my hands (969.25)	not the same (145.62)	you broke my heart (21.91)	it's been so long since (4.30)
900	doo (2167.75)	i tell (879.75)	yes i am (134.25)	me what you want (20.78)	you were the only one (4.06)
1000	seven (1898.75)	s a (802.88)	it was the (126.00)	all that i want (19.69)	just the way it is (3.88)
1500	food (1140.25)	am the (562.75)	a whole lot (95.38)	i wanna thank you (15.59)	mean a thing to me (3.20)
2000	fields (776.75)	caught up (434.12)	give me love (79.25)	got nothing to say (13.09)	a shoulder to cry on (2.78)
2500	vie (575.50)	saturday night (352.00)	yes you are (68.12)	know that you can (11.62)	was it good for you (2.48)
3000	compromise (451.00)	of things (295.38)	all about the (60.12)	is how we do (10.34)	right round like a record (2.27)
3500	couch (363.00)	the white (254.50)	think you can (53.75)	joy to the world (9.38)	your love would be untrue (2.08)
4000	pu (301.25)	they see (223.75)	i can fly (49.00)	if i don't get (8.69)	he was the only one (1.94)
4500	collect (254.75)	we'll have (197.62)	you said you'd (44.81)	give it all to (8.09)	you that we won't stop (1.81)
5000	product (219.25)	you drive (179.12)	want to hold (41.25)	wanna get with you (7.59)	cut me down to size (1.72)
6000	whatchu (169.50)	where you're (149.50)	take a breath (35.94)	your eyes on me (6.78)	round the ole oak tree (1.56)
7000	battered (135.25)	a plane (128.62)	right here in (32.00)	i wish i may (6.19)	move on down the line (1.44)
8000	verloren (111.25)	step out (111.88)	of all that (29.00)	we can make love (5.69)	bow wow wow yippie yo (1.33)
9000	nt (93.25)	fuck what (99.12)	be waiting for (26.44)	who the fuck are (5.25)	to warm a lonely night (1.25)
10000	honda (79.75)	you should've (88.38)	that what you (24.38)	like a loaded gun (4.88)	ain't that what you said (1.19)
15000	fuma (43.75)	little angel (57.50)	i wouldn't mind (17.44)	it's better this way (3.75)	on christmas day in the (0.94)
20000	cooper (28.50)	the undertow (42.00)	the wrong place (13.69)	since she left me (3.09)	and let it all go (0.80)
25000	fishy (20.25)	a major (32.88)	for one last (11.38)	and maybe you can (2.66)	no matter how far away (0.70)
30000	illtown (15.25)	alright baby (27.00)	you should try (9.75)	it take to make (2.34)	t want french fried potatoes (0.62)
35000	ndelo (12.00)	loud enough (22.62)	never give it (8.56)	i came to bring (2.12)	what you gave to me (0.58)
40000	rees (9.75)	view mirror (19.50)	to me a (7.62)	things i'm gonna do (1.94)	love is out the door (0.53)
45000	metaphoric (8.00)	your concern (17.12)	roll roll roll (6.88)	gotta say too much (1.75)	non ci sono solo io (0.50)
50000	memorizing (6.75)	the cancer (15.25)	on the eyes (6.25)	lay on the floor (1.62)	set the floor on fire (0.48)
60000	ajai (5.00)	an' then (12.38)	keep my eye (5.31)	give up on yourself (1.44)	right here next to you (0.44)
70000	aleiki (4.00)	cats be (10.38)	no more runnin' (4.62)	there's only one god (1.31)	gates of the seven seals (0.38)
80000	saatanan (3.25)	blijf ik (8.88)	we'll show them (4.12)	skies from now on (1.19)	we don't even have to (0.38)
90000	sauber (2.75)	yo tell (7.75)	time you say (3.69)	it comes to that (1.09)	ooh when you walk by (0.34)
100000	mosques (2.25)	believe anymore (6.88)	seemed so right (3.38)	but if i leave (1.00)	van de kille stemmen die (0.31)

Table A.4: Example phrases for Music Lyrics extracted by random partitioning.

Appendix B

Context models

B.1 Cross-validation results for missing entry detection

B.1.1 The New York Times

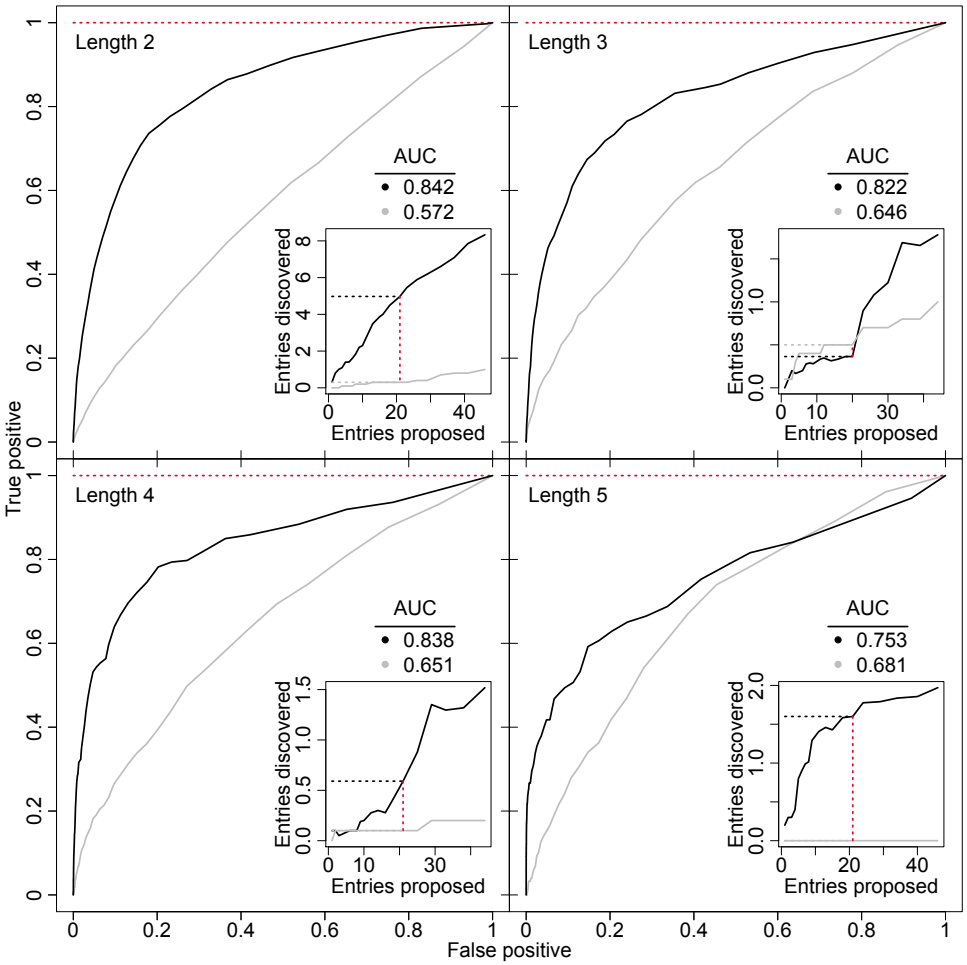


Figure B.1: With data taken from the NYT corpus, we present (10-fold) cross-validation results for the filtration procedures. For each of the lengths 2, 3, 4, and 5, we show the ROC curves (**Main Axes**), comparing true and false positive rates for both the likelihood filters (black), and for the frequency filters (gray). There, we see increased performance in the likelihood classifiers (except possibly for length 5), which is reflected in the AUCs (where an AUC of 1 indicates a perfect classifier). We also monitor the average number of missing entries discovered as a function of the number of entries proposed (**Insets**), for each length. There, the horizontal dotted lines indicate the average numbers of missing entries discovered for both the likelihood filters (black) and for the frequency filters (gray) when short lists of 20 phrases were taken (red dotted vertical lines). From this we see an indication that even the 5-gram likelihood filter is effective at detecting missing entries in short lists, while the frequency filter is not.

B.1.2 Music Lyrics

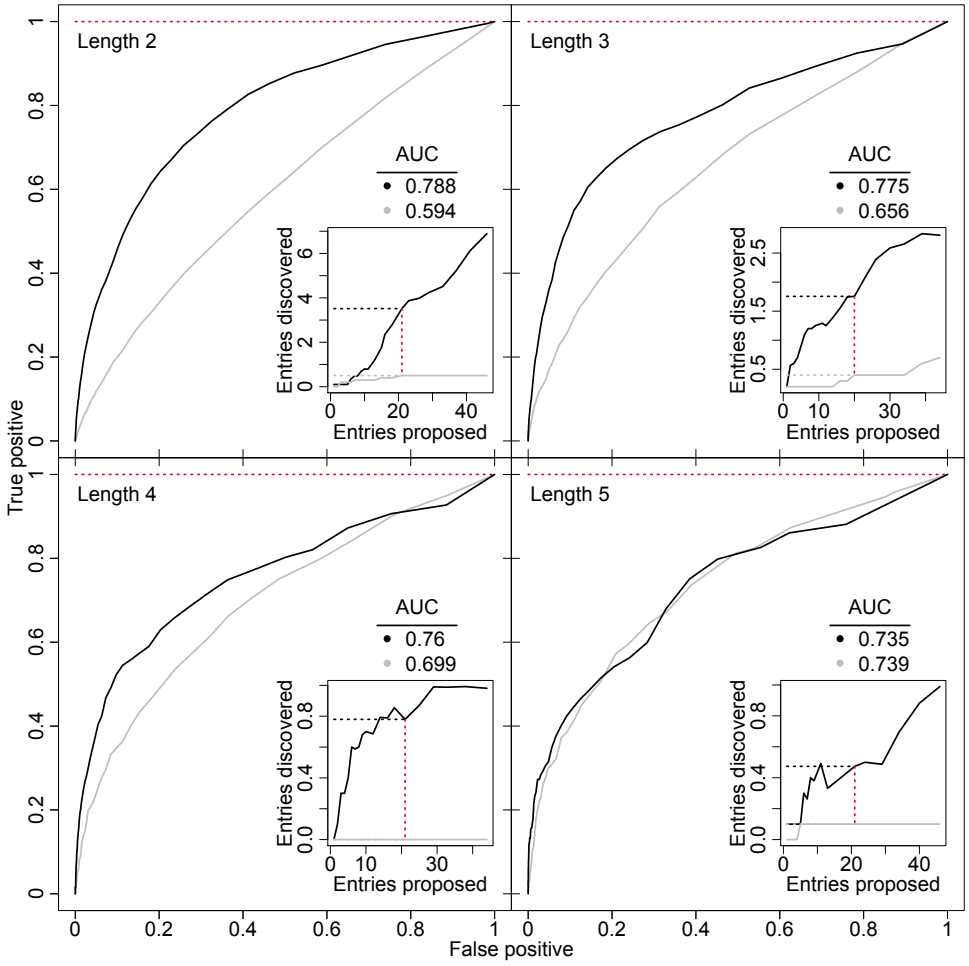


Figure B.2: With data taken from the Lyrics corpus, we present (10-fold) cross-validation results for the filtration procedures. For each of the lengths 2, 3, 4, and 5, we show the ROC curves (**Main Axes**), comparing true and false positive rates for both the likelihood filters (black), and for the frequency filters (gray). There, we see increased performance in the likelihood classifiers, which is reflected in the AUCs (where an AUC of 1 indicates a perfect classifier). We also monitor the average number of missing entries discovered as a function of the number of entries proposed (**Insets**), for each length. There, the horizontal dotted lines indicate the average numbers of missing entries discovered for both the likelihood filters (black) and for the frequency filters (gray), when short lists of 20 phrases were taken (red dotted vertical lines). Here we can see that it may have been advantageous to construct a slightly longer 3 and 4-gram lists.

B.1.3 English Wikipedia

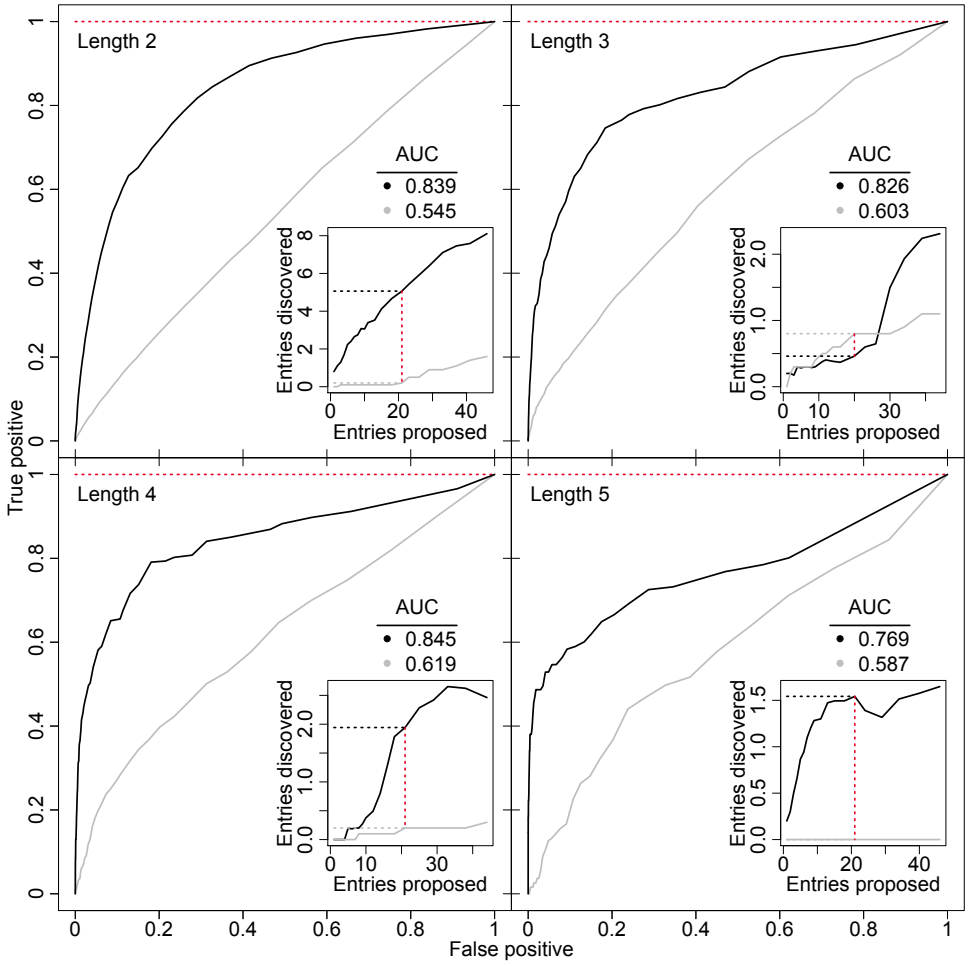


Figure B.3: With data taken from the Wikipedia corpus, we present (10-fold) cross-validation results for the filtration procedures. For each of the lengths 2, 3, 4, and 5, we show the ROC curves (**Main Axes**), comparing true and false positive rates for both the likelihood filters (black), and for the frequency filters (gray). There, we see increased performance in the likelihood classifiers, which is reflected in the AUCs (where an AUC of 1 indicates a perfect classifier). We also monitor the average number of missing entries discovered as a function of the number of entries proposed (**Insets**), for each length. There, the horizontal dotted lines indicate the average numbers of missing entries discovered for both the likelihood filters (black) and for the frequency filters (gray) when short lists of 20 phrases were taken (red dotted vertical lines). Here we can see that it may have been advantageous to construct a slightly longer 3 and 4-gram lists.

B.1.4 Project Gutenberg eBooks

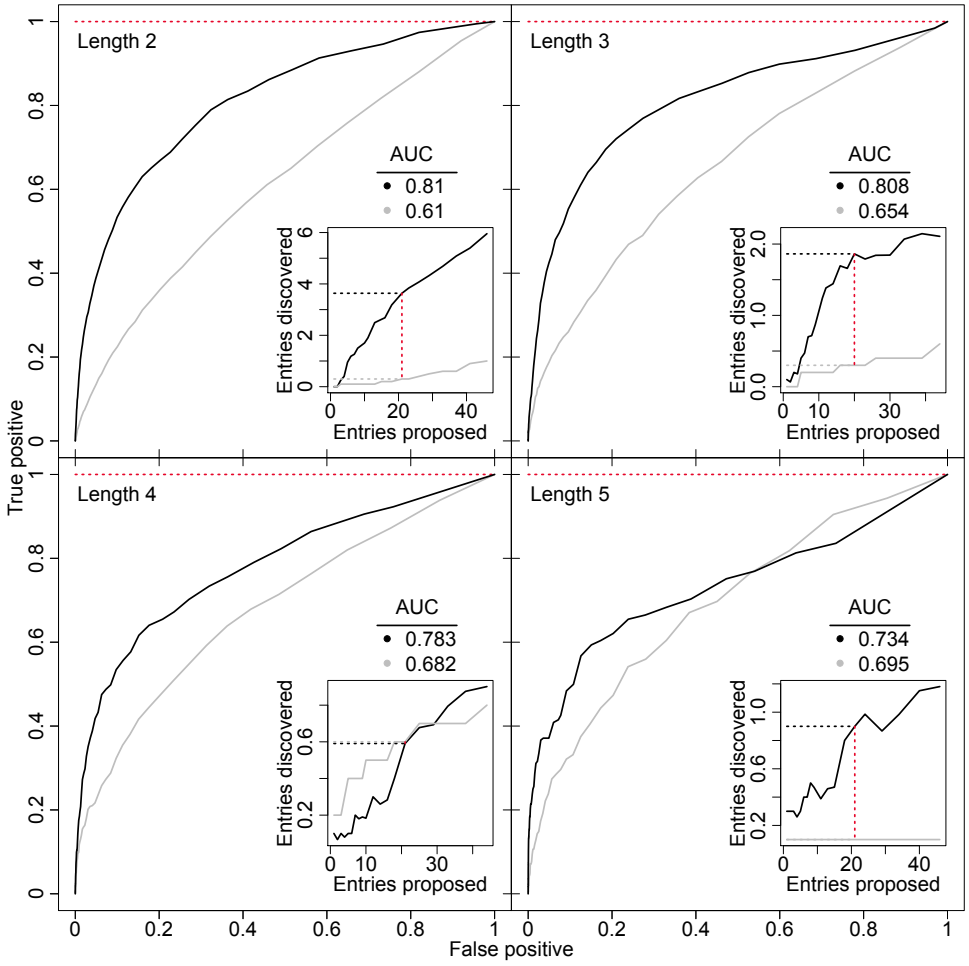


Figure B.4: With data taken from the eBooks corpus, we present (10-fold) cross-validation results for the filtration procedures. For each of the lengths 2, 3, 4, and 5, we show the ROC curves (**Main Axes**), comparing true and false positive rates for both the likelihood filters (black), and for the frequency filters (gray). There, we see increased performance in the likelihood classifiers, which is reflected in the AUCs (where an AUC of 1 indicates a perfect classifier). We also monitor the average number of missing entries discovered as a function of the number of entries proposed (**Insets**), for each length. There, the horizontal dotted lines indicate the average numbers of missing entries discovered for both the likelihood filters (black) and for the frequency filters (gray) when short lists of 20 phrases were taken (red dotted vertical lines). Here we can see that the power of the 4-gram model does not show itself until longer lists are considered.

B.2 Tables of potential missing entries

B.2.1 The New York Times

	rank	2-gram	3-gram	4-gram	5-gram
definition likelihood	1	prime example	as united states	in the same time	when push came to shove
	2	going well	in united states	about the same time	nat. ocean. and atm. admin.
	3	south jersey	by united states	around the same time	all's well that ends well'
	4	north jersey	eastern united states	during the same time	you see what i mean
	5	united front	first united states	roughly the same time	so far as i know
	6	go well	a united states	return to a boil	take it or leave it'
	7	gulf states	to united states	every now and again	gone so far as to
	8	united germany	for united states	at the very time	love it or leave it
	9	dining out	senior united states	nowhere to be seen	as far as we're concerned
	10	north brunswick	of united states	for the long run	as bad as it gets
	11	go far	from united states	over the long run	as far as he's concerned
	12	going away	is a result	why are you doing	days of wine and roses'
	13	there all	and united states	in the last minute	as far as we know
	14	picked out	with united states	to the last minute	state of the county address
	15	go all	that united states	until the last minute	state of the state address
	16	this same	two united states	remains to be done	state of the city address
	17	civil court	its united states	turn of the screw	just a matter of time
	18	good example	assistant united states	turn of the last	be a matter of time
	19	this instance	but united states	turn of the millennium	for the grace of god
	20	how am	western united states	once upon a mattress	short end of the market
	rank	**2-gram**	**3-gram**	**4-gram**	**5-gram**
frequency	1	of the	one of the	in the united states	at the end of the
	2	in the	in new york	for the first time	because of an editing error
	3	he said	the new york	the new york times	the new york stock exchange
	4	and the	some of the	in new york city	for the first time in
	5	for the	part of the	at the end of	he is survived by his
	6	at the	of new york	the end of the	is survived by his wife
	7	in a	president of the	a spokesman for the	an initial public offering of
	8	to be	the end of	at the university of	by the end of the
	9	with the	there is a	one of the most	the end of the year
	10	that the	director of the	of the united states	the securities and exchange commission
	11	it is	it was a	a member of the	for the first time since
	12	from the	according to the	the rest of the	for students and the elderly
	13	she said	in the last	at the age of	beloved wife of the late
	14	by the	the white house	to the united states	he said in an interview
	15	it was	in the united	in lieu of flowers	the dow jones industrial average
	16	as a	the university of	executive director of the	the executive director of the
	17	he was	there is no	the united states and	tonight and tomorrow night at
	18	is a	it is a	is one of the	in the last two years
	19	with a	the first time	of the new york	in the new york times
	20	and a	in the first	by the end of	in the last few years

Table B.1: With data taken from the NYT corpus, we present the top 20 unreferenced phrases considered for definition (in the live experiment) from each of the 2, 3, 4, and 5-gram likelihood filters **(Above)**, and frequency filters **(Below)**. From this corpus we note the juxtaposition of highly idiomatic expressions by the likelihood filter (like "united front"), with the domination of the frequency filters by structural elements of rigid content (e.g., the obituaries). The phrase "united front" is an example of the model's success with this corpus, as it's coverage in a Wikipedia article began in 2006, describing the general Marxist tactic extensively. We also note that we have abbreviated "national oceanographic and atmospheric administration" **(Above)**, for brevity.

B.2.2 Music Lyrics

	rank	2-gram	3-gram	4-gram	5-gram
definition likelihood	1	uh ha	now or later	one of a million	when push come to shove
	2	come aboard	change of mind	made up your mind	come hell of high water
	3	strung up	over and done	every now and again	you see what i mean
	4	fuck am	forth and forth	make up my mind	you know that i mean
	5	iced up	in and down	son of the gun	until death do us part
	6	merry little	now and ever	cry me a river-er	that's a matter of fact
	7	get much	off the air	have a good day	it's a matter of fact
	8	da same	on and go	on way or another	what goes around comes back
	9	messed around	check it check	for the long run	you reap what you sew
	10	old same	stay the fuck	feet on solid ground	to the middle of nowhere
	11	used it	set the mood	feet on the floor	actions speak louder than lies
	12	uh yeah	night to day	between you and i	u know what i mean
	13	uh on	day and every	what in the hell	ya know what i mean
	14	fall around	meant to stay	why are you doing	you'll know what i mean
	15	come one	in love you	you don't think so	you'd know what i mean
	16	out much	upon the shelf	for better or for	y'all know what i mean
	17	last few	up and over	once upon a dream	baby know what i mean
	18	used for	check this shit	over and forever again	like it or leave it
	19	number on	to the brink	knock-knock-knockin' on heaven's door	i know what i mean
	20	come prepared	on the dark	once upon a lifetime	ain't no place like home
	rank	**2-gram**	**3-gram**	**4-gram**	**5-gram**
frequency	1	in the	i want to	la la la la	la la la la la
	2	and i	la la la	i don't want to	na na na na na
	3	i don't	i want you	na na na na	on and on and on
	4	on the	you and me	in love with you	i want you to know
	5	if you	i don't want	i want you to	don't know what to do
	6	to me	i know you	i don't know what	oh oh oh oh oh
	7	to be	i need you	i don't know why	da da da da da
	8	i can	and i know	oh oh oh oh	do do do do do
	9	and the	i don't wanna	i want to be	one more chance at love
	10	but i	i got a	know what to do	i don't want to be
	11	of the	i know that	what can i do	in the middle of the
	12	i can't	you know i	yeah yeah yeah yeah	i don't give a fuck
	13	for you	i can see	you don't have to	yeah yeah yeah yeah yeah
	14	when i	and i don't	i close my eyes	i don't know what to
	15	you can	in your eyes	you want me to	all i want is you
	16	i got	and if you	you make me feel	you know i love you
	17	in my	the way you	i just want to	the middle of the night
	18	all the	na na na	da da da da	the rest of my life
	19	i want	don't you know	if you want to	no no no no no
	20	that i	this is the	come back to me	at the end of the

Table B.2: With data taken from the Lyrics corpus, we present the top 20 unreferenced phrases considered for definition (in the live experiment) from each of the 2, 3, 4, and 5-gram likelihood filters **(Above)**, and frequency filters **(Below)**. From this corpus we note the juxtaposition of highly idiomatic expressions by the likelihood filter (like "iced up"), with the domination of the frequency filters by various onomatopoeiae. The phrase "iced up" is an example of the model's success with this corpus, having had definition in the Urban Dictionary since 2003, indicating that one is "covered in diamonds". Further, though this phrase does have a variant that is defined in the Wiktionary (as early as 2011)—"iced out"—we note that the reference is also made in the Urban Dictionary (as early as 2004), where the phrase has distinguished meaning for one that is so bedecked—ostentatiously.

B.2.3 English Wikipedia

	rank	2-gram	3-gram	4-gram	5-gram
definition likelihood	1	new addition	in respect to	in the other hand	the republic of the congo
	2	african states	as united states	people's republic of poland	so far as i know
	3	less well	was a result	people's republic of korea	going as far as to
	4	south end	walk of fame	in the same time	gone so far as to
	5	dominican order	central united states	the republic of congo	went as far as to
	6	united front	in united states	at this same time	goes as far as to
	7	same-sex couples	eastern united states	at that same time	the federal republic of yugoslavia
	8	baltic states	first united states	approximately the same time	state of the nation address
	9	to york	a united states	about the same time	as far as we know
	10	new kingdom	under united states	around the same time	just a matter of time
	11	east carolina	to united states	during the same time	due to the belief that
	12	due east	of united states	roughly the same time	as far as i'm aware
	13	united church	southern united states	ho chi minh trail	due to the fact it
	14	quarter mile	southeastern united states	lesser general public license	due to the fact he
	15	end date	southwestern united states	in the last minute	due to the fact the
	16	so well	and united states	on the right hand	as a matter of course
	17	olympic medalist	th united states	on the left hand	as a matter of policy
	18	at york	western united states	once upon a mattress	as a matter of principle
	19	go go	for united states	o caetano do sul	or something to that effect
	20	teutonic order	former united states	turn of the screw	as fate would have it
	rank	**2-gram**	**3-gram**	**4-gram**	**5-gram**
frequency	1	of the	one of the	in the united states	years of age or older
	2	in the	part of the	at the age of	the average household size was
	3	and the	the age of	a member of the	were married couples living together
	4	on the	the end of	under the age of	from two or more races
	5	at the	according to the	the end of the	at the end of the
	6	for the	may refer to	at the end of	the median income for a
	7	he was	member of the	as well as the	the result of the debate
	8	it is	the university of	years of age or	of it is land and
	9	with the	in the early	of age or older	the racial makeup of the
	10	as a	a member of	the population density was	has a total area of
	11	it was	in the united	the median age was	the per capita income for
	12	from the	he was a	as of the census	and the average family size
	13	the first	of the population	households out of which	and the median income for
	14	as the	was born in	one of the most	the average family size was
	15	was a	end of the	people per square mile	had a median income of
	16	in a	in the late	at the university of	of all households were made
	17	to be	in addition to	was one of the	at an average density of
	18	one of	it is a	for the first time	males had a median income
	19	during the	such as the	the result of the	housing units at an average
	20	with a	the result was	has a population of	made up of individuals and

Table B.3: With data taken from the Wikipedia corpus, we present the top 20 unreferenced phrases considered for definition (in the live experiment) from each of the 2, 3, 4, and 5-gram likelihood filters **(Above)**, and frequency filters **(Below)**. From this corpus we note the juxtaposition of highly idiomatic expressions by the likelihood filter (like "same-sex couples"), with the domination of the frequency filters by highly-descriptive structural text from the presentations of demographic and numeric data. The phrase "same-sex couples" is an example of the model's success with this corpus, and appears largely because of the existence distinct phrases "same-sex marriage" and "married couples" with definition in the Wiktionary.

B.2.4 Project Gutenberg eBooks

	rank	2-gram	3-gram	4-gram	5-gram
definition likelihood	1	go if	by and bye	i ask your pardon	handsome is that handsome does
	2	come if	purchasing power equivalent	i crave your pardon	for the grace of god
	3	able man	of the contrary	with the other hand	be that as it might
	4	at york	quite the contrary	upon the other hand	be that as it will
	5	going well	of united states	about the same time	up hill and down hill
	6	there once	so well as	and the same time	come to think about it
	7	go well	at a rate	every now and again	is no place like home
	8	so am	point of fact	tu ne sais pas	for the love of me
	9	go all	as you please	quarter of an inch	so far as i'm concerned
	10	picked out	so soon as	quarter of an ounce	you know whom i mean
	11	very same	it a rule	quarter of an hour's	you know who i mean
	12	come all	so to bed	qu'il ne fallait pas	upon the face of it
	13	look well	of a hurry	to the expense of	you understand what i mean
	14	there all	at the rate	be the last time	you see what i mean
	15	how am	such a hurry	and the last time	by the grace of heaven
	16	going away	just the way	was the last time	by the grace of the
	17	going forth	it all means	is the last time	don't know what i mean
	18	get much	you don't know	so help me heaven	be this as it may
	19	why am	greater or less	make up my mind	in a way of speaking
	20	this same	have no means	at the heels of	or something to that effect
	rank	**2-gram**	**3-gram**	**4-gram**	**5-gram**
frequency	1	of the	one of the	for the first time	at the end of the
	2	and the	it was a	at the end of	and at the same time
	3	it was	there was a	of the united states	the other side of the
	4	on the	out of the	the end of the	on the part of the
	5	it is	it is a	the rest of the	distributed proofreading team at http
	6	to be	i do not	one of the most	on the other side of
	7	he was	it is not	on the other side	at the foot of the
	8	at the	and it was	for a long time	percent of vote by party
	9	for the	it would be	it seems to me	at the head of the
	10	with the	he did not	it would have been	as a matter of course
	11	he had	there was no	as well as the	on the morning of the
	12	by the	and in the	i am going to	for the first time in
	13	he said	that he was	as soon as the	it seems to me that
	14	in a	it was not	i should like to	president of the united states
	15	with a	it was the	as a matter of	at the bottom of the
	16	and i	that he had	on the part of	i should like to know
	17	that the	there is no	the middle of the	but at the same time
	18	of his	that it was	the head of the	at the time of the
	19	i have	he had been	at the head of	had it not been for
	20	and he	but it was	the edge of the	at the end of a

Table B.4: With data taken from the eBooks corpus, we present the top 20 unreferenced phrases considered for definition (in the live experiment) from each of the 2, 3, 4, and 5-gram likelihood filters **(Above)**, and frequency filters **(Below)**. From this corpus we note the juxtaposition of many highly idiomatic expresisons by the likelihood filter, with the domination of the frequency filters by highly-structural text. Here, since the texts are all within the public domain, we see that this much-less modern corpus is without the innovation present in the other, but that the likelihood filter does still extract many unreferenced variants of Wiktionary-defined idiomatic forms.