Miller's Monkey Updated: Communicative Efficiency and the Statistics of Words in Natural Language

Spencer Caplan*†[1], Jordan Kodner*†[1], and Charles Yang[1,2]

[1]Department of Linguistics
[2]Department of Computer and Information Science
University of Pennsylvania
3401-C Walnut Street 300C
Philadelphia, PA 19104
{spcaplan,jkodner}@sas.upenn.edu
charles.yang@ling.upenn.edu

May 30, 2020

Abstract

Is language designed for communicative and functional efficiency? G. K. Zipf famously argued that shorter words are more frequent because they are easier to use, thereby resulting in the statistical law that bears his name. Yet, G. A. Miller showed that even a monkey randomly typing at a keyboard, and intermittently striking the space bar, would generate "words" with similar statistical properties. Recent quantitative analyses of human language lexicons, with special focus on the phonological and semantic ambiguities of words (Piantadosi, Tily, & Gibson, 2012), have revived Zipf's functionalist hypothesis. In this study, we update Miller's thought experiment to incorporate empirically motivated phonological and semantic constraints on the creation of words but retain the stochastic mechanism of word generation. Lexicons formed without recourse to functional considerations exhibit the statistical properties that have been previously attributed to communicative efficiency. Furthermore, the updated monkey model provides a good fit for the growth trajectory of English as recorded in the Oxford English Dictionary. Focusing on the history of English words since 1900, we show that the monkey model provides a better embodiment for communicative efficiency than the actual lexicon of English, which further undercuts the functionalist hypothesis. We conclude by arguing for the need to go beyond correlational statistics and to seek direct evidence for the mechanisms that underlie principles of language design.

Keywords: Words; Computational Modeling; Communication; Information Theory

*Correspondence concerning this article should be addressed to both Jordan Kodner (jkodner@sas.upenn.edu) and Spencer Caplan (spcaplan@sas.upenn.edu)

†The authors contributed equally to the work and are listed alphabetically.

1 Introduction

The idea that language functions to facilitate communication has often been met with skepticism in modern linguistics. Chomsky's position on language form and function is well known, starting with the competence-performance distinction (1965). For example, linguistic ambiguity, which can be found at all levels of linguistic structure and poses cognitive processing costs, is regarded as evidence that communicative efficiency is not an essential feature of language (e.g., Berwick & Chomsky, 2016).

The sociolinguistic study of language use, variation, and change in real time provides more direct testing grounds for the role of communicative function. This type of research focuses on specific linguistic processes and is usually complemented by rich information on the social, cultural, and demographic contexts in which these processes take place. It can thus provide a more informative view of the mechanisms of language design than large-scale correlational studies such as the statistical analysis of words in the research tradition reviewed and pursued in the present paper. Here we also find a good dose of skepticism.

Summarizing decades of quantitative research in sociolinguistics, and in a contribution entitled "The overestimation of functionalism", Labov reports little evidence for the role of communicative functions (Labov, 1994). One of the many cases considered by Labov is the deletion of word-final consonants /t/ and /d/ in spoken English (e.g., *walked* is pronounced as "walk"). In a sentence such as "I have walked home", the perfective meaning is doubly expressed by the auxiliary "have" and by the /d/ on the inflected verb. In the simple past "I walked home", by contrast, only the final /d/ on the verb conveys the temporal information. Despite the differences in information content, the rate of /t, d/-deletion does not differ in these contexts (Guy, 1991). Labov's review concludes that "in the stream of speech, one variant or the other is chosen without regard to the maximization of information. On the contrary, the major effects that determine such choices are mechanical: phonetic conditioning and simple repetition of the preceding structures. (pg. 568)". At the same time, Labov finds support for a functionalist interpretation of some, though by no means all, trends in language change. For example, the French feminine plural article *les* (vs. singular *la*) has eliminated the word-final /s/ except when followed by words that begin with a vowel. This results in an increase of ambiguity but the loss of information is partially, though not completely, compensated by an opposition of vowel quality (/le pɔm/ "the apples" vs. /la pɔm/ "the apple").

Indeed, it is in historical studies of language change where functionalist arguments for language are more commonplace; we will return to these insights later in the present paper as well. The 19th-century philologists such as Müller, Schleicher, and others held that the functional pressure of communicative efficiency would gradually shape the structural properties of language, a view that found a receptive audience in Charles Darwin (1888, pg. 91): "A struggle for life is constantly going on amongst the words and grammatical forms in each language. The better, the shorter, the easier forms are constantly gaining the upper hand, and they owe their success to their own inherent virtue." It is interesting to note, however, that the 19th century is also the period that saw the so-called Neogrammarian position rise to dominance in historical linguistics (Campbell, 2013): much of language change proceeds mechanically as it alters the structure of the grammar, affecting all words/units governed by that structure, e.g. a change to a phoneme affects all words/morphemes which contain that phoneme.

It is safe to assume that the relationship between linguistic form and function will continue to occupy language researchers in the years to come. In fact, Chomsky (and Miller)'s own views

on the matter are more nuanced than typically assumed, as we briefly discuss in the concluding portion of this paper. In the current study, we revisit one of the most influential debate on the role of functionalism in language, one which concerns the statistical distribution of words. According to the law that bears his name, G. K. Zipf showed that the rank and frequency of words are approximately inversely correlated (Zipf, 1949). Noting that frequent words are also shorter, Zipf further proposed a *Principle of Least Effort*. Because shorter words are easier to produce, the Principle implies that they will be concentrated in the high frequency range thereby minimizing speaker effort. Critical responses to Zipf were immediate, but the role of communicative function has never ceased to be a focus of research (see Gibson et al. (2019) for a review of recent work). A prominent recent case study of the functional approach (Piantadosi et al., 2012, henceforth PTG) extended Zipf's argument with specific reference to lexical ambiguity. Using English, Dutch, and German lexicons, these authors find that phonologically and semantically ambiguous words tend to be shorter, more frequent, and easier to produce (in a sense to be made clear). Adopting the hypothesis (e.g., Levinson, 2000) that ambiguity can be easily resolved — thereby posing only negligible cognitive cost to the listener — PTG conclude that language is shaped by the function of communicative efficiency.

In this paper, we reevaluate the methods and conclusions of the PTG study in the broader context of functionalism and language. In Section 2, we summarize the main results of the PTG study and the theoretical interpretations provided by these authors. In Section 3, we review Miller's classic response to Zipf, that a monkey striking keys randomly on a typewriter can produce a "lexicon" where shorter words are more frequent, thereby providing a null hypothesis against Zipf's functionalist interpretation of language. Building on linguistic and psychological studies of word well-formedness, we update Miller's argument by presenting a model that incorporates the phonotactic structure of language. The model, dubbed Phonotactic Monkey (PM), generates word forms and assigns meanings to these forms randomly without any communicative/functional considerations, much like Miller's monkey. In Section 4, we show that the PM model exhibits similar statistical properties as the English, Dutch, and German lexicons in PTG. At the same time, while we were able to replicate PTG's results for two of their main results (with respect to word length and frequency), we discuss difficulties with reproducing their results that quantify the effort of articulation ("phonotactic surprisal") and describe the source of the discrepancies which undercuts one of the central results in that study. In Section 5, we augment the PM model with an enhanced semantic component. The resulting Phono-Semantic Monkey (PSM), which closely follows models of word meaning emergence (e.g., Ramiro, Srinivasan, Malt, & Xu, 2018), enables us to directly examine the role of communicative efficiency with historical data. Using words from the Oxford English Dictionary recorded before and after 1900, we show that the PSM model, which generates and assigns word forms and meanings randomly, provides a *better* embodiment of communicative efficiency than the actual words of English: the appeal to communicative efficiency is thus unnecessary to account for linguistic lexicons. In Section 6 we summarize and conclude with a general discussion of the study of functionalism in language.

2 Lexical Ambiguity and Communicative Efficiency: A summary of PTG's study

In this section, we review the methods and findings that PTG take to support the Principle of Least Effort and the hypothesis of communicative efficiency in the lexicon. PTG contends, contrary to

Chomsky's allusion to ambiguity as a dys/anti-functional feature of language but following Zipf's Principle of Least Effort, that ambiguity in fact reflects the efficient use of cognitive resources. PTG makes an important contribution by providing a quantitative and empirically motivated argument for communicative efficiency in language even though it is one which we challenge and disagree with in the end.

2.1 Measures of Ambiguity

The PTG study used three simple and intuitive measures to quantify information and ambiguity in the lexicon. Two of these measures are the familiar notions of *homophony* and *polysemy*. Words with different meanings but coincidentally identical pronunciation are said to be homophonous. For example, *bank* may refer to a financial institution, the side of a river, or the act of turning, meanings that are quite unrelated to each other. Polysemous words, on the other hand, are related semantically, so their identical pronunciations may not be coincidental. For example, *run* in "a run in the park," "a run of wins," "a run in the wool sweater,", and "a salmon run" share some aspect of meaning even though they are not identical. Most of the case studies in PTG used the CELEX corpus for English, Dutch, and German (Baayen, Piepenbrock, & Gulikers, 1995). The CELEX corpus lists the homophonous forms of a word as separate entries; the frequency of a phonological word is the sum of the frequencies for all the entries. Polysemy is computed by counting up the number of senses listed for a word in WordNet (G. Miller, 1998) within part of speech categories (i.e. separately for nouns, verbs, and adjectives). The polysemy study was only conducted for English.

The third measure of ambiguity, syllable informativity, quantifies the information content conveyed by the syllable at the sub-word level. Specifically, syllables that appear in many words are more ambiguous, and thus less informative, than those that appear in fewer words. For instance, the syllable /teɪ/ can appear in the initial position of numerous English words (*ta-ble*, *ta-king*, *tai-lor*, *ta-pir*, etc.) but the syllable /spʌ/ appears as the initial syllable of only one CELEX English lemma (*spu-tter*). Syllable informativity, or the information conveyed by a syllable, is quantified as the number of words it appears in, again using the CELEX corpus for the three languages. While lexical stress disambiguates otherwise similar syllables, PTG removed the stress information on words in their calculations.

2.2 Measures of Communicative Efficiency

Having established three measures of ambiguity, PTG provided three measures of communicative efficiency, two of which are straightforward while the third requires some explanation. The first measure is word length. Shorter words require less articulatory effort, take less time to process, and incur a smaller burden on working memory (Baddeley, Thomson, & Buchanan, 1975; Rayner, 1998), and thus are less costly to produce. The second is frequency, calculated as negative log probability. It is widely known that more frequent words are faster to access and produce (Forster & Chambers, 1973; Murray & Forster, 2004; Rayner, 1998; Whaley, 1978).

Unlike word length and frequency, which are empirical quantities directly estimated from corpora, PTG's third measure of communicative efficiency, *phonotactic surprisal*, is a theoretical construct that relies on the researcher's design choices. Their hypothesis is that as the speaker produces the phonological units of a word in a sequential fashion (Sevald & Dell, 1994), the increased predictability of the next unit given the prior ones will ease articulatory effect. The

phonotactic surprisal of words requires a statistical model of language that assigns probabilities to sequences of phonological units such as phonemes or syllables in a word. Some sequences are impossible, as prohibited by the phonotactic constraints of the language, and among the sequences allowed, some are more common than others. To this end, PTG used a triphone language model that captures phonotactic generalizations statistically. A triphone language model is trained by tabulating the probabilities of strings consisting of three continuous phones in a corpus: more common phone strings will have higher probabilities. For example, consider the triphones /spr/ and /sbr/, the probability of the former is far higher than the latter as /spr/ occurs in many words (*spring*, *express*, *spray*, *mispronounce*, etc.) while /sbr/ can only be found in a few compounds (*housebroken*, *icebreaker*) and some proper nouns. Triphone and similar models have been extensively used in language and speech technology (Jelinek, 1997) and more recently in computational studies of the lexicon (see Section 3.2 for additional discussions). As is standard, smoothing (Chen & Goodman, 1999) is applied to the triphone model to reserve a certain amount of probability mass for triphone strings unattested in the training corpus; the PTG study used the simple *add-one* smoothing to this end. The phonotactic surprisal of a word is defined as the negative logarithm of its probability which, given the multiplicative nature of the triphone model in the assignment of word probabilities, is simply the sum of the negative logarithm of all the triphone components of the word. Note that longer words will have larger values of phonotactic surprisal. To control for this confound, the phonotactic surprisal of words is normalized by its length. PTG's implementation choice in the normalization step critically impacted their results, which will be discussed in Section 4.3.

2.3 Correlating Ambiguity with Communicative Efficiency

To study the role of communicative efficiency in human language, PTG postulated a trade-off between properties of ambiguity, which are represented by homophony, polysemy, and syllable informativity on the one hand, and on the other, properties of efficiency, which are quantified by word length, frequency, and phonotactic surprisal. While the very fact of ambiguity seems at odds with a functional view of communicative efficiency as the listener must work to resolve ambiguity, PTG argue, following Zipf (1949) and Levinson (2000), that the human communicative system is designed to favor the reduction of speaker effort at the expense of the listener. Specifically, PTG assume that the effort involved in language production and comprehension is asymmetrical: the articulatory effort on behalf of the speaker is more costly than the inference needed to resolve ambiguity on behalf of the listener. If so, lexical ambiguity is expected to concentrate in the region of the lexicon that is frequent, short, and phonotactically more probable, thereby favoring the speaker.

Though not the primary focus of the present paper, we note that the assumptions made by these authors are not uncontroversial and may require additional empirical justification. For example, while shorter and more frequent words clearly reduce production effort, there is evidence that words with high phonotactic probabilities are in fact harder to process by both speakers and listeners due to the competition effect from other words in the same (and denser) phonological neighborhood (Luce & Large, 2001; Luce & Pisoni, 1998; Mirman & Magnuson, 2008). Similarly, there is considerable evidence that lexical ambiguity such as homophony does pose significant difficulty on the part of the listener (Rubenstein, Lewis, & Rubenstein, 1971; Van Orden, 1987) even when contextual cues are clear (Boland & Blodgett, 2001). In a naturalistic setting, Labov (2011) also documents evidence that lexical ambiguity, which often results from sound change, can

disrupt linguistic communication and cause misunderstandings. Finally, at least on our reading, Levinson (2000)'s discussion refers to the listener's general success of pragmatic inference so as to avoid gross misunderstandings, and does not directly bear on PTG's hypothesis regarding the real-time trade-off of articulation and comprehension. We nevertheless accept PTG's premises and proceed to evaluate their specific results.

To support the hypothesis of communicative efficiency, PTG computed the correlations between ambiguity and communicative efficiency with linear quasi-Poisson regressions. The main findings are as follows. First, homophony was negatively correlated with word length and frequency in all three languages. It was also negatively correlated with phonotactic surprisal in German and Dutch while a positive but only marginally significant correlation was found for English. Second, polysemy was negatively correlated with word length, frequency, and with phonotactic surprisal for English nouns, verbs, and adjectives. Third, syllable informativity was negatively correlated with its length in phones, frequency, and phonotactic surprisal. Words that are shorter, more frequent, and easier to produce are more ambiguous than words that are longer, less frequent, and harder to produce, which is consistent with the hypothesis of communicative efficiency.

3 Zipf, Miller, and the Phonotactic Monkey

The PTG results are correlational in nature and the authors do not articulate a process by which the trade-off between ambiguity and efficiency is realized in language. As such, the results are open to alternative interpretations: specifically, the observed statistical correlations may be consistent with processes of lexicon formation that make no reference to communicative efficiency. In this section, we first review G. A. Miller's classic monkey thought experiment, a direct response to Zipf's Principle of Least Effort. We then propose a pair of models that mechanically assign word forms to represent word meanings without regard for word length, frequency, processing cost, homophony, or any other measures of communicative efficiency. The later sections explore the statistical properties of the updated monkey models in relation to the hypothesis of communicative efficiency.

3.1 Miller's Monkey

Zipf's Law and its implications for language sciences and technology have been widely recognized (e.g., Baroni, 2005; Jelinek, 1997; Yang, 2013) but the causal factors that result in Zipf's Law have been controversial from the very beginning. As noted earlier, Zipf's own explanation was based on the observation that frequent words tend to be shorter, which he explains functionally by his Principle of Least Effort. Miller's classic paper (1957) is a reassessment of Zipf's statistical result and its purported functional explanation.

Miller proposed the following thought experiment. Imagine a monkey typing away at a keyboard with a fixed probability of hitting the space bar (but never twice in a row) and an equal probability of hitting each of the twenty-six character keys. What would the distribution of these space-delimited "words" look like? It should be clear that short sequences of characters will be more likely than longer sequences of characters, because the probability of the monkey hitting the space bar increases exponentially as the sequence gets longer. Shorter and less effortful words, then, will be more frequent than long words — the very fact for which Zipf's functional principle was proposed. Indeed, the statistical distribution of Zipf's Law can be closely reproduced by random generation processes and has been observed in many (non-linguistic) natural and social

processes (Chomsky, 1958; Conrad & Mitzenmacher, 2004; Kanwal, Smith, Culbertson, & Kirby, 2017; Li, 1992; Piantadosi, 2014).

Miller in fact went further in his critique. Following a mathematical argument by Mandelbrot (1954), Miller claims that randomly generated lexicons are a *better* embodiment of Zipf's Principle of Least Effort than the actual words in languages. Mandelbrot's argument is complex — see G. A. Miller (1954) for an exposition — but it boils down to the following. Because shorter words are more efficient, a truly optimal lexicon should exhaust the space of shorter words before moving on to the space of longer words. Miller's monkey generally better approximates the optimal lexicon as longer monkey words are strictly less likely than shorter words due to the multiplicative nature of word generation. By contrast, the space of shorter words in human language is clearly less than fully saturated. For example, *hap*, *dez*, *gug*, *yesh*, and numerous three-phoneme words do not exist in English while many longer words do. Thus, Miller's monkey can create a more efficient lexicon than actual languages.

Of course no one seriously believes that human language is created by an entirely random process (Howes, 1968), and the "words" generated in Miller's scheme do not at all resemble words in natural language (Piantadosi, 2014). Certainly not all combinations of phonemes or letters fit the phonotactic pattern of a possible word: the "word" *qwsd* may be generated by the monkey but it cannot be a possible word of English as it violates the phonotactic constraints of the language. Similarly, Miller's monkey model has no semantic component: it has no representation for word meanings, never mind how they become associated with word forms (e.g., the monkey's keystroke sequences). Yet the methodological point of Miller's argument remains valid: By providing a null hypothesis, Miller showed that Zipf's Principle of Least Effort is not a unique explanation for the statistics of words, which can be obtained "without appeal to least effort, least cost, maximal information, or any branch of the calculus of variations (pg. 314)." The present study retains the spirit of Miller's argument but updates the monkey model to reflect empirical principles that govern the lexicon. The lexicons generated by the updated monkey model not only exhibit the statistical properties of natural languages uncovered by PTG, but also provide a better embodiment of communicative efficiency than natural languages.

3.2 The Phonotactic Monkey

We now present a new model of lexicon formation, the Phonotactic Monkey (PM) model, that retains the spirit of Miller's classic argument but avoids the artificiality in his random generation process.[1] The PM model, introduced here and enriched and refined in Section 5.1, is intended to capture how new words (form-meaning pairs) are created and conventionalized in language (e.g., Richie, Yang, & Coppola, 2014).

When a new meaning m needs to be expressed, the model randomly generates a phonological item w in the space of possible word forms, which is provided by the phonotactic properties of the language in a sense to be made clear. The model pays no attention to factors such as word length, frequency, phonotactic probability, or any other factor that affects communicative efficiency. If the selected form has already been paired with an existing meaning m', then lexical ambiguity, i.e., a new mapping (m, w) will ensue: homophony, if m and m' are unrelated (e.g., the two senses of the word "bank") or polysemy, if m and m' are related (e.g., the various meanings of "run"

[1] Full implementation and code needed to reproduce all presented analyses available open-source: `https://github.com/jkodner05/ThePhonotacticMonkey`. Additional data and statistical analyses summarized in the present paper but not presented in full can also found there.

reviewed earlier). Note that in both cases, the relatedness between m and m' plays no role at all in the selection of w. If, on the other hand, w has not been previously associated with a meaning, then a new mapping (m, w), i.e., a new word, will be created.

The PM model extends Miller's original thought experiment not only by incorporating a semantic dimension but also through the addition of phonotactics, a fundamental component in a speaker's knowledge of language (Halle, 1978; Hayes & Wilson, 2008). Speakers of English, for instance, will recognize that strings such as *pight*, *clight*, and *zight* could potentially be yet-unknown English words while strings such as *lright*, *dnight*, and *ptight* are decidedly foreign. Phonotactic knowledge is acquired very early and rapidly by children (Chambers, Onishi, & Fisher, 2003; Jusczyk, Luce, & Charles-Luce, 1994) and plays a critical role in both language acquisition (Brent, 1996; Mattys, Jusczyk, Luce, & Morgan, 1999) as well as language processing (Norris, McQueen, Cutler, & Butterfield, 1997; Vitevitch & Luce, 1999). More directly, phonotactics is strongly implicated in the creation of new words. When foreign words are borrowed, they are often adapted to the phonotactic constraints of the native language (Calabrese & Wetzels, 2009; Hyman, 1970). For instance, words of Greek origin such as *pneumatic* and *mnemonic* have nasal consonant clusters which are illicit under English phonotactic constraints. As a result, these words are consistently pronounced with just an initial /n/ rather than /pn/ or /mn/ as their spellings would suggest.

Following PTG, the effect of phonotactic constraints in lexicon formation is captured by a triphone model trained on the CELEX corpus of each corresponding natural language; see Daland et al. (2011) for a summary of recent applications of such models to the psycholinguistic study of phonotactic knowledge. After training, the triphone model is used to generate the phonological words. Specifically, the next phoneme (p_i) is generated probabilistically according to the transitional probability given two immediately preceding phonemes (p_{i-2}, p_{i-1}), i.e. $P(p_i|p_{i-2}, p_{i-1})$. Following the standard practice in computational linguistics, the biphone transitional probability from a START symbol prefixed to the word is used for the generation of the first phoneme in a word. A potential phonological word is completed once the STOP symbol corresponding to a word end is generated, similarly to when Miller's monkey hits the space bar. (We mention the technical implementation involving the START and STOP symbol because they are the cause of a spurious conclusion in the PTG study; see Section 4.3.) The resulting word form is accepted if it follows a minimal word requirement (McCarthy & Prince, 1995) that, for the languages under study, a word must contain at least one syllable which in turn must contain at least one vowel.

The assignment of meaning in the PM model is as follows. Suppose the language has M unique meanings and N unique phonological words. Here we assume $M > N$ as the language permits homophony and/or polysemy. We first generate N unique phonological words by applying the trained triphone model described above. The stochastic generation process may create the same phonological word multiple times, thereby creating the token frequency for that phonological word, again similar to Miller's original proposal. We assume that each phonological word is paired with at least one meaning. For each of $(M - N)$ additional meanings, one of the N phonological words will be chosen to be paired with that meaning, with a probability proportional to its token frequency.

Before proceeding, several technical and methodological remarks about the PM model are in order. First, it is worth emphasizing that our work is not an attempt to reproduce Zipf's Law, whose nature still remains an open question as discussed by the references cited earlier. Likewise, the PTG study does not focus on Zipf's Law *per se* either, but only invokes his more general Principle of Least Effort. Our homage to Miller is more conceptual and methodological: like

Miller, we formulate a null hypothesis devoid of functionalist considerations and demonstrate its compatibility with the quantitative findings identified by PTG and attributed to functionalist considerations.

Second, we have chosen the triphone model as the phonotactic component to be consistent with the PTG study. A plethora of phonotactic models exists in the literature. These models may operate at the level of phonemes and range from simpler uni- and bi-phone models (e.g., Jurafsky and Martin (2009)) to more complex context-free grammar models (e.g., Coleman and Pierrehumbert (1997)), but can also operate at the level of phonological features (e.g., Albright (2009); Hayes and Wilson (2008)). Some models are crafted specifically to reflect the psycholinguistic findings in lexical processing (e.g., Vitevitch and Luce (2004)) while others adopt approaches from other studies of analogy and similarity (e.g., Bailey and Hahn (2001)). It is not our intention to establish the best phonotactic model for language, although we believe that a wide range of phonotactic models will suffice for the purpose of our study and reach similar conclusions. Indeed, we had implemented a simple uniphone — in effect Miller's original monkey — and biphone phonotactic component of the PM model and found that both are capable of reproducing the statistical findings in the PTG study.

Third, several previous studies also used phonotactic models to establish baseline lexicons for their respective studies. For example, Dautriche, Mahowald, Gibson, Christophe, and Piantadosi (2017); Futrell, Albright, Graff, and O'Donnell (2017) and Mahowald, Dautriche, Gibson, and Piantadosi (2018) considered several different phonotactic models for pseudorandom word generation in a process similar to the PM model. But several important differences remain. These models are fitted against the empirical word frequencies of actual languages (e.g., English): as such, they cannot assess the functional relationship between word length and frequency. In addition, these models have no semantic component and thus cannot shed light on the role of polysemy in word formation. Moreover, these models discard any word that has been previously generated, and thus they do not bear on the role of homophony either. The PM model, by contrast, generates word forms, frequencies, and meanings randomly, thereby serving as a baseline null hypothesis against which the communicative efficiency hypothesis can be evaluated.

Lastly, the assignment of word meanings is entirely independent of word forms under the PM model: when a new meaning m is needed, the PM model assigns it to a randomly chosen phonological word. This is obviously unrealistic as new word meanings are often extensions of existing meanings and thus existing phonological words. For example, the meaning of "computer" as a calculating and storage device, which emerged in the 20th century, is clearly related to the meaning of "compute" and thus taking on a similar form. The PSM model, which we present in detail in Section 5.1, implements this process by associating new word meanings to phonological forms whose existing meanings are deemed sufficiently close (e.g., Ramiro et al., 2018). We have chosen to present the PM and PSM models separately for the purpose of clarity as the latter is built on the former. In addition, lexicons generated by both PM and PSM models can reproduce PTG's statistical results. This, along with the insensitivity to the choice of phonotactic models noted above, suggests that the space of possible models compatible with PTG's findings and the communicative efficiency hypothesis is likely very large.

4 Monkey Lexicons Show Communicative Efficiency

4.1 Methods

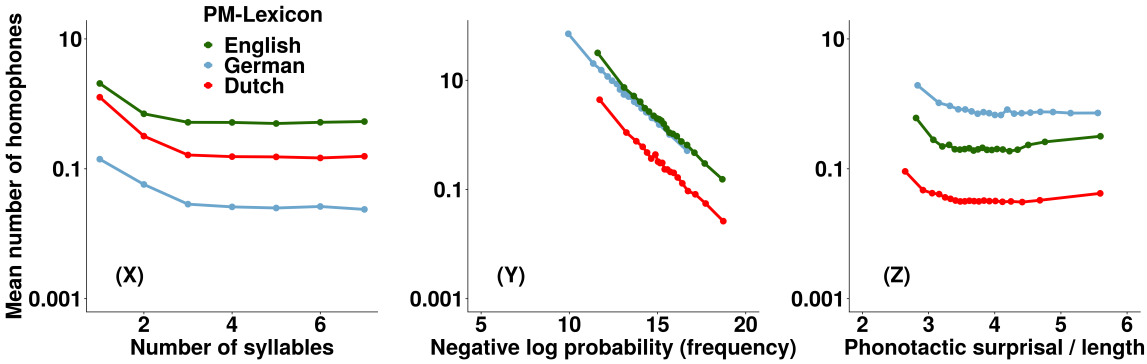
We first sought to reproduce PTG's results over the English, Dutch, and German lexicons using the CELEX database. Following PTG, the polysemy analysis was only carried out for English using WordNet (G. Miller, 1998). Correlations were then calculated between measures of ambiguity (homophony, polysemy, and syllable informativity) on the one hand, and measures of communicative efficiency (word length, frequency, and phonotactic surprisal) on the other. All of the trends reported in PTG were closely reproduced except for phonotactic surprisal which warrants further discussion (Section 4.3).

The main experiment repeated the correlational analysis on the lexicons generated by the PM model as described in Section 3.2. We first trained triphone phonotactic models for English, Dutch, and German using the corresponding CELEX corpora. For each language, we generated a phonological lexicon with the PM described in Section 3.2. Specifically, suppose a language has N phonological words and M meanings as determined via CELEX; the PM was then repeatedly applied until N unique phonological PM-words were generated: the number of times a PM-word is generated will be tallied as its frequency. The M meanings were then distributed randomly across the N words. For the polysemy study (English only), three PM-lexicons were generated in the same way but separately for nouns, verbs, and adjectives, with cardinalities (N's and M's) corresponding to those extracted from the English corpus. For each part-of-speech, we tallied up the total number of senses recorded in WordNet and randomly distributed them over the corresponding PM-lexicon. All PM-lexicons were subjected to the statistical tests in PTG to see if they also exhibited properties attributed to the communicative efficiency hypothesis. For robustness we generated the PM-lexicons ten times using different random seeds; the results were consistent on each run.

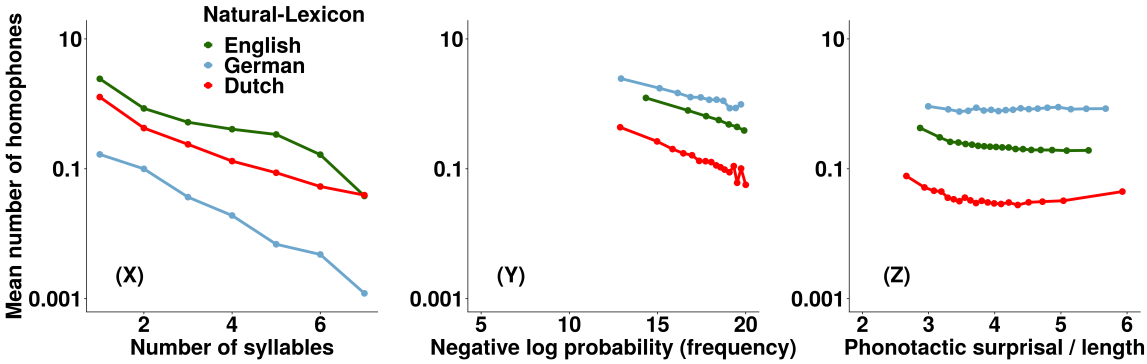
4.2 Results

The PM-lexicons exhibit every significant correlation which PTG uncover in natural lexicons and take as evidence for the communicative efficiency hypothesis. The measures of production ease (length, frequency, phonotactic surprisal) are significantly correlated with ambiguity (homophony, polysemy, syllable informativity) in the PM-lexicons under the same quasi-Poisson regressions applied in the PTG study. For brevity and direct comparability with the PTG's original work, we only report the results from the triphone phonotactic model; as noted earlier, uni- and biphone phonotactic models produced similar results. We summarize these case studies below while more detailed results of the statistical tests are presented in Appendix A.

With respect to homophony, both the natural lexicons and the PM-lexicons show that shorter, more frequent words and those consisting of more common phoneme-sequences are, on average, more likely to be homophonous. The statistical results (Table 1 in Appendix A) are visualized in Figure 1. With respect to polysemy (English only), the regression results on the PM-noun, PM-verb, and PM-adjective lexicons (Table 2 in Appendix A) were all statistically significant and in the same direction as on the actual English noun, verb, and adjective corpora reported in PTG (Figure 2). Finally, syllable informativity was obtained by calculating the number of words each unique syllable appears in. The distributions obtained from the PM-lexicons again showed statistically significant negative correlations for all three measures of ambiguity (Table 3 in Appendix A and



(a) PM-lexicons



(b) Natural-lexicons

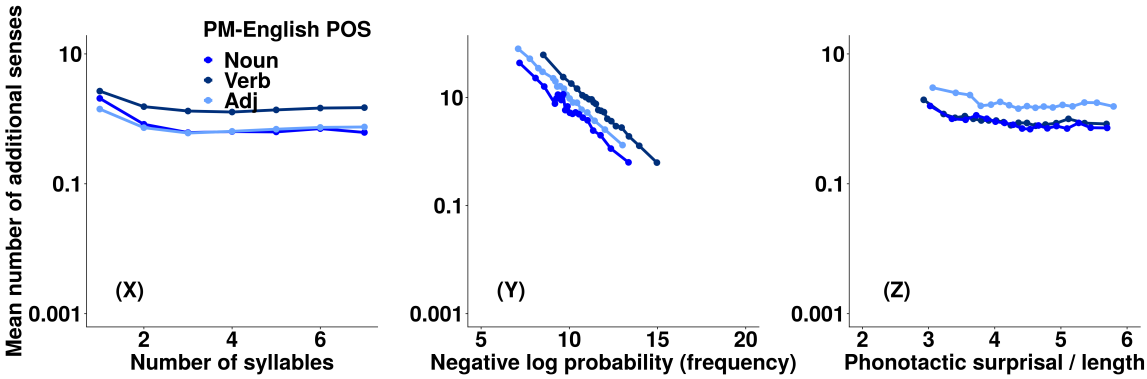
Figure 1: Raw number of additional meanings (homophones) phonological forms have in each corpus, as a function of (X) length, (Y) negative log probability, and (Z) phonotactic surprisal. All y-axes are logarithmically spaced and match the display parameters used in PTG. Our PM-language corpora generated without concern for functional pressures (1a) exhibit the same statistically significant trends observed in natural language (1b).

Figure 3) just like the natural lexicons. Taken together, these results suggest that the statistical distributions of words do not uniquely support the communicative efficiency hypothesis.

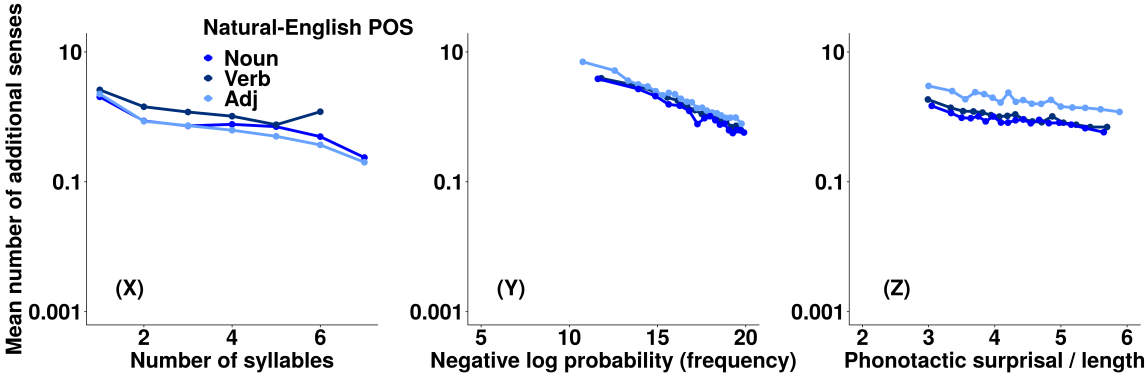
4.3 A Spurious Result for Phonotactic Surprisal

We were initially unable to replicate PTG's results on the effect on phonotactic surprisal. Upon consulting with Steven Piantadosi, the first author of that study, we were able to locate the source of the discrepancies in the way that the phonotactic surprisal measure is calculated. Recall that phonotactic surprisal is a quantitative measure of articulatory ease. Because longer words will have higher values of phonotactic surprisal, it is necessary to normalize the values by word length to control for the confound. We divided each surprisal value by n where n was the length of that word in phones, following the standard practice in the statistical modeling of language in computational linguistics. This is also how word length effects are taken into account in lexical processing research (e.g., Balota et al. (2007)), which has direct bearings on the current and PTG studies that quantitatively measure communicative efficiency. However, PTG's implementation divided all values by $(n + 2)$, which appears to be the length of each word plus the two special START and STOP symbols that were added to the beginning and end of words for the purpose of training the triphone model (reviewed in Section 2.3).

Normalization by $(n + 2)$ rather than word length obviously decreases the calculated surprisal for every word, but the effect is stronger for shorter words than for longer words since the ratio between n and $n + 2$ is smaller for small n. For example, for words with length $n = 3$, the result of normalization by $(n + 2)$ changes the correct result of normalization by n by 40%, but for longer words with $n = 10$, the change is only 16%. Figure 4 plots the difference in phonotactic surprisal calculated with word length and $n + 2$ normalization, demonstrating a significant inverse correlation between word length and difference ($\beta < -0.2$, $t = -297.6$, $p < .001$). The effect is substantial enough to change the correlation between phonotactic surprisal per phoneme and number of homophones. When we normalized phonotactic surprisal using the $(n + 2)$ scheme, we closely matched the PTG results: there is a significant negative correlation between surprisal and homophony or polysemy (For English homophony, $\beta < -0.795$, $t = -22.31$, $p < .001$). However, correct normalization of surprisal by n, the actual word length, fails to produce statistically significant results ($\beta < -0.006$, $t = -1.865$, $p = 0.062$). Thus, PTG's results on phonotactic surprisal are spurious, caused by an unconventional implementation choice and cannot be accepted as a true characterization of natural language lexicons. This technical discussion of an implementation choice should not obscure the larger conceptual point. Should PTG have made an empirically motivated choice of word length normalization and the phonotactic surprisal results been accurate, one still could not draw the conclusion that language follows the functionalist pressure to reuse phonotactically likely words. As discussed in Section 4.2, the PM-lexicons in fact did produce statistically significant correlations between phonotactic surprisal and lexical ambiguity. In other words, the PM-lexicon has higher communicative efficiency than the lexicons in actual language. This result echoes Mandelbrot/Miller's observation that random generation lexicons provide a better embodiment of Zipf's Principle of Least Effort. We expand and strengthen this argument presently.

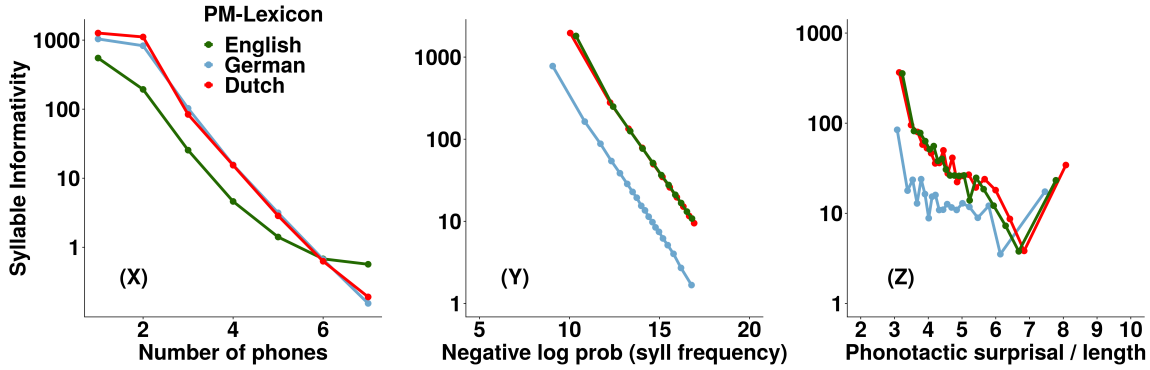


(a) PM-lexicons

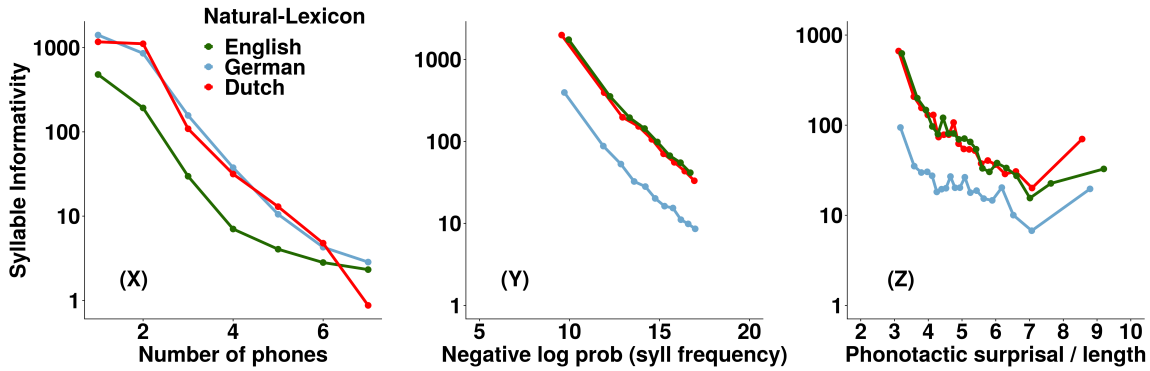


(b) Natural-lexicons

Figure 2: Raw number of additional senses (polysemy) a word has for three part-of-speech categories, as a function of (X) length, (Y) negative log probability, and (Z) phonotactic surprisal. All y-axes are logarithmically spaced and match the display parameters used in PTG. Our PM-language corpora generated without concern for functional pressures (2a) exhibit the same statistically significant trends observed in natural language (2b).



(a) PM-lexicons



(b) Natural-lexicons

Figure 3: Raw number of additional words a syllable appears in (syllable informativity) for each corpus, as a function of the (X) length, (Y) negative log probability, and (X) phonotactic surprisal. All y-axes are logarithmically spaced and match the display parameters used in PTG. Our PM-language corpora generated without concern for functional pressures (3a) exhibit the same statistically significant trends observed in natural language (3b).

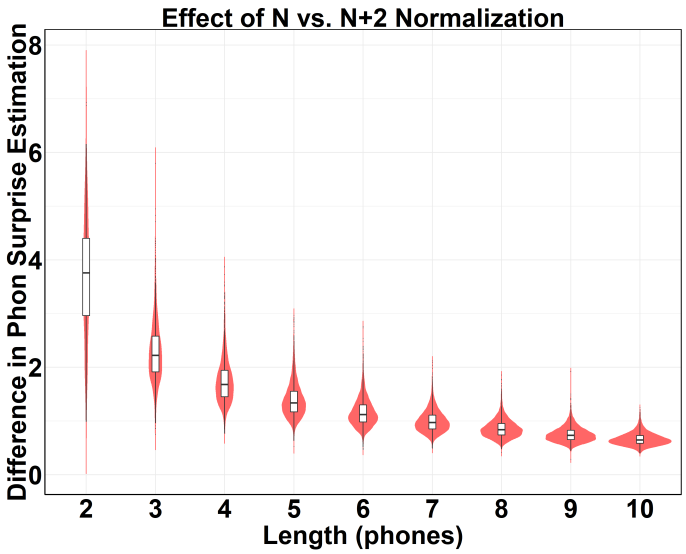


Figure 4: Average difference between phonotactic surprisal values computed using $(n+2)$ normalization and correct n normalization as a function of word length. The error introduced is stronger for shorter words than longer words, introducing a bias into PTG's result.

5 Monkeys Create More Efficient Words than English

The analyses presented so far suggest that the PM model is capable of (re)producing the statistical distributions observed in natural language lexicons that have been attributed to communicative efficiency. It should be pointed out, however, that both PTG's study and our PM model only reflect the *static* properties of language, namely lexicons already fully formed by historical processes. Thanks to the availability of language data with historical depth, we are now in a position to more directly evaluate the communicative efficiency hypothesis. This will enable us to quantify the degree of efficiency in word formation, and compare the results from English words against a null hypothesis in the spirit of Miller and Mandelbrot.

In order to do so, we begin by introducing a semantic extension to the PM model: the Phono-Semantic Monkey (PSM) incorporates an empirically motivated mechanism for the emergence of words that gives rise to polysemy and homophony. In Sections 5.1 and 5.2, we show that the lexicon growth pattern under the PSM model closely tracks the historical trajectory of English words recorded in the Oxford English Dictionary (OED). Having established the plausibility of the PSM model, we can calculate and compare the efficiency of lexicon emergence in "Monkey English". In Section 5.3, we describe the result of a thought experiment in which monkeys got hold of the English lexicon in the year 1900 and proceeded to create new words until present day: we show that the randomly generated Monkey English would have made more efficient use of the resources in the 1900 lexicon than what has transpired in the actual world of English.

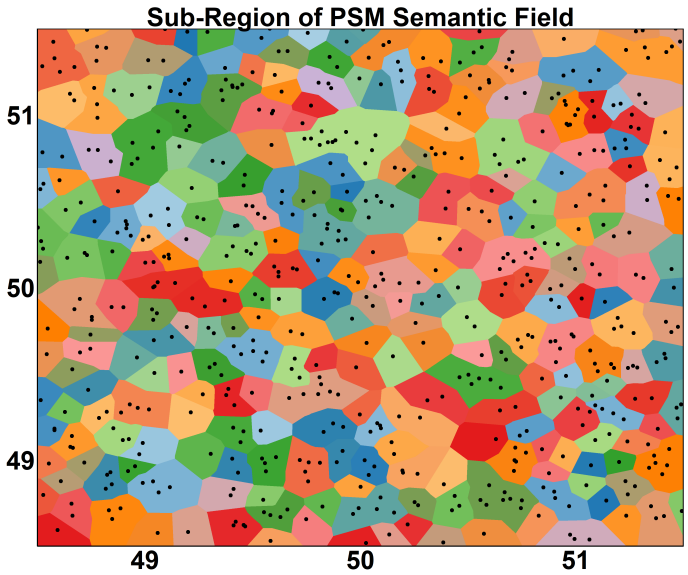


Figure 5: Voronoi diagram depicting a 3×3 region of the semantic field which results from PSM. Black points indicate the locations of individual meanings, and colored polygons indicate the space taken up by sets of polysemes.

5.1 Modeling the Emergence of Word Meanings

The PM model implemented in previous sections has an unrealistic meaning component: a new word meaning is randomly assigned to a word form regardless of the meaning(s) already associated with that word form. The Phono-Semantic Monkey (PSM) model addresses this inadequacy. It builds upon work in the historical study of lexicons (Ramiro et al., 2018; Rodd et al., 2012; A. Xu, Ramiro, & Xu, 2019), with the critical result that new meanings tend to be assigned the forms of existing words with similar meanings. That is, new word senses are more likely to become polysemes with existing word forms with similar meanings than they are to be assigned new word forms. For example, the word *game* has become associated with a large number of similar meanings over the centuries, from 'pastime' and 'jest' in Old English, to a form of 'entertainment,' and 'sport' in Middle English, to a 'method of play,' and 'standard of performance' in Modern English (Ramiro et al., 2018, Fig. 5). These could have been associated exclusively novel word forms, but they were given the same form as *game*.

Word forms are generated in PSM by the same triphone phonotactic model as in PM. Word meanings/senses are represented as points distributed in a 100×100 two-dimensional space of real numbers, where senses closer to each other in the space are said to have more similar meanings. The lexicon is created as follows. To generate each of a specified M new word senses, the PSM model begins by randomly choosing a point in the space, which represents the emergence of a new meaning that will need to be expressed by a word form. If that new sense is close enough to any existing word sense by Euclidean distance, it is given the same word form thereby making

it a polyseme (i.e., a word sharing a form and similar meaning with another). Otherwise, a word form is generated in the same way as the PM model. Like before, the new form may have been used already by an existing word, thereby creating homophony, otherwise an entirely novel word, in both meaning and form, will be created. "Close enough" in this model is defined by a threshold parameter θ, with smaller θ indicating a stricter similarity criterion resulting in fewer polysemes. Unlike in the PM model, the total number of word forms N is not fixed in the PSM model and is instead a function of M and θ. New forms are only generated when new word senses are sufficiently distant from existing ones: the value of θ can be tuned to produce the desired values of N and M as found in the lexicon of actual languages. While the value of θ in the current study is a parameter we adjust, it is ultimately a reflection of semantic divergence that language users tolerate while recycling old word forms before new word forms are deemed necessary; as such, it can be studied further with suitable behavioral and quantitative methods.

The process for calculating sense relationships closely follows Ramiro et al.'s (2018) *nearest-neighbor chaining algorithm*. It is strictly local in that it does not require tracking or optimizing over the entire lexicon or its history: for instance, it does not retain information about which senses were added when, or what polysemous set any existing sense is a part of. The PSM model has the effect of forming clusters of polysemes in the semantic space, and it can also form homophones if two distant meanings in the semantic space happen to receive the same form from the phonotactic model. Figure 5 shows a Voronoi visualization of a 3×3 window into the 100×100 semantic space after 700,000 word senses (156,912 forms, $\theta = 0.1$) are generated, with polygons highlighting polysemous sets. These values of word senses and forms are approximately those of English words up to the year 2000 in the OED; Section 5.2 provides additional details. As θ increases, so does typical cluster size. All the same, the PSM model shares the mechanical aspect of the PM model: the creation of words makes no reference to functional considerations.

5.2 The Growth of Meanings and Forms

We now turn our attention to the historical trajectory of English words and the suitability of the PSM model as a null hypothesis of word formation. The OED provides a straightforward means for empirically measuring the accumulation of English word forms and word senses over time. Each form in the OED lists a date of first attestation as well as an enumeration of word senses deemed distinct by lexicographers. The OED contains 152,698 word forms with 689,166 senses (4.513 senses per form) that were first attested in English before the year 2000, which gives an approximate estimate for the distribution of forms and senses in the English lexicon, with the caveat that not all forms or senses in the OED are still in use today and that its entries do not necessarily make a consistent distinction between polysemy and homophony.

To test the suitability of the PSM model, we generated 700,000 word senses with $\theta = 0.1$, which creates a lexicon that is comparable with the OED words. This resulted in 156,912 forms (4.461 senses per form). Each time a new word was added to the lexicon, its form was tracked so that the accumulation of word senses and forms could be plotted. Figure 6 shows a comparison between the empirical accumulation of forms and senses in the OED by year compared to those generated by the PSM. The OED plot ranges from 1500, the year traditionally chosen to delimit the transition from Middle English to Early Modern English, through 1999, and PSM ranges from the 25,000th to the 700,000th word sense added. Since the OED only presents accumulation at the granularity of years, the PSM sense numbers were re-scaled to be comparable with OED years. Crucially, word senses rise faster than word forms in both, indicating a tendency towards reuse of forms, and

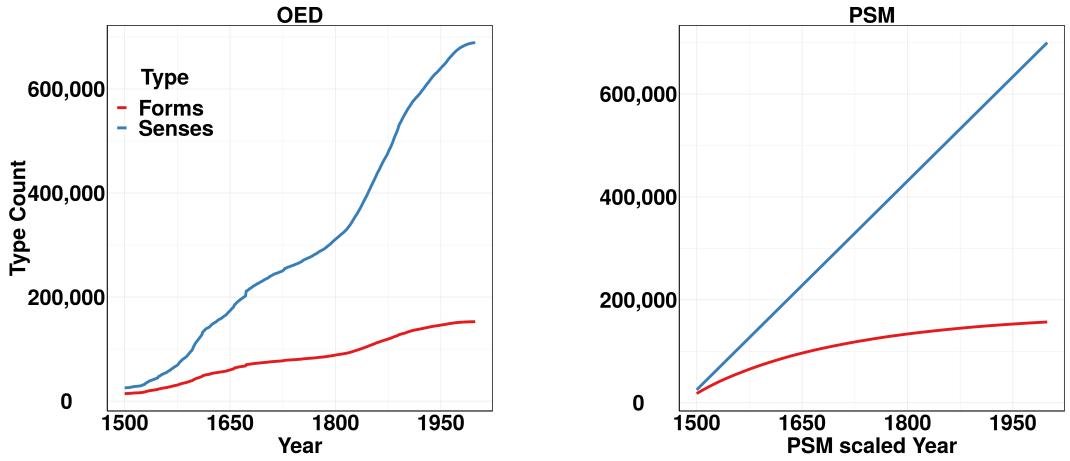
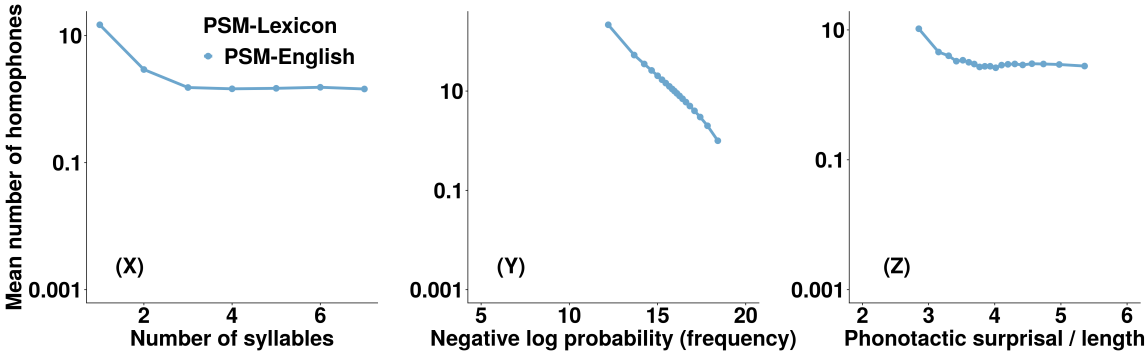


Figure 6: Comparison of the cumulative number of word senses and word forms by year in the OED's entries for modern English (1500-1999) and those generated by the PSM (new senses 25,000-700,000). The x-axis of the PSM plot is scaled in order to be directly comparable with the OED.

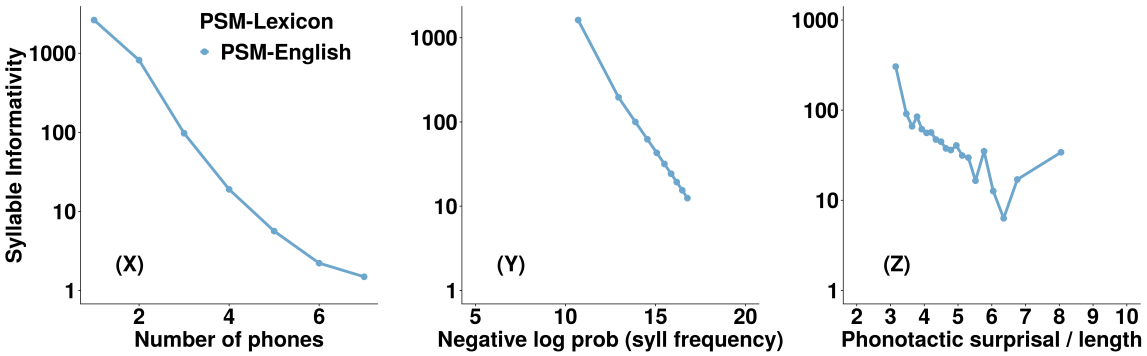
416,666 of 700,000 meanings in the PSM simulation were added as polysemes of existing words, showing that new senses result in polysemy more often than not.

To quantify the similarity between the form/sense ratios over time, we fit a linear model predicting each of the 500 OED by-year form/sense ratios (1500-1999), the PSM ratio at the same normalized time-slice, and their interaction. There is a clear trend between PSM ratio and OED ratio ($\beta = 1.550$, $t = 3.30$, $p < .001$) and the model offers an extremely good fit (Adjusted $R^2 = 0.987$) to the empirical data. Note that only the start and end points of the PSM generation were fixed to correspond to the OED, so the accumulation of PSM word forms could conceivably have followed some other path than it did in between. Instead, this ratio was retained even in the periods where the rate of sense accumulation in the OED fluctuated. Thus, PSM serves as a plausible baseline for the historical accumulation of new words in the lexicon. Furthermore, this analysis also provides support for the mechanisms of word sense assignment in previous work (Ramiro et al., 2018).

Does the PSM model also exhibit the statistical properties of the lexicon attributed to communicative efficiency? The answer is yes. When subject to a statistical analysis, the 700,000 word PSM-lexicon exhibits the same significant trends between all metrics of production ease (length, frequency, phonotactic surprisal) and ambiguity (homophony and syllable informativity) as PM and PTG (Figure 7). Correlations were measured using the same quasi-Poisson regressions as in Section 4.2 with more detailed results of the statistical tests presented in Appendix A. These reiterate the results uncovered by the PM model, demonstrating that its minimal assumptions are sufficient to give an appearance of communicative efficiency. An additional benefit of the PSM's added complexity can be seen in its application to the historical investigation of word formation, which we pursue presently.



(a) PSM-English Homophony



(b) PSM-English Syllables

Figure 7: The PSM-English corpus generated without concern for functional pressures (7a-7b) exhibits the same statistically significant trends observed in natural language and PM with respect to homophony (Figure 1) and syllable metrics (Figure 3).

5.3 The Communicative Efficiency of English vs. Monkey English: 1900-2000

Recall the source of lexical ambiguity under PTG's interpretation: short, frequent, and easy word forms are (re)used when new meanings need to be expressed. The historical record of the OED makes a direct testing ground for this hypothesis, again in contrast to a null hypothesis PSM model.

The present analysis implements the following thought experiment. It is in the spirit of Mandelbrot and Miller, that aims to compare the efficiency of a monkey against Zipf's Principle of Least Effort: as they noted, a monkey would exploit the space of short words more efficiently (Section 3.1). Suppose the English lexicon at some particular time T is given to a community of English speakers and a community of random monkeys (here implemented as the PSM model). As time moves forward, both communities will need to express new meanings with words. Some of these words will take on word forms already available at time T, resulting in lexical ambiguity (i.e., homophony and polysemy), while others will be word forms that do not exist at T and are created *de novo*. After a certain period of time, i.e., at time $T + t$, there will be two lexicons, one created by English speakers and the other by monkeys. We can then quantify the efficiency with which these two lexicons have used (and reused) the lexical resources, i.e., the same starting lexicon at time T. If the monkeys prove to be more adroit, then the communicative efficiency hypothesis loses considerable force.

More specifically, suppose the English lexicon consists of a set of words, i.e., form-meaning pairs, at time T. All new meanings that enter into the language after T will take on one of three possibilities. First, new meanings may reuse word forms that exist before or at T; call these "Reused Form" (RF) words. By contrast, there are word forms that exist at T but do not get reused; call these "Stale Form" (SF) words, for lack of a better term. By hypothesis, RF words, which have become more ambiguous, should be more efficient than SF words in a directly measurable way. Finally, new meanings may take on brand new word forms: call these "New Form" (NF) words. While the NF words obviously do not reuse old word forms, they are still expected to show efficiency: for instance, they should preferentially reuse phoneme sequences with higher phonotactic probabilities and thus lower articulatory effort. Using the OED data, we can extract the three sets of words for English — RF_E, SF_E, and NF_E — and quantify the effect of communicative efficiency on these words. For comparison, we model the word formation process under our monkey model (PSM). This takes the same English lexicon at time T as input and proceeds to assign additional word meanings to forms, thereby creating three alternative sets of RF_{ME}, SF_{ME}, and NF_{ME} words ("ME" for Monkey English). We can then quantify the effect of communicative efficiency on these three sets of words in this alternate universe of Monkey English, again as a baseline in comparison with the actual universe of English.

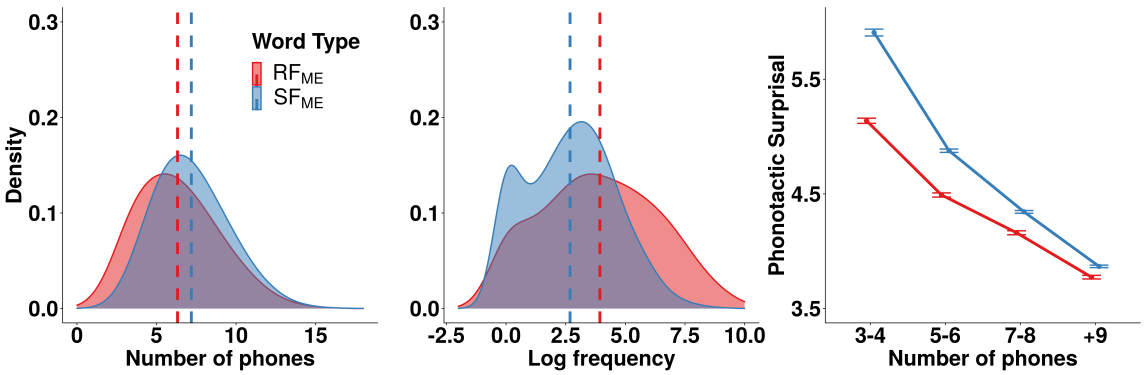
We first obtained the three vocabulary sets from the OED, by setting the time T to be 1900 and t to be 100 (i.e., a century of English words). The first set, the RF_E words, consists of word forms that were already associated with meanings before 1900 but gained at least one new sense in or after 1900. Examples of RF_E words include *plane* which became a flying machine in 1908, *alien* which first referred to extraterrestrials in 1926, and *computer* which became a calculation and storage device in 1946. The similarity between the old and new meanings is presumably responsible for the reuse of the existing word forms as described in Ramiro et al. (2018) and implemented in the PSM model. The second set, the SF_E words, are forms that existed prior to 1900 but did not gain any additional senses. Examples of SF_E words include *the*, a determiner since the earliest attested English, *spence*, either a 'pantry' or 'steward' with no new meanings since the 14th century, and

mauve which received its final definition 'purple aniline dye' in 1859. The last set, the NF_E words, consists of completely new word forms whose first attestations occurred after 1900. Example NF_E words include *riboflavin*, a B vitamin first isolated in 1920, *Malawian*, first used to refer to the people of then-Nyasaland in 1963, and *pulsar*, a class of celestial object discovered in 1968. NF_E words were adapted from other languages or coined *de novo* rather than assigned the form of a similar existing meaning. Because we needed to compute phonotactic probabilities of words, the sets extracted from the OED were additionally intersected with CELEX to obtain the phonological transcription and frequency of each form. This resulted in 1,086 NF_E words, 11,249 RF_E words, and 17,811 SF_E words.

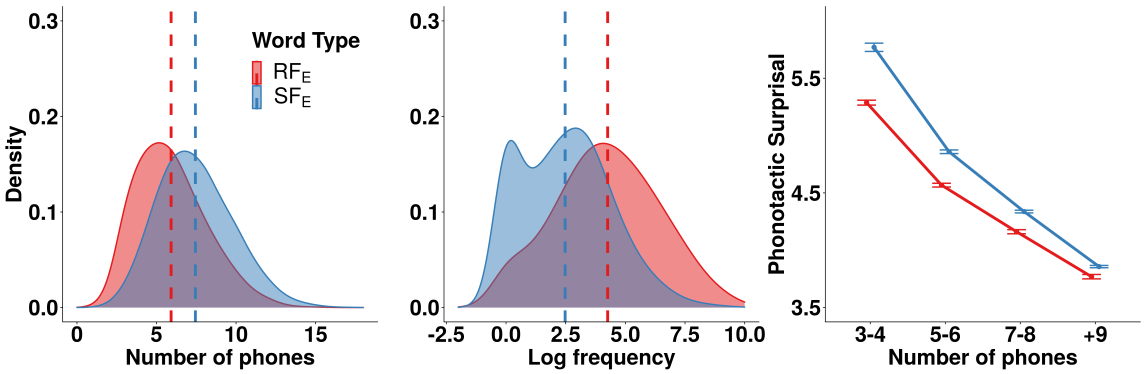
We then proceeded to construct the Monkey English lexicon. The starting point in this alternate universe is also the year 1900. We first provided the PSM model with the English lexicon at this time (i.e., the union of RF_E and SF_E words defined above, 29,060 in total). Because the PSM model makes use of a randomly generated semantic space, we started by populating the semantic space and running the PSM model until it created 29,060 word forms using the semantic distance parameter θ set to 0.1 as discussed in Section 5.2. These word forms, having been randomly generated, would not bear strong resemblance to the actual 29,060 English words forms available in 1900, so we simply replaced the PSM word forms with the actual English word forms while holding word frequency ranks constant. At this point, then, the monkeys would have exactly the same set of word forms and frequencies as the English speakers in 1900. These 29,060 words were then used by the PSM model to create new words in the post-1900 period. As before, new meanings were randomly generated in the semantic space. If a new meaning fell within the neighborhood of an existing meaning, again as controlled by θ at 0.1, then it was assigned the corresponding word form, thereby becoming a reused form (RF_{ME}) and a polysemous word in Monkey English. If the new meaning was far enough from any of the existing meanings, the PSM model proceeded to generate another word form. If this form already existed, then it also became a reused form (RF_{ME}) and a homophonous word was created in Monkey English; otherwise it became a brand new word form (NF_{ME}) in Monkey English.

As time went on, Monkey English grows three vocabulary sets: new forms NF_{ME}, reused forms RF_{ME}, and stale forms SF_{ME}. Because these sets do not grow exactly at the pace of actual English words (NF_E, RF_E, and SF_E), we performed several experiments with different stopping conditions. First, we stopped when Monkey English had reused the same number of words as in actual English (i.e., 11,239, thus $|RF_{ME}| = |RF_E|$) and took all the new forms generated up to that point as NF_{ME}. Second, we stopped until Monkey English had created the same number of new words (i.e., 1,086, thus $NF_{ME}| = |NF_E|$) and took all the reused forms generated until then as RF_{ME}. Finally, we generated post-1900 Monkey English for a sufficiently long time and took the first 1,086 new forms and the first 10,430 reused forms (thus $|RF_{ME}| = |RF_E|$ and $|NF_{ME}| = |NF_E|$). In all experiments, the set of stale forms SF_{ME} is simply the portion of the 1900 lexicon that the monkey model started with but did not get reused, as $RF_{ME} \cup SF_{ME}$ = $RF_E \cup SF_E$. All three experiments reached the same conclusions. For simplicity, we only presented the results from the last experiment with the matching number of RF and NF words between actual and Monkey English.

The communicative efficiency hypothesis makes two predictions. The first concerns a comparison between the RF and SF words: the former should be more efficient than the latter because they became reused while gaining (potentially additional) lexical ambiguity. The second concerns the NF words. Because these words did not exist prior to 1900 but were created anew, they obviously are not a reuse of existing words but should have still made use of the more efficient phonotactic



(a) Monkey English



(b) Actual English

Figure 8: Monkey-English (8a) shows the same statistically significant trends observed in the actual reuse of English forms post-1900 (8b). Reused forms are shorter, more frequent, and phonotactically more probable than forms which gained no additional meanings, independent of functional considerations.

structures available in the language. For instance, a more efficient triphone sequence should be more likely to reappear in brand new words than a less efficient sequence. Following PTG, we measure efficiency with length, frequency, phonotactic surprisal: shorter, more frequent, and more phonotactically probable words (or phone sequences) are more efficient and more likely to be reused.

We first compared RF and SF under post-1900 English and Monkey English. Note the union of these words, in both actual English and Monkey English, are exactly the same set of English words available in 1900 though they were partitioned differently due to reuse. Figure 8 shows the distribution of word length and word frequency for RF and SF words, along with the surprisal of RF and SF words normalized for word length. It is clear that for both actual English and Monkey English, the RF's are more efficient than the SF's on both efficiency measures. We confirmed this statistically by conducting a linear regression to predict each word's phonotactic surprisal from "reuse status" (RF vs. SF) along with length plus an interaction term. SF-status positively correlates with phonotactic surprisal in both Monkey English ($\beta = 0.230, t = 21.875, p < .001$) and actual English ($\beta = 0.210, t = 19.137, p < .001$). In a regression using the same independent variables, SF-status negatively correlates with log-frequency in both Monkey English ($\beta = -0.989, t = -40.15, p < .001$) and actual English ($\beta = -1.384, t = -54.93, p < .001$). The significance of the length difference for RF's is similarly confirmed via a *t-test* for both Monkey English ($t = -29.851, df = 22452, p < .001$) and actual English ($t = -56.223, df = 24951, p < .001$).

We then turned to the comparison of NFs, the newly created English and Monkey English words since 1900. Monkey English made better use of the 1900 lexical resources in the creation of new word forms (Figure 9). The NF_{ME} words (mean of 5.86 phones) are statistically significantly shorter than the NF_{E} words (mean of 7.23 phones) — *t-test*: $t = 11.261, df = 2164.7, p < .001$ — and they are also phonotactically more probable (NF_{E} mean of 5.42 compared with NF_{ME} mean of 5.03). We evaluated this latter point both overall via a *t-test* ($t = 5.986, df = 2170, p < .001$) as well as in a linear regression which controls for the correlation between length and normalized phonotactic surprisal ("Monkey-Status vs. Actual-Status" negatively correlates with phonotactic surprisal: $\beta = -0.599, t = -18.436, p < .001$). Moreover, NF_{ME} (mean of 5.86 phones) are actually shorter than pre-1900 forms (mean of 6.85 phones): via *t-test* $t = -11.568, df = 1145.4, p < .001$. That is, the monkey is correcting for the inefficiencies in the English lexicon by exploiting the space of phonotactically possible short words that was unoccupied in 1900 and afterwards.

6 Discussion

6.1 Summary

Attributing the original insight to Zipf's Principle of Least Effort, the authors of the PTG study claim that language design reflects communicative efficiency to reduce the articulatory effort on part of the speaker: lexical ambiguity (polysemy and homophony) tends to reside in "easier" — shorter, more frequent, and more probable — words. Some reservations about their assumptions about communicative efficiency notwithstanding (e.g., that disambiguation on part of the listener is effort-free, that phonotactically more probable words are easier to pronounce; Section 3.2), we proceeded to replicate the PTG study. Following Miller's monkey argument, we constructed a series of random generation lexicon models (PM) that incorporate phonotactic constraints on the structure of words. In Section 4.2, we report that the statistical distributions of words in English, Dutch, and German, which were conjectured to support the role of communicative efficiency, are

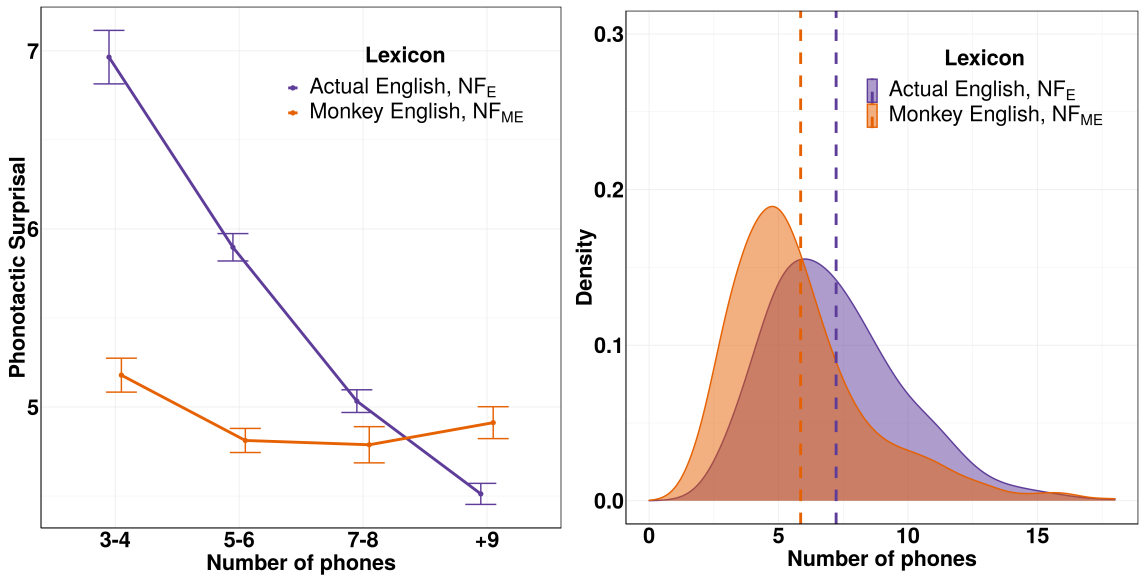


Figure 9: Newly created Monkey-English words are both shorter (right column) and more phonotactically probable (left column) than actual new forms introduced to English post-1900.

in fact consistent with the PM model that assigns word form-meaning pairings without functional considerations. Although we report only the results from the triphone phonotactic model, we had never failed to obtain the appearance of communicative efficiency under numerous other conditions, including alternative phonotactic models (e.g., uni- and bi-phone). In Section 4.3, we discuss a technical design issue in the PTG study. Once corrected to the standard method in natural language processing and psycholinguistics, lexical ambiguity is found not to correlate with phonotactic probability: only the length and frequency effects hold. However, the monkey model consistently distributes lexical ambiguity to more efficient words in all three respects.

That the PM lexicons appear more efficient than actual lexicons echoes Mandelbrot and Miller's observation that Zipf's Principle of Least Effort is better embodied by random generation processes. Partly motivated by this result, and partly driven to more accurately model the process of word emergence, we amended the PM model with a semantic component that closely follows the assignment of word senses to word forms uncovered by recent research (Ramiro et al., 2018). The resulting PSM model, described in Section 5.1, appears to track closely the growth of English words as recorded in the OED (Section 5.2). The historical data from the OED provided an additional and especially revealing test of the communicative efficiency hypothesis against the monkey baseline. In Section 5.3, we compared the outcome of English vocabulary growth since 1900 against a monkey model that started with the same lexical resources, i.e. the 1900 English lexicon. On a variety of measures, the post-1900 Monkey English lexicon proves more efficient than its counterpart in the real world.

It is worth emphasizing that we do not claim the monkey model to be *optimally efficient*. It clearly is not. For instance, in the post-1900 Monkey English experiment, the PSM model produced new word forms to express new meanings. As discussed in Section 5.3, these monkey words are shorter than their English counterpart but they did not exhaust the space of shorter words before moving on to longer words: words such as "fub," "trallist," and "graw," etc., all of which

conform to the English phonotactic constraints, were never created even though other longer forms were. However, we do conclude that in order to make a claim of communicative efficiency, *some* benchmark of efficiency is necessary: at the minimum, one needs to show the human language lexicon to be more efficient than one, such as the PM/PSM model, that does not at all take communicative efficiency into consideration. On this ground, the PTG study does not deliver.

6.2 Future research

Our study can be extended in several directions. The first concerns lexical-statistical research of natural language. The current study used three closely related languages, which have similar phonological and morphological structures and in fact share a good number of cognates. Future quantitative studies of the lexicon should focus on languages that are more representative of the linguistic diversity across the world. On the other hand, fine-grained data from diachronic sources such as our OED study can shed light on the word creation process as it unfolded. It may be especially useful, for example, to study narrower families of related languages whose historical relations are well understood. For instance, the vocabulary of Latin and the lexical divergence that occurred in its descendant Romance languages, all of which are well documented, may provide a unique opportunity for fine-grained analysis of lexicon formation. Finally, existing lexical corpora place a severe limitation on the quantitative study of word meanings. Because the frequency of a word is collected over text corpora, there is no way to distinguish the individual frequencies of the senses associated with a single (orthographic) form. Automatic word sense disambiguation technology has not reached a satisfactory level of precision although recent distributional approaches to meaning such as (contextual) embedding (e.g., Devlin, Chang, Lee, & Toutanova, 2018; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) hold promise for progress especially for languages with an abundance of textual resources.

The second direction concerns computational models of the lexicon. To the best of our knowledge, the present work is the first computational model to integrate well-motivated phonological and semantic constraints for word formation. A major area of improvement is to incorporate a component of morphology, which obviously recycles lexical resources in word formation. The PSM model at the present time can only approximate the effect of morphology with fixed-length phoneme sequences; it may be augmented with state-of-the-art unsupervised learning systems (H. Xu, Kodner, Marcus, & Yang, 2020) to automatically detect, and thus reuse, morphological units. Another pressing task in future work should focus on how lexicons are *actually* formed by language users. For instance, the critical parameter value (θ) that controls the assimilation of new word senses into existing word forms can be empirically investigated with human subjects once the semantic relatedness of words can be accurately quantified and represented in a graph-theoretic fashion. Here again developments in distributional approaches to meanings, which can be plausibly applied to historical corpora (e.g., Hamilton, Leskovec, & Jurafsky, 2016), may provide new avenues of investigation.

The third and final direction concerns the integration of macro-level quantitative lexical research with micro-level empirical studies of language variation, use, and change. As discussed in Section 1, case studies in sociolinguistics provide details of linguistic and social forces that form and mold words as they take place. This is especially relevant when the phonological system in language changes (e.g., the addition and loss of phonemes). The most recent comprehensive survey (Eckert & Labov, 2017) has largely upheld the neogrammarian mechanical view of language change (Labov, 1994), with a minor role attributed to social and cultural factors. This has direct implications of the

functionalist view of the lexicon as sound change inevitably affects the phonotactic structure of the language and the degree of ambiguity in the lexicon. Similarly, we can turn to empirical studies of how new words spontaneously emerge, the conditions under which they are conventionalized (or fall out of usage) in speech communities, and how they are transmitted through language acquisition when lexical ambiguity arises. Such studies need to draw upon the established historical records of languages and the social situations in which words take hold; see Richie et al. (2014) for a study of word conventionalization in Nicaraguan Sign Language with a focus on the role of social networks.

6.3 Functionalism in Language

There is little doubt that language is subject to constraints imposed by the cognitive and perceptual systems it is embedded in and must interact with. On our reading of the history of generative linguistics, such "functionalist" considerations were well recognized and pursued. For all their well-known skepticism toward specific functionalist hypotheses (highlighted by PTG), Chomsky and Miller, both collaboratively and individually, were in fact among its early proponents. For instance, in a foundational contribution, Miller and Chomsky (1963) proposed that the rules in the formal grammar of language may be shaped by the perceptual system (e.g., parsing mechanisms and memory limitations). This theme was revisited in the theoretical literature. Chomsky and Lasnik (1977, pg. 434-438) considered how a parsing strategy that seeks to discharge unresolved syntactic dependency immediately may be the underlying motivation for certain syntactic phenomena typically stated as formal constraints on the grammar; see (Hofmeister & Sag, 2010; Sprouse, Wagers, & Phillips, 2012) for contemporary discussion.

Even though the specific claims of the PTG study are not supported, the results reported in this paper do not necessarily rule out other types of ambiguity or the functionalist approach in general. A convincing argument of any persuasion, however, must go beyond correlational study: the space of models compatible with the observed statistical patterns in the lexicon, as we have demonstrated throughout, is too large to uniquely support any specific hypothesis. To do so requires one to provide precisely formulated and empirically motivated mechanisms of how communicative efficiency does, or does not, shape language. The PM/PSM models take a useful step in the direction: they present a set of specific mechanisms for lexicon formation even though much work remains to be done for further validation. At the very minimum, the updated monkey models provide a well-defined benchmark to compare against. As G. A. Miller and Chomsky (1963) note in their discussion of Zipf's observation of word frequency, the monkey "has something of the status of a null hypothesis, and, like many null hypotheses, it is often more interesting to reject than to accept. (pg. 463)". And that's our invitation.

Acknowledgments

We are grateful to Steven Piantadosi for sharing the source code of their study and for helpful clarifications about the implementation details. We thank Barbara Malt, the associate editor of this journal, and three anonymous reviewers for their useful comments which have improved this work. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing Interests

The authors have no competing interests to declare.

References

Albright, A. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology*, *26*(1), 9–41. doi: 10.1017/s0952675709001705

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The celex lexical database (release 2). *Distributed by the Linguistic Data Consortium, University of Pennsylvania*.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of verbal learning and verbal behavior*, *14*(6), 575–589.

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, *44*(4), 568–591.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... Treiman, R. (2007). The english lexicon project. *Behavior research methods*, *39*(3), 445–459.

Baroni, M. (2005). 39 distributions in text. *Corpus Linguistics: An International Handbook Volume*, *2*, 803–822.

Berwick, R. C., & Chomsky, N. (2016). *Why only us: Language and evolution*. Cambridge, MA: MIT Press.

Boland, J. E., & Blodgett, A. (2001). Understanding the constraints on syntactic generation: Lexical bias and discourse congruency effects on eye movements. *Journal of Memory and Language*, *45*(3), 391–411.

Brent, M. R. (1996). Advances in the computational study of language acquisition. *Cognition*, *61*(1-2), 1–38.

Calabrese, A., & Wetzels, L. (2009). *Loan phonology*. John Benjamins Publishing Company.

Campbell, L. (2013). *Historical linguistics*. Edinburgh University Press.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, *87*(2), B69–B77.

Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, *13*(4), 359–394.

Chomsky, N. (1958). [Review of Belevitch 1956]. *Language*, *34*(1), 99-105.

Chomsky, N. (1965). Aspects of the theory of. *Syntax*, 16–75.

Chomsky, N., & Lasnik, H. (1977). Filters and control. *Linguistic Inquiry*, *8*(3), 425-504.

Coleman, J., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. *arXiv preprint cmp-lg/9707017*.

Conrad, B., & Mitzenmacher, M. (2004). Power laws for monkeys typing randomly: the case of unequal probabilities. *IEEE Transactions on information theory*, *50*(7), 1403–1414.

Daland, R., Hayes, B., White, J., Garellek, M., Davis, A., & Norrmann, I. (2011). Explaining sonority projection effects. *Phonology*, *28*(2), 197–234.

Darwin, C. (1888). *The descent of man and selection in relation to sex* (Vol. 1). Murray.

Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi, S. T. (2017). Words cluster phonetically beyond phonotactic regularities. *Cognition*, *163*, 128–145.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eckert, P., & Labov, W. (2017). Phonetics, phonology and social meaning. *Journal of sociolinguistics*, *21*(4), 467–496.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Memory and Language*, *12*(6), 627.

Futrell, R., Albright, A., Graff, P., & O'Donnell, T. J. (2017). A generative model of phonotactics. *Transactions of the Association for Computational Linguistics*, *5*, 73–86.

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in cognitive sciences*.

Guy, G. R. (1991). Contextual conditioning in variable lexical phonology. *Language variation and change*, *3*(2), 223–239.

Halle, M. (1978). Knowledge unlearned and untaught: What speakers know about the sounds of their language. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic theory and psychological reality* (p. 294-303). Cambridge, MA: MIT Press.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th annual meeting of the association for computational linguistics* (pp. 1489–1501). Berlin, Germany: Association for Computational Linguistics.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, *39*(3), 379–440.

Hofmeister, P., & Sag, I. A. (2010). Cognitive constraints and island effects. *Language*, *86*(2), 366–415. doi: 10.1353/lan.0.0223

Howes, D. (1968). Zipf's law and miller's random-monkey model. *The American Journal of Psychology*, *81*(2), 269–272.

Hyman, L. (1970). The role of borrowing in the justification of phonological grammars. *Studies in African linguistics*, *1*(1), 1.

Jelinek, F. (1997). *Statistical methods for speech recognition*. MIT press.

Jurafsky, D., & Martin, J. H. (2009). Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing. *Upper Saddle River, NJ: Prentice Hall*.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*(5), 630–645.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45–52.

Labov, W. (1994). Principles of linguistic change. vol. 1: Internal features. *Language in Society). Oxford: Blackwell*.

Labov, W. (2011). *Principles of linguistic change: Cognitive and cultural factors*. John Wiley & Sons.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Li, W. (1992). Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Transactions on information theory*, *38*(6), 1842–1845.

Luce, P. A., & Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes*, *16*(5-6), 565–581.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and hearing*, *19*(1), 1.

Mahowald, K., Dautriche, I., Gibson, E., & Piantadosi, S. T. (2018). Word forms are structured for

efficient use. *Cognitive science*, *42*(8), 3116–3134.
Mandelbrot, B. (1954). Structure formelle des textes et communication. *Word*, *10*(1), 1-27.
Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive psychology*, *38*(4), 465–494.
McCarthy, J. J., & Prince, A. (1995). Faithfulness and reduplicative identity. *Linguistics Department Faculty Publication Series*, 10.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).
Miller, G. (1957). Some effects of intermittent silence. *The American journal of psychology*, *70*(2), 311–314.
Miller, G. (1998). *Wordnet: An electronic lexical database*. MIT press.
Miller, G. A. (1954). Communication. *Annual Review of Psychology*, *5*, 401–420.
Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology. Volume II* (p. 419-491). New York: Wiley.
Mirman, D., & Magnuson, J. S. (2008). Attractor dynamics and semantic neighborhood density: processing is slowed by near neighbors and speeded by distant neighbors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 65.
Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, *111*(3), 721.
Norris, D., McQueen, J. M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology*, *34*(3), 191–243.
Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, *21*(5), 1112–1130.
Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.
Ramiro, C., Srinivasan, M., Malt, B. C., & Xu, Y. (2018). Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences*, *115*(10), 2323–2328.
Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, *124*(3), 372.
Richie, R., Yang, C., & Coppola, M. (2014). Modeling the emergence of lexicons in homesign systems. *Topics in cognitive science*, *6*(1), 183–195.
Rodd, J. M., Berriman, R., Landau, M., Lee, T., Ho, C., Gaskell, M. G., & Davis, M. H. (2012). Learning new meanings for old words: Effects of semantic relatedness. *Memory & Cognition*, *40*(7), 1095–1108.
Rubenstein, H., Lewis, S. S., & Rubenstein, M. A. (1971). Evidence for phonemic recoding in visual word recognition. *Journal of verbal learning and verbal behavior*, *10*(6), 645–657.
Sevald, C. A., & Dell, G. S. (1994). The sequential cuing effect in speech production. *Cognition*, *53*(2), 91–127.
Sprouse, J., Wagers, M., & Phillips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. *Language*, 82–123.
Van Orden, G. C. (1987). A rows is a rose: Spelling, sound, and reading. *Memory & cognition*, *15*(3), 181–198.
Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*(3), 374–408.

Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in english. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 481–487.

Whaley, C. P. (1978, apr). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior*, *17*(2), 143–154. doi: 10.1016/s0022-5371(78)90110-x

Xu, A., Ramiro, C., & Xu, Y. (2019). A predictability-distinctiveness trade-off in the historical emergence of word forms. In *Proceedings of the 41st annual meeting of the cognitive science society.*

Xu, H., Kodner, J., Marcus, M., & Yang, C. (2020). Modeling morphological typology for unsupervised learning of language morphology. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (p. (To appear)). Seattle, Washington: Association for Computational Linguistics.

Yang, C. (2013). Who's afraid of george kingsley zipf? or: Do children and chimps have language? *Significance*, *10*(6), 29–34.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*.

A Regression Results for PM/PSM-Lexicons

Table 1: Homophony vs. ease for PM-Dutch, PM-English, and PM-German

Ease Metric	PM-NL	PM-EN	PM-DE
Length	$\beta = -0.075$	$\beta = -0.203$	$\beta = -0.011$
	$t = -24.45$	$t = -24.71$	$t = -24.04$
	$p < .001$	$p < .001$	$p < .001$
Frequency	$\beta = -0.229$	$\beta = -0.361$	$\beta = -0.064$
	$t = -694.0$	$t = -476.5$	$t = -220.73$
	$p < .001$	$p < .001$	$p < .001$
Surprisal	$\beta = -0.020$	$\beta = -0.073$	$\beta = -0.006$
	$t = -6.704$	$t = -9.207$	$t = -12.28$
	$p < .001$	$p < .001$	$p < .001$

Table 2: Polysemy vs. ease for English PM-Adjectives, PM-Nouns, and PM-Verbs

Ease Metric	PM-Adjectives	PM-Nouns	PM-Verbs
Length	$\beta = -0.115$	$\beta = -0.200$	$\beta = -0.145$
	$t = -10.35$	$t = -18.28$	$t = -9.634$
	$p < .001$	$p < .001$	$p < .001$
Frequency	$\beta = -0.284$	$\beta = -0.355$	$\beta = -0.322$
	$t = -102.51$	$t = -251.4$	$t = -121.7$
	$p < .001$	$p < .001$	$p < .001$
Surprisal	$\beta = -0.077$	$\beta = -0.067$	$\beta = -0.091$
	$t = -7.107$	$t = -6.18$	$t = -6.212$
	$p < .001$	$p < .001$	$p < .001$

Table 3: Syllable Informativity vs. ease for PM-Dutch, PM-English, and PM-German

Ease Metric	PM-NL	PM-EN	PM-DE
Length	$\beta = -2.150$	$\beta = -1.851$	$\beta = -1.853$
	$t = -4.174$	$t = -13.868$	$t = -52.23$
	$p < .001$	$p < .001$	$p < .001$
Frequency	$\beta = -1.638$	$\beta = -1.283$	$\beta = -1.700$
	$t = -3073.2$	$t = -1362.2$	$t = -2737.6$
	$p < .001$	$p < .001$	$p < .001$
Surprisal	$\beta = -1.567$	$\beta = -0.565$	$\beta = -1.453$
	$t = -3.966$	$t = -7.162$	$t = -6.281$
	$p < .001$	$p < .001$	$p < .001$

Table 4: Homophony vs. ease and syllable Informativity vs. ease for PSM-English

Ease Metric	Homophony	Syllable Informativity
Length	$\beta = -0.751$	$\beta = -2.166$
	$t = -42.270$	$t = -8.611$
	$p < .001$	$p < .001$
Frequency	$\beta = -0.834$	$\beta = -1.568$
	$t = -5.047e13$	$t = -2.782e13$
	$p < .001$	$p < .001$
Surprisal	$\beta = -0.195$	$\beta = -1.050$
	$t = -11.210$	$t = -9.490$
	$p < .001$	$p < .001$