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

June 26, 2019

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" that follow the same statistical distribution. Recent quantitative analysis of human language lexicons, with special focus on the phonological and semantic ambiguities of words (Piantadosi, Tily, & Gibson, 2012), has revived Zipf's functional hypothesis. In this study, we first report our replication effort, including the identification of a spurious result in that study which undercuts the communicative efficiency hypothesis. Second, an update to Miller's thought experiment that incorporates the phonotactic structure of language shows that lexicons generated without recourse to functional considerations in fact exhibit the statistical properties of words attributed to communicative efficiency. Finally, the statistical distribution of the English words that emerged since 1900 shows that the attested process of lexicon formation is consistent with the updated monkey model but does not support the claim of communicative efficiency. We conclude by arguing for the need to go beyond correlational statistics and to seek direct evidence for the mechanisms that underly principles of language design.

Keywords: Language; Computational Modeling; Information Theory; Zipf's Law.

*Correspondence concerning this article should be addressed to 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 language 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 quantitative study of language use, variation, and change would provide more direct testing grounds for the role of communicative function. However, summarizing decades of quantitative research on language variation and change — in a contribution entitled "The overestimation of the functionalism" — Labov finds little evidence for this position (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 does find support for a functionalist interpretation of some, though by no means all, trends in language change. For example, the French feminine plural article *las* (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 diachronic studies where functionalist arguments for language are more commonplace. The 19th-century philologists such as Müller, Schleicher, and others held that 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 neogrammarian position rise to dominance in historical linguistics (Campbell, 2013): language change proceeds mechanically as it alters the structure of the grammar (e.g., a phoneme), affecting all words/units governed by that structure.

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 the current study, we revisit perhaps 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 furthered 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 case study of the functional approach (Piantadosi et al., 2012, henceforth PTG) extended Zipf's argument with specific reference to ambiguity in the lexicon. 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 on the part of the listener thereby posing only negligible cognitive cost, PTG conclude that language is shaped by the function of communicative efficiency. We term this the *communicative-lexicon* hypothesis (henceforth Comlex).

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, shows that a monkey striking keys randomly on a typewriter can produce "words" whose distribution closely matches Zipf's Law. Building on word well-formedness research from both linguistics and psychology, we update Miller's argument by presenting a model that incorporates the phonotactic structure of language. The model, dubbed Phonotactic Monkey (PM), stochastically assigns semantic meanings to words without any communicative/functional considerations, much like Miller's monkey. In Section 4, we show that our 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 the 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 turn to data that directly bears on the process of lexicon formation. By examining the words that entered in the English language after 1900 as recorded in the Oxford English Dictionary (OED), we show that the statistical distributions of these newly formed words are inconsistent with a process that favors communicative efficiency à la PTG but are consistent with the PM model. Specifically, we show that there does not seem to be pressure influencing the articulatory effort of words based on their degree of semantic ambiguity. To the best of our knowledge, the OED study is the first evaluation of the functionalist approach to the lexicon that makes use of detailed and accurate phonological and semantic data from the perspective of language change. We then 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 their Comlex hypothesis that the lexicon was shaped by the functional pressure of communicative efficiency.

2.1 Measures of Ambiguity

The PTG study used three simple and intuitive measures to quantify information and ambiguity in the lexicon. Two are the familiar notions of *homophony* and *polysemy* at the level of words. 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 (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 few words. For instance, the syllable /teɪ/ can appear in the initial position of numerous English words (*ta-ble*, *ta-king*, *tai-lor*, etc.) but the syllable /gɹʌ/ appears in only one English lemma (*gra-da-tion*). 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 may disambiguate 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 for words, 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. The intuition here 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 of the 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 those allowed, some are far 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. The triphone and similar models have been extensively used in language and speech technology (Jelinek, 1997) and more recently in computational studies of the lexicon (Daland et al., 2011); 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 design 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, and 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 may require additional empirical validation. 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). In addition, 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 (see also Labov (2011)), and does not directly bear on PTG's hypothesis regarding the real-time trade-off of articulation and comprehension.

To support the Comlex hypothesis, PTG computed the correlations between ambiguity and communicative efficiency with linear quasi-Poisson regressions. Here we state their main findings. 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. Taken as a whole, these results are consistent with the Comlex hypothesis and Zipf's original Principle of Least Effort: words that are shorter, more frequent, and easier to produce are more ambiguous than words that are longer, less frequent, and harder to produce.

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 lexical ambiguity and communicative efficiency actually shapes the formation of the lexicon in language. As such, the results are open to alternative interpretations, and 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 George Miller's classic random monkey experiment, a direct response to Zipf's original *Principle of Least Effort*. We then propose a model that mechanically assigns 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 model in relation to the Comlex hypothesis.

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 of his observation 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 Law 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 spacebar increases exponentially as the sequence gets longer. Thus shorter words will be more frequent than long words — the very fact for which Zipf's functional principle was proposed. In fact Miller showed with a simple proof that the expected distribution would match the Zipfian distributions as attested in natural language corpora; see Conrad and Mitzenmacher (2004) for the general case that the keys are struck with unequal probabilities.

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. Yet the methodological point of Miller's argument remains valid: by providing an alternative null hypothesis that makes no reference to communicative efficiency, Miller showed that the statistical distribution of words is not uniquely consistent with Zipf's functional Principle of Least Effort. Around the same time, Mandelbrot (1954) and Chomsky (1958) made similar criticisms of Zipf's Law. For example, Mandelbrot proved that the random placement of spaces in text in fact minimizes the average cost of per unit of information. Chomsky noted that if "words" are defined as strings delimited by the letter *e*, or any other letter, we can obtain an even better fit of Zipf's Law than the actual words of English.

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 is intended to capture how new words, pairs of meaning-form correspondences, are formed and conventionalized in language (e.g., Richie, Yang, & Coppola, 2014). When a new meaning *m* needs to be expressed, the model randomly picks a phonological item

[1] Full implementation and code needed to reproduce all presented analyses available open-source: `https://github.com/jkodner05/ThePhonotacticMonkey`

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 processing or 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" 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 suggest).

The effect of phonotactic constraints in lexicon formation is captured by a triphone model established 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. We use the probabilities of the triphones to generate the phonological words. That is, the next phoneme (p_i) is generated probabilistically according the transitional probability given two immediately preceding phonemes $(p_{i-2}p_{i-1})$, i.e. $P(p_{i-2}p_{i-1} \rightarrow p_i)$. For the special case of the first phoneme, the biphone transitional probability from a START symbol prefixed to the word is used. 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 END symbol because they are the cause of a spurious claim 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. As will be discussed in Section 5 with data from the Oxford English Dictionary (OED), the PM model correctly characterizes the phonotactic distribution of new words which entered English since 1900.

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. Communicative efficiency concerning word length, frequency, articulation or comprehension effort plays no role in the formation of the lexicon.

Like Miller's original proposal, the statistical properties of the PM model should approximate Zipf's Law. Since the length of words will decrease sharply due to the probabilistic generation process of the triphone model, the shorter words will be more frequent than longer words. In what follows, we use the PM model as a baseline null hypothesis to evaluate the Comlex hypothesis in the PTG study.

4 Phonotactic Monkey Gives Appearance of Communicative Efficiency

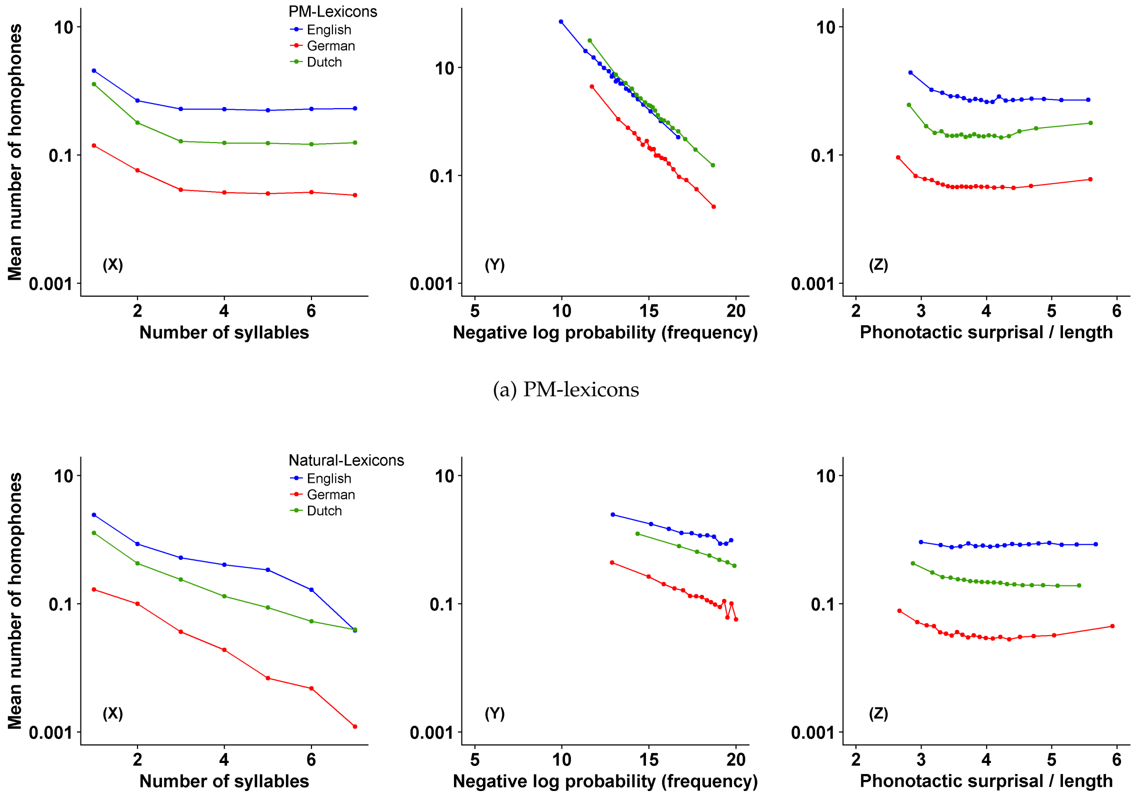
4.1 Methods

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 (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 without regard of any notion of communicative efficiency. 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 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 Comlex 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 Comlex 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 that were applied to the natural data. We summarize these case studies below while more detailed results of the



(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).

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 2 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 3 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. See also Figure 2 in Appendix B. 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 4 in Appendix A and Figure 3 in Appendix B) just like the natural lexicons. Taken together, these results suggest that the statistical distributions of words do not uniquely support the Comlex hypothesis but are compatible with a process of lexicon formation such as the PM model that makes no reference to communicative efficiency.

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, which lies 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. However, PTG's implementation divided all values by $(n+2)$, which was the length of each word plus the two special START and STOP characters that were added to the beginning and end of words for the purpose of training the triphone model.

Normalization by $(n+2)$ rather than n 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%. Table 1 compares phonotactic surprisal outputs on a sample of words under $(n+2)$ and n normalization. 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, $\beta < -0.795$, $t = -22.31$, $p < 0.001$). 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 a technical error of implementation, and cannot be accepted as a true characterization of natural language lexicons. Incidentally, as discussed in Section 4.2, the PM-lexicons did produce statistically significant correlation between phonotactic surprisal and lexical ambiguity. This recalls the classic observation that random placement of spaces (Mandelbrot, 1954) and "words" delimited by the letter *e* (Chomsky, 1958) provide a better fit of Zipf's Law than Zipf's Principle of Least Effort.

word	**transcription**	**# phones**	$n+2$ **normalization**	n **normalization**	**difference**
trustfully	trVst-fU-lI	9	2.803	3.426	22%
sweatshirt	SwEt-S3t	7	4.094	5.264	28%
hand	h5nd	4	4.325	6.487	33%

Table 1: Comparison of phonotactic surprisal values with $(n+2)$ normalization and correct n normalization on sample CELEX word forms. The error introduced is stronger for shorter words than longer words, introducing a bias into PTG's result.

5 Communicative Efficiency and Lexicon Formation: Evidence from the OED

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 and our PM model results only reflect the *static* property of language: as such, they cannot directly assess the role of communicative efficiency, or indeed any other factor, functional or otherwise, in the design of language. As remarked earlier, the quantitative study of language variation and change (Labov, 1994) provides the most direct assessment of the two approaches to language design as represented by Zipf and Miller.

Fortunately, we are now in a position to directly evaluate the predictions of the contrasting approaches thanks to the availability of language data with historical depth. In this section, we examine the statistical distributions of the new words that have entered English since 1990, as recorded in the OED, which are shown to be inconsistent with the Comlex hypothesis but consistent with the PM model.

We extracted two sets of vocabulary from the Oxford English Dictionary Online. The first set, *'New Forms'* (NF), consists of completely new word forms whose first attestations occurred in or after 1900. Examples of the NF words include *riboflavin* (a B vitamin first isolated in 1920), *Malawian* (first used to refer to the people of then-Nyasaland in 1963), or *pulsar* (a class of celestial object discovered in 1968). Thus, the NF words developed de novo and had no pre-existing meaning. By contrast, the second set, *'Old Form, New Meaning'* (OF), consists of word forms that were already associated with meanings before 1900 and recorded as such in the OED but gained at least one new sense in or since 1900. Examples of the OF words include *plane* which became a flying machine in 1908, *alien* as in an extraterrestrial (1926) and *computer* as in the calculation and storage device (1946). The sets extracted from the OED were additionally intersected with CELEX and WordNet to provide the phonological transcription and frequency of each form. This resulted in 901 NF words and 14,374 OF words.

Before comparing the predictions of Comlex and PM on the OED data, we evaluated whether the PM model accurately captures the phonotactic aspects of lexicon formation. To do so, we trained two triphone phonotactic models, one on the NF words alone, which appeared de novo since 1900, and the other one on the words that already existed prior to 1900. The first model will contain a probability for each triphone in the NF words. The second, while not trained on the NF words at all, can also assign a probability to each NF triphone based on the statistics of the pre-existing words. Because the NF and OF sets are different sizes, the models were limited to only the triphones with at least 10 occurrences in each dataset. If the PM model accurately characterizes the phonotactic properties of English lexicon formation, then the triphone probabilities under two models, which were trained on entirely disjoint sets of words, should be similar. Indeed, a paired Wilcox signed-rank test between the triphone probabilities obtained under the two models revealed no statistically significant difference ($p = 0.202$, NF median $= 0.0069$, OF median $= 0.0053$).

Because the NF and OF sets consist of words that gained meanings under unambiguous and ambiguous circumstances respectively, they can be contrasted to investigate whether communicative efficiency is implicated in the emergence of words. Consider word frequency first. Neither hypothesis makes any specific predictions about the frequencies of the NF and OF words in comparison: after all, these depend on the frequencies of meanings in usage (e.g., the usage prevalence of the NF *pulsar* and the OF *plane* in the sense of aircraft). However, both hypotheses make the following prediction with respect to the words that existed before 1900, some of which became OF words while the rest did not (call these ¬OF words). The Comlex hypothesis predicts that at the time of emergence, the OF words should have been more frequent than the ¬OF words, because lexical ambiguity is more likely to arise among the most frequent and “easier" words according to this view. The PM model makes a similar prediction. When a new meaning needs to be expressed, a word form is selected stochastically. Although communicative efficiency plays no role in this process, more frequent word forms are more likely to be chosen — thus becoming OF — than less frequent ones (i.e., ¬OF). Unfortunately this prediction is difficult to validate as it would require the historical frequencies of the OF and ¬OF words, which are not available at the present time.

With respect to word length, the Comlex hypothesis and the PM model also make the same predictions but again for different reasons. Under Comlex, NFs ought to be longer than OFs because, once again, lexical ambiguity should favor shorter words. The PM model makes a similar prediction. This is because the combinatorial space of shorter words is considerably smaller than that of longer words: thus, when a new meaning needs to find a phonological form, the PM model is more likely to generate a shorter word than a longer word — and the shorter word is more likely to have been associated with a previous meaning already. Thus, NFs are also more likely to be longer than OFs under the PM model. Indeed, this is what we found. When measured in the number of syllables, a *t*-test shows that the NF words are significantly longer than the OF words ($p < 0.001$, NF mean = 2.83, OF mean = 2.12).

With respect to phonotactic surprisal, Comlex and PM do make contrasting and testable predictions. Because the OF words were already lexically ambiguous, their forms should be more concentrated on phonotactically less surprising regions of the lexicon — which presumably makes them easier to produce — than the NF words. By contrast, PM makes no such prediction because its assignment of word meanings is completely blind to the communicative functions such as articulatory effort as characterized by its phonotactic surprisal. Here the OED data is consistent with the PM model but inconsistent with the Comlex hypothesis. We applied a Wilcoxon rank-sum test for the phonotactic surprisals between OF and NF words: there is no significant difference between the two distributions ($p = 0.7195$, NF median = 4.57, OF median = 4.46).

6 Discussion

Our results can be summarized concisely. The statistical distributions of words in English, Dutch, and German from the PTG study, which were conjectured to support the role of communicative efficiency in language design, are in fact consistent with the PM model, a stochastic model which forms sound-meaning pairs of words mechanically without any functional considerations. Furthermore, the statistical properties of recently conventionalized English words provide direct evidence against the Comlex hypothesis but in favor of the PM model.

Our studies can be extended in several directions although some foreseeable limitations should also be noted. First, the PTG study used three closely related languages, which have similar phonological and morphological structures and in fact share a good deal of words. Future research should focus on languages that are more representative of the linguistic diversity across the world. Second, diachronic considerations such as our OED study, which we believe is the first assessment of the functionalist hypothesis using detailed historical data, should be extended to additional languages. Along this vein, it may be especially useful, and convenient, 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. Third, 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 with which the word is used. Automatic word sense disambiguation technology has not reached a satisfactory level of precision although recent distributional approaches to meaning such as embedding (e.g., Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) hold promise for progress especially for languages with an abundance of textual resources.

However, the most pressing task in future work should focus on how lexicons are *actually* formed by language users. The PTG study is correlational and does not provide any proposal on how the Comlex hypothesis could be realized in a human linguistic and cognitive system. The PM model does specify a precise mechanism: while it incorporates the well-supported linguistic and psychological principle of phonotactics, direct evidence for the other components of the model would also be desirable. Thus we must turn to the 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.

We conclude by noting that even though the specific claims of Comlex in the PTG study — especially with respect to their phonotactic surprisal calculation (Section 4.3) and the OED results — are not supported, the results reported in this paper do not necessarily rule out the functionalist hypothesis in the general sense. This, we believe, is the spirit of Miller's original critique of Zipf's Law, which can be obtained "without appeal to least effort, least cost, maximal information, or any branch of the calculus of variations (pg. 314)". Judiciously chosen null hypotheses must be formulated and evaluated, a task that PTG failed in their original study.

Acknowledgments

We are grateful to Steven Piantadosi for sharing the source code of their study and for helpful clarifications about the implementation details. 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

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.

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.

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.

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.

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

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.

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

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.

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.

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.

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.

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.

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., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.

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.

Sevald, C. A., & Dell, G. S. (1994). The sequential cuing effect in speech production. *Cognition*, *53*(2), 91–127.

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.

Whaley, C. P. (1978). Word-nonword classification time. *Journal of Memory and Language*, *17*(2), 143.

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 on PM-Lexicons

Table 2: Homophony vs. ease for PM-Dutch, PM-English, and PM-German

Ease Metric	**PM-NL**	**PM-EN**	**PM-DE**
Length	$\beta = -0.075$ $t = -24.45$ $p < 0.001$	$\beta = -0.203$ $t = -24.71$ $p < 0.001$	$\beta = -0.011$ $t = -24.04$ $p < 0.001$
Frequency	$\beta = -0.229$ $t = -694.0$ $p < 0.001$	$\beta = -0.361$ $t = -476.5$ $p < 0.001$	$\beta = -0.064$ $t = -220.73$ $p < 0.001$
Surprisal	$\beta = -0.020$ $t = -6.704$ $p = 2.04e - 11$	$\beta = -0.073$ $t = -9.207$ $p < 0.001$	$\beta = -0.006$ $t = -12.28$ $p < 0.001$

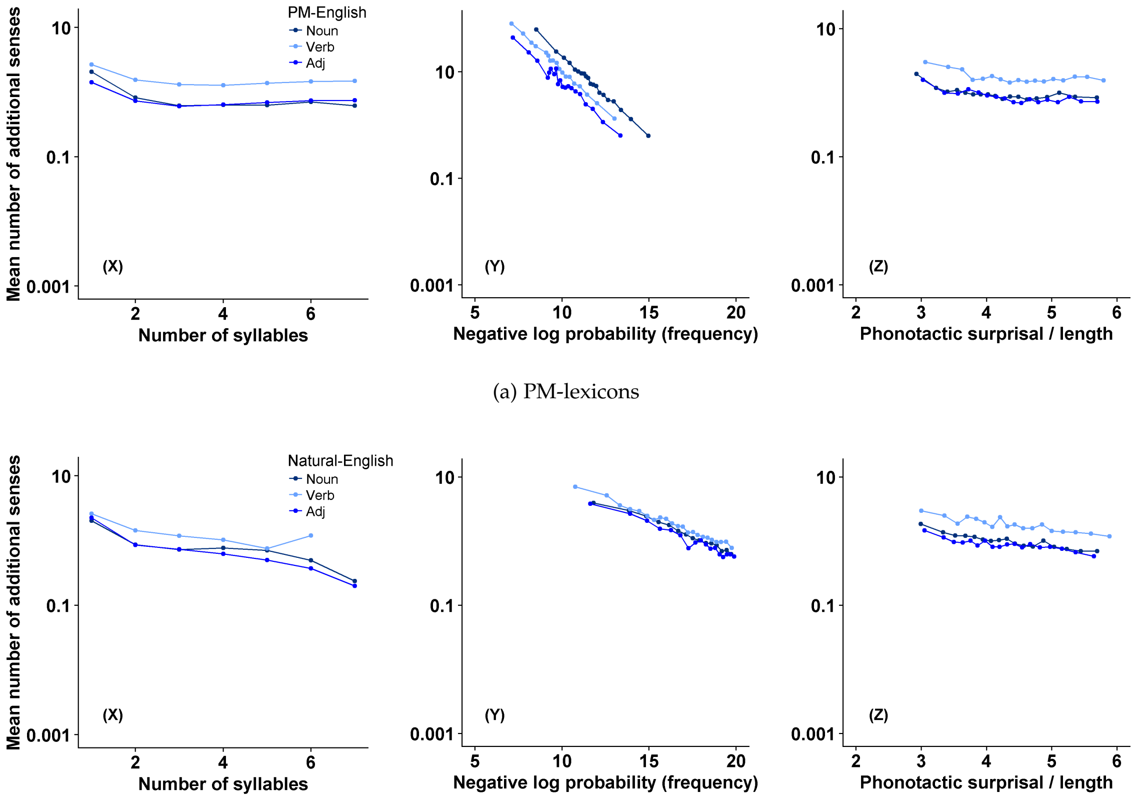
Table 3: Polysemy vs. ease for English PM-Adjectives, PM-Nouns, and PM-Verbs

Ease Metric	**PM-Adjectives**	**PM-Nouns**	**PM-Verbs**
Length	$\beta = -0.115$ $t = -10.35$ $p < 0.001$	$\beta = -0.200$ $t = -18.28$ $p < 0.001$	$\beta = -0.145$ $t = -9.634$ $p < 0.001$
Frequency	$\beta = -0.284$ $t = -102.51$ $p < 0.001$	$\beta = -0.355$ $t = -251.4$ $p < 0.001$	$\beta = -0.322$ $t = -121.7$ $p < 0.001$
Surprisal	$\beta = -0.077$ $t = -7.107$ $p < 0.001$	$\beta = -0.067$ $t = -6.18$ $p < 0.001$	$\beta = -0.091$ $t = -6.212$ $p < 0.001$

Table 4: Syllable Informativity vs. ease for PM-Dutch, PM-English, and PM-German

Ease Metric	**PM-NL**	**PM-EN**	**PM-DE**
Length	$\beta = -2.150$	$\beta = -1.851$	$\beta = -1.853$
	$t = -4.174$	$t = -13.868$	$t = -52.23$
	$p < 0.001$	$p < 0.001$	$p < 0.001$
Frequency	$\beta = -1.638$	$\beta = -1.283$	$\beta = -1.700$
	$t = -3073.2$	$t = -1362.2$	$t = -2737.6$
	$p < 0.001$	$p < 0.001$	$p < 0.001$
Surprisal	$\beta = -1.567$	$\beta = -0.565$	$\beta = -1.453$
	$t = -3.966$	$t = -7.162$	$t = -6.281$
	$p < 0.001$	$p < 0.001$	$p < 0.001$

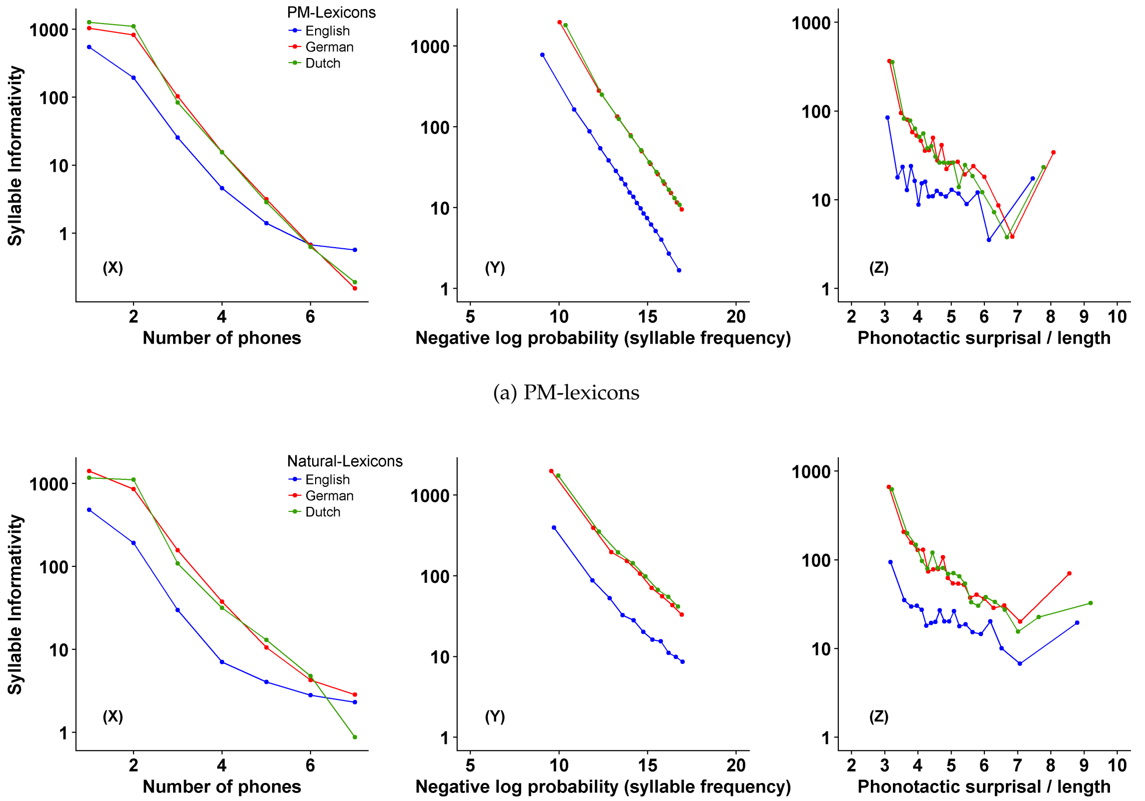
B Correlation Plots for PM and Natural-Lexicons



(a) PM-lexicons

(b) Natural-lexicons

Figure 2: Raw number of additional senses (polysemy) a word has for variable 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).