# Apparent Communicative Efficiency in the Lexicon is Emergent

**Spencer Caplan**[1,3,*], **Jordan Kodner**[1,4,5,*], **Charles Yang**[1,2]
[1] University of Pennsylvania, Department of Linguistics
[2] University of Pennsylvania, Department of Computer and Information Science
[3] Swarthmore College, Department of Computer Science
[4] Stony Brook University, Department of Linguistics
[5] Stony Brook University, Institute for Advanced Computational Science
{spcaplan,jkodner}@sas.upenn.edu
charles.yang@ling.upenn.edu

**Please cite the full version of this work published in *Cognition* (Caplan et al., 2020). What follows is a brief summary for SCiL 2021.**

## 1 Introduction

Recent quantitative analyses of human language lexicons have been interpreted as evidence that language is designed for functional efficiency. For example, Piantadosi et al. (2012; PTG) showed that ambiguous words tend to be short, frequent, and easier to articulate, a trend they interpret as evidence for communicative efficiency as the context of language use often provides cues to overcome lexical ambiguity. This line of argument is not new: Zipf (1949) famously argued that shorter words tend to be more frequent because of a pressure to minimize speaker effort, but Miller (1957) showed that a random typing process, like a monkey typing at a keyboard, could recreate these patterns, suggesting that a functional explanation is premature.

We bring Miller's approach to the modern era: can correlational results like PTG's be attributed to communicative efficiency, or are they merely the byproducts of blind, mechanical processes? Our results from two sets of computational models — Phonotactic Monkey (PM) and Phono-Semantic Monkey (PSM) — support the latter hypothesis.

## 2 Trade-Offs

PTG investigate three measures of ambiguity in the lexicon: **1)** *homophony* (how many distinct meanings a word form has: *river bank* vs. *blood bank*), **2)** *polysemy* (how many semantic variants a form has: *a run in the park* vs. *a run of wins*), and **3)** *syllable informativity* (how many words a syllable appears in: syllables that recur in many words are less informative → more ambiguous). These are balanced with three measures of production ease: **a)** *word length* (shorter words are easier to produce), **b)** *frequency* (actually, negative log probability: more frequent words are faster to access), and **c)** *phonotactic surprisal* (higher probability phonotactic sequences are easier to articulate).

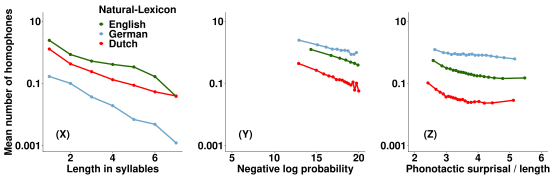


Figure 1 Natural Lexicons

PTG quantitatively investigate the trade-offs between measures (1-3) and (a-c) in CELEX (Baayen et al. 1995) English, German, and Dutch lexicons and report consistent negative correlations between each measure of ambiguity and production ease (plots for homophony vs. (a-c) reproduced in Fig 1.) They argue that such consistent effects would not arise without a pressure towards communicative efficiency.

## 3 The Phonotactic Monkey

But what PTG lack is a baseline. What would a lexicon with no effect of communication, let alone any pressure for communicative efficiency, look like? To test this, we construct pseudo-lexicons for English, Dutch, and German in the spirit of Miller's monkey. Responding to criticism that Miller's monkey does not create naturalistic word forms, we train triphone language models on CELEX's transcriptions for each language and then generate lexicons with the language models, adding a new word to the lexicon each time its STOP token was generated. The number of times the same form is generated is tabulated as its frequency, which follows an emergent Zipfian distribution. Generation is run until each PM-lexicon contains the same number of word types as the corresponding natural lexi-

[*] Denotes equal contribution



*Proceedings of the Society for Computation in Linguistics (SCiL) 2021*, pages 349-350.
Held on-line February 14-19, 2021

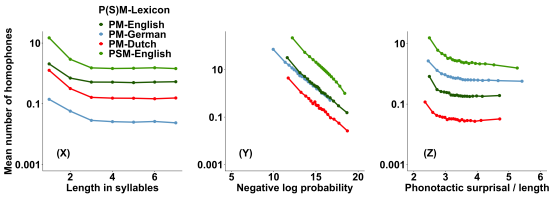


Figure 2 P(S)M-Lexicons

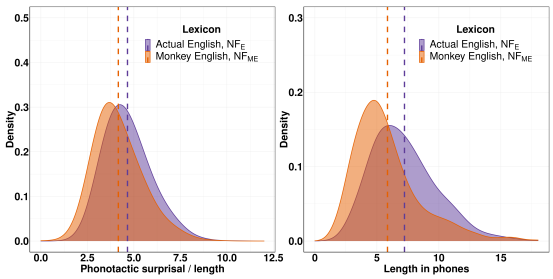


Figure 3 PSM vs Actual OED

con. Next, "meanings" are distributed randomly throughout the lexicon weighted by the token frequency for each word form until the PM-lexicon has been assigned same number of meanings as the matching natural lexicon.

What results is a PM-lexicon matched in phonotactics, number of unique word forms and meanings to its corresponding natural lexicon, but generated via a random process which makes no reference to ambiguity, production ease, or communicative efficiency. The same measures (1-3 and a-c) reported in PTG's results were calculated over the PM-generated pseudo-lexicons, and the same negative correlations were reliably observed (as for homophony in Fig 2). Under PTG's interpretation of such correlations, we would have to conclude that the PM-lexicons were shaped by communicative efficiency, but this cannot be the case. These trends are emergent. PTG's investigation of the correlation between ambiguity and production ease does not not show that languages exhibit communicative efficiency above and beyond a baseline.

## 4 PSM is *more* efficient than natural language

To more directly assess the role of communicative efficiency in the process of lexicon formation, we compared the output of a monkey-model with a more realistic semantic component (Phono-Semantic Monkey) against the historical trajectory of English over the last century (extracted from the OED). Under PSM, meanings are points in 2D space, which lets us define semantic distance and distinguish polysemy from homophony. Meanings enter the PSM-lexicon one by one; a new meaning is assigned the word form of a semantic neighbor (polysemy) if one is sufficiently close, or the word form is generated randomly through PM, which may result in either homophony or a novel form. PSM accurately tracks patterns of empirical semantic development (cf. Ramiro et al 2012) and still reproduces PTG's correlations (Fig. 2).

We then analyzed forms added to the OED by date of attestation, and divided words into several sets based on whether, or when, they gained additional meanings. Of particular interest are novel forms which gained their first sense 1900; since these New Forms were created *de novo*, any forces shaping the lexicon are in principle free to assign such meanings to the most efficient available phonotactic forms. If the development of the lexicon can be attributed to communicative efficiency, *à la* PTG, then New Forms of the last 100 years should have greater production ease than a mechanical baseline such as PSM. To test this we seeded the PSM with precisely the lexicon of English as it existed in 1900 and compared the actual record of New Forms added to the OED post-1900 to an equivalent number of PSM Monkey-English New Forms. We find that, contra the predictions of communicative efficiency, the New Forms generated by PSM are actually better — shorter and more phonotactically probable — than the empirical trajectory of English (as in Fig 3).

## 5 Conclusions

Our claim is that lexical ambiguity in human language is emergent, rather than the result of communicative pressures. A convincing argument, however, should go beyond correlational studies: the space of models compatible with the observed statistical patterns in the lexicon is too large to uniquely support any specific conclusion. To do so requires precisely formulated and empirically motivated mechanisms of how efficiency does, or does not, shape language. P(S)M provides the kernel for future mechanistic accounts of lexicon formation.

## References

Spencer Caplan, Jordan Kodner, and Charles Yang. 2020. Miller's monkey updated: Communicative efficiency and the statistics of words in natural language. *Cognition*, 205:104466.