Cognition 126 (2013) 313–318

Contents lists available at SciVerse ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Brief article

Info/information theory: Speakers choose shorter words in predictive contexts

Kyle Mahowald [a,*], Evelina Fedorenko [a], Steven T. Piantadosi [b], Edward Gibson [a]

[a] *Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, United States*
[b] *Department of Brain and Cognitive Sciences, University of Rochester, United States*

ARTICLE INFO

Article history:
Received 14 May 2012
Revised 9 September 2012
Accepted 15 September 2012
Available online 30 October 2012

Keywords:
Psycholinguistics
Information theory
Language
Efficiency
Communication

ABSTRACT

A major open question in natural language research is the role of communicative efficiency in the origin and on-line processing of language structures. Here, we use word pairs like *chimp/chimpanzee*, which differ in length but have nearly identical meanings, to investigate the communicative properties of lexical systems and the communicative pressures on language users. If language is designed to be information-theoretically optimal, then shorter words should convey less information than their longer counterparts, when controlling for meaning. Consistent with this prediction, a corpus analysis revealed that the short form of our meaning-matched pairs occurs in more predictive contexts than the longer form. Second, a behavioral study showed that language users choose the short form more often in predictive contexts, suggesting that tendencies to be information-theoretically efficient manifest in explicit behavioral choices. Our findings, which demonstrate the prominent role of communicative efficiency in the structure of the lexicon, complement and extend the results of Piantadosi, Tily, and Gibson (2011), who showed that word length is better correlated with Shannon information content than with frequency. Crucially, we show that this effect arises at least in part from active speaker choice.

1. Introduction

Zipf famously showed that word length and frequency are inversely correlated: shorter words tend to be more frequent. The reason for this relationship, according to Zipf, is to maximize efficiency by using short words—which take less effort to produce—more often than long ones (Zipf, 1935, 1949). Zipf's ideas about language and efficiency in some ways anticipated information theory (Shannon, 1948), which provides a mathematical framework for formally characterizing the efficiency of communicative systems. However, the application of such ideas to natural language waned in the second half of the twentieth century, perhaps in response to the rise of generative linguistics, which has typically eschewed efficiency-based models of language. Chomsky, for example, has repeatedly argued against communication-theoretic models of language (e.g., Chomsky, 1975).

Recently, however, there has been renewed interest in studying language as a communication system. Piantadosi et al. (2011) used insights from information theory to build upon Zipf's claim about word length. As applied to language, the information conveyed by a word can be quantified by its *surprisal*, a measure of how unpredictable a word is given its context: $-\log P(W=w|C=c)$ (Hale, 2001; Levy, 2008). This notion captures the intuition that words that are completely predictable from context ($P(W=w|C=c)=1$ convey no ($\log(1)=0$) bits of information. For instance, in the phrase "to be or not to *be*", the final *be* is almost entirely predictable from the preceding context so would have surprisal approaching 0. Conversely, words that are highly unpredictable from context will have surprisal values that tend towards infinity as the probability of the word given its context approaches 0. The final word in "to be or not to *kumquat*" would have a surprisal that, while not infinite, would be high indeed.

* Corresponding author. Address: Brain and Cognitive Sciences, MIT, 77 Massachusetts Avenue, Cambridge, MA 02139, United States. Tel.: +1 954 401 1400.

E-mail address: kylemaho@mit.edu (K. Mahowald).

0010-0277/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cognition.2012.09.010

In any communication system with variable-length communicative units (codes), it is most efficient to assign shorter codes to those elements that convey less information. In language, it is natural to take words as the code units, and study whether shorter words forms are indeed assigned to less informative meanings. Piantadosi et al. (2011) show that this prediction holds—that shorter words tend to convey less information (as measured by an idealized statistical model) than longer words. This result improves on Zipf's theory of frequency-driven word lengths by explicitly considering a word's typical predictability in linguistic context. Like an efficient variable-length code, language is organized such that low-information words—not just more frequent words—are shorter than high-information ones.

An important consequence of this organizational structure is that a lexicon in which word length is a function of information content allows speakers to approach a uniform rate of information conveyance. This *Uniform Information Density* (UID) allows speakers to maximize information conveyed without exceeding the production/perceptual channel capacity (Aylett & Turk, 2004; Frank & Jaeger, 2008; Genzel & Charniak, 2002; Jaeger, 2006, 2010; Levy & Jaeger, 2007; van Son & Pols, 2003). Much previous work on UID has shown that information density can be manipulated by factors outside lexical content: syntactic variation such as *that* omission (Jaeger, 2010; Levy & Jaeger, 2007), phonetic reduction and lengthening (Aylett & Turk, 2004; Bell et al., 2003), and contraction of phrases like *you are* to *you're* (Frank & Jaeger, 2008). A correlation between information content and word length (Piantadosi et al., 2011), however, suggests the possibility that even content words—which are typically perceived as most fundamental to meaning and thus hardest for speakers to manipulate—can be used to control information rate.

Piantadosi et al.'s corpus results alone, however, are not sufficient to draw this conclusion. Because Piantadosi et al. do not attempt to control for meaning and syntactic category, the relationship between word length and information could arise from broad differences among syntactic classes of words. For instance, the effect could be driven by function words being shorter and less informative than content words, by a large-scale difference between nouns and verbs, or by other unforeseen regularities in the corpora.

To evaluate whether the information/word length relationship holds for words of the same class, one would want to measure average information content while varying word length and holding meaning and syntactic category constant. Short/long word pairs, like *chimp*/*chimpanzee*, *math*/*mathematics*, and *exam*/*examination*, offer precisely such a controlled comparison by providing near-synonyms that vary in length.[1] If the information/word length effect holds for words of the same class, shorter words in these pairs are predicted on average to convey less information. We test this prediction in a corpus study and a behavioral experiment.

In addition to determining whether content words can be used to manipulate information rate, there is another important implication of studying the information content of short/long word pairs. A systematic difference in expected surprisal between short and long forms would serve as evidence that the information/word length relationship constitutes part of a speaker's abstract linguistic knowledge and is not solely a product of long-term linguistic evolution. In other words, the most plausible explanation for a systematic difference in surprisal between nearly synonymous noun pairs differing in length would be that speakers are sensitive to the relationship between word length and predictability and thus actively choose word-forms that conform to that relationship during on-line production. For example, they tend to choose *exam* after predictive contexts and *examination* after non-predictive contexts. The absence of such a difference, however, would suggest that Piantadosi et al.'s effect does not constitute part of an individual speaker's knowledge of language. One might instead conclude that the effect arises from differences among classes of words or because of long-term pressure for linguistic efficiency that does not extend to the level of active speaker choice.

2. Materials

Word pairs of the form *exam*/*examination* for both the corpus and behavioral study were selected by generating a list of possible candidates using a combination of CELEX (Baayen, Piepenbrock, & Gulikers, 1995), Wordnet (Fellbaum, 1998), and Marchand (1966). Word pairs were selected to ensure that the short and long form of each pair could be used interchangeably.

Because the corpus does not distinguish between different meanings of identically spelled words, pairs like *ad*/*advertisement* were used only in the behavioral experiment since *ad* is not just an abbreviation for *advertisement* but is also used in Latin expressions like *ad infinitum*, *ad nauseum*, etc. Moreover, multi-word forms, like *United States*, were included in the behavioral experiment but not in the corpus analysis due to limitations of the corpus. Thus, the corpus materials are a subset of the behavioral materials. Every effort was made to include any and all pairs that meet the criteria above.

3. Corpus study: methods and results

In the corpus study, we first used the data from Piantadosi et al. (2011) (an unsmoothed three-gram model from the Google corpus) to obtain average surprisal estimates for 22 short/long word pairs. Using the corpus, surprisal for each word w was estimated by the equation:

$$-\frac{1}{N}\sum_{i=1}^{N}\log P(W = w|C = c_i)$$

where c_i is the context of the ith occurrence of w and N is the total frequency of w. Because the three-gram model was shown to be the most reliable by Piantadosi et al.,

[1] Note that shortenings like these are distinct from the shortenings associated with massive phonological reduction (Johnson, 2004). The present shortenings represent distinct words and thus crucially differ from the phonetic and phonological reductions that have already been shown to play a role in controlling information rate (Aylett & Turk, 2004; Bell et al., 2003; Gahl & Garnsey, 2004; Pluymaekers, Ernestus, & Baayen, 2005). For instance, a careful pronunciation of the word *math* would not sound anything like *mathematics*, whereas Johnson's examples of massively reduced forms (e.g., pronouncing the word *apparently* in two syllables) would all be identical to the unreduced form when pronounced carefully.

context c_i was estimated here as the two words preceding word *w*.

Replicating Piantadosi et al., the mean surprisal for long forms (9.21) was significantly higher than that for short forms (6.90) ($P = .004$ by Wilcoxon signed rank test). Of the 22 pairs, 18 showed higher average surprisal for the long form than for its shorter counterpart. A linear regression (with unscaled variables) modeling difference in log frequency between short and long forms as a predictor for difference in surprisal revealed that the effect held even when controlling for the fact that short words tend to be more frequent than long ones. Although there was indeed an expected significant effect of difference in log frequency on difference in surprisal ($t = -4.67$, $P < .001$), an intercept of 1.45 ($t = 2.76$, $P = 0.01$) indicated that, when there was no difference in frequency between the forms, the mean surprisal of long forms was 1.45 higher than that of short forms.

Fig. 1 shows the difference in average surprisal between the long and short form of each word pair plotted against the log corpus count of the pair's short and long form combined (i.e., the frequency of the pair as a unit). For ease of reading, only the short form is listed on the plot. The key feature of the plot is that most pairs fall above the line drawn at $x = 0$.

These results demonstrate that the long form of a word carries more information on average than its shorter counterpart, and that this effect cannot be explained only by a difference in frequency between short and long forms: predictiveness of context plays an important role.

4. Behavioral study: methods and results

To test whether participants actively choose short forms in predictive contexts, we used software developed by Gibson et al. (2011) for administering surveys on Amazon's Mechanical Turk to present 58 native English speakers with forced-choice sentence completions in which they chose between the short and long form of a word pair based on which sounded more natural. The manipulation of interest was whether the context was predictive of the missing final word (supportive-context condition) or non-predictive (neutral-context condition).

Sample item:

> *Supportive context*: Susan was very bad at algebra, so she hated…
> 1. math 2. mathematics
> *Neutral context*: Susan introduced herself to me as someone who loved…
> 1. math 2. mathematics

The order of the answer choices was balanced across participants and items. Supportive and neutral contexts were matched for length. To avoid any biases from common phrases like "final exam", the key word was never presented as part of a common phrase. Comprehension questions were included to ensure that participants were engaged in the task.

To ensure that our context manipulation was effective, we presented a separate group of native English speakers

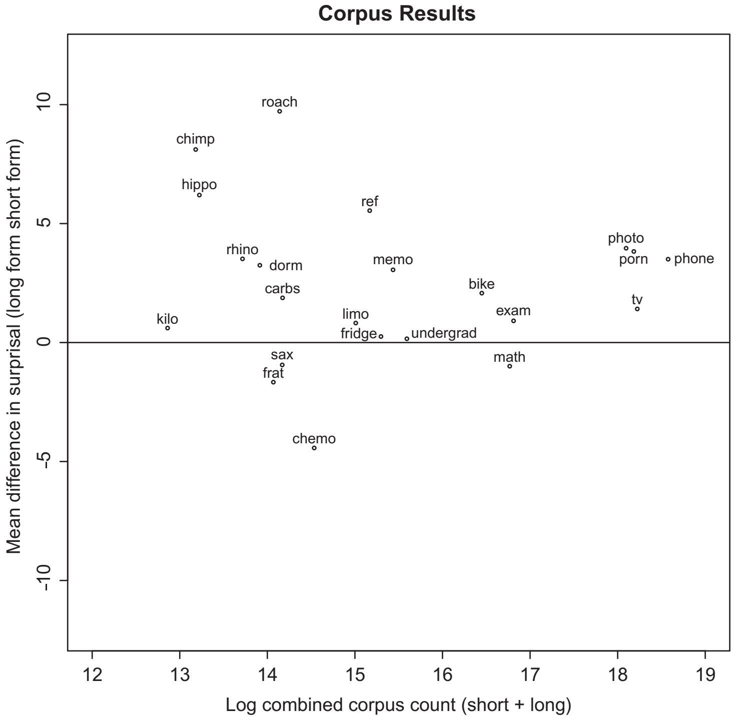

Fig. 1. Difference in mean surprisal between the long and short form (long–short) plotted against log combined corpus count of short and long. The pairs above the line at $x = 0$ show the expected effect whereby long-form surprisal is greater than short-form surprisal.

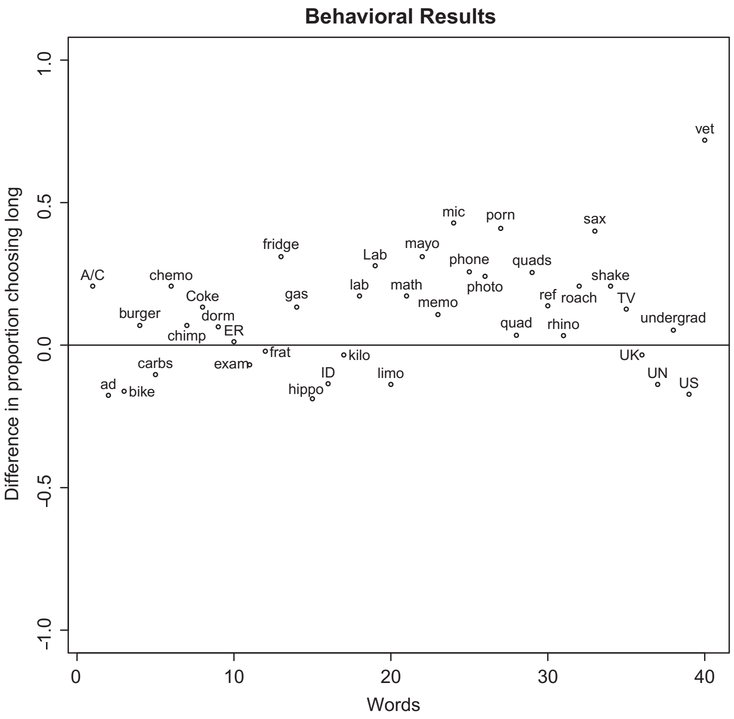

Fig. 2. The *y*-axis shows the proportion of trials where the long form was chosen in predictive contexts subtracted from the proportion of trials for which the long form was chosen in non-predictive contexts. The pairs that fall above the line at $x = 0$ show the expected effect whereby the long form was chosen more often in non-predictive contexts than in predictive contexts. The *x*-axis indicates the alphabetical order of the words.

($n = 80$) with the same sentence preambles and asked them to supply a word of their choosing to complete the sentence. For supportive contexts, the target word (either the short or long form) was chosen 52.4% of the time compared to just 1.6% of the time for neutral sentences. The effect of context on whether the word from the target pair was supplied was highly significant by a mixed-effect logistic regression with item and participant slopes and intercepts ($\beta = 5.22$, $z = 13.09$, $P < 10^{-15}$).

In the critical experiment, in supportive contexts, the short form was chosen more often (67%) than in neutral contexts (56%). This effect of context on choice of form was significant when compared to the overall mean by a mixed-effect logistic regression (Baayen, Davidson, & Bates, 2008; Gelman & Hill, 2007) with item and participant slopes and intercepts ($\beta = .75$, $z = 3.65$, $P < .001$). There was also a significant baseline preference for the short form independent of context ($\beta = .77$, $z = 2.76$, $P < .01$). The context dependence of choice of form suggests that the correlation between word length and informativeness is likely influenced by language production phenomena, where users actively prefer to convey meanings with short forms when the meanings are contextually predictable, even when controlling for syntactic category and meaning.

Fig. 2 shows the proportion of trials for which the long form of a word was selected in supportive contexts subtracted from the proportion of trials for which the long form was chosen in neutral contexts. As expected, the words tend to cluster above 0, which indicates that the long form of a word is chosen more often in neutral contexts.

We conclude that speakers actively select shorter word forms in more predictive contexts and longer forms in less predictive contexts.

5. Discussion

We have shown that the information/word length effect holds for pairs of near synonyms that vary in length. The corpus analysis revealed that short forms in almost all instances have significantly lower information content than long forms. The word-choice experiment further showed that speakers are more likely to choose a short form in a supportive context than in a neutral context. We therefore conclude that, just as phonetic and syntactic factors can be used to manipulate information rate, so too can content words.

By looking at the same phenomenon (word shortening) in both a quantitative corpus analysis and a behavioral experiment with human participants, we have untied two disparate approaches to measuring linguistic context. Our corpus study relied on a precise notion of context: two words immediately preceding the target word. Although this measure of information is a rather impoverished substitute for the rich real-world context that determines a word's surprisal, it has the advantage of being precise and easily quantified. Moreover, n-gram surprisal measures have repeatedly been shown to perform well in predicting many real-world linguistic phenomena like syllable duration (Aylett & Turk, 2006), phonological reduction (Jurafsky, Bell, Gregory, & Raymond, 2000), and word length (Piantadosi et al., 2011). N-grams also serve as an adequate surrogate for real-world context in a wide array of natural language engineering tasks (Jurafsky & Martin, 2009).

That said, n-gram measures of surprisal ignore the semantic and pragmatic cues that play a role in the interpretation of real-world linguistic context. Our behavioral study used an alternative notion of linguistic context in which a context *c* is deemed either "neutral" or "supportive" for a given word *w* based on whether speakers routinely supply word *w* after context *c*. Thus, for the behavioral study, whether or not a context was supportive was usually a product of salient semantic and pragmatic factors. The fact that our results generalized across both the precise n-gram notion of context as well as the fuzzier but more intuitive notion of context in the behavioral study is evidence that the relationship between word length and information content is not merely an artifact of using simplified n-gram-based measures of surprisal.

Moreover, these results suggest that considerations of word length and predictability form part of a speaker's knowledge of language. An important outstanding question concerns the level of abstraction at which this knowledge exists. It is possible that the corpus results presented here arise from learned preferences for each specific shortened form in predictive contexts and for each longer form in non-predictive contexts. Indeed, previous work (Fedzechkina, Jaeger & Newport, 2011) has suggested that acquisition of communicatively optimal forms is easier than acquisition of less efficient forms. One way that this process could work would be for the short form to be preferentially learned in certain highly predictive contexts (*final exam*, *football ref*, etc.). But the behavioral experiment rules out this scenario *in general* since the sentences avoided these types of fixed phrases and presented participants with what were likely novel contexts. Thus, the behavioral results suggest that the corpus results (both the ones presented here and the ones reported by Piantadosi et al. (2011)) arise because of an abstract association between word length and information. Such association is most plausibly due to speakers preferentially varying word length on-line during production. These results thus add to a growing body of work showing that speakers actively control information rate.

Additionally, our behavioral findings provide evidence against recent claims by Ferrer i Cancho and Moscoso del Prado Martín (2011), who argue that the observed correlation between word length and information content in lexical systems may not be meaningful, because such a correlation would also be found in random typing, which is not communicative. First, it should be noted that their observation cannot explain the primary finding of Piantadosi et al. (2011), that word length is better predicted from information content than from frequency since those two measures coincide for random typing models. Second, the present behavioral findings highlight the meaningfulness of Piantadosi et al.'s (2011) results. In particular, speakers' tendency to vary their word choice based on in-context predictability is not explainable by random typing. Instead, the system behaves as UID predicts: speakers appear to have internalized predictive models of language and thus vary word length, at least coarsely, to keep the number of bits of information communicated per unit time relatively constant (Aylett & Turk, 2004; Frank & Jaeger, 2008; Genzel & Charniak, 2002; Jaeger, 2006, 2010; Levy & Jaeger, 2007; van Son & Pols, 2003). This predictive model fits with Piantadosi et al.'s explanation for word length patterns and is not compatible with Ferrer i Cancho and Moscoso del Prado Martin's baseline statistical model, which makes no predictions about how context should affect speaker choice.

This line of research may also have implications for understanding certain types of lexical change. Specifically, if a word's surprisal decreases, one should expect that word to shorten over time. This hypothesis accords with long-held intuitions about language change: "[Shortened forms] originate as terms of a special group, in the intimacy of a milieu where a hint is sufficient to indicate the whole" (Marchand, 1966). Information theory provides a way to formalize this intuition. In future work, it may be possible to model and even predict lexical change based on changes in surprisal.

More broadly, these results demonstrate the power of applying information theory to the study of language. An information-theoretic framework and model was critical in formulating the behavioral experiment, and correctly predicted its outcome. Our results therefore provide further evidence that language—and the cognitive systems that process it—result in part from pressures for efficient communicative design.

Acknowledgments

We thank Rachel Chasin and Melina Flores for contributing to the pilot stages of this project. We also thank Roger Levy, Florian Jaeger, the members of Tedlab, the CUNY 2012 audience, and three anonymous reviewers for helpful input and discussion. This work was supported by a National Science Foundation Grant from the linguistics Program 0844472 (to E.G.).

References

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech, 47*(1), 31–56.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (release 2)[cd-rom]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America, 113*, 1001.

Chomsky, Noam. (1975). *Reflections on language*. New York: Pantheon Books.

Fedzechkina, M., Jaeger, T., & Newport, E. (2011). Functional biases in language learning: Evidence from word order and case-marking interaction. In *Proceedings of the Cognitive Science Society*.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Ferrer i Cancho, R., & Moscoso del Prado Martín, F. (2011). Information content versus word length in random typing. *Journal of Statistical Mechanics: Theory and Experiment, 2011*(12), L12002.

Frank, A., & Jaeger, T. F. (2008). *Speaking rationally: Uniform information density as an optimal strategy for language production*. In *Proceedings of the Cognitive Science Society*.

Gahl, S., & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, 748–775.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models* (Vol. 648). New York: Cambridge University Press.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 199–206).

Gibson, E., Piantadosi, S. T., & Fedorenko, K. (2011). Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistic Compass, 5*, 509–524.

Hale, J. (2001). *A probabilistic earley parser as a psycholinguistic model*. Association for Computational Linguistics, pp. 1–8.

Jaeger, T. F. (2006). *Redundancy and syntactic reduction in spontaneous speech*. Stanford University.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology, 61*(1), 23–62.

Johnson, K. (2004). Massive reduction in conversational American English. In *Spontaneous speech: Data and analysis*. *Proceedings of the 1st session of the 10th international symposium* (pp. 29–54).

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. (2000). Probabilistic relations between words: Evidence from reduction in lexical production. In *Frequency in the emergence of linguistic structure*. Amsterdam: John Benjamins.

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

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Proceedings of the twentieth annual conference on neural information processing systems* (pp. 849–856). Cambridge, MA: MIT Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177.

Marchand, H. (1966). *The categories and types of present-day English word-formation: A synchronic-diachronic approach*. University of Alabama Press.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences, 108*(9), 3526.

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *The Journal of the Acoustical Society of America, 118*, 2561.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423. 623–656.

van Son, R. J. J. H., & Pols, L. C. W. (2003). How efficient is speech? *Proceedings Institute of Phonetic Sciences, University of Amsterdam, 25*, 171–184.

Zipf, G. (1935). *The psychology of language*. NY: Houghton-Mifflin.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley.