Rethinking language: How probabilities shape the words we use

Thomas L. Griffiths[1]
Department of Psychology, University of California, Berkeley, CA 94720-1650

If you think about the classes you expect to take when studying linguistics in graduate school, probability theory is unlikely to be on the list. However, recent work in linguistics and cognitive science has begun to show that probability theory, combined with the methods of computer science and statistics, is surprisingly effective in explaining aspects of how people produce and interpret sentences (1–3), how language might be learned (4–6), and how words change over time (7, 8). The paper by Piantadosi et al. (9) that appears in PNAS adds to this literature, using probabilistic models estimated from large databases to update a classic result about the length of words.

Quantitative Analysis of Language

The classic finding that Piantadosi et al. (9) revisit is Zipf's observation that the length of words is inversely related to their frequency: Words that are used often, such as "the," tend to be short (10). This was one of several results obtained through quantitative analysis of the statistics of language, of which perhaps the most famous is the power-law distribution of word frequencies (known as "Zipf's Law"). Zipf explained these regularities by appealing to a "Principle of Least Effort" (11), which is sufficiently provocative as to have made its way into Pynchon's *Gravity's Rainbow* (12). For the relationship between length and frequency, the idea is that producing longer words requires more effort, so languages should be structured to use such words infrequently. This work has been followed by detailed quantitative studies of the distributions of word frequencies and word lengths (13, 14).

Zipf's analyses were done at a time when mathematical ideas were beginning to be applied to language, including probability theory. Three decades earlier, Markov introduced the idea of modeling a sequence of random variables by assuming that each variable depends only on the preceding variable (a Markov chain) using the example of modeling sequences of letters (15). A simple probabilistic model for a sequence of letters might be to choose each letter independently, with probability proportional to its frequency in the language, like drawing a set of tiles in Scrabble. Unfortunately, as Scrabble players know all too well, putting down these tiles in sequence is unlikely to make a word in English. Imagine if you took tiles from a bag where the probabilities were determined by how often each letter followed the last letter you drew—no more nasty sequences of all vowels or all consonants! A decade after Zipf published his analyses, Shannon (16) used these Markov chains to predict sequences of words, observing that a reasonable approximation to English could be produced if each word was chosen based not just on the previous word but on the last few words, and introduced a mathematical framework for analyzing the information provided by a word. In this framework, the informativeness of a word is given by the negative logarithm of its probability, matching the intuition that less probable words carry more information.

The length of words is not just related to their frequency but to their predictability in context.

Research applying probability theory to language slowed with the rise of Chomskyan linguistics. Chomsky (17) argued convincingly against the ability of Markov chains to capture the structure of sentences. His famous sentence "Colorless green ideas sleep furiously" was constructed, in part, to illustrate that Markov chains cannot be used to determine whether a sentence is grammatical: Each pair of words in this sequence is unlikely to occur together, so its probability should be near zero even though most speakers of English would agree it is grammatical (if a little unusual). This argument against a specific probabilistic model was taken to refute more generally the relevance of probability theory to understanding language, with formal linguistics turning to a mathematical framework that had more in common with logic.

The return of probability theory to linguistics came via work on the more applied problem of making computers process human languages. Probability theory can be used to solve two kinds of problems: making predictions and making inferences. Both are relevant to processing language. If you want to do a good job of interpreting human speech, it helps to have a good model of which sequences of words you are likely to hear. Understanding sentences and learning language are both problems of inductive inference, requiring a leap that goes from the words we hear to an underlying structure, and probability theory (and particularly Bayesian inference) can be used to solve these problems. Computational linguists discovered that ideas from probability theory could improve algorithms for speech recognition (18), identifying the roles that words play in sentences (19) and inferring the structure of those sentences (20), and it is now difficult to understand most papers at a computational linguistics conference without a good education in statistics.

The Rise of Probability

Probability theory has begun to migrate from computational linguistics into other areas of language research. The problems posed by colorless green ideas can be circumvented by using more sophisticated probabilistic models than Markov chains (21), and theorists are beginning to ask whether probabilities appear in linguistic representations (22, 23). Psycholinguists have begun to examine how the predictability of a word influences its production and processing (1–3). Language learning researchers have used probability theory as the basis for theoretical arguments that language can be learned (4), as well as in experiments and models exploring the acquisition of its components (5, 6). Research on how languages change over time now has access to reconstructions of the relationships between languages (and the words themselves) produced using probability theory (7, 8). Supporting these probabilistic models is the availability of large amounts of linguistic data, through databases that are larger and easier to access than ever before.

Piantadosi et al. (9) draw on these resources to conduct a deeper analysis of the factors influencing the length of words. Their basic empirical result is a nice extension of Zipf's original observation, showing that the length of words is not just

Author contributions: T.L.G. wrote the paper.

The author declares no conflict of interest.

See companion article on page 3526 in issue 9 of volume 108.

[1]E-mail: tom_griffiths@berkeley.edu.

related to their frequency but to their predictability in context. By considering the frequency of a word, Zipf measured how predictable that word is if you know nothing else about the words you are likely to encounter. However, Markov chains can be used to compute how probable each instance of a word is based on the last few words, providing a way to measure the predictability of a word in its context. This makes it possible to calculate how much information is contributed by that word, using the metric introduced by Shannon (16). If a word is easy to predict based on context, it contributes little information. Piantadosi et al. (9) find that the average information contributed by a word is better correlated with its length than is its overall frequency, suggesting that the predictability of a word in context is what matters in determining how long that word should be.

This refinement of our understanding of the relationship between the length of a word and its probability is bolstered by a theoretical framework that adds precision to Zipf's Principle of Least Effort and connects the relationship between word length and probability to an idea that has already proven valuable in other areas of psycholinguistics. This framework is based on the "Uniform Information Density" hypothesis: the idea that human languages follow the optimal strategy for communicating information through a noisy channel, by transmitting information at a constant rate that matches the capacity of the channel (2, 24–27). A crude analogy might be to imagine communication in terms of pumping oil along a fragile pipe. If you pump too slowly, it takes too long; pumping too quickly risks breaking the pipe; and varying the rate of flow is either inefficient or dangerous. The best strategy is to pump at a constant level set by the capacity of the pipe. In the case of language, we are pumping words at one another; the time it takes to send a word along the pipe is determined by its length, and the capacity of the pipe is determined by the rate at which we can process linguistic information. The best solution is to send information at a constant rate, which means that less predictable words, those that carry more information, should be longer.

The Uniform Information Density hypothesis shares with the Principle of Least Effort the notion of optimization, making the most of a limited resource, but gives this notion a formal precision that leads to a variety of other interesting predictions. For example, including an additional unnecessary word, such as "that," in a sentence (e.g., "How big is the family that you cook for?") potentially dilutes the information density of the sentence (specifically, the information associated with the clause beginning with "you"). The information density will thus become more uniform if such words are introduced to sentences that carry more information, and people's word choices seem to follow this prediction (2). Explanations framed in terms of information density rather than least effort also make it clearer that we should imagine language as being tailored to fit human minds rather than human laziness.

Providing a formal framework connecting word length and predictability opens the door to further analyses using more sophisticated probabilistic models, and considering other statistics that might be relevant to understanding the lengths of words. There is still a great deal of variance in the length of words that is not explained by their predictability. However, the deeper message behind the results of Piantadosi et al. (9) is that probability and information theory can help us rethink the way that language works, and how it should be studied. Probabilities can augment the classic rule-based representations that are widely used in linguistics, and information theory provides a way to formalize ideas like the Principle of Least Effort in a way that leads to unique predictions about language. Conversely, perhaps judgment should be reserved until Uniform Information Density makes its own appearance in literary fiction.

1. Hale J (2001) A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics, Stroudsburg, PA), pp 159–166.
2. Levy R, Jaeger TF (2007) Speakers optimize information density through syntactic reduction. *Adv Neural Inf Process Syst*, eds Scholkopf B, Platt J, Hoffman T (MIT Press, Cambridge, MA), Vol 19, pp 849–856.
3. Padó U, Crocker MW, Keller F (2009) A probabilistic model of semantic plausibility in sentence processing. *Cogn Sci* 33:794–838.
4. Chater N, Vitányi P (2007) Ideal learning of natural language: Positive results about learning from positive evidence. *J Math Psychol* 51:135–163.
5. Saffran JR, Aslin RN, Newport EL (1996) Statistical learning by 8-month-old infants. *Science* 274:1926–1928.
6. Xu F, Tenenbaum JB (2007) Word learning as Bayesian inference. *Psychol Rev* 114:245–272.
7. Gray RD, Atkinson QD (2003) Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426:435–439.
8. Bouchard-Côté A, Griffiths TL, Klein D (2009) Improved reconstruction of protolanguage word forms. *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL09)* (Association for Computational Linguistics, Stroudsburg, PA), pp 65–73.
9. Piantadosi ST, Tily H, Gibson E (2011) Word lengths are optimized for efficient communication. *Proc Natl Acad Sci USA* 108:3526–3529.
10. Zipf G (1936) *The Psychobiology of Language* (Routledge, London).
11. Zipf G (1949) *Human Behavior and the Principle of Least Effort* (Addison-Wesley, New York).
12. Pynchon T (1973) *Gravity's Rainbow* (Viking, New York).
13. Grzybek P (2007) *Contributions to the Science of Text and Language: Word Length Studies and Related Issues* (Springer, Dordrecht, The Netherlands).
14. Baayen H (2002) *Word Frequency Distributions* (Kluwer, Dordrecht, The Netherlands).
15. Markov AA (1913) An example of statistical investigation of the text Eugene Onegin concerning the connection of samples in chains; trans. Custance G, Link D (2006) *Science in Context* 19:591–600 (Russian).
16. Shannon CE (1948) A mathematical theory of communication. *Bell System Technical Journal* 27:379–423; 623–656.
17. Chomsky N (1957) *Syntactic Structures* (Mouton, The Hague).
18. Rabiner LR (1989) A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc IEEE* 77:257–286.
19. Merialdo B (1994) Tagging English text with a probabilistic model. *Comput Linguist* 20:155–172.
20. Collins M (1996) A new statistical parser based on bigram lexical dependencies. *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Stroudsburg, PA), pp 184–191.
21. Pereira F (2000) Formal grammar and information theory: Together again? *Philos Trans R Soc London* 358:1239–1253.
22. Gahl S, Garnsey S (2004) Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language* 80:748–775.
23. Bod R, Hay J, Jannedy S (2003) *Probabilistic Linguistics* (MIT Press, Cambridge, MA).
24. Aylett M, Turk A (2004) The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Lang Speech* 47:31–56.
25. Levy R (2005) Probabilistic models of word order and syntactic discontinuity. PhD thesis (Stanford University, Palo Alto, CA).
26. Jaeger T (2006) Redundancy and syntactic reduction in spontaneous speech. PhD thesis (Stanford University, Palo Alto, CA).
27. Genzel D, Charniak E (2002) Entropy rate constancy in text. *Proceedings of the Fortieth Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Stroudsburg, PA), pp 199–206.