# Word Frequency Distributions

**R. Harald Baayen**
(University of Nijmegen)

Dordrecht: Kluwer Academic Publishers (Text, speech and language technology series, edited by Nancy Ide and Jean Véronis, volume 18), 2001, xxii+333 pp and CD-ROM; hardbound, ISBN 0-7923-7017-1, $108.00, €117.00, £74.00; paperwork, ISBN 1-4020-0927-5, $48.00, €48.00, £30.00

*Reviewed by*
*Geoffrey Sampson*
*University of Sussex*

Baayen's book must surely in the future become the standard point of departure for statistical studies of vocabulary.

Baayen begins with a puzzle that has troubled many investigators who have studied vocabulary richness, for instance, people hoping to find stylistic constants characteristic of individual authors for use in literary or forensic authorship disputes. Naïvely one imagines that the ratio of number of distinct word types in a document to number of word tokens—the "type/token ratio," or as Baayen prefers, exchanging numerator and denominator, the "mean word frequency"—might be a suitable index. It is not, because it is not independent of sample size. In most domains, sample means fluctuate randomly around population means while getting closer to them as sample sizes increase. In natural language vocabulary studies, mean word frequencies systematically increase with sample size even when samples of tens of millions of words are examined.

To make the point concrete, Baayen compares Lewis Carroll's *Alice in Wonderland* and *Alice through the Looking-Glass*. One might hypothesize that greater experience would lead a writer to use a richer vocabulary in a later book, but mean word frequency is actually higher (i.e., type/token ratio lower) in *Through the Looking-Glass* than in *Wonderland*: 10.09 to 10.00. *Through the Looking-Glass*, however, is a somewhat longer book. If just the first 26,505 words are used (this is the length of the earlier book), the direction of the difference in mean word frequencies is reversed: 9.71 to 10.00. Normally, more data give a more accurate picture (of anything); but here the direction of change in frequency, from 9.71 for 26,505 words to 10.09 for 29,028 words, is usual. Can we conclude that Carroll was using a richer vocabulary in the later book, because of the figures for equal-sized samples? Or that he was using a less rich vocabulary, because of the figures for total available samples? Or can we make no inference either way?

A number of scholars have devised formulae more complex than the simple type/token ratio in an attempt to define characteristic constants that are independent of sample size. Gustav Herdan argued in a series of works that were influential in the 1960s that the ratio of the logarithms of number of types and number of tokens was such a constant. Baayen considers "Herdan's law" and various other proposals in the literature, such as G. K. Zipf's, and shows empirically that each is mistaken: All the measures turn out to be dependent on sample size (though one proposed by Honoré [1979] appears to be less so than the others). Conversely, Baayen quotes Naranan and

Balasubrahmanyan (1998, page 38) as claiming that "a word frequency analysis of a text can reveal nothing about its characteristics." Eventually, Baayen is able to show that this negative position is also unjustified; but between that conclusion and the statement of the puzzle lie some two hundred pages of fairly dense mathematics. (This is certainly not a book for the mathematically fainthearted. Baayen does a great deal, though, to help the reader follow him through the thickets. Not only does each chapter end with a summary of its findings, but—unusually for a work that is not a student textbook—Baayen also gives lists of test questions that the diligent reader can work through to consolidate his understanding of the material.)

What lies behind the unusual relationship between type frequencies and sample sizes in the case of vocabulary? Baayen clarifies the situation by an analogy with die-throwing. Think of repeated throws of a single die as a system generating a sequence over the vocabulary "one, two, . . . , six": Baayen plots a graph showing how the expected frequency spectrum (that is, the number of vocabulary elements observed once, the number of vocabulary elements observed twice, . . .) changes as the sequence is extended. For *hapax legomena* (elements of the vocabulary observed once each), the expected figure rises to a maximum of about 2.5 (I am reading approximate figures off Baayen's plot rather than calculating exact figures for myself) at five throws, and then falls back to near zero by 40 throws. For successive elements of the spectrum, the waves are successively lower and later, but the pattern is similar: for *dis legomena* (types observed twice) the maximum is about 1.8 at about 12 throws and close to zero by about 60 throws, and so on. Meanwhile, a plot on the same graph of expected sample vocabulary size rises rapidly and is close to the population vocabulary (i.e., six) by 40 throws. In most domains to which statistical techniques are applied, sample sizes are large enough to involve areas far out to the right of this kind of graph (a serious examination of possible bias in a die would surely involve hundreds of throws), so the special features of its left-hand end are irrelevant. With natural language vocabulary studies, on the other hand, even the largest practical samples leave us in an area analogous to the extreme left-hand end of the die-throwing graph, with numbers of *hapax legomena* (and consequently also *dis legomena, tris legomena*, etc.), as well as vocabulary size, continuing to grow with increased sample size and showing no sign of leveling out.

Using a term borrowed from Khmaladze (1987), Baayen describes achievable sample sizes in vocabulary studies as falling into the "large number of rare events" (LNRE) zone of the sample-size scale. The intuitive meaning of this is fairly clear, and it is made exact through alternative formal definitions. Much of Baayen's book is about the special mathematical techniques relevant to the study of LNRE distributions. (Using these techniques, it turns out that the growth in vocabulary richness between *Alice in Wonderland* and *Alice Through the Looking-Glass*, after truncation to make their length the same, *is* marginally significant.) Not all of the exposition is original with Baayen. One of the many virtues of his book lies in drawing together in one convenient location a clear statement of relevant analyses by others over several decades, often published relatively obscurely. Baayen's chapter 3 presents three families of LNRE models, which are due respectively to J. B. Carroll (1967), H. S. Sichel (1975), and J. K. Orlov and R. Y. Chitashvili (1983a, 1983b). A point that emerges from the book (and that readers of this review may have begun to infer from names cited) is the extent to which, in the late 20th century, this mathematical approach to natural language was a scholarly specialty of the former Soviet Union; in consequence it was largely unknown in the West. There are other channels through which this work has become accessible to the English-speaking world in recent years, notably the *Journal of Quantitative Linguistics*, but that German-based journal, though published in English,

has to date attracted limited attention in Britain and North America. The book under review may well be the most significant route by which important Soviet research in our area will become known to English-speaking scholars.

It would be beyond the scope of this review to survey all the issues relating to LNRE distributions that Baayen investigates. For linguists, one particularly interesting area concerns departures from the randomness assumption made by the simpler LNRE models. These pretend, for the sake of mathematical convenience, that texts are constructed by drawing successive words blindly out of an urn containing different numbers of tokens of all possible words in the vocabulary, so that the difficulties to be addressed relate only to the vast size of the urn. Real life is not like that, of course: for instance, from the frequency of the word *the*, the urn model predicts that the sequence *the the* should occur once in every couple of pages or so of text, but in practice that sequence is hardly ever encountered.

If we are primarily interested in overall vocabulary size, one problem that is repeatedly produced by the urn model is that inferences from vocabulary size in observed samples to vocabulary sizes for other, so-far-unobserved sample sizes turn out to be overestimates when samples of the relevant size are examined. Many linguists, particularly after the above discussion of *the the*, will be professionally inclined to assume that this problem stems from ignoring syntactic constraints within sentences, as the urn model does. Baayen demonstrates that this is *not* the source of the problem. If the sentences of *Alice in Wonderland* are permuted into a random order (while preserving the sequence of words within each individual sentence), the overestimation bias disappears. Instead, the problem arises because key words (for *Alice in Wonderland*, some examples are *queen, king, turtle,* and *hatter*) are "underdispersed." Different passages of a document deal with different topics, so topic-sensitive words are not distributed evenly through the text.

The bulk of Baayen's book consists of sophisticated mathematical analysis of the kinds of issues considered in the preceding paragraphs. No doubt what Baayen gives us is not always the last word to be said on some of the questions he takes up, but (as already suggested) it is hard to think that future analyses will not treat Baayen as the standard jumping-off point for further exploration.

Baayen's final chapter (chapter 6) concerns applications, and this is arguably something of an anticlimax. It is natural to want to show that the analysis yields implications for concrete topics, but some of the topics investigated do not seem very interesting other than as illustrations of Baayen's techniques, and some of them apparently lack the LNRE quality that gives the bulk of this book its impact. For natural language–processing applications, probably the most significant topic considered is bigram frequency (Baayen's section 6.4.4), but on this the author has only a very limited amount to add to the existing literature. In terms of general human interest, there is much promise in a section that studies the statistical pattern of references in recent newspapers to earlier years from the 13th century onward and finds a striking discontinuity about the year 1935 "suggesting that this is a pivotal period for present-day historical consciousness." But in the first place, this seems disconnected from the body of the book, because the relevant distributions are not LNRE. Furthermore, the only newspaper identified by name is the *Frankfurter Allgemeine Zeitung*, and although we are told that other newspapers show the same pattern, we are not told which newspapers these are. Finding that Germans perceive a unique historical discontinuity in the 1930s might be a very different thing from finding that Europeans, or Westerners in general, do so.

Nevertheless, this last chapter does also contain important findings that relate more closely to the central concerns of the book. In this chapter Baayen illustrates the

sophisticated statistical calculations that he uses in place of naïve type/token ratios, in the quest for characteristic constants of lexical usage. For each of a range of literary works, the calculations yield a curve occupying some portion of a two-dimensional "authorial space" (my phrase rather than Baayen's). With many pairs of separate works by the same author, the resulting curves are satisfactorily close to one another and well separated from curves for other authors: This is true when authors are as different as Henry James (*Confidence* and *The Europeans*) and St. Luke (St. Luke's Gospel and *Acts of the Apostles*). But there are exceptions: H. G. Wells is a case showing that "intra-author variability may be greater than inter-author variability," since the curves for his *War of the Worlds* and *The Invisible Man* are somewhat far apart, and the curves for Jack London's *Sea Wolf* and *The Call of the Wild* are superimposed on one another in the space between the two Wells curves.

The final chapter also contains a number of misprints, which are not self-correcting and may be worth listing here. In a discussion of word length distribution, there are repeated confusions between length 4, length 5, and length 6, on pages 196, 197 (Figure 6.1), 198 (Figure 6.2), and 199; some of the passages indicated may be correct as printed, but they cannot all be correct. On page 204, in a list of Dutch prefixes and suffixes, the prefixes *her-* and *ver-* are shown as suffixes. Page 208 cites "Baayen (1995)," which is not listed in the bibliography (the reference intended may be to the item listed as 1994b). In Table 6.1 (page 211) and the associated Figure 6.9 (page 212), there are mistakes in the codes for different literary works. (In the table, Emily Brontë's *Wuthering Heights* is coded identically to L. F. Baum's *Tip Manufactures a Pumpkinhead*—surely an implausible confusion—but *Wuthering Heights* seems to be "B1" in the figure; two novels by Arthur Conan Doyle are assigned the same code and identical word lengths in the table, whereas *The Hound of the Baskervilles* is probably the item coded "C2" in the figure.)

The volume is accompanied by a CD-ROM containing numerous relevant software programs; these and various data sets are detailed in a series of four appendices to the book.

**Acknowledgment**

I am grateful to fellow members of the Sussex University probabilistic parsing reading group for insights gained during the weeks when the book under review was our target volume. Responsibility for errors and inadequacies in this review is entirely mine.

**References**

Carroll, J. B. 1967. On sampling from a lognormal model of word frequency distribution. In Henry Kučera and W. Nelson Francis, editors, *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI, pages 406–424.

Honoré, A. 1979. Some simple measures of richness of vocabulary. *Association of Literary and Linguistic Computing Bulletin*, 7(2):172–179.

Khmaladze, E. V. 1987. The statistical analysis of large numbers of rare events. Technical Report MS-R8804, Department of Mathematical Sciences, Centrum voor Wiskunde en Informatica. Amsterdam: Centre for Mathematics and Computer Science.

Naranan, S. and V. Balasubrahmanyan. 1998. Models for power law relations in linguistics and information science. *Journal of Quantitative Linguistics*, 5:35–61.

Orlov, J. K. and R. Y. Chitashvili. 1983a. Generalized Z-distribution generating the well-known "rank-distributions." *Bulletin of the Academy of Sciences, Georgia*, 110:269–272.

Orlov, J. K. and R. Y. Chitashvili. 1983b. On the statistical interpretation of Zipf's law. *Bulletin of the Academy of Sciences, Georgia*, 109:505–508.

Sichel, H. S. 1975. On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70:542–547.

*Geoffrey Sampson* is Professor of natural language computing at the University of Sussex. Much of his research has concerned statistical parsing techniques; he has contributed the articles on "Statistical Linguistics" to successive editions of the *Oxford International Encyclopedia of Linguistics*. Sampson's address is School of Cognitive and Computing Sciences, University of Sussex, Falmer, Brighton BN1 9QH, England; e-mail: geoffs@cogs.susx.ac.uk.