Vocabulary: imitative or generative?

András Kornai*
MetaCarta Inc.

Gerald Penn†
University of Toronto

There are two traditionally opposed views of language: under the "imitative" view all linguistic competence is acquired by imitating material already seen, while under the "generative" view the acquisition of language involves the learning of patterns (rules or constraints) that can be recursively substituted in one another. The imitative view naturally leads to an assumption of finite vocabulary, and the generative view to infinite vocabulary. In this review article we provide a highly critical overview of the evidence presented by Baayen (2001) in favor of a finite vocabulary assumption.

Introduction

In the introductory section of the Mahabhashya, Patanjali has just argued that it is simpler to enumerate the correct expressions of the language than to enumerate the incorrect expressions. He then raises the question how the enumeration is to be done. By listing them? No, that would be difficult. For it is said that Brhaspati (the teacher of the gods) taught Indra a work containing all correct expressions of Sanskrit for a thousand divine years (360,000 human years), and still did not come to an end. So, a fortiori, how could it be possible these days, when people live for at most a hundred summers?

Given the acceptance of the Paninian view by generative grammar, and the astounding amount of empirical evidence, starting perhaps with Berko's (1958) "Wug" tests, that have been amassed in favor of the generative view, it is somewhat surprising to find a book (Baayen 2001), that is thoroughly grounded in the imitative view. In this review article, we take issue with this view by looking critically at the data presented by Baayen. We also discuss how the erroneous imitative view impacts the rest of the book.

To the extent feasible we adopt Baayen's notation: in what follows corpus size (number of tokens) is denoted by N, observed vocabulary size (number of types) by $V(N)$, and the number of types with token frequency i by $V(i, N)$. The probability of the r-th word is denoted by π_r, the sample (relative) frequency by $p(r, N)$. However, we would not do justice to Baayen by simply assuming that as N goes to infinity, $p(r, N) \to \pi_r$. This assumption, the empirical foundation of probabilistic arguments, what standard textbooks like Cramér (1955) call the *stability property of frequency ratios,* is clearly not good enough for Baayen, who early on cautions the reader that "the law of large numbers cannot be relied on when dealing with words and their frequencies of use" (p 7). We beg to differ: we will argue that Baayen himself not only fails to provide an alternative foundation for statistics and probability theory, but in fact keeps continually relying on the standard methods whose ultimate empirical justification lie in the central limit theorems he so disdains.

The conceptual confusion that results from Baayen's attempt to defend the indefensible imitative hypothesis extends to the notation as well: he defines the quantities N_i^* as the sample size N at which $V(i, N)$ reaches maximum, and S as the population

* 350 Massachusetts Ave, Cambridge, MA 02139
† 10 King's College Rd. Toronto M5S 3G4

vocabulary size. These numbers are the pot of gold at the end of the rainbow: in practice, no increasing sequence of linguistic samples shows a decrease of $V(i, N)$, ever, so there can be no such maximum, and of course $S = \infty$.

1 The main problems

Here we introduce two simple languages, Marrian, and CoinToss, to bring in sharp relief what we consider to the main problems with Baayen's approach: the use of absurdly small sample sizes and extending conclusions from small sublanguages to language as a whole.

1.1 Marrian

The Marrian language, inspired by the works of Nikolai Yakovlevich Marr (1865-1934), has only four words, *sal, ber, yon* and *rosh*, which occur with probability 1/2, 1/4, 1/6, and 1/12 respectively. For such a language, everything that Baayen proposes makes eminent sense: the number of hapaxes $V(1, N)$ is bounded by $S = 4$, and by the pigeonhole principle as soon as $N \geq 5$, $V(1, N) \leq 3$. Indeed, $V(k, N)$ is expected to be zero for $N >> 12k$, and as N tends to infinity, for any fixed $k, V(k, N)$ tends to zero.

At the heart of Baayen's positive contribution are some probabilistic derivations (which Chitashvili and Baayen 1993 doesn't claim to be original with them) concerning the expected values of $V(N)$ and similar quantities. In particular, by assuming that words are drawn from an urn with probabilities $\pi_1, \pi_2, \ldots \pi_S$ (with replacement, i.e. a multinomial model), he obtains

$$E[V(N)] = \sum_{i=1}^{S} (1 - e^{-N\pi_i}) \tag{1}$$

1.2 CoinToss

The rule of generating CoinToss texts is simple: a fair coin is tossed, and we append the grapheme H to our last word if it comes out heads, and terminate the word if it comes out tails – a new word is started with the letter H on the next toss that results in heads. Therefore, the vocabulary of CoinToss is *H, HH, HHH, HHHH* ... with probabilities 1/2, 1/4, 1/8. 1/16, ... respectively. No matter how large a corpus we collect, words hitherto unseen still enter the sample: at corpus size N we expect $V(N) = \log_2(N)$.

Notice that (1) will give approximately the right result if we terminate the summation at $V(N)$, but of course if we already know where to terminate the summation, we don't need to estimate $V(N)$. Critical to the derivation of (1) is the assumption that the absolute sample frequency f_i of any word can be replaced by $N\pi_i$, an estimate that receives its justification from the the laws of large numbers.

Of note here is the mean word frequency. For Marrian, the mean absolute frequency increases without bounds as $N \to \infty$, but mean relative frequency converges to 0.25. For CoinToss, the mean absolute frequency tends to infinity, but mean relative frequency tends to 0 with $1/V(N) = 1/\log_2(N)$.

1.3 Real languages

The standard view is that real languages are neither like Marrian nor like CoinToss. They are not like Marrian because their vocabulary size is infinite, (a point we shall take some pains to establish to the readers' satisfaction), and they are not like CoinToss because of a more subtle statistical observation known as Zipf's Law.

Zipf's idea (actually going back to much earlier work staring with Pareto 1897 and

Estoup 1916) was to rank the words in order of decreasing frequency, and investigate how *log* probability (log frequency) depends on rank. In CoinToss, $\log p_r = -\log(2)r$, so $\log p_r$ is a linearly decreasing function of r, while in real languages Zipf observed that $\log p_r$ is a linearly decreasing function of *log* rank.

Plotting frequencies against ranks on log-log paper is known as a *Zipf plot,* and it has been commonly observed that such plots are approximately linear. In geometric terms, the intercept of the linear function grows with $\log(N)$ as sample size N grows, and the slope tends to a constant $-B$, with B close to 1.

A closely related empirical observation is Herdan's Law, which states that vocabulary size $V(N)$ grows with a fractional power N^C of sample size N. Plotting number of types against number of tokens on log-log paper is known as a *Herdan plot,* but this time the intercept is zero and the slope is positive. Indeed, for the case of real languages the Zipf parameter B and the Herdan parameter C are related to each other by $C = 1/B$ (for a simple proof see Kornai 1999).

In CoinToss, log of vocab size

1.4 The use of small samples

Figure 1. plots the partial sums $\sum_{i=1}^{k} 1/i\sqrt{i}$ for $k = 1, \ldots, 12$. There is a strong resemblance to Baayen's plots such as his Fig 1.7A (p 18), which he uses to demonstrate the lack of convergence for Zipf's law. Yet $1/i\sqrt{i}$ is convergent, in fact the rate of convergence is not even too bad ($1/\sqrt{n}$).

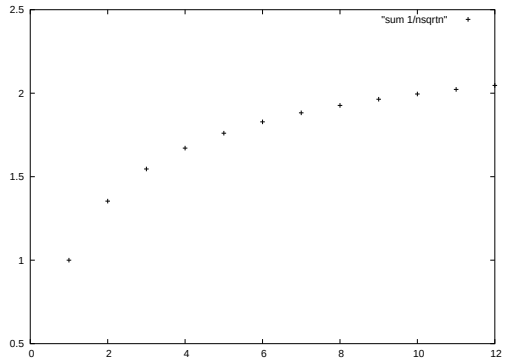

Figure 1
Partial sums in convergent series

What is so striking about Baayen's Figure 1.7 is that it lumps together as panels A and B, apparently entirely on the basis of graphical analogy, two radically different phenomena: the fact that the intercept in Zipf's Law plots grows without bounds (true divergence) and that the slope converges slowly (true convergence that looks as divergence for small numbers, just as our Figure 1.).

The graphical sleight-of-hand is of course a reflection of a far deeper methodological error, one that pervades the whole book, namely the use of tiny samples. In Chapter 1, Baayen illustrates many of his subsequent points with *Alice in Wonderland,* a corpus that has only 27,283 tokens for 2,571 types. Our numbers differ slightly from Baayen's, but as he did not make his raw data public (in spite of well over 640MB of unused space available on the CD-ROM that comes with the book) we have trouble exactly replicating his findings, a matter we shall return to in 1.5. [For our own computations, we used

the Project Gutenberg e-text of *Alice*, but probably our tokenization rules are slightly different. This does not appear to affect the overall conclusion here.] An independent count of contemporary journalistic prose based on a sample over two thousand times bigger (55.5m words from the *Wall Street Journal*) shows more than a dozen words such as *treacle, sluggard*, or *mournfully* that appear only once in this larger corpus, and more importantly, it shows well over a hundred words such as *barrowful, piteous* or *porpoise* that appear in *Alice* but not in the larger sample.

There are significant differences in the age, composition, and authorship of the two samples to be sure. Yet the authors of both belong to the same broad historical period of the same language, Modern English, and we have no doubt that the low frequency (one in a few million, rather than one in a few thousand) of words such as *treacle* or *barrowful* would be borne out by a wordcount restricted to literary prose strictly contemporaneous with Lewis Carroll. The point is simply this: by using a word in a text that is only N words long, the author does not actually endow the word with population frequency $\geq 1/N$. We have every reason to believe (including the evidence from his other published writings) that Lewis Carroll was familiar with many words that do not appear in *Alice,* and therefore *Alice* can only be considered a non-representative sample of his language.

Indeed, to take a first stab at establishing characteristics based on word frequency, we should at the very least consider the core vocabulary listed in abridged dictionaries, somewhere between 30 and 60 thousand words. The tail end of this vocabulary is full of words such as *spinal* which are perfectly familiar to speakers of English, yet have frequency less than one in a million. Therefore, it is unlikely in the extreme that characteristics of English usage could be even vaguely approximated based on samples that are bound by their low size not to contain these words: a more reasonable starting point would be a few million words of text, or two orders of magnitude more. Given this size requirement, Baayen's method of subdividing a tiny corpus into 20 equal parts, and thus losing another order of magnitude, is absurd.

We wish to emphasize here that the above argument, based on the size of abridged dictionaries, makes no appeal to Lewis Carroll's literary genius – if anything, we expect such geniuses to have a wider range of vocabulary than average, and thus requiring even larger samples. This is not to say that studies of literary vocabulary, authorship etc. based on small samples are inherently worthless: only a few writers are prolific enough to bequeath millions of words to posterity, and clearly there are a lot of situations when small samples are all we have. However, generalizing from such small samples to English as a whole is a fatal methodological error.

1.5 Extending conclusions from small sublanguages to language as a whole

Chitashvili and Baayen (1983) use two sublanguages to contrast important behavioral differences between sublanguages as sample size is increased. On the one hand, the sublanguage of stems that occur with the productive suffix *-ness* appears as CoinToss – the more words we add the more new stems we discover. On the other, the sublanguage of stems that occur with the improductive prefix *en-* appears as Marrian: once the variety of words is exhausted all we gain by larger samples is a more precise estimate of the population frequencies, but no change in inventory.

To distinguish the two cases, Chitashvili and Baayen (1983) introduce the notion of the LNRE (Large Number of Rare Events) zone (defined essentially the same way in Baayen 2001 p 55-56, who refers to Khmaladze 1987 – see 1.5 for further discussion). A sample is in this zone as long as it is smaller than N_1^*, the sample size at which $V(1, N)$ reaches maximum. So far, so good. But what is missing from the main line of the discussion is that in general there is no regular progression from the LNRE zone to the post-LNRE zone: not only are our current samples inside the LNRE zone but we have

reasons to believe that no sample will ever saturate and show a decreasing proportion of hapaxes. Sublanguages like words in *en-* are interesting, and there is clearly value in statistical criteria that help us determine whether such a sublanguage is imitative or generative. But such criteria, when applied to English or other natural languages, actually support the generative view, a point that we shall establish in greater detail in 2 below.

It is not that Baayen goes out and explicitly denies the possibility of $S = \infty$, to the contrary, there are snippets sprinkled through the whole text alluding to such a possibility. However, these snippets appear buried in subordinate clauses in the middle of lengthy discussions, while in the main line of argumentation (such as the summary to Chapter 1, p 34) we find statements such as the following:

> The main thrust of this chapter has been to show that for word frequency distributions the sample mean and many other summary measures change in a highly systematic way as a function of sample size. The parameters of the zeta (Zipf) and lognormal distribution are subject to exactly the same kind of systematic dependency on the sample size.

To fully appreciate how misguided this is, one needs to know that by this time Baayen actually considered, and on the basis of a single *Alice* plot dismissed, Herdan's Law of vocabulary growth $V(N) = N^C$ (see his Figure 1.16 and the related discussion on pp 29-30), in spite of the fact that his own estimates of Herdan's constant move only from .798 to .782, i.e. only 2%, for doubling (in his case, halving) the corpus size.

As we shall argue in Section 2, Herdan's Law is correct at least to a first approximation. If this is so, sample mean frequency is approximately N^{1-C} (absolute) or N^{-C} (relative), numbers that will indeed "change in a highly systematic way as a function of N". In the next sentence, Baayen again lumps together, on the basis of graphical similarity, two radically different phenomena: the fact that estimates of the lognormal parameters diverge (noted as early as Carroll 1967:407) , and the fact that Zipfian slope can not be reliably estimated on tiny samples.

Yet the problem with the lognormal hypothesis is a fundamental one, which stems from fact that as average frequency decreases with N^{-C}, average log frequency diverges. Because the lognormal predicts $S = \exp(\sigma^2/2 - \mu)$, i.e. an absolute limit on vocabulary size, there is no arithmetic wiggle room, and the variance estimate must also change with $\sqrt{\log N}$ (for a more detailed derivation of this effect see Kornai 1999). In sharp contrast, the Zipfian hypothesis does not lead to the assumption of a finite (imitative) vocabulary limit: beyond the trivial observation that the intercept in Zipf-style (log frequency vs. log rank) plots must grow with $\log N$, there is no reason to suppose that the critical slope parameter B does not show convergence. We return to this matter in Section 2, where we survey some of the literature and data in support of the standard Zipf model.

1.6 Chitashvili, Khmaladze, Orlov

Baayen dedicates the book to the memory of Rezo Chitashvili, makes many references to their work on LNRE models, and cites no less than nine papers authored or coauthored by Chitashvili, Khmaladze, and Orlov. Yet aside from the early papers of Orlov (which have no discussion of LNRE), and a joint paper by Chitashvili and Baayen, none of this material is available to Western scholars, having been published in places such as the *Bulletin of the Georgian Academy of Sciences* and the *Transactions of the Tbilisi Mathematical Institute.*

The authors have run independent library searches at some of the major academic libraries of North America, including the library systems of Harvard, Stanford, MIT, Indiana University, and University of Toronto, and came up empty. (We actually wrote

to the Razmadze Mathematical Institute in Tbilisi and CWI in Amsterdam in search of the more elusive tech reports, but received no answer).

Sadly, no Western scholar has access to the material. Therefore, when Baayen relegates a discussion of what he calls the Orlov-Chitashvili model to section 3.2.2 of his book (under the rather misleading heading "The Zipfian family of LNRE models"), and refrains from mathematically defining it, he closes off the last avenue that others could have to this material. As with the lack of publishing his raw data, the lack of e.g. a website with reprints of the critical papers creates the impression that Baayen prefers to operate in the lacunae of library budgets rather than fostering (indeed, enabling) scholarly debate.

2 Zipf's and Herdan's Laws

In this section we discuss the significance of Zipf's laws and the closely related Herdan law, and present some empirical evidence in favor of these laws. Finally, we describe some simple experiments that the reader can perform herself, and make some specific predictions concerning the outcome of these.

2.1 Understanding the main terms

When dealing with complex phenomena, the first step is almost invariably to understand the main terms of the laws that govern it. Nobody who understands elasticity would go out on a limb and claim that the displacement of a spring is *strictly* linearly proportional with the force applied to it. Yet it is precisely this statement of linear relationship, Hooke's Law, which lies at the heart of our understanding the phenomenon of elasticity, because for small forces and displacements it provides the main term of the true equation. In fact, for most purposes we don't even need to know the true equation – by careful use of Hooke's Law we can design machinery with elastic components.

Arranging the words in a corpus of size N in order of decreasing frequency, the Rth type (w_R) is said to have *rank* R, and its frequency is denoted by $f(R, N)$. As Estoup (1916) and Zipf (1935) noted, the plot of log frequencies against log ranks shows, at least in the middle range, a reasonably linear relation. Fig. 2 shows this for a sample of the works of Dickens (5.5m words, based on all of Project Gutenberg's Dickens e-texts).

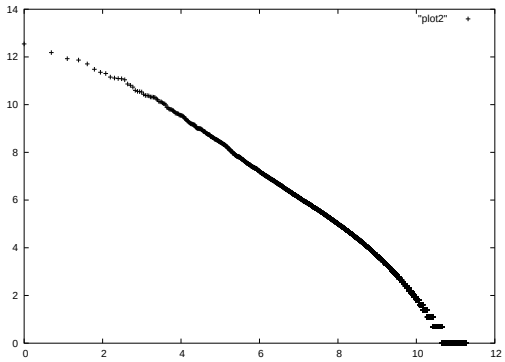

Figure 2
Plot of log frequency as a function of log rank for Dickens

It is obvious that the relationship is not strictly linear. But the first and crudest approx-

imation to the empirical curve is still

$$\log(p(r)) = H_N - B_N \log(r) \tag{2}$$

i.e. Zipf's first law. In (2) we replaced absolute ranks and frequencies by relative units: on the x axis we use relative ranks $r = R/V(N)$ rather than absolute ranks R, and on the y axis relative frequencies $p = f/N$ rather than absolute frequencies f (these correspond to linear shifts as long as both axes are plotted logarithmically). There are two terms to this approximation, the constant term (intercept) H_N and the linear term (slope) B_N. To some extent, these will depend on N, but they are independent of r.

Looking at the highest rank $R = V(N)$, we have $r = 1$ and $p(r) = 1/N$ (assuming at least one hapax in the sample, a matter we return to shortly), which yields $H_N = \log(1/N)$, assuming a perfect fit at the right margin of the plot. Therefore, Baayen is correct in noting that H_N, the Zipf intercept, is not constant. But the conclusion he draws from this, that Zipf's law itself is subject to systematic dependency on sample size, does not follow, since the classical Zipf model is a one-parameter distribution and the key issue is dependence, or lack thereof, of this one parameter on sample size.

Least squares fitting in the log-log plane is in fact not a great way to estimate the slope parameter, because it overly emphasizes the high frequency items, whose exact distribution is highly dependent on tokenization (as already noted in Mandelbrot 1961). A better approximation is obtained from considering the high frequency items, or just the single highest frequency item, only asymptotically. At the lowest rank we find the most frequent word, typically *the*, with (base e) log frequency approximately -3, tending to the constant $\log(\pi_{the})$ as we grow N. For the right-hand side of (2) to tend to constant, we need $H_N = \log(1/N)$ to grow at the same rate as $B_N \log(1/V(N))$. If Herdan's Law holds up, i.e. $V(N)$ grows with N^C, $\log(1/V(N))$ will grow with $C \log(1/N)$, and therefore the required asymptotic condition is met as long as B_N tends to $1/C$.

Another method of deemphasizing the high frequency range is to focus on the $V(i, N)$ that capture, for low i, the primarily the statistics governing the tail. For Zipf's second law, we plot $\log(i)$ against $log(V(i, N))$, and again obtain an approximately linear relation with slope $-D_N$. Figure 3. shows this for Baayen's *The Independent* corpus (8m words):

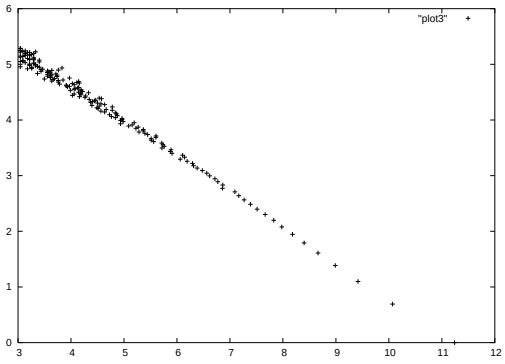

Figure 3
Zipf's Second Law for *The Independent* (8m words)

Again we would have to normalize the axes to account for growing sample size, but the conclusion is the same as above: only the slope parameter is critical, and the equation $D = B/(1 + B)$ connects the slope parameter of the second law to the slope parameter

of the first law. Altogether, the two Zipf's laws and the Herdan law yield asymptotically the same results for the parameter range of practical interest ($C < 1$), and the defining parameters B, C, D can be simply computed from one another.

Returning to Dickens for a moment, taking the eight novels over 300k words that are accessible to us as e-texts, we find that the proportion of hapaxes, $V(1,N)/V(N)$ is remarkably consistent, ranging from a low of 38.1% in *Nicholas Nickleby* to a high of 40.2% in *Martin Chuzzlewit*. We also find that Herdan's C ranges between a low of 0.751 in *David Copperfield* to a high of 0.767 *The Pickwick Club*. The high consistency of these numbers shows that, contrary to Baayen's claims, these are very good measures of vocabulary richness.

work	N	$V(N)$	$V(1,N)$	$V(1,N)/V(N)$	C
Dombey and Son	366649	16397	6355	0.388	0.757
Copperfield	366335	15164	6007	0.396	0.751
Bleak House	357668	15573	6146	0.395	0.755
Little Dorrit	348719	15856	6165	0.389	0.758
Chuzzlewit	348659	16608	6682	0.402	0.761
Our Mutual Friend	338405	16357	6492	0.397	0.762
Nickleby	334971	15932	6069	0.381	0.761
Pickwick Club	313215	16318	6490	0.398	0.767

Once again, we make no claim here that the Zipf or Herdan laws are the final answer to all questions of word frequency distributions. To the contrary, we believe these to be no more than very reasonable first approximations. The problem with Baayen's work is not that he claims these laws to be untrue, for surely they are untrue if held to a standard of exactness. The problem is that he entirely misses their pivotal status as first approximations: more sophisticated statistical models need to be conservative extensions of these, replicating the same first order effects, including infinite (generative) vocabulary.

Since Baayen lacks the guidance offered by asymptotic analysis of the first order, second order, and higher order effects, his discussion of the more sophisticated statistical models (Chapter 3) is entirely haphazard and confusing. He is simply incapable of distinguishing those models such as Waring and lognormal which have no chance of accounting for the first order phenomena, from those models that do, such as negative binomial (Efron and Thisted 1976, Thisted and Efron 1987).

2.2 Experiments on larger corpora

With the sample from *The Independent,* we have reached the outer limits of Baayen's work: all but 4 of the 24 samples he uses are less than 170k words (the remaining are 1m, 3.5m, 6.2m and 7.8m respectively). This may have been reasonable in the seventies, when the best corpora in wide use were the Brown (1m words), London-Lund (.5m words), and LOB (.5m). But by the eighties far larger corpora with $10^7 - 10^8$ were widely disseminated by the Linguistic Data Consortium, and by the nineties billion-word (10^9) probabilistic language models were commonly used e.g. in speech recognition.

The proportion of hapaxes, which we have seen to be about 40% for the longer novels of Dickens, has shown the exact opposite tendency as predicted by Baayen: multi-million word corpora show 50% hapaxes or more, and billion-word corpora often have 60% hapaxes. Monolingual segments of large search engine caches (trillion-word corpora) can have as high as 70%. There is no sign of vocabulary saturation anywhere: to the contrary, we see the elusive S moving farther and farther away.

Our advice is: try this at home. At this point, collecting a billion words is not hard for anyone with a good network connection, and ordinary PCs are perfectly capable of performing frequency counts on these. We predict that *no* near-random sample of *any* natural language will show no hapaxes, or even a tiny fraction of hapaxes: they are really

and truly LNRE.

As scientists, we are obliged to promulgate those hypotheses which are consistent with the observed range of facts, and to discard those that are not supported by any empirical data. We invite the readers to collect their own data and see whose predictions are borne out.

3 Conclusion

At the core of the book, there is a positive contribution to the study of word frequency distributions: some sublanguages appear to be imitative, while others are generative. That such a difference can exist among sublanguages is of course no surprise: for example, strings with length 3 form a sublanguage whose vocabulary size S is bounded by 26^3 (assuming tokenization to lowercase, as Baayen does), while strings composed of nothing but digits will have $V(N)$ as high as there are numbers spelled out in the text, clearly coming from an infinite pool. The value of Baayen's tests comes from distinguishing imitative (closed) and generative (open) sublanguages based on relatively small samples.

Yet this positive contribution is all but destroyed by the vaulting ambition of the book to extend the framework from small controlled sublanguages to language as a whole: in his quest to define LNRE to be central to the study of word frequencies Baayen loses sight of the classical facts easily observable on large corpora, starting with Zipf's Laws. As we have argued here, LNRE distributions are indeed central, but not simply because our samples happen to be in the LNRE zone, but rather because there is no other zone at the end of the rainbow.

This finding should naturally increase interest in the Georgian work that Baayen so persistently alludes to. As we discussed in 1.5 above, we have not had the good fortune to consult this literature, and have no means to pass judgment on it. [For Baayen's custody of this material, see Matthew 25:18 et seq] To the extent we can guess, Khmaladze, Chitashvili, and Orlov do make the all-important distinction between imitative (finite) and generative (infinite) vocabulary.

Many scholars, interested in language but trained more in the tradition of the humanities than the sciences, will look at studies of word frequency as a good entry point into the world of computational linguistics and natural language processing, with its heavy statistical machinery. We must regretfully advise all such readers to steer clear of Baayen's book.

Acknowledgments

We are indebted to Paul Kiparsky and Michael Inman for their help with the Mahabhasya. We thank Project Gutenberg for their wonderful collection of e-texts, and Axelero Internet for access to large monolingual search engine caches.

Baayen, R. H.: 2001, *Word Frequency Distributions* Dordrecht: Kluwer Academic Publishers

Berko, J.: 1958 The child's learning of English morphology. *Word* **14**

Carroll, J. B.: 1967, On Sampling from a lognormal model of word-frequency distribution. In: H. Kucera and W. Francis (eds.): *Computational Analysis of Present-Day American English.* Providence, RI: Brown University Press, pp. 406–424.

Cramér, H.: 1955, *The elements of probability theory.* New York: John Wiley & Sons.

Efron, B. and R. Thisted: 1976, Estimating the number of unseen species: How many words did Shakespeare know?. *Biometrika* **63**, 435–448.

Estoup, J.: 1916, *Gammes Stenographiques.* Paris: Institut Stenographique de France.

Kornai, A.: 1999: Zipf's Law outside the middle range. In: J. Rogers (ed): *Proceedings of the 6th Meeting on Mathematics of*

Language (MOL6), University of Central Florida 347–356

Mandelbrot, B.: 1961, On the thory of word frequencies and on related markovian models of discourse. In: R. Jakobson (ed.): *Structure of language and its mathematical aspects*. American Mathematical Society, pp. 190–219.

Thisted, R. and B. Efron: 1987, Did Shakespeare write a newly-discovered poem?. *Biometrika* **74**, 445–455.

Zipf, G. K.: 1935, *The psycho-biology of language; an introduction to dynamic philology*. Boston: Houghton Mifflin.

Zipf, G. K.: 1949, *Human Behavior and the Principle of Least Effort*. Addison-Wesley.