A STOCHASTIC PROCESS FOR WORD FREQUENCY DISTRIBUTIONS

Harald Baayen*
Max-Planck-Institut für Psycholinguistik
Wundtlaan 1, NL-6525 XD Nijmegen
Internet: baayen@mpi.nl

ABSTRACT

A stochastic model based on insights of Mandelbrot (1953) and Simon (1955) is discussed against the background of new criteria of adequacy that have become available recently as a result of studies of the similarity relations between words as found in large computerized text corpora.

FREQUENCY DISTRIBUTIONS

Various models for word frequency distributions have been developed since Zipf (1935) applied the zeta distribution to describe a wide range of lexical data. Mandelbrot (1953, 1962) extended Zipf's distribution 'law'

$$f_i = \frac{K}{i^A}, \tag{1}$$

where f_i is the sample frequency of the i^{th} type in a ranking according to decreasing frequency, with the parameter B,

$$f_i = \frac{K}{B + i^A}, \tag{2}$$

by means of which fits are obtained that are more accurate with respect to the higher frequency words. Simon (1955, 1960) developed a stochastic process which has the Yule distribution

$$f_i = AB(i, \rho + 1), \tag{3}$$

with the parameter A and $B(i, \rho + 1)$ the Beta function in $(i, \rho + 1)$, as its stationary solutions. For $i \to \infty$, (3) can be written as

$$f_i \sim \Gamma(\rho + 1) i^{-(\rho+1)},$$

in other words, (3) approximates Zipf's law with respect to the lower frequency words, the tail of the distribution. Other models, such as Good (1953), Waring-Herdan (Herdan 1960, Muller 1979) and Sichel (1975), have been put forward, all of which have Zipf's law as some special or limiting form. Unrelated to Zipf's law is the lognormal hypothesis, advanced for word frequency distributions by Carroll (1967, 1969), which gives rise to reasonable fits and is widely used in psycholinguistic research on word frequency effects in mental processing.

A problem that immediately arises in the context of the study of word frequency distributions concerns the fact that these distributions have two important characteristics which they share with other so-called large number of rare events (LNRE) distributions (Orlov and Chitashvili 1983, Chitashvili and Khmaladze 1989), namely that on the one hand a huge number of different word types appears, and that on the other hand it is observed that while some events have reasonably stable frequencies, others occur only once, twice, etc. Crucially, these rare events occupy a significant portion of the list of all types observed. The presence of such large numbers of very low frequency types effects a significant bias between the rank-probability distribution and the rank-frequency distribution, leading to the contradiction of the common mean of the law of large numbers, so that expressions concerning frequencies cannot be taken to approximate expressions concerning probabilities. The fact that for LNRE distributions the rank-probability distributions cannot be reliably estimated on the basis of rank-frequency distributions is one source of the lack of goodness-of-fit often observed when various distribution 'laws' are applied to empirical data. Better results are obtained with Zipfian models when Orlov and Chitashvili's (1983) extended generalized Zipf's law is used.

A second problem which arises when the appropriateness of the various lexical models is

*I am indebted to Klaas van Harn, Richard Gill, Bert Hoeks and Erik Schils for stimulating discussions on the statistical analysis of lexical similarity relations.

considered, the central issue of the present discussion, concerns the similarity relations among words in lexical distributions. These empirical similarity relations, as observed for large corpora of words, impose additional criteria on the adequacy of models for word frequency distributions.

SIMILARITY RELATIONS

There is a growing consensus in psycholinguistic research that word recognition depends not only on properties of the target word (e.g. its length and frequency), but also upon the number and nature of its lexical competitors or neighbors. The first to study similarity relations among lexical competitors in the lexicon in relation to lexical frequency were Landauer and Streeter (1973). Let a *neighbor* be a word that differs in exactly one phoneme (or letter) from a given target string, and let the neighborhood be the set of all neighbors, i.e. the set of all words at Hamming distance 1 from the target. Landauer and Streeter observed that (1) high-frequency words have more neighbors than low-frequency words (the neighborhood density effect), and that (2) high-frequency words have higher-frequency neighbors than low-frequency words (the neighborhood frequency effect). In order to facilitate statistical analysis, it is convenient to restate the neighborhood frequency effect as a correlation between the target's number of neighbors and the frequencies of these neighbors, rather than as a relation between the target's frequency and the frequencies of its neighbors — targets with many neighbors having higher frequency neighbors, and hence a higher mean neighborhood frequency $\bar{f}_n$ than targets with few neighbors. In fact, both the neighborhood density and the neighborhood frequency effect are descriptions of a single property of lexical space, namely that its dense similarity regions are populated by the higher frequency types. A crucial property of word frequency distributions is that the lexical similarity effects occur not only across but also within word lengths.

Figure 1A displays the rank-frequency distribution of Dutch monomorphemic phonologically represented stems, function words excluded, and charts the lexical similarity effects of the subset of words with length 4 by means of boxplots. These show the mean (dotted line), the median, the upper and lower quartiles, the most extreme data points within 1.5 times the interquartile range, and remaining outliers for the number of neighbors ($\#n$) against target frequency (neighborhood density), and for the mean frequency of the neighbors of a target ($\bar{f}_n$) against the number of neighbors of the target (neighborhood frequency), for targets grouped into frequency and density classes respectively. Observe that the rank-frequency distribution of monomorphemic Dutch words does not show up as a straight line in a double logarithmic plot, that there is a small neighborhood density effect and a somewhat more pronounced neighborhood frequency effect. A Spearman rank correlation analysis reveals that the lexical similarity effects of figure 1A are statistically highly significant trends ($p < 0.001$), even though the correlations themselves are quite weak (see table 1, column 1): in the case of lexical density only 6% of the variance is explained.[1]

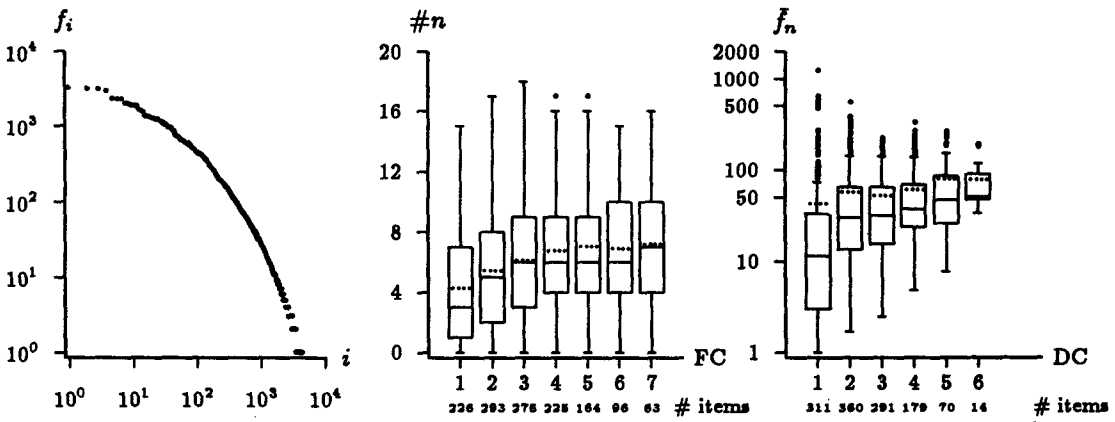
Table 1: Spearman rank correlation analysis of the neighborhood density and frequency effects for empirical and theoretical words of length 4.

		Dutch	Mand.	Mand.-Simon
dens.	r_s	0.24	0.65	0.31
	r_s^2	0.06	0.42	0.10
	t	9.16	68.58	11.97
	df	1340	6423	1348
freq.	r_s	0.51	0.62	0.61
	r_s^2	0.26	0.38	0.37
	t	21.65	63.02	28.22
	df	1340	6423	1348

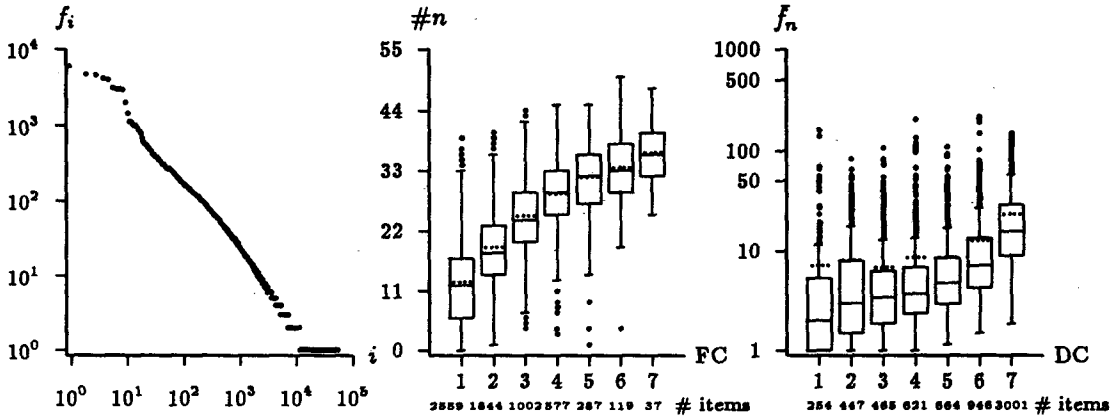
STOCHASTIC MODELLING

By themselves, models of the kind proposed by Zipf, Herdan and Muller or Sichel, even though they may yield reasonable fits to particular word frequency distributions, have no bearing on the similarity relations in the lexicon. The only model that is promising in this respect is that of Mandelbrot (1953, 1962). Mandelbrot derived his modification of Zipf's law (2) on the basis of a Markovian model for generating words as strings of letters, in combination with some assumptions concerning the cost of transmitting the words generated in some optimal code, giving a precise interpretation to Zipf's 'law of abbreviation'. Miller (1957), wishing to avoid a teleological explanation, showed that the Zipf-Mandelbrot law can also be derived under slightly different assumptions. Interestingly, Nusbaum (1985), on the basis of simulation results with a slightly different neighbor definition, reports that the neighborhood density and neighborhood frequency effects occur within

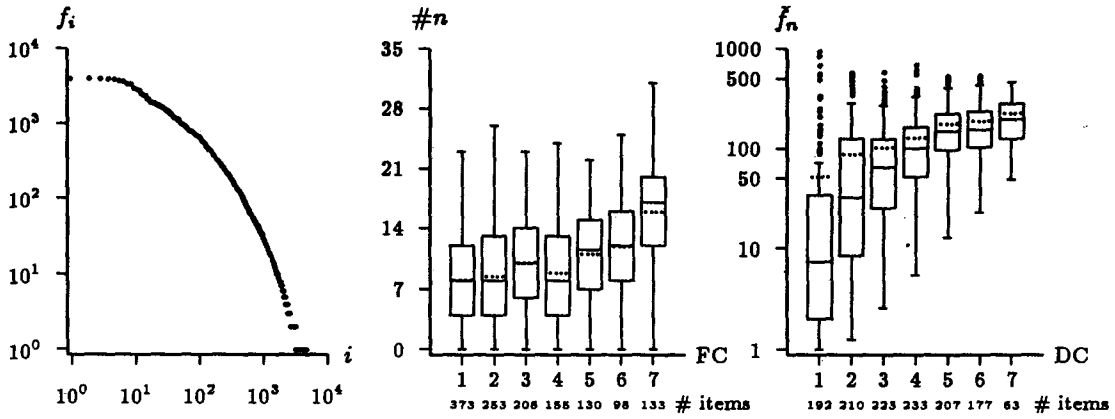
[1] Note that the larger value of r_s^2 for the neighborhood frequency effect is a direct consequence of the fact that the frequencies of the neighbors of each target are *averaged* before they enter into the calculations, masking much of the variance.



A: Dutch monomorphemic stems in the CELEX database, standardized at 1,000,000. For the total distribution, $N = 224567$, $V = 4455$. For strings of length 4, $N = 64854$, $V = 1342$.



B: Simulated Dutch monomorphemic stems, as generated by a Markov process. For the total distribution, $N = 224567$, $V = 58300$. For strings of length 4, $N = 74618$, $V = 6425$.



C: Simulated Dutch monomorphemic stems, as generated by the Mandelbrot-Simon model ($\alpha = 0.01$, $V_C = 2000$). For the total distribution, $N = 291944$, $V = 4848$. For strings of length 4, $N = 123317$, $V = 1350$.

Figure 1: Rank-frequency and lexical similarity characteristics of the empirical and two simulated distributions of Dutch phonological stems. From left to right: double logarithmic plot of rank i versus frequency f_i, boxplot of frequency class FC (1:1;2:2–4;3:5–12;4:13–33;5:34–90;6:91–244;7:245+) versus number of neighbors $\#n$ (length 4), and boxplot of density class DC (1:1–3;2:4–6;3:7–9;4:10–12;5:13–15;6:16–19;7:20+) versus mean frequency of neighbors $\bar{f}_n$ (length 4). (Note that not all axes are scaled equally across the three distributions). N: number of tokens, V: number of types.

a given word length when the transition probabilities are not uniformly distributed. Unfortunately, he leaves unexplained why these effects occur, and to what extent his simulation is a realistic model of lexical items as used in real speech.

In order to come to a more precise understanding of the source and nature of the lexical similarity effects in natural language we studied two stochastic models by means of computer simulations. We first discuss the Markovian model figuring in Mandelbrot's derivation of (2).

Consider a first-order Markov process. Let $A = \{0, 1, \ldots, k\}$ be the set of phonemes of the language, with 0 representing the terminating character space, and let $\mathcal{P} = (p_{ij})_{i,j \in A}$ with $P_{00} = 0$. If X_n is the letter in the n^{th} position of a string, we define $P(X_0 = i) = p_{0i}, i \in A$. Let y be a finite string $(i_0, i_1, \ldots, i_{m-1})$ for $m \in N$ and define $X^{(m)} := (X_0, X_1, \ldots, X_{m-1})$, then

$$p_y := P(X^{(m)} = y) = p_{0i_0} p_{i_0 i_1} \cdots p_{i_{m-2} i_{m-1}}. \tag{4}$$

The string types of varying length m, terminating with the space and without any intervening space characters, constitute the words of the theoretical vocabulary

$$S_m := \{(i_0, i_1, \ldots, i_{m-2}, 0) : i_j \in A \setminus 0, j = 0, 1, \ldots, m-2, m \in N\}.$$

With N_y the token frequency of type y and V the number of different types, the vector $(N_{y_1}, N_{y_2}, \ldots, N_{y_V})$ is multinomially distributed. Focussing on the neighborhood density effect, and defining the neighborhood of a target string y^t for fixed length m as

$$C_t := \{y \in S_m : \exists! i \in \{0, 1, \ldots, m-2\} \text{ such that } y_i \neq y^t\},$$

we have that the expected number of neighbors of y_t equals

$$\mathrm{E}[V(C_t)] = \sum_{y \in C_t} \{1 - (1 - p_y)^N\}, \tag{5}$$

with N denoting the number of trials (i.e. the number of tokens sampled). Note that when the transition matrix $\mathcal{P}$ defines a uniform distribution (all p_{ij} equal), we immediately have that the expected neighborhood density for length m_1 is identical for all targets y_t, while for length $m_2 > m_1$ the expected density will be less than that at length m_1, since $p_y^{(m_2)} < p_y^{(m_1)}$ given (4). With $\mathrm{E}[N_y] = N p_y$, we find that the neighborhood density effect does occur across word lengths, even though the transition probabilities are uniformly distributed.

In order to obtain a realistic, non-trivial theoretical word distribution comparable with the empirical data of figure 1A, the transition matrix $\mathcal{P}$ was constructed such that it generated a subset of phonotactically legal (possible) monomorphematic strings of Dutch by conditioning consonant C_k in the string $X_i X_j C_k$ on X_j and the segmental nature (C or V) of X_i, while vowels were conditioned on the preceding segment only. This procedure allowed us to differentiate between e.g. phonotactically legal word initial *kn* and illegal word final *kn* sequences, at the same time avoiding full conditioning on two preceding segments, which, for four-letter words, would come uncomfortably close to building the probabilities of the individual words in the database into the model.

The rank-frequency distribution of 58300 types and 224567 tokens (disregarding strings of length 1) obtained by means of this (second order) Markov process shows up in a double logarithmic plot as roughly linear (figure 1B). Although the curve has the general Zipfian shape, the deviations at head and tail are present by necessity in the light of Rouault (1978). A comparison with figure 1A reveals that the large surplus of very low frequency types is highly unsatisfactory. The model (given the present transition matrix) fails to replicate the high rate of use of the relatively limited set of words of natural language.

The lexical similarity effects as they emerge for the simulated strings of length 4 are displayed in the boxplots of figure 1B. A very pronounced neighborhood density effect is found, in combination with a subdued neighborhood frequency effect (see table 1, column 2).

The appearance of the neighborhood density effect within a fixed string length in the Markovian scheme with non-uniformly distributed p_{ij} can be readily understood in the simple case of the first order Markov model outlined above. Since neighbors are obtained by substitution of a single element of the phoneme inventory A, two consecutive transitional probabilities of (4) have to be replaced. For increasing target probability p_{y^t}, the constituting transition probabilities p_{ij} must increase, so that, especially for non-trivial m, the neighbors $y \in C_t$ will generally be protected against low probabilities p_y. Consequently, by (5), for fixed length m, higher frequency words will have more neighbors than lower frequency words for non-uniformly distributed transition probabilities.

The fact that the lexical similarity effects emerge for target strings of the same length is a strong point in favour of a Markovian source

for word frequency distributions. Unfortunately, comparing the results of figure 1B with those of figure 1A, it appears that the effects are of the wrong order of magnitude: the neighborhood density effect is far too strong, the neighborhood frequency effect somewhat too weak. The source of this distortion can be traced to the extremely large number of types generated (6425) for a number of tokens (74618) for which the empirical data (64854 tokens) allow only 1342 types. This large surplus of types gives rise to an inflated neighborhood density effect, with the concomitant effect that neighborhood frequency is scaled down. Rather than attempting to address this issue by changing the transition matrix by using a more constrained but less realistic data set, another option is explored here, namely the idea to supplement the Markovian stochastic process with a second stochastic process developed by Simon (1955), by means of which the intensive use can be modelled to which the word types of natural language are put.

Consider the frequency distribution of e.g. a corpus that is being compiled, and assume that at some stage of compilation N word tokens have been observed. Let $n_r^{(N)}$ be the number of word types that have occurred exactly r times in these first N words. If we allow for the possibilities that both new types can be sampled, and old types can be re-used, Simon's model in its simplest form is obtained under the three assumptions that (1) the probability that the $(N+1)$-st word is a type that has appeared exactly r times is proportional to $rn_r^{(N)}$, the summed token frequencies of all types with token frequency r at stage N, that (2) there is a constant probability α that the $(N+1)$-st word represents a new type, and that (3) all frequencies grow proportionaly with N, so that

$$\frac{n_r^{(N+1)}}{n_r^{(N)}} = \frac{N+1}{N} \text{ for all } r, N.$$

Simon (1955) shows that the Yule-distribution (3) follows from these assumptions. When the third assumption is replaced by the assumptions that word types are dropped with a probability proportional to their token frequency, and that old words are dropped at the same rate at which new word types are introduced so that the total number of tokens in the distribution is a constant, the Yule-distribution is again found to follow (Simon 1960).

By itself, this stochastic process has no explanatory value with respect to the similarity relations between words. It specifies use and reuse of word types, without any reference to segmental constituency or length. However, when a Markovian process is fitted as a front end to Simon's stochastic process, a hybrid model results that has the desired properties, since the latter process can be used to force the required high intensity of use on the types of its input distribution. The Markovian front end of the model can be thought of as defining a probability distribution that reflects the ease with which words can be pronounced by the human vocal tract, an implementation of phonotaxis. The second component of the model can be viewed as simulating interfering factors pertaining to language use. Extralinguistic factors codetermine the extent to which words are put to use, independently of the slot occupied by these words in the network of similarity relations,[2] and may effect a substantial reduction of the lexical similarity effects.

Qualitatively satisfying results were obtained with this 'Mandelbrot-Simon' stochastic model, using the transition matrix of figure 1B for the Markovian front end and fixing Simon's birth rate α at 0.01.[3] An additional parameter, V_C, the critical number of types for which the switch from the front end to what we will refer to as the component of use is made, was fixed at 2000. Figure 1C shows that both the general shape of the rank-frequency curve in a double logarithmic grid, as well as the lexical similarity effects (table 1, column 3) are highly similar to the empirical observations (figure 1A). Moreover, the overall number of types (4848) and the number of types of length 4 (1350) closely approximate the empirical numbers of types (4455 and 1342 respectively), and the same holds for the overall numbers of tokens (291944 and 224567) respectively. Only the number of tokens of length 4 is overestimated by a factor 2. Nevertheless, the type-token ratio is far more balanced than in the original Markovian scheme. Given that the transition matrix models only part of the phonotaxis of Dutch, a perfect match between the theoretical and empirical distributions is not to be expected.

The present results were obtained by implementing Simon's stochastic model in a slightly modified form, however. Simon's derivation of the Yule-distribution builds on the assumption that each r grows proportionaly with N, an as-

[2] For instance, the Dutch word *kuip*, 'barrel', is a low-frequency type in the present-day language, due to the fact that its denotatum has almost completely dropped out of use. Nevertheless, it was a high-frequency word in earlier centuries, to which the high frequency of the surname *kuiper* bears witness.

[3] The new types entering the distribution at rate α were generated by means of the transition matrix of figure 1B.

sumption that does not lend itself to implementation in a stochastic process. Without this assumption, rank-frequency distributions are generated that depart significantly from the empirical rank-frequency curve, the highest frequency words attracting a very large proportion of all tokens. By replacing Simon's assumptions 1 and 3 by the 'rule of usage' that

> the probability that the $(N+1)$-st word is a type that has appeared exactly r times is proportional to

$$H_r := -\frac{rn_r}{\sum_r rn_r} \log\left(\frac{rn_r}{\sum_r rn_r}\right), \qquad (6)$$

theoretical rank-frequency distributions of the desired form can be obtained. Writing

$$p(r) := \frac{rn_r}{\sum_r rn_r}$$

for the probability of re-using any type that has been used r times before, H_r can be interpreted as the contribution of all types with frequency r to the total entropy H of the distribution of ranks r, i.e. to the average amount of information

$$H = \sum_r -p(r)\log p(r).$$

Selection of ranks according to (6) rather than proportional to rn_r (Simon's assumption 1) ensures that the highest ranks r have lowered probabilities of being sampled, at the same time slightly raising the probabilities of the intermediate ranks r. For instance, the 58 highest ranks of the distribution of figure 1C have somewhat raised, the complementary 212 ranks somewhat lowered probability of being sampled. The advantage of using (6) is that unnatural rank-frequency distributions in which a small number of types assume exceedingly high token frequencies are avoided.

The proposed rule of usage can be viewed as a means to obtain a better trade-off in the distribution between maximalization of information transmission and optimalization of the cost of coding the information. To see this, consider an individual word type y. In order to minimalize the cost of coding $C(y) = -\log(\Pr(y))$, high-frequency words should be re-used. Unfortunately, these high-frequency words have the lowest information content. However, it can be shown that maximalization of information transmission requires the re-use of the lowest frequency types (H_r is maximal for uniformly distributed $p(r)$). Thus we have two opposing requirements, which balance out in favor of a more intensive use of the lower and intermediate frequency ranges when selection of ranks is proportional to (6).

The 'rule of usage' (6) implies that higher frequency words contribute less to the average amount of information than might be expected on the basis of their relative sample frequencies. Interestingly, there is independent evidence for this prediction. It is well known that the higher-frequency types have more (shades of) meaning(s) than lower-frequency words (see e.g. Reder, Anderson and Bjork 1974, Paivio, Yuille and Madigan 1968). A larger number of meanings is correlated with increased contextual dependency for interpretation. Hence the amount of information contributed by such types out of context (under conditions of statistical independence) is less than what their relative sample frequencies suggest, exactly as modelled by our rule of usage.

Note that this semantic motivation for selection proportional to H_r makes it possible to avoid invoking external principles such as 'least effort' or 'optimal coding' in the mathematical definition of the model, principles that have been criticized as straining one's credulity (Miller 1957).[4]

FUNCTION WORDS

Up till now, we have focused on the modelling of monomorphemic Dutch words, to the exclusion of function words and morphologically complex words. One of the reasons for this approach concerns the way in which the shape of the rank-frequency curves differs substantially depending on which kinds of words are included in the distribution. As shown in figure 2, the curve of monomorphemic words without function words is highly convex. When function words are added, the head of the tail is straightened out, while the addition of complex words brings the tail of the distribution (more or less) in line with Zipf's law. Depending on what kind of distribution is being modelled, different criteria of adequacy have to be met.

Interestingly, function words, — articles, pronouns, conjunctions and prepositions, the so-called closed classes, among which we have also reckoned the auxiliary verbs — typically show up as the shortest and most frequent (Zipf) words in frequency distributions. In fact, they are found with raised frequencies in the the empirical rank-frequency distribution when compared with the curve of content words only, as shown in the first

[4] In this respect, Miller's (1957) alternative derivation of (2) in terms of random spacing is unconvincing in the light of the phonotactic constraints on word structure.

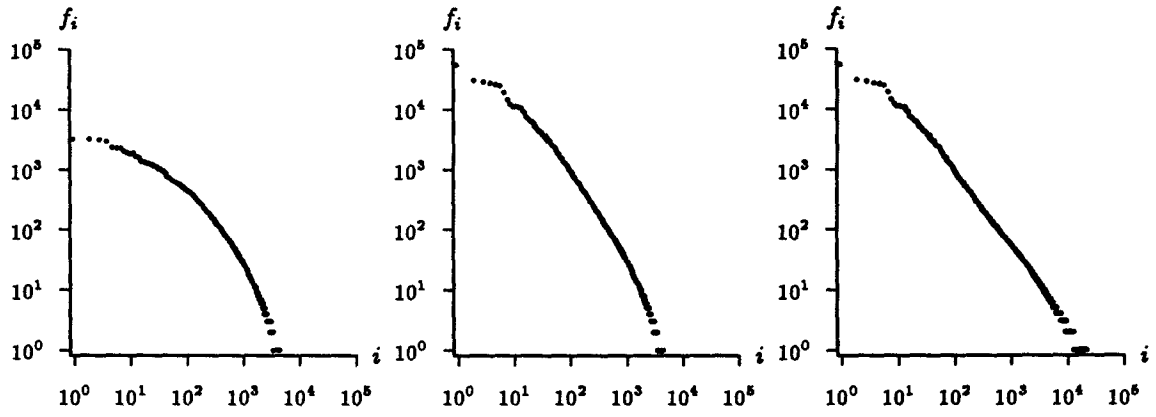


Figure 2: Rank-frequency plots for Dutch phonological stems. From left to right: monomorphemic words without function words, monomorphemic words and function words, complete distribution.

two graphs of figure 2. Miller, Newman & Friedman (1958), discussing the finding that the frequential characteristics of function words differ markedly from those of content words, argued that (1958:385)

> Inasmuch as the division into two classes of words was independent of the frequencies of the words, we might have expected it to simply divide the sample in half, each half retaining the statistical properties of the whole. Since this is clearly not the case, it is obvious that Mandelbrot's approach is incomplete. The general trends for all words combined seem to follow a stochastic pattern, but when we look at syntactic patterns, differences begin to appear which will require linguistic, rather than mere statistical, explanations.

In the Mandelbrot-Simon model developed here, neither the Markovian front end nor the proposed rule of usage are able to model the extremely high intensity of use of these function words correctly without unwished-for side effects on the distribution of content words. However, given that the semantics of function words are not subject to the loss of specificity that characterizes high-frequency content words, function words are not subject to selection proportional to H_r. Instead, some form of selection proportional to rn_r probably is more appropriate here.

MORPHOLOGY

The Mandelbrot-Simon model has a single parameter α that allows new words to enter the distribution. Since the present theory is of a phonological rather than a morphological nature, this parameter models the (occasional) appearance of new simplex words in the language only, and cannot be used to model the influx of morphologically complex words.

First, morphological word formation processes may give rise to consonant clusters that are permitted when they span morpheme boundaries, but that are inadmissible within single morphemes. This difference in phonotactic patterning within and across morphemes already reveales that morphologically complex words have a different source than monomorphemic words.

Second, each word formation process, whether compounding or affixation of suffixes like *-ness* and *-ity*, is characterized by its own degree of productivity. Quantitatively, differences in the degree of productivity amount to differences in the birth rates at which complex words appear in the vocabulary. Typically, such birth rates, which can be expressed as $\frac{E[n_1']}{N'}$, where n_1' and N' denote the number of types occurring once only and the number of tokens of the frequency distributions of the corresponding morphological categories (Baayen 1989), assume values that are significantly higher that the birth rate α of monomorphemic words. Hence it is impossible to model the complete lexical distribution without a worked-out morphological component that specifies the word formation processes of the language and their degrees of productivity.

While actual modelling of the complete distribution is beyond the scope of the present paper, we may note that the addition of birth rates for word formation processes to the model, necessitated by the additional large numbers of rare

words that appear in the complete distribution, ties in nicely with the fact that the frequency distributions of productive morphological categories are prototypical LNRE distributions, for which the large values for the numbers of types occurring once or twice only are characteristic.

With respect to the effect of morphological structure on the lexical similarity effects, we finally note that in the empirical data the longer word lengths show up with sharply diminished neighborhood density. However, it appears that those longer words which do have neighbors are morphologically complex. Morphological structure raises lexical density where the phonotaxis fails to do so: for long monomorphemic words the huge space of possible word types is sampled too sparcely for the lexical similarity effects to emerge.

REFERENCES

Baayen, R.H. 1989. *A Corpus-Based Approach to Morphological Productivity. Statistical Analysis and Psycholinguistic Interpretation.* Diss. Vrije Universiteit, Amsterdam.

Carroll, J.B. 1967. On Sampling from a Lognormal Model of Word Frequency Distribution. In: *H.Kučera & W.N.Francis 1967*, 406–424.

Carroll, J.B. 1969. A Rationale for an Asymptotic Lognormal Form of Word Frequency Distributions. *Research Bulletin — Educational Testing Service*, Princeton, November 1969.

Chitašvili, R.J. & Khmaladze, E.V. 1989. Statistical Analysis of Large Number of Rare Events and Related Problems. *Transactions of the Tbilisi Mathematical Institute.*

Good, I.J. 1953. The population frequencies of species and the estimation of population parameters, *Biometrika* 43, 45–63.

Herdan, G. 1960. *Type-token Mathematics*, The Hague, Mouton.

Kučera, H. & Francis, W.N. 1967. *Computational Analysis of Present-Day American English.* Providence: Brown University Press.

Landauer, T.K. & Streeter, L.A. 1973. Structural differences between common and rare words: failure of equivalence assumptions for theories of word recognition, *Journal of Verbal Learning and Verbal Behavior* 12, 119–131.

Mandelbrot, B. 1953. An informational theory of the statistical structure of language, in: W.Jackson (ed.), *Communication Theory*, Butterworths.

Mandelbrot, B. 1962. On the theory of word frequencies and on related Markovian models of discourse, in: R.Jakobson, *Structure of Language and its Mathematical Aspects. Proceedings of Symposia in Applied Mathematics* Vol XII, Providence, Rhode Island, Americal Mathematical Society, 190–219.

Miller, G.A. 1954. Communication, *Annual Review of Psychology* 5, 401–420.

Miller, G.A. 1957. Some effects of intermittent silence, *The American Journal of Psychology* 52, 311–314.

Miller, G.A., Newman, E.B. & Friedman, E.A. 1958. Length-Frequency Statistics for Written English, *Information and control* 1, 370–389.

Muller, Ch. 1979. Du nouveau sur les distributions lexicales: la formule de Waring-Herdan. In: Ch. Muller, *Langue Française et Linguistique Quantitative.* Genève: Slatkine, 177-195.

Nusbaum, H.C. 1985. A stochastic account of the relationship between lexical density and word frequency, *Research on Speech Perception Report # 11*, Indiana University.

Orlov, J.K. & Chitashvili, R.Y. 1983. Generalized Z-distribution generating the well-known 'rank-distributions', *Bulletin of the Academy of Sciences, Georgia* 110.2, 269–272.

Paivio, A., Yuille, J.C. & Madigan, S. 1968. Concreteness, Imagery and Meaningfulness Values for 925 Nouns. *Journal of Experimental Psychology Monograph* 76, I, Pt. 2.

Reder, L.M., Anderson, J.R. & Bjork, R.A. 1974. A Semantic Interpretation of Encoding Specificity. *Journal of Experimental Psychology* 102: 648–656.

Rouault, A. 1978. Loi de Zipf et sources markoviennes, *Ann. Inst. H.Poincare* 14, 169-188.

Sichel, H.S. 1975. On a Distribution Law for Word Frequencies. *Journal of the American Statistical Association* 70, 542–547.

Simon, H.A. 1955. On a class of skew distribution functions, *Biometrika* 42, 435–440.

Simon, H.A. 1960. Some further notes on a class of skew distribution functions, *Information and Control* 3, 80–88.

Zipf, G.K. 1935. *The Psycho-Biology of Language*, Boston, Houghton Mifflin.