REPRINTED FROM:

# COGNITION

International Journal of
Cognitive Science

Volume 43, No. 1, 1992

# Cross-linguistic regularities in the frequency of number words*

Stanislas Dehaene and Jacques Mehler

*INSERM, CNRS and EHESS, Laboratoire de Sciences Cognitives et Psycholinguistique, 54 Bd Raspail, F-75270 Paris Cédex 06, France*

pp. 1–29



ELSEVIER

*Cognition*, 43 (1992) 1–29

# Cross-linguistic regularities in the frequency of number words*

Stanislas Dehaene and Jacques Mehler
*INSERM, CNRS and EHESS, Laboratoire de Sciences Cognitives et Psycholinguistique, 54 Bd Raspail, F-75270 Paris Cédex 06, France*

Received February 20, 1990, final revision accepted July 29, 1991

*Abstract*

Dehaene, S., and Mehler J., 1992. Cross-linguistic regularities in the frequency of number words. Cognition, 43: 1–29.

*We examine the frequency of numerals and ordinals in seven different languages and/or cultures. Many cross-cultural and cross-linguistic patterns are identified. The most striking is a decrease of frequency with numerical magnitude, with local increases for reference numerals such as 10, 12, 15, 20, 50 or 100. Four explanations are considered for this effect: sampling artifacts, notational regularities, environmental biases and psychological limitations on number representations. The psychological explanation, which appeals to a Fechnerian encoding of numerical magnitudes and to the existence of numerical points of reference, accounts for most of the data. Our finding also has practical importance since it reveals the frequent confound of two experimental variables: numerical magnitude and numeral frequency.*

## Introduction

Cognitive psychologists often tacitly endorse the stable state hypothesis that all

*Correspondence to:* S. Dehaene, INSERM, CNRS and EHESS, Laboratoire de Sciences Cognitives et Psycholinguistique, 54 Bd Raspail, F-75270 Paris Cédex 06, France.

*The frequency data were kindly communicated to us by G. Hatano (Japanese), P. Karanth (Kannada), W.J.M. Levelt (Dutch) and N. Sebastian (Catalan and Spanish). D. Rioux provided us with two issues of the French newspaper *Le Monde* on a computer disk, from which we could compute the frequencies for French in arabic notation. We also acknowledge fruitful discussions with G. Altman, S. Bossini, L. Cohen, E. Dupoux and J. Segui, and extensive commentaries by S. Sokol and one anonymous reviewer. Requests for reprints should be sent to Stanislas Dehaene, Laboratoire de Sciences Cognitives et Psycholinguistique, 54 Bd Raspail, F-75270 Paris Cédex 06, France.
Editor's note: the editorial review process for this article was handled by Steve Pinker.

adult subjects possess, to a first approximation, a similar competence in the domain of study (Chomsky, 1965; Mehler & Bever, 1968). The hypothesis appears plausible in domains that seem largely innate or that are acquired through an innately guided learning process (Fodor, 1983; Gould & Marler, 1987; Jusczyk & Bertoncini, 1988). In humans, the paradigmatic example of the latter is speech acquisition – though even in that case the adult stable state for phonological awareness seems radically different in literates and illiterates (see Bertelson, 1987, and references therein). Conversely, the stable state hypothesis seems less viable for domains which, like chess playing, are usually acquired slowly, require formal teaching, and can be mastered to varying degrees or, for the present matter, not mastered at all.

Is the stable state hypothesis true in mathematics? Certainly, a considerable diversity exists in adult mathematical performance, from illiterates up to Fields medal winners. Not everybody is able to add two seven-digit numbers mentally, and the strategies that would allow for such a performance are likely to vary widely from individual to individual, for example depending on memory span or training with the abacus (Hatano & Osawa, 1983; Stigler, 1984). Even for additions of one-digit numbers, Siegler (1987) has drawn attention to the wide diversity of strategies that children may use; in such cases, averaging data from different subjects may result in a severe distortion of reality.

Given that mathematical notations and calculation procedures are the result of cultural evolution (Ifrah, 1981; Menninger, 1969), it is not surprising that individual variations are found at the level of abstract mathematical knowledge, or even in the application of arithmetical procedures. Yet at a lower level, a minimal set of principles of elementary arithmetics might be universal, possibly regardless of culture, language or level of education. Gallistel and Gelman (1991) postulate that all humans are endowed with a preverbal system of counting and arithmetic reasoning that is shared with a broad range of animals. This view is supported by the existence of pre-numerical discrimination capacities in the newborn infant (e.g., Starkey & Cooper, 1980) as well as in rats or in pigeons (Gallistel, 1990). Adults also show remarkably stable and reproducible effects in some simple number-processing tasks. For example, in numerical comparison, the distance effect – comparison times decrease with increasing distance between the operands – is found with subjects from different linguistic communities (Dehaene, Dupoux, & Mehler, 1990), resists extensive training, and is already present in 6-year-old children (Sekuler & Mierkiewicz, 1977; Duncan & McFarland, 1980).

The case for a stable state in elementary arithmetics would be strengthened if it were possible to demonstrate similar uses of elementary number concepts in many different cultures and linguistic communities. Here we report that the frequencies of number words in spoken or written language are strikingly similar across cultures. We use this fact to analyse number representations.

That number word frequency may provide valuable information about the organization of number concepts is suggested by numerous studies showing the importance of frequency of occurrence in the structuring of mental representations. Word frequency is a primary parameter along which input and output verbal lexicons appear to be organized (Forster & Chambers, 1973; Rubenstein, Garfield, & Millikan, 1970). The face recognition system is also drastically influenced by the relative presentation rates of, for instance, Caucasian versus Asiatic faces (e.g., Shepherd, 1981). Even at the neuronal level, cortical maps show dramatic expansion for biologically relevant and frequently encountered parameter values, for instance the range of frequencies relevant to echo location in the bat (Suga, 1982). Somatosensory maps have been shown to locally expand or contract in the adult monkey depending on the relative frequency of stimulation of afferents (Merzenich, 1987). All these examples document the influence of item frequency in the structuring of internal representations. The converse case – constraints on mental representations affecting, for instance, word frequencies – is rarely considered and will be addressed in this paper.

Before we describe the data, a word on vocabulary. We shall use the term *numeral* for a word referring to a specific numerical quantity, and *ordinal* for a word denoting rank order. Two systems of numerals are distinguished: *arabic numerals*, composed of *digits*, and *verbal numerals*, which may be *spoken* or *written* using language-specific notations (e.g., alphabetic notation for French; Kanji for Japan), and may be decomposed into elementary *number words*.

## Method

We compared the frequency of usage of numerals and ordinals in seven different languages: American English, Catalan, Dutch, French, Japanese, Kannada (a Dravidian language of South India) and Spanish. The following categories of number words were examined: (1) number words from "zero" to "nine"; (2) number words from "ten" to "nineteen"; (3) decade number words from "ten" to "ninety"; (4) number words for powers of ten ("one", "ten", "hundred", "thousand", "million", "billion"); (5) ordinals from "first" to "ninth"; and (6) ordinals from "tenth" to "nineteenth". The categories 1, 2, 3 and 4 include the ones, teens, tens and multiplier classes or stacks which, as neuropsychological evidence suggests, play dissociable roles in number writing and reading (Deloche & Seron, 1984; McCloskey, Sokol, & Goodman, 1986).

The number word frequencies were taken from standard databases (see Table 1 for references). Some number words were omitted because they did not exist in the language (e.g., teens for Japanese), or were not available in the database (e.g., hyphenated words like "soixante-dix" in French), or because word counts were too small to be meaningful. Total word counts were used for synonyms (e.g., "second" and "deuxième" in French) or declined words (e.g., "u", "un",

Table 1. *Summary of languages and databases analysed*

| Language | Corpus size | Source |
|---|---|---|
| American English | 1,014,000 | Francis and Kucera (1982) |
| Catalan | 258,771 | Conesa et al. (1983) |
| Dutch | | |
| All instances | 44,000,000 | CELEX |
| Numerals only | 44,000,000 | CELEX |
| French | | |
| Written number words | 37,653,685 | Imbs (1971) |
| Spoken number words | 312,135 | Gougenheim et al. (1956) |
| Arabic numerals | 86,843 | *Le Monde*, 2nd and 3rd July 1990 |
| Japanese | | |
| Kanji | 2,000,000 | NLRI (1970) |
| Arabic numerals | 2,000,000 | NLRI (1970) |
| Kannada | 100,000 | Ranganatha (1982) |
| Spanish | 500,000 | Juilland and Chang-Rodriguez (1964) |

"una" in Catalan). In French, separate word counts were obtained for written and for spoken number words. The Japanese word counts were for written Kanji notation only. Finally the Dutch and American English databases provided separate counts for the different grammatical uses of words. For both languages, word counts were taken only from the "cardinal numeral" and "ordinal numeral" categories (excluding, for example, the adverbial meaning of "first" in English). Additionally in Dutch we also compiled total word counts, regardless of grammatical usage. These data, labelled "Dutch (all instances)" in the graphs, provide an estimate of the variation that this procedure introduced with respect to languages for which no grammatical classification was available.

We also tabulated the frequencies of occurrence of arabic numerals. These were available only for Japanese and for French. In both cases, only full arabic numerals, not digits, were counted. For instance "31" was counted as one occurrence of arabic numeral 31, not as one occurrence of 3 and one occurrence of 1. Note that for verbal notation, tables of word frequencies give only the frequency of number words (e.g., "twenty", "one"), never that of full verbal numerals such as "twenty-one". This is an important difference between the two systems of numerals: in verbal notation, only the component numbers words were counted, not the whole numerals; in arabic notation, only the whole numerals were counted, not their component digits.

In Japanese, the frequency of arabic numerals was available only for numerals 0–9 and the decades 10–90. For simplicity, these were displayed and analysed together with number words categories (1) and (3) defined above. In French, a frequency analysis program was designed which could evaluate the frequency of any arabic numeral. However, given the small size of our corpus, only the frequencies for numerals 0–20, 30, 40, 50, 60, 70, 80, 90 and 100 were judged reliable. Again, these were included in the categories (1), (2), (3) and (4) defined above.

## Results

The observed frequencies appear in Figure 1. An important cross-linguistic similarity is found regardless of geographic distribution, language and notation. For example, Japanese uses an extremely regular numeration system with no "teens" words; French has irregular words for 11–19, 70, 80 and 90; and Dutch inverts decades and units, saying "one and twenty" for 21. Despite these linguistic differences, the absolute frequencies of numerals and ordinals are almost superposable. Only within-language differences have a sizeable influence: numerals are about four times more frequent in spoken French than in written French, and, for Japanese and French data, in arabic notation than in verbal notation (Kanji or alphabetic).

For the numerals 1–9 and 10–90, as well as for the ordinals 1st to 9th, number word frequency decreases with numerical magnitude. This decreasing pattern is extremely robust and reproducible across languages. Other number word classes show non-decreasing but still cross-linguistically similar patterns. Thus, ordinals from 10th to 19th show a U-shaped curve, possibly reflecting a recency effect for names of centuries in the range 15th to 19th. The frequency of powers of ten is fairly constant, especially over the range $10–10^6$. The frequency of zero is much lower than that of the other numerals. Finally, for teens words 10–19, a decrease in frequency is observed only for American English and Dutch ($p < 0.05$) and is generally masked by two phenomena. First, the frequency of the numerals 12 and 15 is unexpectedly elevated, probably reflecting the continued usage of duodecimal and/or sexagesimal counting principles in some domains like months, days, hours or minutes (this is discussed below). Second, the frequency of the numeral 13 is low relative to 12 or 14. This might reflect the "devil's dozen" superstition, which assigns a maleficent power to number 13, to such a degree that there is no 13th floor in most American buildings! Note the absence of a drop for 13 in Kannada, a language of South India where no such superstition exists. The small size of the Kannada corpus, however, makes it difficult to reach firm conclusions on this minor point.

Figure 2 shows the frequency profiles over the range 1 to a billion (databases for which too many datapoints were missing are not shown). On this log-log scale, a straight-line decrease of frequency with magnitude is observed, at least over the range 1–9 (all $r^2 > 0.78$, $p < 0.01$). Thus, frequency is a power function of magnitude, with an average exponent of $-1.90$ for number words and $-0.87$ for arabic numerals. In addition to the decreasing trend, localized sharp increases are found which occur in all languages for the same numerals 10, 12, 15, 20, 50 and

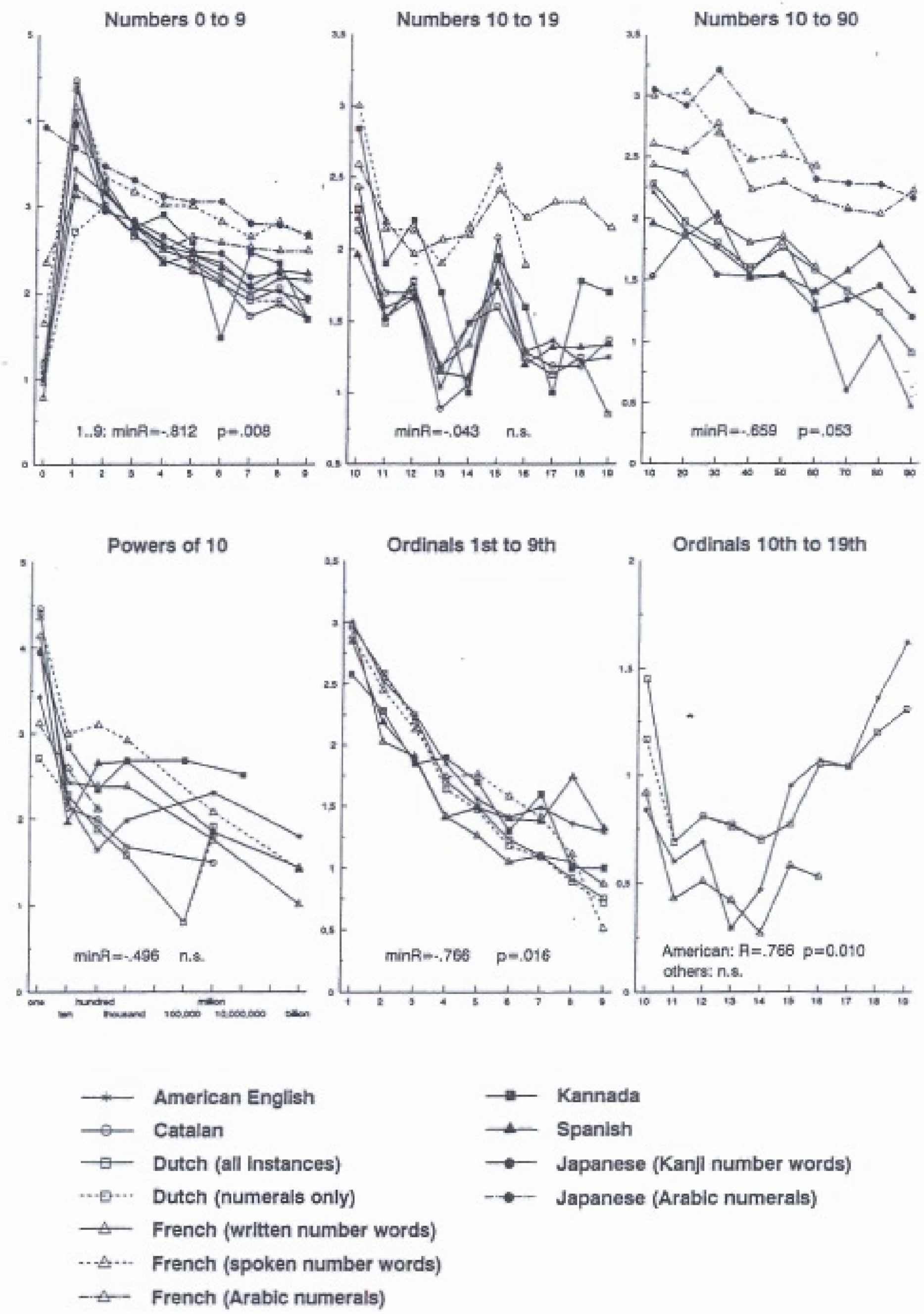


Figure 1. *$Log_{10}$ frequencies (occurrences per million) of numerals and ordinals. Unless otherwise stated, the frequency is for written number words. MinR = minimum correlation coefficient of any data set with the abscissa.*

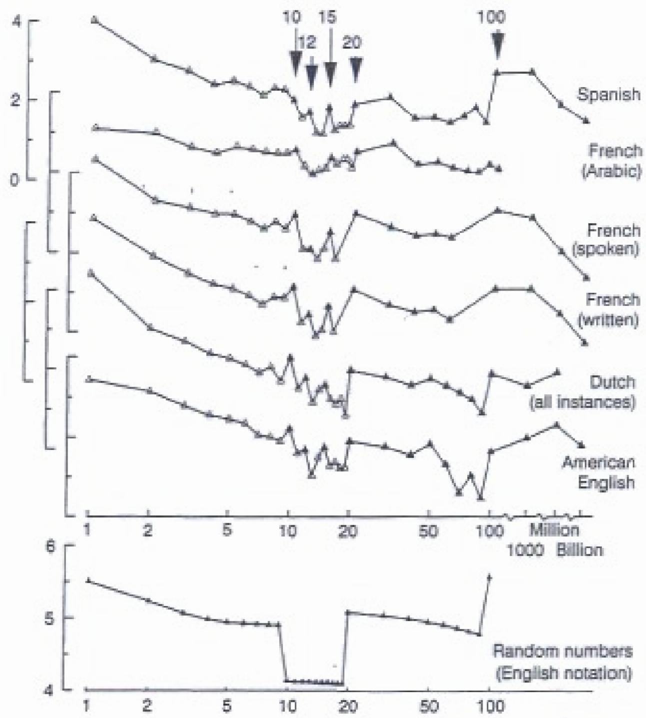


Figure 2. *Top: frequency profiles over the range 1 to a billion (log-log scale). To improve readability, the curves have been displaced by arbitrary amounts along the Y axis. Languages for which data points were missing have been omitted. Filled triangles: numerical points of reference. Open triangles: non-reference numerals. Bottom: frequency profile for random numbers drawn from an exponential distribution with relaxation constant 1/87 and written in English notation. Within-category decreases in frequency are observed, but the frequency of teens is largely underestimated.*

100. The frequency of numerals 30, 40, 60, 70, 80, 90, $10^3$, $10^6$ and $10^9$ is also higher than expected by a linear interpolation on the 1–9 range.

**Discussion**

The expression of a numeral or an ordinal is generally the end product of a complex chain of verbal production, the key elements of which are depicted in Figure 3. In the following four sections, we analyse the non-exclusive factors that might contribute to frequency variations. We focus mainly on the two strongest cross-cultural effects: the sharp decrease of frequency with numerical magnitude, and the local increases for numerals 10, 12, 15, 20, 50 and 100.

In section 1, sampling artifacts are considered. Strictly speaking, frequency tables only provide information about the occurrence of numerals and ordinals in

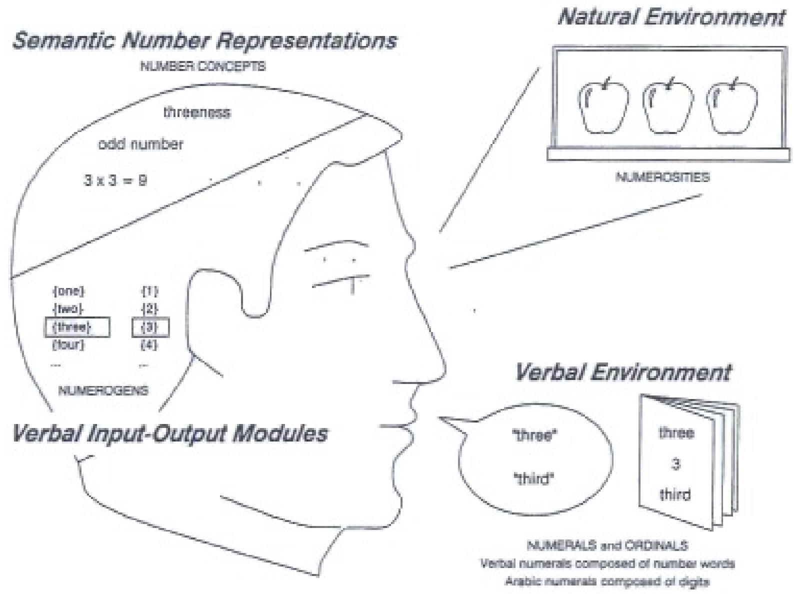


Figure 3. *Four components contributing to the production of a given numeral.*

speech and texts. At this level, biased sampling or notational conventions may contribute to the observed frequency patterns.

In section 2, we examine the possible role of notational regularities. Written or spoken number production engages specialized mental verbal input–output modules (McCloskey et al., 1986). The numerical representations that are accessed at this level, that we call *numerogens*, are notation-specific lexical tokens which serve as "mental tags" for numerals and ordinals. Possibly, there may exist regularities in number notation which imply that some numerogens, and consequently some numerals or ordinals, are activated more often than others when a number needs to be produced.

In section 3, we consider the effect of environmental factors. Possibly some numerical quantities (*numerosities*) occur more often than others in our natural environment, therefore biasing numeral production right from the beginning.

Finally, in section 4, psychological limitations are considered. Perhaps our environment itself is not biased, but our number representation scheme is limited to grasping only small or "round" quantities. Psychological constraints on *number concepts* (abstract mental representations of number which are accessed regardless of input or output number notation) would then account for the observed frequency patterns.

The very existence of genuine *number concepts* is currently debated. Some authors argue that all mental numerical processing is performed on notation-specific codes (i.e., at the level of numerogens) and may vary qualitatively depending on which particular notation is used for input or output (e.g., Campbell & Clark, 1988; Gonzalez & Kolers, 1982, 1987). Others postulate that numerical processing involves a single amodal abstract semantic representation, and that notation-specific modules are used only for input–output operations (e.g., McCloskey et al., 1986; Sokol, Goodman-Schulman, & McCloskey, 1989). It will therefore be important to determine if the observation of notation-independent frequency regularities necessarily implies the existence of abstract number concepts.

## *1. Sampling artifacts*

Might frequency variations arise solely from sampling problems or from conventions of number writing? It is common practice, for instance, to print small numbers in full words and large numbers in arabic notation. Hence, large number words would appear less often in word frequency databases. However, if this artifact had any sizeable influence, large arabic numerals should be *more* frequent than small ones; in fact both arabic numerals and written verbal number words follow decreasing frequency curves. Furthermore, the frequency decrease is found even for spoken verbal numerals, whereas to the best of our knowledge no conventions limit the usage of numerals in spoken French.

If they do not account for the bulk of the data, sampling problems certainly contribute to some of the frequency variations, particularly for numerals "zero" and "one". For instance, the frequency of zero in arabic notation is much lower in French than in Japanese, where it even exceeds the frequency of "1". However, the notation for zero was introduced in Japan about a hundred years ago. In contrast to other number words, the word "zero" is written in Katakana, not in Kanji. Conventions apparently limit the use of this Katakana notation. For instance in writing the date "1905", all digits appear in Kanji, except the zero which generally appears in arabic notation. This and similar conventions may well have resulted in an artificially inflated frequency for the arabic notation of zero in Japanese journals.[1]

Likewise, sampling problems certainly account for the high variability in the frequency of the number word "one". The Japanese Kanji for one, "ichi", is used only in a numerical context. In all other languages that we studied, the word for "one" is ambiguous and can also serve as an indefinite article (Catalan, Dutch,

[1] In principle, only full words, not individual digits, are counted in the Japanese database that we used. Hence, 0 should not have been counted in "1905". However, the automated analysis of Japanese texts faces a problem of segmentation: in journals, the Kanji symbols are aligned without indication of word boundaries. Segmentation algorithms exist that circumvent this problem, but we do not know how they handle words with the non-Kanji symbol "0". Most likely, this symbol was counted each time that it was encountered, whether in a word or not.

French, Spanish), a definite article (Kannada), a personal pronoun (Catalan, English, French, Spanish), etc. The influence of this ambiguity factor is clearly perceptible in Figure 1: languages with an ambiguous word for "one" (Catalan, French, Kannada, Spanish) show a 10- to 100-fold elevation in the frequency of this word, relative to languages with no such ambiguity (Japanese) or for which the ambiguity was lifted by a separation of grammatical uses in the database (American English, Dutch). In Dutch, where a direct comparison is available, usage of "een" in a numerical sense accounts for only 2% of the word counts for this word (this may actually represent an underestimation). The Dutch data also make it clear that this artifact affects only the number word "one". Almost all instances of number words other than "one" are classified as genuine numerical uses of these words.

The ambiguity artifact contributes much to the non-linearity of the frequency curves over the interval 1–9 in Figure 1. However, it does not explain the frequency decrease observed for number words 2–9, 10–90, and for ordinals. In Japanese, which is not contaminated by the "one" artifact, a linearly decreasing curve is still observed over the interval 1–9 on the log-linear plot of Figure 1. This implies that frequency decreases exponentially with numerical magnitude ($F(n) = 1860 \times 10^{-0.15n}$, $r^2 = 0.97$, $p < 0.001$). Alternatively, the Japanese data may also be fitted with a straight line on a log-log plot, implying that frequency decreases as a power function with exponent $-1.25$ ($r^2 = 0.96$, $p < 0.001$). Both mathematical representations of the data confirm the existence of a sharp frequency decrease, for which explanations other than sampling artifacts must be sought.

## 2. *Mathematical regularities in number notation*

The production of well-formed numerals obeys strict grammatical rules (Hurford, 1975) which may, as a side-effect, require the more frequent use of some particular number words or symbols, Benford's (1938) observations seem to support this possibility. Benford examined all kinds of numerical tables, for instance those giving the surface of American lakes, the street addresses of famous scientists, or the square roots of various integers. He found that, regardless of their origin, arabic numerals were about six times more likely to start with digit 1 than with digit 9 (Figure 4). About 31% of numerals started with digit 1, 19% with digit 2, 12% with digit 3, etc. The probability of finding digit $n$ in the first position decreased smoothly with $n$, according to the mathematical law $P(n) = \log_{10}(n+1) - \log_{10}(n)$. At first sight, this law bears considerable similarity to our findings. Could it be that both phenomena are different surface manifestations of a common underlying regularity?

In order to evaluate this possibility, a better understanding of the origins of Benford's law is required. Are small things more numerous than large things in

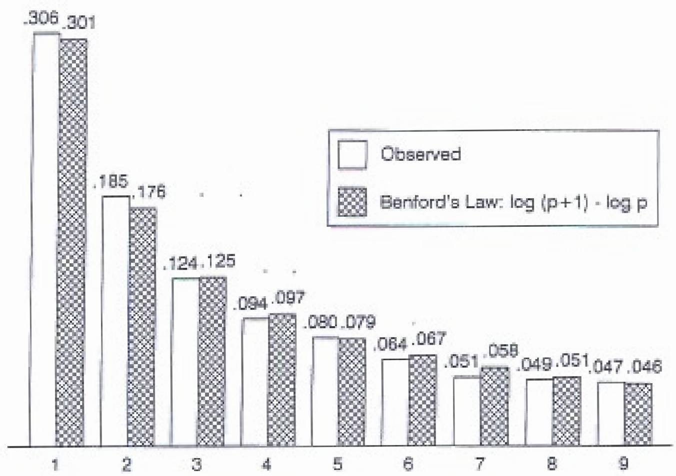


Figure 4. *The distribution of first digits in arabic numerals taken from various numerical tables (adapted from Benford, 1938).*

nature, as Benford (1938) originally suggested (p. 571)? Or does the law describe a more general mathematical regularity (see Raimi, 1969, and the mathematical theorems therein)? In support of the second proposal, we have found empirically that random numbers drawn from almost any smooth distribution with large standard deviation follow a decreasing law of first digits similar to Benford's law. If the random numbers are drawn from a monotonous decreasing distribution, this result is hardly surprising: if 1 is more frequent than 2, 10 more frequent than 20, etc., then 1 will appear more often than 2 in the first position of the random numbers. More importantly, the law continues to hold for non-monotonous distributions, such as a Gaussian. In this situation, qualitative arguments show that only the right-hand part of the distribution, which is necessarily decreasing, really counts. We refer the reader to Appendix A, where the mathematical origins of Benford's law are laid out in greater detail. Unfortunately, our intuitive grasp of the law is hardly increased by these mathematical arguments! At any rate, the fact that the law is verified with random numbers rules out an environmental explanation. The law appears to hold whenever numbers of any origin are written down in arabic notation. It is therefore a notational regularity, not an environmental one.

Benford's law applies only to first digits, however, not to whole arabic numerals or to number words. Extending the law to account for our observations is not a straightforward task. First, consider the case of spoken or written number words. A modified version of Benford's law does hold when we tabulate word frequencies for random numbers written in English notation, instead of first digits in arabic notation. We discovered this empirically using a computer program that draws a random number (e.g., 131), translates it into English notation ("one hundred thirty one"), and counts how frequently each number word like "one", "hundred" or "thirty" is found. The results are shown at the bottom of Figure 2, for direct comparison with the actual English data. Frequency decreases are found within each category of numerals (ones, teens and tens words). Frequency discontinuities are also found for 20 and 100. However, three results are in contradiction with the data: (1) the discontinuities for 12 and 15 are not predicted; (2) a discontinuity is predicted in the wrong direction for 10 (i.e., in random numbers, "ten" is actually *less* frequent than "nine"); and (3) the predicted frequency of teens words is much too low relative to the frequency of ones words; the real data actually show a continuous decrease of frequency from 1 to 19, if one excepts number words 10, 12 and 15.

Like the original Benford law, these phenomena are largely independent of the precise distribution chosen for the random numbers. In order to suppress the three contradictions above, a very specific distribution is needed for the numbers, namely a sharply decreasing one (e.g., a power function with exponent −2) and with peaks for numbers 10, 12, 15 and 50. Of course, we are then back to square one, having to explain why the numbers to be produced obey such a specific distribution.

Similarly, our data on the frequency of arabic numerals cannot be explained by Benford's law without making strong additional assumptions. We measured the frequency of full arabic numerals (e.g., "1", "12"), not that of individual digits (e.g., 1 in "12"). Hence, no notational bias is involved at all in the arabic data (except for zero in Japanese; see the above discussion). The arabic data directly measure the frequency with which a given *numerosity* is produced. And the fact is that this frequency is decreasing, with increases for 10, 12, 15, and the class of numerals 20, 30, 40 . . . 100. Generally speaking, the high similarity of the frequency curves for arabic numerals and for verbal number words in all languages suggests that the effects are not notation-specific, hence do not arise at the level of input–output modules.

To summarize this long argument, notational regularities akin to Benford's law do not begin to explain our data on arabic numerals. They can partly predict the data on verbal numerals, but only at the expense of ad hoc additional assumptions. Therefore notational regularities have a limited explanatory power. Nevertheless, they certainly contribute to some of the differences that can be observed between arabic and verbal notation. For instance the frequency decrease is sharper in verbal notation than in arabic notation (Figure 2). Probably in verbal notation the decrease is accentuated by Benford's law. Another difference is that the frequency of the number word "hundred" is high relative to that of "ninety", but in arabic notation "100" is less frequent than "90" (Figure 2). A likely explanation is that the word "hundred" belongs to the class of multiplier words in verbal notation (Hurford, 1975; Power & Longuet-Higgins, 1978). Therefore it is counted almost each time a number greater than 99 is produced (e.g., in "six *hundred* twenty"). "100" does not play such a role in the arabic notation system, hence its lower frequency.

### 3. *Environmental factors*

If small numbers are produced more often than large ones, as clearly attested by the frequency counts of arabic numerals, it might be because we encounter small numerosities more often than large ones in our natural environment. For instance, the distribution of the number of children in modern families is naturally biased towards small numerosities. Environmental factors may therefore predict a monotonous decrease of numeral frequency with magnitude. It might further be argued that the numerosities 10, 12, 15, 20, 50 or 100 are also more frequent than their neighbours in our man-made environment. For instance, eggs are sold by the dozen, and nails by the hundred. We often encounter such round numerosities, and this may explain the elevated frequency with which we use the words that describe them.

Not all number productions are meant to describe a physical environment, though. Would the frequency regularities still obtain under experimentally controlled conditions in which environmental constraints would play little or no role? Baird, Lewis, and Romer (1970) asked subjects to estimate numerically the ratio of two visually presented lengths, areas or distances. They found that some numerals were more frequently produced than others (e.g., 5, 10, 12, 15, 20, 100), even though the physical stimuli that were presented did not privilege those round ratios. Even more compelling is a study by Baird and Noma (1975; Noma & Baird, 1975), who had subjects freely produce, within a given interval, integers that "they thought other people would produce under the same conditions". The subjects' responses departed widely from randomness, as some numbers were much more likely to be produced than others. We reproduce in Figure 5 the frequencies obtained for the interval [1 . . . 100]. The data are astonishingly similar to those we have compiled, with a slow decrease of frequency from 1 to 19, and peaks for particular numbers such as the multiples of 10 or the numbers 5, 15, 25 and 75.

In these experiments, no direct environmental or notational artifact is involved. However, it might still be argued that Baird and Noma's (1975) adult subjects had a long history of interactions with their environment, and that therefore the ultimate cause of frequency patterns is environmental rather than psychological. One possible example of such a long-term influence of nature on nurture is the widespread use of base ten in numerical notation and calculation, which may be attributed in part to the accidental fact that we have 10 fingers. Could frequency patterns be traced back to a similar remote environmental source?

The distribution of number-prefixed words can be used to demonstrate that, in

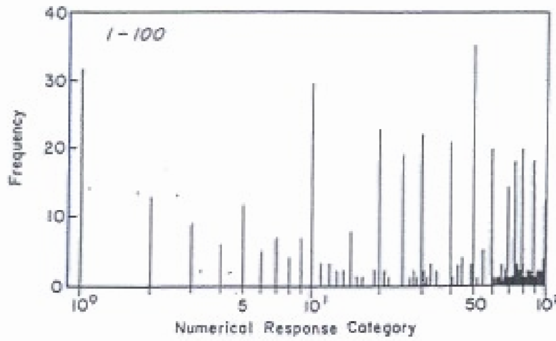


Figure 5. *The frequency of numerals in free production (from Noma & Baird, 1975). Subjects had to produce "numbers that they thought other people would produce under the same conditions".*

fact, the presence of environmental biases is neither necessary nor sufficient to account for linguistic biases. Several words have a prefix which specifies numerosity (e.g., binoculars, triangle). Just like the word "two" is more frequent that "three", words starting with prefix bi-, di- or duo- are more numerous than words starting with prefix tri- in English or French dictionaries (their frequency is generally too low to be meaningful). This pattern could be attributed again to environmental bias, but there are domains in which such an explanation can be rejected. For instance, Table 2 lists the prefixed words for temporal duration or repetition (e.g., bimonthly) in English and in French. Of course environmental events are not more likely to repeat every two months than every seven months.

Table 2. *Prefixed words for temporal intervals in English and in French*

| 2 | 3 | 4 | 5 |
|---|---|---|---|
| biannual, -ly | triennial, -ly | quadrennial, -ly | quinquennial, -ly |
| bicentenary | triennium | quadrennium | quinquennium |
| bicentennial | trimester | quadricentennial | |
| biennial, -ly | trimonthly | quartan | |
| biennium | triweekly | quatercentenary | |
| bimillenary | | | |
| bimonthly | | | |
| biweekly | | | |
| biyearly | | | |
| diestrus, -ous, -ual | | | |
| bicentenaire | tricentennaire | quadriennal | quinquennal |
| biennal | triennal | quadrimestre | quinquennat |
| bihebdomadaire | trimestre, -iel | quatre-temps | |
| bimensuel | trisannuel | | |
| bimestre, -iel | | | |
| bimillénaire | | | |
| bisannuel | | | |

Nevertheless, there are again more words for small temporal intervals than for large ones.

This example shows that a lexical bias for small numerosities can emerge in the absence of any environmental bias. Conversely, cases can be found where clear environmental biases fail to be incorporated into the lexicon. There are obviously more vehicles with four wheels than with two wheels, but we have a number-prefixed word only for the latter (bicycle), not for the former (quadricycle?). Numerical regularities in the environment seem to be incorporated in language only if they concern a small enough numerosity. For instance, we have number-prefixed words for plants with three leaves (e.g., trifoliate, trifolium; trèfle in French), but not for many other plants or flowers with a fixed but large number of leaves or petals. Words like "octopus", which explicitly refer to a relatively large numerosity, are rare. As a final example, *Scolopendra morsitans*, an arthropod with 21 body segments and 42 legs, is commonly called a "mille-pattes" (1000 legs) in French and a centipede (100 feet) in English! Clearly, we pay attention to numerical regularities of nature only inasmuch as they fit with our cognitive apparatus, which seems biased towards small or round numerosities.

In summary, it seems to us that the environmental explanation, while difficult to refute, only postpones the problem without solving it. Strictly speaking, small numerosities are not more frequent than large ones in the environment. Objectively, in any given situation, there is a very large number of things to be perceived and counted. Why do we parcel the world into small sets, rather than perceiving a single group of large numerosity? Why do we perceive a rabbit and not a set of body parts (Quine, 1960)? The impression that the world is essentially composed of small numerosities is an illusion imposed by our perceptual and/or cognitive apparatus, and therefore calls for a psychological explanation.

## 4. *Psychological factors*

According to a psychological explanation, the frequency variations that we have observed reveal the structure and the limitations of our cognitive number-processing system. Three psychological constraints which may relate to the frequency data are considered in turn: (1) the fact that we count; (2) the limitations on our apprehension of numerosity; and (3) our faculty for approximation.

### 4.1 *Counting*

Counting is our fundamental method for assessing numerosity. The very fact that we count may have drastic consequences on the structure of our mental number representation systems. If one counts aloud, then the word "one" is used for every count, "two" is used only for sets of more than two objects, "three" for sets of more than three objects, etc. Therefore, counting offers a ready explanation for why the word "one" is used more often than "two", "three", etc.

This explanation of the frequency decrease works well for situations in which the whole sequence of counting tags is verbalized. This is often the case for ordinals: it is rare to talk of the sixth instance of something without mentioning before the fifth, fourth, etc. However, as far as numerals are concerned, counting aloud is rare in adult behaviour. We often verbalize only the final tag, which gives the cardinal of the counted set. Yet even in that case, the decreasing frequency of numerals may arise from a subtle consequence of counting. Simply because we label things by counting from one up, small numbers are used more often than large ones for labelling things. Consider the example of street numbers. In one street the numbers may go from 1 to 32, in another from 1 to 561, and in yet another from 1 to 813. Overall there are much more houses bearing number 1 or 2 than bearing number 32 or 561. Generally speaking, because we count, we preferentially use numerical intervals whose lower bound is 1 (e.g., hours (1–12), minutes (1–60), hotel rooms, chapters in books . . .). Hence the inflated frequency of small numerals (see also Flehinger, cited in Raimi, 1969).

Despite its seducing simplicity, this explanation has some problems too. First, still no account is provided for the local increases in frequency within the generally decreasing frequency profile (but see below). Second, the frequency decreases predicted by counting alone are too small. Only a linear decrease is predicted if one assumes that we are equally likely, over some interval, to count from 1 up to any number $n$. This is a small decrease in comparison to the power function observed in the data, with an exponent of $-0.87$ for arabic numerals, $-1.25$ for Japanese Kanji, or about $-2$ for other languages (possibly contaminated by a "one" artifact). Surprisingly, in order to obtain an $n^{\alpha}$ frequency decrease, one must suppose that the upper bounds $n$ for the counting process themselves obey an $n^{\alpha-1}$ power law.[2] When the exponent $\alpha$ is $-1$, this hypothesis

[2] In counting, the following numerals are produced:

1 if only one object is counted
1 2 if two objects are counted
1 2 3 if three objects are counted, etc.

If all such counting sequences are considered equiprobable up to some number $N$, then the frequency of numeral $n$ $(0 < n \leq N)$ is

$$F(n) = (N - n + 1)/K, \quad \text{with constant } K = \tfrac{1}{2}N(N+1) \tag{1}$$

Therefore frequency should decrease linearly with $n$. More generally if $p([1 \ldots i])$ is the probability of counting from 1 up to $i$, then the frequency of numeral $n$ becomes

$$F(n) = \frac{1}{K} \sum_{i=n}^{i=\infty} p([1 \ldots i]) \quad \text{for } n > 0 \tag{2}$$

amounts to saying that we count from 1 to 3 four times more often than we count from 1 to 6. Hence, even when biases induced by the counting procedure are taken care of, there still remains an unexplained preference for using small numerical intervals rather than large ones. The fact that we count certainly contributes to the observed frequency variations, but it offers only an incomplete explanation for them.

### *4.2 Limitations on the apprehension of numerosities*

Several experiments have revealed limitations in our grasping of numerical concepts, which may well induce a preference for manipulating small numerosities rather than large ones. First, the process of quantification is easier with small numerosities than with large ones. For sets of less than three or four objects, a fast and accurate quantifying procedure called *subitizing* seems to operate. Conversely, with increasing numerosity, the counting procedure becomes slow and people resort to an inaccurate estimation procedure (Mandler & Shebo, 1982).

Evidence for the privileged status of small quantities in mental processing is not limited to the quantification of physical numerosities, but extends to experiments of comprehension or production of numerals (see Krueger, 1989, for review). For equal numerical distances, two small numbers are judged less similar than two large ones (Shepard, Kilpatrick, & Cunningham, 1975). Small numbers are also processed faster than large numbers in a larger–smaller number comparison task (Buckley & Gillman, 1974; Dehaene, 1989). These and other experiments (e.g., Rule, 1969) indicate that the comprehension of small numerals is faster and/or more refined than the comprehension of larger numerals. Numeral production is similarly biased. When asked to produce random numbers, subjects actually produce more small numbers than large ones (Banks & Hill, 1974). Contrary to the case of subitizing, no sharp limit between small and large numbers is found in these experiments. Rather, the effects are monotonous and continuous over the range of numbers tested.

Taken as a whole, these experiments suggest that Fechner's law holds for the central representation of numerical quantities (Krueger, 1989): small numbers

---

where the constant $K$ is given by

$$K = \sum_{i=1}^{i=\infty} i \sqcup p([1 \ldots i]) \tag{3}$$

Equation (2) is easily transformed into the equivalent form

$$p([1 \ldots n]) = K[F(n) - F(n+1)] \tag{4}$$

which shows that the values of $p([1 \ldots n])$ can be calculated from the measured frequencies $F(n)$. For $F(n)$ equal to $n^{\alpha}$, we obtain that $p([1 \ldots n])$ should be approximately proportional to $\alpha n^{\alpha-1}$.

receive an expanded and more accurate mental representation relative to large numbers. This may explain the biased frequency distribution of numerals. Assuming that the frequency of production of numerals is directly related to the importance of the associated mental number representation or number concept, a decrease in frequency with magnitude is predicted. The local increases in frequency, which are not immediately explained by Fechnerian encoding, are considered below.

### *4.3 Approximation and reference numbers*

The studies cited above have shown that the precision with which we can quantify a set decreases rapidly for large numerosities (see also Buckley & Gillman, 1974). By necessity, then, we often use numerals only to denote approximation. In some cases, however, we do want to convey a sense of exact numerosity, for instance because we have attained a precise knowledge of it by counting. This poses a problem for verbal communication: when is the listener to decide that a given numeral is meant literally, or that it gives only an approximation?

Natural languages offer several solutions to resolve this ambiguity. Thus, words like "exactly" or "about" may be used for disambiguation. In addition, some numerals are conventionally used only for precise denotation of numerosity, whereas others may be used to convey a sense of approximation. Thus the sentence "there are $x$ students in the room" may mean that the number of students is exactly $x$ or approximately $x$, depending on whether $x$ is 19 or 20. Similarly, in French or in German it is natural to speak of "fifteen days" to mean "two weeks", even though the rigorous number of days is 14. Rosch (1975) showed that sentences such as "9 is approximately 10" are preferred over the reversed sentences ("10 is approximately 9"). She concluded that numbers like 10 or 20 function as cognitive reference points in the numerical domain. Following her terminology, we shall term "reference numerals" the numerals that can be used for approximation. Reference numerals may span a whole region of numerical space, in much the same way that the words "possible" or "likely" refer to whole ranges of probability (Jaffe-Katz, Budescu, & Wallsten, 1989).

Let us denote by "numerical span" the range of numerosities that a given numeral may approximate. For example, the span of "nineteen" is 1, since this word denotes the exact quantity 19; the span of "twenty" is about 10, because this word may be used to approximate quantities in the range 15–25. Reference numerals, which have large numerical spans, may be used in many more contexts than numerals that can only be used for precise denomination. Therefore the frequency of reference numerals should be elevated. This explains the local increases in frequency for 10, 12, 15, 20, 50 or 100.

We may only speculate on why the same reference numerals appear in all languages and cultures. These numbers are either multiples of powers of the base

(base 10) or integer divisors of 60. Base 6 and base 60 were introduced by the Babylonians, and they are still privileged today in some domains (hours, minutes, seconds, angular degrees, eggs, etc.). Given a base system like 6, 10 or 60, the reference numerals 10, 12, 15, 20, 50 and 100 may be easier to write, memorize, or calculate with. The choice of a particular base itself is probably governed by converging constraints, some mathematical (a good base should have many divisors), some psychological (memory limitations require the choice of a small base number) and some anecdotal (we have 10 fingers) (see, for example, Ifrah, 1981).

### *4.4 A model of the mental "number line"*

The above arguments suggest that two supposedly universal properties of mental semantic number representations are responsible for the observed frequency patterns:

(1) The internal representation of numbers is compressive and akin to Fechner's law, that is, it is more detailed for small numbers than for large ones.
(2) Some numerals, called reference numerals, are used not only to refer to precise numerosities, but also to approximate wide ranges of numerosities.

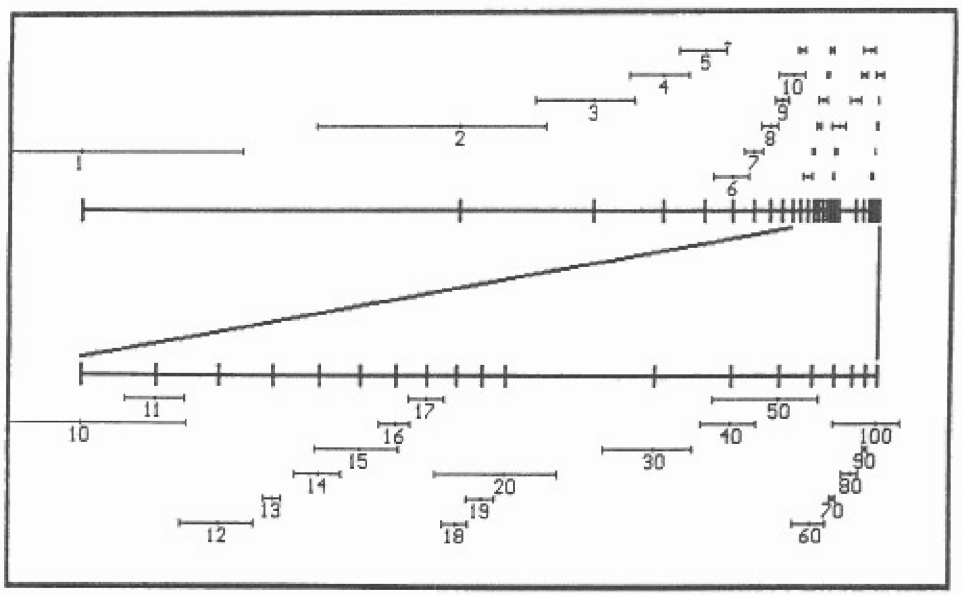


Figure 6. *Number line and numerical spans reconstructed from the American English frequency data (similar representations were obtained for the other data sets). The lower line is an expanded portion of the upper full number line. Segment lengths are proportional to numeral frequency and represent each number's span. Numerical spans are larger for reference numerals such as 10, 12, 15 or 20. See Appendix B for mathematical details.*

If this view is correct, then the frequency of numerals can be used as an indicator of the mental organization of number concepts. Accordingly, we have designed a method that, given numeral frequency data, is able to reconstruct a one-dimensional sketch of the putative mental number space or "number line" (see Appendix B for mathematical details). The method reconstructs both the psychophysical function that maps numerosities onto the number line, and the numerical spans of each numeral, that is, the portion of the number line that each numeral may adequately refer to.

The mental number line reconstructed by this method (Figure 6) shows a Fechnerian compression of internal distances. Furthermore, reference numerals are found to cover larger portions of the number line than other non-reference numerals. Even the quantitative results conform to our intuitions. The numeral "ten" covers the line from about 9 to 11. "Fifteen" spans from 14 to 16. Decade names from "twenty" to "ninety" each cover about their own decade, with little overlap. Only "fifty" spans a slightly larger interval. Finally "hundred" covers an even larger numerical region from about 70 to 130. It will be interesting to see if this mental representation for numbers can be confirmed using on-line experimental methods.

## Conclusion

We have observed recurring cross-linguistic regularities in the frequency of number words. The most striking pattern is a regular decrease of frequency with numerical magnitude, with reproducible local increases for 10, 12, 15, 20 and 100. The extreme reproducibility of this law makes it comparable to other classical statistical findings such as Zipf's law (Zipf, 1936). We discussed four possible sources for variations in numeral frequency: sampling artifacts within our verbal environment, mathematical regularities inherent to number notation, environmental biases, and psychological constraints on mental number representations. The first three sources were found to contribute significantly to several details of the frequency profiles, but they could not satisfactorily explain, by themselves, the frequency decrease and the local increases. The most plausible origin for these two effects lies in the psychological organization of number concepts, structured according to Fechner's law and with local numerical points of reference.

Though we focused mainly on numbers, cross-linguistic frequency regularities are likely to exist in other domains as well. Figure 7 gives the frequency of the four colour words "red", "blue", "green" and "yellow" in French (spoken or written) and in American English. Regularities are apparently found in this domain too. For example, "red" is always about three times more frequent than "yellow". Furthermore, as with numbers, the frequency of colour words is well predicted by the span that each word may cover: the larger the wavelength

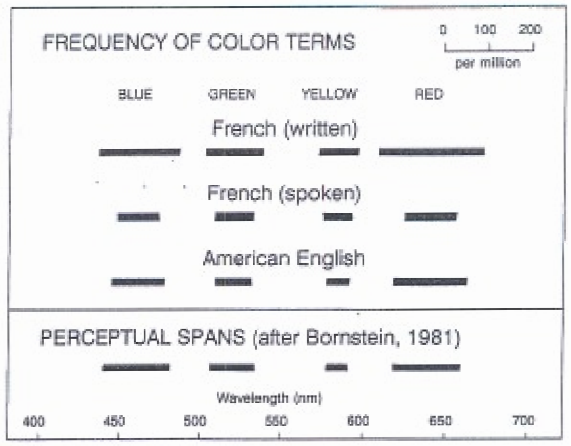


Figure 7. *Frequency analysis for the four most frequent colour words in French and American English. In the upper half, segment lengths are proportional to word frequency in three different databases. The lower half shows the regions of the wavelength spectrum that adults name as blue, green, yellow or red more than 70% of the time (adapted from Bornstein, 1981). Again, a monotonous relation is found between the frequency of a word and the span that the word may cover.*

interval that a colour word may describe, the higher the frequency of that word (Figure 7). Thus, the words "red" and "blue" are comparable to the reference numerals "ten" or "fifteen": both are relatively frequent, and both can apply to a wide region of the underlying wavelength or number continuum. Conversely, the words "green" and "yellow" are comparable to non-reference numerals such as "fourteen": both are relatively infrequent and apply only to a narrow range of their respective continuum. Variations in colour word frequencies therefore seem to reflect the mental processes at work in colour perception. This observation strengthens our hypothesis that variations in number word frequencies partially reveal the architecture of mental representations for numbers.

### *Consequences for experimentation*

That all languages tested possess the same numerical reference points, and show similar decreasing frequency curves, suggests the existence of a universal stable-state core of numerical representations. Thus, testing of groups of subjects and cross-cultural comparisons of data are justified *a posteriori*. However, our research points to a problem that may be encountered in investigations of arithmetic competence. Since number word frequency and numerical magnitude are highly correlated, it will often be impossible to distinguish the effects of these two variables. The word frequency effect has been thoroughly studied in psycholinguistics, and is known to affect processing in many tasks, including word naming (e.g., Forster & Chambers, 1973) and lexical decision (e.g., Rubenstein et al., 1970). It is likely to have contaminated several experiments on numerical processing.

In evaluating past experiments for contamination, it is useful to distinguish effects of *relative magnitude* versus effects of *absolute magnitude*. For instance, numerical comparison experiments have disclosed a distance effect: the larger the numerical distance between two numbers, the faster they are compared (e.g., Moyer & Landauer, 1967). Such effects of relative numerical magnitude are unlikely to be contaminated by frequency artifacts, since it is doubtful that the ratio or difference of two word frequencies would be computed on-line and affect subjects' judgements.

By contrast, calculation experiments in cognitive arithmetics have revealed an effect of problem size: the larger the two operands of an addition or a multiplication, the slower subjects retrieve the answer (e.g., Groen & Parkman, 1972; Geary, Widaman, & Little, 1986). Because this is an effect of absolute numerical magnitude, it might as well be attributed to word frequency. Indeed Ashcraft (1987) has suggested that the problem size effect arises from the frequency with which arithmetical problems are presented in text books and at school.

Numerical stimuli have also been used, mainly in the neuropsychological literature, to study the processes of transduction from one numerical notation to the other. Number word frequency has not been taken into account in this domain of research. Yet processing disruptions would seem likely to affect less frequent numerals before the more frequent ones. Two different transduction pathways might even exist for dealing with frequent versus infrequent numerals. Frequent numerals would rely extensively on lexically stored information, whereas infrequent ones would require step-by-step generic transduction algorithms such as those proposed by McCloskey et al. (1986) or Deloche and Seron (1987). These possibilities remain unexplored. Finally, some researchers have examined the influence of notation on the speed and ease of number processing (e.g., Gonzalez & Kolers, 1982, 1987; Holender & Peereman, 1987; Takahashi & Green, 1983; Vaid & Corina, 1989). In our data, number word frequency varies with a factor of 4 between arabic notation and verbal notation, even if the relative frequency variations are similar in both notations. Therefore, in this situation, controlling for artifacts of frequency is an absolute must.

## Appendix A: Why Benford's law holds for random numbers

Why do first digits of random numbers drawn from most distributions follow a decreasing distribution similar to Benford's law? In this appendix, we propose a series of qualitative mathematical arguments. First, consider random numbers drawn from a flat distribution over an interval $[1, n]$. If $n$ is a number like 9, 99, 999, etc., then the random numbers are equally likely to start with each of the nine digits. For all other values of $n$, smaller digits will appear more often as first digits. This is most clearly seen if you imagine drawing in the interval $[1, 24]$ for example: digits 1 and 2 will be much more likely than the others.

Now consider numbers drawn from a decreasing distribution. This is similar to drawing from several intervals $[1, n_i]$, where $n_i$ can take on several different values. The effects of drawing from a single interval are thus averaged, and a smoothly decreasing digit frequency curve is found. Flehinger (cited by Raimi, 1969) further proved that this frequency profile is *exactly* Benford's law when the initial distribution obeys some additional requirements. These additional assumptions were criticized by Raimi (1969), who argued that they amounted to a hypothesis about the probability of numerosities in the real world. However, our analysis shows that even if Flehinger's assumptions are relaxed, a close approximation to Benford's law still obtains.

Another way of understanding what happens with flat or decreasing distributions of random numbers is shown in Figure 8. This figure gives, for three different distributions, the progressive calculation steps that lead from the full distribution over the range of integers to the distribution of first digits only. The top left graph of each panel gives the shape of the distribution on a standard linear scale, superimposed on a saw-toothed curve giving the value of first digits from 1 to 9. The graph below shows the same distribution plotted on a piecewise linear scale. The scale is linear on each $\log_{10}$ interval (e.g., [1 . . . 9], [10 . . . 99], [100 . . . 999]) and gives equal space to each such interval; therefore, the saw-toothed curve for first digits now appears periodic. Because the scale is not linear any more, the distribution must be weighted by the appropriate scale factor on each $\log_{10}$ interval, so that its integral over any interval still gives the probability of drawing a random number in this interval. Finally the graph to the right gives the resulting distribution of first digits. It is obtained simply by adding the curves within each of the $\log_{10}$ intervals of the previous graph, merging them onto a single interval [1 . . . 9] (technically, the integrals of the curves are added, not the curves themselves).

Let us consider the concrete case of the exponentially decreasing distribution plotted in the middle section of Figure 8. On the top left graph, it is seen that the distribution smoothly decreases towards zero. The graph below shows the behaviour of the distribution over the intervals [1 . . . 9] (1-digit numbers), [10 . . . 99] (2-digit numbers) and [100 . . . 999] (3-digit numbers). Of course the distribution is still decreasing within each of these intervals. The striped area under the curve also makes it clear that 1-digit numbers are much less frequent in the distribution than are 2- or 3-digit numbers. To obtain the distribution of first

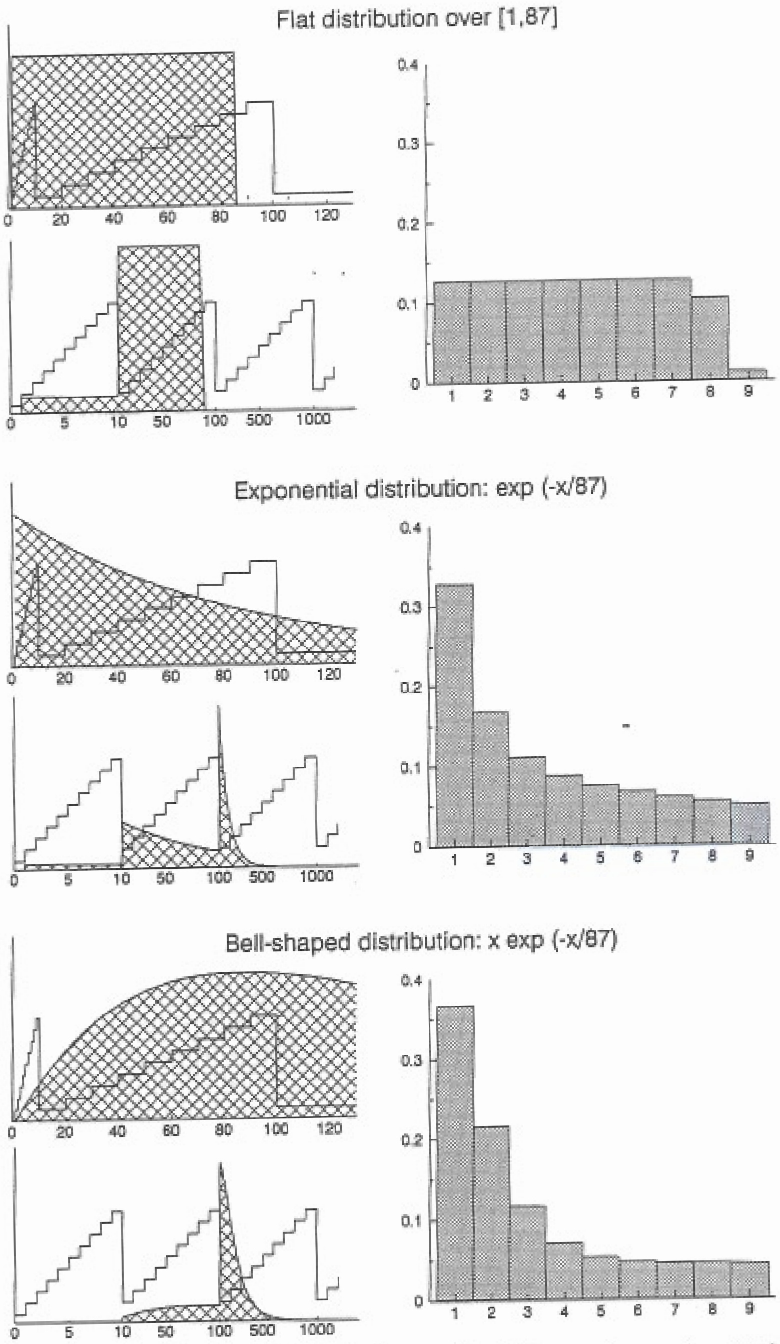


Figure 8. *Why a Benfordian law holds for random numbers. Three random number distributions are considered: flat, exponentially decreasing, and bell-shaped. See Appendix A.*

digits shown on the right, the values corresponding to all numbers starting with digit 1 must be added together, regardless of whether the numbers are 1, 2, or 3 digits long; and similarly for all numbers starting with digits 2, 3, 4, etc. This essentially consists in superimposing the $\log_{10}$ intervals for 1-, 2- and 3-digit numbers. Since over each such interval the distribution is decreasing and convex upward, the average is again a decreasing convex upward curve which bears considerable similarity to the actual Benford law (compare with Figure 4).

The case of a flat distribution is shown at the top of Figure 8. The distribution is flat over all but the last $\log_{10}$ interval that it spans. In this last interval the distribution suddenly drops from a constant to zero, resulting in a step-like (but still decreasing) distribution of first digits.

Finally, what happens for a non-decreasing distribution of random numbers? The bottom section of Figure 8 shows the case of a smooth, bell-shaped distribution. The distribution is increasing over the first $\log_{10}$ interval [1 . . . 9], reaches its maximum within the interval [10 . . . 99], and decreases only over the last interval [100 . . . 999]. However, as shown on the bottom left graph, the contribution of the latter interval to the final distribution of first digits is very important, because there are more 3-digit numbers in the distribution than there are 1- or 2-digit numbers. Also, even though the initial distribution is sharply rising and slowly falling, this is actually reversed on the bottom left graph, because of the compression inherent in the piecewise linear scale (numbers from 100 to 999 occupy the same space as numbers from 1 to 9). As a result, only the decreasing, right-hand side of the distribution really contributes to the final distribution of first digits. The curve obtained is indeed almost identical to that resulting from a strictly decreasing distribution.

Obviously, however, not all non-decreasing distributions of random numbers follow Benford's law. For instance, numbers drawn from a narrow Gaussian distribution centred on 58 are obviously much more likely to start with a 5 or a 6 than with, say, digit 1. Another obvious counter-example is a monotonically increasing distribution over [1 . . . 99]. Given the above examples and counter-examples, we conjecture that the following conditions are sufficient for a Benfordian law to hold with random numbers:

(1) The distribution of numbers spans over several $\log_{10}$ intervals, that is, it includes numbers with differing numbers of digits.
(2) It is smooth enough, that is, it does not have sharp increases and decreases within a single $\log_{10}$ interval.

Note that numerical tables of the kind that Benford used in compiling his data are likely to satisfy these two conditions. We think that this explains the apparent universality of Benford's law.

## Appendix B: Reconstructing the putative mental space of numbers from frequency data

Let us define the numerical span $s_n$ of a numeral $n$ as the width of the interval $I(n)$ of numerosities that it can adequately represent or approximate. This appendix describes a mathematical method for reconstructing approximate numerical spans from numeral frequency data. Let $\Phi$ be the psychophysical function specifying the internal encoding of numerosities on the mental "number line". By hypothesis, we expect the observed frequency $F(n)$ of numeral $n$ to be proportional to the width of the internal interval $\Phi[I(n)]$ on the number line. Since the psychophysical function $\Phi$ is defined up to a linear transformation, one may write:

$$F(n) = \Phi(n + s_n/2) - \Phi(n - s_n/2) \tag{5}$$

For a non-reference numeral $n$, the span $s_n$ is 1 by definition, and thus $F(n) \approx \Phi'(n)$. In Figure 2, we observe that on a log-log scale the frequency of non-reference numerals is linearly related to their numerical magnitude. Hence, by linear regression, parameters $a$ and $b$ can be estimated such that, for non-reference numerals, $\text{Log}\, F(n) = -a \log n + b$. By integrating and letting $\Phi(1) = 0$, one gets

$$\Phi(n) = \frac{e^b}{a-1}(1 - n^{1-a}) \tag{6}$$

The parameters $a$ and $b$ are obtained only from non-reference numerals, but in first approximation we may generalize the equation to reference numerals as well. We may then solve equation (5) numerically by Newton's method to get an estimate of the span $s_n$ for reference numerals. Figure 6 shows the psychophysical function $\Phi$ and the spans $s_n$ obtained when applying this scheme to the American English data. Parameter $a$ was estimated at 1.92. Thus, function $\Phi(n)$ behaved approximately as $1 - 1/n$.

## References

Ashcraft, M.H. (1987). Children's knowledge of simple arithmetic: A developmental model and simulation. In J. Bisanz, C.J. Brainerd, & R. Kail (Eds.), *Formal methods in developmental psychology* (pp. 302–338). Berlin: Springer.

Baird, J.C., Lewis, C., & Romer, D. (1970). Relative frequencies of numerical responses in ratio estimation. *Perception and Psychophysics*, *8*, 358–362.

Baird, J.C., & Noma, E. (1975). Psychophysical study of numbers. I. Generation of numerical responses. *Psychological Research*, *37*, 281–297.

Banks, W.P., & Hill, D.K. (1974). The apparent magnitude of number scaled by random production. *Journal of Experimental Psychology Monographs*, *102*, 353–376.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, *78*, 551–572.

Bertelson, P. (1987). *The onset of literacy*. Cambridge, MA: MIT Press.

Bornstein, M.H. (1981). Psychological studies of color perception in human infants: Habituation, discrimination and categorization, recognition and conceptualization. In L.P. Lipsitt (Ed.), *Advances in infancy research* (Vol. 1). Norwood, NJ: Ablex.

Buckley, P.B., & Gillman, C.B. (1974). Comparison of digits and dot patterns. *Journal of Experimental Psychology*, *103*, 1131–1136.

Campbell, J.I.D., & Clark, J.M. (1988). An encoding complex view of cognitive number processing: Comment on McCloskey, Sokol, & Goodman (1986). *Journal of Experimental Psychology: General*, *117*, 204–214.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Conesa, A., et al. (1983). *Diccionari de freqüencies de produccions orals i escrites en llengua catalana*. Facultat de Psicologia, Universitat de Barcelona.

Dehaene, S. (1989). The psychophysics of numerical comparison: A reexamination of apparently incompatible data. *Perception and Psychophysics*, *45*, 557–566.

Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Anological and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 626–641.

Deloche, G., & Seron, X. (1984). Semantic errors reconsidered in the procedural light of stack concepts. *Brain and Language*, *21*, 59–71.

Deloche, G., & Seron, X. (1987). Numerical transcoding: A general production model. In G. Deloche & X. Seron (Eds.), *Mathematical disabilities: A cognitive neuropsychological perspective* (pp. 137–170). Hillsdale, NJ: Erlbaum.

Duncan, E.M. & McFarland, C.E. (1980). Isolating the effects of symbolic distance and semantic congruity in comparative judgments: An additive-factors analysis. *Memory and Cognition*, *8*, 612–622.

Fodor, J.A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Forster, K.I., & Chambers, S.M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, *12*, 627–635.

Francis, W.N., & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

Gallistel, C.R. (1990). *The organization of learning*. Cambridge, MA: Bradford Books/MIT Press.

Gallistel, C.R., & Gelman, R. (1991). Preverbal and verbal counting and computation. Submitted for publication.

Geary, D.C., Widaman, K.F., & Little, T.D. (1986). Cognitive addition and multiplication: Evidence for a single memory network. *Memory and Cognition*, *14*, 478–487.

Gonzalez, E.G., & Kolers, P.A. (1982). Mental manipulation of arithmetic symbols. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *8*, 308–319.

Gonzalez, E.G., & Kolers, P.A. (1987). Notational constraints on mental operations. In G. Deloche and X. Seron (Eds.), *Mathematical disabilities: A cognitive neuropsychological perspective* (pp. 27–42). Hillsdale, NJ: Erlbaum.

Gougenheim, G., Michea, R., Rivenc, P., & Sauvageot, A. (1956). *L'élaboration du français élémentaire et d'une grammaire de base*. Paris: Didier.

Gould, J.L., & Marler, P. (1987). Learning by instinct. *Scientific American*, 62–73.

Groen, G.J., & Parkman, J.M. (1972). A chronometric analysis of simple addition. *Psychological Review*, *79*, 329–343.

Hatano, G., & Osawa, K. (1983). Digit memory of grand experts in abacus-derived mental calculation. *Cognition*, *15*, 95–110.

Holender, D., & Peereman, R. (1987). Differential processing of phonographic and logographic single-digit numbers by the two hemispheres. In G. Deloche & X. Seron (Eds.), *Mathematical disabilities: A cognitive neuropsychological perspective* (pp. 43–86). Hillsdale, NJ: Erlbaum.

Hurford, J.R. (1975). *The linguistic theory of numerals*. Cambridge University Press.

Ifrah, G. (1981). *Histoire Universelle des Chiffres*. Paris: Seghers.

Imbs, P. (Ed.) (1971). *Trésor de la langue française*. Paris: Editions du CNRS.

Jaffe-Katz, A., Budescu, D.V., & Wallsten, T.S. (1989). Timed magnitude comparison of numerical and non-numerical expressions of uncertainty. *Memory and Cognition, 17*, 249–264.

Juilland, A., & Chang-Rodriguez, E. (1964). *Frequency dictionary of Spanish words*. The Hague: Mouton.

Jusczyk, P.W., & Bertoncini, J. (1988). Viewing the development of speech perception as an innately guided learning process. *Language and Speech, 31*, 217–238.

Krueger, L.E. (1989). Reconciling Fechner and Stevens: Toward a unified psychophysical law. *Behavioral and Brain Sciences, 12*, 251–267.

Mandler, G., & Shebo, B.J. (1982). Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General, 111*, 1–21.

McCloskey, M., Sokol, S.M., & Goodman, R.A. (1986). Cognitive processes in verbal-number production: Inferences from the performance of brain-damaged subjects. *Journal of Experimental Psychology: General, 115*, 307–330.

Mehler, J., & Bever, T.G. (1968). The study of competence in cognitive psychology. *International Journal of Psychology, 3*, 273–280.

Menninger, K. (1969). *Number words and number symbols*. Cambridge, MA: MIT Press.

Merzenich, M.M. (1987). Dynamic neocortical processes and the origins of higher brain functions. In J.P. Changeux & M. Konishi (Eds.), *The neural and molecular bases of learning* (pp. 337–358). Dahlem workshop reports. New York: Wiley.

Moyer, R.S., & Landauer, T.K. (1967). Time required for judgements of numerical inequality. *Nature, 215*, 1519–1520.

NLRI (1970). *Studies on the vocabulary of modern newspapers*. Technical Report, National Language Research Institute (Japan).

Noma, E., & Baird, J.C. (1975). Psychophysical study of numbers. II. Theoretical models of number generation. *Psychological Research, 38*, 81–95.

Power, R.J.D., & Longuet-Higgins, H.C. (1978). Learning to count: A computational model of language acquisition. *Proceedings of the Royal Society (London), B200*, 391–417.

Quine, W.V.O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Raimi, R.A. (1969). The peculiar distribution of first digits. *Scientific American, 221*, 109–120.

Ranganatha, M.R. (1982). *Morphophonemic analysis of the Kannada language*. CIIL Occasional Monograph Series - 17, Mysore.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology, 7*, 532–547.

Rubenstein, H., Garfield, L., & Millikan, J. (1970). Homographics entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior, 9*, 487–494.

Rule, S.J. (1969). Equal discriminability scale of number. *Journal of Experimental Psychology, 79*, 35–38.

Sekuler, R., & Mierkiewicz, D. (1977). Children's judgments of numerical inequality. *Child Development, 48*, 630–633.

Shepard, R.N., Kilpatrick, D.W., & Cunningham, J.P. (1975). The internal representation of numbers. *Cognitive Psychology, 7*, 82–138.

Shepherd, J.W. (1981). Social factors in face recognition. In G. Davis, H.D. Ellis, & J.W. Shepherd (Eds.), *Perceiving and remembering faces*. London: Academic Press.

Siegler, R.S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116*, 250–264.

Sokol, S.M., Goodman-Schulman, R., & McCloskey, M. (1989). In defense of a modular architecture for the number-processing system: Reply to Campbell and Clark. *Journal of Experimental Psychology: General, 118*, 105–110.

Starkey, P., & Cooper, R.G., Jr. (1980). Perception of numbers by human infants. *Science, 210*, 1033–1035.

Stigler, J.W. (1984). "Mental abacus": The effect of abacus training on Chinese children's mental calculation. *Cognitive Psychology, 16*, 145–176.

Suga, N. (1982). Functional organization of the auditory cortex. Representation beyond tonotopy in the bat. In C.N. Woolsey (Ed.), *Cortical sensory organization. Vol. III. Multiple auditory areas*. Clifton, NJ: Humana Press.

Takahashi, A., & Green, D. (1983). Numerical judgments with Kanji and Kana. *Neuropsychologia, 21*, 259–263.

Vaid, J., & Corina, D. (1989). Visual field asymmetries in numerical size comparisons of digits, words and signs. *Brain and Language, 36*, 117–126.

Zipf, G.K. (1936). *The psycho-biology of language*. London: Routledge.