Short note

# Segmental inventory size, word length, and communicative efficiency[1]

DANIEL NETTLE

*Abstract*

*Functional theories of language structure predict that as the number of contrastive segments in a language increases, the average length of a word will decrease. This relationship is found to hold for a sample of ten languages, and to fit the synergetic model* $Y = aX^b$. *The average length of a word is approximately* $7 \pm 2$ *segments. This corresponds to the proposed capacity of working memory.*

## Introduction

One aspect of the general revival of interest in functional explanations in linguistics has been an increase in the use of quantitative models of language structure. Such models have been used to explain the qualities of vowels and consonants found in phonological inventories of a given size (Lindblom 1986; Lindblom and Maddieson 1988), and cross-linguistic preferences for certain phonotactic constraints (Kawasaki-Fukumori 1992). The starting point of these models is the assumption that language is functionally adapted to the needs of efficient communication, which are taken to be the need for articulatory ease and the need for perceptual salience. However, if only these needs are considered, it is unclear why any language should use more than a bare minimum of contrastive segments, as having a larger segmental inventory seems likely to either increase articulatory cost, because more extreme articulatory gestures will be needed, or decrease the ease of decoding, as the perceptual space will be more crowded, or both. The number of segments actually used by natural languages varies a great deal, from 12 to at least 120 (Maddieson 1984).

There has been a great deal of work in quantitative-functional linguistics in the German and Eastern European traditions (see, for example,

Linguistics 33 (1995), 359–367

0024–3949/95/0033–0359

Hammerl and Sambor 1993, and references therein). The general model of language structure arising from this work is the synergetic model (Köhler 1986, 1987, 1993). Synergetic linguistics treats language as a self-organizing, dynamical system that is structured by a number of COMPETING MOTIVATIONS, such as the need for simplicity on the one hand, and the need for economy of coding on the other. The rationale is very similar to that of Haiman (1983) and DuBois (1985).

From this perspective, the aforementioned disadvantages of increasing the inventory size should have the compensatory advantage of allowing shorter linguistic units, and hence greater economy. It has often been hypothesized that languages with larger segmental inventories will have generally shorter words (or morphemes — Saporta 1963), but the effect has not been demonstrated empirically for a sizeable set of languages.

According to Köhler's model (1987: Table 2), the length of a word will be a function of its frequency, the number of segments in the inventory of the language, the number of words in the lexicon of the language, and the degree of security of transmission that the language requires. The last of these factors can be assumed to be a constant, giving (1):

(1) $\text{Length} = a\ (\text{Frequency}^b)\ (\text{Segments}^c)(\text{Lexicon}^d)$

— where a, b, c, and d are constants. This is just an application of the general synergetic model for the relationship of lexical variables, which is as follows:

(2) $X = a\ Y^b$

Although there are known to be cross-language differences in the size of the lexicon, the effects of these will be negligible as long as all lexicons are large and d is small. Lexicon size is therefore treated as a constant for the present purposes, giving (3):

(3) $\text{Length} = a\ (\text{Frequency}^b)\ (\text{Segments}^c)$

It follows that if we take a large sample of words of different frequencies from several languages, the mean word length for each language should follow this function:

(4) $L = a\ S^b$

— where L is the mean word length, and S is the number of segments in the inventory. This hypothesis is tested here for ten languages, using a random sample of lexical entries from a dictionary for each language. The word in its citation form is felt to be the most appropriate level to investigate this relationship, as typological differences make cross-

language comparisons of morphemes or words in discourse much more problematic.

## Method

Ten languages were chosen so as to represent a wide selection of inventory sizes and language families. Phonemic inventories were obtained from standard sources such as Campbell (1991). These were used to calculate for each language a measure S, or the total number of contrastive segments. For vowels, this involved multiplying the number of phonemes by the number of tones or contrastive lengths where appropriate.

For each language, 50 head-words were chosen at random from a dictionary, by choosing the first word on every nth page, with n dependent upon the length of the dictionary. The problem of determining the unit equivalent to the word in Mandarin Chinese was obviated by using an English-Chinese dictionary and taking whatever was given as the translation of the English word, which was sometimes a monosyllable and sometimes a bisyllabic unit. The chosen words were transcribed phonemically on the basis of the relevant inventory. This was possible for all the languages because the dictionaries used either gave phonetic transcriptions or used an orthography predictably interpretable in phonological terms (Turkish, Chinese, !Xũ). The length of each word in segments was then established, and the mean length for the 50 words found. The treatment of diphthongs and geminates was determined by the inventory for each language: if the inventory listed a segment independently, it was counted as one unit, and if not it was counted as two.

One possible problem was that differences in the sizes of the dictionaries were responsible for different average word lengths. A smaller dictionary would contain generally more common, and hence shorter, words. To check for such an effect, an estimate of the size of each dictionary in lexical entries was recorded. Additionally, an investigation of the effect of differing dictionary size on mean word length was conducted for one language, Italian. This was done by finding the mean length of the most common 10, 50, 100, 150, 200, 250, 300, 400, and 500 words of a very large corpus of Italian text (Source: Bortolini et al. 1972), plus the mean length of a random sample of 50 of the most frequent 1000, 2000, and 5000 words. These lengths were compared with the mean length of a random sample of 50 words from three dictionaries containing 20,000, 50,000, and 95,000 lexical entries respectively.

## Results

The mean word lengths (L) and the segmental inventory sizes (S) for the ten languages are given in Table 1. The overall mean word length is 6.20 segments. Means for individual languages range from 3.65 to 8.69.

The mean word lengths for the different dictionary sizes in Italian are given in Table 2. Figure 1 shows the relationship between the two variables for dictionary sizes of 10–2000 and 100–100,000 entries. It is clearly negatively exponential, and the rate of increase in word length is very small once a dictionary size of about 1000 words has been reached. The dictionaries used in the main part of this study have between 3300 and 40,000 lexical entries (the dotted vertical lines in the lower graph of Figure 1). Extrapolating from the graph, the variation in word length due to dictionary size within this range will be at most $\pm 0.5$ segments. The differences in word length observed were much larger than this, and there was no significant correlation between dictionary size and mean word length for the ten languages (Table 1: $r = -0.145$).

Comparing across the ten languages, mean word length is related to segmental inventory size as the model predicts (Figure 2). Curvilinear regression gives the relationship as follows:

(5) $L = 29.35\ S^{-0.43}$
($r^2 = 0.77$; 8 degrees of freedom; significance of the regression: $p < 0.001$)

## Discussion

The mean length of a word in each language is close to the "magic number" of $7 \pm 2$ segments. Similarly, Fenk-Oczlon and Fenk (1985)

Table 1. *The ten languages used and segmental inventory size (S), the mean word length (L), and the size of the dictionary used (D)*

| | S | L | D |
|---|---|---|---|
| Thai | 76 | 3.65 | 30,000 |
| Italian | 30 | 7 | 20,000 |
| Hindi | 41 | 5.57 | 29,000 |
| Hawaiian | 18 | 7.08 | 25,000 |
| !Xũ | 119 | 4.02 | 3300 |
| Turkish | 28 | 6.44 | 25,000 |
| Nahuatl | 23 | 8.69 | 10,500 |
| German | 41 | 6.44 | 40,000 |
| Georgian | 34 | 7.74 | 4500 |
| Mandarin | 53 | 5.4 | 11,250 |

Table 2. *The mean length (L) of a sample of words from Italian word lists and dictionaries of various sizes (D)*

| Number | D | L | Source |
|---|---|---|---|
| 1 | 10 | 2.8 | Bortolini et al. (1972) |
| 2 | 50 | 3.68 | Bortolini et al. (1972) |
| 3 | 100 | 4.2 | Bortolini et al. (1972) |
| 4 | 150 | 4.71 | Bortolini et al. (1972) |
| 5 | 200 | 4.97 | Bortolini et al. (1972) |
| 6 | 250 | 5.13 | Bortolini et al. (1972) |
| 7 | 300 | 5.23 | Bortolini et al. (1972) |
| 8 | 400 | 5.54 | Bortolini et al. (1972) |
| 9 | 500 | 5.82 | Bortolini et al. (1972) |
| 10 | 1000 | 6.56 | Bortolini et al. (1972) |
| 11 | 2000 | 6.68 | Bortolini et al. (1972) |
| 12 | 5000 | 7.44 | Bortolini et al. (1972) |
| 13 | 20,000 | 7 | Oxford Italian Minidictionary |
| 14 | 50,000 | 7.86 | Dizionario Italiano (Rizzoli) |
| 15 | 95,000 | 8.34 | Nuovissimo Dardano Dizionario |

Note: The lists from Bortolini et al. (1972) are based on frequency of occurrence in a large corpus of text, the 10-word list containing the 10 most common words, the 50-word list the 50 most common, and so forth.

have found that the mean length of simple declarative sentences in 27 languages is $7 \pm 2$ syllables. The figure is the proposed capacity of short-term or working memory (Miller 1956). This capacity can be increased by "chunking" or recoding data into larger units, and both data sets are compatible with the view that the hierarchy of units in natural language functions to reduce working-memory load by gathering groups of small units into larger ones.

Cross-language differences in mean word length are not due to the size of the dictionary used. The tradeoff between word length and inventory size occurs as the synergetic model predicts. The relationship is, of course, not perfect but is surprisingly good given the sample size and the simplifying assumptions made.

Inventory size is, therefore, a product of the competing motivations of performance and simplicity. Increasing it increases the difficulty of production and perception but, other things being equal, improves the rate of information transmission. The qualification is important — the global rate of information transmission in a language, as evidenced by the length of propositions, is strongly influenced by syntactic and morphological variables as well as word length (Fenk-Oczlon and Fenk 1985).

Which one of the two motivations should win out in any particular

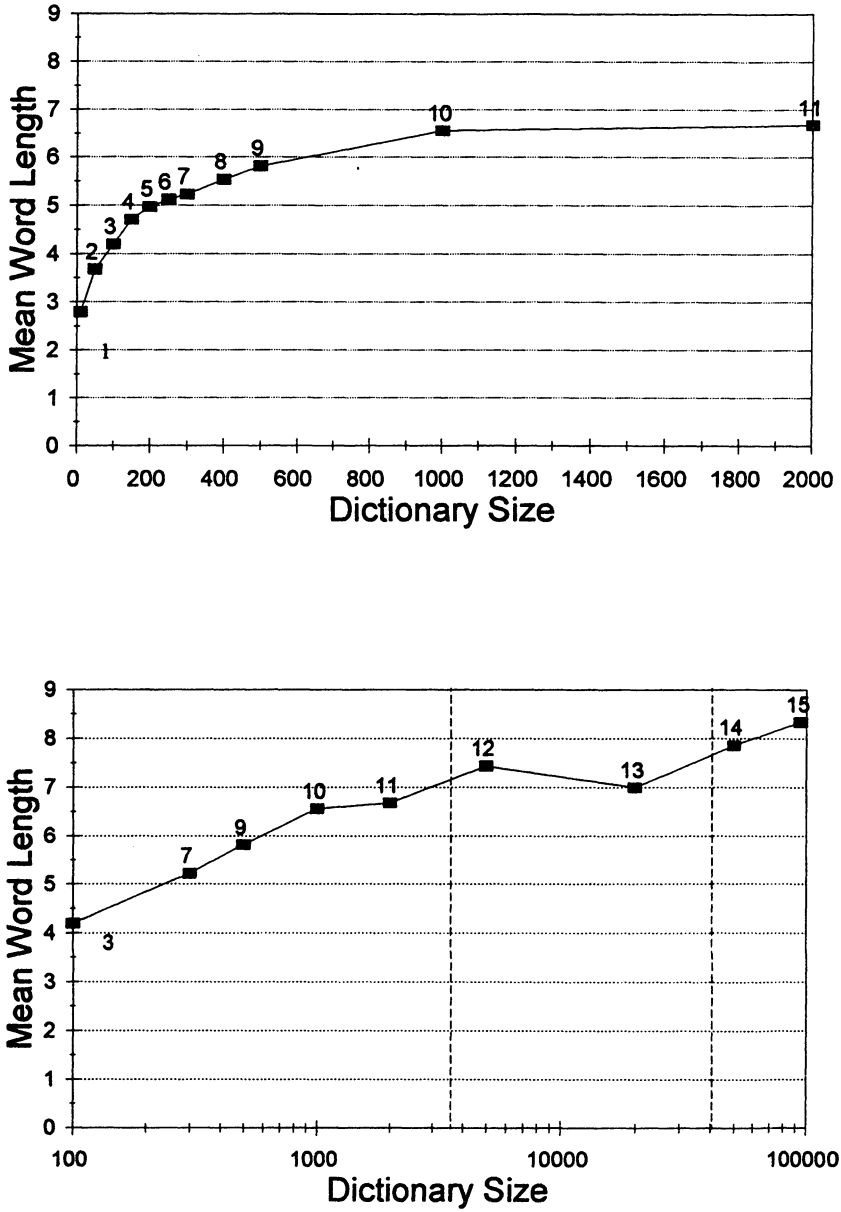


Figure 1. *The mean length of a sample of Italian words from dictionaries and word lists of varying sizes, from 10 to 2000 (upper graph), and from 100 to 100,000 (lower graph) (the numbering refers to Table 2; the dotted vertical lines on the lower graph represent the size of the smallest and largest dictionary used in the main part of this study)*

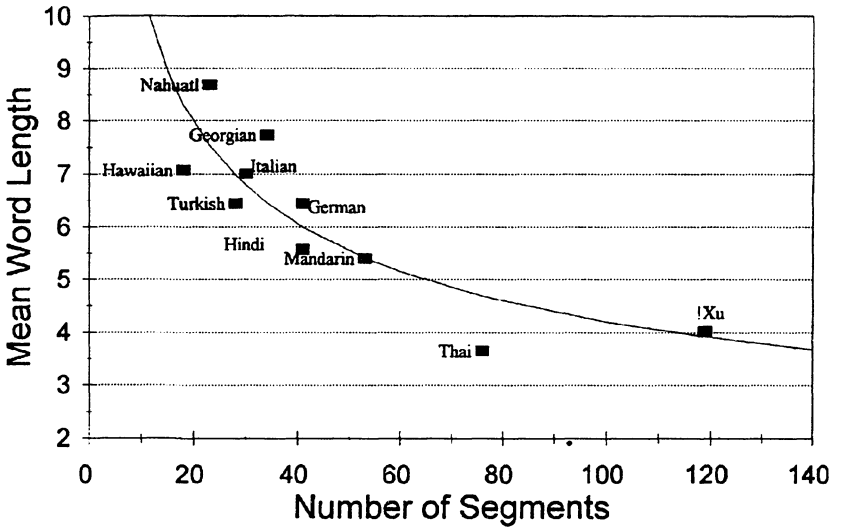


Figure 2. *The mean word length against the segmental inventory size for the ten languages; the curve is given by* $L=29.35\ S^{-0.43}$

case is unclear. Jakobson (1929) suggested that the more diffuse the geographical range of a language, the simpler the system had to be, as ease of learning and discrimination under accent diversity would be preeminently important. Reduction is of course typical in pidginization, where low cost is more important than high performance (Mülhäusler 1986).

Diachronically, the tradeoff implies that where the number of contrasts in the system increases, there will be processes of reduction in the lexicon. Examples of this can be seen in the transition from Latin to French. As the vocalic system expanded, final syllables and syllable-final consonants disappeared. Latin nouns such as *fokus* are more sharply reduced in French (*fø*) than in the other Romance languages that have undergone comparable morphological changes (Italian *foko*). This is both symptom and result of the expanding French inventory.

We can thus posit a cycle in which phonetic changes, the result of social motivations and geographical influences, accrue and cause phonological reinterpretation. An expanded phonological system in turn has a knock-on effect on lexical forms in such a way as to optimize the transmission of information. This relationship underlines the usefulness of seeing languages, or at least lexicons, as dynamical, self-organizing systems. Whether all differences between systems will turn out to be explicable in terms of competing design motivations is, however, far from clear.

*Received 16 June 1994*
*Revised version received*
*4 October 1994*

*University College London*

## Note

1. This research was funded by the UK Medical Research Council. I am grateful to Robin Dunbar, John Harris, and an anonymous referee for useful comments. Correspondence address: Department of Anthropology, University College London, London WC1E 6BT, UK.

## References

Bortolini, U.; Tagliavini, C.; and Zampolli, A. (1972). *Lessico di Frequenza della Lingua Italiana Contemporanea*. Milan: Garzanti.

Campbell, George (1991). *Compendium of the World's Languages*, 2 vols. London: Routledge.

Du Bois, John (1985). Competing motivations. In *Iconicity in Syntax*, John Haiman (ed.), 343–365. Amsterdam: Benjamins.

Fenk-Oczlon, Gertraud; and Fenk, August (1985). The mean length of propositions is $7 \pm 2$ syllables — but the position of languages within this range is not accidental. In *Cognition, Information Processing and Motivation*, Géry d'Ydewalle (ed.), 355–359. Amsterdam: Elsevier.

Haiman, John (1983). Iconic and economic motivation. *Language* 59, 781–789.

Hammerl, Rolf; and Sambor, Jadwiga (1993). Synergetic studies in Polish. In *Contributions to Quantitative Linguistics*, Reinhard Köhler and Burghard Rieger (eds.), 331–359. Dordrecht: Kluwer Academic.

Jakobson, Roman (1929). Remarques sur l'évolution phonologique du russe comparée à celle des autres langues slaves. In *Selected Writings I: Phonological Studies*, Roman Jakobson, 7–116. The Hague: Mouton.

Kawasaki-Fukumori, Haruko (1992). An acoustical basis for universal phonotactic constraints. *Language and Speech* 35 (1,2), 73–86.

Köhler, Reinhard (1986). *Zur Linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

—(1987). Systems theoretical linguistics. *Theoretical Linguistics* 14, 241–257.

—(1993). Synergetic linguistics. In *Contributions to Quantitative Linguistics*, Reinhard Köhler and Burghard Reiger (eds.), 41–52. Dordrecht: Kluwer Academic.

Lindblom, Bjorn (1986). Phonetic universals in vowel systems. In *Experimental Phonology*, John Ohala and Jeri Jaeger (eds.), 13–44. Orlando: Academic Press.

—; and Maddieson, Ian (1988). Phonetic universals in consonant systems. In *Language, Speech and Mind*, Larry Hyman and Charles Li (eds.), 62–80. London: Routledge.

Maddieson, Ian (1984). *Patterns of Sounds*. Cambridge: Cambridge University Press.

Miller, George (1956). The magical number of seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* 63, 81–97.

Mülhäusler, Peter (1986). *Pidgin and Creole Linguistics*. Oxford: Blackwell.

Saporta, Sol (1963). Phonemic distribution and language universals. In *Universals of Language*, Joseph Greenberg (ed.), 48–57. Cambridge, MA: MIT Press.

## *Dictionaries used*

*A Practial English-Chinese Pronouncing Dictionary*. Rutland, VT: Tuttle, 1970.

*Hippocrene Concise English-Georgian Dictionary*. New York: Hippocrene, 1992.

*Langenscheidt Pocket German Dictionary*. London: Hodder and Stoughton, 1958.
*Hawaiian Dictionary*. Honolulu: University of Hawaii Press, 1971.
*A Practical Hindi-English Dictionary*. Delhi: National Publishing House, 1970.
*Oxford Italian Minidictionary*. Oxford: Oxford University Press, 1986.
*Dizionario Italiano*. Milan: Biblioteca Universale Rizzoli, 1988.
*Nuovissimo Dardano Dizionario della Lingua Italiana*. Rome: Armando Curico Editore.
*An Analytical Dictionary of Nahuatl*. Norman: University of Oklahoma Press, 1983.
*Thai Vocabulary*. Washington, D.C.: American Council of Learned Societies, 1955.
*Langenscheidt Pocket Turkish Dictionary*. Berlin: Langenscheidt, 1993.
*Žu|'hõasi Fonologie & Woordeboek*. Capetown: University of Capetown School of African Studies; Rotterdam: A.A. Balkena, 1975.