Relating Word Length to Morphemic Structure: A Morphologically Motivated Class of Discrete Probability Distributions*

Peter Meyer
Georg-August-Universität, Göttingen

ABSTRACT

In this article an alternative way of accounting for the distribution of word length – as measured in terms of number of syllables per word – in texts of certain natural languages will be explored.

A standard synergetic account of word length distribution has been developed and investigated in depth by Altmann, Köhler, Wimmer and others (cf. Wimmer et al., 1994). In this approach, certain mathematical relationships between, for example, neighbouring word length classes are assumed to hold. For instance, in a simple case, the probability P_x assigned to the class with x-syllabic words is taken to be proportional to the probability P_{x-1} of the *preceding* word length class:

$$P_x \sim P_{x-1}, \qquad (1)$$

where the proportionality factor is assumed to be a function $g(x)$ of word length x:

$$P_x = g(x)P_{x-1}. \qquad (2)$$

In the most elementary case $g(x)$ may be taken to have the form of "Menzerath's Law":

$$g(x) = ax^{-b}; \qquad (3)$$

the resulting difference equation will then return the *Conway-Maxwell-Poisson Distribution*.

The approach just sketched has turned out to be a very powerful tool in modelling word length distributions across rather different natural languages. However, the basic assumptions underlying these models are, themselves, in need of theoretical justification. An obstacle to any such attempt at justification is given by the fact that the *parameters* (usually two) of the distributions in question do not admit of any direct *interpretation* in linguistic terms. Roughly speaking, in our equation $g(x) = ax^{-b}$ the parameters a and b may be understood as representing, for example, hearer's vs. speaker's communicative interests or redundancy vs. efficiency of information transmission. Regrettably, however, no method of *measuring* communicative interests or redundancy of transmission is known, as far as natural languages are concerned; and the very idea that human languages are primarily a means of transferring quantifiable chunks of information from speaker to hearer has, of course, its own, well-known philosophical shortcomings. As a consequence, the parameter interpretations proposed include a considerable amount of *a priori* analogical reasoning.

The mathematical model I discuss here provides for a direct interpretation of the distribution parameters in terms of traditional qualitative linguistics without discarding the need for a synergetic approach. It was developed in studying word length in traditional narratives. For the purposes of this paper, let it suffice to say that a

*Address correspondence to: Peter Meyer, Sprachwissenschaftliches Seminar, Georg-August-Universität Göttingen, Theaterstr. 14, D-37073 Göttingen, Germany. E-mail: pmeyer@stud.uni-goettingen.de.

Erschienen in: Journal of Quantitative Linguistics 6 (1999) 1, S. 66-69

'polysynthetic' language is characterised by unbounded recursive morphological left-branching; that is, all stems (pre-ending morpheme sequences) may be enlarged in a *productive* way by suffixing a further morpheme which can be interpreted as the head of the resulting stem and determines, as such, the word class pertinence of the whole 'stem' sequence. For an example, cf. the typical Inuktitut word *ui-qa-ruma-laun[g]-ngit-tunga*, which might be glossed as 'HUSBAND'-'HAVE'-'WANT'-PAST-NEG-1SG:PRES:ITR and translates as 'I didn't want to have a husband'. In this example, the morpheme sequences *ui-*, *uiqa-*, *uiqaruma-* etc. can also be used as inflectable stems on their own. It is important to note that in Inuktitut, morphemes usually contribute a fixed number of syllables (in all but a very few cases, more than zero) to any word they form a part of, despite the all-pervading complex word-internal sandhi processes typical of Inuktitut in general.

The basic idea of the approach proposed here is quite simple: Every word contains a certain number of morphemes; every morpheme, in turn, includes a certain number of syllable nuclei. This leads in a natural fashion to the following two-step approach:

(1) We assume that the *number of morphemes of a given word* is expressed by a random variable N with probability generating function (pgf) $G(t)$, where $E(N)$ is the average number of morphemes per word in the text in question.

(2) We assume that the *number of syllables of any given morpheme* is expressed by a random variable Y with pgf $H(t)$, where $E(Y)$ is the average number of syllables per morpheme in the text in question.

Note that assumption (1) is adequate only for languages that, like Inuktitut, have no principal restrictions on word-internal morphemic complexity. The approach outlined here might, however, also be applicable to languages that show productive recursivity in word formation only in a part of their lexicon, as is the case with German or Chinese nominal composition. This remains to be tested in the light of available data.

The distribution of the *number of syllables in a word* obviously is a *random sum of random variables*, with N giving the number of Y-distributed variables in the sum. Thus, the *total number* of syllables per word is represented by $Y_1+Y_2+Y_3+\ldots+Y_N$. The probability generating function $C(t)$ for such a 'contagious' distribution is calculated as follows (it is assumed that all random variables are mutually independent):

$$C(t) = E(t^{Y_1+Y_2+\ldots+Y_N}) = E_N\,(E(t^{Y_1+Y_2+\ldots+Y_N})|N)$$

$$= E_N\,([H(\mathrm{t})]^N) = G(H(t)). \tag{4}$$

Our result, then, is that the pgf's of the two "composing" distributions simply *concatenate* (functional composition).

In what follows, I shall, for reasons of simplicity, assume that both N and Y_i are Poisson-distributed. Of course, independent reasons for this decision are still needed and will still have to be provided by a synergetic approach as sketched above. Thus, in the case of the simple Poisson distribution, $g(x)$ in (2) will be μ/x, where μ is the expectancy value of the random variable. Since in Inuktitut, any word consists of at least one morpheme and almost all morphemes comprise at least one syllable, it is reasonable to use the simple Poisson distribution in its *one-displaced form* in both cases. Our word length distribution then comes out as

$$G(H(t)) = poi_b^{\,*}(poi_m^{\,*}(t)), \tag{5}$$

where $poi_b^{\,*}(t)$ is the pgf of a one-displaced simple Poisson distribution with parameter b. Note that the two parameters b and m now indeed receive a direct linguistic interpretation: b is the average word length in terms of morphemes *minus* 1, and m is the average morpheme length in terms of syllables *minus* 1. Written out explicitly, we thus have:

$$G(H(t)) = t{\cdot}e^{\,m(t-1)} \cdot e^{\,b(te^{m(t-1)}-1)} \tag{6}$$

If we had chosen the non-displaced variants of the Poisson distribution instead, we would have obtained the well-known two-parameter *Neyman distribution type A*. Note, however, that the distribution in (6) does not belong to the Neyman family of distributions.

Obtaining an explicit representation of the distribution in (6) is a bit more cumbersome. I shall merely give a coarse and very informal outline of how to achieve this here. If we consider, for example, words with 3 morphemes, word length is given as the *sum* of *three* identically distributed, independent random variables Y, each of which is represented by pgf $H(t)$. As the pgf's of *added* independent random variables *multiply*; the probability P (morphemes = 3; morpheme-length-parameter = m; syllables = i) will be

$$P\text{ (morphemes = 3, morpheme-length-par .=} \\ m\text{, syllables} = i) = \{[H(t)]^3\}^{(i)}|_{t=0}/i!. \qquad (7)$$

To obtain the probability that a word has x syllables, we simply sum up these P's for all possible morpheme numbers i multiplied by the probability that a word has in fact i morphemes:

$$P\text{ (word-length-par.} = b\text{, morpheme-length-} \\ \text{par .=} m\text{, syllables} = x) = \mathrm{P}_x^{\,b,m} \qquad (8)$$

$$\sum_{i=1}^{x} (Poisson^{\bullet}_{b}\,(i) \cdot P\ (morphemes = i,\ morph. \\ length.par. = m,\ syllables = x)) =$$

$$\frac{e^{-b}}{X!}\sum_{i=1}^{x}\binom{x}{i} \cdot i \cdot b^{i-1} \cdot (im)^{x-i} \cdot e^{-im}$$

The following recursive representation can be found for the distribution in (8):

$$\mathrm{P}_{x+1}^{\,b,m} = \frac{1}{x}$$

$$\left(mP_x^{\,b,m} + be^{-m}\sum_{k=1}^{x}\frac{(x-k+1)m^{x-k}}{(x-k)!} \cdot P_k^{\,b,m} \right). \qquad (9)$$

I will just mention a further refinement of our mathematics. So far, we have been assuming that morpheme length (in terms of syllables) is *independent* of word length (in terms of morphemes). This assumption, however, runs counter to Menzerath's Law: We should rather expect *shorter* morphemes in *longer* words:

$$m_i = a \cdot e^{-ci}, \qquad (10)$$

where a and c are constants and m_i is the morpheme length parameter in words with i morphemes. If we set a to 1 (for simplicity) and replace m in the second line of (8) by the 'relativised' m_i of (10) we obtain:

$$\mathrm{P}_x^{\,b,c} = \frac{e^{-b}}{x!}\sum_{i=1}^{x}\binom{x}{i} \cdot i^{x-i+1} \cdot b^{i-1} \cdot (e^{-ci})^{x-i} \cdot e^{-i \cdot e^{-ci}} \qquad (11)$$

So far, the distributions (8) and (11) discussed in this paper have been applied only to a rather restricted set of linguistic data. As there is no satisfactory fitting algorithm for the distributions proposed available at the moment, the results listed below must be considered as preliminary and are very likely to be improved considerably as soon as better, iterative fitting techniques are used.

(1) The probability distribution given in (6, 8) could be fitted to 28 out of 33 Inuktitut narratives examined (all taken from Nungak, Arima 1969), where $P(X^2) \geq 0.01$; the distribution was well fittable ($P(X^2) \geq 0.05$) to 21 texts.

(2) The probability distribution given in (6, 8) could be fitted to 19 out of 26 German texts as found in Altmann & Best (1996), where $P(X^2) \geq 0.01$; the distribution was well fittable ($P(X^2) \geq 0.05$) to 18 texts. All texts that did not work contained very long words (more than seven syllables), which possibly points to a specific, as yet unclear, reason for their non-fit.

(3) The probability distribution given in (11) could be fitted to 27 out of 33 Inuktitut narratives counted, where $P(X^2) \geq 0.01$; the distribution was well fittable ($P(X^2) \geq 0.05$) to 23 texts.

The morphology-based account has thus indeed led to some empirically testable mathematical assumptions using so-called multiple or 'contagious' Poisson distributions that may be of interest in other areas of quantitative linguistics as well. It is nevertheless questionable whether the specific theoretical *motivations* given for the above distributions are valid. In order to check this, it will be necessary to calculate the *real average* word and morpheme lengths to be found in the texts examined.

REFERENCES

Altmann, G., & Best, K.-H. (1996). Zur Länge der Wörter in deutschen Texten. In P. Schmidt (Ed.), *Glottometrika 15* (pp. 166–180). Trier: Wissenschaftlicher Verlag Trier.

Nungak, Z., & Arima, E. (1969). *Unikkaatuat sanaugarngnik atyingualiit Puvirngniturngmit. Eskimo Stories from Povungnituk, Quebec, illustrated in soapstone carvings.* Ottawa.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G.. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics, 1*, 98–106.