Journal of Quantitative Linguistics
2005, Vol. 12, No. 1, pp. 53 – 63
DOI: 10.1080/09296170500055335

Routledge
Taylor & Francis Group

Morpheme Length Distribution in Lakota*

Regina Pustet[1], and Gabriel Altmann[2]
[1]University of Munich, [2]University of Bachum

ABSTRACT

In this article it will be shown that morpheme length in Lakota obeys a special multimodal distribution resulting from a difference equation of second order. This is due to the particular structure of Lakota syllables, which provide the building blocks for morphemes.

INTRODUCTION

In this paper, a specific type of discourse analysis of the kind initiated by Zipf (1965a, 1965b) is conducted. The object of investigation is the frequency distribution of linguistic items as defined by their length. The discourse data used come from the Native American language Lakota (Siouan language family, Central North America). The present study differs from most other current Zipf-style analyses in several ways:

(a) The count is taken from a unique text, which warrants its homogeneity. Textual homogeneity is important since mixing (pooling) of texts can cause superposition of frequencies and generation of inadequate data (cf. Altmann, 1992).
(b) The discourse items counted are morphemes rather than words. Previous Zipf-style analyses operate almost exclusively with word counts; morpheme lengths have only been scrutinized in a few cases up to now (cf. Saporta, 1963; Best, 2000, 2001, 2004).
(c) Zero morphemes have been taken into consideration (cf. also Saporta, 1963). This may have consequences for further investiga-

*Address correspondence to: G. Altmann, Stüttinghauser Ringstr. 44, D-58515 Lüdenscheid. E-mail: pustetrm@yahoo.com

tions. When phonemes, syllables and words are examined, this boundary condition is not given.

(d) The empirical distribution is multimodal, which shows that the general theory of length distributions must contain a ceteris paribus condition which is not fulfilled in Lakota morphemes.

It is, of course, possible that not Lakota as a language but only Lakota morphemes contain a special condition which is not incorporated in previous theoretical approaches. Lakota morphemes are composed mostly of CV syllables, and merely a small portion of the morphemes included in the analysed text contain Ø, V, C, or CCV syllables. Morphemes which are between one and four phonemes long have been investigated in detail in this respect. The list of individual morphemes in these length classes which occur in the text comprises 232 syllables which exhibit a CV structure, but only 92 syllables which display other syllable structures, i.e. V, C, and CCV.

These characteristics of Lakota syllable structure automatically lead to a multimodal distribution having the modes at even values of the variable with blurring at higher values of X.

Potential multimodal models have been used several times in length research, e.g. by P. Meyer (1997, 1999), who developed his own model for Inuktitut words, the result being the convolution of the 1-displaced Poisson and the Thomas distribution; the Hermite (or Hirata-Poisson) distribution, being the convolution of the Poisson and the Poisson doublet distributions, has been used by Stark (2001), Dieckmann and Judt (1996), Feldt, Janssen and Kuleisa (1997), Altmann, Best and Wimmer (1996), Knopp (1998), Riedemann (1997); for other than word length purposes it was used by Suhren (2002). Its genesis has been shown in Wimmer and Altmann (1996), but it was not used for an empirical multimodal case because all empirical distributions examined were unimodal. Both distributions are special cases of the generalized Poisson family.

METHOD

The linguistic items which serve as the basic units of the discourse frequency counts conducted in the experiment described below are morphemes. For the purpose of this investigation, the notion of morpheme is defined as follows: a morpheme is an element that is semantically and structurally autonomous in that it constitutes an invariant semantic and structural unit and is thus separable from adjacent elements in discourse. Both grammatical items and lexical roots are classed as morphemes. Lakota is a mildly polysynthetic language, and therefore exhibits quite complex aggregates of morphemes in discourse. Example (1), which is taken from the discourse sample analysed within the present study, can be broken down into ten morphemic constituents:

tákuwe	*c^ha*	*a-má-ya-luš̊tą-pi-Ø-sni*	*hé?*	(1)
why	LK	on-1SG.PAT-2AG-stop.2AG-PL.AG-PRS-NEG	QS	

"why don"t you leave me alone (stop on me)?"

The above definition of the morpheme, which relies heavily on the notion of structural unity, treats fusional elements, i.e. elements which simultaneously express more than a single semantic concept but which cannot be broken down into structural subunits, as monomorphemic. The verb form given as example (2) below codes the semantic concepts of "to say" and "first person singular agent", but the form *ep^há* "I say" is a single, compact structural unit that cannot be analyzed into separate structural elements which convey the meanings "to say" and "first person singular agent", respectively. Put differently, the form *ep^há* "I say" is part of the inflectional paradigm of an irregular verb, i.e. *eyá* "to say".

ep^há (2)
say.1SG.AG
"I say"

The form *lustą* "you finish" in (1) is a parallel example of a fusional structure.

The method of discourse analysis employed in what follows proceeds by first subsuming the linguistic items occurring in the discourse sample into classes on the basis of their length as defined by the number of phonemes they contain. It should not go unmentioned here that in measuring the phonemic complexity of linguistic items, secondary articulatory features, such as aspiration and glottalization for consonants and nasalization for vowels, are not taken into account. Consequently, a phoneme with secondary features is counted as a single phoneme, just like a phoneme that lacks these features.

The next analytical step consists in determining the individual discourse frequencies of all elements contained in each length class for the text sample investigated, and in adding up all figures obtained within each length class. For instance, in the case of the Lakota text dealt with in this study, in the class of morphemes comprising 11 phonemes, the following elements are included:

Table 1. Length Class 11 in the Lakota Text Sample.

	Translation	Length in phnemes	Discourse frquency
héktakiyata	"backwards"	11	1
mninát^{h}*akapi*	"reservoir"	11	1
wágleyutapi	"table"	11	1
wik^{h}*ós̊kalaka*	"young woman"	11	2

As can be gleaned from Table 1, for the text investigated, the length class 11 yields a total of 5 occurrences.

The discourse sample chosen for the purpose of the present study is taken from the Lakota text collection Pustet (forthcoming). It is a story from the life of the narrator, a Lakota full-blood who was 70 years old at the time of data compilation. This discourse sample is composed of 5347 phonemes, which are distributed over a total of 1933 morphemes. The discourse frequency values for the morphemes that make up this

Table 2. Morpheme Length Distribution in Lakota.

x	f_x
0	461
1	57
2	524
3	169
4	370
5	106
6	115
7	41
8	50
9	47
10	12
11	5
12	1
13	1

narrative are presented in Table 2; the morphemes, which are not listed individually, have been subsumed in classes according to their length in phonemes.

Within the grammatical systems of natural languages, zero morphemes are frequently encountered, i.e. morphemes which lack phonetic substance and, therefore, must be ascribed a length value of 0. Zero morphemes occur in Lakota as well. Examples are the markers for non-future tense and for third person agent.

ANALYSIS

The great majority of "classical" approaches to length distributions starts from the assumption that there is a simple control of the length class x by its neighbouring class x-1, namely

$$P_x = g(x)P_{x-1} \tag{3}$$

where $g(x)$ is a proportionality function. Its general form can be found in Wimmer and Altmann (2004). Evidently, this approach holds if the *ceteris paribus* condition is fulfilled, i.e. if there is merely one simple control regime. But in Lakota the form of morphemes imposes another condition, namely the control of P_x also by P_{x-2} because of the predominantly biphonemic form of syllables (see above). This would lead to

$$P_x = g(x)P_{x-1} + h(x)P_{x-2} \tag{4}$$

i.e. a difference equation of second order. From this point on one must proceed inductively because there are still no reasons to choose a special form of $g(x)$ and $h(x)$. One could begin with two constants, i.e. to set $P_x = aP_{x\text{-}1} + bP_{x\text{-}2}$ and complicate the calculation stepwise, but the amount of work required by this procedure would be enormous. Even this simplest formula yields quite complicated results both theoretically and practically. Hence we choose a simpler way and begin with the Hirata-Poisson or Hermite distribution, which, merely being reparametrizations of one another, can be fitted mechanically by means of the existing software (Fitter, 1997).

The probability generating function of this distribution (in Hermite form) is

$$G(t) = \exp[a(t-1) + b(t^2-1)] \tag{5}$$

It is obvious that the simple Poisson-regime given with $G(t) = \exp(a(t\text{-}1))$ is amplified (modified) by the double-Poisson corresponding to the above-mentioned form of morphemes. From this function the probability distribution, the moments, the recurrence formula etc. can easily be derived. The Hirata form follows from the transformation $a = a'(1\text{-}b')$, $b = a'b'$ yielding (written again with a, b)

$$G(t) = \exp\{a[t(1-b) + bt^2 - 1]\}. \tag{6}$$

It should be noted that both distributions are not only convolutions but also compound and generalized distributions, a fact that may facilitate future interpretation and systematisation.

The Lakota data are shown in Table 2. The fitting of (6) to this data yielded a good result, which is graphically presented in Figure 1, but there is a strong discrepancy between the first two classes, though the distribution, all in all, follows the empirical trend. The class $x = 6$ is not very relevant, since here and further down the two regimes flow together. Of course, a chi-square test, when applied to this large sample, signalizes a bad fit, hence we took the analysis a step further and used a more general distribution with three parameters, namely the Gegenbauer distribution, though there are various other possibilities (cf. Wimmer & Altmann, 1999). This distribution is also contained in the software (cf.

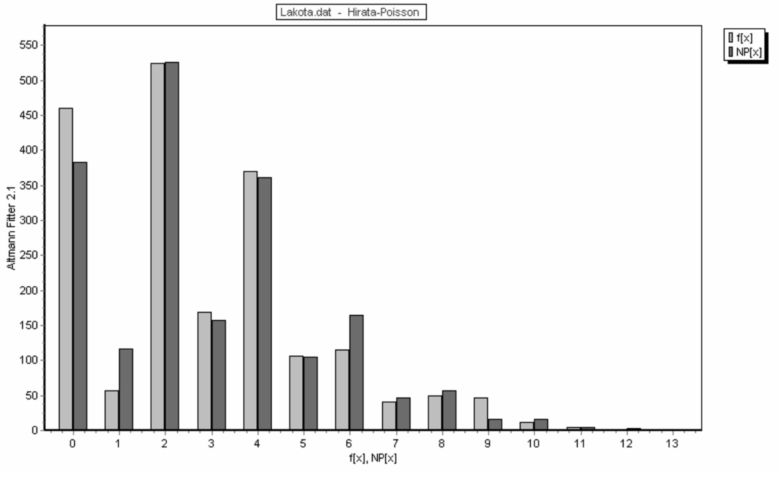

Fig. 1. Fitting the Hermite/Hirata-Poisson distribution to the Lakota data.

Fitter, 1997), but for the sake of further development we show it in more detail in what follows.

The probability generating function of the Gegenbauer distribution is

$$G(t) = (1-a-b)^k(1-at-bt^2)^{-k} \tag{7}$$

and it can be shown that the Hermite and Hirata distributions are its limiting cases (cf. Wimmer & Altmann, 1999). The probability (mass) function is

$$P_x = \begin{cases} (1-a-b)^k, & x = 0 \\ P_0 \sum_{j=0}^{[x/2]} \dfrac{b^j k^{(x-j)} a^{x-2j}}{j!(x-2j)!}, & x = 1, 2, 3, \ldots \end{cases} \tag{8}$$

where $[z]$ is the integer part of z and $k^{(x-j)}$ is the ascending factorial function. The factorial cumulants useful for winning point estimators are given by

$$\kappa_{(r)} = k(r-1)! \sum_{j=0}^{r-1} \binom{r}{j} \frac{(a+2b)^{r-2j} b^j}{(1-a-b)^{r-j}} \tag{9}$$

and the recurrence formula can also be derived from (7) in the following way. The first derivation of (7) is

$$\begin{gathered} G'(t) = \frac{(1-a-b)^k(1-at-bt^2)^{-k}k(a+2bt)}{1-at-bt^2} \Rightarrow \\ \Rightarrow (1-at-bt^2)G'(t) = k(a+2bt)G(t) \end{gathered} \tag{10}$$

Since $G(t) = \sum_x P_x t^x$ and $G'(t) = \sum_x x P_x t^{x-1}$, we can write (10) as

$$\sum_x x P_x t^{x-1} - a \sum_x x P_x t^x - b \sum_x x P_x t^{x+1} = ak \sum_x P_x t^x + 2bk \sum_x P_x t^{x+1}.$$

Equating the coefficients of t^{x-1} on both sides and rearranging yields

$$P_x = \frac{a(k+x-1)P_{x-1} + b(2k+x-2)P_{x-2}}{x}, \quad x = 2, 3, \ldots \tag{11}$$

which is a special case of the assumed difference equation in (4). The first two values are (using 5 or 6)

$$P_0 = (1 - a - b)^k \text{ and } P_1 = (1 - a - b)^k ak. \tag{12}$$

Table 3. Fitting the Gegenbauer Distribution to the Lakota Data.

x	f_x	NP_x
0	461	428.45
1	57	108.52
2	524	528.33
3	169	139.36
4	370	346.29
5	106	94.67
6	115	160.32
7	41	45.23
8	50	58.80
9	47	17.05
10	12	18.17
11	5	5.40
12	1	4.92
13	1	3.49

$a = 0.015283$, $b = 0.072350$, $k = 16.573684$ $C = 0.0289$

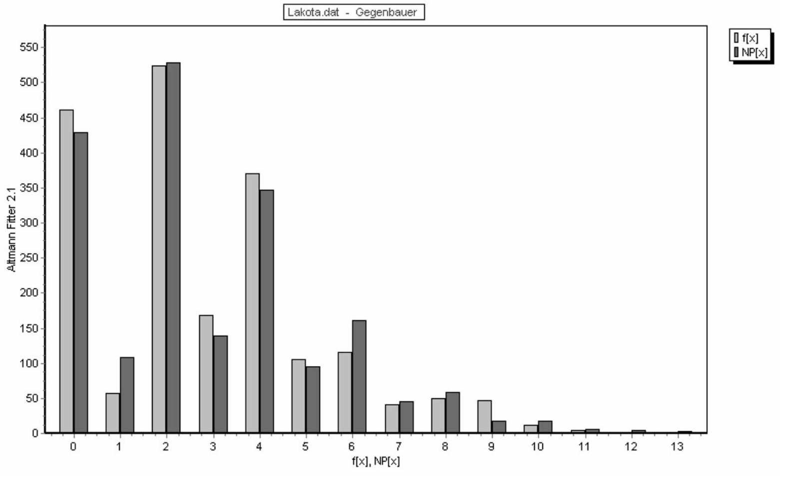

Fig. 2. Fitting the Gegenbauer distribution to the Lakota data.

Since the distribution is contained in the software, which computes the probabilities iteratively, we do not need point estimators. The results are shown in Table 3 and graphically in Figure 2. Though some classes still display a considerable deviation, the results show that the direction this analytical approach takes may be correct.

The following caveats are in order here: (a) The test is preliminary and must be performed on many additional text samples and on languages other than Lakota. (b) In order to obtain a better fit of the first classes, estimators using these classes can be applied. They yield somewhat complex formulas, e.g. using the frequencies f_0 to f_3 we obtain

$$\hat{k} = \frac{1}{2}\,\frac{6P_0P_1P_2 - 3P_1^3 \pm \sqrt{P_1^6 - 12P_0P_1^4P_2 + 36P_0^2P_1^2P_2^2 - 24P_0^2P_1^3P_3}}{6P_0^2P_3 - 6P_0P_1P_2 + 2P_1^3} \tag{13}$$

where P_i is estimated as $P_i = f_i/N$. The other parameters are given as

$$\hat{a} = \frac{P_1}{\hat{k}P_0}, \quad \hat{b} = \frac{2\hat{k}P_0P_2 - \hat{k}P_1^2 - P_1^2}{\hat{k}^2P_0^2} \tag{14}$$

However, these formulas can be used only if the parameters a and b fulfil the condition $0 < a + b < 1$, which in this empirical case is not given.

The chi-square test yields 56.64 with 6 df, which, of course, indicates high significance, but the deviation is, for the most part, caused by a small number of classes; hence we can preliminarily accept the results of this research as an outlook to future projects when more text samples will be available.

Summarizing the results, it can be stated that the original synergetic approach (cf. Altmann & Köhler, 1996) is promising and can easily be extended to a more ample control or combination of boundary conditions yielding superimposed sequences.

ABBREVIATIONS

1	first person
AG	agent
LK	linker
NEG	negative
PAT	patient
PL	plural

PRS	present
QS	question marker
SG	singular

REFERENCES

Altmann, G. (1992). Das Problem der Datenhomogenität. In B. Rieger (Ed.), *Glottometrika 13* (pp. 287 – 298). Bochum: Brockmeyer.

Altmann, G., Best, K.-H., & Wimmer, G. (1996). Wortlänge in romanischen Sprachen. In A. Gather & H. Werner (Eds.), *Semiotische Prozesse und natürliche Sprache. Festschrift für Udo L. Figge* (pp. 1 – 13). Stuttgart: Steiner.

Altmann, G., & Köhler, R. (1996). "Language Forces" and synergetic modelling of language phenomena. In P. Schmidt (Ed.), *Glottometrika 15* (pp. 62 – 76). Trier: Wissenschaftlicher Verlag Trier.

Best, K.-H. (2000). Morphlängen in Fabeln von Pestalozzi. *Göttinger Beiträge zur Sprachwissenschaft*, *3*, 19 – 30.

Best, K.-H. (2001). Zur Länge von Morphen in deutschen Texten. In K.-H. Best (Ed.), *Häufigkeitsverteilungen in Texten* (pp. 1 – 14). Göttingen: Peust & Gutschmidt.

Best, K.-H. (2004). Morphlänge. In G. Altmann, R. Köhler & R. Piotrowski (Eds.), *Quantitative Linguistik—Quantitative Linguistics. Ein internationales Handbuch*. Berlin-N.Y.: de Gruyter (to appear).

Dieckmann, S., & Judt, B. (1996). Untersuchung zur Wortlängenverteilung in französischen Pressetexten und Erzählungen. In P. Schmidt (Ed.), *Glottometrika 15* (pp. 158 – 165). Trier: Wissenschaftlicher Verlag Trier.

Feldt, S., Janssen, M., & Kuleissa, S. (1997). Untersuchung zur Gesetzmäßigkeit von Wortlängenhäufigkeiten in französischen Briefen und Pressetexten. In K.-H. Best (Ed.), *Glottometrika 16* (pp. 145 – 151). Trier:Wissenschaftlicher Verlag Trier.

Knopp, A. (1998). *Wortlängen in französischen Briefen deutscher und französischer Verfasser*. Staatsexamensarbeit, Göttingen.

Meyer, P. (1997). Word length distribution in Inuktitut narratives: empirical and theoretical findings. *Journal of Quantitative Linguistics*, *4*, 143 – 155.

Meyer, P. (1999). Relating word length to morphemic structure: a morphologically motivated class of discrete probability distributions. *Journal of Quantitative Linguistics*, *6*, 66 – 69.

Pustet, R. (forthcoming). *Lakota texts*. Lincoln: University of Nebraska Press.

Riedemann, G. (1997). Wortlängenhäufigkeiten in japanischen Pressetexten. In K.-H. Best (Ed.), *Glottometrika 16* (pp. 180 – 184). Trier: Wissenschaftlicher Verlag Trier.

Saporta, S. (1963). Phoneme distribution and language universals. In J. H. Greenberg (Ed.), *Universals of language* (pp. 61 – 72). Cambridge, MA & London: The MIT Press.

Stark, A.B. (2001). Die Verteilung von Wortlängen in schweizer-deutschen Texten. In K.-H. Best (Ed.), *Häufigkeitsverteilungen in Texten* (pp. 152 – 161). Göttingen: Peust & Gutschmidt.

Suhren, S. (2002). *Untersuchung zum Gesetz von Zwirner, Zwirner und Frumkina am Beispiel des niederdeutschen "De lütte Prinz"*. Staatsexamensarbeit, Göttingen.

Wimmer, G., & Altmann, G. (1996). The theory of word length distribution: Some results and generalizations. In P. Schmidt (Ed.), *Glottometrika 15* (pp. 112–133). Trier: Wissenschaftlicher Verlag Trier.

Wimmer, G., & Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

Wimmer, G., & Altmann, G. (2004). Unified derivation of some linguistic laws. In G. Altmann, R. Köhler & R. G. Piotrowski (Eds.), *Handbook of Quantitative Linguistics*. Berlin: de Gruyter (in print).

Zipf, G. K. (1965a) [1935]. *The psycho-biology of language*. Cambridge, MA: The MIT Press.

Zipf, G. K. (1965b) [1949]. *Human behavior and the principle of least effort*. New York & London: Hafner.

SOFTWARE

Fitter (1997). *Iterative fitting of probability distributions*. Lüdenscheid: RAM-Verlag