Word-Length Distribution in Inuktitut Narratives: Empirical and Theoretical Findings

Peter Meyer
Göttingen

ABSTRACT

This paper deals with the distribution of word length in short native mythological and historical Eskimo narrative texts. To my knowledge, no Eskimo-Aleut data have been the object of quantitative linguistic investigation so far. Due to the strong linguistic and stylistic homogeneity of the examined texts it was assumed that these texts can be subsumed under a single law of word length distribution, if word length distribution of a text is considered as a function of certain of its properties, such as author, language, and genre. So far, word length distribution in texts of a wide variety of languages and genres has been demonstrated to follow distributions of the compound Poisson family of discrete probability distributions. In view of the morphological idiosyncrasies of the Eskimo language in general, which are responsible for an unusually high mean word length of about 4.5 to 5.2 syllables per word in the texts, it is interesting to see whether Eskimo texts show a significantly different behaviour with respect to word length. The results demonstrate that the Eskimo data employed in this study can be fitted well by the Hyperpoisson distribution. Two further discrete probability distributions will be deduced from certain *morphology-based* assumptions about Eskimo. It turns out that most of the Eskimo data can be fitted by these two distributions. The question to what extent these results point to a more grammar-oriented theory of word length is also discussed.

THE DATA

All texts have been taken from a collection of Eskimo Stories from Povungnituk, Quebec (Nungak, & Arima, 1969; the original Eskimo spelling of the toponym is 'Povirngnituq'). A number of very short narratives in the book have not been taken into consideration. The selection of these texts for a quantitative analysis can be justified on several grounds.

First, all texts were produced by native speakers from the town of Povirngnituq, which secures a certain linguistic homogeneity of the texts in question. More specifically, the language of the texts can be determined as belonging to the Nunavik group of the East Canadian Inuktitut branch of the Inuit (Inuktitut) branch of the Eskimo (and, hence, Eskimo-Aleut) language family (cf., the dialectal groupings introduced in Fortescue, Jacobson, & Kaplan, 1994). Apart from geographical considerations, a purely linguistic criterion can be adduced to ascertain the dialectological pertinence to the Nunavik group, viz. the prohibition of two or more subsequent closed syllables (in other words, of sequences CCV(V)CC) within the boundaries of a word, a constraint often referred to as 'Schneider's law'. For example, compare the behaviour of the unpossessed terminalis case ending, the dual number of which is *-nnut*, reducing to *-nut* when following a closed syllable: *arngnaaluunnut maqruunut* ('by two big women', to be found in text 13 in Nungak, & Arima, 1969).

Second, all texts are traditional and usually well-known oral narratives of the area, myths, legends and historical accounts. The texts are either transcriptions of tape recordings or alphabetic transcriptions of text written by the storytellers in standard Eskimo syllabic script. As editorial amendments in the texts were kept to a minimum, the language of the narratives may, with certain restrictions, be said to reflect

Address correspondence to: Peter Meyer, Mittelstraße 2, D-37077 Göttingen, Germany.

Erschienen in: Journal of Quantitative Linguistics 4 (1997) 1-3, S. 143-155

current oral story-telling traditions. Thus, a certain genre-specific, stylistic uniformity of the texts is ensured, although the old story-telling skills have virtually disappeared.

The length of the selected letters is between 47 and 261 words.

ANALYSIS OF THE DATA

General Criteria

Word length is defined here as the number of syllables in a word. In Eskimo, there are several clear-cut and completely coinciding criteria that may serve as defining (token) words in texts. Morphologically, words are either non-inflecting particles to be found as entries in the lexicon or morphologically analyzable entities with a (possibly zero) morph belonging to a closed class of *ending morphemes*. Morphophonologically, contacting morphemes undergo complex word-internal sandhi processes the outcomes of which are, however, not predictable on a purely phonological basis, depending instead on morpheme-specific properties. These contact processes do not occur across word boundaries. Syntactically, inflected words are either nominal or verbal, which determines their role in syntactic structures. In the standard Latin transcription, words are separated by spaces.[1]

Determining the notion of syllable is a more complex matter, however, since the theoretical status of 'diphthongs' (orthographically *ai, au, ia, ua, iu, ui*) and 'long vowels' (orthographically *aa, ii, uu*) is not immediately clear. For example, in the standard Eskimo syllabic sequences like *maa* are denoted by *two* symbols (*ma-a*): does this imply a syllable boundary between the two 'morae' *ma* and *a*? As for the Inuit languages, not much discussion of these questions has taken place so far. Things are different for the Yup'ik branch of Eskimo, where any explanation of ('rhythmic') lengthening and accentuation phenomena hinges crucially on a concept of syllable. Nevertheless, different descriptive approaches are possible even in this case. Thus, Jacobson (1995) offers two different ways of describing suprasegmental phenomena in his grammar of Central Alaskan Yup'ik, which, in a way, may be understood as two different views on the notion of syllable. As the decision should refer to phonological reasoning only, I shall assume here that, in the Nunavik dialect, any orthographical sequence (C)V(V)(C) represents one and only one syllable: that is, 'long vowels' and 'diphthongs' are always tautosyllabic (never have an internal syllable boundary). This takes account of 'Schneider's law' mentioned above, which is thus interpreted as a constraint on possible sequences of syllables. This happens to coincide with the notion of syllable adopted in the above-mentioned Yup'ik grammar by Jacobson (1995) and is accepted, as it seems, by the majority of Eskimo linguists. No satisfactory syllabic interpretation has been found, however, for the exceedingly rare combination of *three* vowel graphemes, *iaa, uii* and *uaa*.[2] I have treated them as bisyllabic. Statistically, these trigraphs occur less than once per text, however.

In all cases, only the running text without headline has been considered.

Findings

With the exception of only one text, all 33 narratives could be fitted by the 1-displaced and 2 -displaced Hyperpoisson distribution, one of the discrete probability distributions used most often for modelling word length. Its formula is as follows:

$$P_x = \frac{a^{x-1}}{b^{(x-1)}{}_1F_1(1;b;a)}, x = 1,2,\ldots$$

and (1)

$$P_x = \frac{a^{x-2}}{b^{(x-2)}{}_1F_1(1;b;a)}, x = 2,3,\ldots$$

[1] In Eskimo, ending morphemes may be followed by one or more *enclitics* belonging to the same word, which are, in Inuit alphabetical rendering, not separated from the word they follow, whereas they are set off by hyphens in standard Yup'ik orthography. These enclitics form a very small closed group in all Eskimo languages and are subject to word internal sandhi

[2] As a matter of fact, sequences of more than two vowels should not occur in the Nunavik dialect at all. In two cases, the trigraphs can indeed be shown to be misprints.

The Altmann-Fitter (1994) was used with a standard chi-square test. A fit with $P(X^2) \geq 0.05$ was considered as good, which is the case for 31 out of 33 texts. Fits with $0.05 \geq P(X^2) \geq 0.01$ were evaluated as still acceptable; this holds for 2 out of 33 texts. Obviously, then, the Hyperpoisson distribution proves to be an acceptable way of modelling word length in the Eskimo text genre in question. Note, however, that in a number of texts extensive pooling of small classes was inevitable to achieve a satisfactory fit, as can be read off the respective number of degrees of freedom. For some Eskimo-internal cross-linguistic comparison, three Central Alaskan Yup'ik narrative texts taken from Jacobson (1995) have been examined (see the last three tables below). Though the Yup'ik and Inuit languages differ considerably in grammar and lexicon, even these control texts could be fitted by the Hyperpoisson distribution, all with $P(X^2) \geq 0.05$. Of course, further investigation with respect to different dialects and text genres will be necessary for a more encompassing picture of Eskimo.

The Inuktitut texts are numbered following Nungak and Arima (1969).

x = word length
f_x = observed frequency
NP_x = the calculated frequency according to the Hyperpoisson distribution
a and b = parameters
X^2 = chi-square test value
DF = degrees of freedom
P = probability

Text 1.

x	f_x	NP_x
2	18	18.20
3	43	39.27
4	47	49.55
5	42	44.18
6	30	30.46
7	18	17.11
8	10	8.11
9	3	3.33
10	1	1.79
Sums	212	212.00

Note. $a = 3.0399$; $b = 1.4093$; $X^2 = 1.436$; $DF = 6$; $P(X^2) = 0.9637$.

Text 2.

x	f_x	NP_x
2	4	2.66
3	7	6.49
4	9	9.82
5	10	10.76
6	6	9.24
7	12	6.52
8	4	3.91
9	1	3.60
Sums	53	53.00

Note. $a = 3.9734$; $b = 1.6273$; $X^2 = 8.429$; $DF = 5$; $P(X^2) = 0.1341$.

Text 3.

x	f_x	NP_x
2	18	16.43
3	20	23.00
4	24	24.57
5	28	21.23
6	13	15.40
7	9	9.62
8	2	5.28
9	4	2.58
10	2	1.89
Sums	120	120.00

Note. $a = 4.5151$; $b = 3.2253$; $X^2 = 5.949$; $DF = 6$; $P(X^2) = 0.4289$.

Text 5.

x	f_x	NP_x
2	7	8.26
3	24	19.33
4	20	23.49
5	20	19.28
6	9	11.94
7	9	5.94
8	2	2.47
9	1	1.29
Sums	92	92.00

Note. $a = 2.5289$; $b = 1.0813$; $X^2 = 4.306$; $DF = 5$; $P(X^2) = 0.5062$.

Text 6.

x	f_x	NP_x
2	5	6.99
3	18	15.28
4	19	18.98
5	16	16.48
6	10	10.99
7	6	5.95
8	4	2.71
11	1	1.62
Sums	79	79.00

Note. $a = 2.8810$; $b = 1.3190$; $X^2 = 1.988$; $DF = 5$; $P(X^2) = 0.8508$.

Text 7.

x	f_x	NP_x
2	7	6.51
3	18	16.74
4	17	20.98
5	18	17.39
6	15	10.76
7	5	5.32
10	1	3.30
Sums81	81.00	

Note. $a = 2.4463$; $b = 0.9513$; $X^2 = 4.175$; $DF = 4$; $P(X^2) = 0.3829$.

Text 8.

x	f_x	NP_x
2	12	11.12
3	22	26.09
4	38	31.70
5	24	25.99
6	15	16.08
7	7	7.99
8	5	3.31
9	0	1.18
10	1	0.54
Sums	124	124.00

Note. $a = 2.521$; $b = 1.0747$; $X^2 = 3.444$; $DF = 5$; $P(X^2) = 0.6319$.

Text 9.

x	f_x	NP_x
1	1	1.07
2	10	10.78
3	22	21.66
4	27	24.20
5	18	18.71
6	8	11.07
7	7	5.30
8	3	3.21
Sums	96	96.00

Note. $a = 2.5150$; $b = 0.2515$; $X^2 = 1.825$; $DF = 5$; $P(X^2) = 0.8728$.

Text 10.

x	f_x	NP_x
2	20	17.70
3	21	20.02
4	17	20.20
5	16	18.39
6	15	15.26
7	12	11.63
8	13	8.19
9	6	5.37
10	4	7.24
Sums	124	124.00

Note. $a = 9.3479$; $b = 8.2656$; $X^2 = 5.479$; $DF = 6$; $P(X^2) = 0.4840$.

Text 11.

x	f_x	NP_x
2	13	11.89
3	18	16.47
4	15	18.31
5	13	17.01
6	21	13.57
7	7	9.48
8	6	5.90
9	4	3.30
10	1	1.68
12	1	1.39
Sums	99	99.00

Note. $a = 5.6470$; $b = 4.0784$; $X^2 = 7.020$; $DF = 7$; $P(X^2) = 0.4268$.

Text 12

x	f_x	NP_x
2	17	15.33
3	22	17.26
4	13	17.85
5	8	17.05
6	16	15.14
7	16	12.55
8	14	9.76
9	11	7.15
10	4	4.95
11	1	3.25
12	2	2.02
13	1	2.69
Sums	125	125.00

Note. $a = 12.5807$; $b = 11.1698$; $X^2 = 15.250$; $DF = 9$; $P(X^2) = 0.0843$.

Text 13.

x	f_x	NP_x
2	19	23.81
3	34	25.59
4	15	24.30
5	21	20.67
6	23	15.92
7	9	11.21
8	10	7.26
9	3	4.36
10	2	2.43
11	1	1.27
12	1	1.18
Sums	138	138.00

Note. $a = 8.1550$; $b = 7.5865$; $X^2 = 12.485$; $DF = 8$; $P(X^2) = 0.1308$.

Text 14.

x	f_x	NP_x
1	2	9.00
2	27	18.02
3	26	26.67
4	34	31.30
5	29	30.43
6	18	25.26
7	24	18.29
8	13	11.74
9	3	6.77
10	5	3.54
11	2	1.70
12	1	1.28
Sums	184	184.00

Note. $a = 5.6711$; $b = 2.8321$; $X^2 = 4.851$; $DF = 5$; $P(X^2) = 0.4344$.

Text 15.

x	f_x	NP_x
2	13	10.88
3	13	17.03
4	19	17.99
5	14	14.34
6	12	9.18
7	5	4.91
8	1	2.25
9	1	1.42
Sums	78	78.00

Note. $a = 3.2508$; $b = 2.0769$; $X^2 = 3.100$; $DF = 5$; $P(X^2) = 0.6846$.

Text 16.

x	f_x	NP_x
2	13	11.78
3	22	19.95
4	20	23.03
5	21	20.18
6	13	14.25
7	8	8.42
8	7	4.28
9	1	3.11
Sums	105	105.00

Note. $a = 3.6349$; $b = 2.1479$; $X^2 = 4.014$; $DF = 5$; $P(X^2) = 0.5474$.

Text 17.

x	f_x	NP_x
2	24	6.63
3	11	10.59
4	5	13.94
5	10	15.63
6	11	15.26
7	8	13.19
8	17	10.23
9	5	7.19
10	7	4.62
11	2	2.74
12	3	2.98
Sums	103	103.00

Note. $a = 7.5444$; $b = 4.7286$; $X^2 = 5.793$; $DF = 1$; $P(X^2) = 0.0161$.

Text 18.

x	f_x	NP_x
2	11	9.70
3	20	17.65
4	18	20.71
5	15	17.94
6	13	12.32
7	11	7.00
8	3	5.68
Sums	91	91.00

Note. $a = 3.3102$; $b = 1.8206$; $X^2 = 4.870$; $DF = 4$; $P(X^2) = 0.3009$.

Text 19.

x	f_x	NP_x
2	5	9.22
3	15	11.10
4	7	10.16
5	4	7.50
6	4	4.63
7	8	2.46
8	4	1.93
Sums	47	47.00

Note. $a = 3.8107$; $b = 3.1645$; $X^2 = 5.026$; $DF = 1$; $P(X^2) = 0.0250$.

Text 21.

x	f_x	NP_x
2	49	41.93
3	45	63.75
4	73	62.77
5	51	45.70
6	24	26.39
7	11	12.63
8	7	5.16
9	0	1.84
10	1	0.83
Sums	261	261.00

Note. $a = 2.7941$; $b = 1.8379$; $X^2 = 9.421$; $DF = 4$; $P(X^2) = 0.0514$.

Text 22.

x	f_x	NP_x
2	22	20.43
3	23	24.47
4	17	23.27
5	27	18.36
6	13	12.37
7	4	7.28
8	4	3.80
9	2	1.78
10	1	1.24
Sums	113	113.00

Note. $a = 4.6247$; $b = 3.8616$; $X^2 = 7.545$; $DF = 6$; $P(X^2) = 0.2733$.

Text 23.

x	f_x	NP_x
2	12	8.95
3	4	10.24
4	10	10.05
5	8	8.62
6	12	6.58
7	7	4.51
8	1	2.81
9	1	3.24
Sums	55	55.00

Note. $a = 6.8540$; $b = 5.9861$; $X^2 = 2.266$; $DF = 1$; $P(X^2) = 0.1322$.

Text 25.

x	f_x	NP_x
2	15	22.93
3	37	36.27
4	51	41.01
5	22	36.09
6	36	26.00
7	11	15.85
8	10	8.37
9	4	3.90
10	5	1.63
11	0	0.61
12	0	0.21
13	2	0.13
Sums	193	193.00

Note. $a = 3.9677$; $b = 2.5086$; $X^2 = 3.633$; $DF = 1$; $P(X^2) = 0.0567$.

Text 28.

x	f_x	NP_x
2	25	21.51
3	16	24.73
4	30	24.67
5	21	21.73
6	17	17.13
7	21	12.23
8	4	7.97
9	1	4.78
10	3	2.65
11	2	2.60
Sums	140	140.00

Note. $a = 7.5258$; $b = 6.5444$; $X^2 = 7.564$; $DF = 4$; $P(X^2) = 0.1089$.

Text 29.

x	f_x	NP_x
2	21	17.50
3	23	27.19
4	24	29.89
5	20	25.43
6	22	17.64
7	21	10.33
8	5	5.23
9	1	3.79
Sums	137	137.00

Note. $a = 3.7595$; $b = 2.4195$; $X^2 = 7.788$; $DF = 3$; $P(X^2) = 0.0506$.

Text 30.

x	f_x	NP_x
2	12	16.65
3	39	31.01
4	39	37.70
5	34	34.02
6	19	24.41
7	12	14.53
8	9	7.39
9	4	3.28
10	3	2.01
Sums	171	171.00

Note. $a = 3.5016$; $b = 1.8801$; $X^2 = 6.093$; $DF = 6$; $P(X^2) = 0.4128$.

Text 31.

x	f_x	NP_x
2	14	11.52
3	12	15.41
4	15	15.81
5	14	13.16
6	12	9.20
7	4	5.56
8	5	5.34
Sums	76	76.00

Note. $a = 4.4037$; $b = 3.2927$; $X^2 = 2.685$; $DF = 4$; $P(X^2) = 0.6118$.

Text 32.

x	f_x	NP_x
2	35	31.21
3	36	37.23
4	36	38.88
5	26	36.11
6	34	30.20
7	32	22.97
8	17	16.02
9	11	10.31
10	5	6.17
11	2	3.44
12	1	1.80
13	1	1.66
Sums	236	236.00

Note. $a = 8.3992$; $b = 7.0426$; $X^2 = 9.092$; $DF = 9$; $P(X^2) = 0.4288$.

Text 34.

x	f_x	NP_x
2	13	2.57
3	6	18.37
4	25	24.14
5	20	17.48
6	10	8.73
7	1	3.32
8	1	1.39
Sums	76	76.00

Note. $a = 1.6112$; $b = 0.2259$; $X^2 = 2.305$; $DF = 2$; $P(X^2) = 0.3158$.

Text 35.

x	f_x	NP_x
2	15	12.87
3	10	15.92
4	16	17.06
5	18	16.14
6	20	13.65
7	14	10.45
8	5	7.31
9	1	4.70
10	1	2.80
11	2	1.55
12	1	0.80
13	0	0.39
14	1	0.36
Sums	104	104.00

Note. $a = 8.0414$; $b = 6.5026$; $X^2 = 12.070$; $DF = 8$; $P(X^2) = 0.1481$.

Text 39.

x	f_x	NP_x
2	6	5.98
3	14	11.02
4	10	14.98
5	18	16.13
6	12	14.37
7	16	10.92
8	5	7.23
9	4	4.25
10	2	2.24
11	1	1.07
12	1	0.81
Sums	89	89.00

Note. $a = 5.1729$; $b = 2.8064$; $X^2 = 6.175$; $DF = 7$; $P(X^2) = 0.5194$.

Text 40.

x	f_x	NP_x
2	6	9.43
3	8	12.92
4	22	14.37
5	13	13.43
6	8	10.83
7	12	7.68
8	1	4.86
9	6	2.78
11	1	1.45
13	1	0.69
17	1	0.56
Sums	79	79.00

Note. $a = 5.8857$; $b = 4.2957$; $X^2 = 7.323$; $DF = 3$; $P(X^2) = 0.0623$.

Text 41.

x	f_x	NP_x
2	12	11.80
3	16	13.04
4	16	13.21
5	9	12.34
6	10	10.71
7	8	8.67
8	4	6.58
9	5	4.69
10	7	3.16
11	2	4.80
Sums	89	89.00

Note. $a = 12.1036$; $b = 10.9510$; $X^2 = 9.528$; $DF = 7$; $P(X^2) = 0.2169$.

Text 45.

x	f_x	NP_x
2	20	8.20
3	4	12.69
4	13	15.64
5	21	16.03
6	8	14.05
7	9	10.77
8	3	7.32
9	13	4.48
10	0	2.49
11	3	2.33
Sums	94	94.00

Note. $a = 6.0762$; $b = 3.9297$; $X^2 = 5.694$; $DF = 3$; $P(X^2) = 0.1275$.

Text for comparison: Yup'ik narrative (Jacobson, 1995, p. 451)

x	f_x	NP_x
1	12	12.75
2	58	51.94
3	68	75.86
4	68	67.50
5	44	43.19
6	24	21.57
7	10	8.83
8	1	3.06
10	1	1.30
Sums	286	286.00

Note. $a = 2.2774$; $b = 0.5593$; $X^2 = 3.453$, $DF = 6$; $P(X^2) = 0.7503$.

Text for comparison: Yup'ik narrative (Jacobson, 1995, p. 454)

x	f_x	NP_x
1	4	30.00
2	102	60.07
3	77	76.79
4	66	72.09
5	43	53.48
6	33	32.79
7	23	17.13
8	4	7.80
9	2	3.14
10	1	1.71
Sums	355	355.00

Note. $a = 3.5353$; $b = 1.7655$; $X^2 = 5.392$; $DF = 2$; $P(X^2) = 0.0675$.

Text for comparison: Yup'ik narrative (Jacobson, 1995, p. 456)

x	f_x	NP_x
1	9	8.71
2	56	51.55
3	77	76.79
4	52	65.43
5	47	39.04
6	22	17.92
7	4	6.68
8	2	2.88
Sums	269	269.00

Note. $a = 1.9910$; $b = 0.3366$; $X^2 = 7.029$; $DF = 5$; $P(X^2) = 0.2185$.

SOME INTERPRETATION: RELATING LINGUISTIC STRUCTURE AND WORD LENGTH

Previous attempts at explaining the adequateness of certain discrete probability distributions for modelling of word length relied to a large extent on considerations that do not refer to specifically *grammatical* features of the languages in question (cf., the summary on possible approaches in Wimmer, Köhler, Grotjahn, & Altmann (1994)). Thus, one may assume that, in a given language, any class of word length is proportional to all other classes of smaller length, arriving thereby, e.g., at generalized Poisson distributions. However, although such an assumption bears a certain plausibility residing mostly in considerations of mathematical simplicity and previous experience with linguistic synergetics in general, it does not reflect any specifically morphological, syntactic or other properties of the language in question.

In what follows, some perspectives on more grammar-oriented models of word length distribution will be outlined. In particular, I shall draw attention to the *morphemic organisation* of words. Indeed, it is just the strongly agglutinative structure of Eskimo words that suggests such a line of reasoning. In general, any Eskimo word can easily be interpreted as a chain of clear-cut morphemes.

Some brief remarks on the principles of word formation in Eskimo may be in order here. Words either belong to a closed class of uninflected 'adverbial' particles or they are inflected, in which case they are either nouns or verbs. An inflected word consists of a (nominal or verbal) *stem* and a (nominal or verbal) *ending* (E). In the most elementary case, a stem just consists of one morpheme called (nominal or verbal) root or *base* (B), symbolically: $[[B]_{stem} - E]$, e.g. *uiga*[3] HUSBAND-MY:ABS 'my husband' (absolutive case). Usually, however, bases are suffixed by one or more out of several hundred available *postbase* morphemes, here symbolized as P, to form successively more complex stems. Word structure might then be characterized as follows:

$$[[\ldots[[[B - P_1]_{stem\,1} - P_2]_{stem\,2} - P_3]_{stem\,3} - \ldots - P_n]_{stem\,n} - E]$$

Thus, stems may be expanded through recursive right-branching, such that each postbase may be characterized as being the (morphological, semantic) 'regens' of the whole preceding stem and hence the 'head' of the stem that consists of the postbase itself and all preceding morphemes. Postbases are either nominal or verbal in that they determine whether the stem they are head of is nominal or verbal. They are

[3] Hyphens indicate morphemic analysis and are not part of standard alphabetic orthography. Examples in these two paragraphs are taken from the phonologically more regular Igloolik dialect of Inuit.

usually also determined as to whether the stem or naked base they are suffixed to must be verbal or nominal. There is no theoretical limit as to the number of admissible postbases in a word. Endings may be followed by an 'enclitic' suffix, such as *-lu* 'and': *uigalu* 'and my husband'.

For an example, we may take the base *ui-* HUSBAND from above as a starting point and add a 'nominal-to-verbal' postbase *-qaq-* HAVE to get the stem *ui-qaq-* HUSBAND-HAVE 'to have a husband'; cf. *ui-qaq-tunga* HUSBAND-HAVE-1SG:PRES:ITR 'I have a husband' with the ending *-junga*. A 'verbal-to-verbal' postbase such as *-juma-* WANT might then be added to this stem to yield an even more complex stem *ui-qa-ruma-* 'to want to have a husband', as in *uiqarumajungalu* 'I want to have a husband, too' (note the word sandhi found in these examples). This may be continued to form *ui-qa-ruma-laun[g]-ngit-tunga* HUSBAND-HAVE-WANT-PAST-NEG-1SG:PRES:ITR 'I didn't want to have a husband' – or even much larger constructions. This way, Eskimo packs much of what would be sentence structure elsewhere into productive and regular word-internal derivational processes.

Two very simple and plausible principles of word length constitution in Eskimo are postulated:

(1) *Word length* in Eskimo as expressed in terms of *morpheme number* (henceforth called *morphemic distribution*) can be modelled by a (possibly one-displaced) simple Poisson distribution with parameter b.

(2) *Morpheme length* in Eskimo as expressed in terms of *syllable number* (henceforth called *syllabic distribution*) can be modelled by a (possibly one-displaced) simple Poisson distribution with parameter m.

Thus, each word has an average of b morphemes (b+1 in the one-displaced case), with a mean syllable number of m (m+1 in the one-displaced case) per morpheme. I assume mutual independence of all random variables here (the length of a morpheme is influenced neither by the length of other syllables nor by the number of syllables in the word).

The advantages of this approach are obvious. First, the two parameters involved receive a *direct* linguistic interpretation and can, at least in principle (but see conclusion) be verified on the data. Second, the assumptions made here may be viewed as governed by a single principle stating that the length of a certain type of linguistic unit as expressed in terms of the number of its subunits is Poisson-distributed.

I shall now give explicit forms for the word length distribution that is determined by the above postulates, starting with the more involved case of a one-displaced syllabic distribution, where zero-syllable morphemes (which are rare in standard analyses of Eskimo morphology) are disallowed, such that a word of n syllables may consist of any number of morphemes between 1 and n.

Let us denote the probability of word length i (i = 1, 2, 3, ...) with given parameters b and m as explained above by $P(b, m, i)$. Let $\pi(a,x) = e^{-a}\frac{a^{x}}{(x-1)!}$ denote the one-displaced simple Poisson distribution with expectancy value a + 1, where x = 1, 2, ... According to the two general principles proposed above, we then have

$$P(b, m, l) = \sum_{i=1}^{l}\left(\pi(b,i)\cdot \sum_{\substack{(n_1,\ldots,n_i)\\ n_x = 1,2,\ldots\\ \sum n_x = l}} \prod_{x} \pi(m,n_x)\right) \quad (2)$$

This is to be understood as follows. To find, e.g., the probability $P(b, m, 5)$ of a word length of five syllables for given values of m and b, one sums up the probabilities of all 'morpheme configurations' possible for words with five syllables, multiplied by the probability $\pi(b, i)$ of the respective number i of morphemes per word as determined by the configuration. Here a 'morpheme configuration' is simply a sequence of morpheme lengths, such as 2-1-2, meaning 'first a two-syllable morpheme, then a monosyllabic one, finally again a bisyllabic one', which must be distinguished from, e.g., 1-2-2. The probability of such configurations is calculated by multiplying the respective morpheme length probabilities, e.g., for 1-2-2 or 2-1-2: $\pi(m, 1) \times \pi(m, 2) \times \pi(m, 2)$.

In (2), the calculation is split into two parts. We summarize over the number of possible morphemes (viz., 1 to l) in a word of l syllables (outer sum). Within each possible number of morphemes i in the word, the morphemic probability for which is $\pi(b, i)$, the inner sum runs over all possible *ordered i-fold partitions* $\langle n_1, \ldots, n_i \rangle$ of l, that is, over all ordered i-tuples of positive, non-zero integers n_x whose sum is l. (Thus, 1-2-2 is an ordered 3-fold partition of 5.) These partitions represent the available arrangements of i morphemes within a word of l syllables. The probability of each arrangement is, of course, the product of the probabilities for the composing morphemes, which, in turn, only depend on the length of the respective morphemes themselves as determined by the partition in question.

Though (2) gives a straightforward account of our distribution, it is neither easy to handle mathematically nor trivial to calculate. As there are 2^l possible partitions of an integer l, computing times explode exponentially for growing l in (2). As a consequence, we had rather look for some more stringent formula. To begin with, the inner sum in (2), henceforth abbreviated as $\varphi(m, l, i)$, can be interpreted as a probability distribution with parameters m and i, evaluated for value l. This distribution obviously is the *sum* of i mutually independent one-displaced Poisson distributed random variables with parameter m. Since the probability generating functions (pgf) of mutually independent distributions multiply when their random variables are summed up, we have (with $s\,e^{m(s-1)}$ as pgf of $\pi(a, x)$):

$$\varphi(m, l, i) = \frac{d^l}{ds^l}\,(s\,e^{m(s-1)})^i\Big|_{s=0}\,\frac{1}{l!} \qquad (3)$$

In order to further simplify (3), we note that $\frac{d^l}{ds^l}(f(s)\cdot g(s)) = \sum_{x=0}^{l}\binom{l}{x}\cdot f^{(x)}(s)\cdot g^{(l-x)}(s)$ for arbitarily often differentiable functions f and g[4] and that $\frac{d^l}{ds^l}\,s^i\Big|_{s=0}$ will be $i!$ for $i=l$ and simply 0 for all other l. We thus obtain

$$\varphi(m, l, i) = \frac{1}{l!}\sum_{x=0}^{i}\binom{l}{x}(s^i)^{(l-x)}\Big|_{s=0}\cdot(im)^x e^{-im} = \frac{1}{l!}\binom{l}{l-i}\,i^l\,(im)^{l-i}\,e^{-im} \qquad (4)$$

Replacing the inner sum $\varphi(m, l, i)$ in (1) with our last formula in (4), we finally have

$$P(b, m, l) = \sum_{i=1}^{l}(\pi(b,i)\cdot\varphi(m,l,i)) = \sum_{i=1}^{l}\frac{e^{-b}b^{i-1}}{(i-1)!}\cdot\frac{1}{l!}\cdot\frac{l!}{i!(l-i)!}\,i^l\cdot(im)^{l-i}\cdot e^{-im} \qquad (5)$$

After some tidying up our result is a surprisingly simple formula with calculation times growing in a *linear* fashion with growing l, viz.

$$P(b, m, l) = \frac{e^{-b}}{l!}\sum_{i=1}^{l}\binom{l}{i}\,i\cdot b^{i-1}(im)^{l-i}\,e^{-im},\quad l=1,2,3,\ldots \qquad (6)$$

The case of a word length distribution with non-displaced syllabic distribution, here denoted as $P^*(b, m, l)$, is much easier to cope with. Linguistically speaking, P^* may be interpreted as allowing for zero-length morphemes. Thus, if a word has two syllables, there is, according to P^*, a certain, albeit small probability that the word is composed of, say, 137 morphemes, 136 of which do not contain a syllable core. In general, any word is allowed to consist of an arbitrarily high number of morphemes. Instead of (2), we start from the following:

$$P^*(b, m, l) = \sum_{i=0}^{\infty}\left(\pi^*(b,i)\cdot\sum_{\substack{n_x=0,1,2,\ldots\\ \sum n_x=l}}\prod_{x=1}^{i}\pi^*(m,n_x)\right) \qquad (7)$$

Here, $\pi^*(a, x)$ denotes the Poisson distribution, $\pi^*(a, x) = e^{-a}\frac{a^x}{x!}$. For the sake of simplicity, I also assume non-displaced morphemic distribution here. The inner sum in (7), henceforth abbreviated as $\varphi^*(m, l, i)$, may, accordingly, be expressed as

[4] Here, $f^{(x)}(s)$ is, of course, the usual abbreviation of $\frac{d^x}{ds^x}f(s)$, and not a descending factorial.

$$\varphi^*(m, l, i) = \frac{d^l}{ds^l}(e^{m(s-1)})^i\Big|_{s=0} \cdot \frac{1}{l!} = \frac{1}{l!}(im)^l e^{-im} \quad (8)$$

Inserted into (7), this immediately yields

$$P^*(b, m, l) = \sum_{i=0}^{\infty} \frac{e^{-b} b^i}{i!} \frac{1}{l!} i^l m^l (e^{-m})^i \quad (9)$$

After a little bit of reordering, the final formula is

$$P^*(b, m, l) = \frac{e^{-b} m^l}{l!} \sum_{i=0} \frac{i^l (be^{-m})^i}{i!},$$

$$l = 0, 1, 2, \ldots \quad (10)$$

However, this is just the formula of the well-known Neyman distribution type A with two parameters, the use of which for modelling word length has thus been motivated partially on linguistic grounds. Note, however, that $P(b, m, l)$ in (10) is not simply some shifting or reparametrization variant of $P^*(b, m, l)$.

It is clearly of paramount importance to investigate the empirical significance of the two proposed distributions, $P(b, m, l)$ and $P^*(b, m, l)$. A preliminary examination has shown that

- out of 20 Eskimo texts checked, at least 18 may be fitted by $P(b, m, l)$, out of these at least 13 well ($P(X^2) \geq 0.05$);[6]
- out of 20 Eskimo texts counted, at least 18 may be fitted to $P^*(b, m, l)$; out of these at least 13 well ($P(X^2) \geq 0.05$);[7]
- for a large range of possible choices for the parameters m and b, $P(b, m, l)$ is a good approximation to the Hyperpoisson distribution;
- for a large range of possible choices for the parameters m and b, $P^*(b, m, l)$ is a good approximation to the Hyperpoisson distribution.

What is the moral to be drawn from these data? First, $P(b, m, l)$ and $P^*(b, m, l)$ can indeed, to some degree of satisfaction, be regarded as mathematical models of word length in the East Canadian Inuktitut narrative texts under investigation. The small amount of texts examined does not suffice to compare these distributions statistically – they seem to do their job equally well. Second, the data hint at an unusual, but perhaps interesting, way of explaining fit of certain other distributions: Distributions such as the Hyperpoisson one may be "explained" by demonstrating that they approximate and are approximated well by another, well-motivated distribution. Third, the above mathematical reasoning as represented by (2) and (7) defines, if loosely, a family of probability distributions that seems to be a good starting point for word length modelling in non-isolating languages. This family is *not* the Neyman family of distributions, since $P(b, m, l)$ as given in (6) does not belong there. This might be an indication as to which of the several possible *motivations* or *interpretations* for the Neyman A distribution should be preferred in the Eskimo case: If several different distributions belonging to a family F turn out to be acceptable models for the phenomenon to be accounted for, then acceptable motivations for each of these distributions should be extendable to F as a whole.

CONCLUSION

The Hyperpoisson distribution turns out to be a remarkably good model for the Eskimo text genre examined here. By way of theoretical reasoning, two further distributions that reflect the morphemic structure of Eskimo words have been proposed. These distributions, too, have proven to be acceptable models of word length in the Eskimo texts. Further empirical investigation will be necessary to determine whether these or related distributions may be used to model word length in other agglutinative languages.

It should be stressed, however, that the models proposed still do not present anything similar to an *explanation* of the regularities of word length in Eskimo. First, many other possible

[6] The one-displaced variant of the distribution has been employed throughout, with the exception of text 14, the only checked text with a class of monosyllabic words. Results for this distribution were bad only where Hyperpoisson fit was also problematic, at least before pooling classes.

[7] Except for text 14, the two-displaced variant of this distribution was used. Results for this distribution were bad also only where Hyperpoisson fit was also problematic, at least before pooling classes.

parameters (attractors) have simply been left out of consideration, such as the intricacies of word formation in Eskimo hinted at in the section on data analysis; the syntactic and phonological subsystems of the language; and many more. Second, no further justification has been adduced for choosing the simple Poisson distribution, instead of, e.g., the Borel distribution, for both the syllabic and the morphemic distribution. Deeper justifications for this choice will probably have to be founded on general synergetic process modelling.

It might seem that a good way to check the validity of these models is to make sure that the interpretation of their respective parameters is empirically valid. As for $P(b, m, l)$, a preliminary fitting of the data gives parameters that indeed sound realistic (between 1.16 and 3.78 for $b+1$, that is the mean number of morphemes per word; between 1.09 and 3.36 for $m+1$, the mean number of syllables per morpheme). However, the usual linguistic analyses of word structure in Eskimo are not very likely to presuppose a notion of *morpheme* which would be relevant for a stochastic investigation since the role that the morpheme concept has to play in the two different frameworks of 'traditional' and synergetic linguistics are not identical. To mention a trivial example: If the presence of a certain "traditional" morph M_1 conditions a very high probability for it to be followed by another morph M_2 in the very same word, then the sequence M_1M_2 might constitute a single morph from the point of view of a stochastic approach. As a consequence, the very notion of *validating* or *justifying* a synergetic or, more general, stochastic model for a 'real-world' linguistic phenomenon will remain highly theory-dependent and theory-laden. Therefore, probably no simple method of relating probabilistic models of language to 'traditional' linguistic descriptions will await us in the near future.

REFERENCES

Fortescue, M., Jacobson, S.A., & Kaplan, L. (1994) *Comparative Eskimo dictionary With Aleut cognates*. Fairbanks.

Jacobson, S.A. (1995). *A practical grammar of the central Alaskan Yup'ik Eskimo language (with Yup'ik readings written by Anna W. Jacobson)*. Fairbanks.

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.

Software:

Altmann-Fitter: 1994, Lüdenscheid.