Pramana – J. Phys. (2019) 92:60
https://doi.org/10.1007/s12043-018-1709-8

A statistical probe into the word frequency and length distributions prevalent in the translations of Bhagavad Gita

NIKHIL KUMAR RAJPUT[1], BHAVYA AHUJA[1,*] and MANOJ KUMAR RIYAL[2]

[1]Department of Computer Science, Ramanujan College, University of Delhi, New Delhi 110 019, India
[2]Department of Physics, Veer Chandra Singh Garhwali Uttarakhand University of Horticulture and Forestry, Tehri Garhwal 246 123, India
*Corresponding author. E-mail: b.ahuja@ramanujan.du.ac.in

MS received 2 July 2018; revised 8 August 2018; accepted 27 August 2018; published online 15 February 2019

Abstract. A statistical study has been conducted on Bhagavad Gita. Four measures have been derived for the original text in Sanskrit and its translations in Hindi, English and French. First, word frequency distributions for the documents were modelled. Power law was observed with the longest tail in the case of Sanskrit. For other versions, the distributions well replicated the Zipf–Mandelbrot pattern. Second, the Kullback–Leibler (KL) divergence between the documents has been computed with the highest value recorded in all three translations from the Sanskrit text. Next, a Shannon entropy-based measure: vocabulary quotient has been calculated, which estimates the vocabulary richness the texts offer; the highest being in the case of Bhagavad Gita in Sanskrit. Finally, word-length distributions were obtained with the longest word length in Sanskrit. The results attribute to the inflectional nature of Sanskrit.

Keywords. Shannon entropy; power law; word frequency distribution; vocabulary quotient; Kullback–Leibler divergence.

PACS Nos **89.75.Da; 89.70.Cf; 89.90.+n**

1. Introduction

Statistical characterisation of languages and literary works has been one of the intriguing domains for physicists, linguists and statisticians [1]. Predominantly, studying the pattern of frequency distributions of the words in a literary document has been one of the areas of priority [2]. The patterns mostly imitate Zipf's law [3,4], which states that for an array of words x, the word frequency distribution varies as an inverse power of x.

Other distributions such as Zipf–Mandelbrot [5], lognormal [6,7], Gauss–Poisson [6], extended generalised Zipf law [6] have also marked their presence. The language corpus studied is not limited to English, but also includes languages such as Mongolian [8], Chinese [9], Japanese [10], Hindi [11] and many others [2].

Another characteristic feature that has been exhaustively studied and applied is the entropic framework due to Shannon [12]. Entropy-based studies have been carried out for symbolic sequence [13], back-off language models [14], constancy rate principle [15] and several other domains. Long-range correlations based on entropy have been found over two literary texts where the mutual information for the pairs of letters and the entropy of the two documents have also been analysed. A power-law decay in scaling laws for the mutual information, inverse square root of the number of subwords for the entropy per letter and stretched exponential for the word numbers were also observed [16]. Statistical analysis of English literary words was carried out resulting in the creation of a cluster of certain groups of words. A relation was established between the English content words and entropy computed over its probability distribution [17].

In this paper, a statistical analysis of 'Bhagavad Gita', one of the most sacred writings of Hindu religion, has been presented. The literature has previously been extensively studied from various perspectives such as management [18], psychiatry [19], theosophy [20], ethics [21] and the plausible application of its teachings in these domains. The present work is aimed at conducting a statistical characterisation of this text. The effort is interdisciplinary in nature in the sense that it utilises the concepts of statistical physics for the analysis of Bhagavad Gita in its four translations. Measures such as entropy are employed to determine the pattern and randomness in the system (text here). Deduction of power-law distribution in word frequency also connotes

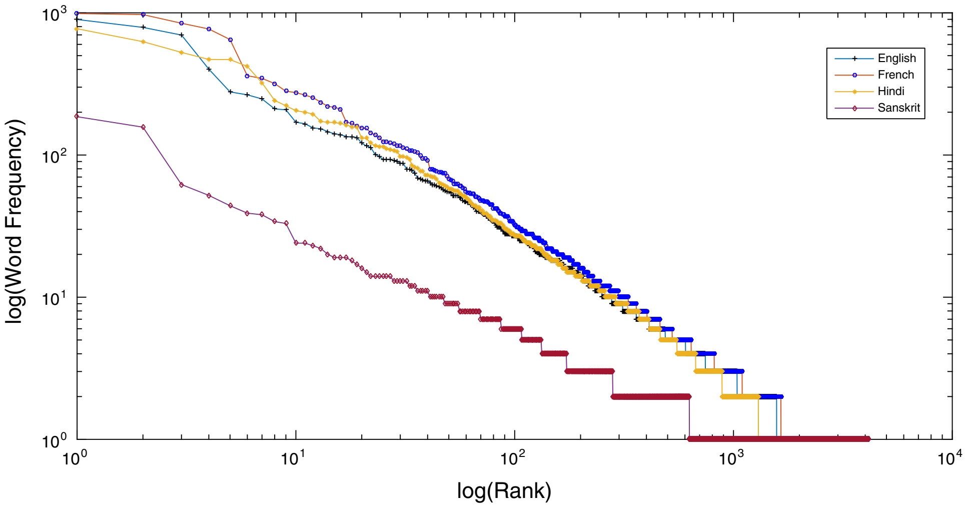

Figure 1. Word frequency distributions (log–log scale).

the statistical patterns in the text. The corpus for the study comprises four versions of the text written in the Indo–European family of languages: Sanskrit, Hindi, French and English. The documents have been obtained from [22] for Sanskrit and Hindi, [23] for French and [24] for English.

2. A statistical study

2.1 *Modelling the word frequency distribution*

Statistical models have been deduced for various literatures and scriptures. For instance, Zipf's law was deployed to model word frequencies in Holy Bible translations for 100 live languages by Mehri and Jamaati which produced a Zipf exponent in the range (0.765–1.442) [25]. The fit of Zipf's law was shown to perform poorly as compared to a number of Pareto-type distributions [26].

The word frequency of the 'Bhagavad Gita' texts in four different languages has been analysed in the present work. Figure 1 depicts the word frequency distributions (on a log–log scale) for the four versions. The curves can be seen to replicate the power-law pattern with the longest tail in the case of Sanskrit version of Bhagavad Gita which can be attributed to its higher number of unique words. The curves representing French, English and Hindi are closer compared to Sanskrit and follow similar pattern of descent as is recorded by the value of Kullback–Leibler (KL) divergence (discussed in the succeeding subsection).

The probability distribution of word frequencies for the four documents has been modelled using four distributions, namely Zipf, Zipf–Mandelbrot [5], Pareto [27] and lognormal [6,7].

Mathematically, Zipf's law can be defined as $p(x) = a/x^b$, where a denotes the normalising constant and b is the Zipf exponent. a and b can be computed empirically for a particular document. Zipf–Mandelbrot as a variation of Zipf's law shows relatively better fit for lower rank values because of the presence of more functional words [28] which can be captured by introducing a new parameter c and takes the form

$$p(x) = \frac{a}{(1+cx)^b}. \tag{1}$$

For comparative analysis, two more distributions have been considered which are Pareto and lognormal distributions. Mathematical form of Pareto distribution in terms of shape parameter a and scale parameter b can be defined as

$$p(x) = \frac{ab^a}{x^{a+1}}, \quad x \geq b \tag{2}$$

and the lognormal distribution with mean μ and standard deviation σ as

$$p(x) = \frac{1}{\sigma\sqrt{(2\pi x)}} \exp \frac{(-\log(x) - \mu)^2}{2\sigma^2}, \quad x > 0. \tag{3}$$

The results for the same have been given in table 1. To validate the goodness of fit for each, sum of squared errors (SSE), R^2 and root mean square errors (RMSE) have been calculated. It was seen that Zipf–Mandelbrot provided the best fit with the lowest values for SSE and RMSE in the case of English, French and Hindi. Also, the value of R^2 obtained was close to 1. In the case of Sanskrit, Zipf model provided the best fit for the data. The value of Zipf's exponent obtained in the four cases was 0.777 for English, 0.754 for French, 0.732 for Hindi and 0.768 for Sanskrit and for Zipf–Mandelbrot, the exponent obtained is close to 1 for English, French

Table 1. Results for modelling of word frequency probability distribution.

	Parameters	SSE	R^2	RMSE
English				
Zipf's Law	$a = 0.05687, b = 0.7766$	0.0003849	0.9511	0.0003363
Zipf–Mandelbrot	$a = 0.08524, b = 0.9747, c = 0.704$	0.0001807	0.977	0.0002304
Pareto	$a = 0.03793, b = 1.945 \times 10^6$	0.001233	0.8433	0.0006018
Lognormal	$\mu = -5.107, \sigma = 1.989$	0.0003315	0.9579	0.0003121
French				
Zipf's Law	$a = 0.05725, b = 0.7537$	0.0008241	0.9075	0.0004795
Zipf–Mandelbrot	$a = 0.062, b = 1.153, c = 0.2741$	0.0001883	0.9789	0.0002292
Pareto	$a = 0.03771, b = 3.087 \times 10^6$	0.002115	0.7628	0.0007681
Lognormal	$\mu = -5.232, \sigma = 2.045$	0.0002417	0.9729	0.0002597
Hindi				
Zipf's Law	$a = 0.05273, b = 0.7322$	0.0005364	0.9305	0.0004345
Zipf–Mandelbrot	$a = 0.05428, b = 1.073, c = 0.2747$	$5.29E{-}05$	0.9931	0.0001365
Pareto	$a = 0.03658, b = 1.751 \times 10^6$	0.001879	0.7566	0.0008131
Lognormal	$\mu = 0.4694, \sigma = 0.0119$	0.008071	−0.04555	0.001685
Sanskrit				
Zipf's Law	$a = 0.02922, b = 0.7679$	$9.96E{-}05$	0.9492	0.0001548
Zipf–Mandelbrot	$a = 6.014, b = 0.7682, c = 1026$	$9.96E{-}05$	0.9491	0.0001548
Pareto	$a = 0.03161, b = 15.53$	0.0002731	0.8606	0.0002563
Lognormal	$\mu = -5.552, \sigma = 1.993$	0.0002798	0.8572	0.0002594

and Hindi. In the case of Hindi, a negative value of R^2 for lognormal shows that the fit was very poor.

2.1.1 *KL divergence among the four versions.* The KL divergence [29,30] gives an asymmetric quantitative measure for the distance between two distributions. For probability distributions p_1 and p_2, the KL divergence D_{KL} is given by

$$D_{\mathrm{KL}}(p_1 \| p_2) = \sum_i p_1(i) \ln \frac{p_1(i)}{p_2(i)}. \tag{4}$$

The KL divergence has been computed between the probability distributions derived for the four versions. Table 2 presents the results obtained for the KL divergence obtained for each pair of languages.

The entries in the last column record the KL divergence of the Sanskrit word frequency distribution from English, French and Hindi which are 0.1638, 0.1913 and 0.2044, respectively. It may be pointed out that these values are larger compared to the KL divergence between English and French; English and Hindi; and French and Hindi.

Table 2. KL divergence for each pair of languages.

	English	French	Hindi	Sanskrit
English	0	0.0071	−0.0044	0.1638
French	0.0060	0	−0.0082	0.1913
Hindi	0.0211	0.0191	0	0.2044
Sanskrit	0.0088	0.0358	0.0172	0

2.2 *Vocabulary quotient*

Vocabulary quotient, a measure used to determine the randomness in the text which is related to the uniqueness in terms of the words used, has been calculated for the documents in the four languages. For computing the entropy, the technique described in [31] is used and further the vocabulary quotient has been derived. To briefly summarise the technique, frequency of each word used in the document is calculated and its probability of occurrence is determined. Using this probability, the Shannon entropy given by

$$S = -\sum p_i \log p_i \tag{5}$$

is computed and the vocabulary quotient is obtained by normalising the entropy value over the maximum possible entropy. The maximum possible entropy was computed as

$$S_{\max} = -\log(1/n) \tag{6}$$

which represents the maximum entropy value that a document can have, given n number of words [31].

Table 3 presents the results obtained for the vocabulary quotient in each document. The column entitled number of words gives the count of total number of

Table 3. Entropy analysis and vocabulary quotient of Bhagavad Gita in four languages.

Language	Number of words	Maximum entropy	Document entropy	Vocabulary quotient
English	18976	4.2782	2.8137	0.6576
Hindi	18680	4.2713	2.7066	0.6336
French	23623	4.3733	2.7341	0.6251
Sanskrit	6458	3.8100	3.3732	0.8853

words in the document. The maximum possible entropy and the entropy of the document computed using the technique prescribed in [31] are recorded in columns 3 and 4, respectively. It can be observed that the maximum document entropy and the highest vocabulary quotient are obtained in the case of the Sanskrit version of the text and may be attributed to the higher number of unique words used in the document. Also, it was seen that there was an extreme usage of fusion words that are formed by the combination of multiple words in Sanskrit. This characteristic was found much more frequently in Sanskrit than in Hindi, English and French which contributes to the vivacity of the document.

2.3 *Word-length distribution*

Word length in a language has been one of the factors in equipping oneself with a new language. Also, it is a marker of the utilisation of the alphabet set of a language and features such as compositions and fusions.

The distributions demonstrated by the word length have been studied by various researchers. The typical pattern has been reported to be Poisson and Ord distribution [32]. A study of word-length distribution on

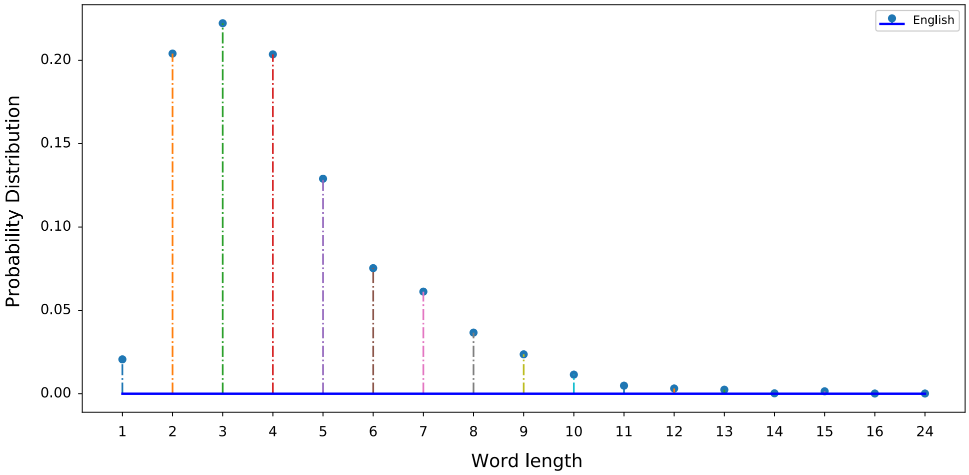

Figure 2. Probability distribution of word lengths in English.

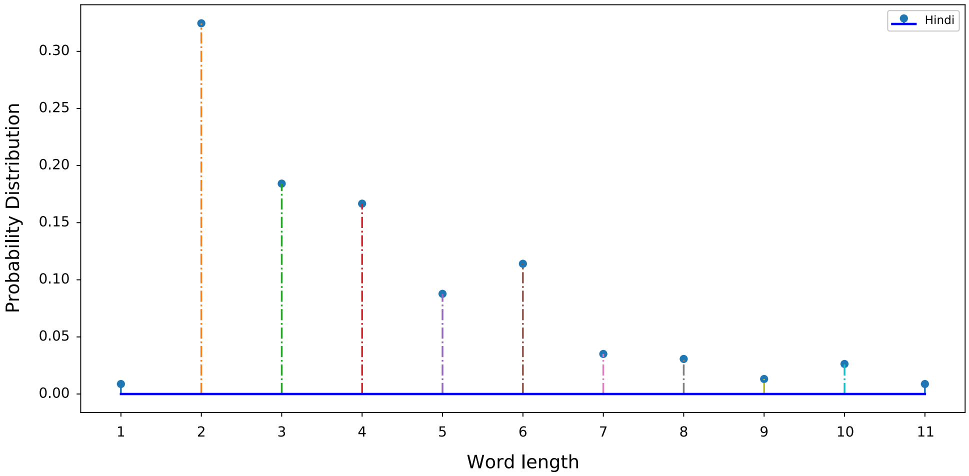

Figure 3. Probability distribution of word lengths in Hindi.

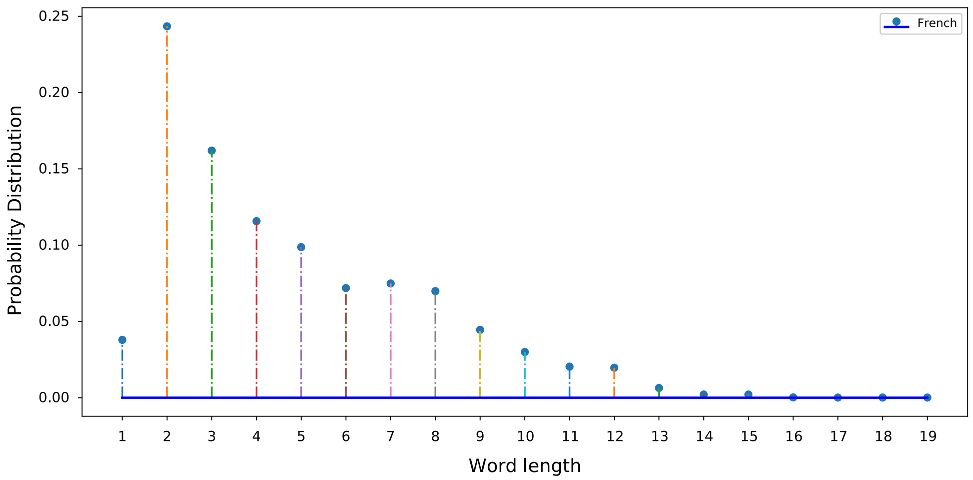

Figure 4. Probability distribution of word lengths in French.

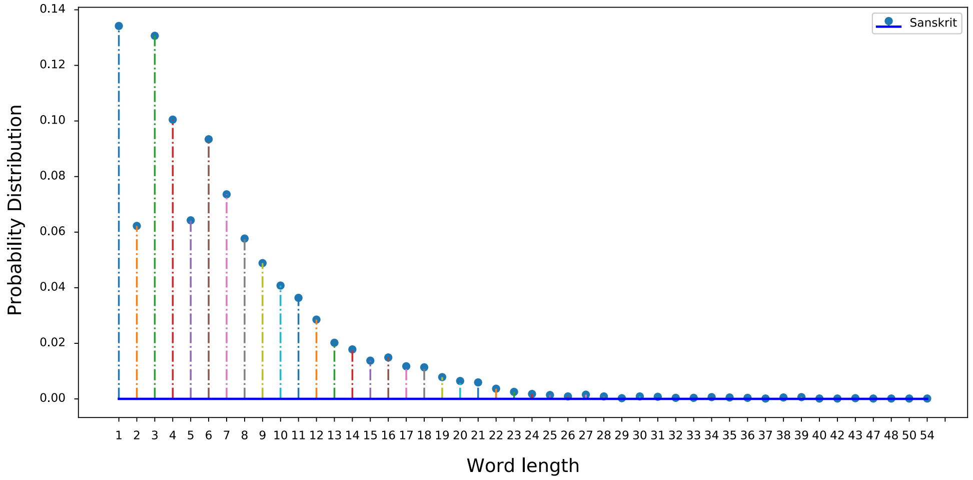

Figure 5. Probability distribution of word lengths in Sanskrit.

Shakespeare's and Bacon's works was carried out [33]. A variant of gamma distribution has been observed in English, Swedish and German language word-length distributions [34]. Word lengths in the Sanskrit version of Bhagavad Gita were found to have typically larger size in comparison to English, Hindi and French versions. The distributions have been depicted in figures 2–5. The highest word length, 54, was reported in the case of Sanskrit while in English it is 24. This highlights the immense presence of fusion in the Sanskrit version of Bhagavad Gita. The most frequent word length in Sanskrit is one due to the maximum occurrence of functional words like 'च' and 'न'.

3. Conclusion

A statistical characterisation of the Bhagavad Gita text has been performed in four languages: English, French, Hindi and Sanskrit. The study has been conducted in four dimensions: building a statistical model based on word frequency distribution, KL divergence between the word frequency distributions of the documents, vocabulary quotient and word-length distributions. The probability distribution of the word frequency was modelled using the Zipf, Zipf–Mandelbrot, Pareto and lognormal distributions. The texts in Hindi, French and English produced a plot that followed the Zipf–Mandelbrot pattern with exponents close to 1. In the case of Sanskrit, the Zipf model fitted better with an exponent of 0.7679. Next, the KL divergence has been computed between the distributions and higher values of KL divergence reflected that there was an immense difference between the probability distribution of Sanskrit and the translated versions. The vocabulary quotients have also been obtained and ranged from 0.6251 to 0.8853 with the highest value for Sanskrit which is an indicator of more number of unique words in the document. Finally, the word-length distributions were plotted and it was noted that the length of the Sanskrit words in Bhagavad Gita was typically quite long compared to other languages.

It was found that word ‘the’ found major occupancy in the document followed by ‘and’ in the case of English. A similar trend was observed in Sanskrit where the word ‘च’ was most frequent which means ‘and’ in English. ‘न’ was second most frequently used in Sanskrit which means ‘not’ in English.

References

[1] C D Manning and H Schütze, *Foundations of statistical natural language processing* (MIT Press, UK, 1999)

[2] R Harald Baayen, *Word frequency distributions* (Springer Science & Business Media, 2001), Vol. 18

[3] G K Zipf, *The psycho-biology of language* (George Routledge & Sons, Ltd., 1936), reprinted in 2002

[4] W Li, *IEEE Trans. Inf. Theory* **38(6)**, 1842 (1992)

[5] B Mandelbrot, *Information theory and psycholinguistics* (BB Wolman and E, USA, 1965)

[6] H Baayen, *Comput. Human.* **26(5–6)**, 347 (1992)

[7] J B Carroll, *Proceedings of the Conference on Language and Language Behavior* edited by E M Zale (Appleton-Century-Crofts, New York, 1968) pp. 213–235

[8] J Narisong Jiang and H Liu, *J. Quant. Linguist.* **21(2)**, 123 (2014)

[9] S Shtrikman, *J. Inf. Sci.* **20(2)**, 142 (1994)

[10] S Miyazima, Y Lee, T Nagamine and H Miyajima, *Phys. A: Stat. Mech. Appl.* **278(1–2)**, 282 (2000)

[11] B D Jayaram and M N Vidya, *J. Quant. Linguist.* **15(4)**, 293 (2008)

[12] C E Shannon, *ACM SIGMOBILE Mob. Comput. Commun. Rev.* **5(1)**, 3 (2001)

[13] W Ebeling and G Nicolis, *Chaos Solitons Fractals* **2(6)**, 635 (1992)

[14] A Stolcke, *Entropy-based pruning of backoff language models*, arXiv:cs/0006025 (2000)

[15] D Genzel and E Charniak, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Philadelphia, 2002) pp. 199–206

[16] W Ebeling and T Pöschel, *Europhys. Lett.* **26(4)**, 241 (1994)

[17] M A Montemurro and D H Zanette, *Adv. Complex Syst.* **5(01)**, 7 (2002)

[18] C C Hoi Hee, *Singapore Manag. Rev.* **29(1)**, 73 (2007)

[19] D V Jeste and I V Vahia, *Psychiatry Interpers. Biol. Process.* **71(3)**, 197 (2008)

[20] W J Johnson, *The Bhagavad Gita* (Oxford University Press, New York, 1994)

[21] S Radakrishnan, *Int. J. Ethics* **21(4)**, 465 (1911)

[22] www.gitasupersite.iitk.ac.in

[23] www.archive.org/stream/LaBhagavadGita-FrenchTranslation

[24] www.gutenberg.org

[25] A Mehri and M Jamaati, *Phys. Lett. A* **381(31)**, 2470 (2017)

[26] M Wiegand, S Nadarajah and Y Si, *Phys. Lett. A* **382**, 621 (2018)

[27] M E J Newman, *Contemp. Phys.* **46(5)**, 323 (2005)

[28] M A Montemurro, *Phys. A: Stat. Mech. Appl.* **300(3–4)**, 567 (2001)

[29] A K Singh *et al*, *IEEE Commun. Lett.* **18(8)**, 1335 (2014)

[30] T M Cover and J A Thomas, *Elements of information theory* (John Wiley & Sons, USA, 2012)

[31] N K Rajput, B Ahuja and M K Riyal, *Digit. Scholarship Human.* **33**, 894 (2018)

[32] G Wimmer, R Köhler, R Grotjahn and G Altmann, *J. Quant. Linguist.* **1(1)**, 98 (1994)

[33] C B Williams, *Biometrika* **62(1)**, 207 (1975)

[34] B Sigurd, M Eeg-Olofsson and J Van Weijer, *Studia Linguist.* **58(1)**, 37 (2004)