Glottometrics 7
2004

RAM-Verlag

ISSN 2625-8226

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitdchrift (2-3 Ausgaben pro Jahr) für die quantitative Erforschung von Sprache und Text.
Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden.
Glottometrics kann aus dem **Internet** heruntergeladen werden (**Open Access**), auf **CD-ROM** (PDF-Format) oder als **Druckversion** bestellt werden.

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals (2-3 times a year).
Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors.
Glottometrics can be downloaded from the **Internet (Open Access)**, obtained on **CD-ROM** (as PDF-file) or in form of **printed copies.**

Herausgeber – Editors

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an

Orders for CD-ROM or printed copies to RAM-Verlag RAM-Verlag@t-online.de

Herunterladen/ Downloading: https://www.ram-verlag.eu/journals-e-journals/glottometrics/

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 7 (2004), Lüdenscheid: RAM-Verlag, 2004. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet (Open Access) unter der Adresse
https://www.ram-verlag.eu/journals-e-journals/glottometrics/ verfügbar.
Bibliographische Deskription nach 7 (2004) **ISSN 2625-8226**

Contents

Articles

History

Review

Glottometrics 7, 2004, 1-24

Some results of quantitative linguistics derived from a structural language model

Wolfgang Hilberg, Darmstadt[1]

Abstract. Recent investigations have shown that a model can be derived from an association matrix, Fig. 1a, which describes a functional language network at the level of words in text. In this model words are gathered in classes, Fig. 1b, in which the number of words increases by a repeated factor of two. It could be shown that when an artificial text is generated with the aid of the model, this text contains maximum entropy. By comparison the conclusion can be drawn that this behaviour is also valid for natural language text. Therefore it is possible to determine some quantitative linguistic values and diagrams immediately and without any difficulty. Recently, this was already shown in the case of Zipf's famous frequency diagrams, Fig. 2 and Fig. 4.

In the paper at hand the following values are derived: text length in dependence on the number of different words, equ. (3), ramification curve, equ. (30), the number of different word pairs in dependence on text length, equ. (34), the average length of sentences measured in words, equ. (35), and the entropy, equ. (38). All these calculations are elementary, based on statistical average values of model parameters and of text generation procedures, and they lead therefore only to first approximations for curves and characteristic values. However more precise results are attainable by simulating the process of text generation in the model step by step. Then, of course, normal distributions can be found, Fig. 7, which lead to Zipf-diagrams with smoothed stairs, Fig. 8a,b, instead of ideal rectangular steps as in Fig. 2. Finally the question of a possible connection between language and prime numbers is discussed.

Keywords: text length, sentence length, word pairs, entropy, ramification curve

1. Introduction

On the basis of measurements of the language network and its representation in a connection matrix with logarithmic axes, called association matrix, see Fig.1a, the underlying functional structure could be obtained and approximated by an idealized model (Hilberg 2000, 2002). It proved to be the common frame for all languages and represents a firm base for an implicit universal grammar (a short discussion of that can be found in Hilberg (2000)). The essential characteristic is the subdivision of the multitude of rank ordered words, numbered from 1 to N

$$n = 1,2,3,...,N \tag{1a}$$

into classes, numbered from 0 to K

$$i = 0,1,2,...,K\,, \tag{1b}$$

[1] Address correspondence to: Wolfgang Hilberg, TU Darmstadt, FG Digitaltechnik. Merckstr. 25, D-64283 Darmstadt, Germany. E-mail: hil@dtro.tu-darmstadt.de

their size increasing by a power of two, see Fig. 1b. That is, the size of a class i is given by

$$2^{i} . \tag{1c}$$

The number of all classes is (K+1) and the number of all different words is N. Beside the possibility of a simple description of the network structure, another advantage is that relations and quantitative results can be derived theoretically, which otherwise could only be obtained by circumstantial long measurements. Even though the model is idealized its use nevertheless results in curves and characteristic values, which are in rather good accordance with measurements. For example we obtain a more precise course of Zipf's frequency curve which is not smoothed, and we are able to calculate the ramification curve, the average length of sentences, the multiple usage of words, the number of word pairs in dependence of the stock of words, and the entropy. Finally we scrutinize the question of possible data compression and of possible connections with prime numbers. In the following sections an overview shall be given over these first quantitative values, which are derived from the structural language model.

2. Zipf's law of frequency

The general type of the language network — not a special existing natural network — was determined with the aid of a random experiment. Starting from an arbitrarily chosen word, in search of the next word one of the (K+1) classes of the structural model was activated by chance and within this class one word was also activated by chance. Starting again from this word the procedure was repeated many times. In doing that, all classes (and all words in it) were called up with equal frequency and an artificial text was generated. The last class of the

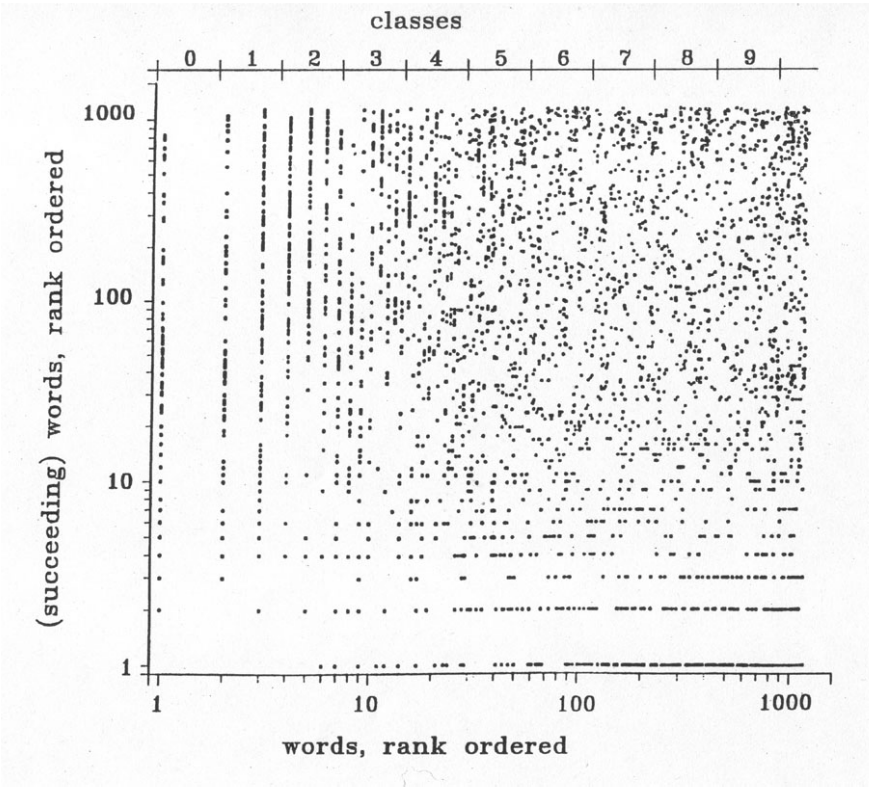


Fig. 1a : The association (connection) matrix (Hilberg 2002). Grouping of nodes in classes is indicated at the upper side.

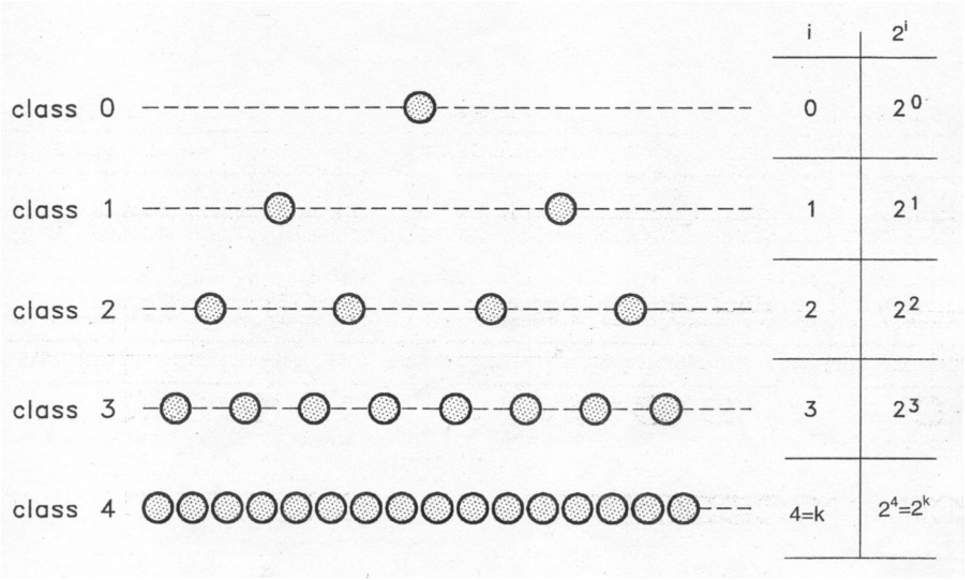


Fig. 1b : Illustration of a basic model where nodes are divided into word classes (Hilberg 2002).

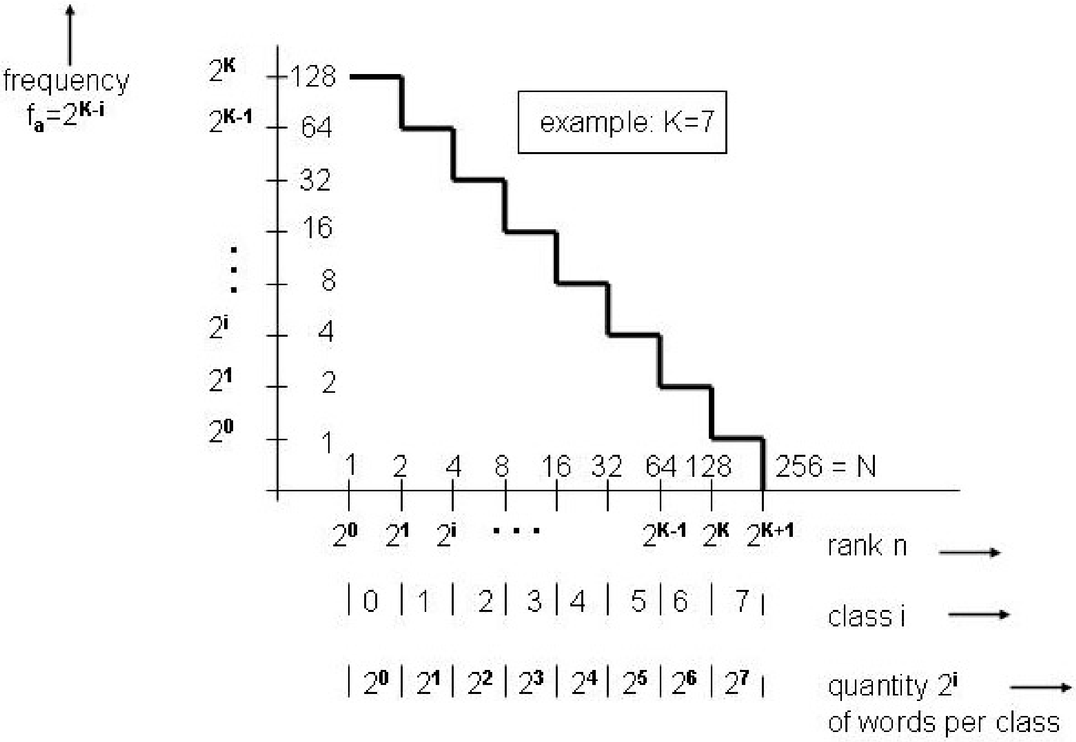


Fig. 2: Principle of the frequency-rank diagram obtained from the model. It is a first structural approximation for the measurement curve in Fig. 4.

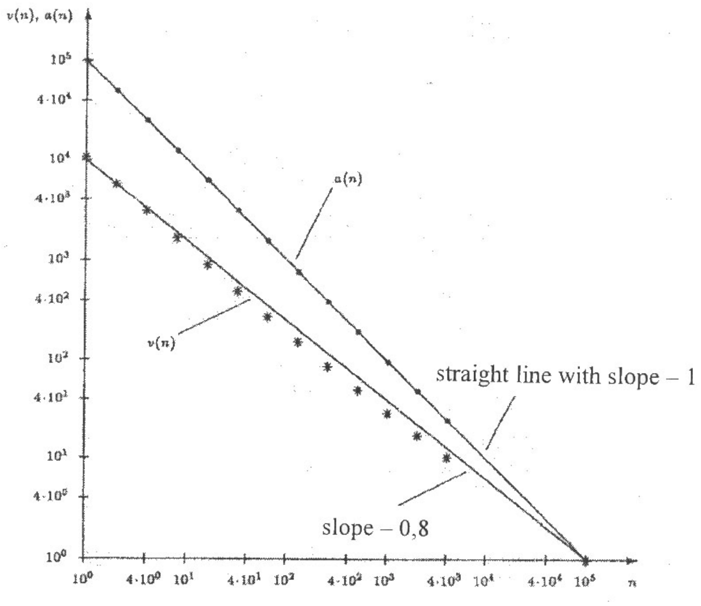


Fig. 3: Meyer's (1989) diagram. Upper curve *a*(*n*): Usual presentation of Zipf's curve as a declining straight line (power law). Lower curve *v*(*n*): Ramification curve (in German: V̲erzweigung) for rank ordered words. (*a*(*n*) and *v*(*n*) in the Fig. are the accumulated frequency f_a and ramification v_j, see also appendix A).

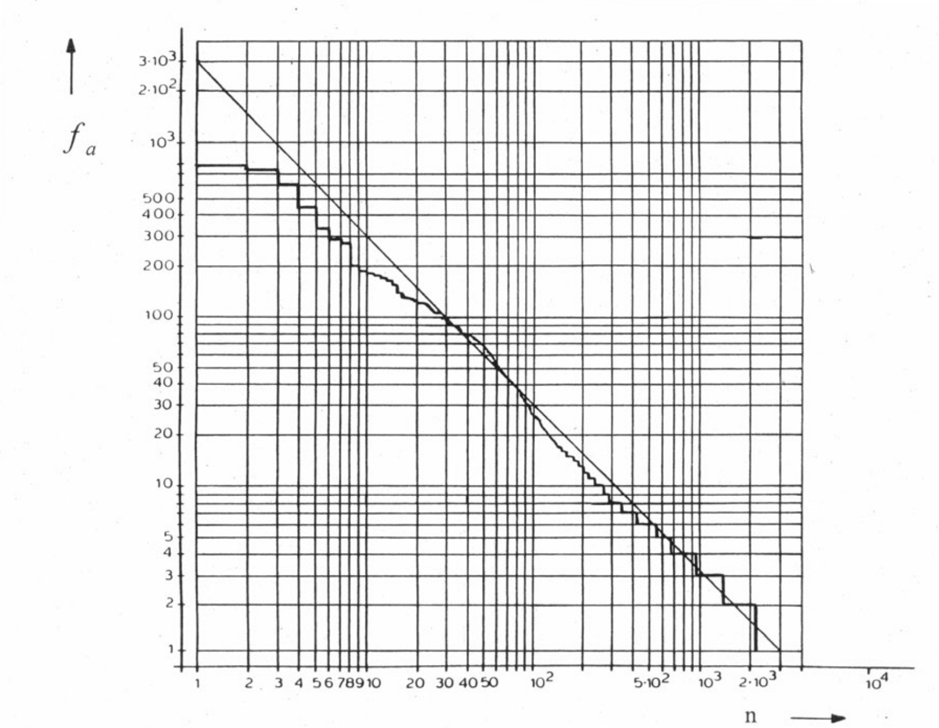


Fig. 4: Comparison of a measured curve, showing (not very regular) stairs, with the theoretical straight line of Zipf's curve (Hilberg 2000).

model with 2^K words, when its words were used but once in the average, was chosen for determining the number of repetitions. (This model was called the homogeneous model. We will take a closer look on in Chapter 10). Thus every class is activated for 2^K times. In an arbitrarily chosen class *i* with word capacity 2^i therefore an accumulated or absolute word frequency f_a results (the problem of using appropriate symbols is discussed in the appendix A)

$$f_a = 2^K / 2^i = 2^{K-i}, \qquad i = 0,1,2,...,K. \tag{1d}$$

Plotted in a diagram, a staircase of frequency results, which corresponds to Zipf's law, Fig. 2. The maximum number of different word-nodes in the model of Fig. 1b, — it is identical with the maximum number of rank in Fig. 2 — is given by

$$N = 1+2+4+8+....2^K = \sum_{i=0}^{K} 2^i = 2^{K+1} - 1 \approx 2^{K+1}. \tag{2a}$$

The result in Fig. 2 is obviously somewhat different to the well known "power law" course of Zipf's law, see the upper curve in Fig. 3, taken from Meyer (1989), where the absolute frequency $a(n)$ of words is presented in dependence on rank *n*. With the denotation f_a which is used here, instead of $a(n)$ in Fig. 3, the power-law is written

$$f_{a,M} = \frac{N}{n}, \qquad n = 1,2,3,...,N. \tag{2b}$$

For better distinction the index *M* (for Meyer 1989) is added to the symbol f_a. Looking at the measured curves, e.g. in Fig. 4, a comparison with equ. (1d) and equ. (2b) shows that in reality Zipf's curve approximates a staircase rather than a straight line even if the former is not exactly a regular staircase. It should also be noted that in reality the value at the beginning is distinctly lower than *N*, which is in agreement with equ. (1d) and equ. (2a). (For power law curves with different slopes which exist in other fields of science see Appendix B).

3. Text length

The sum of all word events in Zipf's frequency measurement equals the text length *L*. In Zipf's diagram this is the result of the sum of the absolute frequencies f_a found in the columns above all word ranks. When we sum up class after class in Fig. 2 , we begin for the sake of simplicity with the highest class at the right edge. In the following equation at first the average frequency of occurrence of a word in this class (here 2^0) is noted and then the number of words in the class (here 2^K). It follows

$$L = 2^o 2^K + 2^1 2^{K-1} + ...+ 2^K 2^0 \tag{3}$$

$$= \sum_{i=0}^{K} 2^i 2^{K-i} = \sum_{i=0}^{K} 2^K = (K+1) \cdot 2^K .$$

For comparison let us look at another value, derived some years ago by J. Meyer (1989), which was based on the formula of Zipf's power-law. Using the notations here and again distinguishing with an index *M* (for Meyer), we find

$$L_M = N \ln N = 2^{K+1} \cdot \ln 2 \cdot \mathrm{ld}\, 2^{K+1} = (K+1)\, 2^K \cdot (2 \ln 2) \qquad (4)$$

Obviously the difference between the values in equ. (3) and equ. (4) is not very large, especially when the assumptions for the idealizations are considered (see also measured values in Fig. 5c).

4. Frequencies of classes and the communication constant

In the modelling process all classes are activated with equal frequency. With the calculated value in equ. (3) the number of activations A_{CL} of any class equals, as was already postulated in words before in chapter 2 :

$$A_{CL} = \frac{L}{K+1} = 2^K \qquad \text{text words per class.} \qquad (5)$$

We obtain the same value by multiplying the number of words per class with the specific frequency number of this class in equ. (1d)

$$K_{COM} = 2^i 2^{K-i} = 2^K . \qquad (6)$$

This constant was named "communication constant" or "communication performance" respectively. From the last two equations it can be concluded

$$A_{CL} = K_{COM} . \qquad (7)$$

That means the simple fact that the frequency of calling up a class (2^K) equals the product of the number of words contained in the class and of its specific word frequency.

5.1 Multiple usage of words

The extent of multiple usage of words in text can be characterized by the relation V_{MULT} , which is given when the text length *L* is divided by the number *N* of different words:

$$V_{MULT} = \frac{L}{N} = \frac{(K+1)2^K}{2^{K+1}} = \frac{(K+1)}{2} . \qquad (8)$$

In the example of the LIMAS-corpus with *K* = 15 classes the words of the word stock (vocabulary) are repeated 8 times in the average. If we consider the simulations of *M.* Steinmann (1996), and assume that the first two classes are not occupied, the words are repeated only 7 times.

5.2 The number of two-word successions in text

The calculation of the number of direct successions of two-word expressions is obviously trivial, provided that repeatedly occurring sequences are also counted repeatedly. This calculation is trivial because for every word called upon in text a successor exists (except at the end). Therefore the number of two-word combinations N_{BIN} is just as in equ. (3) minus one:

$$N_{BIN} = (K+1)\cdot 2^K \text{-}1 \approx (K+1)\cdot 2^K . \qquad (9)$$

In relation to the number *N* of different words and with equ. (2a) we get

$$\frac{N_{BIN}}{N} = \frac{K+1}{2} = \frac{ldN}{2} . \qquad (10)$$

The comparison with equ. (8) still shows (which is an additional proof of equ. 3)

$$N_{BIN} = L = (1/2)\cdot N\cdot \text{ld } N. \qquad (11)$$

It is apparent that a repeatedly activated special word will not always be part of one special two-word combination. Therefore it is interesting that the consideration of manifold word sequences does not have any influence on the calculation above. Therefore we only take

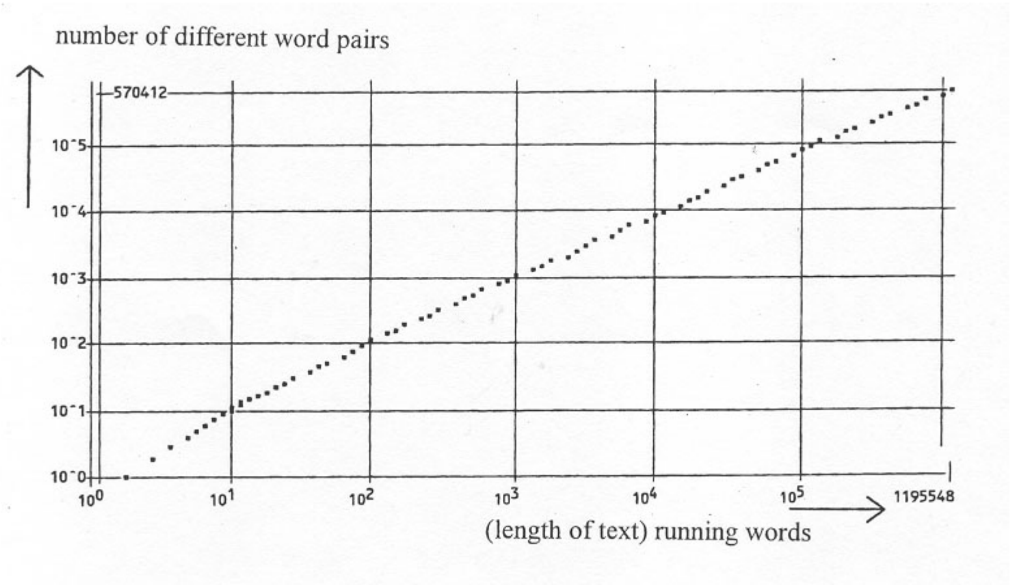


Fig. 5a : Number N_{PAIR} of different word pairs in dependence of the number *L* of running words in text (Burschel 1999).

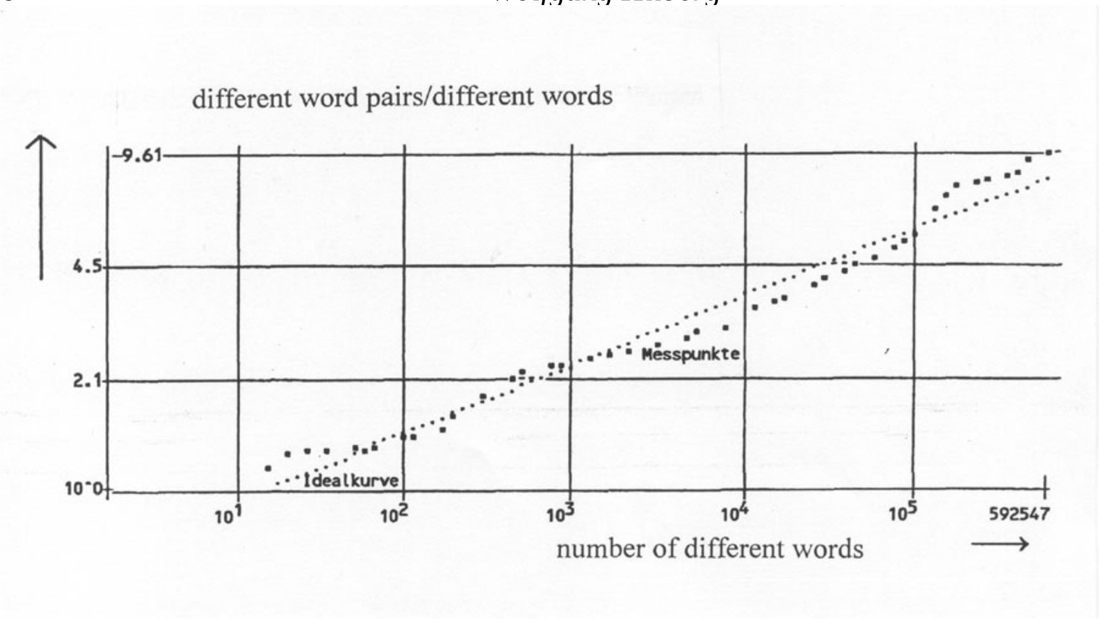


Fig. 5b : Relation N_{PAIR}/N of different word pairs in dependence of the number *N* of different words (Burschel 1999).

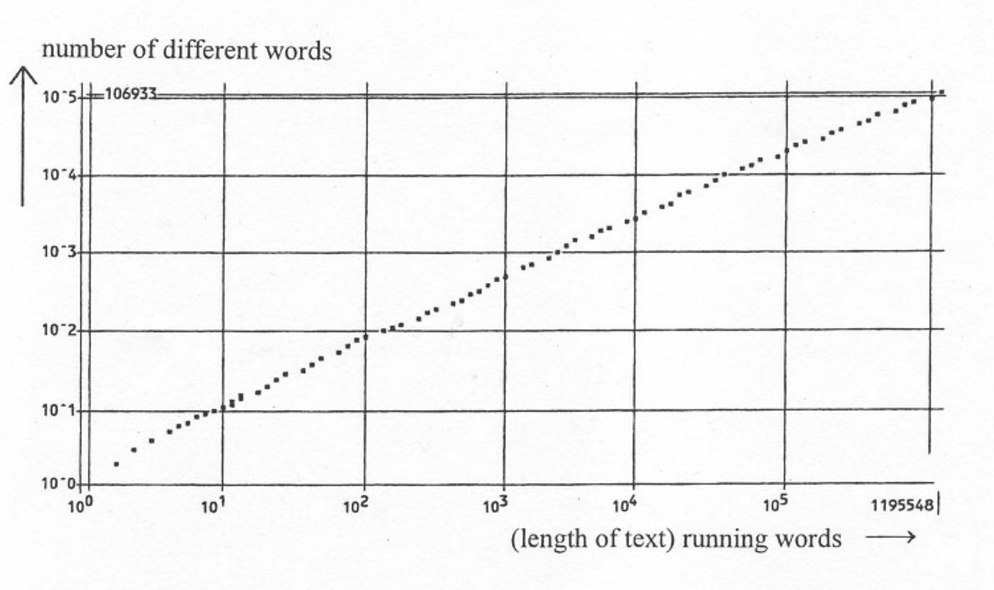


Fig. 5c: Number *N* of different words in dependence of the number *L* of running words in text (Burschel 1999).

note of the fact that the number of two-word combinations increases more than the number of different words, as it was also the case with the number of repeatedly used words.

The number of two-word successions however, when repetitions are not counted, cannot be found in a similarly simple way. The question for the number of such different "word pairs" in text shall be answered immediately in the next chapter.

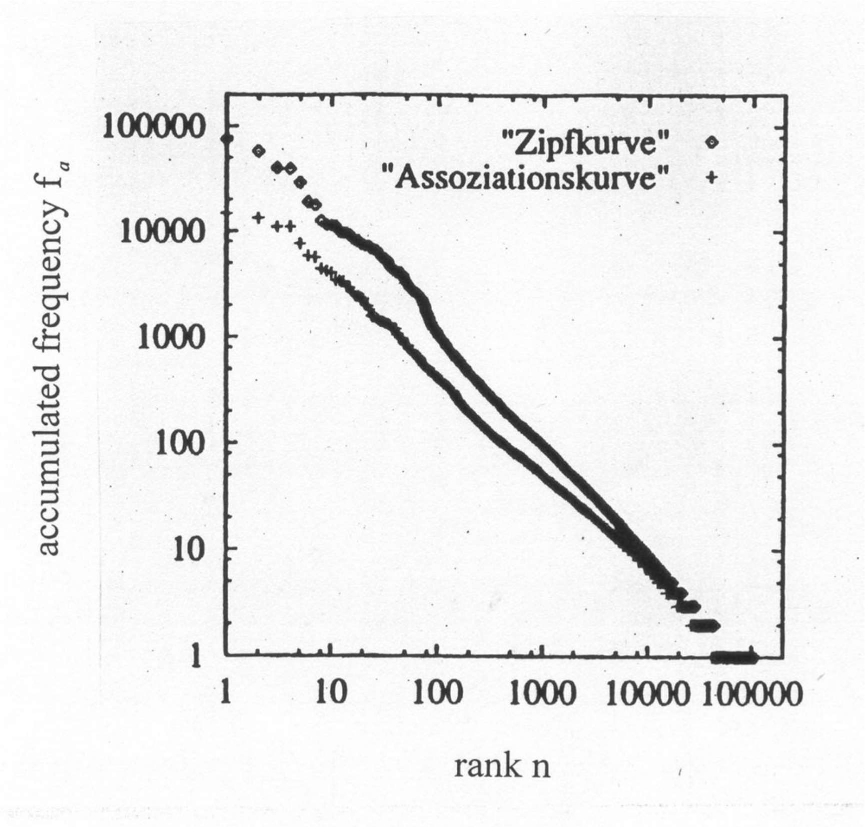


Fig. 6 : Zipf's curve (upper curve) and ramification curve (lower curve) in contrast to Fig.3, now simulated from the structural language model (Steinmann 1996).

6. The ramification curve (number of different word pairs)

By forming the sum of dots in the measured association matrix for a column, we get the ramification number for the corresponding word node in the network (that is the number of connections to succeeding nodes). After forming the sum for every column, the arising curve is called the association curve or, more illustrative, the ramification curve. With regard to Fig. 3 and Fig. 6 this curve lies below Zipf's curve (Meyer 1989; Steinmann 1996; Bassenge 2002). That can be understood because in Zipf's curve the corresponding dots in the matrix columns are counted as often as they appear in text whereas the dots for the ramification curve are structure parameters and are counted only once.

When the question arises of how to derive the ramification curve from the model, the first rough idea may be the following one: The association matrix in Fig.1a shows at first sight a nearly constant dot density in the main part of the area. In a first approximation we could assume a perfect constant density everywhere in the matrix and especially in all the columns and rows. Then in all the stripes, which define classes (Hilberg 2002), the same number of dots would be present. This would imply that all classes contain the same number of connections (remember that the dots stand for connections in the network). The number of ramifications per word then equals the number of dots in the column above its rank number and is given approximately by the value in equ.(1d)

$$2^{K-i} \qquad i = 0,1,2,........K \tag{12}$$

When this value for a word is multiplied with the number of words per class (2^i) we eventually get (compare it with equ. 6)

$$2^{K-i} \cdot 2^i = 2^K . \qquad (13)$$

The idea of using the staircase-function 2^{K-i} as the ramification curve would be possible only for the assumption above that we have a constant dot density everywhere in the matrix. But in fact that is not true. We can see in Fig.1a that in columns and rows near the axes the dots are distributed more sparse than in the upper right corner. Therefore the ramification has to be calculated in a better, second approximation.

6.1 The beginning of the curve

The improved staircase character of the ramification curve can be approached in the following way: equ. (5) shows that when processing a text of length L every class is activated exactly 2^K times. That is, every class receives 2^K times a signal from a preceding word and subsequently puts out a word of its own. We will restrict ourselves to the consideration that the words of any class will search 2^K times a way to a succeeding word in any of the (K+1) existing classes. As the classes are equally probable, the 2^K activations of each class are distributed in equal parts over all the classes. Calling these parts "packets" we get their word content

$$N_{PACK} = 2^K/(K+1). \qquad (14)$$

At first let us look at class 0 somewhat more precisely. The ramification number of its only word is surely the greatest one of all words. But its $2^K/(K+1)$ connections to words of other classes will not always find different words, especially in neighbouring classes with few words. Only in more distant classes ─ let us assume from class κ on ─ there are again more words included in a class than the packet number $2^K/(K+1)$ states. Denoting the number of connections between classes with the symbol v, the connections starting from class 0 can be split up as follows

$$v_0 = v_1 + v_2 \ , \qquad (15)$$

Here v_1 denotes the connections from class $i = 0$ until $i = \kappa$:

$$v_1 = 1 + 2 + 4 + ... + 2^\kappa \qquad (16)$$

and v_2 the connections from class $i = \kappa+1$ until $i = K$:

$$v_2 = \underset{i=\kappa+1}{2^K/(K+1)} + ... + \underset{i=K}{2^K/(K+1)}. \qquad (17)$$

In order to calculate κ, the following condition is valid for the transition place

$$2^\kappa = 2^K/(K+1). \qquad (18)$$

From this we obtain

$$\kappa = \mathrm{ld}(2^{K}/(K+1)) = K - \mathrm{ld}(K+1) \qquad (19)$$

The sum in equ. (16) is

$$v_1 = 2^{\kappa+1} - 1\ . \qquad (20)$$

In equ. (17) the remaining terms are

$$v_2 = (K - \kappa)\cdot 2^{K}/(K+1) \qquad (21)$$

Considering equ. (18) until equ. (21) we have

$$\begin{aligned} v_0 = v_1 + v_2 &= 2^{\kappa+1} - 1 + \frac{K-\kappa}{K+1}\cdot 2^{K} \qquad (22)\\ &= 2^{\kappa+1} - 1 + (K - \kappa)\cdot 2^{\kappa} \\ &= 2^{\kappa}(2 - \frac{1}{2^{\kappa}} + (K - \kappa)). \end{aligned}$$

Finally we find with equ. (18) and equ. (19) the form

$$v_0 = \frac{2^K}{K+1}\,(2 - \frac{K+1}{2^K} + \mathrm{ld}(K{+}1)) \approx \frac{2^K}{K+1}[2{+}\mathrm{ld}(K{+}1)] \qquad (23)$$

As could be supposed, the result is somewhat smaller than the frequency value 2^{K} in equ. (1d) or equ. (12). M. Steinmann (1996) confirmed and extended this result by the possibility of considering also some empty classes at the beginning.

The result above, put in words: the ramification curve starts with a value which is smaller than the frequency value of Zipf's curve. (In the example with K = 15 for the LIMAS-corpus we have a factor of about 3/8). With increasing rank of words the ramification will then decline like a staircase, as the model predicts, right from the calculated value at the beginning until the value 2^{0} in the last class n.

6.2 The whole curve

In order to calculate the height of all stairs we only have to extend the calculation above slightly. At first we assume that the symbol w stands for a word in the fixed class *j* for which we want to calculate the ramification number. Then the first part of the calculation above remains the same, where the number of words in a target class is smaller than the number of branches proceeding from the word w to this class. This may happen again until class κ. Then

$$v_1 = 1 + 2 + 4 + \ldots\ldots + 2^{\kappa} = 2^{\kappa+1} - 1\ , \qquad \kappa < K \qquad (24)$$

For the case that all the branches starting from word w will arrive at all words of class κ the following condition must be met

$$2^{\kappa} = \frac{2^K}{(K+1)} \cdot \frac{1}{2^j} = \frac{2^{K-j}}{(K+1)} \ . \tag{25}$$

Solving for κ yields

$$\kappa = K - j - \text{ld}\,(K + 1), \tag{26}$$

or for j

$$j = K - \kappa - \text{ld}\,(K + 1). \tag{27}$$

Also in the subsequent (K-κ) classes the value at the right side of equ. (25) is not surmounted. Combining the results we get

$$v_2 = (K - \kappa) \cdot \frac{2^{K-j}}{K+1} . \tag{28}$$

The sum is

$$v_j = v_1 + v_2 = (2^{\kappa+1} - 1\,) + (K - \kappa) \cdot \frac{2^{K-j}}{K+1} \cdot \tag{29}$$

Substitution of κ from equ. (25) and equ. (26) yields

$$v_j = (2 \cdot \frac{2^{K-j}}{K+1} -1\,) + (K - K + j + \text{ld}\,(K + 1)) \cdot \frac{2^{K-j}}{K+1} \tag{30}$$

$$= \frac{2^{K-j}}{K+1}\,(2 - \frac{K+1}{2^{K-j}} + j + \text{ld}\,(K + 1)).$$

As a test we may set $j = 0$. Then we find the result in equ. (23). If we set $j = K$, we find

$$v_K = \frac{1}{K+1}(\,2 - K - 1 + K + \text{ld}(K + 1)) = \frac{1}{K+1}\,(\,1 + \text{ld}\,(K + 1)). \tag{31}$$

At first sight this result is somewhat confusing. Using the value K = 15 for comparison, the result is v_K = 5/16 ≈ 1/3 instead of the correct value 1 . Considering all the strong idealizations however and the calculation depending on average values the result may still be seen as satisfying.

7. The number of different word pairs

The total number of different word pairs N_{PAIR} in a text of given length can be obtained from the ramification numbers for single words simply by addition:

$$N_{PAIR} = \sum_{j=0}^{K} \mathrm{v}_j \cdot 2^j \, . \tag{32}$$

Referring to equ. (30) gives

$$\begin{aligned} N_{PAIR} &= \sum_{j=0}^{K} \frac{2^K}{K+1}\left(2 - \frac{K+1}{2^{K-j}} + j + \mathrm{ld}\,(K+1)\right) \\ &= \frac{2^K}{K+1}((K+1)(2 + \mathrm{ld}\,(K+1)) + \sum_{j=0}^{K} (-2^j + \frac{2^K}{K+1} \cdot j) \\ &= 2^K\,(2 + \mathrm{ld}\,(K+1) + (-2^{K+1} + 1 + \frac{2^K}{K+1} \cdot \frac{(K+1)K}{2}) = 2^K\,\mathrm{ld}(K+1) + 1 + K2^{K-1} . \end{aligned} \tag{33}$$

The result is

$$N_{PAIR} = \frac{N}{2}\,\mathrm{ld}(K+1) + 1 + K\frac{N}{4} \approx \mathrm{N}(\frac{ld(K+1)}{2} + \frac{K}{4}) = N(\frac{ldN-1}{4} + \frac{ldldN}{2}). \tag{34}$$

From measurements of Meyer (1989) and Burschel (1999) the value $N_{PAIR}/N \approx 5$ is known for the LIMAS-corpus (see also Figs 5a,b), while the model here yields approximately the value 5.75 under the same assumptions. However here the empty classes at the beginning of the model were not considered. Again the strongly idealized assumptions may also be responsible for the difference.

8. Length of sentences

The average length of sentences in books or newspapers has been measured very often. In German language the values can be found between 11 and 18 words depending on which circle of readers the text shall be addressed to. As had been shown already in Hilberg (2002) the result can be predicted very easily by using the idealized structure model: When a) the first word class is assumed to include not a word but a punctuation mark and when b) according to the model the (K+1) word classes are activated with equal frequency for the generation of text, then on an average a punctuation mark appears after K words. Referring to equ. (2a) that gives an average length of sentences of

$$L_{SENT} = \mathrm{ld}\,(N/2\,) = K \text{ words.} \tag{35}$$

For German text in the LIMAS-corpus with about K = 15 the resulting average length of sentences in the order of 15 words matches astonishingly well with the values mentioned above.

9. Entropy, data compression

The entropy of a word in a text, that is the average information content of a word, can also be calculated very easily with the aid of the model. Starting from an arbitrarily chosen word, we only have to consider the probability for the subsequent random occurrence of a word in a random class of the model and eventually we have to sum up the terms for all words in the class and for all classes. Now, the probability for choosing a class i has the constant value $1/(K+1)$ and the probability for a word in this class has the value $1/2^i$. Together

$$p_1 = 1/[(K+1)2^i]. \tag{36}$$

The entropy according to C.E. Shannon is

$$H = -\sum_i \sum_w p_1 \operatorname{ld} p_1 . \tag{37}$$

We may sum up at first over $w = 2^i$ words of class i and then over $(K + 1)$ classes

$$\begin{aligned} H &= -\sum_{i=0}^{K} 2^i \frac{1}{(K+1)2^i} \operatorname{ld} \frac{1}{(K+1)2^i} \\ &= \frac{1}{(K+1)} \Big(\sum_{i=0}^{K} \operatorname{ld}(K+1) + \operatorname{ld} 2^i\Big) = \frac{1}{K+1} \sum_{i=0}^{K} (\operatorname{ld}(K+1) + i) \\ &= \operatorname{ld}(K+1) + K/2 . \end{aligned} \tag{38}$$

Here the arithmetic series

$$1d + 2d + 3d + \ldots + nd = \frac{n+1}{2} \cdot d \cdot n \tag{39}$$

was used. It is worth mentioning that the entropy per letter in equ. (38) is not a constant value, which is often assumed when only measured values are observed. In theory, however, the entropy increases along with the volume of vocabulary — even though this increase is very slow, see equ. (2a).

This result was obtained already by M. Steinmann (1996). It corresponds well with a value which was calculated by J. Meyer (1989) from the "power-law"-Zipf curve

$$H_M = \frac{ldN}{2} + \operatorname{ld}(\ln N) = \frac{K+1}{2} + \operatorname{ld}((\mathrm{K}+1) \cdot \ln 2). \tag{40}$$

The agreement is obvious, if the last term is neglected for large K. An earlier approximation also existed, the so-called "root law" (Hilberg 1990), which can be written using present symbols

$$H_0 = \operatorname{ld} \sqrt{N} \tag{41}$$

This can be reduced by the introduction of the model parameters

$$H_0 = \operatorname{ld} 2^{(K+1)/2} = (K + 1)/2. \tag{42}$$

Obviously this value is also very close to both values above and it is the same number as in equ. (8) and equ. (10). M. Steinmann (1996), it was mentioned above, considered in the model especially that there may be some empty classes in the beginning. If there are *e* such empty classes, equ. (38) can be supplemented in the following way

$$H_{ST} = \mathrm{ld}\ (K + 1 - \mathrm{e}) + \frac{K - e}{2}. \qquad (43)$$

For the example of $K = 15$ we find from the main equ. (38)

$$H = \mathrm{ld}\ 16 + 7.5 = 11.5 \text{ Bits per word.}$$

For an average word length of about 6 letters in German we get just 1.92 Bits per letter, a number, whose order of magnitude is well known from many text measurements and from some superior data compression procedures. Even in the first measurements which were published about 50 years ago by C.E. Shannon (1951), the well known creator of information theory, a value of about 1.5 Bits per letter was found for text pieces of about 6 letters (average word size). This is not too far away from the 1.92 Bits per letter which were calculated above. Furthermore it may be supposed that Shannon used a text containing a smaller vocabulary which gives a somewhat smaller value. (It is a rare example that Shannon actually made a measurement, because he could not derive the result theoretically).

The most interesting property of the word-entropy is the following one: All the probabilities for the word classes were chosen to be equal, as was also the case for the probabilities for the words inside the classes. Therefore equ. (38) states the maximum value of entropy. That means, on the base of the structure model described it is not possible to generate a text of a given length with higher average information content. Or in other words, text generated by the structural model and its procedures yields maximum information on an average! That is a network property and also the best result we could hope for. Nevertheless we know that in a real natural text there are restricting conditions hidden in a more distant "context". Therefore the pure random law used here is supplemented and, by considering context, further redundancy can be eliminated, leading to lower entropy values. However this cannot be included in the basic network model discussed here, where only direct successions of words are considered. It could be shown that context information can be grasped and stored in additional networks of the same kind, arranged in hierarchical successions of networks with levels of increasing abstraction, see Hilberg (2000; 2003).

10. Normal distribution

Of course it is not very realistic to assume that a random generation of a network will end up in a distribution of connections where all words of a class have the same ramification number as in the approximations above. Theory and simulation rather have shown that in fact by an accurate calculation in every class a normal so-called Gaussian distribution of ramification numbers arises around an average central value, which is identical with the ramification number of the simplified model (called the "homogeneous" model), which was discussed above. M. Steinmann illustrated this in an example, suggested by the author, where the result of a voluminous stochastic simulation was presented in a histogram of ramification numbers which developed in a higher word class, see Fig. 7. The average central value is here approximately $2^7 = 128$, where the frequencies of greater ramification numbers decline to the right and the frequencies of smaller ramification numbers decline to the left (note that the next greater average value is $2^8 = 256$ and the next smaller value $2^6 = 64$, both far away in this diagram).

The precise statistical treatment has of course its influence, which necessitates corrections of the results of the homogeneous model above. These corrections are very obvious e.g. for the shape of the ramification curve or (when proceeding from ramification to frequency) for the shape of Zipf's frequency curve. Here the words are arranged according to their frequency rank, as is well known. Hence we have the rank order by frequency or ramification also within the classes. Considering the result in Fig. 7 the frequency f_a in the middle of a step will be given approximately as in Fig. 8a by the mean value of a normal distribution. Next to this region the corners and edges will be rounded. This is in accordance with well known measured curves in literature and also with the simulated curves for the model in Fig. 8b (Steinmann 1996).

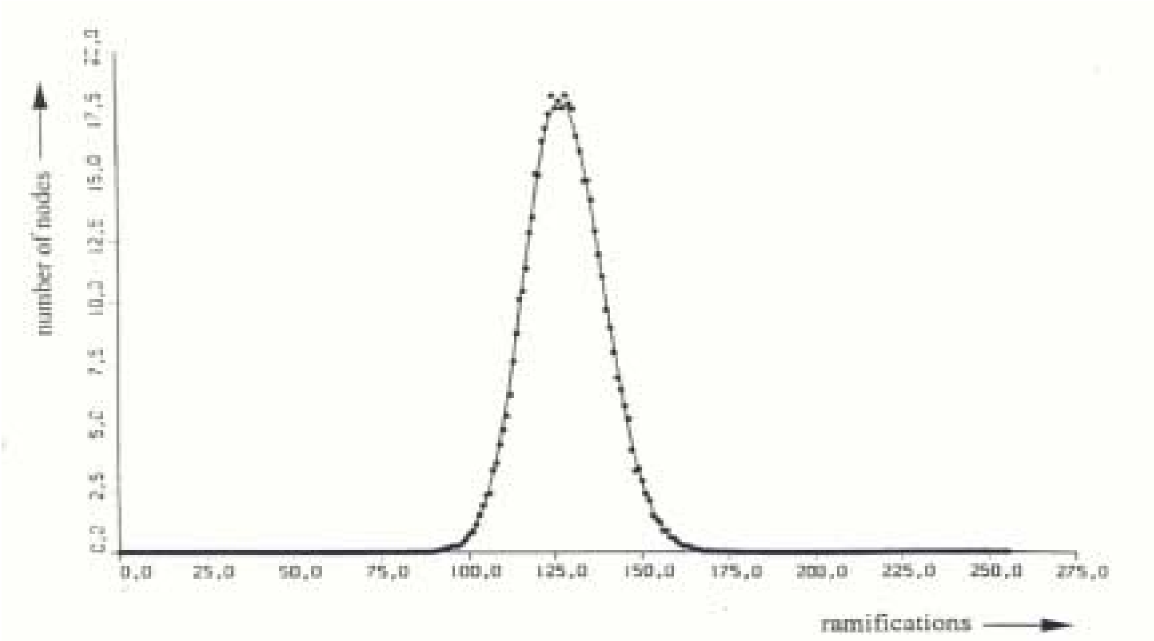


Fig. 7 : The number of word nodes in dependence of the number of ramifications, simulated for a higher class of the structural language model (Hilberg 2000; Steinmann 1996)

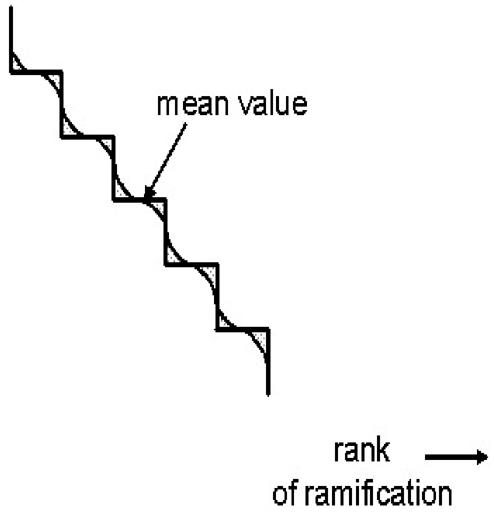


Fig. 8 a : Principle of smoothing the steps if normal distributions are considered.

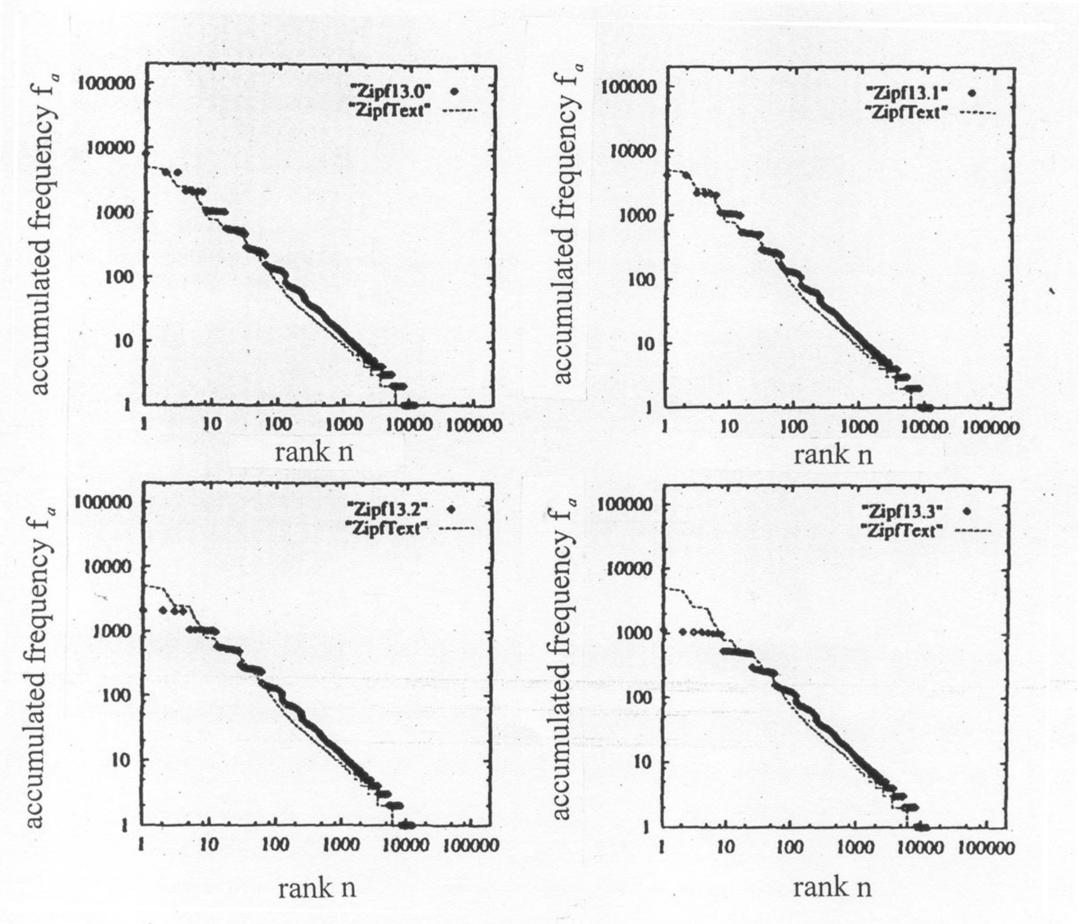


Fig. 8 b : Several examples of Zipf's curves which were simulated with normal distributions in the model classes (Steinmann 1996).

The series of theoretical approximations with growing accuracy for the real curve in Fig. 4 is now obvious: We see the first approximation (power-law) in Fig. 3 as a straight line, the second approximation in Fig. 2 as an ideal staircase and the third approximation (normal distributions) in Fig. 8b as smoothed staircases. The variation in the width of stairs — it corresponds with different class volumes — which can be seen in Fig. 4, is not yet included.

11. Continuing evolution of classes and the diameters

In the basic model of Fig. 1b only constant numbers of word nodes N and classes K were considered. The relation between N and K in equ. (2a) was derived as a consequence of the model structure. However, in natural language texts of increasing length, the number of different words is also increasing. In the model we may consider this behaviour by a way of growth where new classes arise step by step. For every new class the following simple relations are already included in the descriptions and equations above.

In a first step equ. (2a) for the constants N and K may be written as a function

$$N(K) \approx 2^{K+1}. \qquad (44a)$$

When a new class is added, we find

$$N(K+1) \approx 2^{(K+1)+1} = 2 \cdot N(K). \tag{44b}$$

In words: When a new class has been completed the total number of all the different word nodes of the system becomes twice as much as before. One could think that in this case all previous nodes would develop just one new connection to a new word in the additional class (comparable with the construction of a hypercube network). But this conclusion is not consistent with the model structure.

Between the nodes the total number of connections also increases when the network system grows. In a first approximation this increase obeys also a factor of 2 after a step in growth is completed.

The second step: In the discussion above it could be shown that with a fixed number K any class contributes the same packet size of word activities to the whole communication. With an increased number $(K+1)$ the packet size is

$$2^{(K+1)}/((K+1)+1). \tag{44c}$$

Summarizing these packets yields a number of $2^{(K+1)}$ different connections to the words in the last class. Because this is also the number of words in the last class we have on average just one connection per word (note that we already presumed this property at the beginning). At the same time it is obvious that in the evolution of the system previous classes stay their size while their words always get a larger part of connections to other words than new words do (for an example the only word in class 0 should be compared with an arbitrary word in class $(K+1)$.

That has consequences for the performance of communication. When in the network relatively few word nodes with very large ramification numbers exist and amongst them relatively many words are situated with very small ramification numbers, this will result in a high connectivity which can be characterized by a small network diameter. As is well known, a diameter D is defined as the maximum number of steps which are necessary to proceed from a given node to an arbitrarily chosen other node (we may think of words in text which are in a distance of several steps apart). Calculations showed (Hilberg 1991) that such a diameter with $D \approx \text{ld}\, N - 1$ is even somewhat smaller than the diameter of the well known hypercube with $D_H = \text{ld}\, N$. But in some cases the mean diameter D_M is better suited for assessments. Its value $\mathrm{D}_M < (1/2)\text{ld}\, N$ is also somewhat smaller than the value $D_{M,H} = (1/2)\text{ld}\, N$ for hypercubes.

However, these technical values D and D_M are no longer interesting when the actual short steps in a running text from an arbitrarily chosen word only to the next succeeding word are considered. This will be with certainty a word permitted by language. In the "reading-text"-operation along a given text path we thus have the smallest diameter of just $D_{OP} = 1$, which corresponds with the best possible connectivity.

The third step: Last but not least we may remember that another well known relation is valid between the word nodes in the old last class and the word nodes in the adjacent new class. Referring to equ. (1c) we find for the magnification of word contents, if we set $i = K$:

$$\frac{2^{i+1}}{2^i} = \frac{2^{K+1}}{2^K} = 2. \tag{44d}$$

That corresponds with Fig. 1b where a new class always contains twice as much word nodes as the preceding class. This growth of the network system reminds us of the cell division process in biology which perhaps could also take place in the neural networks of the language brain.

Summarising the points: In the view of mathematics three possible mechanisms of growth exist: a) The whole number of word nodes are doubled, b) equal size packets of new connections are sent from all classes to nodes of the new class and c) the nodes of the last class generate twice as much new nodes for the new class.

It seems that the last two points have a certain prospect of physical reality whereas the first point is only a nice conclusion from the model structure.

12. Language and prime numbers

Language still holds many mysteries. For example, the question whether there is a deep connection between Zipf's law (or the structural language model respectively) and prime numbers or not, is still completely open. In any case one cannot disregard a formal mathematical correspondence between both fields which was unknown up to now. The topic shall be sketched out briefly. When we consider an arbitrary coherent text of natural language it is remarkable and well known that with increasing text length ever smaller numbers of new words appear. We find a corresponding fact in mathematics along the number line with increasing natural numbers. The prime numbers included here become more and more rare. Even the same formulae as for the structural language model are valid. The best known fundamental formula for prime numbers was already discovered by the great mathematician Karl Friedrich Gauß at age 14 (Courant 1941; Ingham 1964). He supposed that the quantity G of the prime numbers below the natural number x can be estimated by the formula

$$G = \frac{x}{\ln x}\,. \tag{45}$$

This estimation generates values which are always a little bit too small, if they are compared with the laboriously and precisely counted genuine quantity which is called $\pi\,(x)$. Stimulated by considerations about the structural language model one could also be interested in the inversion formula of equ. (45). It can be found (Hilberg 1988) after taking the logarithm of the last equation

$$\ln G = \ln x - \ln \ln x. \tag{46}$$

Obviously for very large x the following approximation is valid

$$\ln G \approx \ln x. \tag{47}$$

Inserting in equ. (45) gives the complementary approximation (Hilberg 1990):

$$x \approx G \cdot \ln G\,. \tag{48}$$

This estimation generates on the contrary always values which are somewhat too large for the exact quantity of prime numbers. These two results taken together open up the possibility for considerable improvements of the estimations by an interpolation between both values (if the deviations are in the same order of magnitude), e.g. by using the geometric mean (Hilberg 1988). (An example: for $x = 10^6$ the correct value is $\pi\,(x) = 78498$, the Gauß-approximation is G = 72382, the complementary value from equ. (48) is G_{COMPL} = 87847 and the geometric mean is $\overline{G} = \sqrt{GG_{COMPL}}$ = 79741). However this shall not be examined here. Instead we will pursue another idea.

In equ. (4) above J. Meyer found for Zipf's power-law the following relation between the quantity N of different words in text and the text length L. (For the sake of simplicity here we shall omit the indices):

$$L = N \cdot \ln N. \tag{49}$$

Of course the inverse function is again in good approximation (Hilberg 1990):

$$N = \frac{L}{\ln L}. \tag{50}$$

When we compare the two equations (45) and (48) with the two equations (50) and (49) we may suppose that the values G and N and the values x and L correspond. So to speak, the number sequence $1,2,3,...,x$ would be treated as if it was only a peculiar text, where prime numbers take the part of the different words of a vocabulary (rank words) which appear in text.

In this comparison we find both common properties and differences. For example, plotting the numbers including prime number factors over prime numbers, an hyperbolic envelope can be found (power-law with linear scaled axes) (Hilberg 1988), see Fig. 9. Therefore it is not astonishing that the same mathematical formulae result when the curves are integrated or summed up. The maximum difference is surely the fact that we know with certainty that prime numbers appear more and more rarely in the advancing number line and that this behavior never stops because we have an infinite sequence. The same is not exactly true with new words emerging in text, though this behavior could be verified by us for many texts up to lengths in the order of 20 million running words. Moreover, every year about 2000 new words are added to the German language.

Of course there are also the connections to the structural language model above: If the sequence of prime numbers is divided into classes of size 2^i, with $i = 0,1,2,3,...,K$, we can for the most part take over the initial considerations referring to text and words. This is advantageous because we can remove at least the obscurity which the well known mathematician R. Courant (1941) expressed in the following way: "That the average behavior of the prime number distribution can be described by the logarithmic function is a very remarkable discovery, for it is surprising that two mathematical concepts which seem so unrelated should be in fact so intimately connected." Now the values N, L, G, x are no longer given by natural logarithms but by the whole number K and powers of 2 in equ. (3).

In my opinion the formal identity of equations, by which relations in language and in mathematics are estimated, is not due to pure chance. Seen from the statistical point of view, in both cases the next event is always uncertain: Neglecting context, as we can do easily in an artificially generated text, nobody can tell if the next word is a new word from the vocabulary or again a word which was used before. In mathematics, if we proceed along the number line, nobody knows with certainty and without actual testing, whether the next uneven number is a prime number or not. The problems are very similar and the probabilities for a lucky hit decrease in the same way.

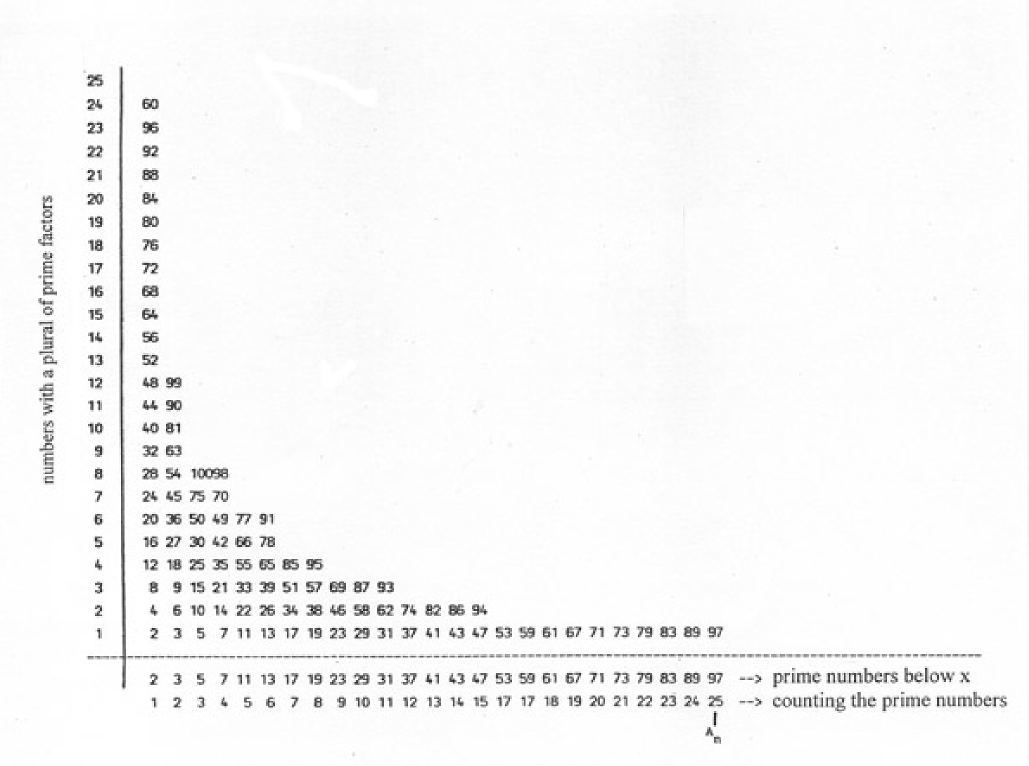


Fig. 9 : The numbers from *x* = 2 to *x* = 100 (including a plural of prime factors) in dependence of the sequence of prime numbers.

References

Bassenge, G. (2002). *Automatische Klassifizierung von Wortformen in Texten der deutschen Gegenwartssprache.* Dissertation TU Darmstadt.

Burschel, H.-D. (1999). *Die meßtechnische Ermittlung von Assoziationen zwischen Worten in kohärentem Text und ihre Nutzung bei Prädiktionen verschiedener Reichweite.* Dissertation TU Darmstadt.

Courant, R. (1941). *What is Mathematics?* New York: Oxford University Press. (German translation: Springer Verlag).

Gell-Mann, M. (1994). *The Quark and the Jaguar. Adventures in the Simple and the Complex.* New York: Freeman.

Hilberg W. (1988). Zur Abschätzung der Primzahlfunktion π (x). *Institutsbericht Nr. 86/88*, TH Darmstadt.

Hilberg, W. (1990). *Die texturale Sprachmaschine als Gegenpol zum Computer*. Groß-Bieberau: Verlag für Sprache und Technik.

Hilberg, W. (1991). Das Netzwerk der menschlichen Sprache und seine technische Realisierung. *Institutsbericht Nr. 126/90,* THD or in *Frequenz 45 (1991) 11-12, pp. 275-284.*

Hilberg, W. (2000). *Große Herausforderungen in der Informationstechnik.* Groß-Bieberau: Verlag Sprache und Technik.

Hilberg, W. (2002). Wie wirklich ist ein Gedanke? — Wittgenstein und die Informationstechnik. *Thema Forschung, TU Darmstadt, 2, 104-109.*

Hilberg, W. (2002). The Unexpected Fundamental Influence of Mathematics upon Language. *Glottometrics 5, 29-50.*

Hilberg, W. (2003). *The Unexpected Fundamental Influence of Mathematics upon Language*. Institutsbericht, FG Digitaltechnik Nr. 227C, 2003.

Hilberg, W. (2003). *Übersichtliche graphische Darstellung von Codierung und Decodierung von Text*. Institutsbericht, FG Digitaltechnik Nr. 243.

Ingham, A.E. (1964). *The Distribution of Prime Numbers*. New York: Stechert-Hafner..

Meyer, J. (1989). *Die Verwendung hierarchisch strukturierter Sprachnetzwerke zur redundanzarmen Codierung von Texten*. Dissertation TH Darmstadt.

Ries, T. (2002) *Über Möglichkeiten einer maschinellen Nacherzählung mit einem konnektionistischen System aus neuronalen Sprachnetzwerken.* Dissertation TU Darmstadt.

Shannon, C.E. (1951). Prediction and Entropy of Printed English. *The Bell System Technical Journal, Jan. 50-64.*

Steinmann, M. (1996). *Netzwerkmodellierung und Segmentierung von Texten sowie Anwendungen zur Informationsverdichtung.* Dissertation TH Darmstadt.

Appendix A. Symbols

The usage of symbols in this field of science is very confusing. Therefore an explanation of the symbols used in this paper may be necessary. The term "frequency" was used once by G.K. Zipf. But this denomination could not be adopted in the world of technology, because "frequency" has been defined here for more than a century as the number of events or cycles per second. Yet Zipf obviously handled the number of special events occuring during the whole measurement. That is simply an accumulation. Thus some years ago instead of "*f*" the symbol "a" was introduced by the engineers in the project "language machine", e.g. in Meyer (1989), which can be read as "absolute frequency" or as "accumulation". Meanwhile it was learned that linguists still prefer the usage of the word "frequency" and the symbol "*f*". Therefore here a compromise is proposed to use the symbol f_a standing for accumulated or absolute frequency.

The second variable in Zipf's diagram is "rank". Words are ordered by (accumulated) frequency and receive corresponding numbers. Rank is often represented by the letter "*r*". (That can be found also in the preceding paper Hilberg 2002). However, usually the running variables are denoted by the symbols "*n*", "*i*" or "*j*". Those are in most cases easily and immediately understandable whereas "rank *r*" induces questions and longer discussions which should be avoided for the present. In papers like this one problems arise when diagrams from various sources in literature shall be shown. Original diagrams should be given in original form. But this will result in misunderstandings. Here, if necessary, the problem shall be solved by giving an explanation of the correlations between the symbols *a, f, r* and the symbols f_a and *n* in the descriptions of the Figures.

Appendix B. General power laws in the shape of staircases

Power laws have been discovered in various fields of science (Gell-Mann 1994). Usually they differ from one another through the slope of their curves. Thus, as is well known, with the aid of a constant exponent ν a generalized power law can be written as follows

$$f_{a,P} = \frac{N}{n^{\nu}} \; . \qquad \text{B(1)}$$

The straight line in the usual diagrams with logarithmic axes then results from this equation in the following way:

$$\log (f_{a,P}) = \log N - \nu \log n \; . \qquad \text{B(2)}$$

Now, if a division of the sequence of numbers *n* and the collection in classes is possible and reasonable, the straight line in a diagram with logarithmic axes can be replaced by a general staircase function:

$$f_{a,G} = 2^{K-\nu i} \;, \qquad \nu = \text{const.}, \qquad i = 0,1,2,... \; . \qquad \text{B(3)}$$

By using the constant ν appropriately the slope of the staircase can be steepened or flattened. When in particular the constants ν = 2,3,4,... and the corresponding limits $i_{\max}$ = *K*/2, *K*/3, *K*/4,... are used, the staircase becomes steeper, and when the constants ν = 1/2, 1/3, 1/4,... with the corresponding limits $i_{\max}$ = 2*K*, 3*K*, 4*K*, ... are used the staircase will be flattened.

If a constant activity in all classes is required in the way as was discussed in chapter 4, we have to accomplish

$$2^{K-\nu i} \cdot 2^{x} \equiv \text{const.} \qquad i = 0, 1, 2,...,K. \qquad \text{B(4)}$$

where *x* is a variable. With the value $x = \nu i$ it follows

$$2^{K-\nu i} 2^{\nu i} = 2^{K} \, . \qquad \text{B(5)}$$

The enlargement of classes becomes with every new step

$$\frac{2^{\nu(i+1)}}{2^{\nu i}} = 2^{\nu} \, . \qquad \text{B(6)}$$

At the same time the amplitude changes like

$$\frac{2^{K-\nu(i+1)}}{2^{K-\nu i}} = 2^{-\nu} = \frac{1}{2^{\nu}} \, . \qquad \text{B(7)}$$

Now a distinction has to be made between the two cases $\nu > 1$ and $\nu < 1$. In the first case the factor 2^{ν} is always an integral number, as can be seen in the example $\nu = 2$ with the factor $2^{2} = 4$. In the second case however, there is no longer an integral number, as can be seen in the example $\nu = 1/2$ with the factor $2^{1/2} = \sqrt{2}$. Then serious problems of interpretation arise immediately.

The case where $\nu = 1$, as in the language model above now proves to be the simplest and also an exceptional good solution, whose realization in real life by men could be accomplished most easily (although in long times). In the directly adjacent case where $\nu = 2$

also no difficulties exist for an extension of the theory and for practical realizations. Therefore it shall not be surprising if a steepened slope for so-called "power laws" with a rather precise value ν = 2 should be met in science more frequently.

Glottometrics 7, 2004, 25-49

Zur Entwicklung des deutschen (Lehn-)Wortschatzes

Helle Körner[1]

Abstract. Logistic laws do not only apply to linguistic but also to medical, biological, demographic etc. developmental phenomena. This study will support the logistic law, known in linguistics as Piotrowski Law, using data from a selected etymological dictionary. In contrast to other studies there has ─ in this process ─ not only been taken account of terms borrowed from other languages but also of words which developed in German language. Special attention has been paid to the Anglo-American terms.

Keywords: Word stock, German, Piotrowski Law, borrowings

1. Einleitung

Die vorliegende Studie untersucht erneut (vgl. z.B. Best 2001, Best 2001a, Best 2003, Best 2003a, Best & Altmann 1986, Müller-Hasemann 1986) Entlehnungsprozesse im Deutschen. Dabei geht es wie in früheren Untersuchungen darum, Gesetzmäßigkeiten des Sprachwandels zu testen. Diese Studie versucht, die Geltung des logistischen Gesetzes, das in der Quantitativen Linguistik auch unter dem Namen „Piotrowski-Gesetz“ bekannt ist, anhand von Daten zu überprüfen, die aus der Auswertung von *Duden. Das Herkunftswörterbuch* (2001)[2] gewonnen werden konnten. Es handelt sich dabei also um einen Ausschnitt des deutschen Wortschatzes, d.h. es liegt keineswegs eine repräsentative Studie des deutschen Wortschatzes vor, wenn auch alle Stichwörter des Duden in dieser Studie berücksichtigt wurden. Das logistische Gesetz ist anwendbar an verschiedene Formen des Sprachwandels, wobei unter Sprachwandel der Veränderungsprozess von Sprachelementen und Sprachsystemen in der Zeit verstanden wird.

Die grundsätzliche Hypothese zum Verlauf des Sprachwandels beruht auf der Annahme, dass ein Mitglied einer Sprachgemeinschaft eine Neuerung verwendet, die von anderen Personen übernommen wird. Ausgehend von dieser These lassen sich drei unterschiedliche Formen des Sprachwandels unterscheiden:

1. der vollständige, bei dem alte Formen vollständig durch die neuen Formen ersetzt werden (z.B. *was* zu *war*)
2. der unvollständige, bei dem sich die neuen Formen und Wörter nur in einem begrenzten Maß durchsetzen (z.B. Fremdwörter) und
3. der reversible Sprachwandel, bei dem neue Formen und Wörter aufkommen, sich ausbreiten und dann wieder verschwinden (z.B. die e-Epithese; vgl. Imsiepen 1983).

Zur Darstellung der Sprachwandelprozesse entwickelte Altmann (1983: 59) ausgehend von Piotrowskis Überlegungen einen Gesetzesvorschlag: „Unter dem Piotrowski-Gesetz verstehen wir die hypothetische Aussage über den zeitlichen Verlauf einer beliebigen sprach-

[1] Address correspondence to: Helle M. Körner, e-mail: hmkoerner@web.de

[2] Wenn im weiteren Verlauf dieser Arbeit der Duden erwähnt wird, so ist damit immer – falls nicht ausdrücklich anders benannt – der *Duden. Das Herkunftswörterbuch* (2001) gemeint.

lichen Entität.“ So sind mittels dieses Gesetzes Voraussagen darüber möglich, wie ein begonnener Sprachwandel weiter verläuft – wenn auch nur unter der Prämisse, dass die Bedingungen, unter denen dieser Sprachwandel stattgefunden hat, sich nicht wesentlich verändern, sondern gleichbleibend sind. Für den unvollständigen Sprachwandel hat Altmann folgende mathematische Funktion entwickelt, wobei *c* die Asymptote und *a* und *b* Parameter darstellen:

$$p_t = \frac{c}{1 + ae^{-bt}} \tag{1}$$

Es handelt sich hierbei um ein Wachstumsmodell vom logistischen Typ, das in vielen Wissenschaften Anwendung findet. Es soll nun ein weiteres Mal geprüft werden, ob dieses Gesetz geeignet ist, Wandelprozesse im deutschen Wortschatz zu modellieren. Dabei werden nicht nur die Fremdwörter berücksichtigt, die ins Deutsche übernommen worden sind, sondern erstmalig auch die Wörter, die zum deutschen Erbwortschatz gehören bzw. Neubildungen im Deutschen sind. So können anhand der Zusammensetzung des deutschen Wortschatzes Rückschlüsse auf die Bedeutung der einzelnen Sprachen für den Wortschatz gezogen werden, und Vorwürfe wie „das Überhandnehmen des Englischen in der deutschen Sprache“ können so vielleicht entkräftet werden.

2. Methodik

Für die Überprüfung des logistischen Gesetzes anhand des deutschen Wortschatzes wurde als Datenbasis *Duden. Das Herkunftswörterbuch* (2001) ausgewählt: Zum einen deshalb, weil er zum Zeitpunkt der Untersuchung eine aktuelle Bearbeitung eines etymologischen Wörterbuchs war (und damit speziell eine größere Anzahl an Anglizismen bzw. Amerikanismen berücksichtigt als frühere Wörterbücher), zum anderen, weil er nach eigenen Angaben 20000 Einträge beinhaltet und damit eine vergleichsweise große Anzahl von Stichwörtern enthält. Das *Etymologische Wörterbuch des Deutschen* von Wolfgang Pfeifer hingegen beinhaltet mit 22000 Wörtern zwar inhaltlich mehr, stammt aber in der Erstauflage schon aus dem Jahr 1989, so dass der Überblick über das letzte Jahrhundert nicht vollständig ist. Für das *Etymologische Wörterbuch der deutschen Sprache* von Friedrich Kluge gilt, dass es in der hier verendeten 23. Auflage zwar erst 1999 erschienen ist, aber nur 13000 Wörter enthält[3].

Bei der Bearbeitung des Duden, d.h. bei der Auswertung der Stichwörter, wurden insgesamt 17828 Wörter ausgezählt. Die Differenz zu der vom Duden angegebenen Anzahl kommt dadurch zustande, dass im Duden auch Sprichwörter, Redewendungen u.ä. berücksichtigt wurden, die aber für eine Erfassung des deutschen Wortschatzes im Rahmen dieser Studie irrelevant waren. Um eine systematische Auswertung zu ermöglichen, waren Angaben über das Jahrhundert der Übernahme bzw. der Bildung im Deutschen und bei Entlehnungen die Nennung der Vermittlersprache notwendig. Von diesen 17828 Wörtern sind 16781 ausreichend genau bestimmbar, d.h. das Jahrhundert der Übernahme sowie die Vermittlersprache sind erkennbar, während 1047 Wörter aufgrund fehlender oder widersprüchlicher Angaben bezüglich der Vermittlersprache bzw. der mangelnden Datierbarkeit nicht zuzuordnen sind.

Im Rahmen dieser Studie wurde von der Richtigkeit der Angaben im Duden ausgegangen. Das heißt, sofern Angaben im Duden gemacht wurden, die eine eindeutige Zuordnung zu (a) einem Jahrhundert und (b) einer Vermittlersprache erlaubten (bzw. klar erkennbar war, dass das Wort im Deutschen gebildet wurde), wurden diese Angaben ohne nochmalige Über-

[3] Für diese erste Einschätzung beziehe ich mich nur auf die Angaben, die die Autoren selbst über den Umfang der Wörterbücher machten.

prüfung durch andere Wörterbücher übernommen. Traten hinsichtlich (a) und (b) jedoch Unstimmigkeiten auf, wurden die Wörterbücher von Kluge (1999) sowie Pfeiffer (2000) hinzugezogen. Stimmten diese beiden überein, wurden ihre Daten übernommen. Wichen die Angaben voneinander ab, wurde wie folgt verfahren: Wenn zwei der drei verwendeten Wörterbücher Übereinstimmungen zeigten, wurde dies wiederum als eindeutige Angabe gewertet. Wenn Widersprüche auftraten, wurde das jeweilige Stichwort als 'nicht zuzuordnen' ausgesondert. Bei Abweichungen hinsichtlich der Datierung wurde die jüngste (d.h. neueste) Zeitangabe berücksichtigt, da davon auszugehen ist, dass das Wort spätestens zu diesem Zeitpunkt tatsächlich im Deutschen existierte.

Als Fremdwort wurde ein Wort nur dann gewertet, wenn es komplett aus einer anderen Sprache als dem Deutschen stammte; einzelne 'fremde' Lexeme wurden – sobald sie einmal im Deutschen vorlagen – nicht ein weiteres Mal berücksichtigt. Daraus folgt, dass ein komplexes Wort dann als im Deutschen gebildet angesehen wird, wenn mindestens eines seiner Lexeme deutsch ist (z.B. gilt *Immatrikulationsbescheinigung* wegen des Bestandteils *–bescheinigung* als deutsches Wort, während *Immatrikulation* als Fremdwort gewertet wird). Dies gilt jedoch nicht für die Wortbildungsmorpheme wie z.B. *–isch* bei Adjektiven oder *–ier* bei Verben. Im Rahmen dieser Arbeit wurde also davon ausgegangen, dass ein fremder Bestandteil zunächst einmal ins Deutsche integriert wurde (teilweise sogar mit kompletter Anpassung sowohl an die deutsche Phonologie als auch an Grammatik und Orthographie), der dann wiederum in der Wortbildung des Deutschen aktiv werden konnte. Sofern Abkürzungen unter dem Stichwort auftraten (wie zum Beispiel *Abi* zu *Abitur*), wurden diese berücksichtigt, wenn eine Datierung möglich war.

2.1. Behandlung von Problemen bei Zeitangaben

Generell wurde bei der Datierung so vorgegangen, wie dies Best (vgl. z.B.: Best, 2001a: 8) in seinen Untersuchungen vorschlägt: Angaben wie „15./16. Jahrhundert" wurden dem erstgenannten Zeitraum zugerechnet, Angaben wie „um 1500" dem 16. Jahrhundert. Für Formulierungen wie „in neuerer/jüngster Zeit" wurden, da sie für eine Datierung nicht ausreichten, die beiden weiteren Wörterbücher herangezogen, was in den meisten Fällen eine ausreichende Datierung erlaubte. Jeweils die erste, d.h. zeitlich früheste Nennung des Wortes war für die Datierung in dieser Studie relevant.

Des weiteren ergaben sich Probleme bei der Datierung des Althochdeutschen bzw. des Mittelhochdeutschen: Wörter in diesen beiden Gruppen waren in den seltensten Fällen im Duden tatsächlich auf ein Jahrhundert datiert. Daher wurden zunächst einmal *Althochdeutsch* bzw. *Mittelhochdeutsch* als grobe Zeitangabe berücksichtigt, die nach den im Duden aufgenommenen Tabellen zur zeitlichen Gliederung des Hochdeutschen (Duden, 2001: 259) dann festgelegt wurden. Somit wurden alle Wörter, die der Gruppe Althochdeutsch zuzuordnen und mit keiner weiteren Datierung versehen waren, mit der Zeitangabe 11. Jahrhundert eingeordnet, da der Duden für das Althochdeutsche den Zeitraum von 700-1050 angibt; d.h. spätestens im 11. Jahrhundert müssen alle diese Wörter in der deutschen Sprache vorhanden gewesen sein. Analog wurde mit der Zuordnung Mittelhochdeutsch verfahren (laut Duden von 1050-1450); hier wurden die Wörter also in das 15. Jahrhundert datiert. Beim Frühneuhochdeutschen wurde als Datierung das 17. Jahrhundert bestimmt. Da eine Zuordnung zu einem Jahrhundert somit gegeben war, wurde auf eine nochmalige Überprüfung anhand der weiteren Wörterbücher verzichtet.

Bei den im Deutschen gebildeten Wörtern traten neben den bereits angesprochenen Schwierigkeiten noch zwei weitere Gruppen von Problemfällen auf. Bei der Bedeutungsvermischung kreuzten sich zwei unterschiedliche Wörter. Dies gilt beispielsweise für *ersticken*:

So wurden *irsticken* aus dem Althochdeutschen und *erstecken* aus dem Mittelhochdeutschen zusammengezogen; hierbei wurde davon ausgegangen, dass es sich eher um einen Bedeutungswandel sowie lautlichen Wandel handelt als tatsächlich um die Bildung eines neuen Wortes. Damit wurde das Wort auf das 11. Jahrhundert (also althochdeutsch) datiert. Bei Zusammenrückungen (z.B. althochdeutsch *in bor* wurde zu mittelhochdeutsch *empor*) wurde analog verfahren.

Ebenso wurde bei Bedeutungsverschiebungen (d.h. die Bedeutung eines Wortes hat sich beispielsweise von *erfahren* im Sinne von 'reisen' zu dem heutigen Gebrauch in der Bedeutung 'erforschen' geändert) vorgegangen: War eine klare Linie der Bedeutungsveränderung erkennbar, wurde ebenfalls die älteste Datierung herangezogen und das Wort nicht als Neubildung betrachtet, da nicht nur neue Wörter übernommen oder gebildet, sondern auch bereits vorhandene je nach Absicht oder Notwendigkeit modifiziert wurden – sowohl in semantischer wie auch phonetischer Hinsicht. Daraus folgt, je älter ein Wort ist, desto größer sind auch seine etymologischen Abweichungen in Bezug auf das 'Ursprungswort'. Dies gilt zumindest für die meisten Wörter; dabei ist es unerheblich, ob sie im Althochdeutschen gebildet wurden oder zu althochdeutscher Zeit aus anderen Sprachen übernommen wurden.

2.2. Probleme bei der Vermittlersprache

Für die Auswertung etymologischer Wörterbücher gibt es zwei Herangehensweisen: Entweder wird die Vermittlersprache, d.h. die Sprache, über die ein Wort ins Deutsche gelangt ist, berücksichtigt, oder aber die Herkunftssprache, d.h. die Sprache, aus der ein Wort ursprünglich stammt. Je nach Verfahren ergeben sich also andere Zuordnungen: Das Wort *Alabaster* wird im ersten Fall (Vermittlersprache) als aus dem Lateinischen übernommen betrachtet, im zweiten Fall (Herkunftssprache) jedoch als Wort griechischer Herkunft. In Anlehnung an Best (2001a: 9) war für diese Auswertung lediglich die Vermittlersprache ausschlaggebend. Das heißt, dass Angaben wie „unter französischem Einfluß" nicht berücksichtigt wurden, wenn die Angabe der Vermittlersprache klar war – es sei denn, es gab berechtigte Zweifel, da die beiden anderen Wörterbücher (Pfeifer und Kluge) übereinstimmten, dass das Wort tatsächlich aus dem Französischen übernommen wurde (wie z.B. bei dem Wort *arrogant*).

Da aus bestimmten Einzelsprachen zu wenige Belege für eine Auswertung vorlagen, sind sie zu 'Sprachgruppen' zusammengefasst worden. Die Möglichkeit der Zusammenfassung wurde dadurch forciert, dass im Duden teilweise statt einer Einzelsprache lediglich die Sprachfamilie als Entlehnungsquelle angegeben wurde. Die Zusammenfassung betrifft vor allem die Sprachen der slawischen Sprachfamilie, die demzufolge als *slawisch* bezeichnet wurden (russisch, tschechisch, polnisch, ukrainisch, serbokroatisch, slowakisch, kaschubisch, slowenisch). So konnte ein Datensatz mit 44 Belegen erstellt werden. Eine umfangreichere Zusammenstellung slawischer Entlehnungen findet sich in Best 2003a. Auch aus dem Dänischen, Schwedischen, Norwegischen und Isländischen wurde so eine Sprachgruppe gebildet, die als *nordgermanisch* bezeichnet wurde. Trotz dieser Umstrukturierung waren es in diesem Fall zu wenig Belege für eine Auswertung. Das gleiche gilt für die Gruppe von Sprachen, die unter dem Begriff *keltisch* zusammengefasst wurden (Irisch, Gälisch, Bretonisch).

Weitere Zweifelsfälle wurden wie folgt behandelt: Zusammenrückungen aus dem Lateinischen (wie z.B. lat. *aqua vitae* – dt. *Aquavit*) wurden als in der fremden Sprache gebildet angesehen, ebenso wie die botanischen Pflanzenbezeichnungen. Wörter, deren Herkunft als „Englisch", „Amerikanisch" oder „Englisch-Amerikanisch" angegeben wurde, sind alle unter „Englisch" notiert, da man das Amerikanische als eine Varietät des Englischen betrachten kann (vgl. z.B.: Stickel, 2001a: 2). Ähnlich wurde auch bei Angaben wie „Schweizerisch" bzw. „vom Wiener Hof" verfahren: Weder das schweizerische noch das österreichische

Deutsch wurden als eigene Sprachen gewertet. Das Niederdeutsche dagegen wurde als selbständige Sprache behandelt; denn so wie sich durch die erste Lautverschiebung das Deutsche vom Germanischen getrennt hat, so erfolgte durch die zweite Lautverschiebung die Trennung von Hoch- bzw. Mitteldeutsch und Niederdeutsch.

Anders als bei Best (2001a) und Kirkness (1988) sind die „deutschen Fremdwörter“ (d.h. Wörter, die aus fremdsprachlichen Lexemen gebildet werden, in dieser Form aber gar nicht in der vermeintlichen Gebersprache existieren wie z.B. *Handy*) im Rahmen dieser Untersuchung nicht gesondert berücksichtigt worden, sondern wurden unter der Rubrik ‘Im Deutschen gebildete Wörter’ aufgeführt. Insgesamt konnten anhand dieses Vorgehens Datensätze für 11 Sprachen gewonnen werden, an die das von Altmann entwickelte Modell für den unvollständigen Sprachwandel angepasst werden konnte. Ein zwölfter Datensatz ergab sich aus dem Gesamtwortschatz des Deutschen.

3. Auswertung

Einen ersten Gesamteindruck soll folgende Tabelle vermitteln, die zunächst eine Gegenüberstellung der ins Deutsche übernommenen Fremdwörter bzw. der im Deutschen gebildeten Wörter zeigt.

Tabelle 1
Verteilung des deutschen Wortschatzes auf die verschiedenen Sprachen

Sprache	**Anzahl**	**Prozent**	**Sprache**	**Anzahl**	**Prozent**
Deutsch	11537	68.750	Arabisch	5	0.030
Latein	2031	12.103	Portugiesisch	5	0.030
Französisch	1424	8.502	Keltisch	4	0.024
Niederdeutsch	545	3.248	Eskimo	3	0.018
Englisch	519	3.092	Gotisch	3	0.018
Italienisch	286	1.704	Indisch	3	0.018
Griechisch	144	0.858	Afrikaans	2	0.012
Niederländisch	87	0.518	Malaiisch	2	0.012
Slawisch	44	0.262	Chinesisch	1	0.006
Spanisch	43	0.256	Finnisch	1	0.006
Rotwelsch	41	0.244	Hebräisch	1	0.006
Nordgermanisch	12	0.072	Hunnisch	1	0.006
Japanisch	10	0.060	Ladinisch	1	0.006
Ungarisch	10	0.060	Persisch	1	0.006
Türkisch	8	0.048	Polynesisch	1	0.006
Jiddisch	6	0.036	**Summe:**	**16781**	**100.000**

Wie die Tabelle zeigt, stehen im *Duden. Das Herkunftswörterbuch* den 68.75% im Deutschen gebildeten Wörtern 31.25% Fremdwörter gegenüber. Die englischen Fremdwörter machen gerade einmal etwas über 3% am Gesamtwortschatz aus[4], so dass der Vorwurf von Sprachgesellschaften, die deutsche Sprache werde durch die wachsende Übernahme gerade englischer

[4] Wenn im weiteren von Wortschatz die Rede ist, ist damit immer der im *Duden. Das Herkunftswörterbuch* verzeichnete Wortschatz, nicht aber der nicht genau bestimmbare Gesamtwortschatz des Deutschen gemeint.

Fremdwörter ausgedünnt, sich als definitiv nicht zutreffend erweist[5]. Die meisten Fremdwörter wurden aus dem Lateinischen übernommen, wobei es sich nicht immer um tatsächlich lateinische Wörter handelt. Auch griechische Wörter wurden häufig über das Lateinische vermittelt, woraus sich wiederum der relativ geringe Anteil der griechischen Wörter am in dieser Studie erfassten deutschen Wortschatz erklärt. Bei der weiteren Auswertung der einzelnen Sprachen wird prinzipiell nach der Reihenfolge dieser Tabelle vorgegangen, mit der Ausnahme, dass das Deutsche am Ende des Auswertungsteils behandelt wird.

Lässt man nun einmal die im Deutschen ererbten oder neu gebildeten Wörter außer acht, ergibt sich für die Anteile der einzelnen Vermittlersprachen folgendes Bild:

Tabelle 2
Verteilung der Fremdwörter im Deutschen auf die Vermittlersprachen

Sprache	Anzahl	Prozent	Sprache	Anzahl	Prozent
Latein	2031	38.730	Portugiesisch	5	0.095
Französisch	1424	27.155	Keltisch	4	0.076
Niederdeutsch	545	10.393	Eskimo	3	0.057
Englisch	519	9.897	Gotisch	3	0.057
Italienisch	286	5.454	Indisch	3	0.057
Griechisch	144	2.746	Afrikaans	2	0.038
Niederländisch	87	1.659	Isländisch	2	0.038
Slawisch	44	0.839	Malaiisch	2	0.038
Spanisch	43	0.820	Chinesisch	1	0.019
Rotwelsch	41	0.782	Finnisch	1	0.019
Japanisch	10	0.191	Hebräisch	1	0.019
Nordgermanisch	10	0.191	Hunnisch	1	0.019
Ungarisch	10	0.191	Ladinisch	1	0.019
Türkisch	8	0.153	Persisch	1	0.019
Jiddisch	6	0.114	Polynesisch	1	0.019
Arabisch	5	0.095	**Summe:**	**5244**	**100.000**

38.730% aller ins Deutsche übernommenen Fremdwörter sind über das Lateinische vermittelt worden; das ist der größte Anteil einer einzelnen Sprache. Danach folgt das Französische mit 27.136%. Die weiteren prozentualen Werte nehmen rapide ab: Der Anteil niederdeutscher Wörter beträgt nur noch 10.393%, das Englische liegt knapp unter 10%.

Als erstes wird das logistische Gesetz an die Entwicklung des gesamten datierbaren Wortschatzes angepasst[6]. So ergibt sich dafür Tabelle 3. *n* gibt die anhand der Auszählung gewonnenen Daten an; dementsprechend werden diese Daten unter *n* (kumuliert) zusammengefasst, während *p* (ber.) die berechneten Zahlen wiedergibt[7]. *a* und *b* sind dabei Parameter, *t* steht für die Zeiteinheit von 100 Jahren. *c* gibt den berechneten Grenzwert für den Sprach-

[5] Vgl. zu dieser Thematik z.B.: www.deutsche-sprachwelt.de/berichte/gesetz/dey03.shtml, Stand vom 14.10.02.

[6] Wenn im weiteren Verlauf dieser Auswertung vom logistischen Gesetz die Rede ist, so ist damit – wenn nicht ausdrücklich anders erwähnt – immer das entsprechende Piotrowski-Gesetz in der Form des unvollständigen Sprachwandels gemeint.

[7] Diese Benennungen gelten auch für die weiteren Tabellen. Die Berechnung erfolgte mittels der Software NLREG.

wandel an[8], der anzeigt, gegen welchen Wert der Sprachwandel strebt. Dabei ist unter *c* nicht ein absoluter Wert zu verstehen, der tatsächlich angibt, wie viele Wörter im Höchstfall aus der jeweiligen Sprache übernommen werden, sondern nur eine Tendenz, die je nach Datenbasis variiert. *D* ist der Determinationskoeffizient. Je größer *D* ist, desto besser gelingt die Anpassung: Es soll mindestens $D \geq 0.80$ sein, um anzuzeigen, dass das Modell den Sprachwandelprozess in annehmbarer Weise wiedergibt. Es handelt sich also mit dem Wert von D = 0.98 um eine sehr gute Anpassung.

Tabelle 3
Entwicklung des deutschen Gesamtwortschatzes

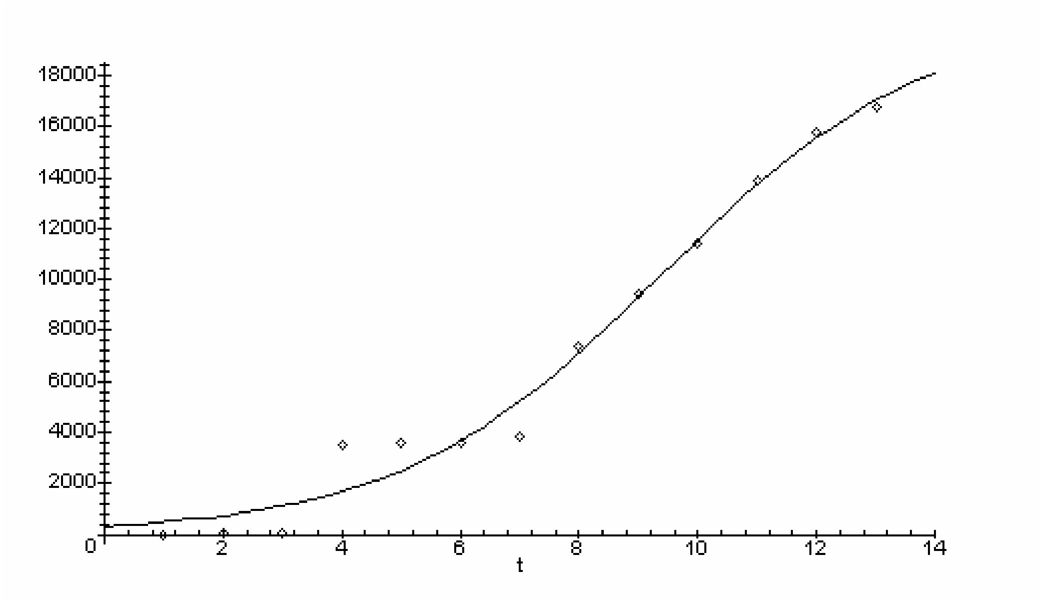
Jhd.	***t***	***n***	***n* (kumuliert)**	***p* (ber.)**
8.	1	2	2	474.58
9.	2	1	3	730.77
10.	3	3	6	1117.54
11.	4	3520	3526	1691.48
12.	5	23	3549	2521.80
13.	6	68	3617	3679.90
14.	7	199	3816	5215.42
15.	8	3574	7390	7120.20
16.	9	2070	9460	9297.08
17.	10	1984	11444	11563.64
18.	11	2435	13879	13705.84
19.	12	1902	15781	15553.03
20.	13	1000	16781	17023.98
a = 65.7674 b = 0.4446 c = 20484.8539 D = 0.98				

Die Abweichungen der beobachteten und der gemessenen Werte im Zeitraum bis zum 15. Jahrhundert sind auf die bereits angesprochenen Schwierigkeiten bei der Datierung jener Angaben zurückzuführen, die lediglich „Althochdeutsch“ bzw. „Mittelhochdeutsch“ als Zeitpunkt der Übernahme oder der Entstehung erkennen ließen. Für die Anpassung an den im Duden enthaltenen Wortschatz kann das logistische Gesetz dennoch bestätigt werden. Aus den errechneten Parametern und dem ermittelten Grenzwert ergibt sich für die Übernahme der Fremdwörter bzw. die Bildung von Wörtern im Deutschen folgender Term:

$$p_t = \frac{20484.8539}{1 + 65.7674e^{-0.4446t}} \tag{2}$$

Dieses Beispiel ist exemplarisch für die Einsetzung der durch Berechnung gewonnenen Daten in die Formel für das logistische Gesetz; daher werden bei den weiteren Tabellen diese Formeln nicht mehr gesondert aufgeführt. Die Graphik spiegelt die gute Annäherung zwischen den berechneten und den beobachteten Daten wider (s. Graphik zu Tabelle 3).

[8] Diese Parameter *a*, *b* und *c* bezeichnen auch bei allen weiteren Tabellen die Parameter, die zum logistischen Gesetz gehören.



Graphik zu Tabelle 3: Entwicklung des deutschen Gesamtwortschatzes

Auf der y-Achse ist dabei die Anzahl der Wörter, auf der x-Achse die Zeit (in Jahrhunderten) abgetragen worden. Die durchgehende Linie gibt in Übereinstimmung mit Formel 2 den Verlauf der berechneten Werte an, die Punkte die tatsächlich gemessenen Werte[9]. Es erfolgt eine asymptotische Annäherung an den Grenzwert *c*, der nach den vorliegenden Daten ungefähr bei 20000 anzusetzen ist. So nimmt die Entwicklung des gesamten Wortschatzes den typischen Verlauf vieler Wachstumsprozesse, wie er z.B. aus der Biologie oder der Bevölkerungsdynamik bereits seit langem bekannt ist.

3.1. Latein

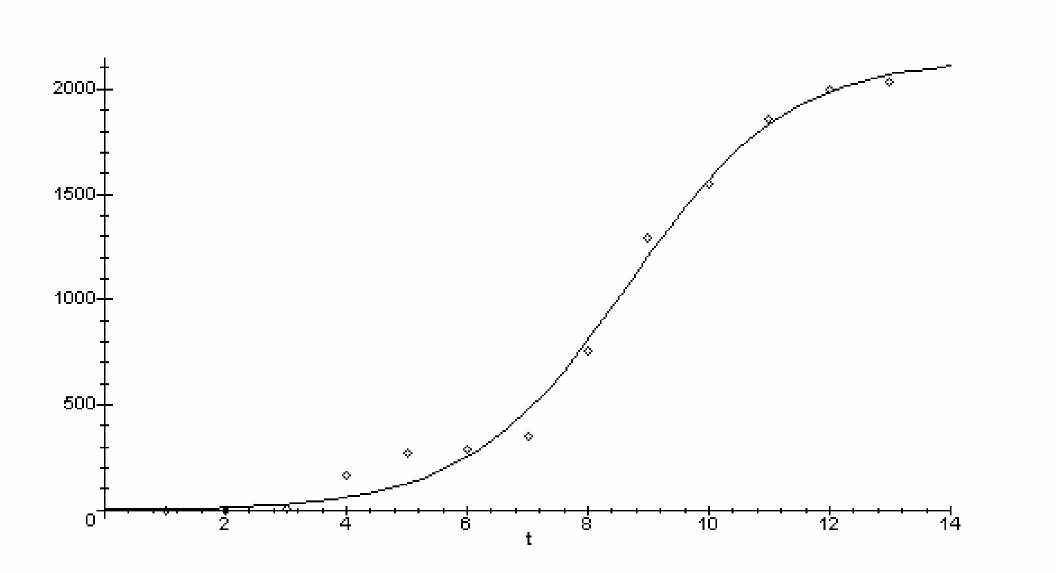
Anhand der Tabelle 4 lässt sich bereits feststellen, dass dieser Prozess der Übernahme lateinischer Wörter anscheinend nahezu abgeschlossen ist; die wenigen Neubildungen, die in neuerer Zeit noch hinzugekommen sind, können größtenteils unter dem Stichwort Internationalismen zusammengefasst werden:

Tabelle 4
Die Übernahme lateinischer Wörter ins Deutsche

Jhd.	***t***	***n***	***n* (kumuliert)**	***p* (ber.)**
8.	1	2	2	6.57
9.	2	1	3	13.92
10.	3	1	4	29.40
11.	4	264	268	61.57
12.	5	2	270	126.86
13.	6	20	290	252.95

[9] Bei den weiteren Graphiken gilt diese Legende ebenso wie die Beschriftung der Achsen, wobei die Zuordnung von *t* zu den Jahrhunderten jeweils aus den Tabellen abgelesen werden kann.

14.	7	57	347	474.91
15.	8	412	759	808.46
16.	9	533	1292	1207.09
17.	10	256	1548	1571.39
18.	11	309	1857	1831.26
19.	12	144	2001	1985.66
20.	13	30	2031	2067.63
a = 692.6991 b = 0.7546 c = 2146.2605 D = 0.99				



Graphik zu Tabelle 4. Die Übernahme lateinischer Wörter ins Deutsche

Der Kurvenverlauf für die Entlehnung lateinischer Wörter verdeutlicht, dass dieser Prozess weitestgehend abgeschlossen ist: Seit dem 18. Jahrhundert nähert sich die Kurve an die Asymptote des Grenzwerts *c* stark an. Die etwas größeren Abweichungen im Zeitraum vom 11. bis zum 15. Jahrhundert lassen sich hauptsächlich darauf zurückführen, dass die Datierung bei den meisten dieser Belege nur mit „in mittelhochdeutscher Zeit“ bzw. „in althochdeutscher Zeit“ angegeben wurde. Dabei spiegelt die Kurve den typischen Verlauf der Fremdwortübernahme in eine Sprache wieder, wie dies bereits u.a. von Best und Altmann (1986) und Best (2001, 2001a) gezeigt wurde. Das Testergebnis ist mit *D* = 0.99 sehr gut.

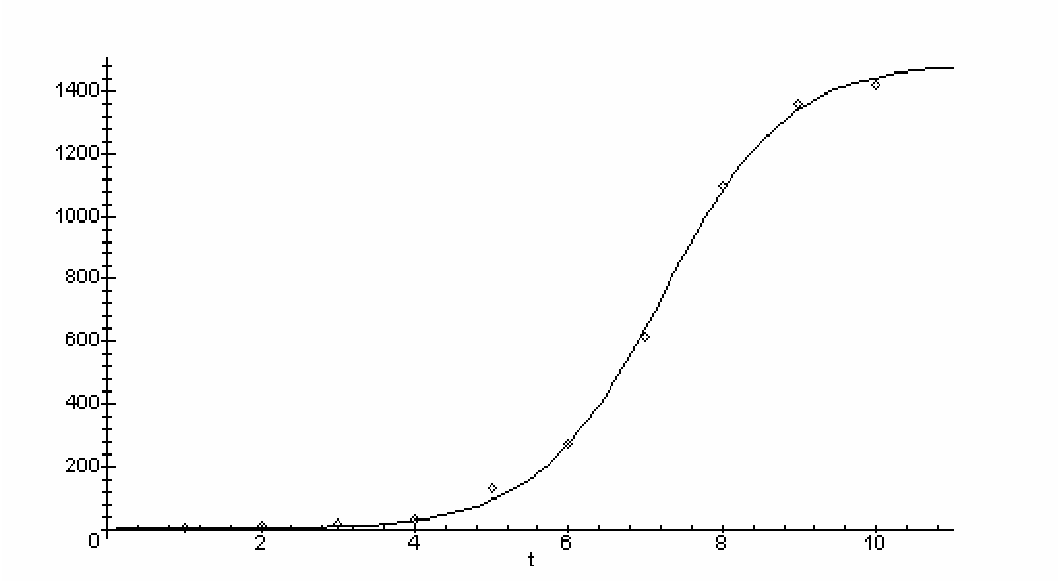
3.2. Französisch

Beim Französischen werden nach der Auszählung die Daten wie in Tabelle 5 ermittelt. Anhand dieser Tabelle fällt die sehr gute Übereinstimmung zwischen gemessenen und berechneten Werten auf, was sich auch in dem sehr guten Determinationskoeffizienten *D* = 0.99 (abgerundet) widerspiegelt.

Tabelle 5
Die Übernahme französischer Wörter ins Deutsche

Jhd.	***t***	***n***	***n* (kumuliert)**	***p* (ber.)**
11.	1	5	5	0.68
12.	2	5	10	2.34
13.	3	9	19	8.04
14.	4	11	30	27.36
15.	5	102	132	90.26
16.	6	139	271	270.95
17.	7	346	617	646.40
18.	8	482	1099	1081.02
19.	9	264	1363	1343.03
20.	10	60	1424	1444.61
$a = 7546.0427$ $b = 1.2374$ $c = 1490.6980$ $D = 0.99$				

Dadurch, dass *c* relativ nah an dem letzten gemessenen Wert liegt, lässt sich auch hier annehmen, dass der Prozess der Übernahme französischer Fremdwörter weitestgehend beendet ist. Bei Betrachtung der Graphik wird diese Übereinstimmung noch deutlicher:



Graphik zu Tabelle 5. Die Übernahme französischer Wörter ins Deutsche

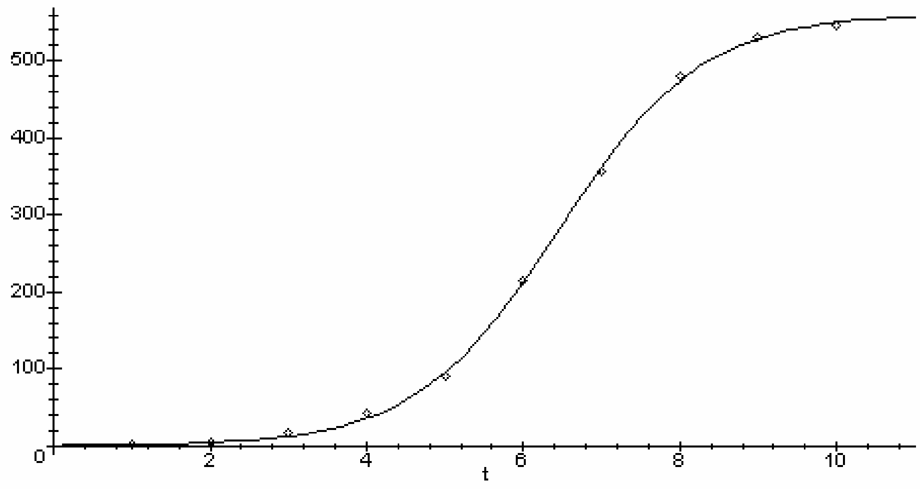
Die Abweichungen wegen der teilweise vagen Datierung fallen hier selbst im Zeitraum zwischen dem 11. und 15. Jahrhundert weniger ins Gewicht, als es im Lateinischen der Fall ist. Dies ist besonders dadurch bedingt, dass es in diesem Zeitraum weit weniger Übernahmen aus dem Französischen als aus dem Lateinischen gab. Ab dem 18. Jahrhundert verläuft die Kurve asymptotisch, was auf die Abgeschlossenheit dieses Übernahmeprozesses hindeutet. Auch im Falle der Übernahme französischer Fremdwörter ins Deutsche kann somit das logistische Gesetz bestätigt werden. Wenn die Entwicklung weiterhin in dieser Richtung stattfindet, ist nicht davon auszugehen, dass der angenommene Grenzwert deutlich überschritten wird.

3.3. Niederdeutsch

Nach Auswertung und Berechnung ergeben sich für das Niederdeutsche folgende Daten:

Tabelle 6
Die Übernahme niederdeutscher Wörter ins Deutsche

Jhd.	***t***	***n***	***n* (kumuliert)**	***p (ber.)***
11.	1	1	1	1.46
12.	2	3	4	4.19
13.	3	12	16	12.39
14.	4	26	42	35.57
15.	5	48	90	94.67
16.	6	125	215	212.27
17.	7	142	357	362.45
18.	8	123	480	474.41
19.	9	52	532	528.89
20.	10	13	545	549.96
a = 1193.7068 b = 1.0980 c = 561.1502 D = 0.99				



Graphik zu Tabelle 6. Die Übernahme niederdeutscher Wörter ins Deutsche

So kann auch für das Niederdeutsche mit D = 0.99 (abgerundet) eine hervorragende Anpassung des logistischen Gesetzes erreicht werden. Dabei besteht ein besonderes Problem darin, dass 17 nieder-deutsche Wörter aufgrund fehlender weiterführender Daten nur mit *Mittelniederdeutsch* zeitlich eingeordnet werden konnten und somit keine ausreichend genaue Datierung möglich war. Stellmacher (1990) datiert Mittelniederdeutsch auf das 13.-17. Jahrhundert; bei einer neuen Berechnung könnte man also diese 17 Wörter auf das 17. Jahrhundert datieren, da sie spätestens zu diesem Zeitpunkt in der Sprache existiert haben müssen.

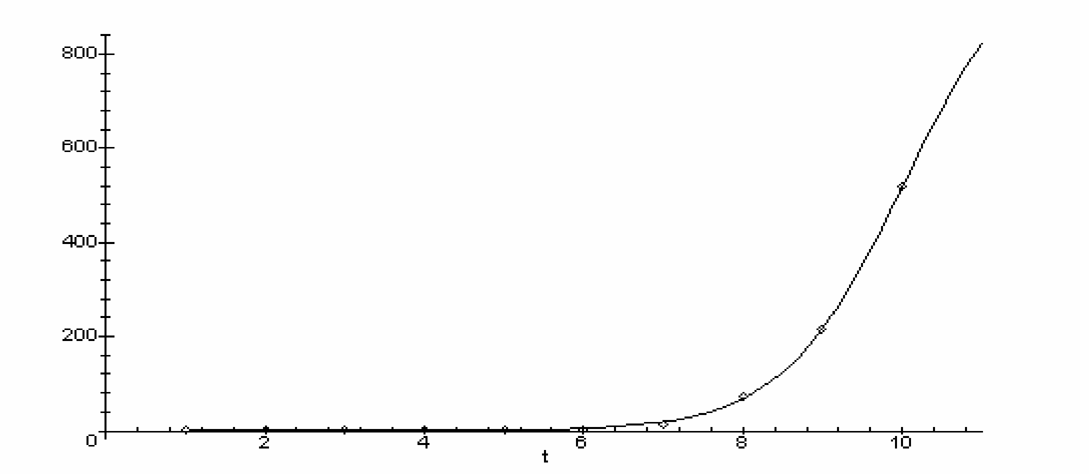
In diesem Fall liegen der berechnete Grenzwert und der gemessene Wert noch näher beieinander als dies im Französischen der Fall ist. Betrachtet man die Graphik, so stellt man auch hier eine asymptotische Annäherung fest. Diese Annäherung beginnt ungefähr mit dem 18. Jahrhundert; ab diesem Zeitpunkt verflacht die Kurve zunehmend. Dies verweist wie auch schon beim Französischen oder Lateinischen darauf, dass bei gleichbleibenden Umständen in der Zukunft vermutlich kaum noch Wörter aus dem Niederdeutschen ins Hochdeutsche übernommen werden.

3.4. Englisch

Die Übernahme englischer Fremdwörter ins Deutsche geschieht tabellarisch folgendermaßen:

Tabelle 7
Die Übernahme englischer Wörter ins Deutsche

Jhd.	***t***	***n***	***n* (kumuliert)**	***p* (ber.)**
11.	1	1	1	0.01
12.	2	0	1	0.03
13.	3	0	1	0.01
14.	4	0	1	0.37
15.	5	0	1	1.38
16.	6	2	3	5.17
17.	7	10	13	19.12
18.	8	60	73	68.29
19.	9	143	216	217.23
20.	10	303	519	518.86
a = 562.5224 b = 1.3222 c = 1047.4253 D = 0.99				



Graphik zu Tabelle 7. Die Übernahme englischer Wörter ins Deutsche

Die Anpassung des Modell ist mit D = 0.99 (abgerundet) wieder sehr gut. Wie schon beim Blick auf die Tabelle deutlich wird, liegt hier die besondere Schwierigkeit darin, dass ein englisches Wort sehr früh – in der Zeit des Althochdeutschen – übernommen worden ist, und danach für eine Zeitspanne von vier Jahrhunderten hinweg keine weiteren Entlehnungen dokumentiert sind. Aus der Tabelle wird deutlich, dass es sich hierbei um einen momentan noch nicht abgeschlossenen Sprachwandelprozess handelt, sondern im Gegenteil um einen höchst aktiven. Der Grenzwert *c* und der letzte gemessene Wert divergieren sehr stark, und an der Graphik ist noch keine mögliche Asymptote erkennbar.

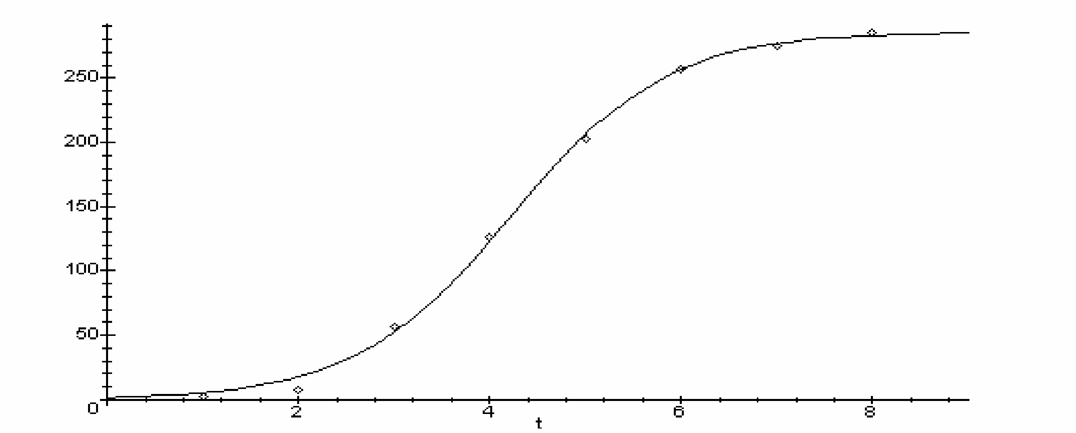
Die Übernahme englischer Fremdwörter befindet sich somit vermutlich gerade auf dem Höhepunkt, aber es ist noch nicht abzusehen, ob die Übernahme nicht weiter ansteigt (was durch Faktoren wie Globalisierung oder technische Neuerungen, die gegenwärtig meist eher aus dem amerikanischen Raum kommen, noch begünstigt wird). Im Gegensatz zu den Entlehnungen aus anderen Sprachen ist eine Abflachung der Kurve beim Englischen nicht sichtbar, was wiederum dagegen spricht, dass der Höhepunkt bereits überschritten ist. Natürlich ist es aber auch möglich, dass der Wendepunkt der Entwicklung bereits erreicht ist, dies jedoch noch nicht ersichtlich ist.

3.5. Italienisch

Die Besonderheit beim Italienischen gegenüber den meisten bislang dargestellten Sprachen und den dazugehörigen Übernahmeprozessen liegt darin, dass die Entlehnungen erst relativ spät beginnen, wie die folgende Tabelle zeigt:

Tabelle 8
Die Übernahme italienischer Wörter ins Deutsche

Jhd.	*t*	*n*	*n* (kumuliert)	*p* (ber.)
13.	1	2	2	5.38
14.	2	6	8	17.50
15.	3	48	56	52.00
16.	4	70	126	123.27
17.	5	77	203	206.18
18.	6	54	257	256.86
19.	7	18	275	276.83
20.	8	11	286	283.29
a = 178.2562 b = 1.2263 c = 286.0662 D = 0.99				



Graphik zu Tabelle 8. Die Übernahme italienischer Wörter ins Deutsche

Das Testergebnis ist wieder mit D = 0.99 (abgerundet) sehr gut. Der Einfluss des Italienischen erstarkt somit später als bei den bisher dargestellten Sprachen mit Ausnahme des Englischen. Nach den Werten in der Tabelle und in der Graphik scheint es auch hier so, dass der Prozess schon fast zum Erliegen gekommen ist, wie die deutliche Abnahme von n zeigt. Auffallend ist hierbei weiterhin, dass der letzte gemessene Wert für n (kumuliert) und der Grenzwert c sehr genau übereinstimmen, was diese Hypothese stützt. Die Tendenz ist hier (gleichbleibende Umstände vorausgesetzt) eindeutig: Ab dem 17. Jahrhundert beginnt schon die Annäherung an den Grenzwert c, die sich seit dem 19. Jahrhundert kaum noch verändert. Es ist also in Zukunft nicht zu erwarten, dass noch viele Wörter aus dem Italienischen ins Deutsche übernommen werden.

3.6. Griechisch

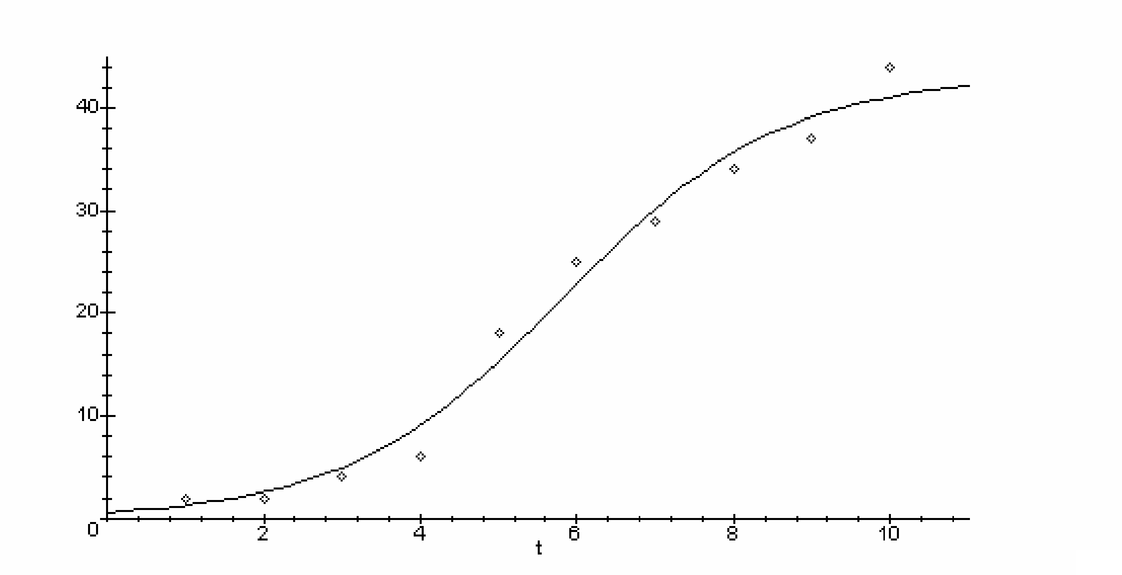
In neuerer Zeit spielt das Altgriechische im Bereich von Wissenschaft und Technik erneut bei der Wortbildung eine gewisse Rolle (die sogenannten Internationalismen). Letztendlich ergeben sich für das Griechische die Daten wie folgt:

Tabelle 9
Die Übernahme griechischer Wörter ins Deutsche

Jhd.	***t***	***n***	***n* (kumuliert)**	***p* (ber.)**
11.	1	3	3	0.08
12.	2	0	3	0.24
13.	3	0	3	0.75
14.	4	1	4	2.33
15.	5	4	8	7.07
16.	6	19	27	20.19
17.	7	16	43	49.65
18.	8	50	93	93.11
19.	9	42	135	129.32
20.	10	9	144	147.69
a = 6381.5058 b = 1.1397 c = 158.2711 D = 0.99				

Die Zahlen der Tabelle lassen auch hier wiederum vermuten, dass der Sprachwandelprozess zum Stillstand gekommen ist; so liegen der Grenzwert c und der letzte gemessene Wert nicht weit auseinander. Die Graphik zeigt diesen Zusammenhang zwischen berechneten und beobachteten Werten.

Die Abflachung der Kurve setzt später ein als beispielsweise beim Lateinischen; erst ab dem 19. Jahrhundert scheint diese Annäherung stattzufinden. Dennoch kann hier wiederum davon ausgegangen werden, dass bei gleichbleibenden Umständen keine verstärkte Übernahme griechischer Fremdwörter mehr erfolgen wird.



Graphik zu Tabelle 9. Die Übernahme griechischer Wörter ins Deutsche

3.7. Niederländisch

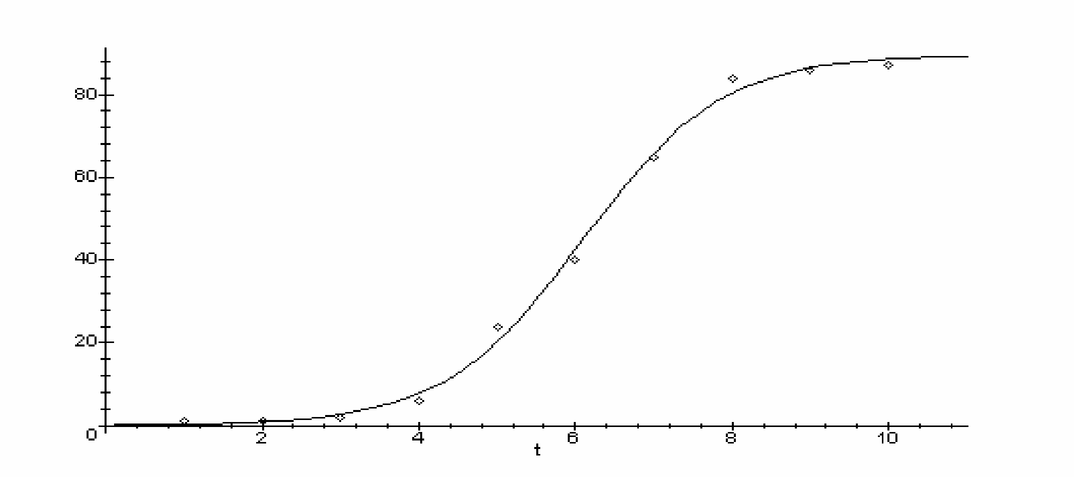
Das Niederländische war relativ häufig die Vermittlersprache für Entlehnungen aus dem Romanischen – insbesondere für Entlehnungen aus dem Französischen. Für die niederländischen Entlehnungen kommen nach der Auszählung des Dudens und der anschließenden Berechnung folgende Werte zustande:

Tabelle 10
Die Übernahme niederländischer Wörter ins Deutsche

Jhd.	***t***	***n***	***n* (kumuliert)**	***p* (ber.)**
11.	1	1	1	0.28
12.	2	0	1	0.87
13.	3	1	2	2.64
14.	4	4	6	7.70
15.	5	18	24	20.21
16.	6	16	40	42.49
17.	7	25	65	66.01
18.	8	19	84	80.37
19.	9	2	86	86.45
20.	10	1	87	88.61
$a = 978.5187$ $b = 1.1302$ $c = 89.6817$ $D = 0.99$				

Daraus resultiert der Determinationskoeffizient von D = 0.99 (abgerundet), der wiederum eine sehr gute Anpassung des logistischen Gesetzes bestätigt.

Auch hier ist wieder der typische Verlauf eines unvollständigen Sprachwandelprozesses zu beobachten: Ab dem 18. Jahrhundert ist die Asymptote ersichtlich, der sich die Kurve nähert. Daraus lässt sich wiederum der Schluss ziehen, dass dieser Prozess weitestgehend beendet ist.



Graphik zu Tabelle 10: Die Übernahme niederländischer Wörter ins Deutsche

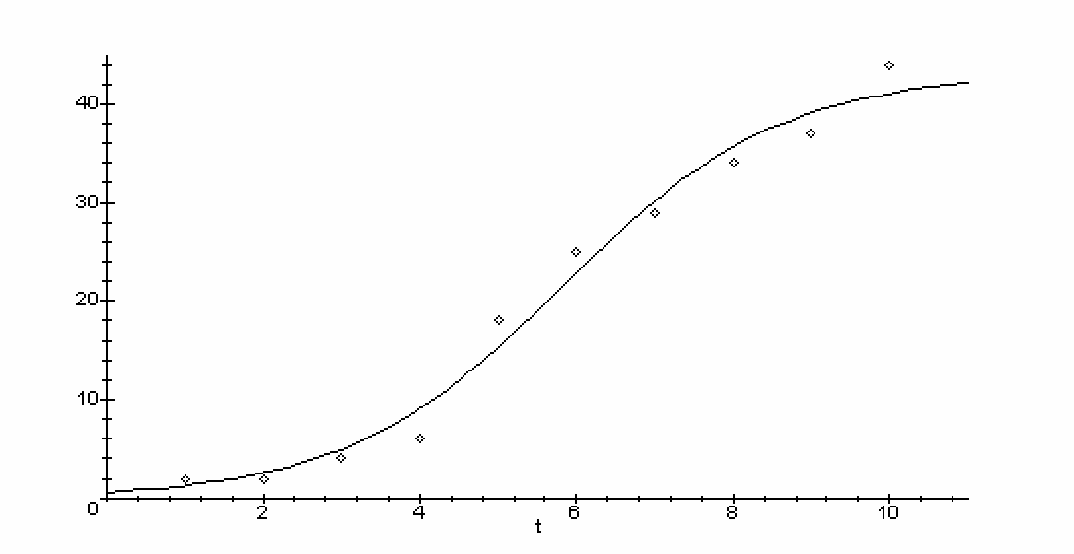
3.8. Slawisch

Wie schon erwähnt, war es in diesem Falle nötig, sämtliche slawischen Sprachen, aus denen Wörter entlehnt wurden, unter dem Sammelbegriff *Slawisch* zusammenzufassen, um eine Datenauswertung zu ermöglichen. Andernfalls wären für jede einzelne dieser Sprachen zu wenig Belege vorhanden gewesen.

Tabelle 11
Die Übernahme slawischer Wörter ins Deutsche

Jhd.	***t***	***n***	***n* (kumuliert)**	***p* (ber.)**
11.	1	2	2	1.28
12.	2	0	2	2.56
13.	3	2	4	4.96
14.	4	2	6	9.09
15.	5	12	18	15.29
16.	6	7	25	22.88
17.	7	4	29	30.15
18.	8	5	34	35.65
19.	9	3	37	39.13
20.	10	7	44	41.07
$a = 67.0599$ $b = 0.7216$ $c = 43.0918$ $D = 0.98$				

Obwohl mehrere Sprachen zusammengefasst sind, ergibt sich auch hier eine einheitliche Tendenz, die auf den Abschluss dieses Prozesses verweist.



Graphik zu Tabelle 11. Die Übernahme slawischer Wörter ins Deutsche

Trotz des relativ geringen Datensatzes fällt die Anpassung sehr gut aus, wenn auch nicht ganz so gut wie beim Niederdeutschen oder Englischen. Dies kann an der geringen Größe des Datensatzes liegen und/oder daran, dass mehrere Sprachen gebündelt wurden. Der Trend jedoch ist letztlich eindeutig, und es kann wohl davon ausgegangen werden, dass sich diese Tendenz auch in den einzelnen slawischen Sprachen wiederfinden lässt – unter der Voraussetzung, dass anhand eines anderen Korpus ein größerer Datensatz erstellt werden kann.

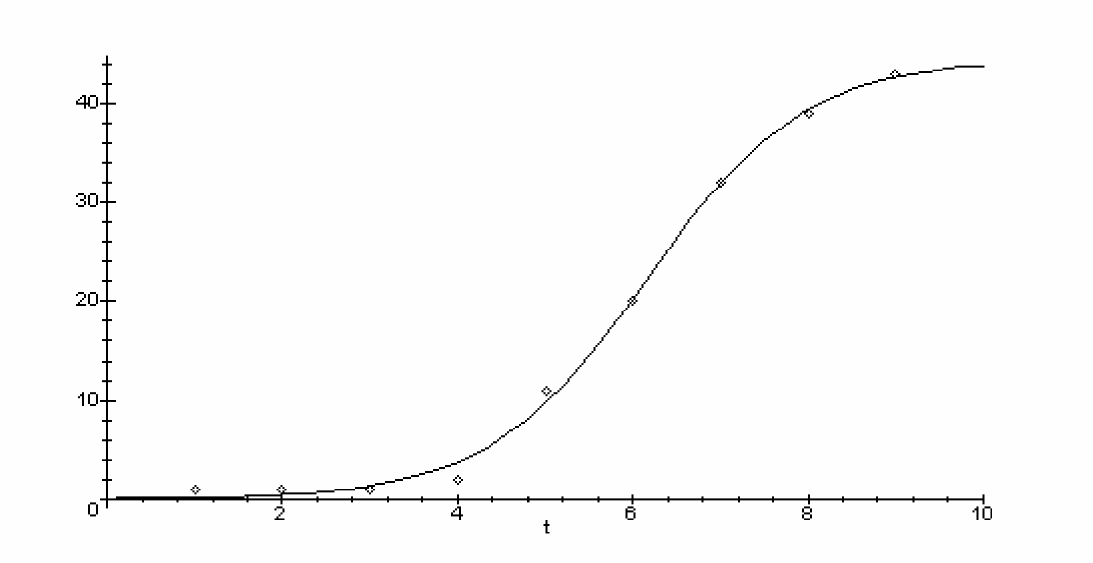
In einer Untersuchung, die speziell den slawischen Entlehnungen gewidmet war, konnte Best (2003a) für slawische Wörter insgesamt und russische speziell ähnliche Entwicklungen nachweisen. Auffallend ist bei den Slawismen, dass der Beobachtungswert für das 20. Jahrhundert immer über dem berechneten Trend liegt.

3.9. Spanisch

Für das Spanische gilt, dass diese Sprache weitaus häufiger Herkunftssprache als Vermittlersprache ist. Daraus resultiert ein relativ kleiner Datensatz.

Tabelle 12
Die Übernahme spanischer Wörter ins Deutsche

Jhd.	***t***	***n***	***n* (kumuliert)**	***p* (ber.)**
12.	1	1	1	0.15
13.	2	0	1	0.44
14.	3	0	1	1.31
15.	4	1	2	3.74
16.	5	9	11	9.66
17.	6	9	20	20.31
18.	7	12	32	31.94
19.	8	7	39	39.39
20.	9	4	43	42.68
a = 915.6131 b = 1.1074 c = 44.5178 D = 0.99				



Graphik zu Tabelle 12. Die Übernahme spanischer Wörter ins Deutsche

Auch hier liegt mit einem Determinationskoeffizienten von *D* = 0.9974 (abgerundet) wiederum eine sehr gute Anpassung des logistischen Gesetzes vor. Dies spiegelt sich auch in der Graphik wider.

Es lässt sich erneut der typische Verlauf beobachten; ab dem 18. Jahrhundert folgt die Kurve langsam dem Grenzwert *c*. Auch in diesem Fall ist es unwahrscheinlich, dass noch viele Wörter aus dieser Sprache Eingang ins Deutsche finden.

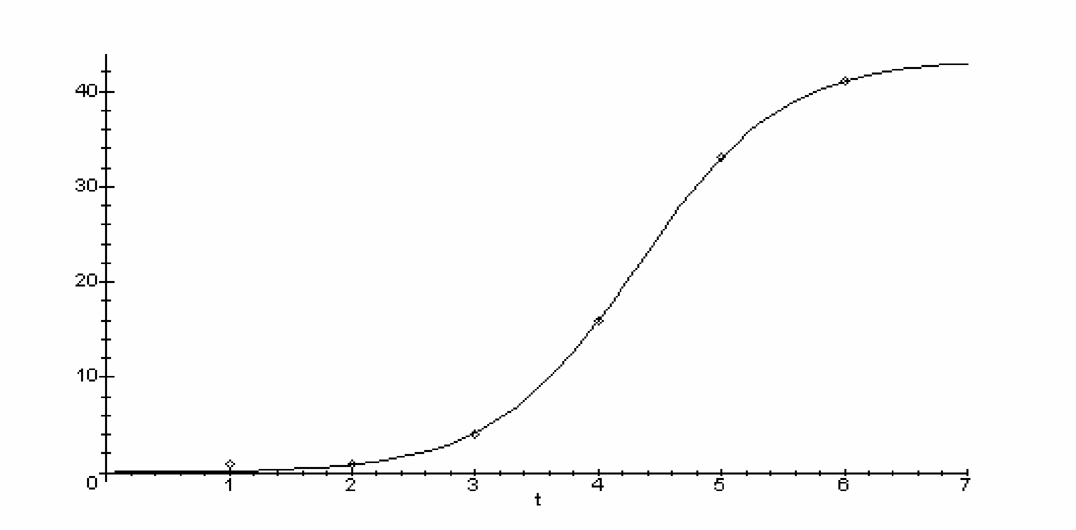
3.10. Rotwelsch

Beim Rotwelschen war bis zu diesem Test nicht klar, ob die Datenmenge überhaupt zur Überprüfung des logistischen Gesetzes ausreicht. In Best & Altmann (1986) war die Anpassung des Modells mangels hinreichender Daten noch misslungen.

Tabelle 13
Die Übernahme rotwelscher Wörter ins Deutsche

Jhd.	***t***	***n***	***n* (kumuliert)**	***p* (ber.)**
15.	1	1	1	0.15
16.	2	0	1	0.83
17.	3	3	4	4.18
18.	4	12	16	15.96
19.	5	17	33	33.00
20.	6	8	41	41.01
a = 1530.4863 *b* = 1.6982 *c* = 43.3640 *D* = 0.99				

Die Anpassung ist mit *D* = 0.99 (abgerundet) sehr gut. Der früheste Übernahmezeitpunkt liegt hier schon relativ spät; meistens wurden die ersten Wörter aus anderen Sprachen bereits zur Zeit des Althochdeutschen, d.h. spätestens im 11. Jahrhundert entlehnt. Obwohl der ge-



Graphik zu Tabelle 13. Die Übernahme rotwelscher Wörter ins Deutsche

samte Zeitraum, in dem Wörter aus dem Rotwelschen übernommen wurden, klein ist, ist es trotz der niedrigen Anzahl der Belege möglich, die Anpassung durchzuführen, die entgegen den Erwartungen sogar extrem gut gelungen ist.

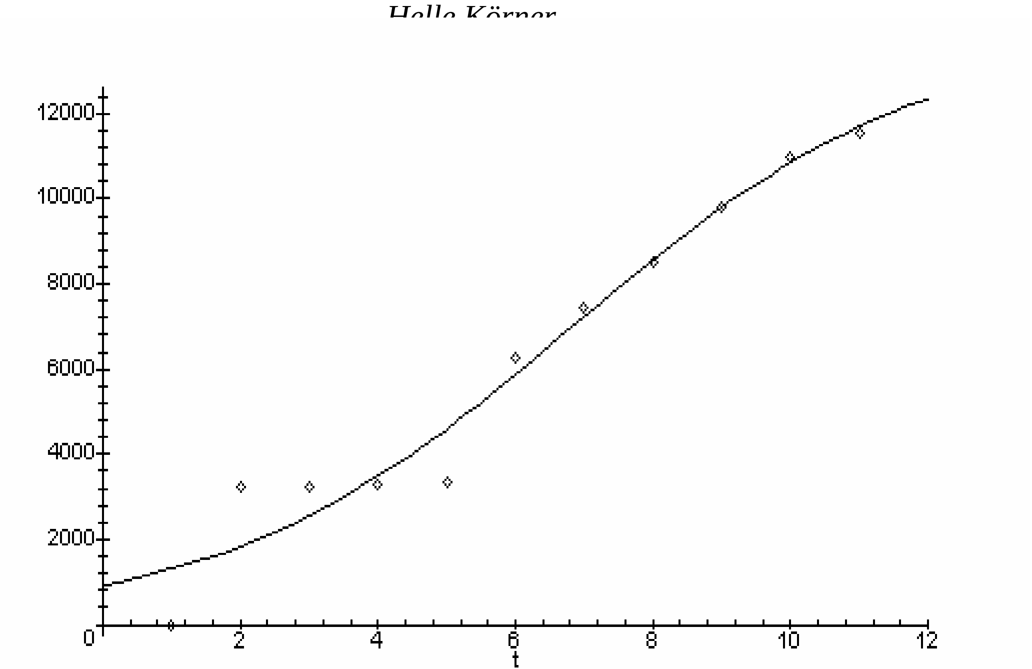
Die Kurve beginnt ab dem 19. Jahrhundert sich der Asymptote anzunähern; ein Hinweis darauf, dass sich auch dieser Sprachwandel bereits in seinem ‘Endstadium’ befindet.

3.11. Deutsch

Im Gegensatz zu anderen Arbeiten, die sich hauptsächlich mit der Betrachtung der im Deutschen vorhandenen Fremdwörter beschäftigen, soll in dieser Arbeit auch der Zuwachs von deutschen Wörtern betrachtet werden. Insofern bietet diese Studie das erste Mal auf der Grundlage des Dudens Erkenntnisse über die Entwicklung von Wörtern in der deutschen Sprache und damit für das Entstehen des deutschen Gesamtwortschatzes. Für die im Deutschen gebildeten Wörter ergibt sich nach der Datenauswertung folgendes Bild:

Tabelle 14
Zuwachs einheimischer deutscher Wörter

Jhd.	***t***	***n***	***n* (kumuliert)**	***p* (ber.)**
10.	1	2	2	1300.96
11.	2	3234	3236	1840.17
12.	3	12	3248	2557.87
13.	4	22	3270	3475.48
14.	5	92	3362	4590.21
15.	6	2924	6286	5863.29
16.	7	1141	7427	7218.59
17.	8	1085	8512	8557.69
18.	9	1299	9811	9786.70
19.	10	1192	11003	10840.63
20.	11	534	11537	11693.07
a = 14.4195 b = 0.3902 c = 14000 D = 0.96				



Graphik zu Tabelle 14. Der Zuwachs einheimischer deutscher Wörter

Die Häufungen im 11. und 15. Jahrhundert sind auf die bereits erwähnten Schwierigkeiten bezüglich der Einordnung des Althochdeutschen bzw. Mittelhochdeutschen zurückzuführen.

Bei angenommenen c = 14000 erhält man mit D = 0.96 eine sehr gute Anpassung des logistischen Gesetzes. In der Graphik zur Tabelle wird wiederum deutlich, dass die Abweichungen zwischen den gemessenen und den berechneten Daten im Zeitraum vom 10. bis zum 14. Jahrhundert relativ groß sind.

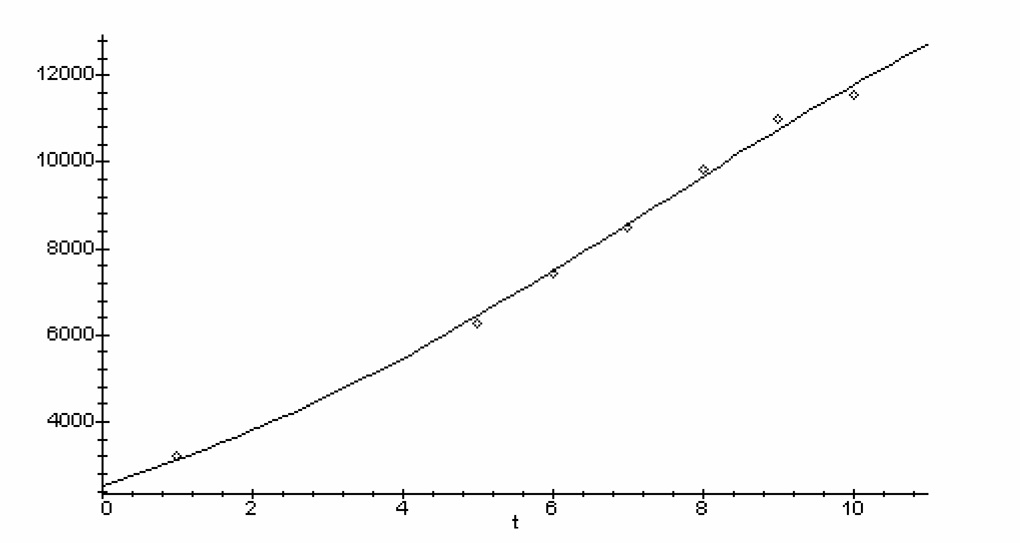
So deutet sich etwa ab dem 18. Jahrhundert eine Verflachung der Kurve an. Im Vergleich zu nahezu abgeschlossenen Sprachwandeln (wie sie u.a. für Latein oder Französisch gezeigt wurden) liegt hier jedoch eine relativ ‘schwache’, d.h. undeutlich erkennbare Asymptote vor. Was sich an diesen Daten allerdings eindeutig ablesen lässt, ist (bislang jedenfalls) der Trend zu einer Vergrößerung des Wortschatzes.

Eine bessere Anpassung kann durch eine Neuinterpretation der vorliegenden Daten erreicht werden. Dafür wurde nur das 11. Jahrhundert für alle Wörter, die mit der Angabe *Althochdeutsch* versehen waren, berücksichtigt, und nur das 15. Jahrhundert für die Wörter, die zur Zeit des *Mittelhochdeutschen* gebildet wurden. Für die späteren Jahrhunderte war überwiegend eine eindeutige Zuordnung möglich; also wurden die weiteren Daten übernommen. Daraus ergeben sich folgende Werte:

Tabelle 15
Der Zuwachs einheimischer deutscher Wörter (bereinigt)

Jhd.	***t***	***n***	***n* (kumuliert)**	***p* (ber.)**
11.	1	3236	3236	3126.75
15.	5	3050	6286	6444.06
16.	6	1141	7427	7483.88
17.	7	1085	8512	8567.09
18.	8	1299	9811	9663.02
19.	9	1192	11003	10739.53
20.	10	534	11537	11766.70
a = 6.0706 b = 0.2439 c = 18000 D = 0.99				

Während bei der ersten Anpassung für die im deutschen gebildeten Wörter der Determinationskoeffizient $D = 0.96$ betrug, kann man ihn durch die Neuinterpretation der Daten auf $D = 0.99$ (abgerundet) steigern. Diese Verbesserung spiegelt sich auch in der folgenden Graphik wieder:



Graphik zu Tabelle 15. Der Zuwachs einheimischer deutscher Wörter (bereinigt)

Der Verlauf ist für beide Graphiken im Wesentlichen derselbe: Es deutet sich in beiden Fällen eine spätere asymptotische Annäherung an. Ebenfalls gemeinsam ist beiden Graphiken, dass sie deutlich den Trend zur Vergrößerung des Wortschatzes anzeigen.

4. Diskussion der vorliegenden Daten

Das logistische Gesetz in der unvollständigen Form konnte mit sehr gutem Ergebnis an jeden einzelnen Datensatz angepasst werden. Dies entspricht den theoretischen Annahmen zu den diversen Sprachwandelprozessen. Problematisch ist dabei die Betrachtung des Grenzwertes *c*. Auch Best und Altmann (1986: 38) sprechen dies an: „Die Werte (...) sind natürlich nicht als absolute Obergrenze zu interpretieren, sondern lediglich als Obergrenze für das vorliegende Datenmaterial. Eine Hochrechnung auf die tatsächlichen Werte setzt eine Einschätzung der Repräsentativität der Angaben des Etymologie-Duden für den Gesamtwortschatz voraus." Zur Schwierigkeit der Interpretation des Grenzwertes *c* sagt Best (2001c: 14) wieterhin: „Es spricht daher tatsächlich alles dagegen, die Schätzwerte für *c* als genaue Werte für den Zuwachs zu verstehen. Sie sind rechnerische Größen, die sich ergeben, wenn man untersucht, ob die Formel für den unvollständigen Sprachwandel ein geeignetes Modell für die jeweilige Da Datenbasis darstellt. Wenn *c* interpretiert werden soll, so immer nur bezogen auf die Wörterbücher, die die Daten für den Entlehnungsprozess geliefert haben. Ein Schluss auf das Lexikon der Sprache insgesamt ist nur denkbar, wenn man berücksichtigt, dass jedes Wörterbuch einen unterschiedlichen Ausschnitt aus dem Vokabular der Sprache darbietet und wenn man diesem Wörterbuch eine gewisse Repräsentativität für die Sprache zubilligen kann."

Auch wenn diese Studie nicht als repräsentativ gelten kann, so kann man doch annehmen, dass sie die Trends hinsichtlich der Entwicklung des Wortschatzes einigermaßen zutreffend wiedergibt. Besonders das Verhältnis der einzelnen Vermittlersprachen zueinander scheint in diesem Ausschnittsbereich des deutschen Wortschatzes annähernd richtig abgebildet zu sein:

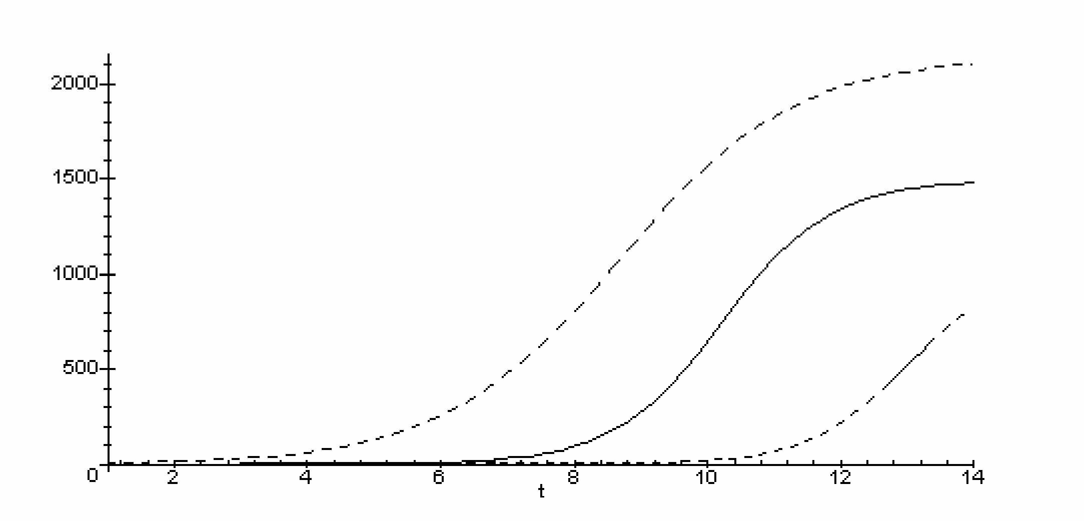
Die wichtigste Gebersprache ist Latein, gefolgt von Französisch. Das Englische scheint erst im 20. Jahrhundert zur „wichtigsten“ Gebersprache geworden zu sein („wichtig“ ist hier bezogen auf die Entwicklungsdynamik).

4.1. Entlehnungen aus dem anglo-amerikanischen Sprachraum

Zur Verdeutlichung des aktuellen Diskussionsstandes in Bezug auf die Anglizismen soll folgendes Zitat einer Internetseite zum Thema „Reinerhaltung der deutschen Sprache“ dienen: „Über die Disziplinlosigkeit im Sprachgebrauch darf man nicht mit dem Sofaargument hinwegsehen, ‘dass es sich hier um eine Modeerscheinung handele, die von selbst wieder verschwinde’. Nein, der Bestand der deutschen Sprache – gewiß, ein politisch vollkommen inkorrekter Ausdruck, über den sich Antifaschisten und andere Gutmenschen wieder echauffieren werden –, ist akut gefährdet“(nordbruch.tripod.com/artikel/aShutup.html Stand vom 23.10.02). Generell lässt sich sagen: „In der öffentlichen Diskussion ist zur Zeit wieder von Überschwemmung, Überflutung, Verwässerung der deutschen Sprache durch Anglizismen die Rede“ (Stickel, 2001a: 3).

Die Anzahl der im Deutschen benutzten englischen Wörter ist natürlich größer als die in den Duden aufgenommene, wie auch das Anglizismen-Wörterbuch von Carstensen & Busse zeigt. Meist handelt es sich dabei allerdings um Modewörter, die recht schnell wieder aus der Sprache verschwinden (man denke nur an Entlehnungen wie *Boots*, *Kids* etc.). Auf der anderen Seite muss berücksichtigt werden, dass eine ‘Umstrukturierung’ des Wortschatzes meist erst mit einer gewissen Verspätung erfolgt, da mündlich durchaus schon gebräuchliche Formen erst in die Schriftsprache (und damit in die Wörterbücher) Eingang finden müssen; so kann insgesamt wohl von einer etwas höheren Anzahl von Anglizismen als in den derzeitigen Wörterbüchern verzeichnet ausgegangen werden. Insgesamt betrachtet machen die Anglizismen aber einen eher geringen Anteil aus, wie Müller-Hasemann anhand seiner Studie zum Anteil englischer Fremdwörter im SPIEGEL und im Quelle-Katalog (1983: 159) zeigt: „Den Pessimisten, die befürchten, dass der Anteil an Fremdelementen in der deutschen Sprache über alle Grenzen wachsen wird, ist nicht Recht zu geben. Dies macht der niedrige Wert der Konstanten *c* deutlich (*SPIEGEL: c* = 0.012153*; Quelle-Katalog: c* = 0.040458 – *H.K.*), durch die der Grenzwert gegeben ist, dem sich der Anteil englischer Fremdelemente asymptotisch annähert.“ In Prozentzahlen ausgedrückt wird dies noch deutlicher: Beim SPIEGEL beträgt der Anteil der englischen Wörter an allen Wörtern 1,2% (Müller-Hasemann, 1983: 155). Beim Quelle-Katalog liegt dieser Anteil ein wenig höher mit 4,2% (Ebd.: 158).

Bei Betrachtung der Graphik, die das Lateinische, das Französische und das Englische nebeneinander stellt (jeweils die berechneten Werte von t_1 = 8. Jhd. bis t_{13} = 20. Jhd. sind in der Graphik berücksichtigt worden), ist die Gemeinsamkeit dieser Sprachwandel nicht zu übersehen. Bislang zeigen sich keine Tendenzen, die darauf hinweisen, dass dieser Sprachwandel beim Englischen anders verläuft als die beiden nahezu abgeschlossenen Entlehnungsprozesse.



Übernahme lateinischer, französischer und englischer Wörter ins Deutsche

Die obere gestrichelte Linie gibt auf der Basis von Duden (2001) die Entwicklung der Übernahmen aus dem Lateinischen wieder, die durchgehende steht für das Französische und die untere gestrichelte Linie zeigt diese Entwicklung für Übernahmen aus dem Englischen.

Anhand dieser vergleichenden Graphik lässt sich erkennen, dass es um so leichter ist, eine Prognose über den weiteren Verlauf eines Sprachwandels zu erstellen, je mehr Daten aus unterschiedlichen Jahrhunderten vorliegen. So können für das Lateinische und das Französische recht eindeutige Schlussfolgerungen bezüglich der weiteren Entwicklung des Sprachwandels gezogen werden, wohingegen beim Englischen die Datenlage lediglich Vermutungen über den weiteren Verlauf zulässt. So scheint es, dass die Übernahme englischer Fremdwörter zu einer ähnlichen ‘Welle’ gehört wie vormals die Übernahme der französischen Fremdwörter, was zu der damaligen Zeit ähnlich scharf kritisiert wurde wie heutzutage die Entlehnungen aus dem englischsprachigen Bereich (vgl.: Stickel, 2001a: 3).

4.2. Ausblick

Wie anhand dieser Daten deutlich geworden ist, gewinnt man bei der Analyse des deutschen Wortschatzes somit nicht nur einen rein historischen Überblick, sondern es ist auch möglich, über potentielle Weiterentwicklungen einzelner Entlehnungsprozesse einen Ausblick zu geben. Im Vergleich mit anderen Studien (vgl. z.B. Best 2001, Best 2001a, Kirkness 1988) konnten zudem die Trends dieser Untersuchung bestätigt werden: Das Lateinische ist die wichtigste Gebersprache, gefolgt vom Französischen und (mittlerweile) dem Englischen. Da den Studien unterschiedliche Wörterbücher zugrunde lagen, ist davon auszugehen, dass diese Trends auch tatsächlich für den deutschen Gesamtwortschatz repräsentativ sind.

Eine weitere Überprüfung des logistischen Gesetzes könnte durch das neueste, 2003 erschienene etymologische Wörterbuch Friedrich Kluges erfolgen – aber auch in diesem Falle ist davon auszugehen, dass die allgemeinen Trends bestätigt werden können.

Literaturverzeichnis

Altmann, Gabriel (1983): Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, Karl-Heinz, & Kohlhase, Jörg (Hrsg.): *Exakte Sprachwandelforschung*. Göttingen:

edition herodot. S. 54-90.

Best, Karl-Heinz (2001): Ein Beitrag zur Fremdwortdiskussion. In: Schierholz, Stefan J. (Hrsg.): *Die deutsche Sprache in der Gegenwart. Festschrift für Dieter Cherubim zum 60. Geburtstag*. Frankfurt am Main; Berlin; Bern; Bruxelles; New York; Oxford; Wien: Lang. S. 263-270.

Best, Karl-Heinz (2001a): Wo kommen die deutschen Fremdwörter her? *Göttinger Beiträge zur Sprachwissenschaft 5*: 7-20.

Best, Karl-Heinz (2001b): *Quantitative Linguistik. Eine Annäherung.* Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (2001c): Der Zuwachs der Wörter auf *–ical* im Deutschen. In: *Glottometrics 2*: 11-16.

Best, Karl-Heinz (2003): Anglizismen – quantitativ. *Göttinger Beiträge zur Sprachwissenschaft 8*: 7-23.

Best, Karl-Heinz (2003a): Slawische Entlehnungen im Deutschen. In: Kempgen, Sebastian (Hrsg.): *Rusistika – Slavistika – Lingvistika. Festschrift für Werner Lehfeldt zum 60. Geburtstag.* München: Sagner. S. 464-473.

Best, Karl-Heinz, & Altmann, Gabriel (1986): Untersuchungen zur Gesetzmäßigkeit von Entlehnungsprozessen im Deutschen. *Folia Linguistica Historica 7*: 31-41.

Best, Karl-Heinz, & Kohlhase, Jörg (Hrsg.) (1983): *Exakte Sprachwandelforschung. Theoretische Beiträge, statistische Analysen und Arbeitsberichte.* Göttingen: edition herodot.

Carstensen, Broder, & Busse, Ulrich (1996): *Anglizismen-Wörterbuch. Der Einfluß des Englischen nach 1945.* Berlin u.a.: de Gruyter.

Duden. *Das Herkunftswörterbuch* (2001). Mannheim; Leipzig; Wien; Zürich: Bibliographisches Institut – Dudenverlag.

Imsiepen, Ulrike (1983): Die *e*-Epithese bei starken Verben im Deutschen. In: Best, Karl-Heinz, & Kohlhase, Jörg (Hrsg.): *Exakte Sprachwandelforschung*. Göttingen: edition herodot. S. 119-142.

Kirkness, **Alan** (Hrsg.) (1988): *Deutsches Fremdwörterbuch* (1913-1988). Begründet v. Hans Schulz, fortgeführt v. Otto Basler, weitergeführt im Institut für deutsche Sprache. Bd.7: Quellenverzeichnis, Wortregister, Nachwort. Berlin; New York: de Gruyter.

Kluge, Friedrich (1999, 23. Auflage): *Etymologisches Wörterbuch der deutschen Sprache.* Berlin; New York: de Gruyter.

Müller-Hasemann, Wolfgang (1983): Das Eindringen englischer Wörter ins Deutsche ab 1945. In: Best, Karl-Heinz & Kohlhase, Jörg (Hrsg.): *Exakte Sprachwandelforschung.* Göttingen: edition herodot. S. 143-160.

Osgood, Charles E., & Sebeok, Thomas A. (1954) (ed.): *Psycholinguistics. A survey of theory and research problems.* Baltimore: Waverly Press Inc.

Pfeifer, Wolfgang (2000, 5. Auflage): *Etymologisches Wörterbuch des Deutschen.* München: Deutscher Taschenbuch Verlag.

Piotrowski, R., Bektaev, K. & Piotrowskaja, A. (1985): *Mathematische Linguistik.* Bochum: Brockmeyer.

Stickel, Gerhard (Hrsg.) (2001): *Neues und Fremdes im deutschen Wortschatz. Aktueller lexikalischer Wandel.* Berlin; New York: de Gruyter.

Internetquellen:

www.deutsche-sprachwelt.de/berichte/gesetz/dey03.shtml Stand vom 14.10.2002.
nordbruch.tripod.com/artikel/aShutup.html Stand vom 23.10.02.

Software:

Altmann-Fitter (1997). *Iterative Fitting of Probability Distributions*. Lüdenscheid: RAM-Verlag.

MAPLE V Release 4 (1996). Berlin u.a.: Springer.

NLREG. *Nonlinear Regression Analysis Program*. Ph.H. Sherrod. Copyright © 1991 – 2001.

The Lerchianness plot

Osama Abdelaziz Hussien[1]

Abstract. Hoaglin (1980) introduced the Poissonness plot to detect departures of data from a hypothesized Poisson model. Hoaglin and Tukey (1985) extended the use of the Poissonness plot to other one parameter discrete distributions include geometric and logarithmic distributions. On the other hand, in many applications the Zipf plot is used to verify that the data obeys the Zipf's law. We present a unified presentation of these plots and extend its use for 3-parameter families. In particular, the Lerch family of discrete distributions, which includes as special cases the Zipf, Zipf-Mandelbrot, the logarithmic and the polylogarithmic distributions. A comparison with other types of plots for discrete distributions shows the resistance and power of this plot, the Lerchianness plot. We apply this plot to some of the Hoaglin and Tukey data sets and show it gives better fits. In addition, an application to subject and letter frequencies and to Egyptian city sizes has been presented.

Key words: EDF, Chi-square goodness-of-fit tests, Lerch transcendent, polylogarithm, frequency ratio plot, Poissonness plot, Poisson, binomial, geometric, logarithmic, power series, polylogarithmic and Lerch distributions, Zipf's law.

1. Introduction

Graphical techniques are generally used in conjunction with numerical techniques to assess the fit of a hypothesized probability model. In particular, probability plots provide effective ways to display departures from the proposed model. The probability plots (like the Probability-Probability (P-P) plots and the Quantile-Quantile (Q-Q) plots) assume the underlying continuous distribution is a location-scale family. No members of the Lerch family are location-scale, even many members are one parameter distribution. Hoaglin (1980) introduces the Poissonness plot to detect departures of data from a hypothesized Poisson model. Hoaglin and Tukey (1985) extend the use of this plot to other one parameter discrete distributions include Poisson, geometric and logarithmic distributions. On the other hand, many man made and naturally occurring phenomena, including city sizes, internet traffic, firm sizes, word frequencies, animals survival and dispersal processes, are assumed to be distributed according to the Zipf's power-law distribution. The Zipf plot is a plot of the logarithm of the frequencies of all events against the rank of these events and it should be a straight line for the Zipf law to be satisfied. Zipf's law states that the size of the r'th largest occurrence of the event is inversely proportional to it's rank r:

$$y \sim r^{-\alpha}, \qquad \text{with } \alpha \text{ close to unity.}$$

The webpage (**http://linkage.rockefeller.edu/wli/zipf/**) gives a comprehensive list of literature on the applications of the Zipf's law in several areas. Mandelbrot-Zipf distribution is adding an extra parameter ν

[1] Address correspondence to: Osama A. Hussien, Department of Statistics, Faculty of Commerce, Alexandria University, Alexandria, Egypt. e-mail: usama1@globalnet.com.eg.

$$y \sim (r+v)^{-\alpha},$$

to better fit the data when the Zipf distribution does not give a good fit.

Good (1953) introduced a converging factor q into the Zipf distribution:

$$y \sim q^r r^{-\alpha},$$

where $q \in (0,1)$ to better capture the shape of the data. Sichel (1975) stated that neither Good nor any other author have fitted the Good's distribution to any real data except for values of q very close to one. Kulasekera and Tonkyn (1992) reintroduced the Good's distribution and use it to model survival processes. Kemp (1998) showed that the Good's distribution (she renamed it the polylogarithmic distribution) with $\alpha > 0$ fits the long tailed distributions much better than the commonly used logarithmic distribution ($y \sim q^x / x$).

Zörnig and Altmann (1995) presented a three parameters generalization to the Zipf-Mandelbort and the polylogarithmic distribution. The probability mass function (pmf) is

$$P(X=x) = \frac{q^x}{b(\alpha,q,v)(x+v)^{\alpha}}, \quad x = 1,2,3,\ldots \qquad (1.1)$$

$$v > 0,\ 0 < q < 1$$

where

$$b(\alpha,q,v) = \sum_{x=1}^{\infty} \frac{q^x}{(x+v)^{\alpha}} = \Phi(\alpha,q,v) - v^{-\alpha}\ .$$

The function

$$\Phi(\alpha,q,v) = \sum_{x=0}^{\infty} \frac{q^x}{(x+v)^{\alpha}}$$

is called the Lerch transcendent (cf. Erdelyi et al. 1981).

In this work, we present a unified presentation of Hoaglin and Tukey plots and the Zipf plot and extend its use for 3-parameter families, specially the Lerch family. A comparison with other types of plots for discrete distributions is presented. Finally, applications of the Lerchianness plot to the Holy Quraan subject and letter frequencies and to Egyptian city sizes.

2. One-Parameter Exponential-Family Plot

A family $\{P_\theta\}$ of distributions is said to form a k-dimensional full rank exponential family if the distributions P_θ have densities of the form

$$P_\theta(x) = h(x)\ \exp\left\{\sum_{i=1}^{k} \theta_i T_i(x) - B(\theta)\right\} \qquad (2.1)$$

where $\boldsymbol{B}$ is a real valued function of the natural parameters $\boldsymbol{\theta} = (\theta_1,\theta_2,\ldots,\theta_s)$ and the T_i are real-valued statistics, and $\boldsymbol{x}$ is a point in the sample space (see Lindsey 1996: 33). Given a random sample of size N from a discrete distribution with pmf given by (1.1), the expected frequency at $X = x$ is given by

$$\eta_x = NP_\theta(X = x).$$

Thus,

$$\ln\left(\frac{\eta_x}{Nh(x)}\right) = \sum_{i=1}^{k} \theta_i T_i(x) - B(\theta). \tag{2.2}$$

Estimating η_x by n_x, the observed frequency of $X = x$, the function

$$\phi(\mathrm{n_x}) = \ln\left(\frac{n_x}{N\ h(x)}\right) \tag{2.3}$$

is called the *count metameter***.** For one dimensional distribution a plot of $\phi(n_x)$ versus $T(x)$ should be a straight line with slope θ_o and intercept $B(\theta_o)$ if the sample is drawn from $F_{\theta o}$ ($\equiv P(X \le x \mid \theta = \theta_o)$). Deviations from the line indicate types of departures from the assumed model $F_{\theta o}$. We refer to this plot as the *one-parameter exponential family plot.*

Given a random sample of size N from a one-parameter exponential discrete family the statistic $W = \sum_{i=1}^{N} T(x_i)$ is a complete sufficient statistic for θ, the Uniformly Minimum Variance Unbiased (UMVU) estimate of θ is $g(W(\mathbf{x}))$ such that $\mathrm{E}(g(W(\mathbf{X}))) = \theta$, if it exists (see Lehman and Casella 1998: 88). Thus, computing $\hat{\theta} = g(W(\mathbf{x}))$ an estimated distribution metameter may be defined as

$$\phi(\hat{\eta}_x) = \hat{\theta}T(x) - B(\hat{\theta}). \tag{2.4}$$

Note that

$$\frac{d\phi(\eta_X)}{d\theta} = T(x) - \frac{dB(\theta)}{d\theta}. \tag{2.5}$$

Thus, for values of θ near $\hat{\theta}$

$$\phi(n_x) - \phi(\eta_x) \approx \phi(n_x) - \phi(\hat{\eta}_x) + (\theta - \hat{\theta})\left(T(x) - \frac{dB(\hat{\theta})}{d\theta}\right). \tag{2.6}$$

Asking for values of θ that makes $\phi(n_x)$ "equal" $\phi(\hat{\eta}_x)$ is (approximately) equivalent to writing

$$0 = \phi(n_x) - \phi(\hat{\eta}_x) + (\theta - \hat{\theta})\left(T(x) - \frac{dB(\hat{\theta})}{d\theta}\right). \tag{2.7}$$

So, to describe the change in $\phi(n_x)$ - $\phi(\hat{\eta}_x)$ in relation to the change in θ we plot the above equation with $\phi(n_x)$ - $\phi(\hat{\eta}_x)$ on the vertical axis and $T(x) - \frac{dB(\hat{\theta})}{d\theta}$ on the horizontal axis. This plot is related to the *indicated-parameter-change plot* of Hoaglin and Tukey (1985) presented for the Poisson distribution with $\hat{\theta}$ replaced by some chosen value θ_o. They refer to a quantity similar to $T(x) - \frac{dB(\theta)}{d\theta}$ as the *leverage* of point x. Note that $\frac{dB(\theta)}{d\theta}$ = $E(T(\mathbf{X}))$ i.e. the leverage of a point x is the deviation of a point $T(x)$ from its expected value $E(T(\mathbf{X}))$. Another suggested plot is to construct a 100(1-α)% confidence band for $\phi(\hat{\eta}_x)$ and judge $\phi(n_x)$ as a good fit if it lies within the band.

Table 1 gives the parameters for the one-parameter exponential family plot for the Power series family distributions, the Zipf distribution and the Zipf-Mandelbrot distribution, with known parameter v. Hoaglin and Tukey (1985) consider only distributions belonging to the power series family, namely, the Poisson, the binomial, the logarithmic and the negative binomial including the geometric distributions. Considering the one parameter exponential family we get better estimates of θ used in constructing more meaningful indicated-parameter-change plot and a clear interpretation of the leverage of a point x. In addition, it gives a statistical meaning to the Zipf's law plots used in many applications. In fact, the data examples considered by Hoaglin and Tukey (1985) can be better fitted by the Zipf, or the more general Lerch distribution as we will see. Another important fact about the one-parameter exponential family plot is that it can be easily generalized to a 2-parameters exponential family plot. That is, we graph the count metameter $\phi(n_x)$ on the vertical axis and $T_1(x)$ and $T_2(x)$ on the horizontal axes. Also, the indicated-parameter-change plot can be extended by plotting $\phi(n_x)$ - $\phi(\hat{\eta}_x)$ on the vertical axis and $T_1(x)$ - $\frac{\partial B(\theta)}{\partial \theta_1}$ and $T_2(x)$ - $\frac{\partial B(\theta)}{\partial \theta_2}$ on the horizontal axes.

Table 1
One Parameter Exponential family members

Distribution	**Pmf**	**h(x)**	**θ**	**T(x)**	**B(θ)**	**$\phi(n_x)$**
Poisson	$e^{-\lambda}\lambda^x/x!$	$1/x!$	ln(λ)	x	e^θ	$\ln(n_x x!/N)$
Geometric	$q^x(1-q)/q$	1	$\ln(q)$	x	$\ln[e^\theta/(1-e^\theta)]$	$\ln(n_x/N)$
Logarithmic series	$q^x/(x(-\ln(1-q)))$	$1/x$	$\ln(q)$	x	$\ln(-\ln(1-e^\theta)$	$\ln(n_x x/N)$
Power series	$a(x)\, q^x/b(0,q,0)$	$a(x)$	$\ln(q)$	x	$\ln(b(0,e^\theta,0))$	$\ln(n_x/Na(x))$
Zipf	$x^{-\alpha}/b(0,1,\alpha)$	1	-α	$\ln(x)$	$\ln(b(0,1,-\theta)$	$\ln(n_x/N)$
Zipf-Mandelbrot v^* is unkown	$(x+v^*)^{-\alpha}/b(\alpha,1,\nu^*)$	1	-α	$\ln(x+v^*)$	$\ln((b(-\theta,1,\nu))$	$\ln(n_x/\mathrm{N})$

The UMVUE for θ for the power series family is $A(t-1, N)/A(t,N)$ where $A(t,N)$ is the coefficient of θ^t in the power expansion of $(b(0,q,0))^N$, $t = x_1 + x_2 + \ldots + x_N$ (See Lehmann and Casella 1998: 105).

3. The Frequency Ratio Plot

Ord (1967) showed that a relationship of the form

$$x\, P(X = x)/P(X = x - 1) = c_o + c_1\, x \qquad (3.1)$$

holds for some members of the power series distribution. He called the plot of xn_x/n_{x-1} against x "the frequency ratio plot". For the Lerch distribution, it is easy to see that

$$\frac{P(X=x+1)}{P(X=x)}=q\left(\frac{x+v}{x+v+1}\right)^{\alpha}. \tag{3.2}$$

Thus,

$$\ln\left(\frac{P(X=x+1)}{P(X=x)}\right)=\ln q+\alpha\ln\left(\frac{x+v}{x+v+1}\right). \tag{3.3}$$

A frequency ratio plot for the Lerch distribution will be a plot of $\ln(n_{x+1}/n_x)$ on the vertical axis versus $\ln\left(\frac{x+v}{x+v+1}\right)$, where the parameter v is estimated (by trial and error) before drawing the plot. The frequency ratio plot lacks resistance, since a single discrepant frequency affects the points for x and x+1. Moreover, the sampling variance of $\ln(n_{x+1}/n_x)$ fluctuates widely (see Friendly 2000). This drawback is illustrated in Figure 1 where we plot $\ln\left(\frac{P(X=x+1)}{P(X=x)}\right)$ versus $\ln\left(\frac{x+v}{x+v+1}\right)$ and $P(X = x)$ is calculated for $q = 0.5$, $\alpha = 1$ and $v = 10$. A random sample of size 1000 is drawn from the same distribution using the Aksenov and Savageau (2002) random number generator, and $\ln(n_{x+1}/n_x)$ was computed and plotted versus $\ln\left(\frac{x+v}{x+v+1}\right)$. The graph shows clearly the nonresistance of the frequency ratio plot.

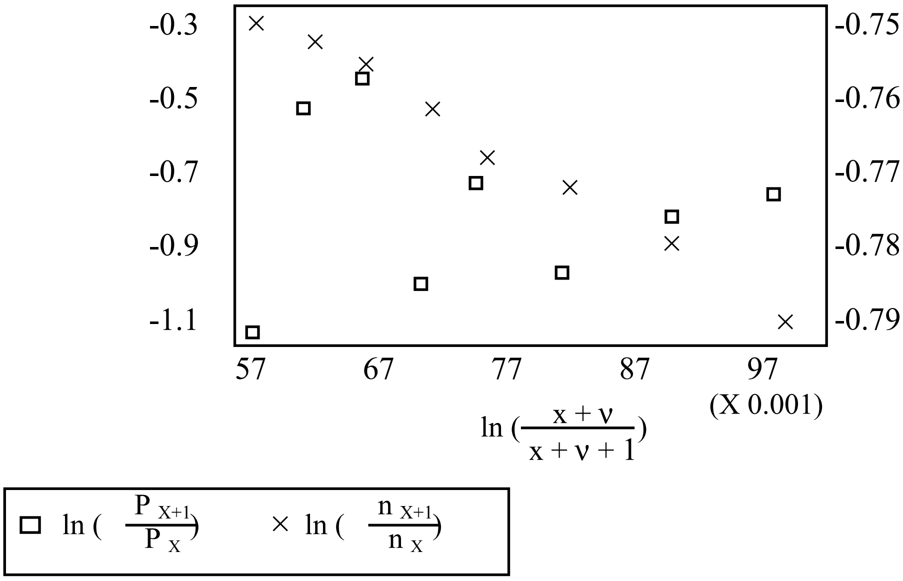


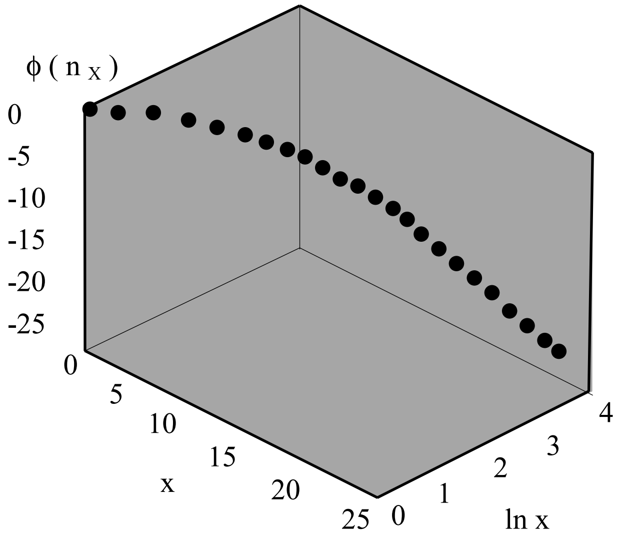
Figure 1. Nonresistance of the frequency ratio plot $\alpha = 1$, $v = 10$, $q = 0.5$

4. The 3-D Lerchianness Plot

For the Lerch distribution (1.1) with known parameter $\nu = \nu^*$, and $x \geq 1$

$$\phi(\eta_x) = \ln P(X = x) \;= x \ln q - \alpha \ln (x + \nu^*) - \ln b(\alpha, q, \nu^*) \qquad (4.1)$$

with special cases the Zipf distribution when $q = 1$ and $\nu = 0$, the Zipf-Mandelbrot distribution when $q = 1$, the logarithmic distribution when $\nu = \alpha = 0$, and the geometric distribution when $\nu = 0$ and $\alpha = 1$. So, one may think of the Lerch model as a model with extra parameter to better fit the data in all of the above one-parameter models. Given $\theta = (\alpha^*, q^*, \nu^*)$ the 3D plot of $\phi(\eta_x)$ on the vertical axis and x and $\ln(x + \nu^*)$ on the horizontal axes is a straight line. We call this plot the 3-D Lerchianness Plot. See Figure 2 for the theoretical Lerchianness plot with some chosen values of θ.



□

Figure 2a. Lerchianness plot for generated data N = 1000, α = 1

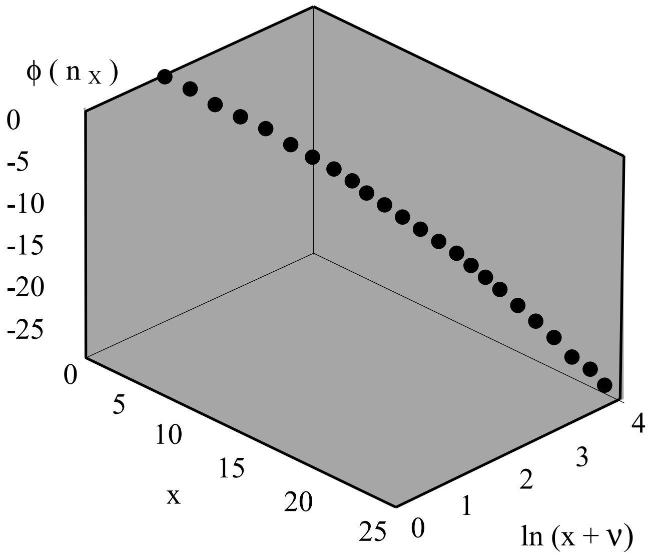


Figure 2b. Lerchianness plot for generated data
$N = 1000 \quad \alpha = 1 \quad \nu = 5 \quad q = 0.5$

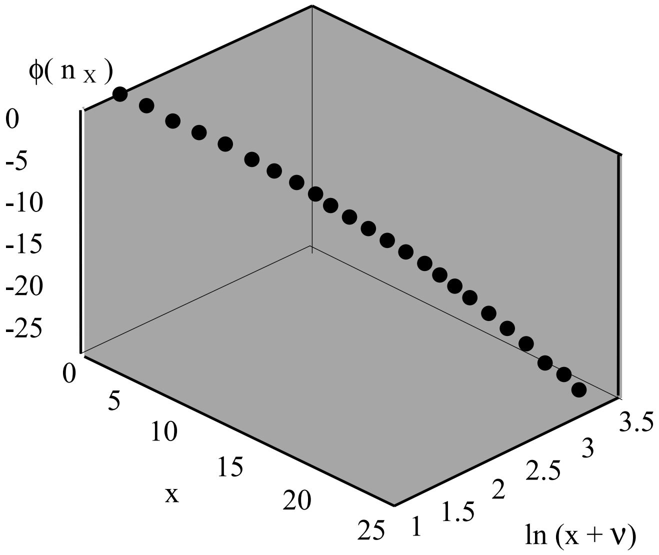


Figure 2c. Lerchianness plot for generated data
$N = 1000 \quad \alpha = 1 \quad \nu = 10 \quad q = 0.5$

Given a random sample, one can estimate θ to get $\phi(\hat{\eta}_x)$. Given an initial estimate of $\nu = v^*$, if the plot $\phi(n_x)$ versus to x and $\ln(x + v^*)$ is a straight line then the data can be fitted by the Lerch distribution. A comparison of the plot of $\phi(\hat{\eta}_x)$ and $\phi(n_x)$ indicate departures from the model. One may construct a 100(1-α)% confidence region for $\phi(\hat{\eta}_x)$ and judge $\phi(n_x)$ as a good fit if it lies inside this region. The indicated-parameter-change plot can be defined by plotting $\phi(n_x)$ - $\phi(\hat{\eta}_x)$ on the vertical axis and $(x - \hat{\mu})$ and $(\ln x - \hat{\mu}_{\ln x})$ on the horizontal axes, where $\mu = E(X)$ and $\mu_{\ln x} = E(\ln X)$.

One can choose v^* that maximizes the coefficient of determination R^2 of the least squares fit of $\phi(n_x)$ on x and $\ln(x + v)$. Note that the least squares estimates of $\ln(q)$ and α are sensitive to small changes in observed counts n_x, but remember we use this plot as a diagnostic tool only.

As a strategy, one can try a one-parameter plot like the power series or Zipf, by plotting $\ln(n_x)$ versus x or $\ln(x)$. If the Zipf plot gives better fit, try plotting $\ln(n_x)$ versus $\ln(x+v)$ for some values of the parameter v. Move to the Zipf-Mandelbrot if adding the parameter v will substantially increase R^2. Move to the Lerch model if adding the extra parameter q will increase R^2 again.

4. Hoaglin and Tukey Examples

The first example considered by Hoaglin and Tukey (1985: 351) is the international terrorism data. The data present number of incidents of international terrorism from January 1968 through April 1974. The data indicate an outlying observation, a month with 12 accidents while all other months with less than five accidents. We calculate R^2_{adj} for all Lerch models with v changes in range 1 to 100 with step 1, to get Figure 3. From the graph we choose $v^* = 40$. This example shows the resistance of the 3D Lerchianness plot, since we get a fit with high $R^2_{\text{adj}} = 0.93$ while the one-parameter plots, including the ones presented by Hoaglin and

Tukey, gives $R^2_{\text{adj}} < 0.3$. That is due to the effect of the outlying observation. Figure 4 shows that all data follows the Lerch model with one outlying observation.

The second example they consider is a word frequency data, page 391. Using the Poissonness plot and the negative binomial plot they have indicated that both distributions cannot provide a good fit. The Lerchianness plot indicates that the Lerch distribution gives a better fit. See Figure 5, where $v^* = 10$ will maximize R^2 adjusted. Also, the Zipf-Mandelbrot plot with $v^* = 10$ indicates a good fit.

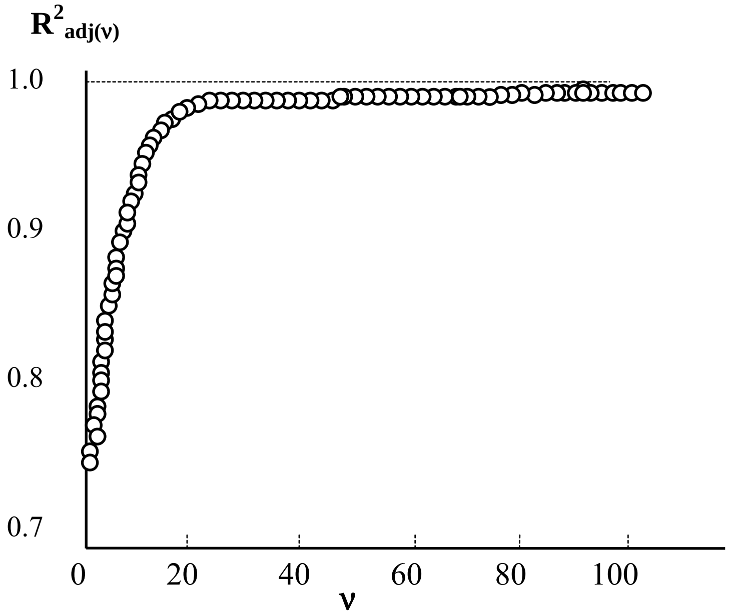


Figure 3. Plot of $R^2_{\text{adj(v)}}$ vs. v Terrorism data

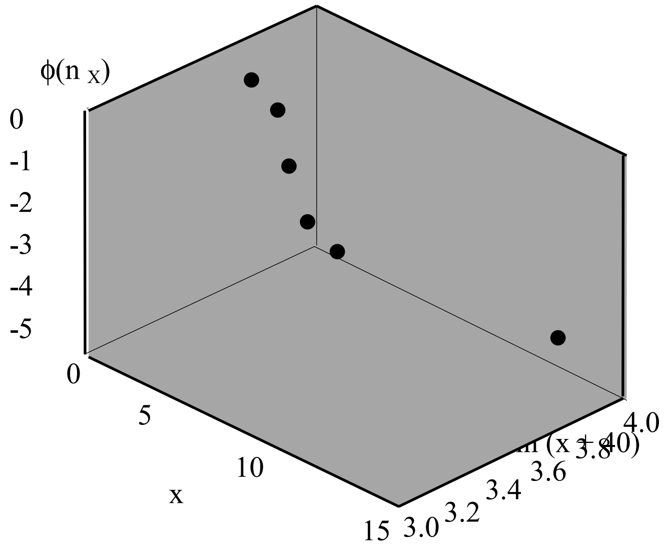


Figure 4. Lerchianness plot for the Terrorism data

$\hat{\beta}_1 = \log q = -6.578$ $\hat{\beta}_2 = -\hat{\alpha} = -316.376$

$R^2_{\text{adj}} = 0.93$ MSE = 0.183

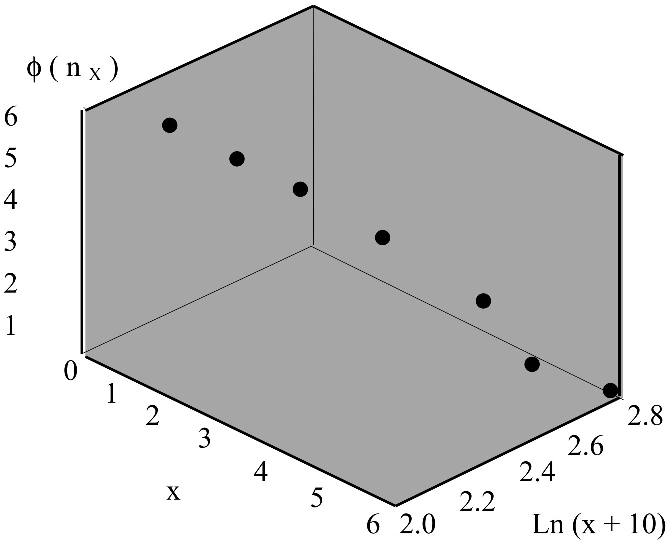


Figure 5 : Lerchianness plot for the word frequency data

$\hat{\alpha} = 9.132 \quad \hat{q} = 0.8226 \quad v^* = 10$

$R^2_{adj} = 0.9731 \quad \text{MSE} = 0.9731$

5. Egyptian City Size Data

Why do cities exist, and why do they vary in size? These fundamental questions have received a considerable amount of attention from regional and urban economists in recent years. In economics notation, the rank-size distribution states that there is an inverse linear relationship between the logarithmic size of a city and its logarithmic rank. A special case of the rank-size distribution is the Zipf's Law.

$$P(\text{city size} = S) = b / S^{\alpha} \text{ where } \alpha \approx 1.$$

Gabaix (1999) proposes the modification

$$P(\text{city size} = S) = b / (S + v)^{\alpha} \text{ for some parameter } v.$$

Even though the rank-size distribution does not hold exactly in reality, it does perform surprisingly well for the (historical) size distribution of cities in most industrialized countries, (Brakman et al., 1999), and even in other less developed countries like India in 1911 (see Zipf 1949: 432).

We use the Lerchianness plot to fit the city size of Egyptian data. The data represents the population of the largest 27 cites (population greater than 100,000) from 1986 and 1996 censuses, see Table 2. The sources of the data are the UN Demographic Yearbook at http://unstats.un.org/unsd/citydata/ and Thomas Brinkhof: City Population at http://www.citypopulation.de. The ranking for the two data sets are almost identical. Moreover, the Ordinary Least Squares (OLS) estimates are almost the same. This indicates the rate of growth is almost constant.

Table 2
Egyptian City Size

City	**Pop86**	**Pop96**
Cairo	6,068,695	6,789,479
Alexandria	2,926,859	3,328,196
Giza	1,883,189	2,221,868
Shubra-el-khema	714,594	870,716
Port Said	401,172	469,533
Suez	327,717	417,610
El-mahalla el-kubra	306,509	395,402
Tanta	336,517	371,010
Mansûra	317,508	369,621
Assyût	272,986	343,498
Zagazig	244,354	267,351
Faiyûm	213,070	260,964
Ismailia	158,045	254,477
Kafr-el-dwar	196,244	231,978
Aswan	190,579	219,017
Damanhûr	188,939	212,203
Menia	179,060	201,360
Beni-suef	152,476	172,032
Kena	119,917	171,275
Sohag	132,649	170,125
Shebin-el-kom	132,209	159,909
Luxer	126,160	*153,758*
Banha	115,701	145,792
Mallawi	98,632	119,283
Bilbays	96,511	113,608
Mit ghamr	91,927	101,803
Al-arish	67,337	100,447

The OLS results for the Zipf and the Lerch models are given in Table 3. The Zipf plot for 1996 is given in Figure 6. The Lerchianness plot for 1996 is given in Figure 7. The highest ranks are for Cairo, Alexandria then the two Greater Cairo areas. Again, the 3D graph is more resistance, since it detects only the Greater Cairo ranks as outliers (1,3,5), while the Zipf plot detect 1,5,6 where observations 5 and 6 are not an outlier. The 3D plot gives a clear distinction between observations 1,2,3,4 and the rest of the data. If we sum the population of the Greater Cairo then it satisfies what is called the law of primate city: "the size of the largest city is greater than the combined size of the second and the third largest ones". In fact, the size of Greater Cairo (9,882,063 in 1996) is greater than the size of all other cities combined (8,750,252 in 1996).

Table 3
OLS results of the city size data

	Lerch Model[*]					Zipf Model				Hill[**]
	$\hat{\beta}_o$	$\hat{\alpha}$	Ln $\hat{q}$	MSE	R^2_{adj}	$\hat{\beta}_o$	$\hat{\alpha}$	MSE	R^2_{adj}	$\hat{\alpha}$
1986	15.6 (0.1425)	1.518 (0.125)	-.03027 (0.0125)	0.0375	0.967	15.417 (0.1249)	1.244 (0.049)	0.053	0.9606	0.08532
1996	15.75 (0.1359)	1.486 (0.112)	-.0314 (0.0189)	0.0342	0.9684	15.536	1.213	0.0424	0.9608	0.0834

The Lerch model is given by (4.1), so $\hat{\beta}_o$ = lnb($\hat{\alpha}, \hat{q}$,v). The Zipf model is given by (4.1) with q=1. The number in parentheses under each estimate is its standard error.

** Hill estimator of α is a conditional maximum likelihood estimator given by

$$\frac{n-1}{\sum_{i=2}^{n} \ln(x_{(i)}) - \ln(x_{(1)})} \quad . \text{ (See Hill (1975)).}$$

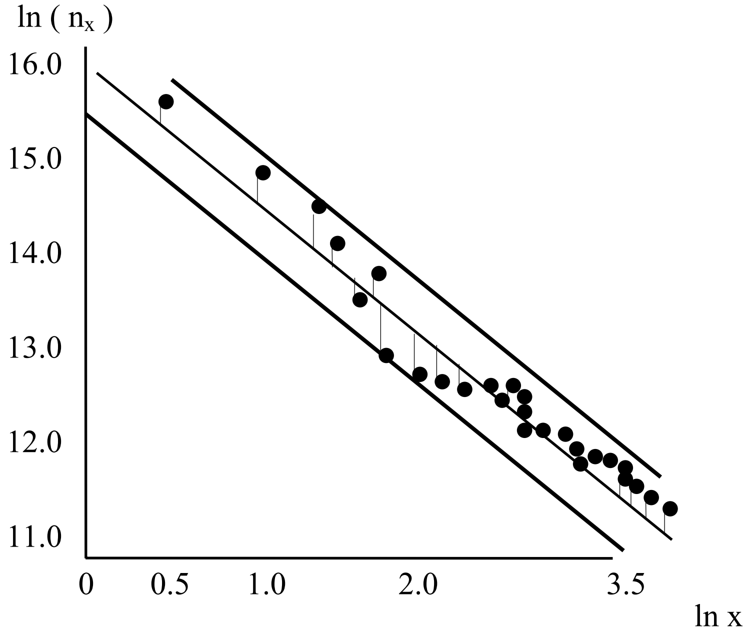


Figure 6. The Zipf plot for the city size data 1996

$\hat{\beta}_0 = 15.539$ $\hat{\alpha} = 1.2213$

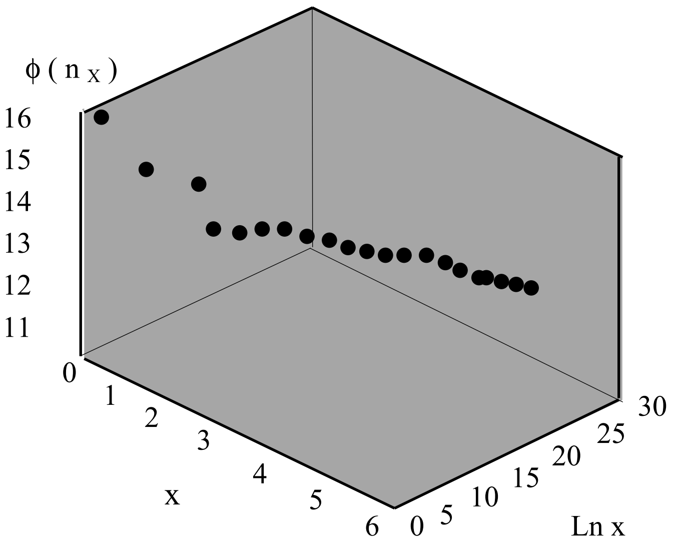


Figure 7. Lerchianness plot for city size data

6. Letter Frequency

Analysis of word frequencies have been used to establish the identity of the authors of literary work, assessing linguistic styles and to study multiple psychological dimensions of speech and text. See Pennebaker and King (1999). The Zipf distribution and the Zipf-Mandelbrot distribution have been used extensively as a description of frequency distribution of word in human languages and as general indicator of quality of texts. Good (1953) introduces a "convergence" factor "q", $0 < q < 1$ to the Zipf distribution, that is to use the Good distribution to model word frequencies.

Adding the modifications of Mandelbrot and Good to the Zipf distribution, we suggest using the Lerch distribution as a model for word frequencies. We illustrate this by drawing the 3D Lerchianness plot to the frequency of Arabic letters in the Holly Quraan. This plot would be an ideal standard for the high quality of any Arabic text. The frequencies of the letters are given in Table 4. See the Zipf-Mandelrot plot in Figure 8, where $\hat{\alpha} = 12.7596$; $v^* = 83$; $R^2_{\text{adj}} = 0.99$ and MSE = 0.0125. For the Lerchianness plot, Figure 9, we get $\hat{\alpha} = 0.09$ (0.062), $v^* = 0$, $\hat{q} = 0.9$ (0.00619), $R^2_{\text{adj}} = 0.9907$ and MSE = 0.0117.

Table 4
Letter frequencies in the Holy Quraan

letter	**freq**	**rank**
"أ"Alef	48800	1
"ل" Lam	33522	2
"م"Meem	26565	3
Wawo"و"	26565	4
Noon"ن"	26354	5
"ي"yaa	25909	6
"هـ"haa	19070	7
"ك"kaaf	14555	8

"ر"raa	11793	9
"ب"baa	11202	10
"ت"taa	10199	11
"ع"een	9020	12
"ف"faa	8499	13
"ق"khaaf	6813	14
"س"seen	5890	15
"د"daal	5642	16
"ذ"thal	4699	17
"لا"lamalef	4099	18
"ح"haa	3990	19
"ج"geem	3273	20
"ط"ttaa	3174	21
"خ"khaa	2416	22
"ض"ddaa	2293	23
"ش"sheen	2253	24
"غ"kheen	2208	25
"ز"zean	1570	26
"ث"tehh	1276	27
"ص"ssaad	1180	28
"ظ"zzaa	842	29

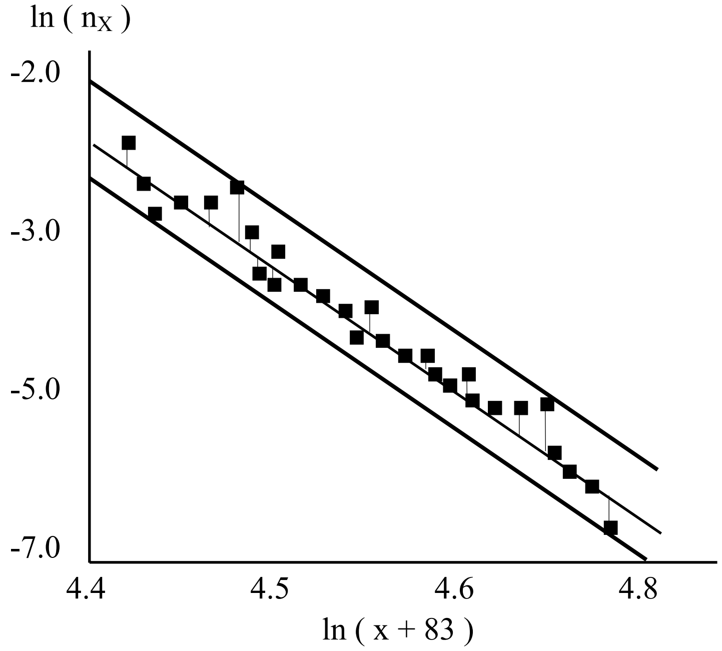


Figure 8. The Zipf-Mandelbrot plot for letter frequency data

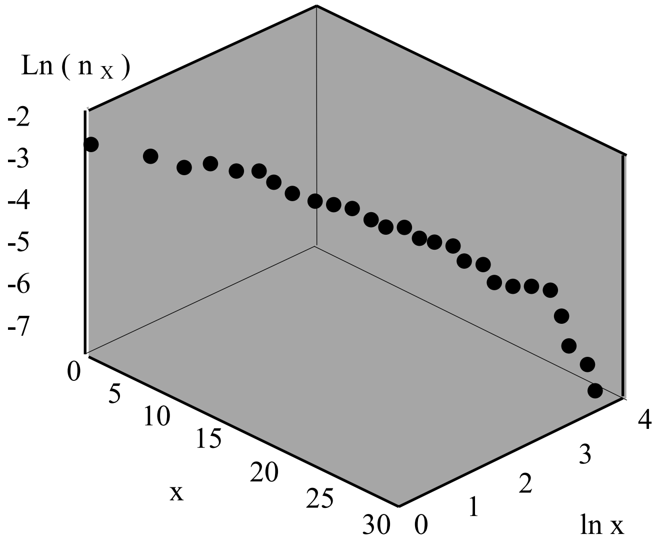


Figure 9. The Lerchiannes plot for the letter frequency data

References

Askenov, S., Savageau, M. (2002). Parameter estimation and random number generation from a Zipf-related Lerch distribution. Submitted to *Computational Statistics and Data Analysis.*

Brakman, S., Garretsen, H., Marrewijk, C., Berg, M. (1999). The return of Zipf: towards a further understanding of the rank-size distribution. *Journal of Regional Science 39, 183-213.*

Erdélyi, A., Magnus, W., Oberthettinger, F., Tricomi, **F.G.** (1981). *Higher Transcendental Functions, Vol 1.* New York: Krieger.

Gabaix, X. (1999). Zipf's law for cities: an explanation. *Quarterly Journal of Economics 111 .739-767.*

Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika, 40(3), 237-264.*

Friendly, M. (2000). *Visualizing categorical data*. Cary, N.C.: SAS Institute.

Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics 3(5), 1163-1174.*

Hoaglin, D.C. (1980). A Poissonness plot. *The American Statistician, 34(3), 146-149.*

Hoaglin, D.C., Tukey, J.W. (1985). Checking the shape of discrete distributions. In: Hoaglin, D.C., Mosteller, F. and Tukey, J.W. (eds). *Exploring Data Tables, Trends and Shapes*. New York: Wiley.

Kemp, A. (1998). *Polylogarithmic distributions*. Supplement to Encyclopedia of Statistical Sciences. (eds. Kotz, S., N.L. Johnson, C.B. Read). New York: Wiley.

Kulasekera, K.B., Tonkyn, D.W. (1992). A new discrete distribution, with applications to survival, dispersal and dispersion. *Communications in Statistics-Simulation 21(2), 499-518.*

Lehmann, E.L., Casella, G. (1998). *Theory of point estimation, 2ed ed.* New York: Springer-Verlag.

Lindsey, J.K. (1996). *Parametric statistical inference*. Oxford: Oxford Science Publications.

Ord, J.K. (1967). Graphical methods for a class of discrete distributions. *Journal of the Royal Statistical Society A 130, 232-238.*

Pennebaker, J.W., King, L.A. (1999). Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology 77(6), 1296-1312.*

Sichel, H.S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association 70(351), 542-547.*

Zörnig, P., Altmann, G. (1995). Unified representation of Zipf distribution. *Computational Statistics and Data Analysis 19, 461-473.*

Zipf, G.K. (1949). *Human behavior and the Principle of Least Effort*. New York: Addison-Wesley.

Glottometrics 7, 2004, 65-84

Symmetry of Japanese Kanji Lexical Productivity on the Left- and Right-hand Sides

Katsuo Tamaoka (Hiroshima, Japan)[1]
Gabriel Altmann (Lüdenscheid, Germany)

Abstract: Japanese kanji combine with other kanji to produce various two-kanji compound words. First, the present study examined whether the extent of left-hand and right-hand productivity of the Japanese 1,945 basic kanji abides by an 'honest' distribution. The result showed that kanji compound building (or kanji lexical productivity) was depicted by a birth-and-death process leading to the negative binomial and/or the Waring distribution. Second, the study investigated whether these basic kanji display symmetry on the left- and right-side lexical productivity. Analysis of these kanji suggested that although each kanji displayed symmetry in lexical productivity, there is no tendency among the basic kanji to produce their compound words to the same extent on the left or the right side on the whole.

Key words: Japanese kanji, kanji lexical productivity, birth-and-death process, symmetry

1. Two-kanji compound words in a Japanese Dictionary

In the written Japanese, *kanji*, which adopted Chinese characters for expressing various words, often combine with another kanji to produce a new meaning. In this sense, the unit of kanji refers to 'morpheme', the smallest unit of meanings. Yokosawa and Umeda (1988) reported that approximately 70 percent of 51,962 words listed in a particular Japanese dictionary were composed of two kanji or two morpheme combinations. In 1981, the Ministry of Education, Science, Sports and Culture, Government of Japan (hereafter simply called the 'Japanese Ministry of Education' except in quotes) published a list of the *1,945 Basic Japanese Kanji* (for detailed information, see Kato, 1989; Yasunaga, 1981). This list established a standard for kanji usage in printed and written Japanese texts (Ministry of Education, Science, Sports and Culture, Government of Japan, 1987, 1998). The National Institute for Japanese Language (1976) conducted a survey on the frequency of kanji in print, and found that 2,000 kanji encompassed 99.6 percent of all kanji used in three major Japanese newspapers, *Asahi*, *Mainichi* and *Yomiuri* published during the year 1966. Although the 1,945 Basic Japanese Kanji and the 2,000 kanji mentioned above were not identical, it is roughly estimated that these basic kanji cover approximately 99 percent of all kanji used in Japanese newspapers. Combining this figure with 70 percent of two-kanji compound words in a dictionary, 1,945 kanji will provide us a reasonable estimate on Japanese kanji and their productivity.

[1] Address correspondence to: Katsuo Tamaoka, International Student Center, Hiroshima University 1-1, 1-chome, Higashihiroshima, Japan 739-8524. E-mail: ktamaoka@hiroshima-u.ac.jp

2. Counting of kanji lexical productivity and purpose of the present study

As Yokosawa and Umeda (1988) pointed out, the majority of lexical items in a Japanese dictionary are words consisting of two kanji units. The number of lexical productivity is simple counting (i.e., type frequency, not token frequency) of all possible two kanji combinations. Interestingly, various complex compound words are often constructed by two kanji compound words such as 経済政策 (/keizai seisaku/[2], ‘economic policy’), 宇宙遊泳 (/utjuR juRei/, ‘space walk’), and 冷凍食品 (/reitoR sjokuhiN/, ‘frozen food’). Furthermore, the word ‘frozen food’ can be combined with another two-kanji compound word 貯蔵 (/tjo zoR/, ‘storage’) to make a six-kanji compound word, 冷凍食品貯蔵 (/reitoR sjokuhiN tjozoR/, ‘storage of frozen food’) consisting of three two-kanji compound words. Thus, two-kanji compound words are a sensible unit to count how many words each kanji can produce.

A single kanji can produce two-kanji compound words in two ways, produced by the combination of kanji placed on the left-hand and right-hand side positions of two-kanji compound words. For example, the kanji 学 /gaku/ meaning ‘to learn’ or ‘learning’ is combined with another kanji on the right-hand side position such as in 学校 (/gaQ koR/, ‘school’), 学生 (/gaku sei/, ‘student’) and 学者 (/gaku sja/, ‘scholar’). Combinations with other kanji on the left-hand side position are also possible such as in 入学 (/njuR gaku/, ‘school admission’), 文学 (/bun gaku/, ‘literature’) and 私学 (/si gaku/, ‘private school’). Kanji productivity of two-kanji compound words refers to the unit of one kanji combined with another kanji to create two-kanji compound words. Therefore, the concept of this term could be understood as a linguistic concept of kanji ‘lexical productivity’ (details see, Hayashi, 1987; Nomoto, 1989; Nomura, 1988, 1989), which can be calculated by the left-side, right-side and both sides together. The present paper investigated two questions: (1) how two-kanji compound words were produced by a single kanji on the left-hand and the right-hand sides, and (2) how symmetric they are on both sides.

3. Kanji lexical productivity

The present study used a lexical corpus of 341,771 words that was established from newspapers containing 287,792,797 words by Amano and Kondo (2000), all of which were taken from the *Asahi Newspaper* printed from 1985 to 1998. At present, this is the largest and the most up-to-date word corpus created from calculating the word frequency of occurrence in Japanese written texts. For counting kanji lexical productivity, the programming language of MacJPerl 5.15r4J for Macintosh was used to run a calculation procedure. Type frequency counts of kanji lexical productivity on the left-hand, right-hand and both sides were arranged for all the 1,945 basic kanji. The beginning 10 kanji of the raw data are presented in Table 1.

[2] The pronunciation in this paper is transcribed using Japanese phonemic symbols which indicate three special sounds in Japanese: /N/ for nasal, /Q/ for geminate and /R/ for long vowel.

Table 1
The first ten kanji and their lexical productivity on the left- and right-hand sides

No.	Kanji	In phonemes	Left-hand side	Right-hand side	Both sides together
1	亜	/a/	3	13	16
2	哀	/ai/	1	19	20
3	愛	/ai/	28	55	83
4	悪	/aku/	35	80	115
5	握	/aku/	4	4	8
6	圧	/atu/	42	21	63
7	扱	/atu/	1	1	2
8	安	/aN/	24	40	64
9	案	/aN/	49	6	55
10	暗	/aN/	5	40	45

Note: A common kanji pronunciations are used to arrange kanji in this table.

Each of the chosen kanji develops its own individual left and right lexical productivity. For example, the eighth kanji 安 (/aN/, ‘rest’, ‘relax’, ‘cheap’, ‘peaceful’, etc.) produced 24 two-kanji compound words by adding kanji on the left-hand side while producing 40 compounds by adding kanji on the right-hand side. The total lexical productivity becomes 64. By counting the number of kanji with *x* compounds, the results presented in Table 2 for kanji on the left-hand side and in Table 3 for kanji on the right-hand side. Two-step investigation is conducted using data of kanji lexical productivity in the present study. The first step is to conjecture the mechanism generating the distribution of compounds. The second step is the eventual modification of the result. This second step must be made because the basic kanji were taken from a ready list, not at random from the dictionary. Thus the bias can be quite systematic.

Table 2
Distribution of productivity of the 1,945 basic kanji on the left-hand side

x	f_x	x	f_x	x	f_x	x	f_x
0	108	35	10	71	2	112	1
1	135	37	14	72	1	114	1
2	109	38	16	73	1	116	1
3	128	39	5	75	5	117	1
4	82	40	9	76	1	118	4
5	84	41	13	77	3	120	2
6	84	42	6	78	2	121	2
7	61	43	8	79	2	122	1
8	73	44	10	80	2	123	1

9	67	45	9	82	4	126	1
10	59	46	8	83	2	127	1
11	53	47	5	84	2	128	1
12	48	48	8	85	2	129	1
13	49	49	6	86	1	130	1
14	36	50	5	87	1	131	2
15	37	51	2	88	1	132	1
16	37	52	7	89	2	135	2
17	39	53	9	91	1	136	1
18	33	54	3	92	1	138	1
19	29	55	6	93	2	140	1
20	17	56	2	94	1	142	1
21	34	57	5	96	2	146	1
22	28	58	3	97	1	148	1
23	15	59	5	98	1	151	1
24	17	60	3	99	1	153	1
25	22	61	3	100	1	162	1
26	21	62	3	101	1	170	1
27	19	63	3	102	1	171	1
28	16	64	6	103	1	180	1
29	13	65	6	104	1	213	1
30	16	66	3	106	4	246	1
31	15	67	8	108	1	268	1
32	13	68	3	109	2		
33	8	69	3	110	1		
34	11	70	3	111	1		

Note: x = number of compounds, f_x = number of kanji producing x compounds.

Table 3
Distribution of productivity of the 1,945 basic kanji on the right-hand side

x	f_x	x	f_x	x	f_x	x	f_x
0	72	33	17	67	3	111	2
1	91	34	16	68	1	114	1
2	115	35	7	69	2	116	1
3	121	36	13	70	4	118	1
4	90	37	13	71	4	119	1
5	93	38	13	72	5	120	1
6	92	39	7	73	3	126	1
7	87	40	11	74	1	127	1
8	80	41	11	75	3	132	1
9	61	42	15	76	2	135	2
10	78	43	9	77	6	137	2
11	51	44	12	78	1	138	3
12	57	45	4	79	3	145	1
13	49	46	6	80	4	146	1
14	51	47	10	82	1	148	1

15	34	48	5	83	2	149	1
16	45	49	7	86	1	151	1
17	32	50	8	87	3	154	2
18	35	51	5	88	1	156	1
19	31	52	6	90	2	157	1
20	29	53	4	91	1	158	1
21	22	54	8	92	1	160	1
22	21	55	4	95	2	163	1
23	18	56	2	96	1	170	1
24	22	57	4	97	4	179	1
25	11	58	1	99	1	194	1
26	21	59	6	101	1	195	1
27	15	60	3	103	2	293	1
28	17	61	5	104	1	350	1
29	12	62	4	106	3	399	1
30	14	63	5	107	1		
31	15	64	3	109	1		
32	15	66	1	110	1		

Note: x = number of compounds, f_x = number of kanji producing x compounds.

4. A birth-and-death process of kanji compound words

Lexical productivity in languages, consisting of two kanji in the present study, always faces to the birth of new forms (e.g., compounds) and their death. The process of the birth and death has two general features in languages. First, this process is incessant in any language since newly-formed words are influenced steadily by various linguistic environments. As a result, some newly-produced words die immediately, some live longer, and some survive for a very long time. Second, this process is, nevertheless, in steady state and is balanced by language self-regulation (Köhler, 1986). Thus, having these two features together, lexical productivity is consistently modeled in the manner of an incessant steady-state birth-and-death process (Altmann 1985; Wimmer & Altmann, 1995).

Considering f_x the number of kanji building exactly x compounds, P_x is the relative number corresponding to f_x, or the probability in the model. A kanji leaves class x if either a compound “dies” (the kanji goes in class x-1) or a new compound arises (it goes to class x+1). A kanji enters class x if in class x-1 a new compound is born or in the class x+1 a compound dies. Let the birth ratio be λ_x and the death ratio μ_x. Then, considering the probability of birth or death of two or more compounds at a time as zero, the process can be written in form of simple equations:

$$(1) \qquad \begin{aligned} \lambda_0 P_0 &= \mu_1 P_1 \\ (\lambda_x + \mu_x) P_x &= \lambda_{x-1} P_{x-1} + \mu_{x+1} P_{x+1}\ , x = 1,2,\ldots \end{aligned}$$

The equation (1) can be solved stepwise but the solution is known as

$$(2) \qquad P_x = P_0 \frac{\lambda_0 \lambda_1 \ldots \lambda_{x-1}}{\mu_1 \mu_2 \ldots \mu_x}\ .$$

Using simple linear functions for the rates, namely

$\lambda_x = a+bx$ (a = constant birth coefficient, b = coefficient of assertion against x rivals)
$\mu_x = cx$ (c = death coefficient) .

The equation (3) is obtained by substituting them in (2)

$$(3)\qquad P_x = P_0 \frac{a(a+b)(a+2b)...[a+(x-1)b]}{x!c^x}.$$

The rest is simple manipulation. Factoring out b and requiring $c > b$ (because of convergence), and substituting

$b/c = q$ $(0 < q < 1)$
$a/b = r$

the equation (4) is obtained:

$$(4)\qquad P_x = P_0 \frac{r(r+1)...(r+x-1)}{x!} q^x = P_0 \binom{r+x-1}{x} q^x .$$

Since P_0 is the normalizing constant following from $\Sigma P_x = 1$, we obtain at last (with p = 1-q)

$$(5)\qquad P_x = \binom{r+x-1}{x} p^r q^x \ , \quad x = 0,1,2,\ldots$$

representing the usual negative binomial distribution, symbolized as $NB(r,p)$. Now fitting this distribution to the data, it can be seen easily that it is not adequate because of the possible bias. The data can even be smoothed by pooling the empirical classes but we rather seek a modified model. The model can be found in different ways. The study will present two of them.

4.1. Fitting the mixed negative binomial distribution for kanji lexical productivity on the left-hand side

Since the kanji are taken from the ready list which may contain different classes of kanji – whatever kinds of classes may be concerned – it is, in the first step possible to add to the equation (5) a second component with different parameters and mixing the components in different proportions, say α and 1-α. Considering this specification, we obtain from the equation (5) the model in the form

$$(6)\qquad P_x = \alpha \binom{r+x-2}{x-1} p_1^r q_1^{x-1} + (1-\alpha) \binom{r+x-2}{x-1} p_2^r q_2^{x-1},\ x = 1,2,3,...$$

where p_1, p_2, r and α are coefficients interpreted above. In our case $k = 1.2269$, $p_1 = 0.1159$, $p_2 = 0.0270$, $\alpha = 0.672$. As can be seen, the Chi-square is 121.52 which with 143 *DF* signalizes a very good fit, $P = 0.90$. The last theoretical value (at $x = 268$) was computed as $1-\sum_{x=0}^{267} P_x$. The graphical representation displaying it can be found in Fig. 1. The observed and theoretical values are shown together with class pooling $NP_x > 1$ in Table A of the Appendix.

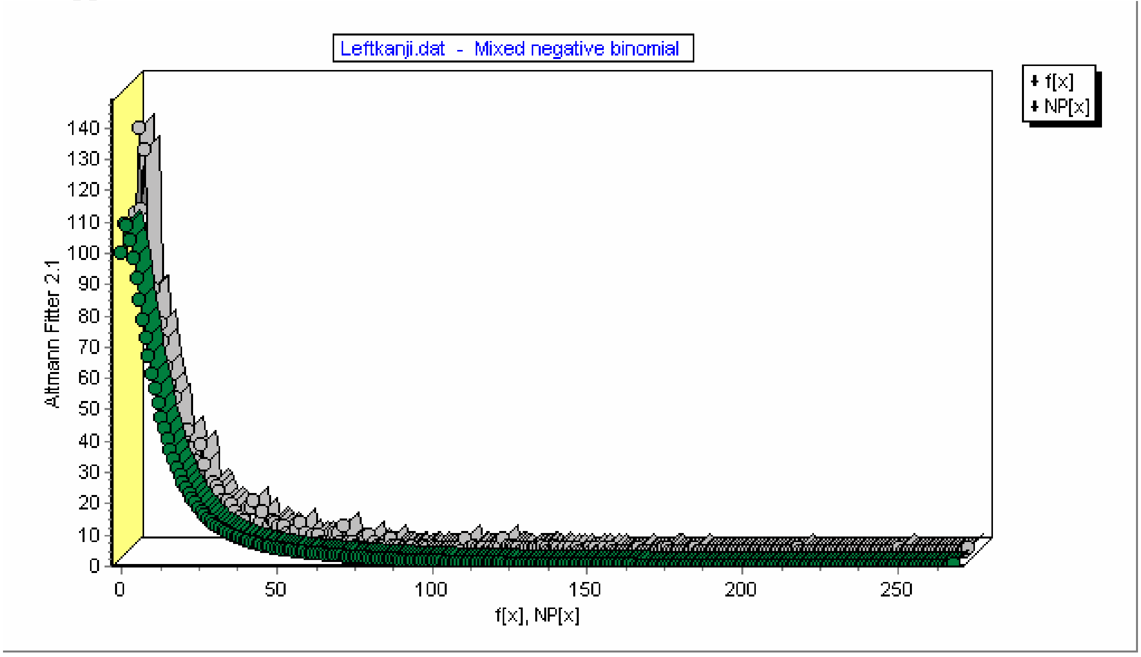


Fig 1. Fitting the mixed negative binomial distribution to left-hand kanji compounding

4.2. Fitting the Waring distribution for kanji lexical productivity on the left-hand side

Another method is the randomization of the parameter p yielding a compound distribution. One of the many possibilities is, symbolically

$$(7) \qquad NB(\mathrm{r},p) \underset{p}{\wedge} beta(b,1)$$

where beta(b,1) is the beta distribution with parameters b and 1, that is

$$(8) \qquad f(p) = \frac{p^{b-1}}{B(b,1)}, \quad 0<p<1, \quad q = 1-p.$$

Multiplying (5) by (8) and integrating the product according to $p \in <0,1>$, one obtains the Waring distribution, introduced in linguistics by G. Herdan, yielding

$$(9) \qquad P_x = \frac{b}{b+r}\frac{r^{(x)}}{(b+r+1)^{(x)}}, \quad x = 0,1,2,...$$

However, this distribution can be obtained directly from the birth-and-death process with birth-rate $\lambda_x = r+x$ and death-rate $\mu_x = b+r+x$. This distribution holds equally well for the data in Table 2: Parameters: b = 3.4961, r = 53.6656, *DF* =136, Chi-square = 124.70, *DF* = 136, *P* = 0.75. Thus, one can derive the Waring distribution either directly from the birth-and-death process or indirectly by randomisation. Assuming that more extended examination could conceal some surprise, we preliminarily adhere to the negative binomial distribution.

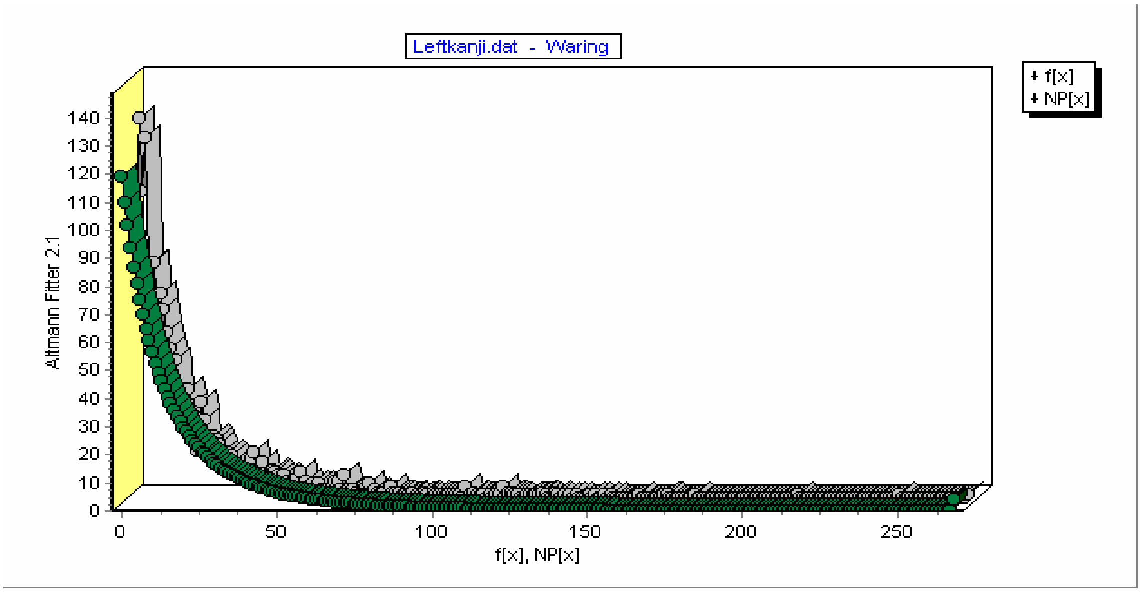


Fig. 2. Fitting the Waring distribution to left-hand kanji compounding

4.3. Kanji lexical productivity on the right-hand side

For the right-hand compounding in Table 3, the mixed negative binomial yields a X^2 = 134.56, *DF* = 136, *P* = 0.52, the fitting is shown graphically in Fig. 3. The Waring distribution (with parameters b = 3.4224, r = 50.9697) is somewhat weaker (X^2 = 162.19, *DF* = 137, *P* = 0.07) but still acceptable. Different pooling of data can improve the fitting (up to P = 0.11). In Figure 4, one finds the fitting of (6) displayed graphically.

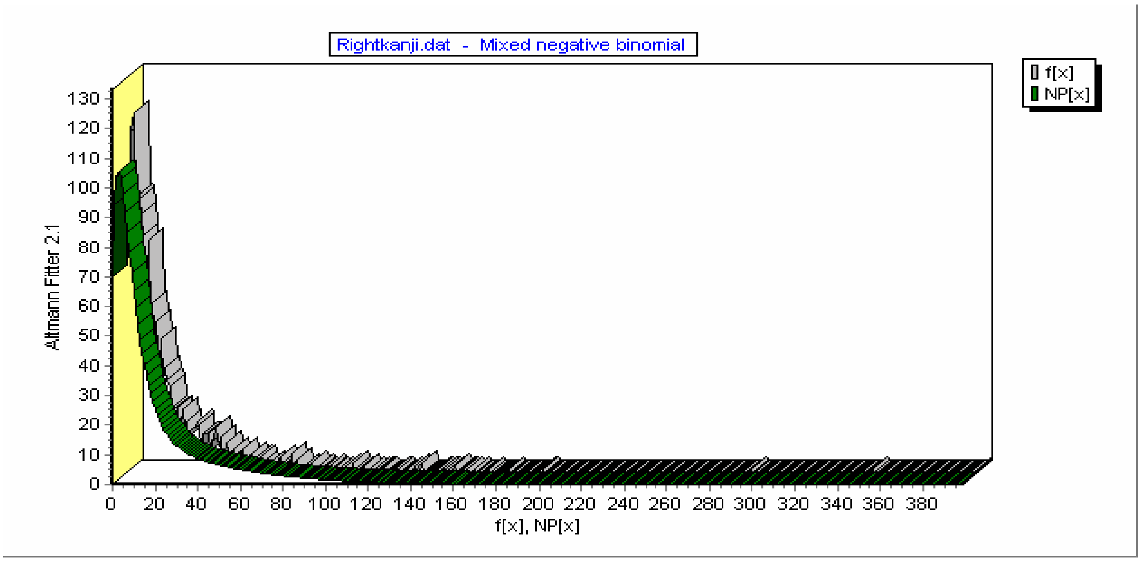


Fig. 3. Fitting the mixed negative binomial d. to right-hand kanji compounding

It must be remarked that the pooling of theoretical values has been performed so that each class contained at least $NP_x > 1.00$. The empirical values were pooled a posteriori but the computation was made for each frequency class explicitly.

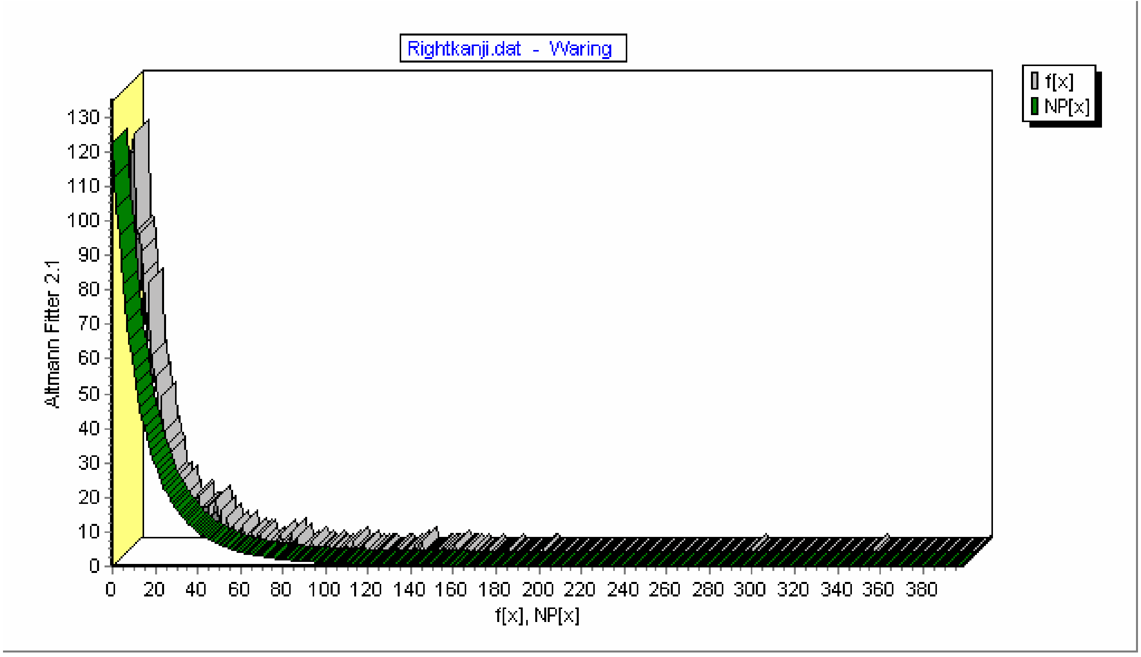


Fig. 4. Fitting the Waring distribution to right-hand kanji compounding

5. Symmetry of kanji lexical productivity

Looking at Table 1, one would intuitively say that left-hand and right-hand side compounding is not symmetrical. In order to test the symmetry (a), one can use for individual kanji the binomial test setting $p = 1/2$, n = total compounding, and compute the probability of, say, the left-hand or the more extreme x. (b) For the whole table one can use the Bowker-test for symmetry (Bowker 1948), leaving out all kanji for which $x_{left} = x_{right}$ or simply perform the test for homogeneity of two sides. (c) Making a quadratic contingency table of the data in Table 1, one can test the marginal homogeneity using Stuart´s test (Stuart 1955).

But if one constructs a two-dimensional distribution with X = left-hand, Y = right-hand compounding, one obtains an enormous matrix which is neither well processable nor very interesting because the great majority of cells is empty. Thus, pooling is necessary. However, in that case, some differences between left and right compounding (within the pooling interval) disappear and the problem of symmetry gets a new face. Problem (a) is eliminated, problem (b) changes thoroughly, problem (c) will display greater homogeneity, but a new aspect arises, namely (d) the possibility to test whether the diagonal of the contingency matrix is preferred; that is, whether there is a tendency to make left-hand and right-hand compounding similar or not. Using test (a), one gets a result concerning individual kanjis, test (b) ignores the diagonal, but test (d) can show whether there is a tendency to make the productivity on the two sides of kanji similar (not identical).

5.1. The compounding symmetry of individual kanji

In Table 1, kanji No. 1 has 16 compound words out of which it is 3 times on the left side (i.e., it is *progressive*) and 13 times on the right side (i.e., it is *regressive*) of the compounds. One can ask whether this ratio is random or whether there is a tendency to build compounds of a special kind. If there is no tendency, the probability that the compounding part will stay right or left is p = 0.5. One finds n (here 16) compounds. What is the probability that exactly x compounds will be progressive (regressive respectively)? Since there are merely two possibilities, the compounds are distributed binomially; thus,

$$(10) \quad P_x = \binom{n}{x} p^x q^{n-x} = \binom{n}{x} (1/2)^n$$

in this case. However, we need the sum of all probabilities being smaller than 0.025 or greater than 0.975; that is, we seek a number k_1 for which $P(X \le k_1) \le 0.025$ and a number k_2 for which $P(X \ge k_2) \le 0.025$. These can be computed by means of the criteria

$$(11) \quad P(X \le k_1) = \sum_{x=0}^{k_1} \binom{n}{x} (1/2)^n \le 0.025$$

$$(12) \quad P(X \ge k_2) = \sum_{x=k_2}^{n} \binom{n}{x} (1/2)^n \le 0.025.$$

Thus, we obtain a kind of confidence interval [k_1; k_2] within which the basic kanji is symmetric, and below or above the interval there is a significant tendency to construct progressive or regressive compounds.

For our example, we have n = 16. Computing the probabilities from 0 according to (11) we see that k_1 = 4 and according to (3) k_2 = 12 because the sum of the probabilities from x = 0 to x = 3 is smaller than 0.025, and that from 13 to 16 is smaller than 0.025, i.e. our interval is [4; 12]. Since kanji No.1 has 13 regressive compounds, it lies outside of the interval and can be classified as a compositionally regressive kanji.

Consider kanji No. 6 with n = 63 compounds. The interval is [24; 39]; since the numbers in Table 1 lie outside this interval and the left-hand side productivity is greater, this kanji has a progressive compounding. However kanji No. 8 with n = 64 and interval [24; 40] lies still in the interval and can be classified as symmetric. The other values are shown in the last column of Table 4. All intervals for n = 6 to 400 are in the Appendix, Table C.

An asymptotic test for symmetry of individual kanji can be performed as follows: Let the number of left-hand side compounds be n_L, that of right-hand side ones n_R and $n_L + n_R = n$. Then, under the hypothesis of equality of both sides, the expected value is n/2. The asymptotic Chi-square criterion is

$$(13) \quad X^2 = \frac{\left(n_L - \frac{n}{2}\right)^2}{\frac{n}{2}} + \frac{\left(n_R - \frac{n}{2}\right)^2}{\frac{n}{2}} = \frac{\left(\frac{n_L - n_R}{2}\right)^2}{\frac{n}{2}} + \frac{\left(\frac{n_R - n_L}{2}\right)^2}{\frac{n}{2}}$$

$$= \frac{(n_L - n_R)^2}{n_L + n_R}$$

which is distributed as a chi-square with 1 degree of freedom. At the α = 0.05 level it must be greater than 3.84 in order to be significant. For example, let $n_L = 2$ and $n = 10$, then (13) yields

$$X^2 = \frac{(2-8)^2}{2+8} = 3.6$$

which is not significant, but for $n_L = 1$, we obtain $(1\text{-}9)^2/10 = 6.4$ which is significant. In most cases this test yields the same results as the exact (binomial) test. Instead of n_L one can insert n_R. The interpretation is the same.

All the 1,945 kanji were tested for symmetry. There were 227 kanji (11.67% of the total) with less than 5 of the total kanji lexical productivity putting both the left-hand and right-hand sides together. Excluding these kanji for symmetry, 902 kanji (46.38%) were judged to be symmetric represented by ‘S’ in Table 4 for the examples of ten kanji. When the left-hand side productivity was greater than the right-hand side, a kanji was judged as progressively asymmetric presented by ‘P’. 403 kanji (20.72%) fell into this category. When the right-hand side productivity was greater than the left-hand side, a kanji was judged as regressively asymmetric presented by ‘R’. 413 kanji (21.23%) were counted in this category. Simply looking at the percentages, a set of the whole 1,945 basic kanji seems to display symmetric pattern of lexical productivity, being 20.72% for progressive, 46.38% for symmetry and 21.23% for regressive. The next section investigates the pattern of the 1,945 basic kanji.

Table 4
Symmetric classification of the first ten kanji in lexical productivity

No.	Kanji	In phonemes	Left-hand side	Right-hand side	Both sides together	Classification
1	亜	/a/	3	13	16	R
2	哀	/ai/	1	19	20	R
3	愛	/ai/	28	55	83	R
4	悪	/aku/	35	80	115	R
5	握	/aku/	4	4	8	S
6	圧	/atu/	42	21	63	P
7	扱	/atu/	1	1	2	S
8	安	/aN/	24	40	64	S
9	案	/aN/	49	6	55	P
10	暗	/aN/	5	40	45	R

5.2. The symmetry of the entire compounding of 1,945 basic kanji

In order to ascertain whether the whole field of basic kanji displays symmetric compounding, one simply builds the sum of (13) which is identical with Bowker´s test for symmetry in a contingency table, i.e.

$$(14) \quad X^2 = \sum_{i=1}^{K} \frac{(n_{L,i} - n_{R,i})^2}{n_{L,i} + n_{R,i}}$$

where K is the number of comparisons (here 1945). The number of degrees of freedom is DF = (K – number of cases where n_L and n_R are equal).

An asymptotically equivalent test can be derived using information statistics, namely

$$(15) \quad 2I = 2\sum_{i=1}^{K}\sum_{j=1}^{2} n_{ij} \ln \frac{n_{ij}}{n_i/2} = 2N\ln 2 + 2\sum_{i=1}^{K}\sum_{j=1}^{2} n_{ij} \ln n_{ij} - 2\sum_{i=1}^{K} n_i \ln n_i$$

where N is the total sum of all frequencies, i is the identification number (ID in Table 4), j = 1,2 are the indices for left and right kanji respectively, $n_i = n_{i1} + n_{i2}$ (i.e. the marginal sum for the ith kanji) and 0 ln 0 = 0 by definition.

Using criterion (14) we obtain X^2 = 16,720.30 with 1,859 degree of freedom for the 1,945 kanji and using criterion (15) we obtain 2I = 2(76,485)ln 2 + 2(138,221.6192 + 138,093.1164) – 2(319,903.0166) = 18,854.1622. Then subtracting 1 for each zero in the table (complete Table 4 where there are 374 zeros) we obtain 2I = 18,483.16 with 1,944 degree of freedom. Though the difference between the test results seems to be great, both of them indicated the same result[3]. The kanji productivity of compounds according to position before or behind the kanji is extremely significant, suggesting an ‘asymmetric’ pattern of kanji lexical productivity on both sides. Consequently, although individual kanji may display symmetry, there is no such symmetric tendency in the whole 1,945 basic kanji.

6. Summary

The present paper investigated the symmetry of the Japanese 1,945 basic kanji when producing two-kanji compound words. The results indicated the following two findings. First, construction of kanji compounds (or kanji lexical productivity) was represented by a birth-and-death process leading to the negative binomial and/or the Waring distribution. Second, although about a half of individual kanji displayed symmetry in lexical productivity on the left-hand and right-hand sides, there is no tendency on the whole for basic kanji to produce their compound words to the same extent on the left or the right side. In summary, lexical productivity of two-kanji compound words was highly asymmetric among the 1,945 basic kanji.

[3] However, it is to be noted that this is not identical with the test for homogeneity for the left- and right-hand production.

References[4]

Altmann, G. (1985). Die Entstehung diatopischer Varianten. Ein stochastisches Modell. *Zeitschrift für Sprachwissenschaft 4, 139-155.*

Amano, N. & Kondo, K. (2000). *Nihongo-no goi tokusei [Lexical properties of Japanese]*. Tokyo: Sanseido.

Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association 43, 572-574.*

Hayashi, S. (1987). *Kanji, goi, bunshoo no kenkyuu e [Studies of kanji, words and sentences]*. Tokyo: Meiji Shoin.

Kato, M. (1989). Kakushu kanji seigenan oyobi gen 'Jooyoo Kanji' wo meguru shojikoo ichiranhyoo [A proposal for the restriction of usage of various kanji and a list of information regarding the present 'Jooyoo Kanji']. In K. Sato (ed.), *Kanji kooza Vol. 11 – Kanji to kokugo mondai [Kanji lecture series Vol. 11 – Kanji and problems of the national language]* (pp. 210-228). Tokyo: Meiji Shoin.

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

Ministry of Education, Science, Sports and Culture, Government of Japan (1987). *Shoogakkoo shidoosho – Kokugohen [The Japanese language - the course of study at elementary school]*. Osaka: Osaka Shoseki.

Ministry of Education, Science, Sports and Culture, Government of Japan (1998). *Monbushoo kokuji – Shoogakkoo gakushuu shidoo yooryoo [The announcement of the elementary school course of study by the Ministry of Education, Science and Culture, Government of Japan.]*. Tokyo: Gyosei.

Nomoto, K. (1989). Miraishakai to kanji [Kanji in the future society]. In K. Sato (ed.), *Kanji kooza Vol. 11 – Kanji to kokugo mondai [Kanji lecture series Vol. 11 – Kanji and problems of national language]* (pp. 210-228). Tokyo: Meiji Shoin.

Nomura, M. (1988). Niji kango no koozoo [Structure of two-kanji compound words]. *Nihongogaku [Study on the Japanese language]*, **7**(5), 44-55.

Nomura, M. (1989). Kanji no zoogo ryoku [Productivity of kanji]. In K. Sato (ed.), *Kanji kooza Vol. 1 – Kanji towa [Kanji lecture series Vol. 1 – What is kanji?]* (pp. 193-217). Tokyo: Meiji Shoin.

National Institute for Japanese Language (1976). *Gendai Shinbun no Kanji [Japanese kanji characters in modern newspapers]*. Tokyo: National Language Research Institute.

Stuart, A. (1955). A test for homogeneity of the marginal distribution in a two-way classification. *Biometrika 12, 412-416***.**

Tamaoka, K., Kirsner, K., Yanase, Y., Miyaoka, Y. & Kawakami, M. (2002) A Web-accessible database of characteristics of the 1,945 basic Japanese kanji. *Behavior Research Methods, Instruments & Computers 34(2), 260-275.*

Wimmer, G. & Altmann, G. (1995). A model for morphological productivity. *J. of Quantitative Linguistics 2, 212-216.*

Yasunaga, M. (1981). Jooyoo kanjihyoo ga umarerumade [A background history of the Jooyoo Kanji List]. *Gengo Seikatsu*, *355*, 24-31.

[4] In this paper, including references, an alphabetic description of Japanese names follows the commonly used Hepburn style. As the Hepburn style does not distinguish between long and short vowels (for example, the proper name of 'Kondo' is pronounced /koNdoR/ with a long vowel at the end), this paper uses the spelling of 'Kondo', not 'Kondoo'. However, to represent precise sounds, Japanese titles of research papers which include long vowels are shown by repeating the same vowels twice, such as 'oo'.

Yokosawa, K., & Umeda, M. (1988). Processes in human Kanji-word recognition. *Proceedings of the 1988 IEEE international conference on systems, man, and cybernetics* (pp. 377-380). August 8-12, 1988, Beijing and Shenyang, China.

Appendixes

Table A
Observed and computed values of the formula (6) for kanji lexical productivity (frequency counts of two-kanji compounds) on the left-hand with class pooling

X[i]	F[i]	NP[i]
0	108	99.8942
1	135	109.1996
2	109	108.4107
3	128	104.0518
4	82	98.1938
5	84	91.7526
6	84	85.1954
7	61	78.7763
8	73	72.6353
9	67	66.8469
10	59	61.4462
11	53	56.4438
12	48	51.8353
13	49	47.6072
14	36	43.7403
15	37	40.2123
16	37	36.9996
17	39	34.0784
18	33	31.4251
19	29	29.0171
20	17	26.8330
21	34	24.8526
22	28	23.0571
23	15	21.4292
24	17	19.9530
25	22	18.6137
26	21	17.3981
27	19	16.2939
28	16	15.2902
29	13	14.3770
30	16	13.5451
31	15	12.7865
32	13	12.0937
33	8	11.4602
34	11	10.8800
35	10	10.3476
36	14	9.8584
37	14	9.4079
38	16	8.9923
39	5	8.6081
40	9	8.2522
41	13	7.9218
42	6	7.6144
43	8	7.3278
44	10	7.0600
45	9	6.8092
46	8	6.5738
47	5	6.3523
48	8	6.1436
49	6	5.9463
50	5	5.7596
51	2	5.5824
52	7	5.4141
53	9	5.2537
54	3	5.1008
55	6	4.9546
56	2	4.8147
57	5	4.6805
58	3	4.5517
59	5	4.4279
60	3	4.3087
61	3	4.1939
62	3	4.0830
63	3	3.9760
64	6	3.8725
65	6	3.7723
66	3	3.6753
67	8	3.5813
68	3	3.4902
69	3	3.4017
70	3	3.3158
71	2	3.2323
72	1	3.1512
73	1	3.0723
74	0	2.9955
75	5	2.9209
76	1	2.8482
77	3	2.7774
78	2	2.7084
79	2	2.6412
80	2	2.5758
81	0	2.5120
82	4	2.4498
83	2	2.3892
84	2	2.3301
85	2	2.2725
86	1	2.2163
87	1	2.1614
88	1	2.1080
89	2	2.0558
90	0	2.0049
91	1	1.9553
92	1	1.9069
93	2	1.8596
94	1	1.8135
95	0	1.7685
96	2	1.7247
97	1	1.6818
98	1	1.6401
99	1	1.5993
100	1	1.5596
101	1	1.5208
102	1	1.4829
103	1	1.4460
104	1	1.4099
105	0	1.3748
106	4	1.3404
107	0	1.3070
108	1	1.2743
109	2	1.2425
110	1	1.2114
111	1	1.1811
112	1	1.1515
113	0	1.1226
114	1	1.0945
115	0	1.0670
116	1	1.0402
117	1	1.0141
118	4	0.9886
119	0	0.9637
120	2	0.9395
121	2	0.9158
122	1	0.8928
123	1	0.8703
124	0	0.8483
125	0	0.8269
126	1	0.8060
127	1	0.7857
128	1	0.7658
129	1	0.7464
130	1	0.7276
131	2	0.7091
132	1	0.6912
133	0	0.6737

134	0	0.6566
135	2	0.6400
136	1	0.6237
137	0	0.6079
138	1	0.5924
139	0	0.5774
140	1	0.5627
141	0	0.5484
142	1	0.5345
143	0	0.5209
144	0	0.5076
145	0	0.4947
146	1	0.4821
147	0	0.4698
148	1	0.4578
149	0	0.4461
150	0	0.4347
151	1	0.4236
152	0	0.4128
153	1	0.4023
154	0	0.3920
155	0	0.3820
156	0	0.3722
157	0	0.3627
158	0	0.3534
159	0	0.3443
160	0	0.3355
161	0	0.3269
162	1	0.3185
163	0	0.3104
164	0	0.3024
165	0	0.2947
166	0	0.2871
167	0	0.2797
168	0	0.2725
169	0	0.2655
170	1	0.2587
171	1	0.2521
172	0	0.2456
173	0	0.2393
174	0	0.2331
175	0	0.2271
176	0	0.2213
177	0	0.2156
178	0	0.2100
179	0	0.2046
180	1	0.1994
181	0	0.1942
182	0	0.1892
183	0	0.1843
184	0	0.1796
185	0	0.1749
186	0	0.1704
187	0	0.1660
188	0	0.1617
189	0	0.1576
190	0	0.1535
191	0	0.1495
192	0	0.1457
193	0	0.1419
194	0	0.1382
195	0	0.1347
196	0	0.1312
197	0	0.1278
198	0	0.1245
199	0	0.1213
200	0	0.1181
201	0	0.1151
202	0	0.1121
203	0	0.1092
204	0	0.1063
205	0	0.1036
206	0	0.1009
207	0	0.0983
208	0	0.0957
209	0	0.0933
210	0	0.0908
211	0	0.0885
212	0	0.0862
213	1	0.0839
214	0	0.0818
215	0	0.0796
216	0	0.0776
217	0	0.0756
218	0	0.0736
219	0	0.0717
220	0	0.0698
221	0	0.0680
222	0	0.0662
223	0	0.0645
224	0	0.0628
225	0	0.0612
226	0	0.0596
227	0	0.0581
228	0	0.0566
229	0	0.0551
230	0	0.0536
231	0	0.0522
232	0	0.0509
233	0	0.0496
234	0	0.0483
235	0	0.0470
236	0	0.0458
237	0	0.0446
238	0	0.0434
239	0	0.0423
240	0	0.0412
241	0	0.0401
242	0	0.0391
243	0	0.0381
244	0	0.0371
245	0	0.0361
246	1	0.0352
247	0	0.0342
248	0	0.0333
249	0	0.0325
250	0	0.0316
251	0	0.0308
252	0	0.0300
253	0	0.0292
254	0	0.0284
255	0	0.0277
256	0	0.0270
257	0	0.0263
258	0	0.0256
259	0	0.0249
260	0	0.0243
261	0	0.0236
262	0	0.0230
263	0	0.0224
264	0	0.0218
265	0	0.0213
266	0	0.0207
267	0	0.0202
268	1	0.7476

Table B

Fitting the formula (6) to kanji lexical productivity on the right-hand side

X[i]	F[i]	NP[i]
0	72	70.0911
1	91	94.3658
2	115	103.6675
3	121	105.3928
4	90	102.8910
5	93	98.0122
6	92	91.8696
7	87	85.1598
8	80	78.3235
9	61	71.6374
10	78	65.2700
11	51	59.3174
12	57	53.8274
13	49	48.8154
14	51	44.2755
15	34	40.1885
16	45	36.5269
17	32	33.2588
18	35	30.3507
19	31	27.7686
20	29	25.4797
21	22	23.4529
22	21	21.6591
23	18	20.0716
24	22	18.6662
25	11	17.4210
26	21	16.3162
27	15	15.3343
28	17	14.4599
29	12	13.6792
30	14	12.9801
31	15	12.3521
32	15	11.7858
33	17	11.2733
34	16	10.8074
35	7	10.3820
36	13	9.9920
37	13	9.6326
38	13	9.2999
39	7	8.9906
40	11	8.7016
41	11	8.4305
42	15	8.1751
43	9	7.9334
44	12	7.7040
45	4	7.4854
46	6	7.2764
47	10	7.0762
48	5	6.8837
49	7	6.6982
50	8	6.5192
51	5	6.3460
52	6	6.1783
53	4	6.0156
54	8	5.8576
55	4	5.7040
56	2	5.5545
57	4	5.4089
58	1	5.2670
59	6	5.1288
60	3	4.9939
61	5	4.8624
62	4	4.7340
63	5	4.6087
64	3	4.4864
65	0	4.3670
66	1	4.2504
67	3	4.1366
68	1	4.0255
69	2	3.9170
70	2	3.8112
71	4	3.7078
72	5	3.6070
73	3	3.5085
74	1	3.4125
75	3	3.3188
76	2	3.2273
77	6	3.1381
78	1	3.0512
79	3	2.9663
80	4	2.8836
81	0	2.8030
82	1	2.7244
83	2	2.6478
84	0	2.5732
85	0	2.5004
86	1	2.4295
87	3	2.3605
88	1	2.2933
89	0	2.2278
90	2	2.1640
91	1	2.1019
92	1	2.0415
93	0	1.9826
94	0	1.9254
95	2	1.8696
96	1	1.8154
97	1	1.7626
98	0	1.7113
99	1	1.6613
100	0	1.6127
101	1	1.5655
102	0	1.5195
103	2	1.4748
104	1	1.4314
105	0	1.3891
106	3	1.3481
107	1	1.3082
108	0	1.2694
109	1	1.2317
110	1	1.1950
111	2	1.1594
112	0	1.1248
113	0	1.0912
114	1	1.0585
115	0	1.0268
116	1	0.9959
117	0	0.9660
118	1	0.9369
119	1	0.9087
120	1	0.8813
121	0	0.8546
122	0	0.8288
123	0	0.8037
124	0	0.7793
125	0	0.7556
126	1	0.7326
127	1	0.7103
128	0	0.6887
129	0	0.6677
130	0	0.6473
131	0	0.6275
132	1	0.6083
133	0	0.5896
134	0	0.5715
135	2	0.5540
136	0	0.5370
137	2	0.5204
138	3	0.5044
139	0	0.4889
140	0	0.4738
141	0	0.4591
142	0	0.4450
143	0	0.4312
144	0	0.4178
145	1	0.4049
146	1	0.3923
147	0	0.3801
148	1	0.3683
149	1	0.3569
150	0	0.3458
151	1	0.3350
152	0	0.3245
153	0	0.3144
154	2	0.3046
155	0	0.2951
156	1	0.2859
157	1	0.2769

158	1	0.2682
159	0	0.2598
160	1	0.2517
161	0	0.2438
162	0	0.2361
163	1	0.2287
164	0	0.2215
165	0	0.2145
166	0	0.2078
167	0	0.2012
168	0	0.1949
169	0	0.1887
170	1	0.1828
171	0	0.1770
172	0	0.1714
173	0	0.1660
174	0	0.1607
175	0	0.1557
176	0	0.1507
177	0	0.1459
178	0	0.1413
179	1	0.1368
180	0	0.1325
181	0	0.1283
182	0	0.1242
183	0	0.1202
184	0	0.1164
185	0	0.1127
186	0	0.1091
187	0	0.1056
188	0	0.1023
189	0	0.0990
190	0	0.0959
191	0	0.0928
192	0	0.0898
193	0	0.0870
194	1	0.0842
195	1	0.0815
196	0	0.0789
197	0	0.0763
198	0	0.0739
199	0	0.0715
200	0	0.0692
201	0	0.0670
202	0	0.0649
203	0	0.0628
204	0	0.0608
205	0	0.0588
206	0	0.0569
207	0	0.0551
208	0	0.0533
209	0	0.0516
210	0	0.0499
211	0	0.0483
212	0	0.0468
213	0	0.0453
214	0	0.0438
215	0	0.0424
216	0	0.0410
217	0	0.0397
218	0	0.0384
219	0	0.0372
220	0	0.0360
221	0	0.0348
222	0	0.0337
223	0	0.0326
224	0	0.0315
225	0	0.0305
226	0	0.0295
227	0	0.0286
228	0	0.0276
229	0	0.0267
230	0	0.0259
231	0	0.0250
232	0	0.0242
233	0	0.0234
234	0	0.0227
235	0	0.0219
236	0	0.0212
237	0	0.0205
238	0	0.0199
239	0	0.0192
240	0	0.0186
241	0	0.0180
242	0	0.0174
243	0	0.0168
244	0	0.0163
245	0	0.0158
246	0	0.0152
247	0	0.0148
248	0	0.0143
249	0	0.0138
250	0	0.0134
251	0	0.0129
252	0	0.0125
253	0	0.0121
254	0	0.0117
255	0	0.0113
256	0	0.0109
257	0	0.0106
258	0	0.0102
259	0	0.0099
260	0	0.0096
261	0	0.0093
262	0	0.0090
263	0	0.0087
264	0	0.0084
265	0	0.0081
266	0	0.0078
267	0	0.0076
268	0	0.0073
269	0	0.0071
270	0	0.0069
271	0	0.0066
272	0	0.0064
273	0	0.0062
274	0	0.0060
275	0	0.0058
276	0	0.0056
277	0	0.0054
278	0	0.0053
279	0	0.0051
280	0	0.0049
281	0	0.0048
282	0	0.0046
283	0	0.0044
284	0	0.0043
285	0	0.0042
286	0	0.0040
287	0	0.0039
288	0	0.0038
289	0	0.0036
290	0	0.0035
291	0	0.0034
292	0	0.0033
293	1	0.0032
294	0	0.0031
295	0	0.0030
296	0	0.0029
297	0	0.0028
298	0	0.0027
299	0	0.0026
300	0	0.0025
301	0	0.0024
302	0	0.0024
303	0	0.0023
304	0	0.0022
305	0	0.0021
306	0	0.0021
307	0	0.0020
308	0	0.0019
309	0	0.0019
310	0	0.0018
311	0	0.0017
312	0	0.0017
313	0	0.0016
314	0	0.0016
315	0	0.0015
316	0	0.0015
317	0	0.0014
318	0	0.0014
319	0	0.0013
320	0	0.0013
321	0	0.0012
322	0	0.0012
323	0	0.0012
324	0	0.0011
325	0	0.0011
326	0	0.0010
327	0	0.0010
328	0	0.0010

329	0	0.0009
330	0	0.0009
331	0	0.0009
332	0	0.0009
333	0	0.0008
334	0	0.0008
335	0	0.0008
336	0	0.0007
337	0	0.0007
338	0	0.0007
339	0	0.0007
340	0	0.0007
341	0	0.0006
342	0	0.0006
343	0	0.0006
344	0	0.0006
345	0	0.0006
346	0	0.0005
347	0	0.0005
348	0	0.0005
349	0	0.0005
350	1	0.0005
351	0	0.0005
352	0	0.0004
353	0	0.0004
354	0	0.0004
355	0	0.0004
356	0	0.0004
357	0	0.0004
358	0	0.0004
359	0	0.0003
360	0	0.0003
361	0	0.0003
362	0	0.0003
363	0	0.0003
364	0	0.0003
365	0	0.0003
366	0	0.0003
367	0	0.0003
368	0	0.0003
369	0	0.0002
370	0	0.0002
371	0	0.0002
372	0	0.0002
373	0	0.0002
374	0	0.0002
375	0	0.0002
376	0	0.0002
377	0	0.0002
378	0	0.0002
379	0	0.0002
380	0	0.0002
381	0	0.0002
382	0	0.0002
383	0	0.0002
384	0	0.0001
385	0	0.0001
386	0	0.0001
387	0	0.0001
388	0	0.0001
389	0	0.0001
390	0	0.0001
391	0	0.0001
392	0	0.0001
393	0	0.0001
394	0	0.0001
395	0	0.0001
396	0	0.0001
397	0	0.0001
398	0	0.0001
399	1	0.0026

Table C
The intervals for the symmetry of compound building

n k1 k2
6, [1, 5]
7, [1, 6]
8, [1, 7]
9, [2, 7]
10, [2, 8]
11, [2, 9]
12, [3, 9]
13, [3, 10]
14, [3, 11]
15, [4, 11]
16, [4, 12]
17, [5, 12]
18, [5, 13]
19, [5, 14]
20, [6, 14]
21, [6, 15]
22, [6, 16]
23, [7, 16]
24, [7, 17]
25, [8, 17]
26, [8, 18]
27, [8, 19]
28, [9, 19]
29, [9, 20]
30, [10, 20]
31, [10, 21]
32, [10, 22]
33, [11, 22]
34, [11, 23]
35, [12, 23]
36, [12, 24]
37, [13, 24]
38, [13, 25]
39, [13, 26]
40, [14, 26]
41, [14, 27]
42, [15, 27]
43, [15, 28]
44, [16, 28]
45, [16, 29]
46, [16, 30]
47, [17, 30]
48, [17, 31]
49, [18, 31]
50, [18, 32]
51, [19, 32]
52, [19, 33]
53, [19, 34]
54, [20, 34]
55, [20, 35]
56, [21, 35]
57, [21, 36]
58, [22, 36]
59, [22, 37]
60, [22, 38]
61, [23, 38]
62, [23, 39]
63, [24, 39]
64, [24, 40]
65, [25, 40]
66, [25, 41]
67, [26, 41]
68, [26, 42]
69, [26, 43]
70, [27, 43]
71, [27, 44]
72, [28, 44]
73, [28, 45]
74, [29, 45]
75, [29, 46]
76, [29, 47]
77, [30, 47]
78, [30, 48]
79, [31, 48]
80, [31, 49]
81, [32, 49]
82, [32, 50]
83, [33, 50]
84, [33, 51]
85, [33, 52]
86, [34, 52]
87, [34, 53]
88, [35, 53]
89, [35, 54]
90, [36, 54]
91, [36, 55]
92, [37, 55]
93, [37, 56]
94, [38, 56]
95, [38, 57]
96, [38, 58]
97, [39, 58]
98, [39, 59]
99, [40, 59]
100, [40, 60]
101, [41, 60]
102, [41, 61]
103, [42, 61]
104, [42, 62]
105, [42, 63]
106, [43, 63]
107, [43, 64]
108, [44, 64]
109, [44, 65]
110, [45, 65]
111, [45, 66]
112, [46, 66]
113, [46, 67]
114, [47, 67]
115, [47, 68]
116, [47, 69]
117, [48, 69]
118, [48, 70]
119, [49, 70]
120, [49, 71]

121, [50, 71]
122, [50, 72]
123, [51, 72]
124, [51, 73]
125, [52, 73]
126, [52, 74]
127, [52, 75]
128, [53, 75]
129, [53, 76]
130, [54, 76]
131, [54, 77]
132, [55, 77]
133, [55, 78]
134, [56, 78]
135, [56, 79]
136, [57, 79]
137, [57, 80]
138, [58, 80]
139, [58, 81]
140, [58, 82]
141, [59, 82]
142, [59, 83]
143, [60, 83]
144, [60, 84]
145, [61, 84]
146, [61, 85]
147, [62, 85]
148, [62, 86]
149, [63, 86]
150, [63, 87]
151, [63, 88]
152, [64, 88]
153, [64, 89]
154, [65, 89]
155, [65, 90]
156, [66, 90]
157, [66, 91]
158, [67, 91]
159, [67, 92]
160, [68, 92]
161, [68, 93]
162, [69, 93]
163, [69, 94]
164, [69, 95]
165, [70, 95]
166, [70, 96]
167, [71, 96]
168, [71, 97]
169, [72, 97]
170, [72, 98]
171, [73, 98]
172, [73, 99]
173, [74, 99]
174, [74, 100]
175, [75, 100]
176, [75, 101]
177, [75, 102]
178, [76, 102]
179, [76, 103]
180, [77, 103]
181, [77, 104]
182, [78, 104]
183, [78, 105]
184, [79, 105]
185, [79, 106]
186, [80, 106]
187, [80, 107]
188, [81, 107]
189, [81, 108]
190, [82, 108]
191, [82, 109]
192, [82, 110]
193, [83, 110]
194, [83, 111]
195, [84, 111]
196, [84, 112]
197, [85, 112]
198, [85, 113]
199, [86, 113]
200, [86, 114]
201, [87, 114]
202, [87, 115]
203, [88, 115]
204, [88, 116]
205, [88, 117]
206, [89, 117]
207, [89, 118]
208, [90, 118]
209, [90, 119]
210, [91, 119]
211, [91, 120]
212, [92, 120]
213, [92, 121]
214, [93, 121]
215, [93, 122]
216, [94, 122]
217, [94, 123]
218, [95, 123]
219, [95, 124]
220, [95, 125]
221, [96, 125]
222, [96, 126]
223, [97, 126]
224, [97, 127]
225, [98, 127]
226, [98, 128]
227, [99, 128]
228, [99, 129]
229, [100, 129]
230, [100, 130]
231, [101, 130]
232, [101, 131]
233, [102, 131]
234, [102, 132]
235, [102, 133]
236, [103, 133]
237, [103, 134]
238, [104, 134]
239, [104, 135]
240, [105, 135]
241, [105, 136]
242, [106, 136]
243, [106, 137]
244, [107, 137]
245, [107, 138]
246, [108, 138]
247, [108, 139]
248, [109, 139]
249, [109, 140]
250, [110, 140]
251, [110, 141]
252, [110, 142]
253, [111, 142]
254, [111, 143]
255, [112, 143]
256, [112, 144]
257, [113, 144]
258, [113, 145]
259, [114, 145]
260, [114, 146]
261, [115, 146]
262, [115, 147]
263, [116, 147]
264, [116, 148]
265, [117, 148]
266, [117, 149]
267, [118, 149]
268, [118, 150]
269, [118, 151]
270, [119, 151]
271, [119, 152]
272, [120, 152]
273, [120, 153]
274, [121, 153]
275, [121, 154]
276, [122, 154]
277, [122, 155]
278, [123, 155]
279, [123, 156]
280, [124, 156]
281, [124, 157]
282, [125, 157]
283, [125, 158]
284, [125, 159]
285, [126, 159]
286, [126, 160]
287, [127, 160]
288, [127, 161]
289, [128, 161]
290, [128, 162]
291, [129, 162]
292, [129, 163]
293, [130, 163]
294, [130, 164]
295, [131, 164]
296, [131, 165]
297, [132, 165]
298, [132, 166]
299, [133, 166]
300, [133, 167]
301, [134, 167]
302, [134, 168]
303, [134, 169]
304, [135, 169]
305, [135, 170]
306, [136, 170]
307, [136, 171]
308, [137, 171]
309, [137, 172]
310, [138, 172]
311, [138, 173]
312, [139, 173]
313, [139, 174]
314, [140, 174]
315, [140, 175]
316, [141, 175]
317, [141, 176]
318, [142, 176]
319, [142, 177]
320, [142, 178]
321, [143, 178]
322, [143, 179]
323, [144, 179]
324, [144, 180]
325, [145, 180]
326, [145, 181]
327, [146, 181]
328, [146, 182]
329, [147, 182]
330, [147, 183]
331, [148, 183]
332, [148, 184]
333, [149, 184]
334, [149, 185]
335, [150, 185]
336, [150, 186]
337, [151, 186]
338, [151, 187]
339, [151, 188]
340, [152, 188]
341, [152, 189]
342, [153, 189]
343, [153, 190]
344, [154, 190]
345, [154, 191]
346, [155, 191]
347, [155, 192]
348, [156, 192]

349, [156, 193]
350, [157, 193]
351, [157, 194]
352, [158, 194]
353, [158, 195]
354, [159, 195]
355, [159, 196]
356, [160, 196]
357, [160, 197]
358, [160, 198]
359, [161, 198]
360, [161, 199]
361, [162, 199]
362, [162, 200]
363, [163, 200]
364, [163, 201]
365, [164, 201]
366, [164, 202]
367, [165, 202]
368, [165, 203]
369, [166, 203]
370, [166, 204]
371, [167, 204]
372, [167, 205]
373, [168, 205]
374, [168, 206]
375, [169, 206]
376, [169, 207]
377, [169, 208]
378, [170, 208]
379, [170, 209]
380, [171, 209]
381, [171, 210]
382, [172, 210]
383, [172, 211]
384, [173, 211]
385, [173, 212]
386, [174, 212]
387, [174, 213]
388, [175, 213]
389, [175, 214]
390, [176, 214]
391, [176, 215]
392, [177, 215]
393, [177, 216]
394, [178, 216]
395, [178, 217]
396, [179, 217]
397, [179, 218]
398, [179, 219]
399, [180, 219]
400, [180, 220]

Glottometrics 7, 2004, 85-90

Pluralallomorphe in Briefen Heinrich von Kleists

Nina Brüers, Anne Heeren[1]

Abstract. The paper brings a further corroboration of the hypothesis that entities of linguistic classes are rank-ordered. Here the geometric distribution has been fitted to the ranked distribution of plural morphemes in letters of H. von Kleist.

Keywords: Diversification, geometric distribution, German

1. Einleitung: Die Verteilung von Pluralallomorphen in deutschen Texten

Die vorliegende Untersuchung entstand im Rahmen des „Göttinger Projekts zur Quantitativen Linguistik“, welches sich mit der Suche nach Sprachgesetzen und ihrer Überprüfung beschäftigt. Dabei werden hauptsächlich die Häufigkeitsverteilungen sprachlicher Einheiten in Texten untersucht. Eine der bekanntesten Erscheinungen dieser Art ist die Diversifikation sprachlicher Entitäten, die durch phonetische, morphologische, semantische, dialektale usw. Variation zustande kommt. Nach einer in der quantitativen Linguistik verbreiteten Hypothese werden Varianten nicht gleichmäßig verwendet, sondern bilden die Taxate, geordnet nach Rang, eine reguläre Verteilung. Durch einen Rückschluss kann man sogar feststellen, ob eine Klasse von Erscheinungen „korrekt“ etabliert wurde (vgl. Rothe 1991). Als ein Beispiel zur formalen Diversifikation führt Best (2001a: 82) die unterschiedliche Realisierung des Plurals der deutschen Substantive an, welche er anhand der Daten eines Textes veranschaulicht.

Um die Diversifikationshypothese und Bests Spezialfall weiter zu prüfen, wurden die neun Pluralmorphe des Deutschen anhand von Briefen Heinrich von Kleists (1777-1811) aus den Jahren 1800 bis 1801 die Häufigkeit ihres Vorkommens ausgewertet.

Für die Untersuchungen wurden Briefe ausgewählt, da diese im Allgemeinen von nur einem Autor stammen, der den Text ohne Unterbrechung verfasst hat. Sie zeichnen sich durch einen relativ hohen Grad an Spontaneität und thematischer Geschlossenheit aus. Gerade wegen dieser Homogenität (vgl. Altmann 1992) erscheinen Briefe als gut geeignet für quantitative Bearbeitungen. Auf der anderen Seite ist der Umfang der Stichproben etwas zu klein, was aber das Testen nicht beeinträchtigt hat.

Die Adressaten der behandelten Briefe sind Kleists Verlobte Wilhelmine von Zenge, seine Halbschwester Ulrike von Kleist sowie Ludwig von Brockes und Adolfine von Werdeck. Das Ungewöhnliche an den Briefen Kleists ist, dass sie sich stilistisch stark von den Novellen des Dichters unterscheiden. Während diese zu lesen ein großes Vergnügen bereitet, befremden die Briefe durch ihren Ton. Gerade jene an Ulrike stecken voller Überheblichkeit, Arroganz und Zynismus. Diese Tatsachen ließen die Briefe für uns interessant erscheinen.

Die neun Pluralallomorphe sind: {-e, -en, -er, -n, -s, -0, Umlaut, Umlaut + -e, Umlaut + -er}. Bei dem Nullallomorph sind keine Unterschiede zwischen Singular und Plural des Substantivs vorhanden, wie an dem Beispiel von *Mädchen* (Sg. und Pl.) erkennbar ist. Das

[1] Address correspondence to: Nina Brüers, Seminar für deutsche Philologie, Käte-Hamburger-Weg 3, D-37073 Göttingen, Germany

Allomorph -{s} tauchte in den insgesamt 25 Briefen nur einmal auf, als Endung des Fremdwortes *Tableau*.

2. Das Verteilungsmodell

Die Kürze der Briefe hatte den Nachteil, dass die von Best angewandte Hyperpoisson-Verteilung nicht in allen Fällen angewandt werden konnte. Sie enthält zwei Parameter, so dass in einigen Fällen keine Freiheitsgrade übrig blieben. Da die Hyperpoisson-Verteilung gegen die geometrische Verteilung, die nur einen Parameter hat, konvergiert (vgl. Wimmer, Altmann 1999), wurde diese in allen Fällen verwendet. Es wurde die in der Rangierung übliche 1-verschobene Form

$$P_x = pq^{x-1},\ x = 1,2,3,... \qquad (x \text{ ist der Rang})$$

verwendet. Der letzte theoretische Wert umfasst alle größeren Ränge bis ∞. Vier von den ausgezählten Briefe enthielten zu wenig verschiedene Pluralallomorphe, um überhaupt getestet werden zu können, zum Beispiel ein Brief an Wilhelmine von Zenge aus Berlin vom 28. März 1801, in dem lediglich ein Allomorph dreimal auftauchte.

Die Berechnungen wurden mit dem Altmann-Fitter (1997) durchgeführt. Die Anpassung der Verteilung an die Daten wird als zufriedenstellend betrachtet, wenn $P \geq 0.05$ ist.

In den Tabellen bedeutet:

x	: den Rang der Pluralallomorphklasse
n_x	: die beobachtete Häufigkeit, mit der die Pluralallomorphklassen im Text vorkommen
NP_x	: die nach der 1-verschobenen geometrischen Verteilung berechnete Häufigkeit
X^2	: das Chiquadrat
FG	: die Freiheitsgrade
P	: die Überschreitungswahrscheinlichkeit des Chiquadrats
U	: Umlaut

3. Anpassung des Modells an die Briefdateien

Die Anpassung der 1-verschobenen geometrischen Verteilung an die Daten der 21 Briefe (in Streller 1978) von Heinrich von Kleist hat folgende Ergebnisse erbracht:

Brief 1
Würzburg, 13. September 1800, an Wilhelmine von Zenge

Rang	x	n_x	NPx
1	-{n}	17	15.92
2	-{e}	7	9.42
3	-{en}	6	5.58
4	U + -{e}	5	3.30
5	-{0}	3	1.95
6	-{er}	1	2.83

$p = 0.4081$ $FG = 4$
$X^2 = 3.350$ $P = 0.50$

Brief 2
Würzburg, 14. September 1800, an Wilhelmine von Zenge

Rang	x	n_x	NPx
1	-{n}	8	8.97
2	-{en}	6	5.75
3	U + -{er}	4	3.69
4	U + -{e}	3	2.36
5	-{er}	2	1.52┐
6	-{e}	1	0.97┘
7	-{0}	1	1.74

$p = 0.3588$ $FG = 4$
$X^2 = 0.731$ $P = 0.95$

Brief 3
Würzburg, 15. September 1800, an Wilhelmine von Zenge

Rang	x	n_x	NPx
1	-{e}	10	7.84
2	-{n}	3	4.61
3	U + -{e}	2	2.70
4	-{en}	1	1.59┐
5	-{0}	1	0.93┘
6	U + -{er}	1	0.55┐
7	-{er}	1	0.78┘

$p = 0.4128$ $FG = 3$
$X^2 = 1.786$ $P = 0.62$

Brief 4
Würzburg, 18. September 1800, an Wilhelmine von Zenge

Rang	*x*	n_x	*NPx*
1	-{e}	6	6.39
2	-{en}	3	2.31
3	-{n}	1	1.31
p = 0.6385			FG = 1
X^2 = 0.303			P = 0.58

Brief 5
Würzburg, 19. September 1800, an Wilhelmine von Zenge

Rang	*x*	n_x	*NPx*
1	-{e}	12	14.77
2	-{n}	12	8.54
3	-{en}	4	4.94
4	U + -{er}	4	2.85
5	U + -{e}	2	1.65
6	-{0}	1	2.26
p = 0.4219			*FG* = 4
X^2 = 3.339			*P* = 0.50

Brief 6
Liechstal, 23. Dezember 1801, an Heinrich Lohse

Rang	*x*	n_x	*NPx*
1	-{n}	6	6.22
2	-{e}	3	2.30
3	U + -{e}	2	1.44
4	-{0}	1	1.34
p = 0.5181			*FG* = 2
X^2 = 0.309			*P* = 0.86

Brief 7
Berlin, 27. Oktober 1800, an Ulrike von Kleist

Rang	*x*	n_x	*NPx*
1	-{0}	5	6.11
2	-{e}	4	3.62
3	-{n}	2	2.15
4	-{en}	2	3.12
p = 0.4073			*FG* = 2
X^2 = 2.246			*P*= 0.32

Brief 8
Berlin, 22. November 1800, an Wilhelmine von Zenge

Rang	*x*	n_x	*NPx*
1	U + -{e}	9	8.71
2	-{en}	3	3.65
3	-{e}	2	1.53
4	-{n}	1	1.11
p = 0.5806			*FG* = 2
X^2 = 0.280			*P* = 0.87

Brief 9
Berlin, 25. November 1800, an Ulrike von Kleist

Rang	*x*	n_x	*NPx*
1	-{en}	11	11.33
2	-{e}	7	7.44
3	U + -{e}	5	4.89
4	-{0}	4	3.21
5	-{n}	4	2.11
6	U + -{er}	2	4.03
p = 0.3433			*FG* = 4
X^2 = 2.957			*P* = 0.57

Brief 10
Berlin, 21. Januar 1801, an Wilhelmine von Zenge

Rang	*x*	n_x	*NPx*
1	-{e}	6	6.90
2	-{en}	4	4.10
3	-{n}	4	2.44
4	-{0}	2	1.45
5	U + -{er}	1	2.12
p = 0.4061			*FG* = 3
X^2 = 1.926			*P* = 0.59

Brief 11
Berlin, 23. März 1801, an Ulrike von Kleist

Rang	*x*	n_x	*NPx*
1	-{e}	5	4.46
2	-{en}	2	3.13
3	-{0}	2	2.20
4	-{n}	2	1.55
5	U + -{e}	2	1.09
6	U + -{er}	2	2.57
p = 0.2975			*FG* = 4
X^2 = 1.519			*P* = 0.82

Brief 12
Berlin, 1. April 1801, an Ulrike von Kleist

Rang	*x*	n_x	*NPx*
1	-{0}	16	16.24
2	-{e}	10	9.30
3	-{n}	4	5.33
4	-{er}	3	3.05
5	U + -{er}	2	1.75
6	U + -{e}	2	1.00
7	-{en}	1	1.34
p = 0.4274			*FG* = 4
X^2 = 1.046			*P* = 0.90

Brief 13
Berlin, 14. April 1801, an Wilhelmine von Zenge

Rang	x	n_x	NPx
1	-{n}	8	9.11
2	-{e}	6	4.96
3	-{en}	3	2.70
4	U + -{e}	2	1.47
5	-{0}	1	1.76

$p = 0.4554$ $FG = 3$

$X^2 = 0.904$ $P = 0.82$

Brief 14
Dresden, 4. Mai 1801, an Wilhelmine von Zenge

Rang	x	n_x	NPx
1	-{e}	4	5.35
2	U + -{e}	4	3.30
3	-{n}	3	2.04
4	-{en}	2	1.26
5	U + -{er}	1	2.04

$p = 0.3818$ $FG = 3$

$X^2 = 1.898$ $P = 0.59$

Brief 15
Sraßburg, 28. Juni 1801, an Wilhelmine von Zenge

Rang	x	n_x	NPx
1	-{e}	7	8.00
2	-{n}	5	3.73
3	-{0}	2	1.74
4	U + -{e}	1	1.52

$p = 0.5334$ $FG = 2$

$X^2 = 0.773$ $P = 0.68$

Brief 16
Paris, 27. Oktober 1801, an Wilhelmine von Zenge

Rang	x	n_x	NPx
1	-{en}	6	6.17
2	-{e}	4	3.79
3	-{0}	2	2.33
4	-{n}	2	1.43
5	U + -{e}	2	2.28

$p = 0.3858$ $FG = 3$

$X^2 = 0.323$ $P = 0.96$

Brief 17
Paris, November 1801, (wahrscheinlich) an Ludwig von Brockes

Rang	x	n_x	NPx
1	U + -{e}	7	8.47
2	-{n}	6	4.88
3	-{en}	4	2.82
4	-{e}	3	3.83

$p = 0.4234$ $FG = 2$

$X^2 = 1.190$ $P = 0.55$

Brief 18
Paris, November 1801, an Adolfine von Werdeck

Rang	x	n_x	NPx
1	-{en}	18	20.35
2	-{e}	17	13.98
3	-{n}	10	9.60
4	-{s}	5	6.60
5	U + -{e}	5	4.53
6	-{0}	4	3.11
7	U + -{er}	4	2.14
8	-{er}	2	4.69

$p = 0.3131$ $FG = 6$

$X^2 = 4.794$ $P = 0.57$

Brief 19
Frankfurt a.M., 29. November 1801, an Adolfine von Werdeck

Rang	x	n_x	NPx
1	-{en}	3	4.25
2	-{e}	3	2.86
3	-{n}	3	1.93
4	U + -{e}	3	1.30
5	U	1	2.67

$p = 0.3270$ $FG = 3$

$X^2 = 4.258$ $P = 0.23$

Brief 20
Basel, 16. Dezember 1801, an Ulrike von Kleist

Rang	x	n_x	NPx
1	-{e}	7	6.76
2	-{en}	2	2.95
3	-{0}	1	1.29
4	-{n}	1	0.56
5	U + -{e}	1	0.44

$p = 0.5632$ $FG = 1$

$X^2 = 0.536$ $P = 0.46$

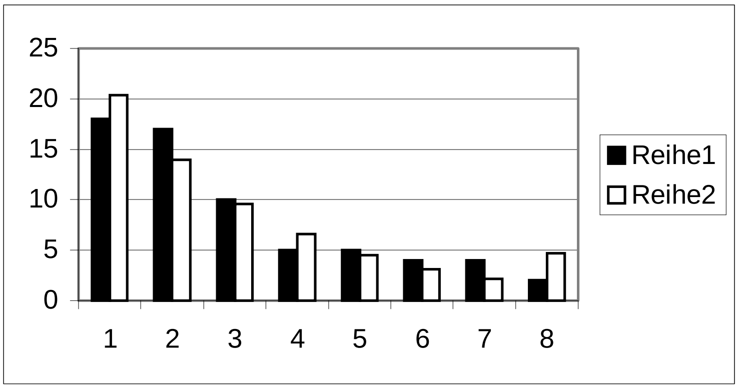
Brief 21
Würzburg, 20. September 1800, an Wilhelmine von Zenge

Rang	x	n_x	NPx
1	-{n}	23	21.14
2	-{en}	10	13.57
3	-{e}	8	8.70
4	U + -{e}	8	5.59
5	-{0}	6	3.58
6	U + -{er}	4	6.42

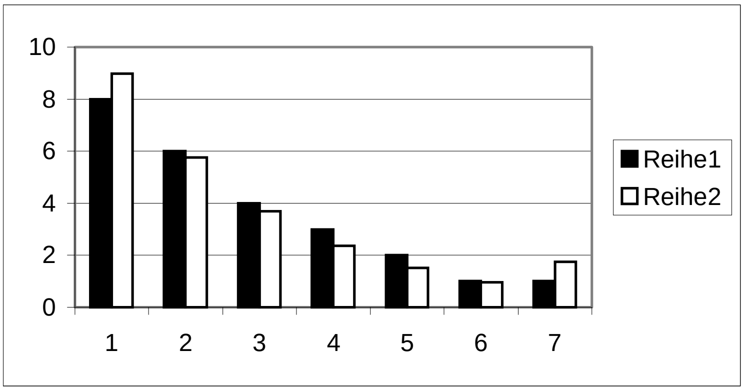
$p = 0.3584$ $FG = 4$

$X^2 = 4.741$ $P = 0.31$

Zur Veranschaulichung sollen abschließend zwei Anpassungen auch graphisch vorgestellt werden:



Beispielgraphik zu Brief 18



Beispielgraphik zu Brief 2

4. Zusammenfassung

Abschließend lässt sich feststellen, dass sich alle ausgezählten Texte mit der 1-verschobenen geometrischen Verteilung modellieren lassen und diese somit ein geeignetes Modell für die Untersuchung der Verteilung von Pluralallomorphen der Substantive des Deutschen darzustellen scheint. Allerdings lässt sich dieses Ergebnis noch nicht verallgemeinern, da es bis dato keine weiteren Untersuchungen zu diesem Thema gibt.

Literatur

Altmann, G. (1992). Das Problem der Datenhomogenität. *Glottometrika 13*, *287-298.*

Best, K.-H. (2001a). *Quantitative Linguistik: Eine Annäherung*. Göttingen: Peust & Gutschmidt.

Best, K.-H. (2001b). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt.

Rothe, U. (Hrsg.), *Diversification Processes in Language: Grammar*. Hagen: Rottmann.

Streller, S. (1978). *Heinrich von Kleist. Werke und Briefe, Band 4.* Berlin und Weimar: Aufbau Verlag.

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

www.hagenstedt.de/rezensionen/25Kleist.html

Software:

Altmann-Fitter (1997). *Iterative Fitting of Probability Distributions.* Lüdenscheid: RAM-Verlag.

Glottometrics 7, 2004, 91-96

History of Quantitative Linguistics

Since a historiography of quantitative linguistics does not exist as yet, we shall present in this column short statements on researchers, ideas and findings of the past – usually forgotten – in order to establish a tradition and to complete our knowledge of history. Contributions are welcome and should be sent to Peter Grzybek, peter.grzybek@uni-graz.at.

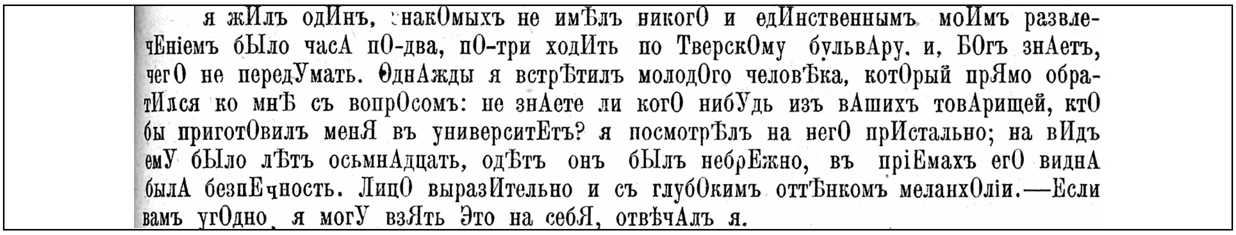
III. Nikolaj Gavrilovič Černyševskij – A Forerunner of Quantitative Stylistics in Russia

Nikolaj Gavrilovič Černyševskij (1828-1889) is one of the best known Russian journalists, writers, and literary critics of the 19th century. He is also known for his philosophical and economical writings, with which he became one of the most important theoreticians of radical social democracy. He is less known, however, for having introduced quantitative arguments to the study of poetic language. Although the relevant passages are rather scarce, they deserve mentioning, since ultimately, they make Černyševskij a forerunner of quantitative stylistics.

Being the son of a Russian priest from Saratov, Černyševskij was a typical representative of the so-called *raznočincy*, a member of the young generation from the middle class who, in the Tsarist Russia of the mid-nineteenth century, started fighting for social rights and democracy. From 1842 to 1846, Černyševskij was educated in the local priest seminar, before he went to St. Petersburg to enter the historical-philological faculty of St. Petersburg University. In 1846, he passed the entrance exams in mathematics, physics, logic, and literature, as well as in Latin and French, and he was accepted a student.

Having finished his university career, Černyševskij returned to Saratov and became a senior teacher at the secondary school of his home town. In 1853, after his marriage, he decided to go back to Petersburg, where he started a journalistic career. In the very same year, Černyševskij began to cooperate with two important journals of that time, the *Sovremennik* (N.A. Nekrasov), on the one hand, and the *Otečestvennye Zapiski* (A.A. Kraevskij), on the other. Černyševskij soon concentrated on the *Sovremennik* only, from 1854 being responsible for the criticism and bibliography sections. In 1856, he handed the criticism section to N.A. Dobroljubov, himself concentrating on politics, philosophy, and economy. He was a fervent defender of a materialist aesthetic theory, best expressed in his M.A. Thesis "*Esthetic Relations of Art to Reality*. His successful defence of this thesis in 1855 caused public uproar, and it was accepted by the ministry only three years later. By and by, Černyševskij – who was much influenced by social utopians like Saint Simon, Fourier, and Feuerbach – became somewhat like a leader of radical and social parts of society; as a consequence, he soon attracted the attention of Secret Police. In 1862, *Sovremmenik* was closed for eight months, Černyševskij was arrested and imprisoned in the Fortress of Sts. Peter and Paul. During imprisonment, he wrote his famous novel *Što delat'* [What is to be Done], which was allowed to be published by mistake in 1863. Černyševskij was finally sentenced in 1864, deprived of all rights, and sent to Siberia. He was allowed to return to the European part of Russia only in 1883. In 1889, he came back to his home town Saratov, where he died the same year.

As to the introduction of quantitative arguments to literary criticism, one can say that Černyševskij indeed was the first to bring up the question as to the relation of stressed and unstressed syllables in poetry and prose. The question was raised by Černyševskij (1855) in the journal *Sovremennik*, when he discussed P.V. Annenkov's (1812-87) edition of A.S. Puškin's *Collected Works*. In his critique, Černyševskij was struck by the observation that the vast majority of poems were characterized by a binary meter: 179 poems written in iambs (xx́), followed by 29 poems in trochees (x́). Only very few poems included ternary meters, i.e. 7 poems with amphibrachs (xx́x), 6 with dactyls (x́xx), and only one with anapests (xxx́). Given the clear dominance of binary meter, Černyševskij (1855: 286) raised the question if the iamb might be "the most natural meter for the Russian language". As a first indication that this might not be the case, Černyševskij objected that in Russian, as compared to German, words rather tend to be multi-syllabic, that prepositions and pronouns usually do not have any stress, and that all other words have one stress only. Černyševskij (1855: 287) then submitted this hypothesis to an empirical test: Counting stressed and unstressed syllables in three randomly chosen prose passages by A.F. Pisemskij (1821-1881), Černyševskij obtained a relations of 66 stressed vs. 193 unstressed syllables:



я жИлъ одИнъ, знакОмыхъ не имѢлъ никогО и едИнственнымъ моИмъ развлечЕніемъ бЫло часА пО-два, пО-три ходИть по ТверскОму бульвАру. и, БОгъ знАетъ, чегО не передУмать. ѲднАжды я встрѢтилъ молодОго человѢка, котОрый прЯмо обратИлся ко мнѢ съ вопрОсомъ: не знАете ли когО нибУдь изъ вАшихъ товАрищей, ктО бы приготОвилъ менЯ въ университЕтъ? я посмотрѢлъ на негО прИстально; на вИдъ емУ бЫло лѢтъ осьмнАдцать, одѢтъ онъ бЫлъ небрЕжно, въ пріЕмахъ егО виднА былА безпЕчность. ЛицО выразИтельно и съ глубОкимъ оттѢнкомъ меланхОліи.—Если вамъ угОдно, я могУ взЯть Это на себЯ, отвѢчАлъ я.

In two other passages, he found similar relations of 25 vs. 75, and 27 vs. 83. Although these passages are rather short, in all three cases (as well as in the combination of them) the relation of stressed and unstressed syllables displays an "astonishing closeness to each other", which is characterized by an almost constant proportion of 1:3. This caused Černyševskij (1855: 287) to "draw the inevitable conclusion that, in Russian, iambs and trochees, which demand 15 stresses within 30 syllables, are by far not as natural a dactyls, amphibrachs, or anapests, which demand 10 stresses within 30 syllables.

Černyševskij's innovative approach remained forgotten for a long time. Almost 20 years later, Dmitrij A. Averkiev (1836-1905), a well-known writer, literary critic and journalist, referred to Černyševskij's ideas in his own ruminations *On Drama* (first published in the journal *Russkij vestnik* in 1877/78, later edited in book form in 1893 and 1907). Averkiev had graduated from the department of natural sciences of the physical-mathematical faculty (St. Petersburg University) in 1859, where he had close contact with important persons from the literary scene, such as Apollon A. Grigor'ev (1822-64) or Nikolaj N. Strachov (1828-96) who, curiously enough, also had studied and taught mathematics in St. Petersburg. In his ruminations on dramatic language, Averkiev harshly criticized the fundamentals of Černyševskij's approach. According to him, "clean iambs or trochees are a rarity in Russian: "In iambs or trochees with four ictuses, four stresses are a great exception; rather, there are two or three of them, and in iambs with five ictuses, three or four. Thus, there are two or three stresses within 8 or 9 syllables, three or four within 10 or 11 syllables, and four or five within 12 or 13 syllables." With reference to Černyševskij's analysis, Averkiev thus concluded that in all cases, these are "just as many as the author of that unfortunate article obtained as a result of his study on Russian prose".

It would take some decades more, until Boris V. Tomaševskij (1919/23) would pick up the problem and discuss it critically from a methodological point of view. Anyway, N.G. Černyševskij can be credited for having pointed out that the problem of poetic language is not an isolated one, but should be studied in comparison to prosaic language, and for having done a first step towards an empirical, quantitative study of this problem.

References

Averkiev, D.A. (1877/78). *O drame*. St. Peterburg, 1907.

Černyševskij, N.G. (1855a). „[Rez.:] Sočinenija Puškina“. In: Ibd., *Polnoe sobranie sočinenij*. *Tom I: Kritika i bibliografija.* St. Peterburg, 1906. (245-330).

Tomaševskij, B.V. (1919/23). „Pjatistopnyj jamb Puškina.“ In: Ibd., *O stiche*. Leningrad, 1929. (138-253).

Peter Grzybek

IV. Anton Semënovič Budilovič (1846-1908) – A Forerunner of Quantitative Linguistics in Russia?

In the history of quantitative linguistics, Anton Semënovič Budilovič (1846-1908) has repeatedly been referred to. Papp, for example, in his well-known monography on *The History of Mathematical Linguistics in the Soviet Union* (1966), presents Budilovič as an important 19th century advocate of statistical-mathematical methods in Russia. In fact, Papp credits Budilovič for having published the first letter frequency statistics of Old Church Slavonic (OCS), as published in his *Sketch of a Church Slavonic Grammar With Regard to a General Theory of Russian and Other Related Languages* [Načertanie cerkovnoslavjanskoj grammatiki priměnitel'no k obščej teorii russkago i drugich rodstvennych jazykov], published by Budilovič in 1883 (not 1881, as wrongly quoted by Papp). Contemporary scholars, too, such as Kempgen (1995, 1999), or Grzybek/Kelih (2004) continue to refer to this work. Given these facts, it seems worthwhile taking a somewhat closer look at the ouevre and activities of this Russian philologist, linguist and cultural theoretician.

A.S. Budilovič was born on May 24, 1846 in Komotovo (now in Belorussia). Having finished the Lithuanian Priest seminar, he studied at the Historical-Philological Faculty of St. Petersburg University. This was the time, when I.I. Sreznvevskij (1812-1880) was a professor for Russian linguistics here, who was well familiar with statistical methods (cf. Grzybek/Kelih 2004), and who probably is responsible for having aroused an interest in these methods in A.S. Budilovič (cf. Karskij 1909: 151). In 1871, he defended his dissertation, *A Linguistic Study of the Old Slavonic Translation of Grigorij Bogoslov's 13th Word on the Basis of the Manuscript of the 16th century Tsarist Public Library* [Isslědovanie jazyka drevneslavjanskago perevoda XIII slov Grigorija Bogoslova po rukopisi Imperatorskoj Publičnoj biblioteki XI veka]. During his subsequent journeys to countries of the Austrian-Hungarian monarchy, to Germany and Turkey, he collected statistical material about the Slavic inhabitants of these countries (cf. Budilovič 1875a, b)[1]. This ethnographic material as such, and the facts and data it provides, are of high value, still today. Additionally, the way Budilovič, while being Dean and Rector of Warsaw University from 1881 to 1892, applied statistical methods upon these data, deserves attention. Thus, in 1883, in the above-mentioned *Načertanie*, he published a list of OCS graphemes which often is held to be the first grapheme statistics in this linguistic field. In fact, a first summarizing table of letter OCS letter frequencies can be found here, separately for vowels and consonants. The percentual frequencies of the vowels can be taken from Table 1 (cf. Budilovič 1883: 67):

Vowels	%	Vowels	%
i	20,4	ê	6,1
ß	13,9	ì	6
e	13,4	ã	4,3
o	13,4	õ	4,3
a	12,7	y	3
u	2,5		

[1] Unfortunately, we have not yet had the chance of getting access to the material quoted ad oculos. The complete bibliography of A.S. Budilovič's works can be found in *Novyj sbornik statej po slavjanověděniju, sostavlennyj i izdannyj učenikami V.I. Lamanskago pri učastii ich učenikov* (St. Petersburg, 1905).

Likewise, the percentual frequencies of consonants can be taken from Table 2: (ibd., 97):

Consonants	%	Consonants	%
t	10,7	w	5,1
v	9,1	g	4,3
n	8,7	k	3,9
s	8,7	<	3,9
й	6,9	b	3,5
r	6,8	p	3,2
d	6,4	c	2,2/0,8*
m	6,3	j	1,8
l	6,2	x	1,5
÷	0,8/2,2*		

As has been described elsewhere (cf. Grzybek/Kelih 2004), it is important to note that we are not concerned here with original data. Though this is, in fact, the first statistics of OCS graphemes published in Russian, the data are taken from an analysis by August Schleicher, who presented them in his 1852 study *Die Formenlehre der kirchenslawischen Sprache, erklärend und vergleichend dargestellt*. Unfortunately, Budilovič confused the percentage of two consonants; the original data given by Schleicher are marked by an asterisk in Table 2. Interestingly enough, Schleicher himself explicitly understood his study as a complement to previous studies by Ernst Förstemann (1846, 1852), who had established corresponding statistics on Greek, Latin and Gothic sounds (as he said). As to OCS, Schleicher's (1852: 19) calculations are based on merely orthographical criteria („…überhaupt alles auf die organische orthographie reducirt…"). Analyzing different sentences from the Ostromir Gospel (omitting proper names), the percentual relation of vowels and consonants equals 50.50% vs. 49.50%. Subsequently comparing these data to Latin, Greek and Gothic, Schleicher arrives at no consistent interpretation, since the percentual differences between vowels and consonants of more or less closely related languages results in no common picture. As opposed to this, A.S. Budilovič (1883:117) suggest an interpretation, saying that, due to the balanced proportions of vowels and consonants, OCS fulfills all conditions of a „high-standard musical evolution" („hohe musikalische Entwicklung"), and therefore is not inferior as compared to other Indo-European languages.

Obviously, no further applications of (linguo-)statistical publications can be traced back to A.S. Budilovič's work. The reason for this seems to be his primary interest in cultural and ethnographic questions (ethnogenesis of the Slavs, the Greek-Byzantine influence on OCS, etc.); also he became increasingly involved in administrative and political activities. Thus, he was Rector of Tartu University from 1892 to 1901, where he played an important role in the russification of Dorpat University, which was renamed to Jur'ev in 1883 (cf. Siilivask 1985: 150). Subsequently, until his death on December 12, 1908, Budilovič's activities are characterized by a strong political and journalistic engagement, being the editor of two important journals, *Slavjanskoe obozrenie* and *Moskovskie vedomosti*.

It seem to be these activities, which are responsible for the fact that today, at least in Russian speaking circles, A.S. Budilovič is better known for his Slavophile concepts of language and culture, rather than for his knowledge of quantitative approaches (cf. Smirnov/ Dmitriev/Safronov 1991:10, Dobrodomov 1996).

References

Budilovič, A.S. (1875a): Něskol'ko dannych i soobraženii iz oblasti obščestvennoj i ėkonomičeskoj statistiki Čechii, Moravii i Avstrijskoj Silezii za poslědnie gody. In: *Slavjanskij sbornik, I.*

Budilovič,A.S. (1875b): *Statističeskija tablicy raspredělenija slavjan: a) po gosudarstvam i narodnostjam i b) po věroispovědanijam, azbukam i literaturnym jazykam (naręčijam), s ob''jasnitel'noju zapiskoju.* Sankt Peterburg.

Budilovič, A.S. (1883): *Načertanie cerkovnoslavjanskoj grammatiki, priměnitel'no k obščej teorii russkago i drugich rodstvennych jazykov.* Varšava.

Dobrodomov, I.G. (1996): Anton Semenovič Budilovič. In: *Russkij jazyk v škole 2, 102-105.*

Förstemann, E. (1852): Numerische Lautverhältnisse im Griechischen, Lateinischen und Deutschen. In: *Germanische Zeitschrift für Vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen 1, 163-179.*

Förstemann, E. (1846): Über die nummerischen Lautverhältnisse im Deutschen. In: *Germania,* (Hrsg. von der Berlinischen Gesellschaft für deutsche Sprache und Altertumskunde); *Bd. 7. 83-90.*

Grzybek, P., Kelih, E. (2003): Graphemhäufigkeiten (am Beispiel des Russischen). Teil 1: Methodologische Vor-Bemerkungen und Anmerkungen zur Geschichte der Erforschung von Graphemhäufigkeiten im Russischen. In: *Anzeiger für Slavische Philologie (XXXI).* 131-162.

Grzybek, P. Kelih, E. (2004): Zur Vorgeschichte quantitativer Ansätze in der russischen Sprach- und Literaturwissenschaft. In: *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch/An International Handbook* (Herausgegeben von Gabriel Altmann; Reinhard Köhler, R. Piotrowski). New York: de Gruyter, 2004. [= Handbücher zur Sprach- und Kommunikationswissenschaft]

Karskij, E. (1909): Pamjati A.S. Budiloviča. In: *Russkij filologičeskij vestnik, LXI,* 149-161.

Kempgen, S. (1995): *Russische Sprachstatistik. Systematischer Überblick und Bibliographie.* München. [= Vorträge und Abhandlungen zur Slavistik, Band 26]. München.

Schleicher, August (1852): Die Formenlehre der kirchenslawischen Sprache, erklärend und vergleichend dargestellt. *Bonn.*

Papp, F. (1966): *Mathematical Linguistics in the Soviet Union.* [= Janua Linguarum, Series Minor, XL]. The Hague.

Siilivask, K. (ed.) (1985): *History of Tartu University 1632-1982.* Tallinn.

Smirnov, S.V., Dmitriev, P.A., Safronov, G.I. (1991): *Russkoe i slavjanskoe jazykoznanie v Rossii serediny XIX - načala vv.* Leningrad.

Peter Grzybek, Emmerich Kelih

Book review

Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S., Wimmerová, S., *Úvod do analýzy textov* [*Introduction to Text Analysis*]. Bratislava: Veda 2003, 344 pp.

Although the research on quantitative text analysis is about 150 years old and the number of research papers in this domain is enormous (cf. Köhler 1995), there has not been any textbook on the subject until now. Since the research is very variegated, the authors were forced to restrict the scope and bring merely selected themes like: phonics, word, sentence, content analysis, supra-sentence structures, sequential structures, text comparison and denotative text analysis.

The book concentrates on setting up models of textual phenomena. All models are derived, thoroughly explained and exemplified in detail using data from different languages. At the end of each sub-chapter one finds tasks that stimulate the reader to develop the given problems further or to organize research projects. The book was designed for teaching purposes. It can be mastered in one or two semesters and even autodidacts with moderate mathematical knowledge can use it.

The book begins with a very extensive exposé of the problems of quantitative analysis, its individual steps, aims and elimination of possible objections. Then the methodological background of this kind of analysis is explicated (units, data, samples, properties, indices, hypotheses, laws, kinds of organization, modeling, theory, science). As a matter of fact, this part is an excellent introduction to the philosophy of science as applied to text analysis. It shows the philosophical background, the starting point of quantitative linguistics striving for a theory. Of course, the book uses mathematical methods, without which no modeling, no testing and no theory building are possible.

For each problem in the following chapters there is at least one statistical testing procedure but in the majority of cases the authors provide models, try to derive them from plausible assumptions and propose different law-candidates. Some of them are already known, e.g. the power law, Zipf-Mandelbrot law etc., but the authors themselves have derived many of them. As a matter of fact, this is up to now the greatest collection of textual models, hypotheses and laws.

The second chapter on phonic phenomena (repetition, euphony, alliteration, assonance, aggregation) tries to capture our intuition of the phonic impression of text quantitatively. The methods used here are the binomial and the multinomial distributions, Cochran´s test and variance analysis, etc. A section is devoted also to monitoring of the Slovak rhyme development in the time period throughout 1848 –1960.

The next chapter concerns the level of words. Since there are about 200 word definitions, it can be operationally determined merely by setting up unequivocal criteria that are sometimes counter-intuitive and sometimes "counter-grammatical". There is no unique criterion that would hold for all languages. Words have an enormous number of properties, none of which stays in isolation, all of them are joined with at least one other property. The best-known property, which at the same time can easily be defined operationally, is its length (usually syllabic length). Up to now a number of models based on Wimmer-Altmann´s approach has been derived for these purposes and the research on word length enjoys a great popularity (cf. http://www.gdwg.de/~kbest/projekt.htm where one can find almost the complete literature). Further, there is a short description of Menzerath-Altmann´s law, word frequency in text (rank-frequency and frequency spectrum), vocabulary richness and its problems, and Frumkina´s law of the distribution of words in text-passages. Though some of these problems have been discussed very thoroughly in the literature – there are several books about some of the problems – here one can find a basic introduction preparing the reader for the mathematically more advanced papers.

The next chapter on sentence begins again with sentence length, the oldest problem of text analysis, generalizes Menzerath-Altmann´s law, shows its discrete derivation and places it in the family of power laws which were derived and formulated differently in other sciences. The syntactic properties of sentence (depth, width, centralization, asymmetry) are examined on the basis of dependence grammar using the results of Altmann and Lehfeldt (1973).

The chapter on content analysis, which is a summary term for a multitude of heterogeneous problems, is somewhat shorter. In this domain mathematical reasoning is not very popular, thus merely better developed problems are shown. There is the famous verb-adjective ratio introduced by Busemann, its sampling properties and proposals for asymptotic tests. The interpretation of the index differs slightly from that of other authors. In order to study word associations in text the book's authors propose a method and a way of constructing association nets. Text readability, which is a rather practical problem, is merely touched on. It is shown how to set up tests for the difference of readabilities using Tuldava´s characteristic.

In the sixth chapter suprasentence structures are dealt with. They are based on Hřebíček´s discovery of "aggregates" or "sentence aggregate", called today "hrebs" consisting of a collection of sentences containing the same word (morpheme, phrase, sign,…). Hrebs are considered to be the units of suprasentence text level. From this viewpoint we break the text down to hrebs rather than to sentences. Analyses of texts in this area have been mostly conducted on Turkish texts. The chapter is rather theoretical and explores the possibility of considering language as a dynamical system, its fractal structure, phase transitions and text compactness. It contains a number of hypotheses which can have a great impact on future theories of language.

The seventh chapter, which is relatively long, can evoke the impression that the study of the sequential text structure is an enormous problem having numerous aspects. Sentence length is studied here once again from the point of view of chaos theory. Ljapunov´s coefficient and Hurst´s indicator are shown as possible characteristics. For other problems Fourier-analysis and theory of runs can be used, which are explained using illustrations. Some models are shown for clumps and for the distribution of intervals between equal entities using among others Markov chains. There is a number of hypotheses concerning co-reference and text cohesion. Positional analysis scrutinizes the shaping of parts of a unit (e.g. verse).

The eighth chapter is devoted to text comparison. It contains tests for homogeneity and curve comparison. Though the methods are a little more sophisticated, they are well exemplified and can mechanically be used with the help of a software.

The last chapter contains a version of denotation analysis, developed by Ziegler and Altmann (2001). Here the text is partitioned in hrebs of a special kind and its properties (core, compactness, centrality, diffusion, coherence, connotative concentration, denotation width etc.) are defined and computed. Graph theory is used as method. There are already different versions (morpheme, word, phrase,…) of this approach.

The book not only explains complex problems in detail but presents new vistas in each chapter and some parts of the book are quite new. It is a pity that it was written in Slovak – a language not belonging to the best known ones - but since all of the authors (a mathematician, three linguists and a programmer) have at least some affinity to it, they decided to write it this way. Many examples are taken from Slovak but one finds also German, Turkish and Indonesian ones. It is our hope that the book will be soon translated into English.

The book cannot be read hastily. It must be studied. For the reader it is recommended to compute all examples either “with pencil and paper” or if possible, to program each problem. It is important to solve at least one of the tasks at the end of the sub-chapters.

Jana Kusendová

References

Altmann, G., Lehfeldt, W. (1973). *Allgemeine Sprachtypologie*. München: Fink.
Köhler, R. (1995). *Bibliography of Quantitative Linguistics*. Amsterdam: Benjamins.
Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse*. Wien: Praesens Verlag.