Glottometrics 2, 2002

RAM - Verlag

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitschrift für die quantitative Erforschung von Sprache und Text
Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungsystem (vorrangig WORD) geschickt werden
Glottometrics kann aus dem **Internet** heruntergeladen, auf **CD-ROM** (in PDF Format) oder in **Buchform** bestellt werden

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals
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**, obtained on **CD-ROM** (in PDF) or in form of **printed copies**

Herausgeber - Editors

G. Altmann 02351973070-0001@t-online.de
K.-H. Best kbest@gwdg.de
L. Hřebíček hrebicek@orient.cas.cz
R. Köhler koehler@uni-trier.de
O. Rottmann otto.rottmann@t-online.de
G. Wimmer wimmer@mat.savba.sk
A. Ziegler arneziegler@compuserve.de

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an
Orders for CD-ROM´s or printed copies to

RAM-Verlag RAM-Verlag@t-online.de

Herunterladen / Downloading: http://www.ram-verlag.de

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Glottometrics. – 2 (2002) –. – Lüdenscheid : RAM-Verl., 2002
Erscheint unregelmäßig. – Auch im Internet als elektronische Ressource unter der Adresse http://www.ram-verlag.de verfügbar. –
Bibliographische Deskription nach 2 (2002)
ISSN 1617-8351

Contents

Glottometrics 2, 2002, 1-10

The case of Czech possessive adjectives and their head nouns: some distributional properties

Ludmila Uhlířová, Prague[1]

Abstract. If a word class is defined on the basis of its co-occurrence with another word class in texts, then its ranking distribution may be modelled well by the negative hypergeometric probability distribution, and the interaction of the two classes can be described by the same type of model as well as with the help of some similar procedures, such as the non-linear regression. Possessive adjectives and their head nouns in a Czech corpus are taken to demonstrate those properties.

Key words: Possessive adjectives, ranking, negative hypergeometric distribution

Empirical experience made so far in quantitative studies has convincingly shown that the distributional statistics of linguistic entities sorted out as **classes** regularly leads to **skewed** distributions. The skewness manifests itself independently of the kind of constituents as well as of their level of complexity. Following the theoretical argumentation brought in by synergetic linguistics, we take the distributional skewness as a property that is grounded in the very nature of human language. With the universal principles of self-organisation and self-regulation being the two underlying control principles of language mechanisms can explain it. For a detailed argumentation, which puts synergetic linguistics into the neighbourhood of systems theory, chaos theory, game theory and other exact disciplines, see Köhler (1986), Altmann & Köhler (1995), Altmann (1995, 1999), Altmann & Koch (1998), Hřebíček (1997, 2000). Let us remember that the development of linguistics towards the science of language in the above mentioned sense has deep roots in the history of the field. As Patrick Sériot (1999:22) claims in his paper devoted to the Prague Linguistic Circle, "...seeing a structure as a goal oriented totality was a necessary step backward which made possible a fantastic step forward toward a theory of systems like Bertalanffy´s after World War II..."

It is only natural that explorative statistical analyses into various classes should go on so as to bring further arguments in support of what has been already achieved empirically and deduced from the theory.

In this paper, distributional features of Czech possessive adjectives and their head nouns are studied. Let us stress that it is by no means crucial that the language under study is Czech, and not, e.g. any other Slavic language or perhaps another language with a similar means of possessivity. The class of possessive adjectives has been chosen as the subject of analysis because of its specific, even "exclusive" features - they are listed below. We shall inquire into how a class with such features behaves statistically.

The relevant features are as follows:

Firstly, the Czech possessive adjectives have **very low frequencies** of occurrence in texts. A large corpus is needed to obtain a sufficient amount of data.

Secondly, it is a class with **fuzzy semantics**. Not all possessive adjectives express the meaning of "possessor". There are many other meanings of "possessive" adjectives as well,

[1] Address correspondence to: L. Uhlířová, Ústav pro jazyk český AV ČR, Letenská 4, CZ-118 51 Prague, Czech Republic. E-mail: uhlirova@ujc.cas.cz

which, in fact, are not possessive at all.

Thirdly, the class of possessive adjectives is **unambiguously defined** on the basis of their form. Czech possessive adjectives are derived only from certain (not from all) nouns by special suffixes *-ův, -ova, -ovo* and *-in, ina, -ino*, e.g. *ministrův tajemník* 'the minister's secretary', *lékařova ruka* 'the doctor's hand', *Janův společník* 'John's company', *babiččin obličej* 'granny's face' etc. No other word class is formed by these suffixes in Czech (for a detailed description of word formation rules including various rather strong restrictive conditions see any Czech grammar, e.g. Mluvnice češtiny 1, 1986). Therefore, no question arises whether a word "still" may be assigned to the class of possessive adjectives, or "rather" to another word class. Possessive adjectives are a class with clear-cut formal boundaries (nonetheless, this does not mean that they are quite easily identified in a text "automatically", with the help of an algorithm; some cases of wrong classification in our corpus had to be corrected.)

Fourthly, possessive adjectives take only **one syntactic position** in the clause, as evidenced in the corpus: they modify a noun. If there is a possessive adjective in a clause, its head noun is there as well.

Fifthly, possessive adjectives are a **recessive** word class. Studying the historical development of Czech one can see that they are used less and less frequently, and more and more restrictive conditions apply to its usage. Alternative linguistic means of expression of possessivity have been winning as the time goes on. Below, main competing means will be listed and their frequencies of occurrence in texts compared with those of possessive adjectives.

The question to be answered in this paper is as follows: How does a low frequent and recessive word class, with fuzzy semantics, but with clear-cut word-formation markers and with just one syntactic role **interact** with the practically unlimited class of nouns as their potential heads?

If we want to speak about **inter**action, we have, logically, to approach the topic from two directions:

(1) ***Pos→N***. It has been said above that if there is a possessive adjective (*Pos*) in a clause, its head noun (*N*) is found there as well. Statistically, the interaction is unique: the head *N* occurs with the probability that practically equals 1 (ellipsis of the head noun is rare).

(2) ***N→Pos***. Considering the opposite direction, we have to ask the following question: If there is an *N* in a clause, what is the probability that it is modified by a *Pos*? On the basis of common experience we may say straight away that the implication *N→Pos* belongs to those characterised by Altmann & Wimmer (2001:110) as "so weak that their frequencies are not even significantly correlated".

Apparently, the interaction differs dramatically if seen from each of the two opposite directions. If we want to know more about it, we have to divide the task into several successive steps and further specify and limit the empirical data to be analysed.

First step. We start from a corpus consisting of 1 262 738 word forms. In the corpus there are about 37 thousands of *N* and only 657 *Pos*. From this amount of *Pos* we choose one semantic subclass of *Pos*, namely those *Pos* which are derived from an *N* with the meaning "possessor or a member of a group/collective", and, which modify an *N* with the meaning "possession or a group/collective". E.g. *manželův společník* 'the husband´s company'*, Janův otec* 'John´s father'*, prezidentova ochranka* 'the president´s security guard'*, Masarykův nástupce* 'Masaryk´s successor'*, Sennův rival* 'Senna´s rival'*, autorův dvojník* 'the author´s double'*, ministrův tajemník* 'the minister´s secretary' etc. In the following we shall deal with this semantically defined subclass of *Pos* only. We consider it necessary to work with a subclass of *Pos* which semantically is relatively homogeneous. The reason for it is that many

"possessive" *NP*s have other meanings than proper possessive (see, e.g. Heine, 1997, for a detailed classification), and, therefore, they display different behaviour in texts accordingly.

Our subclass of *Pos→N* occurred 90 times in the corpus. Let us have a look at *N* from this subclass. The list of *N* is given in Table 1, first column. The total number of different *N* is 37. The frequencies of concrete *N* co-occurring with *Pos* are very low, see column 2; their frequency is ≥ 5 with four *N* only. Comparing the frequencies of *N* co-occurring with *Pos* (given in the second column) with the total frequencies of N in the whole corpus, as given in the seventh column, we can see no salient correlation between the total frequency of individual *N* in the corpus (ranging in the interval from 295 to 3 occurrences) and its respective frequency of co-occurrence with *Pos*. Hence, the result of this "surface" comparison is negative, which is actually what we expected. However, we shall return to the table later on.

Table 1

The frequencies of *N* in possessive constructions *Apos→N, PronPos→N, mít→N, N→Ngen*. Total frequency of the possessive constructions; frequency of *N* in the whole corpus and the relative frequency of *N* in possessive constructions in %.

N	Apos →N	PronPos →N	*mít* →N	N →Ngen	total frequency	frequency of N in the corpus	relative frequency of N in possessive constructions
rodina family	6	20	11	17	54	295	18,3
ruka hand	1	6	4	41	52	280	18,6
hlava head	3	5	3	40	51	238	21,4
hlas voice	2	6	4	39	51	225	22,7
manažer manager	1	2	3	42	48	216	22,2
mluvčí speaker	12	6	1	152	171	213	80,3
náměstek assistant	2	9	1	119	131	158	82,9
tvář face	5	0	2	26	33	142	23,2
tělo body	8	13	0	24	45	139	32,4
soupeř rival	2	10	5	7	24	128	18,8
úředník officer	1	4	3	19	27	96	28,1
otec father	4	12	0	17	33	90	36,7
noha leg	2	1	1	4	8	88	9,1
poradce advisor	4	2	1	23	30	81	37,0
paní lady	1	1	0	2	4	77	5,2
obránce defender	1	1	1	12	15	77	19,5
tajemník secretary	2	0	1	58	61	76	80,3
záda back	4	2	0	7	13	61	21,3
agent agent	1	2	0	13	16	51	31,4
doprovod company	2	2	2	13	19	43	44,2
nástupce successor	2	15	3	7	27	38	71,1
společník partner	1	6	0	4	11	37	29,7
lebka cranium	2	0	0	8	10	27	37,0
pravice right wing	2	1	0	1	4	25	16,0
bok side	1	1	0	7	9	24	37,5
obličej face	1	0	0	3	4	23	17,4
protějšek opposition	1	7	0	2	10	22	45,5
předchůdce predecessor	3	7	0	4	14	19	73,7
ostatky relics	2	0	0	6	8	14	57,1
vyslanec envoy	1	1	0	9	11	12	91,7

ochranka body guards	2	3	0	2	7	11	63,6
rival rival	2	1	0	1	4	9	44,4
choť spouse	2	1	0	1	4	7	57,1
stratég strategist	1	1	0	1	3	6	50,0
dvojník double	1	0	0	0	1	4	25,0
následovník successor	1	2	0	0	3	4	75,0
ctitel fan	1	0	0	1	2	3	66,7
total	90	150	46	732	1018	3059	

Second step. Let us return to *N* and examine their distributional properties in more detail. Let us remind that the subclass of 37 *N* has been chosen from the corpus on the basis of a single **shared property**. The property is their co-occurrence with *Pos* in the corpus. Let us arrange *N* in the order of their decreasing frequency in the corpus, as it is done in Table 1, column 7, and repeatedly in Table 2 (second column - rank, third column - frequency). Now we ask whether the empirical ranking distribution may be modelled by one of the "usual" discrete probability models already known from quantitative analyses. Experimentally (using Altmann´s Fitter, 1997) we find out that there is a very good fit of our data with the negative hypergeometric distribution.

The negative hypergeometric distribution has the following form (Wimmer, Altmann, 1999):

$$(1)\quad P_x = \frac{\binom{M+x-1}{x}\binom{K-M+n-x-1}{n-x}}{\binom{K+n-1}{n}},$$

where *K, M* and *n* are the parameters of the distribution. The values calculated for the 1-displaced negative hypergeometric distribution are given in Table 2, third column. The chi-square test shows a very good fit:

$$K = 4.08,\quad M = 0.88,\ n = 37,\quad DF = 33,\quad X^2 = 33.70,\quad P(X^2) = 0.43.$$

The good fit can be clearly seen also from Fig. 1.

Table 2
Frequency distribution of those *N* in the corpus, which are heads of *Pos*.
The probability model of the distribution.

N	rank	empirical frequencies	calculated values
rodina	1	295	311.96
ruka	2	280	259.72
hlava	3	238	230.71
hlas	4	225	209.01
manažer	5	216	191.04
mluvčí	6	213	175.41
náměstek	7	158	161.43
tvář	8	142	148.74
tělo	9	139	137.07
soupeř	10	128	126.27
úředník	11	96	116.22
otec	12	90	106.84
noha	13	88	98.06
poradce	14	81	89.83
paní	15	77	82.11
obránce	16	77	74.86
tajemník	17	76	68.05

záda	18	61	61.67
agent	19	51	55.68
doprovod	20	43	50.08
nástupce	21	38	44.84
společník	22	37	39.96
lebka	23	27	35.41
pravice	24	25	31.19
bok	25	24	27.28
obličej	26	23	23.67
protějšek	27	22	20.36
předchůdce	28	19	17.33
ostatky	29	14	14.58
vyslanec	30	12	12.10
ochranka	31	11	9.88
rival	32	9	7.90
choť	33	7	6.17
stratég	34	6	4.68
dvojník	35	4	3.41
následovník	36	4	2.36
ctitel	37	3	2.93
total		3059	

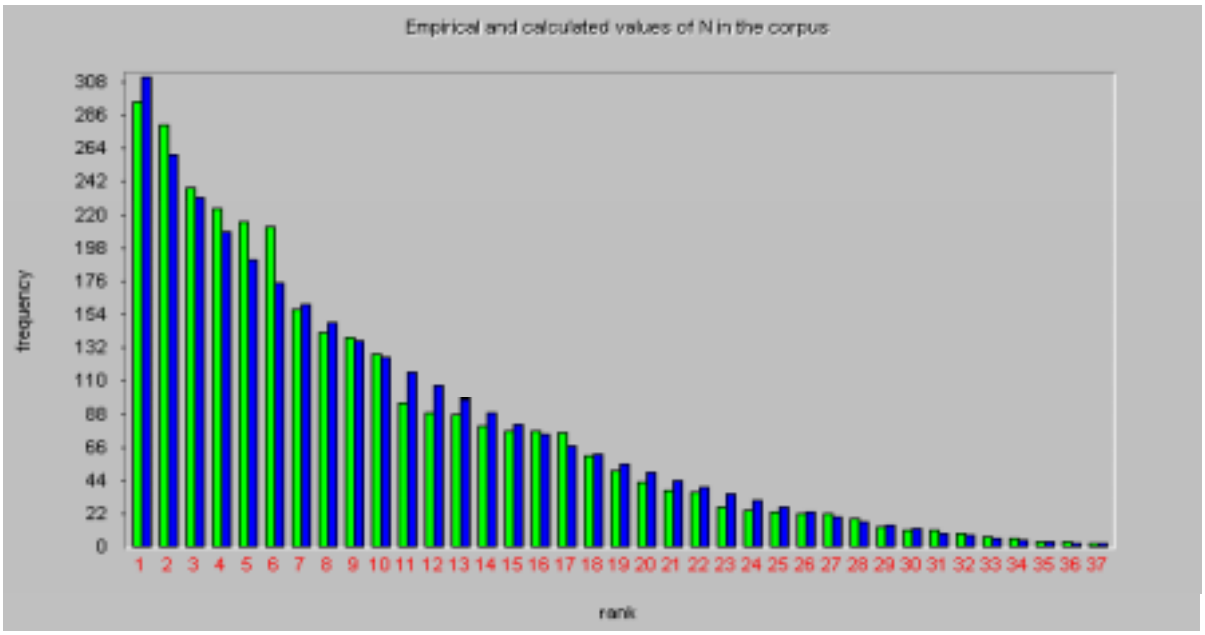

Fig. 1. Empirical and calculated frequencies of *N* in the corpus.

This type of probability distribution is well known in quantitative linguistics. Altmann (1988: 117-121,179) showed that associative repetition of word pairs in texts abide by this distribution. The same holds for word repetition in text blocks (Altmann & Burdinski 1982). Best's recent demonstration (2000: 60-61) that it is exactly this distribution that applies when parts of speech are studied from the point of view of ranking distribution is of special importance for our task. Last but not least, Köhler & Martináková-Rendeková (1998) successfully applied this model for ranking to musicology, investigating pitch, duration and intensity values in Chopin´s Etude op. 25, No. 11. Generally, the negative hypergeometric distribution can be obtained from the stochastic process of birth and death, which is a process underlying both verbal and musical "texts".

The result is positive and will be utilized in further steps.

Third step. As already said, *Pos* are a recessive word class, and if seen from the historical perspective, both its inventory and their usage in texts are more and more restricted. Restrictions concerning their semantics, word formation, syntax, referentiality, individual style, genre etc. could be adduced, but it is not the aim of this paper to do so and we leave this point

aside; just for illustration, let us say that in present-day Czech (in contrast to Old Czech) it is not possible to modify *Pos* and to say e. g. **tohoto autorův subjektivní názor* 'this author´s subjective opinion'. Instead of *Pos→N* alternative syntactic patterns are more and more frequently used. There are three main syntactic competitors: **PosPronoun→N**, **verb *mít* (have) →N** and **N→Ngenitive**. E.g., in addition to *minister´s secretary* we have constructions such as *his secretary* (in continuous text), *the minister has a secretary*, or *secretary of the minister*. More syntactic competitors might be added, but the three suffice for our purpose.

Let us return to our list of 37 *N* now and ask how often these *N* occur in the three above-mentioned competing patterns in the corpus. The frequencies are given in the relevant columns, Table 1. What do the data show? Firstly, the frequency of the verbal pattern is relatively low. Probably, some other "possessive" verbs should have been taken into account. Secondly, *PosPronoun* is significantly more frequent than *Pos→N*, which is clear evidence of text cohesion. Thirdly, the most important competitor is *N→Ngenitive*, as expected. In all, the data confirm that the recessive pattern *Pos→N* does have its regular competitors in texts, and these competitors have, taken together, high frequencies, outcoming the frequencies of *Pos→N* .

The overall frequency of the four competing patterns together (i.e. including *Pos→N* in the corpus is given in column 6, Table 1. Let us examine their distributional properties, similarly as it was done for *N* in the second step above.

Let us remind once again that the four "possessive" patterns were chosen from the corpus on the basis of a single shared property. The property is that they are patterns competing as syntactic periphrases of *Pos→N* and that a *N* from the set of 37 *N* occurs in each case. Let us arrange the total frequencies of the four competing patterns (see table 1, column 6) in the order of their decreasing frequency, as it is done in Table 3, column 3. Now we ask whether the ranking distribution may be modelled by one of the "usual" discrete probability models already known from quantitative linguistics. Experimentally we find out again that there is a very good fit of negative hypergeometric distribution to our data. The calculated values are given in Table 3, column 4; the parameters are as follows:

$K = 2.70, \quad M = 0.57, \text{n} = 39, \quad DF = 32, \quad X^2 = 27.13, \quad P(X^2) = 0.71.$

This result confirms once again that if a constituent class is defined on the basis of a shared categorial property - this time on the basis of the property "a competing pattern", its ranking distribution may be modelled by the negative hypergeometric distribution. The good fit can also be observed in Fig. 2.

Table 3
Ranking distribution of "possessive" constructions with individual *N* - empirical frequencies and theoretical values according to the negative hypergeometric distribution

N	rank	empirical frequencies	calculated frequencies
mluvčí	1	171	184.9
náměstek	2	131	102.7
tajemník	3	61	78.3
rodina	4	54	65.0
ruka	5	52	56.1
hlava	6	51	49.6
hlas	7	51	44.4
manažer	8	48	40.2
tělo	9	45	36.6
tvář	10	33	33.5
otec	11	33	30.8
poradce	12	30	28.4
úředník	13	27	26.2
nástupce	14	27	24.2
soupeř	15	24	22.3

doprovod	16	19	20.6
agent	17	16	19.1
obránce	18	15	17.6
předchůdce	19	14	16.2
záda	20	13	14.9
společník	21	11	13.7
vyslanec	22	11	12.5
lebka	23	10	11.4
protějšek	24	10	10.4
bok	25	9	9.4
noha	26	8	8.4

ostatky	27	8	7.5
ochranka	28	7	6.6
paní	29	4	5.8
pravice	30	4	5.0
obličej	31	4	4.2
rival	32	4	3.5
choť	33	4	2.8
stratég	34	3	2.2
následovník	35	3	1.6
ctitel	36	2	1.0
dvojník	37	1	0.5

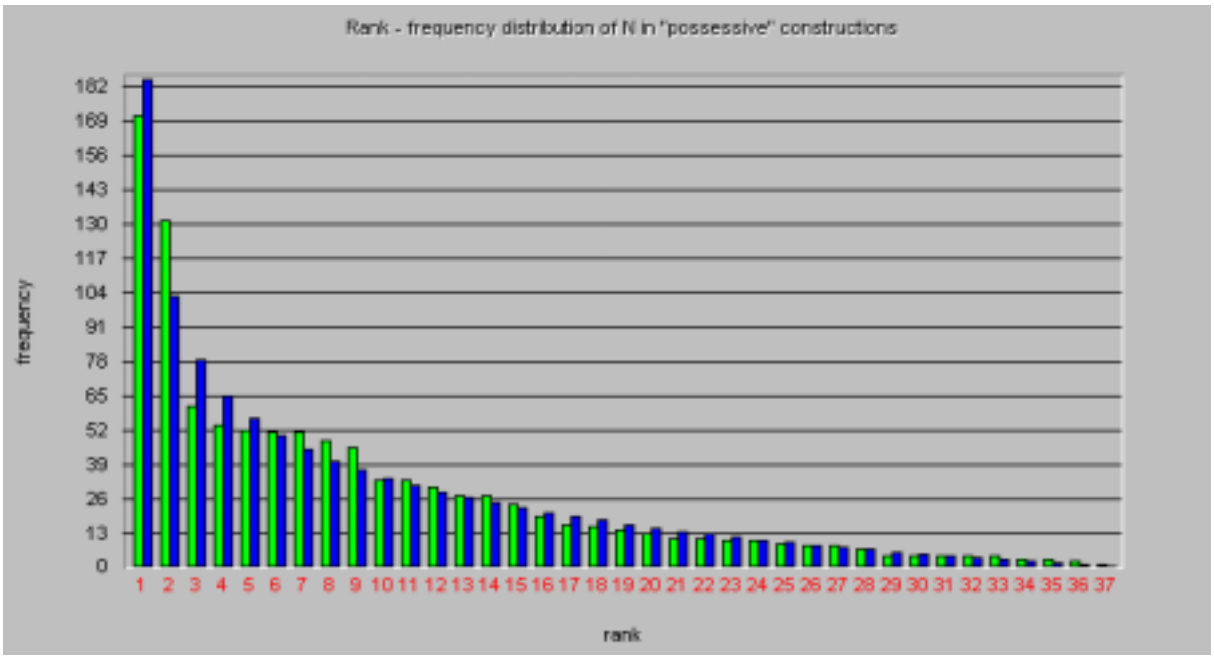

Fig. 2. Rank-frequency distribution of *N* in possessive constructions

Fourth step. Let us return to Table 1 once again. Now we can answer the question what the three preceding steps aim at. The question is: Is the occurrence of *N* in the four competing types of "possessive" patterns just accidental, or is it significant (typical)? In other words: Are there such *N* which significantly "attract" a constituent with which they form together a possessive pattern?

Let us compare the distribution of the four competing "possessive" patterns in Table 1, given in column 6, with the distribution of *N* in the corpus, as given in column 7. If the values in column 6 are divided by the values in column 7 in the analogous rows, then the ratio, given in the last column of Table 1, determines a measure of significance of occurrences of *N* in possessive constructions. As the numbers of occurrences of *N* in the corpus differ for each *N* (in each row of the table), the ratios are calculated with different accuracy. However, this fact does not neglect the result substantially - it still can be used for the given purpose.

What does the result show? The average ratio 1018:3059 (the last row) says that one third of all occurrences of *N* in the corpus occurs in "possessive" constructions. This value is sub-

stantially higher than random. However, there are big differences among individual *N*: the interval between the lowest and highest value of the ratio is <5.2% - 91.7%>. This dispersion is huge, and shows that, practically, the average value has no interpretative strength. But, if we consider individual ratios, we can see that they are higher than 5% with all *N*. If we take 5% as a significance boundary, we can say that the answer to the question posed above is positive for all *N*: It is possible to make predictions of occurrences of *N* (from our set of *N*) in possessive constructions.

Now let us arrange the values from Table 3 (column 8) in the decreasing order, as it is done in Table 4. These values do not represent any random variable, but just a numerical characteristic calculated as the product of two random variables, rank and frequency. What can we say about those values?

The sequence of these values cannot be considered a distribution because it contains proportions. However, a series analogous to the negative hypergeometric can be fitted using iterative non-linear regression. Writing the binomial coefficients as Gamma-functions using the parameters K = 2.45, M = 0.82, n = 37 and a constant C = 9.9020 we obtain

$$(2)\quad y = 9.9020\frac{\Gamma(0.82+x-1)\Gamma(2.45-0.82+37-x+1)}{(x-1)!(37-x+1)!\Gamma(0.82)\Gamma(2.45-0.82)} = \frac{9.902\Gamma(x-0.18)\Gamma(39.63-x)}{(x-1)!(38-x)!\Gamma(0.82)\Gamma(1.63)}$$

yielding the results in column 4 of Table 4. The determination coefficient is D = 0.95 signalising a very good result. The fit is displayed in Fig. 3.

Conclusion

(1) If a word class is defined with the help of a categorial criterion as in our case, i.e. if a word class of *N* is defined on the basis of the co-occurrence with *Pos* (syntagm *Pos*→*N*), then the ranking distribution of *N* in the corpus may be modelled by the negative hypergeometric distribution. This holds good in spite of the defining criterion being weak due to the facts that *Pos* is a rare category and the opposite correlation *N*→*Pos* is not statistically significant at all.

Table 4
The proportion of possessive constructions in the overall frequency of individual *N* in the corpus. Theoretical values according to (2)

N	rank	proportion in %	theoretical
vyslanec	1	91.7	108.8
náměstek	2	82.9	87.7
mluvčí	3	80.3	78.5
tajemník	4	80.3	72.5
následovník	5	75.0	67.9
předchůdce	6	73.7	64.3
nástupce	7	71.1	61.0
ctitel	8	66.7	58.4
ochranka	9	63.6	55.9
ostatky	10	57.1	53.6
choť	11	57.1	51.5
stratég	12	50.0	49.5
protějšek	13	45.5	47.6
rival	14	44.4	45.8
doprovod	15	44.2	44.0
bok	16	37.5	42.3
poradce	17	37.0	40.7
lebka	18	37.0	39.1
otec	19	36.7	37.5
tělo	20	32.4	36.0
agent	21	31.4	34.5
společník	22	29.7	32.9
úředník	23	28.1	31.4
dvojník	24	25.0	29.9
tvář	25	23.2	28.4

hlas	26	22.7	26.9
manažer	27	22.2	24.4
hlava	28	21.4	23.9
záda	29	21.3	22.3
obránce	30	19.5	20.7
soupeř	31	18.8	19.1
ruka	32	18.6	17.4

rodina	33	18.3	15.7
obličej	34	17.4	13.8
pravice	35	16.0	11.9
noha	36	9.1	9.8
paní	37	5.2	7.4

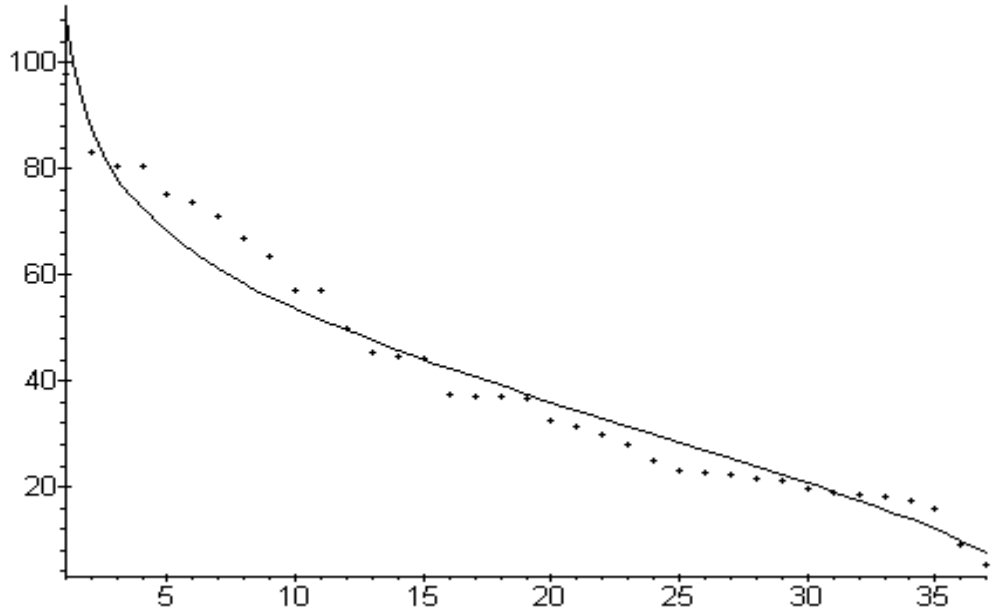

Fig 3. Fitting (2) to proportions

(2) If a constituent class is defined on the basis of a shared property - this time on the basis of the property "competing possessive pattern", containing an *N* from a given class, its ranking distribution may be modelled by the negative hypergeometric distribution.

These two results are in good agreement with what is known about the theoretical properties of this distribution as well as of its implementation on linguistic data, as have been described by Altmann (1988) and Best (2000), mentioned above. Our data are nothing but an example of an application of this model.

(3) Whereas *N→Pos* is statistically negligible, *N→"possessive pattern"* is significant: There exists a class of those *N* that associate with "possessive" patterns in a significant way. The probability that an *N* from the given class, defined on the basis of its "possessive" semantics, will behave in this way, was presented at a general level, by means of a non-linear regressive curve, not just as a single numerical characteristic, and the significance was tested by the coefficient of determination. From the linguistic point of view, it is, hopefully, a modest piece of evidence that lexical approaches to grammar such as, e.g**.** Francis, Hunston & Manning (1998) are fully justified and applicable to typologically diverse languages.

Note

This paper was supported by the Project LN00A063.

References

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Altmann, G. (1997). *Altmann-Fitter.* Lüdenscheid: RAM-Verlag.

Altmann, G., Burdinski, V. (1982). Towards a law of word repetitions in text-blocks. *Glottometrika 4, 147-167.*

Altmann, G. (1999). Review of M. P. Oakes: Statistics for Corpus Linguistics. Edinburgh: Edinburgh University Press 1998. *Journal of Quantitative Linguistics 6, 269-270.*

Altmann, G. (1996). The nature of linguistic units. *Journal of Quantitative Linguistics* 3, 1-7.

Altmann, G., **Koch, W. A.** (1998) W. A. (Eds.), *Systems. New paradigms for the human sciences. 614-645.* Berlin: de Gruyter.

Altmann, G., Köhler, R. (1995). "Language forces" and synergetic modelling of language phenomena. *Glottometrika 15, 62-76.*

Francis, G., Hunston, S., Manning, E. (1998) (Eds.). *Grammar Patterns*. Birmingham: HarperCollins Publishers.

Heine, B. (1997). *Possession. Cognitive sources, forces and grammaticalization*. Cambridge: UP.

Hřebíček, L. (1997). *Lectures on text theory*. Prague: Oriental Institute.

Hřebíček, L. (2000). *Variation in sequences*. Prague: Oriental Institute.

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

Köhler, R. & Martináková-Rendeková, Z. (1998). A systems theoretical approach to language and music. In: Altmann, G., Koch, W. A. (Eds.), *Systems. New paradigms for the human sciences. 614-645.* Berlin: de Gruyter.

Mluvnice češtiny I (1986). Praha: Academia.

Sériot, P. (1999). The impact of Czech and Russian biology on the linguistic thought of the Prague Linguistic Circle. *Travaux du Circle Linguistique de Prague, NS, 3, 15-24.* Amsterdam: John Benjamins.

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

Glottometrics 2, 2002, 11-16

Der Zuwachs der Wörter auf *-ical* im Deutschen

Karl-Heinz Best, Göttingen[1]

Abstract. Since 1955 the number of German words having the ending *–ical* has been increasing. The paper shows that this process abides by the Piotrowski law.

Key words: Piotrowski law, language change

Einleitung

Ein wichtiges Arbeitsfeld der Quantitativen Linguistik ist der Sprachwandel. Viele unterschiedliche Aspekte sind hier zu bearbeiten, u.a. die Frage, wie ein einmal ausgelöster Sprachwandel verläuft. Hierzu haben Altmann (1983) sowie Altmann u.a. (1983) im Anschluss an Vorschläge Piotrowskis ein theoretisches Konzept entwickelt, das sie deshalb Piotrowski-Gesetz nennen und das sich inzwischen in einer Reihe von Untersuchungen bewährt hat. Ein wesentlicher Aspekt des Piotrowski-Gesetzes besteht darin, dass es verschiedene Formen annimmt, je nach dem, ob es sich um einen vollständigen, einen unvollständigen oder einen reversiblen Sprachwandel handelt. Vollständige Sprachwandel liegen dann vor, wenn eine Einheit ganz und gar durch eine neue ersetzt wird; unvollständige dann, wenn ältere Formen teilweise ersetzt oder durch neue ergänzt werden; reversible dann, wenn ein Sprachwandel zunimmt, einen Gipfel erreicht, um dann wieder teilweise oder gar ganz zu verschwinden. Alle diese Formen von Sprachwandel konnten beobachtet werden; in allen bisher beobachteten Fällen konnten die verschiedenen Formen des Piotrowski-Gesetzes mit Erfolg als Modell bestätigt werden.

In der hier vorliegenden Arbeit soll ein weiterer Fall von unvollständigem Sprachwandel untersucht werden. Unvollständige Sprachwandel stellen z.B. die Entlehnungen von Wörtern aus einer „Geber"-Sprache in die aufnehmende Sprache dar. Solche Prozesse konnten bisher für die deutschen, lateinischen und slawischen Entlehnungen ins Ungarische (Beöthy & Altmann 1982), für die arabischen Entlehnungen im Persischen (Altmann u.a. 1983) sowie für eine Reihe gut belegter Entlehnungen ins Deutsche (Best 2001, 2001a, 2001b, 2001c, 2002, Best & Altmann 1986) getestet werden. Auch der Zuwachs des Wortschatzes einer Sprache lässt sich als unvollständiger Sprachwandel modellieren, wie am Beispiel des Englischen (Best 2001c: 108f.) und des Estnischen (Tuldava 1998: 136ff.) gezeigt werden konnte.

Zu den unvollständigen Sprachwandelprozessen gehört aber auch ein etwas anderer Fall: die Übernahme eines Affixes bzw. einer Kombination von Affixen. Das einzige bisher untersuchte und getestete Beispiel hierfür ist die Einbürgerung von *–ität* im Deutschen; es konnte gezeigt werden, dass dieser Prozess, der sich über etliche Jahrhunderte erstreckt, ebenfalls dem Piotrowski-Gesetz folgt (Best 2001c: 107). Eine neue Untersuchung von Kirkness (2001) bietet nun die Möglichkeit, mit der Übernahme und Ausbreitung von *-ical* (ausgehend von *Musical*) einen weiteren solchen Prozess zu testen. Anders als bei *-ität* vollzieht sich die Einbürgerung von *-ical* in nur wenigen Jahrzehnten, von den 50er bis zu den beginnenden 90er Jahren. Dies ist außerdem auch deshalb eine Besonderheit, weil mit der Umgestaltung des

[1] Address correpsondence to: K.-H. Best, Im Siebigsfeld 17, D-37115 Duderstadt. E-mail: kbest@gwdg.de

Genitiv Plurals im Russischen (Altmann u.a. 1983) und dem Wandel des Gebrauchs von *ward/ wurde* bei einem Autor (Kohlhase 1983) bisher nur sehr wenige derart kurzfristige Sprachwandel hinsichtlich ihres Verlaufs beobachtet und getestet wurden.

Die Wörter auf *-ical* im Deutschen

Kirkness (2001) erörtert am Beispiel der Wörter, die sich in der Nachfolge von *Musical* im Deutschen ausgebreitet haben, welchen Einfluss direkte Entlehnungen aus dem Englischen spielen und wie groß der Anteil der Wörter ist, die - auf der Basis fremdsprachiger Wörter oder Wortteile - im Deutschen selbst gebildet wurden. Als Beispiel für diese Erörterung dienen ihm die Wörter, die im *Anglizismen-Wörterbuch* (Carstensen, Busse 1993-1996) als *-ical*-Bildungen aufgeführt sind. Diese Wörter werden datiert in einer Tabelle zusammengestellt, die hier als Datenbasis verwendet wird.

Zur Datenaufnahme: Alle von Kirkness (2001) aufgeführten Wörter aus dem *Anglizismen-Wörterbuch*, die eine *-ical*-Konstituente enthalten, werden berücksichtigt. In den meisten Fällen handelt es sich dabei um Substantive mit *-ical* als letzter Konstituente (*Grusical*, *Erotical*, *Friesical*...), manchmal aber ist es auch Konstituente eines Determinans (*Musical-Publikum*, *Musical-Göttin*). Kirkness unterscheidet *Musical* als polysemes Wort in *Musical1* (Bezeichnung eines Theatergenres) von *Musical2* (Charakterisierung eines Theaterstücks oder Films als zu diesem Genre gehörig); in diesem Fall werden hier ebenso wie bei *Grusical1* und *Grusical2* zwei unterschiedliche Einheiten angesetzt. Eine solche Differenzierung polysemer Wörter wird bei den Lexemen mit einer anderen Basis nicht vorgeschlagen; es handelt sich insgesamt nur um drei Wörter.

Unterschiedliche Schreibweisen (*Grusical* vs. *Grusikal; Finanz-Grusical* vs. *Finanzgrusical*) werden nicht als Hinweis darauf verstanden, dass es sich dabei um verschiedene Wörter handelt, wohl aber verschiedene Wortbildungen (*US-Sexical* vs. *Sexical*). Belege aus andern Quellen als dem *Anglizismen-Wör-terbuch*, die Kirkness vereinzelt anführt, bleiben unberücksichtigt.

Die folgende Tabelle gibt eine Übersicht über die Anzahl der Wörter, die nach Kirkness (2001) eine *-ical*-Konstituente enthalten, und zeigt, in welchem Zeitraum sie im Deutschen auftreten; unter „berechnet“ stehen die Werte, die sich bei der Anpassung des Piotrowski-Gesetzes in der Form des unvollständigen Sprachwandels

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

an die kumulierten Werte der Tabelle 1 ergeben. (Zur Begründung und Ableitung des Modells vgl. Altmann 1983: 60f.). Außerdem gibt *c* an, gegen welchen Wert der Sprachwandel strebt; *t* steht für die Zeit; *a* und *b* sind Parameter.

Die Graphik in Abb. 1 macht schon deutlich, dass in diesem Fall eine hervorragende Übereinstimmung zwischen Theorie und Beobachtung erzielt werden konnte. Dies wird durch den Determinationskoeffizienten $D = 0.99$ bestätigt.

Tabelle 1
Zuwachs der Wörter auf *-ical* im Deutschen (n. Kirkness 2001)

t	Zeitraum	**absolut**	**kumuliert**	**berechnet**
1	1955-59	2	2	7.43
2	1960-64	13	15	13.79

3	1965-69	11	26	23.98
4	1970-74	14	40	37.88
5	1975-79	11	51	53.17
6	1980-84	13	64	66.49
7	1985-89	15	79	75.94
8	1990-91	2	81	81.70
	a = 21.95	*b* = -0.70	*c* = 88.29	*D* = 0.99

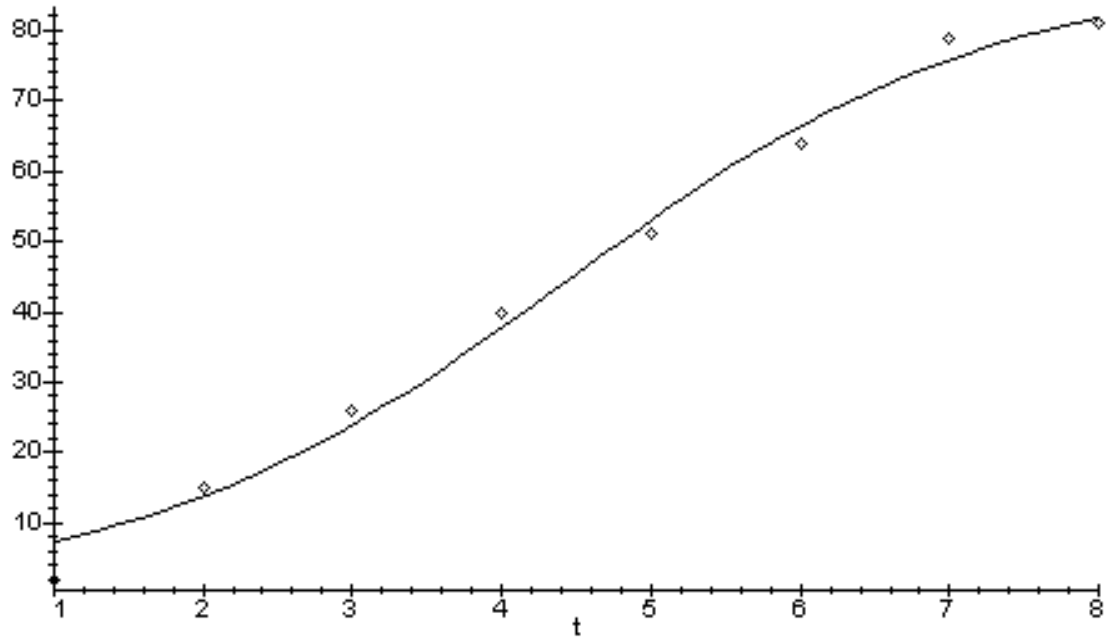

Abbildung 1. Zuwachs von Wörter auf *-ical*

Zusammenfassung

Zunächst einmal darf festgestellt werden, dass auch dieser weitere Fall von Sprachwandel das Piotrowski-Gesetz stützt. Es bleibt dabei, dass bisher kein einziger hinreichend dokumentierter Fall ein negatives Testergebnis provoziert hat. Dies gilt auch für den ursprünglich problematischen Fall von *darft/ darfst* (Best 1983: 112f.), der inzwischen auf verbesserter Datenbasis als „gesetzeskonform" nachgewiesen werden konnte (Best 2001c: 105).

Ein Interpretationsproblem: Parameter *c*

Ein spezielles Problem verdient jedoch weitere Beachtung und Diskussion: Es handelt sich dabei um die Deutungen, die man dem Parameter *c* in der Formel des Piotrowski-Gesetzes für den unvollständigen Sprachwandel geben kann. Dieser Parameter gibt an, auf welche Zielgröße hin sich der Zuwachs der Entlehnungen entwickelt. Beöthy & Altmann (1982: 173) billigen ihm keine allzu große Bedeutung zu: „Schätzt man I_B [= Parameter *c* der angesprochenen Formel. Verf.] aus den Daten, so bekommt man sicherlich eine Zahl, die die Anpas-

sung erleichtert, aber gleichzeitig eine schwache Interpretation hat...“ Kempgen (1990: 112) misst diesem Wert als Obergrenze für den Zuwachs von Entlehnungen überhaupt keine Bedeutung bei.

Nun ist völlig klar, dass *c* nicht als absoluter Wert verstanden werden darf, der tatsächlich das Ende einer Zunahme der Entlehnungen von Wörtern oder Konstituenten wie *-ical* und *-ität* markiert. Dies wird unmittelbar einsichtig, wenn man einmal *c* für die französischen Entlehnungen im Deutschen durch Schätzung bestimmt: Auf der Basis der Daten, die sich aus *Duden. Etymologie (1963)* gewinnen ließen, erhält man *c* = 1401 (Best & Altmann 1986: 36), auf der Basis von Telling (1987) dagegen *c* = 1983 (Best 2001: 264). Für lateinische Entlehnungen im Deutschen ergab sich auf der Basis von *Duden. Etymologie (1963)* in einer neuen Berechnung *c* = 1285, auf der Basis von Kirkness ([Hrsg.] 1988) *c* = 2576 (Best 2001a).

Die Schätzwerte für *c* ändern sich auch schon, wenn man die Datenbasis variiert; so hat sich bei rechnerischen Experimenten ergeben, dass der Schätzwert für *c* im Falle der französischen Entlehnungen ins Deutsche größer wird, wenn man die Daten für das 20. Jhd. einmal beiseite lässt. Dies kann verschiedene Ursachen haben, darunter sicher die Tatsache, dass die Schätzungen dieses Parameters von den Trends abhängig sind, die sich bis zum Abschluss der Datenerhebung zeigten. Die Entwicklungen in der Vergangenheit ließen offenbar für das 20. Jhd. einen größeren Zuwachs französischer Entlehnungen erwarten, als dann tatsächlich eingetreten ist. Hier ist allerdings auch noch zu berücksichtigen, dass diese Wörterbücher ja nicht das gesamte 20. Jhd. erfassen konnten, wenn sie schon 1963 bzw. 1987 erschienen sind.

Betrachtet man einmal die Schätzwerte für *c* als Prognosen für zukünftige Entwicklungen, so ist auch in diesem Fall Vorsicht geboten: Der Wert für *c* kann nur dann Aussagekraft entwickeln, wenn man die bereits erwähnten Probleme beachtet und wenn zusätzlich davon ausgegangen werden darf, dass die Randbedingungen, die sich auf den Zuwachs auswirken, gleich bleiben. Hiermit ist nicht generell zu rechnen; man erinnere sich nur an die Auswirkungen der politischen Entwicklungen auf die slawischen bzw. speziell die russischen Entlehnungen im Deutschen (Best 2002) und Ungarischen (Beöthy & Altmann 1982; Best 2001) im 20. Jhd. In beiden Fällen liegen die empirisch bestimmten Werte deutlich oberhalb der Kurve, die sich aufgrund der Berechnungen ergeben; *c* gibt also für den aktuellen Stand in diesem Fall anders als z.B. im Fall des Französischen zu niedrige Werte an.

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 Datenbasis darstellt. Wenn *c* interpretiert werden soll, so immer nur bezogen auf die Wörterbücher oder andere Quellen, 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 Voka-bular der Sprache darbietet und wenn man diesem Wörterbuch eine gewisse Repräsentativität für die Sprache zubilligen kann. Dabei treten viele Probleme auf, etwa die Frage, wie umfangreich das Lexikon einer Sprache überhaupt ist (vgl. zum Deutschen: Best 2001c: 14ff.), damit man dann evt. bestimmen kann, welchen Anteil daran die Entlehnungen aus den verschiedenen Sprachen haben. Man müsste *c* also in einem solchen Fall „hochrechnen“, indem man seinen berechneten Wert unter Berücksichtigung der Repräsentativität der Datenbasis, aufgrund deren er gewonnen wurde, in Beziehung setzt zur Grundgesamtheit, dem Lexikon der betreffenden Sprache. Gewiss werden die Meinungen dazu auseinander gehen, mit welcher Treffsicherheit dies durchführbar ist. Eine andere sinnvolle Interpretation für *c* ist aber sicher möglich: Wenn man wie in Best & Altmann (1986) auf gleicher Wörterbuchgrundlage bestimmen kann, gegen welchen Wert die Entlehnungen für die verschiedenen „Geber“-Sprachen streben, so lässt sich mit einiger Vorsicht immerhin ihr relativer Einfluss auf die aufnehmende Sprache mit Hilfe von *c* charakterisieren. Mit Kempgen (1989: 112) kann man diesem Index Aufschlusswert

über die „*Stärke des Kontaktes*“ zwischen zwei Sprachen zubilligen, damit aber auch die Stärke des Kontakts einer Sprache mit der zu mehreren anderen Sprachen vergleichen.

Im Fall der Ableitungen auf *-ical* und *-ität* stellt sich das Problem der Interpretation von *c* etwas anders dar, wenigstens, was die *-ical*-Bildungen betrifft. Ein Blick auf die Liste, die Kirkness zusammengestellt hat, macht den Eindruck, als ob es sich dabei in etlichen Fällen um recht kurzfristige Erscheinungen handeln sollte. In diesem Fall könnte sich der Schätzwert $c = 88$ zumindest in seiner Dimension als einigermaßen realistisch erweisen, wenn man einmal annimmt, dass zwar noch weitere, bisher nicht verzeichnete *-ical*-Bildungen zu erwarten sind, andererseits aber etliche der aufgelisteten Formen keine besondere Stabilität erwarten lassen. Im Unterschied dazu sind die *-ität*-Bildungen über Jahrhunderte hinweg entstanden und haben sich also teilweise schon lange Zeit im Lexikon des Deutschen gehalten. Nichtsdestoweniger kann sich ein einmal eingeschlagener Trend bei veränderten Randbedingungen auch umkehren, wie die reversiblen Sprachwandel lehren (Imsiepen 1983).

Literatur

Altmann, Gabriel (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, Karl-Heinz, & Kohlhase, Jörg (Hrsg.), *Exakte Sprachwandelforschung: 54-90.* Göttingen: edition herodot.

Altmann, Gabriel, von Buttlar, H., Rott, W., & Strauß, U. (1983). A law of change in language. In: Brainerd, B. (ed.), *Historical linguistics: 104-115*. Bochum: Brockmeyer.

Beöthy, Erzsébeth, & Altmann, Gabriel (1982). Das Piotrowski-Gesetz und der Lehnwortschatz. *Zeitschrift für Sprachwissenschaft 1*: 171-178.

Best, Karl-Heinz (1983). Zum morphologischen Wandel einiger deutscher Verben. In: Best, Karl-Heinz, & Kohlhase, Jörg (Hrsg.), *Exakte Sprachwandelforschung* (S. 107-118). Göttingen: edition herodot.

Best, Karl-Heinz (2001). Ein Beitrag zur Fremdwortdiskussion. In: *Die deutsche Sprache in der Gegenwart. Festschrift für Dieter Cherubim zum 60. Geburtstag.* Hrsg. v. Stefan J. Schierholz in Zusammenarbeit mit Eilika Fobbe, Stefan Goes u. Rainer Knirsch (S. 263-270). Frankfurt u.a.: Lang.

Best, Karl-Heinz (2001a). Das Fremdwort aus der Sicht der Quantitativen Linguistik. 3. Kolloquium Transferwissenschaften: Theorie, Steuerung und Medien des Wissenstransfers, Göttingen, 5.-7.9.01.

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

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

Best, Karl-Heinz (2002). Slawische Entlehnungen im Deutschen. Ms. (eingereicht)

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

Carstensen, Broder, & Busse, Ulrich (1993-96) *Anglizismen-Wörterbuch. Der Einfluß des Englischen auf den deutschen Wortschatz nach 1945.* 3 Bde. Begründet von Broder Carstensen, fortgeführt von Ulrich Busse. Berlin/ New York: de Gruyter.

Duden. Etymologie (1963). Mannheim: Bibliographisches Institut - Dudenverlag.

Imsiepen, Ulrike. 1983. Die e-Epithese bei starken Verben im Deutschen. In: Best, Karl-Heinz, & Kohlhase, Jörg (Hrsg.), *Exakte Sprachwandelforschung* (S. 119-141). Göttingen: edition herodot.

Kempgen, Sebastian (1990). Zur Modellierung von Lehnbeziehungen. In: Walter Breu (Hrsg.), *Slavistische Linguistik 1989: 99-116.* München: Sagner.

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.

Kirkness, Alan (2001). Anglicisms, Borrowings and Pseudo-Borrowings in German: *-ical* Revisited. In: *Proper Words in Proper Places. Studies in Lexicology and Lexicography in Honour of William Jervis Jones: 320-333.* Ed. by Maire C. Davies, John L. Flood and David N. Yeandle. Stuttgart: Verlag Hans-Dieter Heinz Akademischer Verlag.

Kohlhase, Jörg (1983). Der Wandel von *ward* zu *wurde* beim Nürnberger Chronisten Heinrich Deichsler. Als ein Nachtrag zum Vorigen. In: Best, Karl-Heinz, & Kohlhase, Jörg (Hrsg.), *Exakte Sprachwandelforschung: 103-106.* Göttingen: edition herodot.

Telling, Rudolf (1987). *Französisch im deutschen Wortschatz.* Berlin: Volk und Wissen.

Tuldava, Juhan (1998). *Probleme und Methoden der quantitativ-systemischen Lexikologie.* Trier: Wissenschaftlicher Verlag Trier.

Glottometrics 2, 2002, 17-32

The elements of symmetry in text structures

Luděk Hřebíček, Prague[1]

Abstract. The present paper refers to general text theory which is based on the Menzerath-Altmann law. This law defines the mutual relationships between the language levels inside a given text. They have important semantic consequences. The structure is observed from the viewpoint of possible changes (transformations, movements) in the treatment of the structural relations. We want to know what is invariant when such changes are applied. This approach follows the classical work on symmetry by Hermann Weyl (1952/1968).

Key words: Menzerath-Altmann's law, general text structure, symmetries (automorphisms), Turkish

1. Introduction

1.1. The expression 'text structure' need not be understood as an obscure phrase. It is possible to relate it to the recent attempts at the formulation of a general text theory which is based on principles respected in quantitative linguistics. Those principles are applied to all empirical scientific branches, their typical trait being testability. It means that the core of each theory is a law (= scientific conjecture) formulated such that it can be proved by testing.

An attempt to construct such a general text theory was recently formulated; 'general' here means 'valid for an arbitrary natural language'. This theoretical construction is based on two linguistic laws:

1. Menzerath-Altmann's law using the notions of *language construct* and its *constituent(s)*, and
2. the theory of word length which – according to certain experience - can be expanded to some other language units as well.

The literature concerning the first theory today exhibits a large list of items, from which we quote the basic works only: Menzerath (1928, 1954), Altmann (1980), Altmann, Schwibbe et al. (1989), Köhler (1984), Best (2001: 94-96). The second theory is connected with some families of theoretical distributions; it is explained in the works by Wimmer et al. (1994), Altmann & Köhler (1995), Altmann & Best (1996) and Wimmer & Altmann (1996); it has been applied to many languages by K.-H. Best (see, for example, Best 1996) and by a group of authors presenting their works under Best's editorship. The basic idea of this theory refers to the assumption of two "forces" with one of them belonging to text producers and the other to text recipients.

Both special theories pertain to the size of language units (constructs); both treat this property of language units from different aspects: the first one can be characterized as a semantic theory, the other one as a communicative text theory. The common application of these two language functions can be referred to as functional interference and understood as one of the reasons causing the high complexity of language systems and structures.

In the present paper, however, we are mainly interested in text structures which are

[1] Address correspondence to: L. Hřebíček, Junácká 17, CZ – 16900 Praha 6, Czech Republic. E-mail: hrebicek@orient.cas.cz

elucidated by the first theory which defines the common properties of language levels. Its relevance was accented by the introduction of a new language level encompassing supra-sentence structures, see Hřebíček (1989, 1997). In short, the present paper concentrates on the structures which appear to be consequences of Menzerath-Altmann's law and its application to texts. It offers several addenda to the experiments and their results described elsewhere (see Hřebíček 1992, 1995, 1997, 2000).

1.2. It is generally accepted in linguistics to understand language as a phenomenon comprising structures. Instead of 'phenomenon' the term 'system' can be used. Systems are usually described as sets of points or items constituting the individual systems, and sets of relationships defined at those points. With respect to their mutual relationships, points can be assorted into sets or classes forming the hierarchy of points and relations. The synonymic expression for 'hierarchy' is *structure*. On the other hand, the word 'structure' can be translated as 'configuration' or 'arrangement'. Which is then the reason for introducing the concept of *symmetry*?

Intuitively, the points and their relations forming systems fill in certain abstract spaces. Seeking symmetries should help obtain a more definite notion of each individual space. For the purpose of the better understanding of the concept of symmetry, we refer to the classical work by Weyl (1952/1968).

One can imagine that the sense of this highly abstract concept consists of a better characterization of individual specific structures. Consequently, the notion of symmetry can express more minute distinctions between items and the relationships characterizing systems. Together with symmetry, some general concepts enter the systems and their structures. At first, let us mention the concept of *invariance* of a configuration under certain transformation. This trait is also called *automorphism,* the transformation (or change) under which the character of the respective space structure is maintained. Something in a system is changed, but its important traits (expressed, for example, by scientific laws) remain preserved. Two worlds, one of them obtained from the other through an automorphic transformation which continually maintains general laws of nature, has to be treated as one and the same world. Another general notion is the *act of decision* applied to the case of relative properties of systems such as the relation of *left* and *right* (for example, the left or right orientation of a thread depending on the selected direction to the respective poles). Sometimes, for example in geometry, spatial transformations are considered to be *similarities*. Together with the *identical transformation I,* the reverse transformation I^{-1} is introduced into mathematics. Under the well-known conditions, when

configuration C is similar to itself,

C^a is similar to C^b , and then C^b is also similar to C^a, and

C^a is similar to C^b, C^b to C^c, and then C^a is also similar to C^c,

the respective transformations form a *group.* Certain similarity transformations do not change proportions of a supposed configuration and they can be called *movements*. Sometimes certain configurations have an *iterative* character and then it is not supposed to be a movement, but a repeated operation with a model or pattern.

In the following attempt at seeking the elements of symmetry in text structures different kinds of transformation (or changes) of their treatment are discussed:

1. the direction in which the relationship of two language levels is observed, i.e. from a lower level to a higher one, or in the opposite direction;
2. different substitutions of the units of measurement when the size of the language formants is observed;
3. transposition to the lower levels of the pattern of observation used for the purpose of inquiry of the supra-sentence structure;

4. changing the treatment of a text as the wholeness of the mutually nested (= imbedded) language units to their sequential understanding.

Hermann Weyl expressed his opinion that whenever an object distinguished by some structure is examined, one should try to state the group of its automorphisms, i.e. the group of transformations preserving without change all its structural relations, see Weyl (1952/1968: 159-160).

1.3. Let us resume in short: Menzerath-Altmann´s law explains certain basic properties of text structures. For many centuries language units belonging to different levels have been studied in linguistics. Customarily, their descriptions have the form of grammatical rules and lists of lexical units corresponding to individual languages. The main purpose of these descriptions consists of applications: they help construct correct sentences. Sentence represents the highest language unit (and level) of classical linguistics. Menzerath-Altmann's law used as the core of a text theory formulates the principle which connects sentences and their structures to form a higher structure.

Paul Menzerath observed one characteristic relationship of words and syllables. He formulated the following assertion: The greater the total (for example, word) the smaller its parts (e.g. syllables). Gabriel Altmann generalized this property using the concepts of *language construct* and its *constituent(s)*:

The longer the language construct the shorter its constituents.

Both terms comprise units of different language levels. Thus we can say that the entire language structure (including text) is based on the relation 'to consist of'. Altmann derived the mathematical form of this relation which in its most condensed form obtained the appearance of a non-linear power function:

$$y = Ax^{-b}, \tag{1}$$

where y = constituent, x = construct, and A and b are parameters.

Each of the traditional language levels (phones, syllables, morphs, word forms, syntactic constructions including clauses, and sentences) is organized in accordance with this law, as was confirmed many times in different types of texts and different types of natural languages.

The linguistic reason for this form of language organization can easily be understood when a new supra-sentence language level is introduced as a structural component of languages.

This new level cannot be observed anywhere else but in texts. Let us assume all such sentence(s), in which a given lexical unit occurs in a text. Let us call this set LC (= larger context of a lexical unit). This set forms a construct LC of size x counted in the number of sentences contained in LC; each such sentence represents NC (= narrower context) of the same lexical unit having the size y which is the mean sentence length of LC (measured, for example, by the number of words). The relationship of LC to NC is similar to the relationship between construct and its constituent(s). After many text analyses it can be said that their relation is prescribed by (1), i.e. they are real constructs and their constituents.

The precise meaning of the word form, i.e. the actual semantic value of each word form in a text, is determined by the collocation with other lexical units in NC and LC. "Collocation" should be evaluated as being also the case when a given LC contains only one constituent; this means that LC contains only one sentence and the frequency of the respective lexical unit equals 1. Its lexical meaning is put more precisely by co-oc-currence with certain other units, at least in one sentence. This evidently represents lan-guage economy: if a lexical unit is more frequently determined by occurring in a higher number of sentences (and its frequency is higher than 1), the length of the respective sentences need not be high and the mean size of

the constituents diminishes. Consequently, the reason for the existence of Menzerath-Altmann's law consists in economy (i.e. in the lowering of the redundancy of the language/text system). This holds for the supra-sentence level.

Then we may ask why, for example, with the growing length of words in the number of morphs, the mean morph length declines? In the light thrown to the problem by the highest text level, we can see that the same economy also has an effect on the lower traditional language levels. Generally, it can be said that the principle of semantic precision which is valid for the highest level grew through the system at all language levels. Thus, the linguistic aspect of the law is evident and one can imagine, why language structure is organized according to the principle of 'being a unit nested inside a higher unit'. Each unit is a part of a higher unit, the highest one being text.

Now: what is text? For the sake of the analyses of its structure it must be stressed that text length is limited upward and downward. The upper limit of a text size, as was proved by observation, is given by the contextual ability of lexical units. The lower limit is dictated by the ability of statistical methods to detect the structural properties of extremely short texts.

Thus we can observe that the movement up or down the ladder of the language levels results in the sort of symmetry which represents similarity; the mutually similar items are located inside each other. This is the characteristic property of fractals. Text, or language, can be represented as a set distinguished by self-similarity. Menzerath-Altmann's law embodying this principle remains valid regardless of the movement from one level to another one. The fractal character of language sets and their mathematical qualities is analyzed by Leopold (2001).

2. Direction

2.1. Let us assume the dynamics of the language/text system. Generally, each function including (1) indicates how it changes when its independent variable is changed. Let us assume construct x increasing in one: $x + 1$. If the Menzerath-Altmann principle in the first approximation is

$$\log y = -b \log x \quad ,$$

the proportion representing the increase of the constituent when the construct increases in one is:

$$\frac{y_{x+1}}{y_x} \cong \frac{(x+1)^{-b}}{x^{-b}} \quad , \quad x = 1, 2, ...$$

Then the following expression is also correct:

$$y_{x+1} = y_x \left(\frac{x+1}{x} \right)^{-b} = y_x \left(\frac{x}{x+1} \right)^{b} \quad . \tag{2}$$

If the assumed increased construct ($x + 1$) is related to a given size of the same construct (x), their mutual relation should be the same as in the case of the opposite direction. The opposite direction is expressed by the sign of parameter b. The symmetry of the two assumed directions can be proved by two progressions, each unfolded according to one of the right-hand sides of equation (2); the value of y_1 is defined as A:

$$y_1 = A$$

$$y_2 = A\left(\frac{1}{2}\right)^b$$

$$y_3 = A\left(\frac{1}{2}\right)^b\left(\frac{2}{3}\right)^b$$

$$y_4 = A\left(\frac{1}{2}\right)^b\left(\frac{2}{3}\right)^b\left(\frac{3}{4}\right)^b$$

...

$$y_1 = A$$

$$y_2 = A\left(\frac{2}{1}\right)^{-b}$$

$$y_3 = A\left(\frac{2}{1}\right)^{-b}\left(\frac{3}{2}\right)^{-b}$$

$$y_4 = A\left(\frac{2}{1}\right)^{-b}\left(\frac{3}{2}\right)^{-b}\left(\frac{4}{3}\right)^{-b}$$

...

General formulae for these progressions are:

$$y_x = A\left(\frac{(x-1)!}{x!}\right)^b \quad , \qquad y_x = A\left(\frac{x!}{(x-1)!}\right)^{-b} \quad .$$

This is nothing but a trivial proof of the identity which is contained in (2):

$$y_x = A\left(\frac{1}{x}\right)^b \equiv Ax^{-b} \quad .$$

Thus we can see that there is a symmetry between two situations obtained through the transformation of the language system; the direction describing the increase or decrease of the independent variable does not change the respective function.

The range of function $f(x) = y_x$ is defined by the following limits: max y_1 = A and min $y_x \approx 1$. Both limits express the mean values of the subsets corresponding to the function. This is a postulated assumption, and we may ask whether it corresponds to reality. For example, in a Turkish text the relation between word length x in the number of morphs, and y, the mean morph length in the number of phones, was observed. It corresponds to the function:

$$y_x = 3.995x^{-0.3584}.$$

The min $y = 1$ approximately corresponds to $x = 48$, as is evident from Figure 1.

We can hardly imagine a word containing 47 or 48 morphs not only in Turkish but in any natural language, this is completely unrealistic. Equally unpractical would be the assumption basing on the normal distribution of morphs in words: Mean $\langle y \rangle = \Sigma y/\Sigma x = 1923/756 = 2.54$ in the observed text; with standard deviation = 1.34, the 5% confidence limits for this mean are 5.17 and –0.09; the negative value of the number of morphs per word is not acceptable similarly to the overestimated x. The observed min y_x is 1.78 which corresponds to $x = 9$.

We can assume that there are more complicated conditions for word length (this time, in the

number of morphs), namely, the conditions stated by the theory of language forces that are active in text construction.

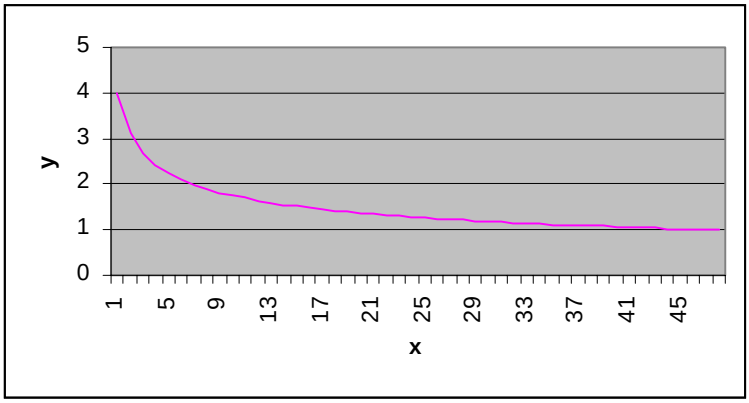

Figure 1. Function $y_x = 3.995x^{-0.3584}$

There is no reason to reject the hypothesis of insignificant difference between the observed and expected function presented in Table 1.

Table 1. Word length in morphs (x) and morph length in phones (y) in a Turkish text

x	**Observed y**	**Expected: $y = 3.99*x^{-0.3584}$**
1	3.995	3.995
2	3.041	3.116
3	2.650	2.695
4	2.328	2.431
5	2.200	2.244
6	2.367	2.102
7	2.179	1.989
8	1.750	1.896
9	1.778	1.818

Menzerath-Altmann´s function: parameters $A = 3.995$, $b = -0.3584$.
Wilcoxon test: $T = 15 > T_{0.05}(8) = 3$.

2.2. A more important type of symmetry is represented by the movement from one level to another. This characteristic inherently belongs to text structures, as is testified by the transformation of all the relationships 'construct – constituent' into sequences, see 4.1. The units of this relationship refer to different levels. The movement from level to level can be understood by an iterative application of Menzerath-Altmann's function. This means that this function formulated in (1) as generating the structure $f(x) = y$ should also be treated as an iteration, i.e. corresponding to the form $f(a) = a$. Menzerath-Altmann's law represents certain iterative pattern applied to (or, relevant for) the entire space of individual languages and texts.

With reference to the mentioned *level* → *level* movement, formula (1) can be rewritten as $x = Ax^{-b}$, from which it follows that

$$x = A^{\frac{1}{1+b}} \quad . \tag{3}$$

Hence it follows that in the language system a language level is defined by its parameters. Let us stress that for linguists this is not a surprise, because in the very distant past the term *hapax legomenon* was introduced as a concept of relevance to texts and lexical units. In the discussed text theory it is expressed as $y_1 = A$, and thus it is valid for each language level. On the supra-sentence level, parameter A represents constructs defined by lexical units approximately occurring with $x = 1$, i.e. with frequency 1. The respective supra-sentence constructs contain one sentence only.

Formula (3) defines the point of stability for the respective level, or the attractor around which the units of the level are clustered. One should not forget that in practical applications quantities A and b are estimated from the observed distributions obtained from individual texts; this means that their values are burdened with observational faults.

When text analyses include not only two levels corresponding to constructs and their constituents, but a sequence of adjacent levels, first, it is necessary to distinguish the levels by subscripts, and second, to construct a chained fraction. Let us transcribe formula (1) with x_1 instead of y and x_2 instead of x. This means that x_1 represents the lower level of constituents, and x_2 the higher level of the constituent-related construct. Similarly, the higher levels with the appropriate subscripts are defined according to the law. Let us assume five levels where subscript 1 represents the lowest level, for example, phones. Their logarithmic form is as follows:

$$\begin{aligned}
\log x_1 &= \log A_1 - b_1 \log x_2 \\
\log x_2 &= \log A_2 - b_2 \log x_3 \\
\log x_3 &= \log A_3 - b_3 \log x_4 \\
\log x_4 &= \log A_4 - b_4 \log x_5 \\
&\ldots
\end{aligned}$$

They can be chained into the following single equation:

$$\log x_1 = \log A_1 - b_1 \log A_2 + b_1 b_2 \log A_3 - b_1 b_2 b_3 \log A_4 + b_1 b_2 b_3 b_4 \log x_5 \ . \tag{4}$$

From the configuration of the quantities in (4) it is evident what the respective general formula for m levels looks like:

$$\begin{aligned}
\log x_1 = \log A_1 - b_1 \log A_2 + b_1 b_2 \log A_3 - \ldots \pm b_1 b_2 \ldots b_{m-2} \log A_{m-1} \pm \\
\pm b_1 b_2 \ldots b_{m-1} \log x_m \ , \ i = 1,2,\ldots,m, \quad m > 1.
\end{aligned} \tag{5}$$

Now let us suppose the transformation of the basic formula understanding construct x as a function of constituent y:

$$x = \left(\frac{A}{y} \right)^{\frac{1}{b}} \ .$$

If again y is substituted by symbol x as a general label of an arbitrary language level, we obtain the same point of stability as for the ascending movement between the levels:

$$x = A^{\frac{1}{b+1}} \ ,$$

see (3).

When subscripts are added, this time $i = 1$ means the highest level, for example the supra-sentence level of LC, and the ascending higher subscripts indicate the lower levels. The same procedure as above begins with the equation:

$$x_1 = \left(\frac{A_1}{x_2} \right)^{1/b_1}$$

and ends with the general equation which is analogous to (5):

$$\log x_1 = b_1^{-1} \log A_1 - (b_1 b_2)^{-1} \log A_2 + \ \ldots \ \pm (b_1 b_2 \ldots b_{m-1})^{m-1} \log A_{m-1} \pm$$
$$\pm (b_1 b_2 \ldots b_{m-1})^{-1} \log x_m \quad , \ i = 1,2,\ldots,m, \quad m > 1 . \qquad (6)$$

It is easy to compare (5) and (6); their differences consist of the subordination of A to the power expressed by parameter b and by the reversed products of bs in (6), i.e. in the case of descending movement among the levels. The terms on the right-hand side of the general equations (with the exceptions of the last terms) represent modular expressions constructed with the help of the parameters. The *modular measure* for mth level is the product $(b_1 b_2 \ldots b_m)^{-1}$, or $(b_1 b_2 \ldots b_{m-1})$ in the case of (6).

The inverse relationship of bs in those modules is given by the basic orientation of the movement between the levels. This orientation is given as one of the symmetric characteristics of the language system. In linguistic analyses, the orientation is an object of the analyst's choice. As a linguistic fact it can be understood as a trait of compactness of the language system. The differences between (5) and (6) are similar to the differences between the right or left orientation of threads: the direction of the movement of the globe remains the same, but the label of its direction depends on the location of the North Pole. Consequently, text structure appears to be symmetric with respect to the moving direction between the levels: upstairs or downstairs.

Example:

In a Turkish text selected at random the following values of parameters and variables were observed:

Level	A	b	Construct/units vs. constituents/units
1	73.6	0.0746	Aggregate/sentence vs. sentence/phone
2	30.2	0.2785	Sentence/clause vs. clause/phone
3	7.6	0.0634	Clause/word vs. word/phone
4	2.6	0.0991	Word/syllable vs. syllable/phone

At the same time the following mean values of units were observed:

Mean aggregate/sentence	$x_1 = 3.53$	Corrected values x_i :
Mean sentence/phone	$x_2 = 66.76$	66.99
Mean clause/phone	$x_3 = 26.84$	9.37 (!)
Mean word/phone	$x_4 = 6.92$	6.17
Mean syllable/phone	$x_5 = 2.35$	2.15

The observed parameters and means were substituted into the general formula (6) for the neighboring levels, i.e. the formulae of the type:

$$\log x_1 = b_1^{-1} \log A_1 - b_1^{-1} \log x_2 \quad , \quad \log x_2 = b_2^{-1} \log A_2 - b_2^{-1} \log x_3 \quad ,$$

etc.

In this way, however, certain inequalities instead of equations were obtained. For this reason the observed values of x_i were corrected in order to obtain equations corresponding to the formulae. The above presented corrected values x_i are the result of those operations. Those corrected values, with one exception, insignificantly differ from the observed means, as can be proved by testing. Only the corrected value of x_3 is significantly different. After the review of the observed data we can say that the applied syntactic analysis of sentences appears to be incorrect. When, however, we maintain the observed value of A_2, the correct value of parameter b_2 is 0.0281 instead of 0.2785.

With the corrected b_2 and x_5 = 2.17, the respective general equation (6) turns into an approximate equation.

We can, however, exclude the level of clauses with its observation loaded with too great an error and chain the rest of the levels (for which we change the subscripts: A_3 becomes A_2 , A_4 becomes A_3, etc.; and similarly we change the subscripts of b and x). After that correction we obtain the value 2.18 instead of x_4 = 2.35.

It is evident that this approach is sensitive to the applied structural analysis of texts.

The same text was evaluated on the basis of the general formula (5). The substituted values are equal to the above presented values, only labeling changed:

(Mean syllable) x_1 = 2.35,	A_1 = 2.6,	b_1 = 0.0991.
(Mean word) x_2 = 6.92,	A_2 = 7.6,	b_2 = 0.0634.
(Mean sentence) x_3 = 66.76,	A_3 = 73.6,	b_4 = 0.0746.
(Mean LC) x_4 = 3.53.		

The substitution into (5) gives the following result:

$$0.371067862 \neq 0.414973348 - 0.087288627 + 0.011729481 - 0.000256746 = 0.339157456.$$

The analysis offers, instead of the observed value of x_1 = 2.35, the corrected value 2.18. We can see that the result is the same as above.

3. Transformation of the Units of Measurement

3.1. The relationship between Menzerath-Altmann's law and the units of measurement is an important point of the topic of symmetry. The normal approach to the acquisition of data from texts proceeds in agreement with the following paradigm:

[language constructs at level Q are measured by the number of constituents belonging to level R]

⇒ [units of level R (= constituents of Q) are measured by the number of a lower level S (= constituents of R)].

Q is a higher level than R, and R is higher than S. Under this condition the law "works" in accordance with its formulation. For example, when Q is sentence, R represents the level of words and words are measured by the number of morphs, syllables or phones, each of which is S.

One may presuppose two possible transformations of this measurement:

(a) Both Q and R are measured by S, i.e. both constructs and constituents are measured with the same units being used; for example, sentence and word length are measured by the number of phones, morphs or syllables (the last two levels represent parallel constituents);

(b) Q is measured by S, and R is measured by (R + i) units; for example, sentence is measured by phones, and words as constituents of sentences are measured by morphs which are higher than phones.

In Table 2 the standard measurement of constructs 'sentences' and the distribution of their constituents 'words' are presented. From the whole distribution only its highest data of sentence length from 1 to 8 are taken into account, because those observations

Table 2. Standard measurement of constructs (i.e. sentences) and constituents (i.e. words) in a Turkish text.

Sentence / words	Word / phones	Expected
1	7.40	7.40
2	6.15	6.72
3	6.19	6.36
4	6.28	6.11
5	6.06	5.93
6	6.17	5.78
7	5.71	5.66

are not bent out of shape by fluctuations as is the case with the other parts of the distribution. The expected values are computed according to the Menzerath-Altmann formula with parameters estimated from observation.

As to the the type of transformation (a), an example is presented in Figure 3. The length of each word form of a Turkish text was given in the number of phones and its respective sentence length in the number of phones was coordinated to them. The total number of such numerical pairs was 2087. Those pairs were ordered ascendingly with respect to sentence length. The ordered values were organized into classes of sentence length such that value 1 represents sentence length from 1 to 10 phones, value 2 from 11 to 20, etc. The coordinated values of word length are averages of word length corresponding to each respective class. From Figure 3 it is evident that the mean values slightly depend on sentence length. From the

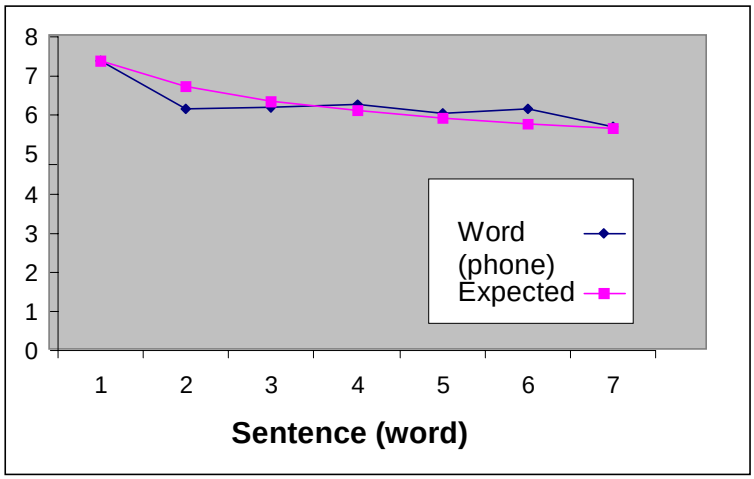

Figure 2. The distribution presented in Table 2.

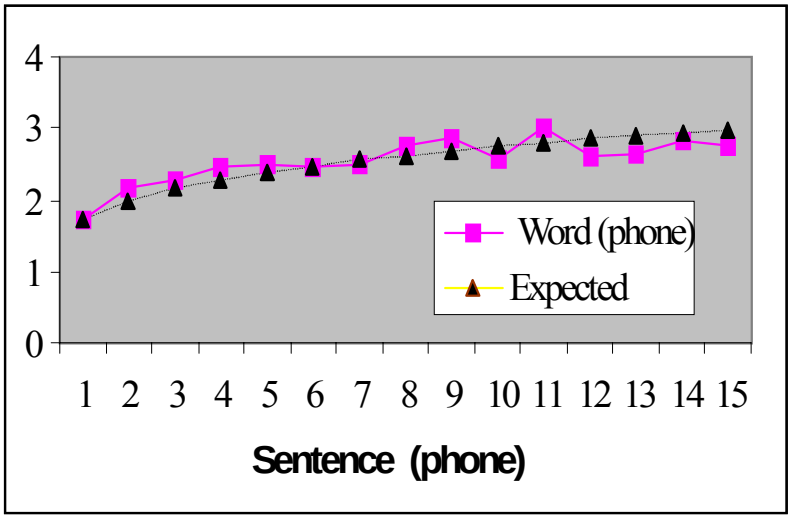

Figure 3. Constructs (sentences) and constituents (words), both measured by phones.

sentence length greater than 8 phones the situation changes and word length in phones begins to fluctuate around the mean word value.

A similar experiment was arranged with the units of measurement as described in (b). The result is presented in Figure 4. The last two figures indicate that the non-standard way of measurement ensues in a behavior of the function which is not in agreement with Menzerath-Altmann's law, or better: with its word-for-word understanding. The function is formulated for *lengths*, or *sizes*, both of constructs and constituents. The transformation of the units of measurement sub (a) and (b), however, darkens the structural relation which is contained in the concepts of construct and its constituents. Menzerath-Altmann's law does not simply refer to some amounts (length, size) of language items; the law is related to phenomena which are nested inside higher phenomena. G. Altmann (2001) in a personal communication formulated this property in the following way: 'If the measurement units are very small (phones), the dependence weakens, the structuring gets lost, the distribution tends to normality', also see Altmann (1988).

It can be concluded that there is no symmetry controlling the appearance of text and the size of its units of different levels without taking their mutual nesting into account. The law cannot be understood literally and without consideration of the structure formed by the levels.

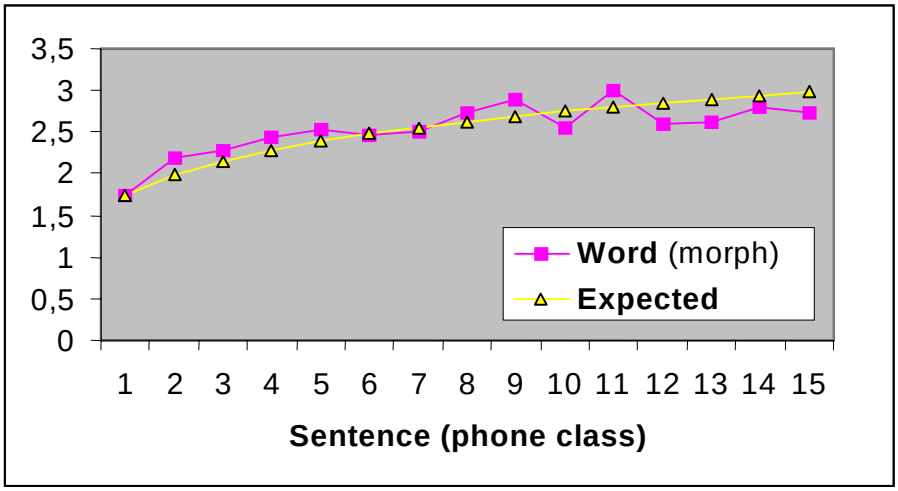

Figure 4. Transformation of the units of measurement of the type (b).
The expected values correspond to: $1.73(\text{Sentence})^{0.2005}$.

4. LC at different levels

4.1. When the ladder of the language levels surpasses sentence and enters the supra-sentence functional space, as we already know, this space has semantic traits and is quite specific in comparison with the lower levels. Irrespective of this fact, Menzerath-Altmann's law functions at all levels, it is the criterion for the existence of the language levels. The differences between the functional spaces came into view when corrections of the input data were required. Can we use a similar operation (like the one executed with the purpose to obtain the level of LC) at the lower levels as well?

First, let us mention what this approach really means: the existence of a list of lexical units (or, a text vocabulary) is assumed; it contains units forming the lexical ground of each text. This list is the result of a semantic interpretation by language users. Each item of the list is then used for the purpose of forming the supra-sentence items LC. They are constructs and their constituents are sentences forming their NC. Consequently, each lexical unit points at one or more sentences of the text.

Now we ask the question whether syllables (or morphemes, the parallel word con-stituents) form *word aggregates* at word level like contextual analogies of LCs do at sentence level; in other words, whether the following lower units:

[a given syllable] ⇒ [word form in which the syllable occurs] ⇒ [a group of words called word aggregate]

can be substituted into the scheme

[a given lexical unit] ⇒ [sentence(s)] ⇒ [supra-sentence construct LC having sen-tence(s) as its constituents in the sense of Menzerath-Altmann's law]

The solution of the problem seems to be: Is the group of words obtained that way a text item, a language construct having the respective words as its constituents?

An analogous question is related to word groups selected with the help of given morphs instead of syllables.

The analytical procedure was as follows: a Turkish text was rewritten as a sequence of syllables; the length of the respective word form *L,* in which the syllable occurs, was attributed to each syllable; then its frequency (= the number of word forms *W* in which the given syllable occurs) was also attributed to each syllable. The result obtained was a distribution of mean values <*L*> with *W*.

Table 3. Mean word length <L> (in syllables)
distributed with word aggregates W (in the number of words)

Word aggregate W	Observed <L>	Expected <L>	Word aggregate W	Observed <L>	Expected <L>
1	3.24	3.20	**10**	3.17	3.68
2	3.30	3.34	**11**	3.68	3.70
3	3.20	3.42	**12**	3.93	3.72
4	3.08	3.48	**13**	3.98	3.74
5	3.04	3.53	**14**	3.70	3.75
6	3.56	3.57	**15**	4.43	3.77
7	3.43	3.60	**16**	4.31	3.79
8	3.56	3.63	**17 –**	3.68	3.80
9	4.11	3.66			

The expected values are computed as $3.20\ W^{0.0609}$.

Table 4. Mean word length <L> (in morphs) distributed with word aggregates W (in the number of words)

Word aggregate W	Observed <L>	Expected <L>	Word aggregate W	Observed <L>	Expected <L>
1	2.50	2.50	**9**	3.73	3.12
2	2.54	2.68	**10**	2.53	3.15
3	2.86	2.79	**11**	2.96	3.18
4	2.81	2.87	**12**	3.17	3.21
5	2.69	2.94	**13**	2.92	3.23
6	3.14	2.99	**14**	3.57	3.26
7	2.82	3.04	**15**	4.33	3.28
8	2.46	3.08	**16 –**	3.59	3.30

The expected values are computed as $2.50\ W^{0.1004}$.

Table 5. Mean sentence length (in phones) distributed with LC (measured by sentences) in the same Turkish text analyzed in Tables 3 and 4

LC/ sentences	Mean sentence / phones	Expected
1	73.62	73.62
2	72.32	69.91
3	68.25	67.82
4	62.08	66.38
5	75.00	65.29
6	62.91	64.40
7	55.29	63.67
8	64.67	63.04

The expected values are computed as $72.62(LC)^{-0.0746}$.

An analogous operation was accomplished with the same Turkish text rewritten as a sequence of morphs.

The results of both observations are presented in Tables 3 and 4. We can see that – irrespective of the fluctuations - the observed values of $\langle L \rangle$ slightly increase in both Tables. This is evident when the distributions of aggregates in Tables 3 and 4 are compared with the distribution of LC presented in Table 5 which, on the contrary, evidently abides by Menzerath-Altmann's law. The obtained *word aggregates*, based on syllables or morphs, do not abide by this law. Expressed in a condensed way, there are no such items in text structures at all.

Let us note that similar transposition of the scheme can be applied to the individual phones ⇒ syllables or morphs ⇒ *syllable* or *morph aggregates*. Soon after the beginning of the appropriate experiment it has become clear that there are no such aggregates in text structures as well.

4.2. The above presented negative result seems to be of a certain importance for the treatment of the general text structure. Of course, strictly formulated, it has been proved for Turkish texts only (the same result was obtained from a corpus of nine Turkish texts). The asymmetry, however, testified in one language stands in contrast to the existence of LCs in many different other languages, see the above quoted references, and it can be conjectured that the same asymmetry is proper to the general text structure.

The approach to language structures by structuralists was based on the *emic* treatment of language units belonging to different levels. Perhaps this treatment can be taken as a sort of

reductionism; its principle is based on the observations of a certain similarity between language levels when language formants are classified. Now we can see that there is one substantial difference inside this similarity. *Lexeme* on the one hand, and *morpheme* (and *syllableme*?) on the other are to be interpreted quite diversely.

Thus syllables as well as morphs cannot be treated as units forming analogous subsets of words as is the case with LCs consisting of sentences. The sets of lexical units forming the vocabularies of texts have a specific position in the language structures. This position is based on a certain abstraction of meanings, on a semantic interpretation of word forms and their contexts. It encompasses the denotation of synonyms and homonyms together with the other semantic identities and differences during its production and reception. It is important that the theory enables us to assume a certain analogy between word connections in texts and in language users' human minds. In the linguistic sense, the discussed difference is the difference between the phase of grammar (together with its semantics) and the phase of semantics which is based on interpretation and on the intuition of minds operating with meanings. Regardless of the asymmetry and the fact that the respective level is understood as a field of brain functioning, this higher phase of the text structure belongs to language and is an object of linguistic analyses.

Both symmetries and asymmetries represent a superstructure permitting a better understanding of the character of text structures.

5. Vertical vs. horizontal symmetry

The last transformation to be discussed in the present paper is the transformation of the unit 'text', taken as an organized agglomeration of language formants, into the sequence of such language formants as it really occurs in texts. These two viewpoints can be imagined as vertical (agglomerative) and horizontal (sequential) comprehension. The sequential viewpoint was analyzed in Hřebíček (2000), and here we restrict ourselves to recalling the main results of this investigation.

At an arbitrary level from phones up to sentences, any text can be rewritten as a sequence of the respective units. Thus text is described as a sequence of phones, morphs or syllables, word forms, clauses and sentences. Parallel numerical sequences were then created, each value indicating the respective size of a given unit of a sequence. Each unit has a rank number according to its position in the respective sequence.

Let us define *distance* as the absolute value of the difference between the rank numbers of the identical units located in the sequences in the closest positions of pairs to each other. Then the distances of the identical values in sequences were investigated and evaluated as distributions of mean sizes with distances. The statistical experiments led to the conclusion that regardless of the chaotic appearance of the individual sequences, Menzerath-Altmann's law also works when texts are transformed into horizontal forms.

This refers to all language levels up to sentences. And what about the supra-sentence level of LCs? An LC is a unit of specific character, it is not an additive kind of unit, one cannot measure text length in the number of LCs, as it is possible with phones, morphs, etc. up to sentences (with the exception of the imbedded syntactic constructions). Therefore we took several individual LCs based on more frequent lexical units separately; the distances between the next occurrences of a respective lexical unit in the sequence was measured. And surprisingly, the distributions of distances abide by Menzerath-Altmann's law.

Thus we can conclude that text structure described by Menzerath-Altmann's law is symmetrical under the transformation from vertical to horizontal form.

6. Conclusions

The investigation of the character of text structure concerns Menzerath-Altmann's text model and its automorphisms.

Mathematical formalism of Menzerath-Altmann's law qualifies its basic idea to be an automorphic property. This means that in the relationship between language constructs and constituents neither is prior or superior to the other. This is not a surprise, such quality is intuitively expected.

Another automorphism can be detected when assumptions do not center around two levels only, but more than two or the whole string of the levels occurring in texts. Menzerath-Altmann's law holds for such strings regardless of the two opposite viewpoints taken up by the observer: from a lower level to a higher one, or *vice versa*.

The text model discussed is not symmetric under the transformation of the units of measurement. This structural property is connected with the nesting of text units inside higher units, irrespective of the symmetric character of this relation. Measurements applied to text analyses must consider this fact.

The supra-sentence level of LCs constitutes specific asymmetry which is explained by the contextual aptitude of lexical units. It has been proved that there are no LCs at the level of word forms.

The automorphism between vertical and horizontal (sequential) treatments of texts is also based on Menzerath-Altmann's law and it testifies the importance of this law. It represents a principle based on meanings and the fact that it is valid not only for the vivacious supra-sentence structures being fully controlled by language users, but also for the ossified levels regulated by the grammatical and other types of language rules.

One evident symmetry has not been discussed: the one between the position of text producer and text recipient. Both language users deal with text interpretation, but symmetry is not complete, very often the results of interpretation (= understanding texts) are not identical in real language communication.

References

Altmann, G. (1980). Prolegomena to Menzerath´s law. *Glottometrika 2: 1-10.*

Altmann, G. (1988). Verteilungen der Satzlängen. *Glottometrika 9*, *147-170*.

Altmann, G. & Best, K.-H. (1996). Zur Länge der Wörter in deutschen Texten. *Glottometrika 15*, *166-180.*

Altmann, G. & Köhler, R. (1995). "Language forces" and synergetic modelling of language phenomena. *Glottometrika 15*, *62-76.*

Altmann, G. (2001). Personal communication.

Altmann, G. & Schwibbe, M. H. et al. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim-Zürich-New York: Olms.

Best, K.-H. (1996). Zur Wortlängenhäufigkeit in schwedischen Pressetexten. *Glottometrika 15*, *147-157*.

Best, K.-H. (2001). *Quantitative Linguistik.* (Göttinger Linguistische Abhandlungen 3.) Göttingen: Peust & Gutschmidt.

Hřebíček, L. (1989). The Menzerath-Altmann law on the semantic level. *Glottometrika 11, 47-56.*

Hřebíček, L. (1992). *Text in communication: supra-sentence structures*. Bochum: Brockmeyer.

Hřebíček, L. (1995). *Text levels. Language constructs, constituents and the Menzerath-*

Altmann law. Trier: Wissenschaftlicher Verlag Trier.
Hřebíček, L. (1997). *Lectures on text theory*. Prague, Oriental Institute.
Hřebíček, L. (2000). *Variation in sequences. (Contributions to general text theory)*. Prague: Oriental Institute.
Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.
Leopold, E. (2001). Fractal structures in language. The question of the imbedding space. In: Uhlířová, L., Wimmer, G., Altmann, G. & Köhler, R. (eds.): *Text as a linguistic paradigm: levels, constituents, constructs: 163-176.* Trier: Wissenschftlicher Verlag Trier.
Menzerath, P. (1928). Über einige phonetischen Probleme. *Act du premier congrès international de linguistes: 104-105*. Leiden, Sijthoff,
Menzerath, P. (1954). *Die Architektonik des deutschen Wortschatzes.* Bonn: Dümmler.
Weyl, H. (1952/1968). *Symmetry.* Oxford University Press. [The Russian edition is quoted: *Simmetriya,* Moscow, Nauka, 1968].
Wimmer, G., Grotjahn, R., Köhler, R., & Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics 1, 98-106*.
Wimmer, G. & Altmann, G. (1996). The theory of word length: some results and generalizations. *Glottometrika 15, 112-133.*

Software:

Altmann-Fitter (1994). Lüdenscheid, RAM-Verlag.

Two Turkish texts used for illustration in tables and figures:

Necati Cumalı (1979). İnsanlar Kardeştir. In: N. Cumalı, *Revizyonist: 42-46.* Istanbul: Tekin.
Necati Cumalı (1973). *Ya˘gmurlar ve Topraklar: 5-15.* Istanbul: Cem.

Glottometrics 2, 2002, 33-44

Der altrussische Jerwandel

Werner Lehfeldt, Göttingen[1]
Gabriel Altmann, Lüdenscheid

Abstract. This article shows that the fall of the reduced vowels (ь, ъ) in weak positions in Old Russian is the consequence of a process controlled by Menzerath´s law holding for the relationship between construct (phonetic word) and constituent (syllable). That process reduces excess redundancy.

Key words: jers, Old Russian, Menzerath´s law

1. In der Geschichte sämtlicher slavischer Sprachen hat sich vor ca. 1000 Jahren ein Lautwandel abgespielt, der zu einem Umbau der einzelsprachlichen Phonemsysteme geführt und auch die Phonemkombinatorik tiefgreifend umgestaltet hat. Gemeint ist der in der Überschrift genannte Jerwandel. Dieser Wandel betraf die beiden Vokalphoneme vorderes Jer (ь) und hinteres Jer (ъ), die häufig auch als „reduzierte Vokale“ bezeichnet werden. Bei ihnen handelte es sich im späten Gemeinslavischen um „basically high lax vowels“ (Lunt 1974: 26), „überkurze Laute“ (Bräuer 1961: 111). In sogenannter schwacher Position (im weiteren mit – gekennzeichnet) verstummten ь und ъ, wohingegen sie in sogenannter starker Position (im weiteren mit + gekennzeichnet) mit anderen Vokalen zusammenfielen. Von diesem Zusammenfall waren in den slavischen Einzelsprachen unterschiedliche Vokale betroffen, woraus schon ersichtlich ist, daß die Folgen des Jerwandels bis heute die Physiognomie der slavischen Sprachen wesentlich mitprägen. „Die gesamte historische Epoche der Entwicklung der slavischen Sprachen stellt in ihren bestimmenden Zügen die Entfaltung der Prozesse dar, die mit den Folgen des Jerwandels verknüpft waren“ (Gasparov, Sigalov 1974: 187). Betrachten wir hier zur Verdeutlichung lediglich das Russische und das Serbokroatische. Im Russischen fiel ь in starker Position mit *e*, ъ hingegen mit *o* zusammen, während im Serbokroatischen beide Jers mit *a* zusammenfielen; vgl. folgende Beispiele:

Altruss.	Russ.	Serbokr.	Bedeutung
dьnь	d’en’	dan	Tag
+ –			
sъnъ	son	san	Schlaf
+ –			
otьcь	ot’ec	otac	Vater
+ –			
nizъkъ	n’izok	nizak	niedrig
+ –			
tьmьnica	t’emn’ica	tamnica	Gefängnis
+ –			

[1] Address correspondence to: W. Lehfeldt, Seminar für Slavische Philologie, Humboldtallee 19, D-37073 Göttingen. E-mail: wlehfel@gwdg.de

Angesichts der grundlegenden Bedeutung des Jerwandels für die Lautgeschichte sämtlicher slavischer Sprachen erstaunt es nicht, daß die slavistische Sprachwissenschaft sich seit langem darum bemüht, ihn so detailliert wie möglich zu erfassen, in seinem Verlauf und in seinen Konsequenzen zu beschreiben und für ihn eine Erklärung zu finden. Während dieser Sprachwandel als umfassend dokumentiert und gut beschrieben gelten kann (vgl. z.B. Markov 1964; Sidorov 1966; Zaliznjak 1993), ist die Erklärungsproblematik bis heute umstritten. Es kann an dieser Stelle nicht darum gehen, sämtliche bisher vorgelegten Erklärungsversuche aufzuführen und zu analysieren. Wir wollen nur diejenigen Ansätze näher betrachten, die sich, wie das auch in diesem Beitrag geschehen soll, quantitativer Untersuchungsmethoden bedienen.

(a) Der vermutlich erste Forscher, der eine „statistische Interpretation" des Jerwandels vorgelegt hat, war R. Abernathy (1956). Der Autor geht von einer Hypothese zur Sprachevolution aus, derzufolge der Entropiewert $H = - K\Sigma\ p_i\ log\ p_i$ einer bestimmten Menge sprachlicher Signale dazu tendiert, im Laufe der Zeit immer geringer zu werden. Anschließend berechnet er aufgrund einer Untersuchung von Auszügen aus dem altkirchenslavischen Codex Zographensis die *H*-Werte für die Vokale jeweils in starker und in schwacher Position. Weiter werden für beide Positionen die verschiedenen *H*-Werte bestimmt, die sich ergeben, wenn ъ und ь mit irgendeinem der anderen Vokale einschließlich Zero zusammenfallen. Dabei erweist sich, daß *H* dann den geringsten Wert annimmt, wenn ъ/ь in starker Position mit *o*/*e* zusammenfallen und in schwacher Position verstummen. Genau dieser Prozeß hat sich im Makedonischen tatsächlich abgespielt. Der Wandel also, der stattgefunden hat, scheint darauf gerichtet gewesen zu sein, *H* zu minimieren, was der von R. Abernathy zu Beginn seiner Arbeit formulierten Hypothese entspricht.

Abgesehen davon, daß die genannte Hypothese zur Sprachevolution generell als falsifiziert zu gelten hat (die Entropie hängt vom Inventarumfang ab; vgl. Altmann, Lehfeldt 1980: 166-178), wirft Abernathys Versuch, den Jerwandel zu erklären, noch einige weitere Fragen auf. Die von dem Autor angestellten Berechnungen gelten unter der Voraussetzung, daß es gerade die Jers sind, die sich ändern. Warum aber haben sich diese Vokale überhaupt verändert? Wenn die behauptete Tendenz tatsächlich gültig sein sollte, so müßte man überprüfen, ob sich nicht noch kleinere *H*-Werte ergeben, wenn man — hypothetisch — andere Vokale sich verändern läßt, wenn man also beispielsweise *a*/*ě* mit ъ/ь in starker Position zusammenfallen und in schwacher Position verstummen läßt. Solange eine solche Untersuchung nicht durchgeführt ist, haben wir bestenfalls eine Erklärung dafür, daß sich ъ/ь so verändert haben, wie sie sich im Makedonischen verändert haben, nicht aber dafür, weshalb sie sich überhaupt verändert haben (vgl. auch noch Abernathy 1963).

Das von R. Abernathy für die Untersuchung des Codex Zographensis angewandte Verfahren wurde 1964 von V. V. Kolesov auf die Analyse einiger altrussischer Sprachdenkmäler aus dem 10. und dem 11. Jahrhundert übertragen. Kolesov akzeptiert auch Abernathys allgemeine Hypothese zur Sprachevolution. Die von ihm erhobenen Daten sind aber in einer Reihe von Fällen geeignet, die Behauptung von Abernathy in Frage zu stellen, daß mit Hilfe seines Ansatzes die Art und Weise der Veränderung der Jers erklärt werden könne. So kommt manchmal der *H*-Wert für den hypothetischen Zusammenfall von ъ/ь mit *a*/*ě* in starker Position demjenigen für den tatsächlichen Zusammenfall von ъ/ь mit *o*/*e* sehr nahe, und so fragt man sich, warum ъ/ь statt mit *o*/*e* nicht mit *a*/*ě* zusammengefallen sind. Eine weitere Schwierigkeit für Abernathys Erklärungsmodell ergibt sich daraus, daß für die schwache Position *H* dann den geringsten Wert annimmt, wenn ъ/ь nicht, wie es tatsächlich geschehen ist, verstummen, sondern mit *i*/*y* oder mit *o*/*e* zusammenfallen. Entgegen der ursprünglichen Absicht ihres Autors ist die datenreiche Arbeit von Kolesov eher geeignet, Abernathys Vorschlag zu unterminieren, statt ihn zu bekräftigen.

(b) Noch weniger als der Ansatz von R. Abernathy vermag die „kybernetische“ Analyse von N. D. Rusinov (1979) zu überzeugen. Dieser Autor berechnet für das altrussische Sprachdenkmal Izbornik 1076 g. verschiedene Entropie- und Redundanzwerte, ohne daß dazu irgendeine Hypothese formuliert worden wäre. Erst anschließend wird der Jerwandel „kybernetisch“ interpretiert. Rusinov stellt fest, daß die Jers nur in schwacher Position eine hohe Auftretenswahrscheinlichkeit aufgewiesen hätten, während sie in starker Position selten verwendet worden seien. Dies habe das Schicksal von ъ und ь besiegelt; denn die Jers in starker Position hätten wegen ihrer negativen Redundanz die in schwacher Position stehenden nicht „unterstützen“ können, weshalb letztere „fortgefahren“ hätten, sich abzuschwächen und somit „zum Verschwinden verurteilt“ gewesen seien, während ъ und ь in starker Position „unausweichlich“ der analogischen Einwirkung anderer, artikulatorisch ähnlicher Vokale anheimgefallen seien. Keine dieser Behauptungen wird auch nur ansatzweise mit einer Begründung versehen.

2. Die vorliegende Untersuchung bewegt sich, ebenso wie die referierten Untersuchungen, im Rahmen der Quantitativen Linguistik, schlägt aber von vorneherein einen grundsätzlich anderen Weg ein. Wir werden uns hier auf den Jerwandel im Altrussischen beschränken, wo dieser Vorgang gegen Ende des 12. Jahrhunderts abgeschlossen war. Für andere slavische Sprachen stehen uns bisweilen noch keine entsprechenden Daten zur Verfügung.

Den Ausgangspunkt unserer Überlegungen bildet die — seit langem bekannte — Einsicht, daß der Jerwandel zu einem tiefgreifenden Umbau der Silbenstruktur geführt hat. Vor diesem Wandel war im Altrussischen der zulässige Silbenumfang durch zwei Faktoren relativ beschränkt:

(a) Auf den vokalischen Silbenkern konnte kein Konsonant folgen, d.h., alle Silben waren offen. Konsonanten und Konsonantenkombinationen kamen nur vor dem Silbenkern vor: CV, CCV, CCCV, z.B o|tь|cь| ‘Vater’, pro|stъ|ļu|di|nъ| ‘einfacher Mensch’, kъ|to| ‘wer’, lъ|žь|kъ| ‘Löffel’ G. Pl., pri|tъ|ča| ‘Unglücksfall’, spě|ši|ti| ‘eilen’, stra|chъ| ‘Furcht’.

(b) Die Kombinierbarkeit der Konsonanten in prävokalischer Position unterlag einer strengen Regelhaftigkeit, durch die sehr viele Konsonantenverbindungen ausgeschlossen wurden „Zugelassen waren nur Folgen mit steigendem phonematischem ‘Rang’ des Konsonanten — vom Spiranten zum Explosivkonsonanten und vom Explosivkonsonanten zum Sonor“ (Gasparov, Sigalov 1974: 73), d.h., in dreielementigen Konsonantenverbindungen war nur die Anordnung Spirant–Explosivkonsonant–Sonor möglich, während in zweielementigen Verbindungen die Anordnungen Spirant–Explosivkonsonant, Spirant–Sonor und Explosivkonsonant–Sonor in Frage kamen; vgl.: vgl. stra|chъ| ‘Furcht’, slo|vo| ‘Wort’, sta|ti| ‘werden’, ve|sti| ‘führen’, spě|chъ| ‘Eile’, pra|vь|da| ‘Wahrheit’.

Durch diese Beschränkungen waren die Silben des Altrussischen im Durchschnitt relativ kurz und standen überhaupt nur relativ wenige Silben zur Verfügung.

Der geschilderte Umstand hat zur Folge, daß die Wortformen des Altrussischen, um „durch genügend Redundanz auch unter Störungseinfluß“ (Altmann, Schwibbe 1989: 6) von den jeweils anderen Wortformen bzw. Taktgruppen hinreichend unterscheidbar zu sein, im Durchschnitt relativ viele Silben umfassen müssen. Da aber die Silbe eine recht rigide Struktur hat, erzeugt sie auf ihrer Ebene allzuviel Redundanz, d.h., das Anstrengungsgleichgewicht zwischen Artikulation und Dekodierung (d.h. Sprecher vs. Hörer) ist stark zugunsten des Hörers ausgeprägt. In derartigen extremen Fällen kann man mit der Zeit (die sehr lang sein kann)

eine Veränderung direkt an der kritischen Stelle (hier der Silbenstruktur) oder an einer anderen, die mit dieser funktional verbunden ist, erwarten. Mit anderen Worten, an dieser Stelle entsteht im Altrussischen eine selbstorganisierte Kritikalität, die zu einer „Katastrophe“ führen muß (vgl. Bak 1999).

Der zweite Umstand, der zu einer Änderung antreibt, ist die Tatsache, daß bei einem störungsfreien Gleichgewicht zwischen Konstrukt- und Komponentengröße (hier Wort- und Silbenlänge) die einfache Form des Menzerathschen Gesetzes gelten muß (vgl. Altmann 1980; Altmann, Schwibbe 1989), das durch eine *monoton fallende* Potenzkurve ausgedrückt wird. Gemäß diesem Gesetz gilt, daß, je größer bzw. komplexer ein sprachliches Konstrukt ist, seine Konstituenten um so kleiner bzw. einfacher sind. Den uns interessierenden Spezialfall dieses Gesetzes können wir mit den Worten von Menzerath selbst formulieren: „Die relative Lautzahl nimmt mit steigender Silbenzahl ab, oder mit anderer Formel gesagt: je mehr Silben ein Wort hat, um so (relativ) kürzer (lautärmer) ist es“ (Menzerath 1954: 100).

Es ist wichtig, sich von vorneherein klarzumachen, daß, wie jedes Gesetz, so auch das Menzerathsche Gesetz nur unter der Voraussetzung gültig ist, daß die notwendigen Bedingungen gegeben sind. Auf den uns interessierenden Fall angewandt, bedeutet dies, daß bei Zunahme der Silbenzahl von Wortformen der durchschnittliche Silbenumfang nur dann monoton sinken kann, wenn die Regeln der Phonemkombinatorik die Bildung hinreichend vieler voneinander unterschiedener Silben gegebener Größe zulassen. Wenn das nicht der Fall ist, wird die monotone Verringerung des jeweiligen durchschnittlichen Silbenumfangs verhindert. Wenn wir uns die strenge Reglementierung der Konsonantenkombinatorik im Altrussischen vor dem Jerwandel in Erinnerung rufen, so wird klar, daß wir für diese Periode nicht ohne weiteres die Gültigkeit des *einfachen* Menzerathschen Gesetzes postulieren dürfen. Auf der anderen Seite kann man sich nur schwer vorstellen, daß es im Altrussischen einen einfachen Attraktor gegeben haben sollte, der mathematisch nicht modellierbar wäre, oder daß wir es hier mit Strukturen zu tun haben sollten, die keinem Gesetz folgen, denn „everything abides by laws“ (Bunge 1977: 17). Eher vermuten wir, daß das Altrussische einer anderen, durch extreme Randbedingungen hervorgerufenen Silbenstrukturdynamik gehorchte.

Tatsächlich müssen wir im Falle der altrussischen Silbenstruktur vor dem Jerwandel von vorneherein mit einer gewissen Beschränkung des von der Menzerathschen Version beschriebenen Mechanismus durch empirische Randbedingungen rechnen: Die Menge der einsilbigen Wortformen ist relativ klein, und zwar erstens wegen der genannten Restriktionen, denen die Konsonantenkombinierbarkeit unterliegt, und zweitens, spezieller, deshalb, weil einsilbige Wortformen im wesentlichen auf die Struktur V bzw. CV beschränkt sind (vgl. allerdings die Präpositionen *pro* und *pri* sowie solche Aoristformen wie *sta*, *spě* , *ply*, *plu*, *slu*, *zna*, *kri*). Die durchschnittliche Silbenlänge, gemessen als Anzahl der Phoneme, wird also vermutlich im Maximalfall 2.00 betragen, in der Regel wohl noch unter diesem Wert liegen. Bei zwei- und mehrsilbigen Wortformen kommen auch Silben der Struktur CCV und CCCV vor, so daß insbesondere bei Wortformen, die jeweils nur wenige Silben umfassen, die durchschnittliche Silbenlänge größer als 2 sein kann. Wir können also nicht, wie schon erwähnt, von vorneherein mit einem vollständig monotonen Verlauf des durch die Menzerathsche Hypothese postulierten Trends rechnen, das ist erst ab der zweiten Position — Wortformen aus zwei Silben — zu erwarten.

Das Verstummen von ъ und ь in schwacher Position — der sogenannte Jerausfall — war bekanntlich für die Konsonantenkombinatorik und damit für die Silbenstruktur von großer Bedeutung; denn

(a) entstanden jetzt geschlossene Silben, d.h., kamen Konsonanten und Konsonantenkombinationen auch nach dem vokalischen Silbenkern vor — vgl. dь|nь| >

d'en'|, sъ|nъ| > son|, o|tь|cь| > o|t'ec|, ni|zъ|kъ| > n'i|zok|, tь|mь|ni|ca| > t'em|n'i|cą|—, und

(b) entstanden jetzt viele Kombinationen von Konsonanten, die vorher „verboten" gewesen waren (vgl. dazu Gasparov, Sigalov 1974: 182) — vgl. beispielsweise Kombinationen aus zwei Verschlußkonsonanten wie in pt'i|ca| < pъ|ti|ca|, kto| < kъ|to|, aus zwei Spiranten wie in scho|d'i|t'i| < sъ|cho|di|ti|, aus Explosivkonsonant und Spirant wie in psy| < pь|sy|, aus Explosivkonsonant und Nasal wie in o|kno| < o|kъ|no|, aus zwei Nasalen wie in mno|go| < mъ|no|go|, aus Sonant und Explosivkonsonant wie in lga|t'i| < lъ|ga|ti| usw.

Dieser Umstand bewirkte, daß nach dem Jerausfall viel mehr längere Silben vorhanden waren, wodurch das Silbeninventar des Russischen insgesamt vergrößert wurde. Dies wiederum ermöglichte die Bildung von relativ vielen kurzen Wortformen mit jeweils relativ langen Silben: „In kurzen Konstrukten müssen … längere Konstituenten verwendet werden" (Altmann, Schwibbe 1989: 6). Man wird also mit einer Verschiebung zu rechnen haben: Die Menge der einsilbigen Wortformen wird auf Kosten der bisher zweisilbigen Wortformen anwachsen, und anwachsen wird auch die durchschnittliche Länge der Silben einsilbiger Wortformen. Der Schwund, den die bisher zweisilbigen Wortformen durch diesen Vorgang erleiden, wird jedenfalls bis zu einem gewissen Grad dadurch kompensiert, daß ihr Inventar durch Zuzügler aus der Menge der bisher dreisilbigen Wortformen aufgefüllt wird; vgl. etwa ni|kъ|to| > n'i|kto|, ži|vo|tъ| > ži|vot| usw. Auch hier wird die durchschnittliche Silbenlänge anwachsen. Analog wird es sich bei den weiteren Längen verhalten.

Wir gelangen durch die vorgetragenen Beobachtungen und Überlegungen zu folgender Hypothese: Eine wichtige Folge des Jerausfalls besteht darin, daß die durchschnittliche Silbenlänge von Wortformen, die aus i (i = 1, 2, 3,…) Silben bestehen, signifikant erhöht wird. Wir dürfen jetzt auch mit einem vollständig monotonen Verlauf des Trends rechnen, d.h. ab der ersten Position, da bereits einsilbige Wortformen zahlreiche Konsonantenkombinationen aufweisen können und tatsächlich auch aufweisen.

Mit Anomalien und Randbedingungen/Restriktionen, die den monotonen Verlauf der Kurve stören, wurde bereits in der ursprünglichen Ableitung gerechnet, wo Altmann (1980) einen „Störfaktor" a in die Differentialgleichung einbaute, d.h.

$$\frac{dy}{y} = -\left(a + \frac{b}{x}\right)dx \tag{1}$$

setzte und

$$y = K\, x^{-b} e^{-ax} \tag{2}$$

erhielt. Zieht man jedoch noch weitere Restriktionen in Betracht, so scheint heutzutage eher die Hypothese „die Silbenlänge ist eine Funktion der Wortlänge" zu gelten. Um jedoch auch solch extremen Bedingungen wie denjenigen im Altrussischen gerecht zu werden, führen wir in die Differentialgleichung (1) auf der rechten Seite noch ein weiteres Glied für phonemkombinatorische Restriktionen ein und bekommen somit einen Spezialfall des Ansatzes von Wimmer, Köhler, Altmann (2003), der auch die Theorien von Naranan und Balasubrahmanyan (z.B. 1998) umfaßt und bereits von Geršić und Altmann (1988) zur Modellierung der Vokaldauer benutzt wurde, nämlich

$$\frac{dy}{y} = -\left(a + \frac{b}{x} + \frac{c}{x^2}\right)dx \tag{3}$$

deren Lösung

$$y = Kx^{-b}e^{-ax}e^{c/x} \quad (4)$$

ergibt.

3. Um unsere Hypothese zu überprüfen, wurde folgende vergleichende Untersuchung ins Werk gesetzt. Aus dem im Jahre 1056 vollendeten Ostromir-Evangelium (OE), dem bis vor kurzem (vgl. Zaliznjak, Janin 2001) ältesten überlieferten russisch-kirchenslavischen Sprachdenkmal, das den Zustand vor dem Jerwandel ziemlich getreu widerspiegelt, wurden anhand der Ausgabe von Vostokov (1843) die ersten 25 Seiten ausgewertet. Ähnlich wie in den zu Eingang referierten quantitativen Untersuchungen zum Jerausfall bildete die sogenannte Taktgruppe die Analyseeinheit. Im Normalfall umfaßt eine Taktgruppe eine einzige Wortform. Wenn aber eine Wortform von einem Klitikon oder mehreren Klitika umgeben ist, dann bildet sie zusammen mit diesen eine Taktgruppe bzw. — in anderer Terminologie — ein phonologisches Wort; vgl. etwa на слово, слово же, и не за князя, и не по закону ли (vgl. Zaliznjak 1985: 119). Der Text mußte also im ersten Untersuchungsschritt in Taktgruppen zerlegt werden. Dies wiederum erforderte eine Bestimmung der Klitika des Altrussischen. Eine solche Bestimmung ist von A.A. Zaliznjak (1985: 145 f.) vorgenommen worden. Sie wurde unserer Un-tersuchung zugrundegelegt. Zur Verdeutlichung unseres Vorgehens sei die Taktgruppenzerlegung des Anfangs des OE angeführt:

> Iskoni | bě | slovo | i-slovo | bě | otъ-boga | i-bogъ | bě | slovo | se | bě | iskoni | u-boga | i-těmь | vsę | byšę | i-bez-nego | ničьtože | ne-bystь | ježe | bystь | vъ-tomь | životъ | bě | i-životъ | bě | světъ | člověkomъ | i-světъ | vъ-tьmě | svьtitь-sę | i-tьma | jego | ne-obętъ | …

Im Anschluß an die Taktgruppenzerlegung konnten für jede Taktgruppe deren Silbenzahl und Phonemzahl ermittelt werden. Dazu wurde zunächst folgende Textmodifizierung vorgenommen: Wenn in einer Taktgruppe ein etymologisches *ь* oder *ъ* „fehlte“, d.h., wenn sich der Beginn des Jerausfalls graphisch manifestierte, wurde *ь* oder *ъ* restituiert, sofern die gleiche Taktgruppe oder eine Taktgruppe mit dem gleichen Morphem auch mit *ь* oder *ъ* im Text belegt ist; vgl. in der zitierten Anfangspassage die Form *vsę*, die zu *vьsę* umgewandelt wurde, weil Formen wie *vьsi*, *vьsěmi*, *vьsěxъ*, *vьsę* u.a. reichlich belegt sind.

Eine weitere Modifikation bestand darin, daß sämtliche Personen- und Ortsnamen gestrichen und damit bei der Zählung nicht berücksichtigt wurden. Damit sollte sichergestellt werden, daß nach Möglichkeit nur „typisch altrussische“ Silben erfaßt wurden. Diese Absicht hätte sich nicht verwirklichen lassen, wenn die zahlreich und häufig vorkommenden Namen hebräischen und griechischen Ursprungs nicht ausgeschlossen worden wären; vgl. etwa den Namen *Ioanъ* mit seiner im Altrussischen „unmöglichen“ Aufeinanderfolge von drei Vokalen.

Die Phonemzählung reduzierte sich selbstverständlich nicht auf eine mechanische Buchstabenzählung, sondern setzte eine phonologische Interpretation der Buchstabensequenzen sämtlicher Taktgruppen voraus; vgl. die Akkusativ-Singularform *i*, die phonologisch als *jь* zu werten ist; *istinьnyi* N. Sg. m. = *istinьnъjь*, *ichъ, imъ* = *jichъ, jimъ*; *tьi* N. Sg. m. = *tъjь*; *ijudei* N. Pl. = *ijudeji*; *posъlavъšiimъ* D. Pl. = *posъlavъšijimъ*; *ide* 3. Ps. Sg. Aor. = *jьde*; *srьdьcьmь* I. Sg. = *sьrdьcьmь*.

Nach der Durchführung der geschilderten Vorbereitungsschritte wurden die Silben- und die Phonemzahl sämtlicher Taktgruppen ermittelt. Insgesamt umfaßt unsere Stichprobe 1572 ein-, zwei-, drei-, vier-, fünf- bzw. sechssilbige Taktgruppen. 22 sieben- und 11 achtsilbige

Taktgruppen wurden nicht berücksichtigt, da diese Zahlen nicht als hinlänglich repräsentativ gelten können. Es ergaben sich die in Tabelle 1 angegeben Werte.

Tabelle 1. Silbenlänge in Taktgruppen vor dem Jerwandel

x_i	n_i	y_i
1	46	1.978
2	537	2.075
3	480	2.017
4	311	1.982
5	132	1.967
6	50	1.940

x_i : Taktgruppe des Silbenumfangs i
n_i : Anzahl der Taktgruppen des Silbenumfangs i
y_i : mittlere Anzahl von Phonemen pro Silbe in den Taktgruppen des Silbenumfangs i

Wie vorausgesagt, ist y_2 größer als y_1, d.h., das Altrusssische folgt offensichtlich nicht der einfachen „Menzerathschen" Dynamik.

Im zweiten Untersuchungsschritt wurde die gegen Ende des 15. Jahrhunderts in Novgorod entstandene Gennadius-Bibel herangezogen, die den Zustand nach dem Jerwandel repräsentiert. Und zwar wurden aus dieser Handschrift, die als Facsimile publiziert vorliegt (Biblija 1499 goda), diejenigen Partien aus dem Neuen Testament ausgewählt, die den im OE untersuchten entsprechen. So sollte eine möglichst weitgehende Kommensurabilität der beiden Vergleichsgrößen gewährleistet werden. Die Taktgruppeneinteilung und die philologische Vorbereitung des Textes wurden analog wie beim OE vorgenommen. Insgesamt umfaßt die Stichprobe 1431 ein-, zwei-, drei-, vier- bzw. fünfsilbige Taktgruppen. Unberücksichtigt blieben 10 sechs- und 5 siebensilbige Taktgruppen. Die Auswertung führte zu den Werten von Tabelle 2.

Bereits bei einem ersten, noch nicht mit mathematischen Methoden durchgeführten Vergleich der beiden Tabellen lassen sich einige interessante Feststellungen treffen:

(a) die Anzahl der einsilbigen Taktgruppen ist in der Gennadius-Bibel, d.h. nach dem Jerwandel, weit größer als in der entsprechenden Stichprobe aus dem OE, wie vorhergesagt;
(b) in der Stichprobe aus der Gennadius-Bibel liegen die y_i-Werte jedesmal über den entsprechenden Werten, die wir bei der Auswertung der OE-Stichprobe gewonnen haben;
(c) der Trendverlauf ist fast vollkommen monoton.

Tabelle 2. Silbenlänge in Taktgruppen nach dem Jerwandel

x_i	n_i	y_i
1	225	3.182
2	562	2.244
3	390	2.173
4	187	2.136
5	67	2.149

Im nächsten Schritt wurde für beide Stichproben die Vermutung überprüft, daß sich die empirischen Daten mit Hilfe des Menzerathschen Gesetzes modellieren lassen. Die Resultate

der Anpassung der Daten von Tabelle 1 mit Hilfe von Kurve (2) ist in der dritten, die Anpassung mit Kurve (4) in der vierten Spalte von Tabelle 3 zu sehen.

Tabelle 3. Anpassung der Kurven (2) und (4) an die Daten von Tabelle 1

x_i	y_i	Y_i (2)	Y_i (4)
1	1.978	1.991	1.982
2	2.075	2.036	2.055
3	2.107	2.031	2.032
4	1.982	2.005	2.000
5	1.967	1.969	1.962
6	1.940	1.926	1.931
		K = 2.0679, b = -0.0871, a = 0.0379, D = 0.77	K = 2.5728, b = 0.1358, c = -0.2609, D = 0.92

Da die empirischen Werte nicht monoton fallend sind, läßt sich die einfachste Version des Gesetzes $y = Kx^{-b}$ überhaupt nicht anpassen (D = 0.20). Die „störungsgeladene“ Version (2) ergibt bereits 77% an erklärter Varianz, aber wie man sieht, ist die „freie“ Störung nicht besonders relevant. Dafür aber ist die Bedingung, die mit dem quadratischen Glied in (3) ausgedrückt wird, so relevant, daß man den Parameter a hier gleich Null setzen kann. In der vierten Spalte sind die Werte von $y = Kx^{-b}e^{c/x}$ enthalten, die mit D = 0.92 zeigen, daß das „Menzerathsche Regime“ der Silbenlängen vor dem Jerwandel durch die Strenge der Phonemdistribution modifiziert wird. Die Anpassungen kann man in graphischer Form in Abb. 1 sehen.

Des weiteren wurde die Kurvenanpassung für die Daten aus der Gennadius-Bibel-Stichprobe vorgenommen, zunächst mit Hilfe der Formel

(5) $$y = Kx^{-b},$$

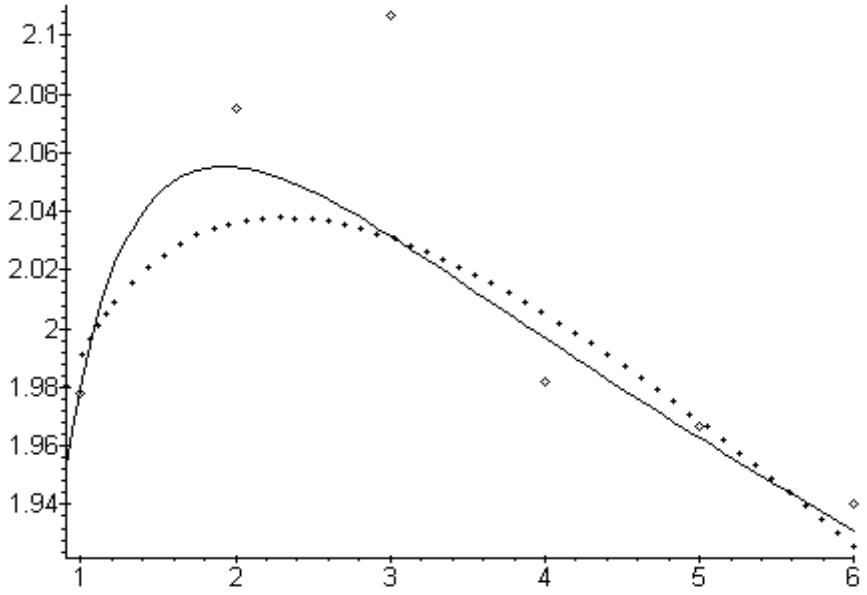

Abb. 1. Anpassung der Kurven (2) (gepunktete Linie) und (4) (volle Linie) an die Daten von Tabelle 1

die ein Spezialfall sowohl von (2) als auch von (4) ist und einen monoton fallenden Verlauf darstellt. Das Resultat ist in der dritten Spalte von Tabelle 4 zu sehen. Wie ersichtlich, haben wir diesmal eine befriedigende Anpassung erhalten. Ein noch besseres Resultat ergibt sich bei der Anpassung mit Hilfe der Formel (2), die in der vierten Spalte zu sehen ist.

Tabelle 4. Anpassung der Kurven (5) und (2) an die Daten von Tabelle 2

x_i	y_i	Y_i (5)	Y_i (2)
1	3.182	3.062	3.159
2	2.244	2.499	2.350
3	2.173	2.218	2.110
4	2.136	2.039	2.046
5	2.149	1.910	2.069
		K = 3.0622, b = 0.2934 D = 0.89	K = 2.6991, b = 0.6536, a = -0.1572, D = 0.97

Das Resultat der Anpassung kann man in graphischer Form in Abb. 2 sehen. Wie ersichtlich, drückt (die gepunktete) Kurve (2) auch die Konvexität am Ende der Daten aus.

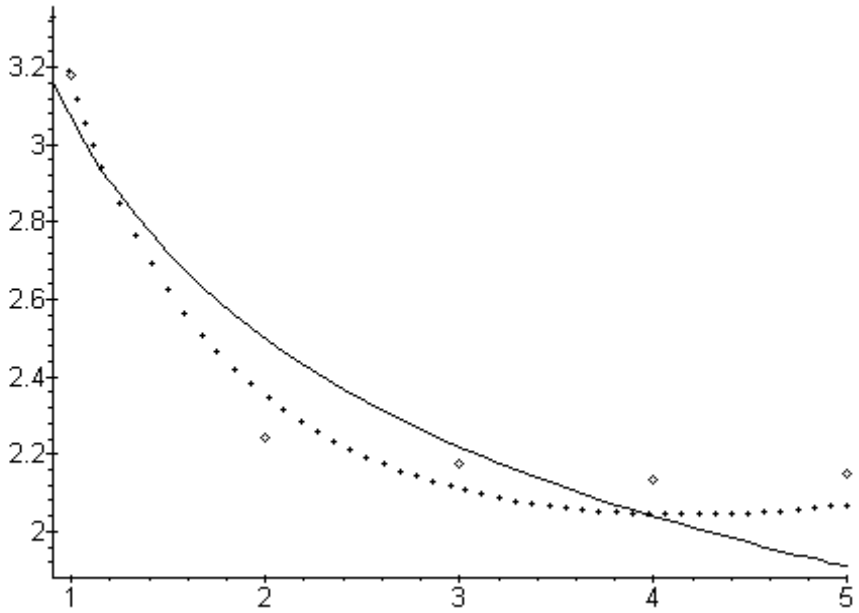

Abb. 2. Anpassung der Kurven (5) (volle Linie) und (2) (gepunktete Linie) an die Daten von Tabelle 2

Um unsere Untersuchung zu verbreitern, d.h. um die Repräsentativität der bei der Analyse der Gennadij-Bibel gewonnenen Ergebnisse zu überprüfen, wurden zusätzlich die Seiten 2v-6, also acht Seiten, aus einer weiteren Handschrift gemäß dem oben erläuterten Verfahren ausgewertet, insgesamt 393 Taktgruppen. Bei dieser Handschrift, die als Facsimile vorliegt (Dianova 1980), handelt es sich um eine aus dem 17. Jhdt. stammende Kopie des *Skazanie o Mamaevom poboišče,* das gegen Ende des 14. Jhdts. entstanden ist. Die Resultate dieser Analyse sowie die Anpassung dieser empirischen Ergebnisse mit Hilfe der Kurven (5) und (2)

sind in Tabelle 5 zu finden (vgl. Abb. 3). Wie ersichtlich, wird durch sie das Bild, das wir durch die Analyse der Sprache der Gennadij-Bibel gewonnen haben, in einer fast ideal zu nennenden Weise bestätigt.

Tabelle 5. Daten aus dem *Skazanie o Mamaevom poboišče*

x_i	y_i	Y_i (5)	Y_i (2)
1	3.163	3.036	3.129
2	2.366	2.591	2.488
3	2.290	2.362	2.258
4	2.259	2.212	2.164
5	2.163	2.102	2.135
6	2.083	2.016	2.147
		K = 3.0355 b = 0.2283 D = 0.89	K = 2.8614 b = 0.4596 a = -0.0894 D = 0.96

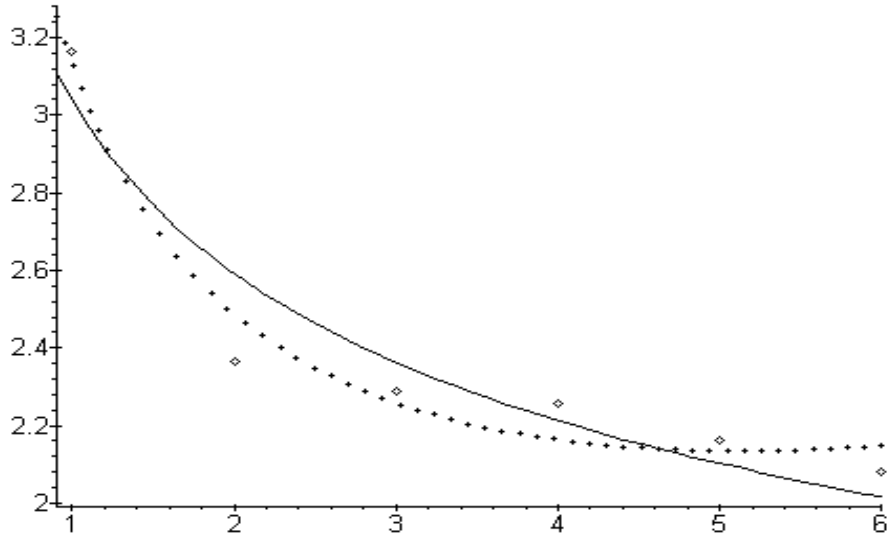

Abb. 3. Anpassung der Kurven (5) (volle Linie) und (2) (gepunktete Linie) an die Daten von Tabelle 5

4. Fazit. Das einfache Menzerathsche Gesetz, ausgedrückt durch (5), ist ein starker Attraktor, der die gesamte Hierarchie der Beziehungen zwischen Konstrukt- und Komponentengröße steuert (vgl. Hřebíček 1997). Die Emergenz dieses Gesetzes hängt damit zusammen, daß die menschliche Sprache nach verschiedenen Ebenen gegliedert ist, und zwischen diesen Ebenen sorgt das Menzerathsche Gesetz für eine bestimmte Art des Gleichgewichts. Es kann im Laufe der Entwicklung geschehen, daß eine Sprachkomponente sich zu einem Extrem entwickelt und in diesem Zustand verweilt. Diesen Zustand betrachten wir als ein lokales Minimum, als einen Attraktor, den die Sprache in ihrer Entwicklung gefunden hat. Da sich aber im lokalen Minimum nicht alle Komponenten gleichmäßig entwickeln können, bedeutet dies eine Stasis,

die irgendwann aufgebrochen werden muß. Der Bruch (Katastrophe) entsteht gerade an der extremen Stelle, im Altrussischen im Bereich der übermäßigen, durch die Jers hervorgerufenen Redundanz auf der Silbenebene, die durch den Jerwandel optimal abgebaut werden kann. Die slavischen Sprachen verlassen schnell Attraktor (4) (Rottmann, pers. Mitteilung) und gehen zu dem Attraktor (2) über, sehr oft sogar zu (5), der ein globales Minimum, einen optimalen Zustand darstellt.

Für das Modellieren der Spracherscheinungen ist dieser Fall sehr lehrreich, denn er zeigt, daß

(a) Daten nicht einfach gegeben sind, sondern von uns konstruiert werden, auch wenn nur eine embryonale Theorie existiert;

(b) Gesetze, d.h. allgemeine Aussagen über Mechanismen, nur gelten, wenn entsprechende Bedingungen erfüllt sind;

(c) bei „anomalen" Bedingungen nicht immer eine Parametervariation im herrschenden Modell ausreicht, sondern die Hintergrundannahmen selbst modifiziert werden müssen;

(d) es Attraktoren in der Sprache gibt, die eine starke Anziehungskraft ausüben und durch die jede Sprache einmal gehen muß. Extreme Zustände können lange Zeit dauern, aber letztlich kreuzt die Sprache in ihrer Entwicklung immer wieder den zentralen Attraktor, den in unserem Falle das Menzerathsche Gesetz darstellt;

(e) wenn eine Spracherscheinung nicht einem bewährten Gesetz folgt, dieses keinesfall eine chaotische Gesetzeslosigkeit bedeutet, sondern uns zu verschiedenen Maßnahmen zwingt (vgl. Bunge 1967), in jedem Fall zu einer Verbesserung unserer Theorie.

(f) je mehr Parameter eine Kurve hat, es um so leichter ist, eine gute Anpassung zu erhalten. Man soll dieser Verlockung widerstehen, denn wenn es nur darum ginge, wären Polynome am besten geeignet. Sie sind aber um so schwieriger in eine Theorie einzubetten, je höher ihr Grad ist. Im allgemeinen soll man bei dem Aufbau einer Theorie auf Polynome verzichten, es sei denn, sie entstehen aus begründeten Annahmen.

Резюме
Падение редуцированных в древнерусском языке

Авторы ставят перед собой цель разработать новое объяснение причин, приведших к падению редуцированных гласных ь и ъ в слабой позиции в древнерусском языке. В основу предлагаемого ими подхода положен закон П. Менцерата, согласно которому с возрастанием числа слогов в словоформах снижается среднее число фонем в слогах. Авторы исходят из предположения о том, что до падения редуцированных закон Менцерата не мог действовать, потому что не было необходимых для этого условый, а именно, в силу строго регламентированной сочетаемости согласных в рамках слогов. Это предположение подтверждется в результате исследования начала Остромирова Евангелия. Далее авторы предполагают, что после падения редуцированных закон Менцерата вступил в действие, потому что теперь указанная строгая регламентированность сочетаемости согласных в слогах не действовала. И это преположение подтверждается в результате анализа начала Геннадиевской Библии и начала Сказания о Мамаевом побоище. Решение исходной проблемы выглядит следующим образом. Закон Менцерата является сильным „аттрактором", который всегда ищет пути к „победе". „Ломка" происходит обычно в экстремальной точке. В древнерусском такой точкой являлась чрезмерная избыточность структуры слогов, вызванная как раз наличием редуцированных. Она наилучшим образом была устранена падением редуцированных в слабой позиции, ведь именно этот процесс привел к появлению тех условий, которые необходимы для действия закона Менцерата. Проведенные тесты подтвердили правомерность этой гипотезы.

Literatur

Abernathy, R. (1956). The fall of the jers: a statistical interpretation. *For Roman Jakobson: 13-18.* The Hague: Mouton.

Abernathy, R. (1963). Some theories of Slavic linguistic evolution. In: *American Contributions to the Fifth International Congress of Slavists, Vol. I: Linguistic Contributions: 8-26*. The Hague: Mouton.

Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika 2, 1-10.*

Altmann, G., Lehfeldt, W. (1980). *Einführung in die quantitative Phonologie.* Bochum: Brockmeyer.

Altmann, G., Schwibbe, M. (1988). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim: Olms.

Bak, P. (1999). *How Nature Works. The Science of Self-organized Criticality*. New York: Copernicus–Springer.

Biblija 1499 goda i Biblija v sinodal´nom perevode. S illjustracijami. V desjati tomach. Tom 7: Moskva 1992.

Bräuer, H. (1961). *Slavische Sprachwissenschaft. I. Einleitung, Lautlehre*. Berlin: de Gruyter (Sammlung Göschen Band 1191/1191a).

Bunge, M. (1967). *Scientific Research I, II*. Berlin: Springer.

Bunge, M. (1977). *Treatise on Basic Philosophy, Vol. 3*. Dordrecht: Reidel.

Dianova, T.V. (1980). *Skazanie o Mamaevom poboišče*. Licevaja rukopis´ XVII veka iz sobranija Gosudarstvennogo Istoričeskogo muzeja. Al´bom. Moskva: Sovestskaja Rossija.

Gasparov, B.M., Sigalov, P.S. (1974). *Sravniteľnaja grammatika slavjanskich jazykov, Tom I.* Tartu: Tartuskij gosudarstvennyj universitet.

Geršić, S., Altmann, G. (1988). Ein Modell für die Variabilität der Vokaldauer. *Glottometrika 9, 49-58.*

Hřebíček, L. (1997). *Lectures on Text Theory*. Prague: Oriental Institute.

Kolesov, V.V. (1964). Padenie reducirovannych v statističeskoj interpretacii. *Voprosy jazykoznanija, vyp. 1964/2, 30-44.*

Lunt, H.G. (1974). *Old Church Slavonic Grammar. Sixth edition*. The Hague, Paris: Mouton.

Markov, V.M. (1964). *K istorii reducirovannych glasnych v russkom jazyke.* Kazan´.

Menzerath, P. (1954). *Die Architektonik des deutschen Wortschatzes.* Bonn: Dümmler.

Naranan, S., Balasubrahmanyan, V.K. (1998). Models of power law relations in linguistics and information science. *Journal of Quantitative Linguistics 5, 35-61*.

Rusinov, N.D. (1979). Istorija drevneslavjanskich reducirovannych glasnych v kibernetičeskoj interpretacii. *Ėvoljucija i predystorija russkogo jazykovogo stroja: 3-24*. Gor´kij.

Sidorov, V.N. (1966). Reducirovannye glasnye ъ i ь v drevnerusskom jazyke XI v. In: Sidorov, V.N., *Iz istorii zvukov russkogo jazyka: 5-37*. Moskva: Nauka.

Vostokov, A. (izd.) (1843). *Ostromirovo evangelije 1056-57 goda*…, izdannoe A. Vostokovym. Sankt-Peterburg (Nachdruck Wiesbaden: Harrasowitz 1964).

Wimmer, G., Köhler, R., Altmann, G. (erscheint 2003). Unified derivation of some linguistic laws.

Zaliznjak, A.A. (1985). *Ot praslavjanskoj akcentuacii k russkoj*. Moskva: Nauka.

Zaliznjak, A.A. (1993). Padenie reducirovannych po dannym berestjanych gramot. In: Janin, V.L., Zaliznjak A.A., *Novogorodskie gramoty na bereste (iz raskopok 1984-1989 gg.): 241-270*. Moskva: Nauka.

Zaliznjak, A.A., Janin, V.L. (2001). Novgorodskij kodeks pervoj četverti XI v. – drevnejšaja kniga Rusi. *Voprosy jazykoznanija, vyp. 2001/5, 3-25.*

Glottometrics 2, 2002, 45-52

Freedom of choice and psychological interpretation of word frequencies in texts

Simone Andersen, Hamburg[1]

Abstract. To what extent are word frequencies in texts manipulable by text producers? The expected word frequency is regarded as a conditioned probability. This allows showing that observed frequencies are only partly frequencies of choice. When comparing frequencies, confoundation with frequency of conditions, here called occasions, has to be avoided. The amount of free choice is discussed.

Key words: Word frequency

1.

There are many investigations in Quantitative Linguistics requiring that the observed frequencies of linguistic units or classes of units are compared with the corresponding expected frequencies. For some of those hypotheses it is correct to expect under hypothesis H_0 the state of equal probabilities; for another kind of studies related to the individual text or text producers' specific properties it is reasonable to take under H_0 the differing observed frequencies of units as they can be counted in language as a whole: e.g. the different word frequencies as noted down in frequency lexica.

An important group of questions, however, makes it necessary to define expected frequencies more precisely.

Fundamental laws in Quantitative Linguistics (Zipf 1949; Altmann 1988a,b; Wimmer, Altmann 1996; Orlov 1982a,b) describe proportionality phenomena related to frequencies of units or of classes made up by features of them, e.g. by their lengths. For example, word lengths - as probably the lengths of any linguistic units - in a text always occur in frequencies that follow a special theoretical distribution (cf. Altmann 1988a,b; Wimmer et al. 1994; Wimmer, Altmann 1996, Best 1997, 2001). Another example is the closedness of texts (Orlov 1982a,b): According to Orlov, there is only one size for any text, where the rank frequency distribution follows the Zipf-Mandelbrot-law, indicating that this is the "true" size of the text.

In search for causal hypotheses for these phenomena one has to ask what frequencies occurring in texts mean and by what they are determined. Many reasons are conceivable, from the supposition of pure randomness to the idea that text producers compose their texts according to an "ideal" distribution etc. Are those distributions of frequencies due to influences working during text production, as for example the text producers' conscious or unconscious decisions - or to other forces?

The frequency of a unit occurring in a text - should it be a single text, a group of texts or

[1] Address correspondence to: Simone Andersen, Loogestieg 19, D-20249 Hamburg.
E-mail: AndersenSC@aol.com

the total of all texts in language as a whole - can be regarded as the result of some choice: the text producer chose it (or did not) when producing the text.

An important question, however, is the degree of freedom in this choice: how much freedom of variation is there in the actual use of a linguistic unit? To what extent are frequencies in texts manipulable by the text producer? How large is the amount of frequency which is due to voluntary use, the amount of real "frequency of choice"?

Here we will concentrate on the frequency of words, because using a word in a text is intuitively supposed to have an especially high degree of free choice, compared to the use of other linguistic units.

2.

We suppose the following:
There are 20 color cards with shades of the spectrum, ranging from red over yellow and green to blue. There are also five color names available: "red", "orange", "yellow", "green" and "blue".

An individual gets the task to draw by random one color card and choose a name for it out of the five color names available. (The individual has to draw the card by random and has to be prevented from looking at all cards at one time which could induce drawing lines in advance, categorizing the shades into five groups etc.)

Now we ask:

What is the probability that the word "blue" is chosen, after a card has been drawn? It is neither p = 1/5 nor p = 1/20 nor p = 1/5 x 1/20.

Instead of that, you can think of two kinds of probability distributions: for each color name there is a specific probability distribution over the color spectrum that answers the question: What is the shade you are thinking of after hearing this name? And reverse, there is also a specific conditional probability distribution over the possible names for every color shade that answers the question: What is the best fitting name for this shade? This second conditional probability distribution is of interest here.

x is the special occasion - the color card - and $p(y|x)$ denotes the conditional probability that the name "y" is chosen, if x happens. Or: p("blue" is chosen$|x$).

The situation (fictitious example) is depicted in Fig 1.a. On the x-axis you have the color shades - here only 6 shades are shown instead of 20 -, on the y-axis you have the probability (or better: likelihood) p. For each x you get the 5 color names with their $p(y|x)$. (For better visualization, the points are connected by lines. In the case of color shades, however, one can think of a continuous variable x, so five curves would yield.)

Some - probably very few - cards surely will be called "blue", if they are drawn ($p(y_b|x_r) = 1$), in other words, "blue" will be the one and only way to denote this shade. Others surely will not be called "blue" ($p(y_b|x_s) = 0$). Some cards maybe will be called "blue", with $p(y_b|x_t) = 0.9$, $p(y_b|x_u) = 0.7$ or $p(y_b|x_v) = 0.5$ etc.

For every x: $\sum_y p(y|x) = 1$

This is illustrated in Fig.1.b.

To find the probability that the word "blue" will be spoken or chosen one has to multiply for every color card x_i the corresponding value of the conditional probability related to "blue" - the value $p(y =$ "blue"$|x_i)$ - by the probability of its occurrence $p(x_i)$ (in this case: 1/20 for all cards) and sum them up.

The probability for saying "blue" is (for example)

p("blue") = (1/20·0) + (1/20·0) + ... (1/20·0.2) + (1/20·0.5) + (1/20·0.98) + (1/20·1) + (1/20·1) = (1/20)·$\sum_x$ p("blue"|color shade x).

In general:

$p(y) = \sum_x p(x)\cdot p(y|x) = \sum_x p(x,y)$

So the probability of occurrence depends on two other probabilities:

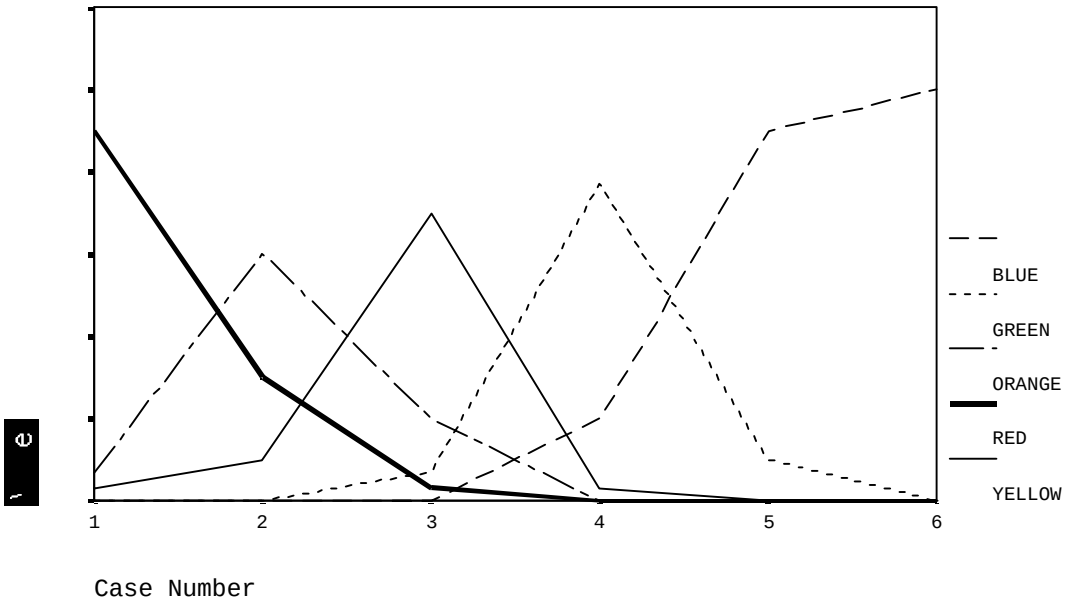

Fig. 1.a. Fictitious example: Six color shades (cases) x with their likelihoods (values) $p(y|x)$ for five different words (color names).

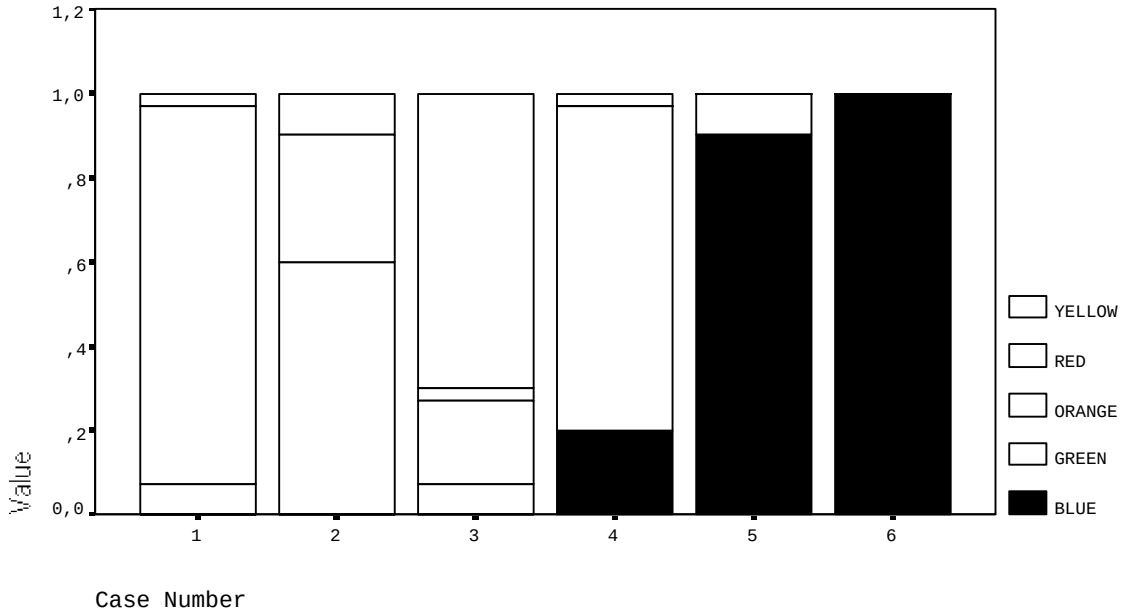

Fig.1.b. Sums of conditional probabilities, showing that for any shade (case) x the values of $p(y|x)$ add to 1 in different partitioning.

1. The probability $p(x)$ that the special color shade is drawn (in this case: 1/20 for all shades), the "probability of occasion".
2. The conditional probability that this x will be called y: $p(y|x)$, the "probability of name application".

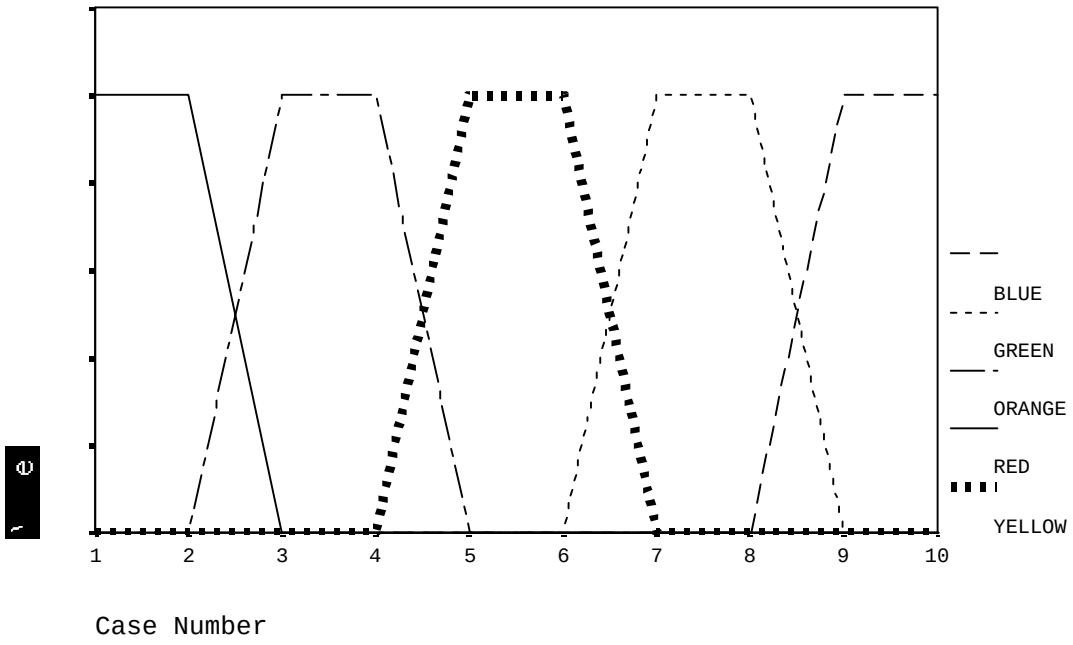

Fig.2.a. Special case: equal partitioning of the spectrum and values of conditional probabilities always being $p(y|x) = 1$

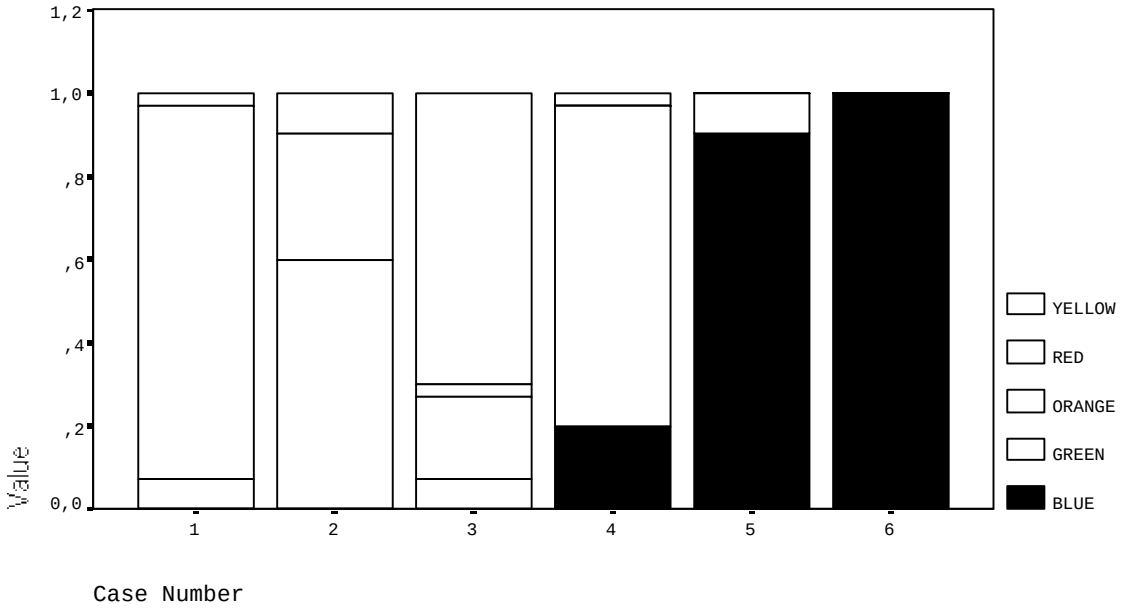

Fig.2.b. Standardized sums of likelihoods for the special case in 2.a., to compare with Fig.1.b.

If you generalize the example, in most cases the $p(x)$ are not equal and the spectrum in question is not partitioned in an "ideal way" by the different y_i.

In cases where all x would have equal probabilities, and the y_i would divide the possible occasions into parts of equal number with the conditional probabilities always being $p(y|x) = 1$ the result would be the special case (Fig.2.a,b) that $p(y) = \text{const}$ (in the above example 1/5).

This case, in spite of its unrealistic assumption, is implicitly supposed in many interpretations of word frequencies. Neglecting this when comparing frequencies means taking the risk of confoundation.

The idea that meaning is not dichotomous - i.e. the probability for choosing a word is not either 1 or 0, because the limits of extension are not sharply determined, so that word curves may overlap - has been already shown and highly developed in fuzzy set theory (Zadeh 1965, 1982; Zimmermann 1991; Klir, Folger 1988) and especially its application to linguistics (see the fuzzy semantics approach by Rieger 1989, 1999). Some of the graphs in the present paper may resemble to “fuzzy partitioning”, but there is a difference in the focus of attention. While the fuzzy approach focuses on the fact that there are no rectangles beneath the curves, we look at the differing amounts of area, given the different quantities and the different frequencies of occasions. The fuzzy approach in linguistics or cognitive psychology looks at semantics: the kind of assigning relation between words and objects (referents). It can be applied to lexicon investigation. In the present paper, however, we are interested in the frequencies occurring in language use. This is not only a problem of semantics. The explanations here are without sense for the lexicon, more than that, they show the difference between lexicon and texts – related to the interpretation of frequencies. Semantics which indeed has to be modeled fuzzy-wise is only one of the constraints modulating a priori probabilities. The difference in perspective becomes visible in the fact that we get two kinds of conditional probability distributions here – as explained above –, while there is only one graph for each example in the fuzzy approach.

The example of color naming has been chosen here because color shades show ideal properties in the role of "occasions". The variation of objects can be taken out intersubjectively. Their identity vs. difference as referents is clearly - physically - definable, in terms of wave lengths. For any color term the set of alternatives can be limited and well-defined. The application of one of the basic color terms excludes the application of the others. (As color coding is one of the easiest possible ways of modelling reference it is also one of the favourite paradigms in the field of concept formation.) But theoretically you can generalize the example such that it represents any communication situation. The occasions represent the "forces of reality", those aspects of reality which have to be talked about. One can consider the occasions x to be requirements or what Köhler (1990) called "Kodierungsbedürfnis". When a text is produced, some properties or forces of the context (grammatical habits etc.) can also be considered occasions x. Anyway the frequency of x is given and cannot be manipulated by the speaker.

The probability of occurrence for y is the sum of all products of the probabilities of the occasions x and their corresponding probabilities of name application (with $p(y_i|x) = 0$ for many x, of course).

Likewise, the frequency of occurrence depends on two other frequencies:

1. The frequency - for example - at which a special color shade is drawn, i.e. the "frequency of occasion".

2. The frequency at which this x has been called y, the "frequency of name application".

Thus, the frequency of a word has to be read as a function of the frequencies of occasions and their corresponding conditional probabilities (propensities) of application of this word. The frequencies of words in real texts have to be interpreted as a certain amount of inform-

ation about the frequencies of occasions "in the world". So word frequencies listed in frequency lexica also reveal evidence of frequencies found in the reality language users live in.

3.

How much freedom of choice is left?

Considering $p(y|x)$ the "probability of name application" is still not yet the probability of choice. It is only the probability that a certain subject will perceive - or identify - occasion x as a case of "y_w " and not "y_v", "y_u", "y_s", "y_t",....

In our example, the y_i exclude one another: if a particular individual has made the decision that the color shade is a case of "yellow", then at the same time he excludes "red" or "green". For example, if a text producer decides that he has to speak of a case of "influenza", he does not speak of "heart attack". This force could be named the force of individual perception of application.

So the second kind of information we find expressed by frequency is the frequency of perception or the frequency at which the text producer decided that this x was a case of y_w

If we give him for each y_w a reservoir of expressions e_i, he then can make his choice: He has to choose an expression e for y_w for which we could consider the chance to be $p(e_i|(y_w|x))$.

The equivalent - or even synonymous - alternative expressions do not necessarily have to consist of the same number of words, and theoretically the set of possible alternatives is infinite. But for a given individual at a given moment t under psychologically realistic conditions (for example, memory capacities) the number of alternatives will be very finite, if not small.

It would be interesting to investigate the number of equivalent ways to express something that subjects can produce within certain time limits.

If we look at word frequencies in language, we can theoretically determine a mean and a variance. We claim that, if we partition this variance, only a smaller amount of the variance is explained by a free choice of words or intentional expression.

4.

Now we turn to an example related to classes of words and mentioned at the beginning: the question of word lengths and the choice of words.

Word lengths in texts always occur in frequencies that follow a theoretical distribution belonging to a special distribution family arising from a simple proportionality recurrence relation (cf. Wimmer et al. 1994; Wimmer, Altmann 1996, Best 1997, 2001). How much freedom is left to text producers to compose their texts according to the relevant distribution?

If we look at the most frequent words (one-syllable words like "a", "the", "and", "or", "if"), probably few alternative expressions will exist.

But for the majority of words we can postulate the following: For every occasion an individual has a set of possible expressions available. Those expressions, however, will have different probabilities. So we could think of a probability hierarchy of expressions e_i with their corresponding $p(e_i|(y_w|x))$.

We suppose that the shorter words (expressions) will be those with the higher probability. What could induce a text producer to choose a longer word (or an expression with longer words) in spite of its lower probability?

Here we remember Hull's theory of drives (1943, 1952) and his construct of a hierarchy of reactions (habits). We consider the occasion to be the stimulus which is followed by

reactions of different probability, which are the different expressions. According to Hull the specific reaction depends on the degree of arousal: under high arousal the reaction with the highest probability will be performed.

Is it possible to identify factors in text production that correspond to the construct of arousal? Further investigation on this question has to be done.

Altmann proposes (2001; personal notification) that the arousals could follow a classifying principle already holding in psychophysiology just according to the proportionality law. This would mean, in other words, a brain's preference for a special form of categorizing: Our ability for building classes could work in a way that we concentrate on those features that occur according to the preferred proportionality distribution: a hypothesis to be tested in cognitive neuropsychology.

Anyway, the amount of free choice related to frequencies now seems to be smaller than supposed. So another proposition would be that - given the small amount of free choice - the validity of the proportionality law supplies evidence that frequencies are to a great extent not caused by influences working during text production. Other than in musical composition, in the case of verbal text production the decisive processes determining unit or unit feature frequencies are to a great extent out of reach for the individual.

References

Altmann, G. (1988a). Verteilungen der Satzlängen. In K.-P. Schulz (ed), *Glottometrika 9, 147-169.* Bochum: Brockmeyer.

Altmann, G. (1988b). *Wiederholungen in Texten.* Bochum: Brockmeyer.

Altmann, G. (1991). Modelling diversification phenomena in language. In U. Rothe (ed), *Diversification processes in language: Grammar: 33-46,* Hagen: Margit Rottmann Medienverlag.

Best, K.H. (ed.) (1997). *The distribution of word and sentence length.* (= *Glottometrika 16*). Trier: WVT.

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

Hull, C.L. (1943). *Principles of behavior*. New York: Appleton Century Crofts.

Hull, C.L. (1952). *A behavior system.* New Haven: Yale University Press.

Klir, G., Folger, T.A. (1988). *Fuzzy sets, uncertainty and information.* Englewood Cliffs: Prentice Hall.

Köhler, R. (1990). Elemente der synergetischen Linguistik. In R. Hammerl (ed), *Glottometrika 12, 179-187*. Bochum: Brockmeyer.

Orlov, Ju.K. (1982a). Dynamik der Häufigkeitsstrukturen. In H. Guiter & M.V. Arapov (eds), *Studies on Zipf's Law: 116-153*. Bochum: Brockmeyer.

Orlov, Ju.K. (1982b). Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? In Ju.K. Orlov, M.G. Boroda & I.Š. Nadarejšvili (eds), *Sprache, Text, Kunst. Quantitative Analysen: 1-55*. Bochum: Brockmeyer.

Rieger, B. (1989). *Unscharfe Semantik: die empirische Analyse, quantitative Beschreibung, formale Repräsentation und prozedurale Modellierung vager Wortbedeutungen in Texten*. Frankfurt am Main – Bern - New York – Paris: Lang.

Rieger, B. (1999). Computing Fuzzy semantic granules from natural language texts. A computational semiotics approach to understanding word meanings. In: M.H. Hamza (ed),

Artificial Intelligence and soft computing, Proceedings of the IASTED International Conference: 475-479. Anaheim; Calgary; Zürich: IASTED Acta Press.

Wimmer, G., Köhler, R., Grotjahn, R., Altmann, G. (1994). Towards a theory of word length distributions. *J. of Quantitative Linguistics 1, 98-106.*

Wimmer, G., Altmann, G. (1996). The theory of word length: some results and generalizations. *Glottometrika 15, 112-133*.

Zadeh, L. (1965). Fuzzy sets. *Information and Control 8, 338-353*.

Zadeh, L. (1982). A note on prototype theory and fuzzy sets. *Cognition, 12, 291-297*.

Zimmermann, H.-J. (1991). *Fuzzy set theory and its applications.* Dordrecht: Kluwer.

Zipf, G.K. (1949). *Human behavior and the principle of least effort.* Cambridge, Massachussetts: Addison-Wesley Press.

Glottometrics 2, 2002, 53-81

Subjektive Bewertung von Vorkommenshäufigkeiten: Methode und Ergebnisse

Marion Krause, Bochum[1]

Abstract. Word frequency effects play an important role in speech production and perception as well as in language change processes. In most cases, frequency data are taken from text corpora. After discussing some results of the traditional procedure, the paper presents an alternative method of gathering subjective frequency data. This method was developed by R. Frumkina and her colleagues in the 1960s - 1970s. It is directed to the so-called "frequency index" that is assumed to treat frequency data from input and output and store it with the lexical representations in the individual's internal lexicon. The paper describes the experimental procedure and the underlying statistical methods and illustrates them on the basis of two experiments conducted with German words. Their outcome - two word frequency lists comprising 642 and 144 words resp. - is listed according to frequency as well as in alphabetic order.

Key words: Psycholinguistics, word frequency, experimental methods

1. Vorkommenshäufigkeit als Komponente des mentalen Lexikons

Die Struktur des mentalen Lexikons, der Aufbau von Einträgen und ihre Verknüpfung miteinander bilden seit Jahren einen der interessantesten Forschungsgegenstände jener Wissenschaftsdisziplinen, die sich mit dem menschlichen Sprachvermögen einerseits und seiner Modellierung für technologische Anwendungen andererseits befassen. Seitens der Linguistik liegt der Schwerpunkt auf der Untersuchung semantischer Repräsentationen und ihrer prozessualen Leistungen in der Sprachproduktion und -rezeption. In jüngster Zeit konzentriert sich die Forschung u.a. auf die Potenzen der Lexikon-Syntax-Schnittstellen, um die Frage, wie viel Syntax im Lexikon angelegt ist (z.B. Schönefeld 2001). Im Hintergrund setzt sich dabei immer die wissenschaftliche Kontroverse zwischen modularen und konnektionistischen Modellen der Sprachverarbeitung fort.

In die Debatte über die Struktur des mentalen Lexikons soll hier nicht eingegriffen werden. Allerdings wird mit dem vorliegenden Aufsatz ein Aspekt der Repräsentation von Einträgen hervorgehoben, der in den aktuellen, semantisch orientierten Modellierungen des mentalen Lexikons zu wenig Beachtung findet. Es handelt sich um den *quantitativen* Aspekt der Vorkommenshäufigkeit. Er spielt bei der Organisation des Lexikons offenbar eine wichtige Rolle. Diese Annahme stützt sich auf eine Vielzahl empirischer Untersuchungen zur Sprachverarbeitung. Sie belegen, dass die Vorkommenshäufigkeit von Wörtern und Strukturen die Schnelligkeit und die Art des Zugriffs auf sprachliche Information beeinflusst: sowohl lexikalische Entscheidungsaufgaben als auch Experimente zur Spracherkennung (u.a. Savin 1963; Frumkina, Vasilevič, Gerganov 1971; Bock 1978; Štern 1982; Marslen-Wilson 1989; Krause 1989, 1992; Čugaeva 1989; Štern 1992) verifizierten den Frequenzeffekt.

[1] Address correspondence to: Marion Krause, Seminar für Slawistik, Universität GB3/8, Universitätstr. 150, D-44780 Bochum. E-mail: Marion.Krause@ruhr-uni-bochum.de

Auch bei der Erklärung von Erscheinungen des Sprachwandels bieten Frequenzdaten einen wichtigen Anhaltspunkt. Gemeinsam mit anderen Faktoren erschließen sie sowohl Beharrungs- als auch Veränderungsphänomene. Dennoch muss Ruoff in diesem Zusammenhang feststellen: „Die seit über hundert Jahren kräftig zum Durchbruch drängende Erkenntnis, dass zu jeder systematischen Behandlung von Sprachformen, dass zu jeder Erkenntnis des sprachlichen Systems notwendig die Kenntnis von Gebrauchshäufigkeiten gehört, hat nur selten zu deren Untersuchung geführt“ (Ruoff 1990, 8).

Das quantitative Merkmal lässt sich ganz allgemein als „Häufigkeitsindex“, der einem Lexikoneintrag zugeordnet ist, charakterisieren (Frumkina, Vasilevič 1971, 7-8). Es ergänzt die qualitative (semantische) Beschreibungsebene und ist mit allen gängigen, auf der Netzwerkmetapher operierenden Modellen des internen Lexikons vereinbar: als hoch gelegener, dicker Knoten in einem hierarchischen Netzwerk oder aber als „ausgetretener Pfad“ (Aitchison 1997, 296) in einem mindestens zweidimensionalen Modell. Als Index lässt es sich auch an taxonomische Modelle knüpfen. Manche Forscher nehmen sogar separate Speicher für frequente und weniger frequente Einträge an (vgl. Aitchison 1997, 277).

2. Zur Typologie von Frequenzbewertungen und -wörterbüchern

Wie ist nun die Vorkommenshäufigkeit zu erfassen? Es gibt grundsätzlich zwei Wege der Frequenzdatenerhebung: die Korpusauszählung und das psycholinguistische Experiment.

2.1. Texte

Die am weitesten verbreitete Methode beruht auf der **Auszählung von Texten**. Noch vor 10-15 Jahren erforderte dieses Verfahren einen enormen Arbeitsaufwand. Ein beredtes Zeugnis dafür liefert das bereits 1898 von Kaeding veröffentlichte „Häufigkeitswörterbuch der deutschen Sprache“ (Kaeding 1898). Für dieses Wörterbuch wurden von über 100 Arbeitsgruppen Texte im Umfang von insgesamt 10,9 Millionen Wörtern (Token[2]) ausgezäht.

Die meisten Frequenzwörterbücher stehen auf einer bescheideneren Materialbasis. Für das Frequenzwörterbuch von E. Štejnfeldt (1963), das als ein Standardnachschlagewerk des Russischen gilt, wurden Texte im Gesamtumfang von ca. 400000 Wörtern (wahrscheinlich Token) ausgezählt. Das Wörterbuch enthält die 2500 häufigsten Wörter (Lemmata[3]) des Russischen.

Heute erlaubt die Arbeit mit großen Textkorpora eine weitaus umfangreichere und damit auch repräsentativere Datenerfassung. Das von Lönngren (1993) publizierte Frequenzwörterbuch des Russischen basiert auf ca. 1 Million Token aus dem Uppsala-Korpus. Wie die meisten Datenkorpora beruht das Uppsala-Korpus auf klaren Auswahlkriterien (vgl. Lönngren 1993, 10f.). Es dominieren schriftsprachliche Texte aus der Belletristik und den Printmedien; die Umgangssprache bleibt bewusst ausgeblendet. Damit wird ein wesentlicher Nachteil der meisten zur Zeit existierenden Frequenzwörterbücher deutlich: der alltägliche Sprachgebrauch findet nur ungenügend Niederschlag.

In der Einführung zum „Häufigkeitswörterbuch der gesprochenen deutschen Sprache“ (Ruoff 1981[1],1990[2]) wird auf diesen Mangel hingewiesen und die Notwendigkeit der Arbeit mit gesprochener Sprache unterstrichen. Das Häufigkeitswörterbuch entstand im Rahmen eines breit angelegten Projektes zur Erforschung der „Sprache in Südwestdeutschland“. Seine

[2] Der Begriff *Token* steht für das individuelle Vorkommen eines Wortes – genauer: einer Wortform. Wiederholungen ein und desselben Lexems, ja ein und derselben Wortform werden somit einzeln gezählt.

[3] Der Begriff *Lemma* stammt aus der Lexikologie. Gemeint ist, dass im Wörterbuch eine festgelegte Grundform alle Formen des Wortes (Lexems) repräsentiert; für Substantive wählt man üblicherweise den Nominativ Singular, für Verben den Infinitiv.

Materialbasis bilden Aufnahmen der ländlichen Bevölkerung Baden-Württembergs. Diese regionale und soziale Einschränkung wird ausführlich begründet und ist ohne Weiteres nachzuvollziehen. Leider findet sie im Titel keine Erwähnung, obwohl der Herausgeber selbst einräumt, dass die „Korpusbedingtheit" eines Frequenzwörterbuchs gesprochener Sprache größer sei als diejenige eines auf schriftsprachlichem Material beruhenden Wörterbuchs (Ruoff 1990, 14). Als sensitive Faktoren zeichnen sich die landschaftliche und soziale Differenzierung der Sprache wie auch das Vorhandensein von Gattungsunterschieden zwischen geschriebener und gesprochener Sprache ab (vgl. die Gegenüberstellung von „literacy" und „orality", u.a. Michaels, Collins 1984; Sappok 1998).

Im „Häufigkeitswörterbuch der gesprochenen deutschen Sprache" erfolgt die Auflistung der Vorkommenshäufigkeiten innerhalb der Wortarten. Es werden auch rückläufige Listen angeboten. Bestimmte Anwendungen sind durch diese Anordnung der Auszählungsergebnisse jedoch erschwert. Es fehlt eine „absolute" Liste, die allein auf Frequenzdaten beruht und von der Wortartzugehörigkeit der Lemmata abstrahiert.

Überregional angelegte Datenkorpora (verschriftlichter) gesprochener Sprache (z.B. das Mannheimer Korpus der gesprochenen Sprache) können die Basis für aktualisierte Frequenzdaten bilden und ermöglichen zudem einen Vergleich mit Ruoffs regional verankerten Vorkommenshäufigkeiten. Modernes Informationsretrieval und große Speicheraufkommen, die nur noch geringen technischen Aufwand erfordern, erleichtern das Erstellen von Frequenzdaten aus Korpora. Dennoch bleibt ein grundsätzliches Problem bestehen: kein Korpus ist allumfassend! Es werden immer Restriktionen in Bezug auf die Textauswahl, die zugrundegelegten Textsorten, Kommunikationssituationen und Interaktionsarten getroffen werden müssen, um dem Material eine Systematik zu erhalten. Die Arbeit mit Korpora birgt eigene Risiken, insbesondere dann, wenn ein Korpus nicht allzu groß ist. In jedem Fall hängt es von der Aufgabenstellung ab, ob man sich auf ein Korpus und auf Korpusanalysen beschränken kann.[4]

2.2. Bewertung

Das zweite Verfahren zur Bestimmung von Vorkommenshäufigkeiten ist experimenteller Natur. Es besteht im – durch eine Aufgabenstellung vermittelten – Zugriff auf das interne Lexikon von SprecherInnen/HörerInnen einer Sprache. Das Verfahren beruht auf **Häufigkeitsbewertungen**, die Probanden in Bezug auf bestimmte Stimuli vornehmen, und erfordert die statistische Aufbereitung der Daten. Deshalb sprechen die Autoren der Methode auch von „subjektiver Vorkommenshäufigkeit (F_{sub})" und unterscheiden sie von der an Korpora ermittelten sogenannten „objektiven Vorkommenshäufigkeit (F_{ob})".

Das Verfahren wurde in den 1960-er Jahren von einer Moskauer Arbeitsgruppe um R.M. Frumkina auf der Basis der psychometrischen Arbeiten von Guilford (1954) entwickelt. Im Mittelpunkt des Projektes standen wahrscheinlichkeitsprognostische Aspekte der Sprachverarbeitung. Die Vorkommenshäufigkeit wurde als zentrale Determinante der Wahrscheinlichkeitsprognose herausgearbeitet und detailliert untersucht.

---

[4] Zur Illustration dieser These: In einer sehr ausführlichen Arbeit zu den kognitiven Grundlagen der epistemischen Modalität wurde die epistemische und performative (in einem weiten Sinne) Bedeutung der Verben *denken* und *glauben* im Deutschen untersucht (Nuyts 2000). Die schwache epistemische Belegung des Verbs *denken* im untersuchten Korpus (2 Fälle in einem Korpus von 60000 Wörtern) führte den Autor zur Annahme, das deutsche Verb *denken* hätte – im Unterschied zum englischen *think* und zum niederländischen *denken* – wohl kaum eine qualifikationale (epistemische) Lesart entwickelt (ibd., 120). Auf der Grundlage dieser – quantitativ korpusgestützten – Beobachtung formuliert der Verfasser weitreichende theoretische Annahmen über systemimmanente semantische Entwicklungen. Deutsche MuttersprachlerInnen sind über das Fazit eher verwundert, denn Sätze wie *Ich denke, er kommt heute später. – Ich dachte, du kommst heute später.* sind ihnen sehr vertraut, und zwar in epistemischer, die Wahrscheinlichkeit eines Sachverhalts spezifizierender Lesart.

Derartige wahrscheinlichkeitsprognostische („prediction-derived“) Ansätze spielen in komplexen, multiplen Modellen der Sprachverarbeitung eine entscheidende Rolle (vgl. Aitchison 1997, 282-289). Hervorzuheben ist jedoch, dass wahrscheinlichkeitsprognostische Annahmen – im Sinne von Frumkina und ihren NachfolgerInnen – auf allen Ebenen der Sprachverarbeitung getroffen werden; sie sind nicht per se mit höheren, als Top-Down-Stränge wirkenden Verarbeitungsprozessen (Semantik, Kontext usw.) identisch.

In den Ländern der ehemaligen Sowjetunion wurden die von Frumkina et al. entwickelten Methoden für verschiedene psycholinguistische Fragestellungen fruchtbar gemacht. Varianzanalytisch angelegte Untersuchungen bestätigten, dass die subjektive Vorkommenshäufigkeit dem „Frequenzindex“ näher kommt als jene Daten, die an Textkorpora erhoben wurden. U.a. wurde nachgewiesen, dass das (auditive) Erkennen von sprachlichen Stimuli unter verschiedenen Konditionen stärker vom Faktor „F_{sub}“ determiniert wird als vom Faktor „F_{ob}“ (Štern 1982; Krause 1989; Čugaeva 1989).

Leider ist der Forschungsansatz von Frumkina et al. im Westen nicht rezipiert worden. Im Folgenden sollen daher die theoretischen und methodischen Grundlagen der Methode genauer vorgestellt und ihre Vorzüge und Schwachstellen diskutiert werden.

3. Subjektive Häufigkeitsbewertung

3.1. Hierarchische Struktur, Komplexität und Dynamik

Das Verfahren basiert auf der Hypothese, dass es einen Zusammenhang zwischen der sprachlichen Erfahrung (Input und Output) eines Individuums und der Struktur seines internen Lexikons gibt. „Sprachliche Erfahrung“ als Resultat der Verarbeitung von In- und Output wird als **integrales** Produkt angesehen. Es umfasst die gesamte sprachliche Tätigkeit des Individuums – Lesen wie Schreiben, Sprechen wie Hören.[5] Ein Urteil, das auf diesem Produkt integraler Verarbeitung beruht, ist selbstverständlich **komplexer** als jede Korpusauszählung.

Der Zusammenhang zwischen Spracherfahrung und Lexikon wurde zunächst folgendermaßen formuliert (Frumkina, Vasilevič 1971, 7):

1. Das interne Lexikon hat eine hierarchische Struktur.

2. Diese Struktur ergibt sich u.a. aus den Vorkommenshäufigkeiten der Einheiten[6] in der sprachlichen Tätigkeit des Sprechenden/Hörenden. Der Platz, den ein Eintrag in der Hierarchie des Lexikons einnimmt, wird als Funktion des Frequenzindex betrachtet. Je häufiger ein Wort vorkommt, um so höher in der Hierarchie ist es angesiedelt und um so schneller kann darauf zugegriffen werden.

Gleichzeitig wird von einer weitgehenden interpersonalen Übereinstimmung der Struktur interner Lexika ausgegangen, Folgender Zusammenhang wird angenommen: Je homogener die Gruppe, an der das interne Lexikon untersucht wird, umso stärker die Übereinstimmung der Häufigkeitsbewertungen (ebd., 8).

---

[5] Ėl'kin, Štern (1992, 160) weisen in diesem Zusammenhang darauf hin, dass die Basis für Frequenzbewertungen nicht allein im sprachlich belegbaren In- und Output zu suchen ist. Vielmehr sind Frequenzurteile auch sensitiv gegenüber dem – außersprachlichen – Vorkommen der Denotate. Das scheint vor allem damit zusammenzuhängen, dass mit der Wahrnehmung und Identifikation von Dingen und Sachverhalten durch den Menschen in der Regel auch das Bestreben nach verbaler Zuordnung (Nomination) verbunden ist – ohne dass dies bewusst oder gar ausgesprochen wird. Insofern weisen subjektive Frequenzdaten natürlich über die sprachliche Domäne hinaus.

[6] Die Formulierung „Einheit“ wurde bewusst gewählt, weil neben Lexemen u.a. auch Buchstabencluster und Phonemverbindungen getestet wurden (u.a. Frumkina, Vasilevič, Gerganov 1971). Die folgende Darstellung wird sich allerdings auf lexikalische Einträge, auf Wörter – oder, in strengerem Sinne: auf Lemmata – beschränken.

Sowohl der Frequenzindex selbst als auch die Art des Zugriffs werden in diesem Modell nicht als statische Größen begriffen. Der Index unterliegt Dynamiken, die mit Veränderungen im Kommunikations- und Lebensumfeld des Individuums einhergehen und verändertes Weltwissen reflektieren. Diese Dynamiken lassen sich sowohl im Längs- als auch im Querschnittsvergleich verschiedener Probandengruppen nachweisen.

So untersuchten Ėl'kin, Štern (1992) und Ovčinnikova et al. (2000) anhand von Häufigkeitsbewertungen die Entwicklung des mentalen Lexikons bei Kindern im Vorschul- und frühen Schulalter. Für die Bewertungen von Erwachsenen und Kindern wurden Rangkorrelationskoeffizienten von +0,65 (Ėl'kin, Štern 1992, 158) bzw. +0,82 (Ovčinnikova et al. 2000, 53) berechnet. Diese Werte verweisen sowohl auf die bereits früh bestehenden Übereinstimmungen (beide Arbeiten beginnen mit der Untersuchung 6-Jähriger) wie auch auf die existierenden Unterschiede. Die Untersuchungen zur ontogenetischen Entwicklung beschränken sich bisher auf jüngere Schulkinder; wünschenswert wäre ein Vergleich über größere Entwicklungszeiträume hinweg. Gegenwärtig ist eine Studie in Arbeit, die neben dem Faktor „Alter" auch sprach- und kulturbedingte Unterschiede in den Bewertungen der Vorkommenshäufigkeit zum Gegenstand hat (Uglanova, in Vorbereitung).

3.2. Die experimentelle Methode

Subjektive Frequenzbewertungen können auf verschiedene Weise erhoben werden. Eine Möglichkeit bietet der Paarvergleich von jeweils zwei Stimuli. Dabei ergibt sich die Notwendigkeit, jedes Zielwort mit jedem anderen zu kombinieren. Die Anzahl der möglichen Testwörter ist daher von vornherein relativ begrenzt.

Ein weitere Methode besteht in der Ermittlung einer Rangfolge. Soll dabei der gegenseitige Bezug der Stimuli für die Probanden fassbar bleiben, dann ist auch bei dieser Methode die Anzahl der einzubeziehenden Stimuli eingeschränkt.

Das dritte Verfahren, die sog. „Skalierung in aufeinanderfolgende Intervalle", erwies sich für die Aufgabenstellung als der am besten geeignete Weg. Den Probanden werden zur Bewertung der Stimuli mehrere Kategorien vorgegeben, die sich hinsichtlich der Intensität des von ihnen repräsentierten Merkmals unterscheiden. Als sinnvoll erwies sich die Verwendung von fünf bzw. sieben Kategorien. Für die Arbeit mit sieben Graduierungen wurden folgende, dem Symmetrieprinzip entsprechende Bezeichnungen empfohlen: „auf Schritt und Tritt" [7], „sehr häufig" [6], „eher häufig als selten" [5], „weder selten noch häufig" [4], „eher selten als häufig" [3], sehr selten [2], „niemals" [1] (Frumkina, Vasilevič 1971, 20).

Die Versuchspersonen (Vpn.) haben die Aufgabe, jedes Stimuluswort einer Bewertungskategorie zuzuordnen. In der Instruktion wird gebeten einzuschätzen, wie oft das Wort nach der Erfahrung des Probanden in der jeweiligen Sprache vorkommt.

Die Wörter werden in der Nennform, also als Lemmata, vorgelegt. Die Zuordnung erfolgt, ohne dass ein *direkter* Bezug zu den anderen Stimuli hergestellt werden muss. Insofern bleibt die Autonomie jedes einzelnen Stimulus gewahrt, obwohl natürlich eine grundsätzliche Kontextsensitivität gegenüber den anderen Stimuli eingeräumt werden muss. Dieser Tatsache wird u.a. dadurch Rechnung getragen, dass die Vpn. einmal getroffene Urteile revidieren dürfen. Gleichzeitig werden sie dazu aufgefordert, ihre Entscheidung ohne langes Überlegen zu treffen.

3.3. Statistische Verfahren: Auswertung

Aus psychologischer Sicht entspricht die dem Verfahren zugrundeliegende Bewertungsskala eine Intervallskala. Da aber nicht gesichert ist, dass die Intervalle auf dieser Skala gleich groß sind, wird sie bei der statistischen Auswertung als Rangskala betrachtet. Diese Interpretation birgt eine Reihe von Konsequenzen für die statistische Bearbeitung in sich. Beispielsweise ist es nicht zulässig, aus den Antworten der Probanden zu jedem einzelnen Stimulus das arithme-

tische Mittel zu berechnen. Als Mittelwert wird statt dessen der Median bestimmt. Der Median Me ist derjenige Wert in einer nach der Größe geordneten Reihenfolge von Messwerten, der diese Reihe halbiert. Dabei kommt folgende Formel zur Anwendung:

$$Me = L_i + k_i(N/2 - N_1)/ N_2 , \quad (1)$$

wobei i die Codezahl der Kategorie ist, die den gesuchten Median enthält. L_i ist die untere Grenze dieser i-ten Kategorie. Sie liegt jeweils bei i - 0,5. Bei der Festlegung der unteren Kategoriengrenze weiche ich von Frumkina/Vasilevič ab. Sie setzten in der Regel eine ganze Zahl als untere Kategoriengrenze an, wobei es in ihren Arbeiten in Bezug auf diese Größe gewisse Unregelmäßigkeiten gibt. In Anlehnung an Guilford (1954, 225f) wähle ich die untere Grenze so, dass die Codezahlen die Klassenmitte bilden; k_i ist die Breite der Kategorie. Es wird $k_i = 1$ angenommen (vgl. Guilford 1954, 120f.). N ist die Gesamtzahl der Probanden, N_1 ist die Zahl derjenigen Vpn., die den Stimulus unterhalb der Kategorie, in die der Median fällt, einordnen. N_2 ist die Anzahl der Vpn., deren Reaktionen in die Kategorie fallen, die den Median enthält. Die auf diese Weise berechneten Medianwerte sind keine ganzen, sondern gebrochene Zahlen. Sie erlauben die Festlegung einer differenzierten Rangfolge.

Als Streuungsmaß wird der durch 2 geteilte Quartilabstand berechnet. Die Quartile werden analog zum Median gebildet:

$$Q_1 = L_i + k_i (N/4 - N_1)/N_2 \quad (2)$$
$$Q_3 = L_i + k_i (3N/4 - N_1)/N_2. \quad (3)$$

Der Quartilabstand wird dann nach folgender Formel bestimmt:

$$QuA = (Q_3 - Q_1)/2. \quad (4)$$

Der Quartilabstand gibt an, inwieweit die Probandenurteile übereinstimmen. Zur qualitativen Bewertung des Streuungsmaßes errechneten Frumkina, Vasilevič (1971, 24) für eine 7-stufige Skala die folgenden kritischen Werte:

QuA/2 -Werte	Qualitative Bewertung der Streuung
$0{,}25 \le QuA/2 \le 0{,}60$	Einhellige Bewertung
$0{,}61 \le QuA/2 \le 0{,}90$	gute Übereinstimmung
$0{,}91 \le QuA/2 \le 1{,}10$	mittelmäßige Übereinstimmung
$1{,}11 \le QuA/2 \le 1{,}80$	geringe Übereinstimmung
$1{,}81 \le QuA/2 \le 2{,}00$	geringe Übereinstimmung mit Tendenz zu bimodaler Verteilung
$2{,}01 \le QuA/2 \le 2{,}50$	bimodale Verteilung der Bewertungen

3.4. Vorzüge und Nachteile der Methode

Ein grundlegender Vorteil der vorgestellten Methode besteht in ihrer Handhabbarkeit. Im Vorfeld von Untersuchungen, die eine Kontrolle des Faktors „Vorkommenshäufigkeit“ erfordern, lassen sich die notwendigen Frequenzdaten relativ einfach und an kleinen Probandengruppen erheben. Im Vergleich zu herkömmlichen korpusbasierten Häufigkeitsangaben hat das Verfahren außerdem den Vorzug, die aus der Komplexität der sprachlichen Tätigkeit der Probanden resultierenden Strukturen und Hierarchien integral zu erfassen.

Andererseits folgt aus dem komplexen und gleichzeitig dynamischen Charakter der Probandenurteile eine erhöhte Sensitivität der ermittelten Frequenzdaten gegenüber der Stichprobenauswahl – sowohl in Bezug auf die Probanden wie auch in Bezug auf die Testwörter. Die-

ses Problem wurde von der Arbeitsgruppe um Frumkina ausführlich untersucht, denn es hat eine ganze Reihe praktischer und methodischer Implikationen.

Die Stabilität von Bewertungen unter den Bedingungen unterschiedlich zusammengesetzter Materialkorpora (und Probandengruppen) wurde von Vasilevič (1971) speziell getestet. Er stellte fest, dass die Daten, die für ein und dieselben Stimuli in verschiedenen Stichproben erhoben wurden, gut miteinander korrelieren: die Rangkorrelationskoeffizienten lagen bei 0,90 bis 0,95 ($p<0,001$) (Vasilevič 1971, 51). Lediglich bei 6,5 % aller Stimuli betrugen die Abweichungen mehr als eine Graduierung (1,00); in 55 % der Fälle war sie kleiner als 0,5; in 33 % der Fälle betrug die Abweichung nicht mehr als 0,3 Graduierungen (ebd.). Diese Ergebnisse haben praktische Bedeutung für den möglichen Umfang von Frequenzlisten, denn auch bei der vorliegenden Skalierungsmethode sind dem Material quantitative Grenzen gesetzt. Vasilevič (1971) geht von 100–150 Wörtern pro Bewertungsdurchgang aus; ich selbst habe mit ca. 200 Wörtern gearbeitet. Eine Frequenzliste sollte also zwischen 100–200 Einträgen umfassen. Unter Umständen ist dieser Umfang in Hinblick auf weitere, am Material zu bearbeitende Fragestellungen zu gering. In einem solchen Fall sind mehrere Versuchsserien erforderlich. Sie sollten nach Möglichkeit mit ein und denselben Testpersonen durchgeführt werden. Die Ergebnisse zu den Abweichungen suggerieren, dass der Fehler, der durch die strukturellen Eigentümlichkeiten verschiedener Materialkorpora entstehen und zu einer Heterogenität zusammengefasster Frequenzlisten führen könnte, relativ gering ist. Seine Existenz ist jedoch prinzipiell nicht zu negieren.

Ebenso sind Schwankungen im Urteil ein und derselben Personen über die Zeit nicht auszuschließen – hier wirkt das dynamische Prinzip der internen Strukturierung des Lexikons. Als über die Zeit besonders stabil erweisen sich jene Stimuli, für die eine einhellige bis gute Übereinstimmung der Probandenurteile ermittelt wurde. Stärkere Divergenzen konzentrieren sich auf jene Wörter, für welche die Bewertungen der Vpn. stark streuen (Frumkina, Vasilevič 1971, 26-27).

Gleichzeitig ist zu beachten, dass all jene Mittelwerte (Mediane) in ihrer Aussagekraft und Verallgemeinerbarkeit am zuverlässigsten sind, die an den Polen der Häufigkeitsskala lokalisiert sind. Für Wörter, deren Medianwerte in die mittleren Graduierungen fallen, ist zu klären, ob das Ergebnis einer breiten Streuung der Probandenurteile geschuldet ist oder aber aus der überzufälligen Einordnung des Stimulus in eine bestimmte Kategorie resultiert.

Die „Korpussensitivität“, die bereits im Zusammenhang mit der Erfassung von Vorkommenshäufigkeiten in der gesprochenen Sprache angedeutet wurde, ist eine der Gemeinsamkeiten zwischen korpus- und urteilsbasierten Frequenzwörterbüchern. Sie betrifft auch die Auswahl der Probanden, deren Urteile in Abhängigkeit von Faktoren wie Alter, Beruf, Lebenssphäre, Geschlecht variieren können. Systematische Untersuchungen gab es dazu bisher nur mit Blick auf Entwicklungen in der Ontogenese (vgl. 3.1). Wie bei der Arbeit mit gesprochener Sprache ist daher eine möglichst genaue Charakterisierung der Probandenpopulation, auf die sich die statistischen Erhebungen und Verallgemeinerungen beziehen, wünschenswert. Die Arbeit mit 25 Probanden, so hat sich gezeigt, liefert bereits eine aussagekräftige Statistik.

4. Deutsche Frequenzlisten

4.1. Frequenzliste 1

Nach der Methode von Frumkina et al. führte ich Mitte der 1980-er Jahre eine erste Serie von Experimenten mit deutschen Stimuli durch. Die Materialauswahl umfasste 642 Lexeme. Drei davon waren 3 Nonsenswörter (*linnig*, *Trumpe*, *preinen*). Sie wurden eingebaut, um die Kategorie „niemals“ [1] zu stützen. Die Zusammenstellung des Materials erfolgte unter den Gesichtspunkten einer nachfolgenden Untersuchung. Es wurden primär Substantive, Verben und

Adjektive in der Kurzform ausgewählt. Außerdem fanden die Faktoren „Länge in Silben“, „rhythmische Struktur“ und „betonter Vokal“ Beachtung.

Die Wörter wurden in drei Serien zu jeweils 214 Wörtern von 25 Versuchspersonen bewertet. Den Vpn. standen die in 3.2 beschriebenen sieben Bewertungskategorien zur Verfügung. Das Alter der TeilnehmerInnen lag zwischen 18 und 45 Jahren. Sie entstammten einem Teil der Bevölkerung, den man heute als gebildete Mittelschicht betrachten würde. Alle Versuchspersonen hatten Mittelschulabschluss bzw. Hochschulreife, unterschieden sich jedoch stark in ihrer beruflichen Tätigkeit. Zum Zeitpunkt der Experimente lebten sie im Süden der ehemaligen DDR, im sächsischen Sprachraum; sie sprachen jedoch alle Hochsprache bzw. einen regional gefärbten Standard.

Die Ergebnisse der drei Serien wurden in *einer* Frequenzliste zusammengeführt. Im Anhang sind die Wörter sowohl nach der Größe der berechneten Me-Werte (Tabelle 1) als auch nach ihrer alphabetischen Reihenfolge (Tabelle 2) geordnet.

Die Liste 1 weist einige Besonderheiten auf. Zum einen enthält sie einige Lemmata, die heute sicher anders bewertet werden würden als zum Zeitpunkt der Experimente (z.B. *Kommunist*, *Kollektiv*, *Marxist*, *Wandzeitung*, *Staatsrat*, *Sputnik*). Hier widerspiegeln sich auf sprachlicher Ebene die einschneidenden gesellschaftlichen Veränderungen, die seit 1989 nicht nur im Osten Deutschlands stattgefunden haben: Wörter werden verdrängt, andere erobern ihren Platz, oder aber es werden andere Bedeutungen aktualisiert.

Einige wenige Stimuli der Liste 1 lassen Homonyme zu, z.B. *Kiefer*, *spicken*. Ohne zusätzliche Erhebungen (z.B. Assoziationsversuche) oder Spezifikationen der Stimuli ist im Nachhinein nicht klar zu bestimmen, auf welches Lexem sich die Häufigkeitsbewertungen bezogen. Dieses Problem wurde in den späteren Versuchen beachtet.

Die Analyse der Streuungsdaten zeigt, dass die Stimuli von den Probanden verhältnismäßig homogen bewertet wurden. Für 25 % aller Stimuli kann von relativ einhelligen Urteilen ausgegangen werden; das Streuungsmaß beträgt in diesen Fällen $0{,}25 \leq QuA/2 \leq 0{,}60$ (vgl. 3.3). Unter den einhellig bewerteten Stimuli dominieren jene, die an den Polen der Skala lokalisiert sind. Besonders hoch ist die Übereinstimmung der Urteile für Wörter mit geringer Häufigkeitsbewertung (Graduierungen „niemals“ [1] bis „eher selten als häufig“ [3]): auf sie entfallen 55 % der weitgehend einhelligen Bewertungen.

Lediglich 3 % der Stimuli weisen eine geringe Übereinstimmung der Probandenurteile auf; in diesen Fällen liegen die Streuungswerte in der Stichprobe zwischen $QuA/2 = 1{,}11$ und $QuA/2 = 1{,}33$ (für den Stimulus *Gleichung*). Die relativ hohen Streuungswerte konzentrieren sich auf Wörter, die in der Mitte der Häufigkeitsskala verortet wurden.

4.2. Frequenzliste 2

Die Frequenzliste 2 besteht aus 144 deutschen Stimuli, deren russische Äquivalente für ein Perzeptionsexperiment mit deutschen Probanden benutzt wurden (Krause 1989). Wiederum enthalten sind einige wenige Nonsenswörter. Ein Teil der Substantive hat sein Pendant im russischsprachigen subjektiven Frequenzwörterbuch von Vasilevič (1971, 57-68). Die Auswahl wurde zu Vergleichszwecken erstellt. Sie enthält auch einige umgangssprachliche Übersetzungsäquivalente. So wurde russ. *ručka* als *Kugelschreiber* und als *Kuli* übertragen und getrennt zur Bewertung vorgelegt.[7] Einige Stimuli wurden näher spezifiziert, um das Problem der Homonymie zu kontrollieren (*Stab (militärisch)*, *(Park-)Bank*).

Die Stimuli wurden von 25 Slavistikstudenten und -studentinnen der Friedrich-Schiller-

[7] Für den Stimulus *Kuli* wurden allerdings nur von 14 Vpn. Angaben erhoben. Der Unterschied in der Bewertung der Synonyme (Me = 6,17 für *Kuli* und Me = 4,00 für *Kugelschreiber*) zeigt, dass tatsächlich vor allem das versprachlichte Vorkommen bewertet wurde. In der Frequenzliste 1 (Tabellen 1 und 2 im Anhang) wurde für *Kuli* Me= 5,77 bestimmt; die Abweichung liegt bei 0,5.

Universität in einem Durchgang bewertet. Zum Zeitpunkt der Experimente (1986/1987) waren die Vpn. zwischen 18 und 20 Jahre alt. Den Testpersonen standen wiederum sieben Graduierungen zur Verfügung (vgl. 3.2).

Hinsichtlich der Homogenität und Verlässlichkeit der Probandenurteile lassen sich ähnliche Schlüsse ziehen wie für die erste Liste. Für 7 % der Stimuli streuen die Urteile relativ stark ($1{,}11 \leq QuA/2 \leq 1{,}34$ (1,34 = maximaler Streuungswert in dieser Stichprobe)), wobei wiederum Wörter dominieren, die in den mittleren Bewertungskategorien zu finden sind. 20 % der Stimuli weisen weitgehend einhellige Bewertungen auf, wobei die Randkategorien ebenfalls wieder stärker vertreten sind als die mittleren Graduierungen.

5. Zusammenfassung

Der vorliegende Beitrag sollte einen ausführlichen Überblick über ein experimentelles Verfahren geben, das neben der Erhebung von Angaben über das subjektive Empfinden von Vorkommenshäufigkeiten auch bei vollkommen anderen (psycho-)linguistischen Fragestellungen Anwendung finden kann, z.B. bei der Untersuchung der epistemischen Modalität (Krause 1998).

Zur Veröffentlichung der Frequenzlisten wurde ich durch die Arbeiten von I. Uglanova und I. Ovčinnikova angeregt. Sie hätte sicher zu einem früheren Zeitpunkt erfolgen sollen. Dennoch glaube ich, dass die Frequenzlisten im Wesentlichen nichts an empirischem Wert eingebüßt haben und nach wie vor einen guten Materialfundus für Untersuchungen bieten, in denen der Faktor „Vorkommenshäufigkeit“ kontrolliert werden muss.

Literatur

Aitchinson J. (1997). *Wörter im Kopf: Eine Einführung in das mentale Lexikon*. Tübingen: Niemeyer.

Bock M. (1978). *Wort-, Satz- und Textverarbeitung*. Stuttgart: Kohlhammer.

Čugaeva T.N. (1989). *Mechanizmy audirovanija rodnoj i inojazyčnoj reči.* Dissertacija na soiskanie učenoj stepeni kandidata filologičeskich nauk. Leningrad. [Auditive Mechanismen in der Mutter- und der Fremdsprache. Dissertation.]

Ėl'kin Ju. A., Štern A.S. (1992). Slovar' sub''ektivnych častot slov kak otraženie mira rebenka. In: Cejtlin S.N. (ed.), *Detskaja reč': lingvističeskij aspekt: 153-164.* Skt. Peterburg: Obrazovanie. [Das subjektive Frequenzwörterbuch als Spiegel der Welt des Kindes. In: Kindersprache: der linguistische Aspekt.]

Frumkina R.M., Vasilevič A.P. (1971). Polučenie ocenok verojatnostej slov psichometričeskimi metodami. In: Frumkina R.M. (ed.), *Verojatnostnoe prognozirovanie v reči: 7-28.* Moskva: Nauka. [Die Erhebung von Bewertungen der Vorkommenshäufigkeit von Wörtern mit psychometrischen Methoden. In: Wahrscheinlichkeitsprognose in der Sprache.]

Frumkina R.M., Vasilevič A.P., Gerganov E.N. (1971). Sub''ektivnye ocenki častot ėlementov teksta kak prognozirujuščij faktor. In: Frumkina R.M. (ed.) *Verojatnostnoe prognozirovanie v reči: 70-93.* Moskva: Nauka. [Die subjektive Häufigkeitsbewertung von Textelementen als prognostizierender Faktor. In: Wahrscheinlichkeitsprognose in der Sprache.]

Guilford J. P. (1954). *Psychometric Methods*. New York: McGraw-Hill.

Kaeding F.W. (1898). *Häufigkeitswörterbuch der deutschen Sprache.* Steglitz bei Berlin: Selbstverlag.

Krause M. (1989). *Dinamika mechanizma vosprijatija slova pri različnych uslovijach ovla-*

denija inostrannym jazykom. Dissertacija na soiskanie učenoj stepeni kandidata filologičeskich nauk. Leningrad. [Dynamik des Mechanismus der Wortwahrnehmung unter verschiedenen Bedingungen des Fremdsprachenerwerbs. Dissertation.]

Krause M. (1992). Ein Modell zur Beschreibung linguistischer Faktoren der Wortwahrnehmung. In: Hess W., Sendlmeier W.F. (ed.), *Beiträge zur angewandten und experimentellen Phonetik: 56-69.* Stuttgart: Steiner.

Krause M. (1998). Prosodic correlates of certainty – uncertainty in utterances with modal words. In: *Proceedings of the 14th International Congress of Phonetic Sciences. San Francisco, 1313-1315.*

Lönngren L. (1993). *Častotnyj slovar' russkogo jazyka*. Uppsala: Almquist & Wiksell. (Frequenzwörterbuch des Russischen.)

Marslen-Wilson W. (1989). Access and Integration: Projecting Sound onto Meaning. In: Marslen-Wilson W. (ed.), *Lexical Representation and Process: 3-24*. Cambridge: MIT Press.

Michaels S., Collins S. (1984). Oral discourse styles: classroom interactions and the acquisition of literacy. In: Tannen D. (ed.), *Coherence in Spoken and Written Discourse: 219-244*. Norwood, N.J.: Ablex.

Nuyts J. (2000). *Epistemic Modality, Language, and Conceptualization*. Amsterdam, Philadelphia: Benjamins.

Ovčinnikova I.G., Beresneva N.I., Dubrovskaja L.A., Penjagina E.B. (2000). *Leksikon mladšego škol'nika.* Perm': Izdatel'stvo Permskogo universiteta. [Das Lexikon des jüngeren Schulkinds.]

Ruoff A. (ed.) (1981[1], 1990[2]). *Häufigkeitswörterbuch der gesprochenen deutschen Sprache*. Tübingen: Niemeyer.

Sappok Ch. (1998). Der dialogisch organisierte Dialekttext aus diskursiver, auditiver und gattungsbezogener Sicht. In: *Wiener Slawistischer Almanach 41, 263-288*.

Savin H.R. (1963).Word frequency effect and errors in the perception of speech. *Journal of the Acoustic Society of America 35, 200-206.*

Schönefeld D. (2001). *Where Lexicon and Syntax Meet*. Berlin.

Štejnfeldt E.V. (1963). *Častotnyj slovar' sovremennogo russkogo literaturnogo jazyka*. Tallinn: NII Pedagogiki ĖSSR. [Frequenzwörterbuch des modernen Standardrussischen.]

Štern A.S. (1992). *Perceptivnyj aspekt rečevoj dejatel'nosti*. Skt. Peterburg: Izdatel'stvo Skt. Peterburgskogo universiteta. [Der perzeptive Aspekt der sprachlichen Tätigkeit.]

Štern A.S. (1982). *Vlijanie lingvističeskich faktorov na vosprijatie reči*. Dissertacija na soiskanie učenoj stepeni kandidata filologičeskich nauk. Leningrad. [Der Einfluss linguistischer Faktoren auf die Sprachwahrnehmung.]

Uglanova I.A. (in Vorbereitung) *Frequenzbewertungen und mentales Lexikon im Grundschulalter: empirische Untersuchung im Sprachvergleich.*

Vasilevič A.P. (1971). K voprosu ob ispol'zovanii sub''ektivnych ocenok kak istočnika svedenij o častote slov-stimulov. In: Frumkina R.M. (ed.), *Verojatnostnoe prognozirovanie v reči: 44-69.* Moskva: Nauka. [Zur Frage der Nutzung subjektiver Bewertungen als Informationsquelle über die Vorkommenshäufigkeit von Stimuluswörtern. In: Wahrscheinlichkeitsprognose in der Sprache.]

Anhang: Frequenzlisten 1 und 2

Tabelle 1. Frequenzliste 1, sortiert nach Median (Me) und halbiertem Quartilabstand (QuA/2)

Nr.	Lexem	Me	QuA/2
1	haben	6,93	0,28
2	gut	6,84	0,33
3	Buch	6,67	0,53
4	Zeit	6,67	0,53
5	Frieden	6,67	0,56
6	Uhr	6,36	0,50
7	studieren	6,33	0,65
8	fragen	6,13	0,57
9	Butter	6,11	0,47
10	gerade	6,11	0,56
11	Partei	6,07	0,42
12	Milch	6,06	0,42
13	politisch	6,06	0,80
14	Flasche	6,00	0,42
15	sitzen	6,00	0,42
16	gleich	6,00	0,62
17	frei	5,96	0,45
18	hart	5,94	0,96
19	Dienstag	5,92	0,51
20	Gabel	5,92	0,55
21	Sprache	5,90	0,42
22	Gruppe	5,89	0,47
23	Inhalt	5,88	0,37
24	Hand	5,88	0,59
25	Republik	5,87	0,49
26	teilnehmen	5,85	0,54
27	gemeinsam	5,85	0,72
28	Zucker	5,85	0,76
29	Glas	5,82	0,32
30	praktisch	5,82	0,51
31	lustig	5,82	0,64
32	Programm	5,80	0,53
33	Fuß	5,79	0,64
34	Kino	5,79	0,68
35	Haar	5,78	0,80
36	international	5,77	0,55
37	Kuli	5,77	0,60
38	Kommunist	5,77	0,72
39	Radio	5,77	0,72
40	stattfinden	5,75	0,52
41	Grund	5,75	0,67
42	Käse	5,73	0,67
43	Mittag	5,73	0,71
44	Resultat	5,71	0,63
45	genug	5,71	0,89
46	Junge	5,69	0,65
47	Kollektiv	5,67	0,91
48	Industrie	5,65	0,65
49	Gruß	5,64	0,65
50	probieren	5,62	0,62
51	Tier	5,62	0,72
52	Sitzung	5,61	0,54
53	spielen	5,61	0,69
54	schlimm	5,56	0,49
55	gefallen	5,55	0,93
56	Marxist	5,55	0,72
57	Blume	5,54	0,60
58	tief	5,54	0,62
59	Strumpf	5,54	0,67
60	Bericht	5,54	0,79
61	fleißig	5,45	0,52
62	Sinn	5,44	0,61
63	Institut	5,44	0,68
64	Gesundheit	5,44	0,74
65	schief	5,43	0,77
66	schneiden	5,43	0,84
67	funktionieren	5,38	0,54
68	Blödsinn	5,37	1,08
69	Erziehung	5,36	0,55
70	Kultur	5,36	0,65
71	bitten	5,36	0,75
72	Bild	5,35	0,60
73	Planung	5,35	0,62
74	Hund	5,33	0,65
75	Gleichung	5,33	1,33
76	Rand	5,32	1,07
77	Ball	5,31	0,72
78	billig	5,31	0,75
79	Freiheit	5,29	0,55
80	demokratisch	5,28	0,80

Nr.	Lexem	Me	QuA/2
81	Film	5,27	0,67
82	Spiegel	5,27	0,71
83	Kleinigkeit	5,27	0,82
84	wach	5,27	0,86
85	Disziplin	5,25	0,67
86	Kampf	5,25	0,68
87	beeinflussen	5,25	0,70
88	Spiel	5,25	0,86
89	privat	5,22	0,75
90	Fach	5,21	0,58
91	salzig	5,20	1,08
92	hierbleiben	5,19	0,83
93	dumm	5,18	0,62
94	Gedanke	5,18	0,64
95	Politik	5,14	0,64
96	putzen	5,14	0,87
97	reisen	5,12	0,59
98	Schluck	5,11	0,87
99	spazieren	5,09	0,39
100	Gedicht	5,09	0,60
101	Samstag	5,08	1,17
102	Knie	5,06	1,02
103	munter	5,05	0,70
104	Feier	5,05	0,71
105	ziehen	5,04	0,55
106	hinweisen	5,04	0,56
107	Klub	5,04	0,58
108	erkundigen	5,00	0,54
109	Temperatur	5,00	0,64
110	siegen	5,00	0,64
111	gewinnen	5,00	0,74
112	bearbeiten	5,00	0,77
113	heiraten	5,00	0,89
114	Blut	5,00	0,90
115	Geist	5,00	0,93
116	Druck	5,00	1,08
117	Kapital	5,00	1,14
118	aufmerksam	4,96	0,55
119	nachweisen	4,96	0,57
120	Fisch	4,95	0,74
121	Klima	4,94	0,90
122	empfangen	4,93	0,84
123	Jahrhundert	4,92	0,88
124	Gas	4,91	0,71
125	Darstellung	4,89	0,52
126	Demokrat	4,89	0,83
127	Tanz	4,88	0,60
128	Pfund	4,88	0,80
129	Fall	4,86	0,89
130	zeichnen	4,86	0,89
131	gemeinsam	4,86	0,94
132	außerordentlich	4,85	0,59
133	reich	4,85	0,59
134	glatt	4,85	0,64
135	Ideal	4,85	0,66
136	Soldat	4,85	0,72
137	Struktur	4,85	0,76
138	Kandidat	4,85	0,84
139	Aggressor	4,82	0,51
140	Wut	4,82	0,64
141	demonstrieren	4,82	0,67
142	Produkt	4,82	0,67
143	Rennerei	4,80	0,98
144	Autobus	4,80	1,30
145	geschmacklos	4,79	0,60
146	schießen	4,78	0,75
147	gewaltig	4,78	0,80
148	Geburt	4,77	0,53
149	Kindheit	4,77	0,82
150	beauftragen	4,75	0,52
151	grau	4,75	0,52
152	Faschist	4,75	0,75
153	Gardine	4,75	0,91
154	Gliederung	4,75	0,91
155	Dienst	4,75	0,98
156	reichen	4,75	1,00
157	Streik	4,75	1,08
158	Minimum	4,75	1,12
159	Phantasie	4,73	0,61
160	wandern	4,73	0,61
161	sparsam	4,73	0,62
162	Beleuchtung	4,71	0,63
163	Offizier	4,71	0,63
164	hauen	4,71	0,68
165	Thematik	4,71	0,82
166	Fabrik	4,71	0,96

Nr.	Lexem	Me	QuA/2
167	gering	4,71	0,96
168	Grippe	4,69	0,72
169	Genuss	4,69	0,82
170	springen	4,67	0,69
171	Polizei	4,67	0,73
172	Ring	4,65	0,62
173	bitter	4,65	0,65
174	Sitz	4,65	0,65
175	schlucken	4,65	0,68
176	flach	4,65	0,73
177	giftig	4,65	0,73
178	Gulasch	4,65	0,78
179	Schiff	4,65	1,10
180	Geheimnis	4,62	0,62
181	Gemeinsamkeit	4,62	0,62
182	gratulieren	4,58	0,91
183	reinigen	4,58	0,91
184	Regal	4,58	0,96
185	Wandzeitung	4,57	0,77
186	Bilanz	4,57	0,96
187	zart	4,56	0,78
188	Palast	4,56	0,83
189	Abhängigkeit	4,56	0,83
190	streicheln	4,56	0,95
191	konkurrieren	4,56	1,02
192	Hindernis	4,55	0,66
193	heiraten	4,55	0,85
194	Feigheit	4,53	0,82
195	schmal	4,45	0,66
196	geduldig	4,45	0,67
197	Urteil	4,45	0,75
198	mutig	4,45	0,82
199	betonen	4,44	0,55
200	Duft	4,44	0,66
201	Besichtigung	4,44	0,78
202	Getreide	4,44	0,78
203	Gummi	4,44	0,78
204	steil	4,44	0,83
205	spendieren	4,44	0,85
206	stinken	4,44	1,02
207	Alkoholiker	4,43	0,78
208	hindurch	4,43	0,96
209	illegal	4,43	1,03
210	Dreieck	4,43	1,10
211	Scheidung	4,42	1,05
212	Wein	4,40	0,59
213	konservativ	4,40	1,01
214	Impuls	4,40	1,12
215	klassisch	4,38	0,61
216	Klinik	4,38	0,66
217	Gleichgültigkeit	4,36	0,65
218	Schicksal	4,33	0,73
219	Absatz	4,33	0,82
220	Geige	4,33	0,82
221	Bauer	4,31	0,82
222	kahl	4,31	0,88
223	Schnitt	4,31	0,88
224	Teilung	4,31	1,04
225	Bescheidenheit	4,29	0,60
226	Umgang	4,29	0,63
227	gönnen	4,29	0,67
228	Fieber	4,29	0,74
229	wiegen	4,29	0,83
230	kitzlig	4,29	0,89
231	stricken	4,29	0,92
232	gießen	4,28	1,04
233	geradlinig	4,28	1,15
234	erleichtern	4,27	0,61
235	schuldig	4,25	0,65
236	Frist	4,25	0,75
237	Malerei	4,25	0,75
238	halbieren	4,25	0,81
239	Jahreszeit	4,25	0,94
240	Reichtum	4,25	0,94
241	schieben	4,25	0,95
242	Sandale	4,22	0,71
243	speichern	4,22	0,75
244	Staatsrat	4,21	0,75
245	schweigsam	4,19	0,81
246	graben	4,19	0,83
247	Drogerie	4,19	0,86
248	reinigen	4,19	0,89
249	Imbiss	4,19	0,94
250	wild	4,18	0,60
251	Sinfonie	4,18	0,61
252	Geheimnis	4,18	0,64

Nr.	Lexem	Me	QuA/2
253	Muskel	4,18	0,72
254	ermuntern	4,15	0,64
255	Galerie	4,15	0,68
256	Sack	4,15	0,74
257	gehorchen	4,15	0,95
258	Stil	4,14	0,86
259	Fruchtsaft	4,14	0,91
260	tanken	4,14	0,94
261	schematisch	4,14	1,05
262	diktieren	4,12	0,56
263	Gehalt	4,12	0,57
264	nackt	4,12	0,57
265	formal	4,12	1,00
266	Geleier	4,12	1,13
267	fundamental	4,11	0,72
268	reiben	4,11	0,91
269	standhalten	4,11	0,92
270	Profil	4,09	0,61
271	Schiedsrichter	4,09	0,61
272	Spruch	4,09	0,61
273	ratsam	4,09	0,62
274	Leidenschaft	4,09	0,63
275	Urkunde	4,09	0,63
276	dramatisch	4,09	0,70
277	Gliederung	4,08	1,16
278	ergiebig	4,06	0,78
279	turnen	4,06	0,81
280	Geiz	4,06	0,85
281	liberal	4,06	0,85
282	Teich	4,06	0,91
283	kitschig	4,06	0,92
284	betonen	4,04	0,45
285	Tal	4,00	0,42
286	Debatte	4,00	0,69
287	Signal	4,00	0,70
288	Spurt	4,00	0,75
289	Humanist	4,00	0,85
290	Kinderlied	4,00	0,86
291	spannen	4,00	0,89
292	Kaninchen	4,00	0,90
293	organisch	4,00	0,92
294	Gesang	4,00	1,00
295	speisen	4,00	1,13
296	Gemeinde	4,00	1,14
297	stumm	3,96	0,54
298	Besorgnis	3,95	0,71
299	Musikant	3,95	0,72
300	knallen	3,94	0,80
301	Solist	3,94	0,97
302	heikel	3,92	1,07
303	belichten	3,92	1,10
304	Demonstrant	3,92	1,17
305	nachahmen	3,89	0,80
306	spionieren	3,88	0,57
307	nahrhaft	3,88	0,60
308	behutsam	3,88	0,63
309	gleiten	3,88	0,62
310	Kiefer	3,88	0,91
311	Glut	3,86	0,85
312	General	3,85	0,28
313	Acker	3,85	0,56
314	weigern	3,85	0,64
315	Mandarine	3,85	0,72
316	spicken	3,85	0,72
317	ticken	3,85	0,95
318	Krug	3,82	0,62
319	umstimmen	3,82	0,64
320	undicht	3,82	0,64
321	Bär	3,82	0,67
322	Spirale	3,82	0,67
323	spitzfindig	3,82	0,67
324	Pfannkuchen	3,82	0,72
325	Register	3,82	0,72
326	schmunzeln	3,81	0,78
327	knurren	3,81	0,88
328	Gift	3,80	0,94
329	dokumentarisch	3,80	0,99
330	stickig	3,79	0,58
331	zersplittern	3,79	0,60
332	Gesinnung	3,79	0,64
333	Kanal	3,78	0,70
334	flink	3,78	0,71
335	Konditorei	3,78	0,72
336	bockig	3,78	0,76
337	Gips	3,75	0,67
338	kommunal	3,75	0,71

Nr.	Lexem	Me	QuA/2
339	Schwimmerin	3,75	0,71
340	intim	3,75	0,78
341	jammern	3,75	0,87
342	Klavier	3,75	0,91
343	Sputnik	3,75	0,91
344	Hudelei	3,75	1,15
345	klagen	3,73	0,67
346	Mahnmal	3,73	0,71
347	umkreisen	3,73	0,78
348	mahnen	3,73	0,78
349	Hubschrauber	3,73	1,12
350	Grab	3,71	0,75
351	Grube	3,71	0,79
352	belustigen	3,71	0,82
353	Gegebenheit	3,71	0,82
354	Inland	3,71	0,83
355	Hirn	3,69	0,75
356	Meile	3,69	0,75
357	entsinnen	3,69	0,82
358	Skandal	3,69	0,82
359	frisieren	3,69	0,88
360	Fluch	3,69	0,89
361	glitzern	3,67	0,68
362	schwindlig	3,67	0,70
363	krümmen	3,65	0,65
364	Druckerei	3,65	0,78
365	Jurist	3,64	0,69
366	zivil	3,64	0,69
367	Umzug	3,64	0,83
368	Beistand	3,63	0,75
369	schneiden	3,63	0,90
370	rascheln	3,63	0,92
371	grafisch	3,63	1,16
372	heilsam	3,61	0,54
373	handelsüblich	3,61	0,75
374	Intoleranz	3,57	0,80
375	plump	3,57	0,80
376	Glanz	3,56	0,64
377	lieblich	3,56	0,74
378	streichen	3,56	0,74
379	unliebsam	3,56	0,74
380	Rache	3,56	0,76
381	textil	3,56	0,78
382	Wachs	3,56	0,83
383	knirschen	3,56	0,90
384	administrativ	3,56	0,98
385	plausibel	3,55	0,62
386	lateinisch	3,55	0,64
387	rudern	3,55	0,66
388	Getriebe	3,55	0,93
389	inhuman	3,55	0,95
390	Weisheit	3,55	0,57
391	platzieren	3,54	0,67
392	schmierig	3,46	0,60
393	spurten	3,46	0,76
394	spritzig	3,45	0,59
395	knicken	3,45	0,66
396	Tumult	3,45	0,66
397	mannigfaltig	3,45	0,70
398	glimmen	3,45	0,72
399	Lichtung	3,45	0,75
400	reiten	3,44	0,66
401	schlurfen	3,44	0,73
402	rieseln	3,44	0,74
403	schlicht	3,44	0,78
404	Skala	3,44	0,79
405	Tusche	3,44	0,79
406	Eintrag	3,44	0,90
407	kreuzen	3,44	0,94
408	Hamsterei	3,44	1,07
409	keimen	3,43	0,80
410	Skulptur	3,43	0,80
411	Doktrin	3,43	0,84
412	Gelatine	3,42	0,84
413	Keil	3,42	0,91
414	Schwur	3,38	0,62
415	kneifen	3,38	0,68
416	Schifffahrt	3,36	0,65
417	abscheulich	3,36	0,75
418	Schmeichelei	3,36	0,75
419	Ballon	3,35	0,65
420	Halunke	3,35	0,78
421	hinderlich	3,33	0,66
422	silbern	3,33	0,69
423	Krone	3,33	0,77
424	Sklaverei	3,33	0,82

Nr.	Lexem	Me	QuA/2
425	Rasur	3,33	0,91
426	gruselig	3,32	0,62
427	Disput	3,31	0,75
428	Prise	3,31	0,81
429	Kathedrale	3,31	0,82
430	Zufuhr	3,31	0,82
431	Fabel	3,29	0,60
432	enthaltsam	3,29	0,63
433	Spinat	3,29	0,85
434	schmeichlerisch	3,28	0,79
435	sprudeln	3,27	0,66
436	Hummel	3,27	0,67
437	beschlagnahmen	3,25	0,58
438	Hieb	3,25	0,62
439	Pudel	3,25	0,66
440	Damm	3,25	0,67
441	Frachter	3,25	0,67
442	Granit	3,25	0,70
443	schinden	3,25	0,71
444	Genick	3,25	0,75
445	Limit	3,25	0,80
447	Schall	3,23	0,40
446	Knabe	3,23	0,72
448	Habsucht	3,22	0,69
449	Mais	3,22	0,73
450	grimmig	3,22	0,80
451	mildern	3,21	0,60
452	weichen	3,21	0,60
453	reinlich	3,20	0,84
454	Diener	3,20	1,21
455	Schimpanse	3,19	0,79
456	Intrige	3,19	0,82
457	friedfertig	3,18	0,51
458	Statur	3,18	0,64
459	Fabrikant	3,18	0,67
460	Gratulant	3,18	0,67
461	Skrupel	3,18	0,67
462	Gießerei	3,18	0,79
463	Gerippe	3,15	0,43
464	abprallen	3,15	0,56
465	Spekulant	3,15	0,59
466	Schlupfwinkel	3,15	0,64
467	Schleier	3,15	0,66
468	dreist	3,15	0,94
469	steril	3,12	0,59
470	schlaff	3,12	0,66
471	fatal	3,11	0,78
472	panieren	3,09	0,61
473	Geschwulst	3,09	0,63
474	Mineral	3,09	0,66
475	Span	3,09	0,71
476	Kaliber	3,09	0,77
477	geschmeidig	3,08	0,35
478	Giraffe	3,06	0,67
479	Eid	3,06	0,78
480	geleiten	3,06	0,80
481	gigantisch	3,06	1,17
482	Dynamit	3,05	0,69
483	Heide	3,05	0,70
484	hissen	3,05	0,70
485	Stimulus	3,05	0,71
486	Zank	3,05	0,77
487	Gurgel	3,04	0,45
488	Katholik	3,04	0,54
489	Diamant	3,04	0,58
490	erhaben	3,00	0,42
491	gastieren	3,00	0,48
492	klumpig	3,00	0,48
493	preisen	3,00	0,48
494	beben	3,00	0,62
495	Graphiker	3,00	0,62
496	Distel	3,00	0,63
497	spreizen	3,00	0,63
498	blutarm	3,00	0,82
499	Schmuckkasten	3,00	0,85
500	filtrieren	3,00	0,87
501	Klausur	3,00	0,87
502	Spelunke	3,00	0,88
503	schrumpfen	2,96	0,54
504	Absolutismus	2,96	0,55
505	Humorist	2,96	0,55
506	gesunden	2,95	0,70
507	Sklavin	2,95	0,70
508	Tagesanbruch	2,94	0,87
509	deklarieren	2,92	0,96
510	Gleichnis	2,91	0,39

Nr.	Lexem	Me	QuA/2
511	karg	2,91	0,39
512	peinigen	2,91	0,62
513	Fabrikant	2,89	0,73
514	gebirgig	2,89	0,76
515	Pult	2,89	0,76
516	Giebel	2,86	0,83
517	Schwimmmeister	2,85	0,52
518	fleischig	2,85	0,64
519	schrill	2,85	0,64
520	geschwind	2,85	0,66
521	steinern	2,85	0,69
522	Pirat	2,82	0,49
523	Humus	2,82	0,60
524	schlitzäugig	2,81	0,47
525	bleiern	2,81	0,78
526	Bunsenbrenner	2,81	0,78
527	Kiesel	2,80	0,89
528	hegen	2,79	0,56
529	Dung	2,79	0,58
530	Erlass	2,78	0,69
531	Gigant	2,78	0,69
532	Golf	2,75	0,39
533	definitiv	2,75	0,65
534	destruktiv	2,75	0,65
535	Flussbett	2,75	0,65
536	Tribunal	2,73	0,63
537	dumpf	2,73	0,77
538	Glatteis	2,72	0,58
539	Radierung	2,72	0,58
540	Brut	2,71	0,55
541	Geweih	2,71	0,55
542	barmherzig	2,71	0,57
543	Greisin	2,71	0,79
544	buschig	2,69	0,72
545	Literat	2,69	0,72
546	Katapult	2,67	0,65
547	katzbuckeln	2,67	0,65
548	gravieren	2,67	0,69
549	wachrufen	2,67	0,69
550	gluckern	2,65	0,62
551	weichherzig	2,65	0,62
552	Filtrat	2,65	0,62
553	glucksen	2,64	0,62
554	kraus	2,62	0,56
555	ehrerbietig	2,60	0,83
556	splittern	2,59	0,34
557	Strudel	2,58	0,83
558	Getier	2,58	0,84
559	Gerassel	2,57	0,74
560	Epigramm	2,57	0,76
561	Gesuch	2,56	0,64
562	Galaxis	2,56	0,66
563	Gefilde	2,56	0,66
564	Jupiter	2,56	0,66
565	maritim	2,56	0,66
566	sinnwidrig	2,56	0,66
567	Glaserei	2,56	0,68
568	simultan	2,56	0,71
569	Sintflut	2,56	0,61
570	sittsam	2,56	0,61
571	Wucherer	2,55	0,58
572	Gestirn	2,55	0,60
573	Deich	2,55	0,64
574	Frühling	2,55	0,66
575	wurmstichig	2,54	0,52
576	Fremdling	2,54	0,55
577	salben	2,46	0,57
578	dominant	2,46	0,60
579	Harpune	2,46	0,64
580	Reisig	2,46	0,67
581	altdeutsch	2,46	0,74
582	Gusseisen	2,46	0,74
583	gallig	2,45	0,59
584	lukullisch	2,45	0,63
585	feminin	2,45	0,78
586	fuchsig	2,45	0,93
587	Eichhorn	2,44	0,79
588	Schwindsucht	2,39	0,57
589	skalpieren	2,39	0,57
590	Kadaver	2,38	0,54
591	Kalkulator	2,38	0,66
592	Staffelei	2,33	0,56
593	absorbieren	2,33	1,00
594	Hausierer	2,32	0,49
595	Stickerin	2,32	0,52
596	Gardist	2,32	0,58

Nr.	Lexem	Me	QuA/2
597	Gelübde	2,31	0,57
598	Hain	2,27	0,50
599	zagen	2,27	0,50
600	Hornisse	2,27	0,60
601	linieren	2,27	0,77
602	dublieren	2,25	0,52
603	Fossil	2,25	0,55
604	Schalmei	2,24	0,55
605	internieren	2,23	0,57
606	Geheiß	2,23	0,71
607	Skalp	2,19	0,39
608	Fraktur	2,18	0,43
609	spektral	2,18	0,53
610	Kellerei	2,18	0,61
611	sporadisch	2,18	0,64
612	Bache	2,13	0,44
613	feist	2,13	0,46
614	Lagune	2,12	0,31
615	Abtei	2,11	0,47
616	Gebaren	2,08	0,35
617	Gevatter	2,07	0,28
618	Drangsal	2,04	0,45
619	Gnu	2,03	0,37
620	gummieren	2,00	0,48
621	Abszess	1,97	0,37
622	Brunnenkresse	1,96	0,48
623	Kabale	1,95	0,33
624	steifen	1,93	0,42
625	Inspirator	1,91	0,39
626	halbseiden	1,85	0,52
627	Falbe	1,85	0,52
628	habil	1,82	0,50
629	Bassist	1,75	1,03
630	Kubismus	1,73	0,59
631	Broschur	1,69	0,70
632	Garnele	1,69	0,70
633	drainieren	1,61	0,51
634	Treiber	1,55	0,59
635	lasieren	1,46	0,50
636	dental	1,46	0,62
637	Binge	1,19	0,38
638	Adjunkt	1,12	0,31
639	Aphasie	1,09	0,30
640	linnig	1,07	0,28
641	Trumpe	1,02	0,26
642	preinen	1,00	0,25

Tabelle 2. Frequenzliste 1, alphabetisch sortiert

Lexem	Me	QuA/2
Abhängigkeit	4,56	0,83
abprallen	3,15	0,56
Absatz	4,33	0,82
abscheulich	3,36	0,75
Absolutismus	2,96	0,55
absorbieren	2,33	1,00
Abszess	1,97	0,37
Abtei	2,11	0,47
Acker	3,85	0,56
Adjunkt	1,12	0,31
administrativ	3,56	0,98
Aggressor	4,82	0,51
Alkoholiker	4,43	0,78
altdeutsch	2,46	0,74
Aphasie	1,09	0,30
aufmerksam	4,96	0,55
außerordentlich	4,85	0,59
Autobus	4,80	1,30
Bache	2,13	0,44
Ball	5,31	0,72
Ballon	3,35	0,65
Bär	3,82	0,67
barmherzig	2,71	0,57
Bassist	1,75	1,03
Bauer	4,31	0,82
bearbeiten	5,00	0,77
beauftragen	4,75	0,52
beben	3,00	0,62
beeinflussen	5,25	0,70
behutsam	3,88	0,63
Beistand	3,63	0,75
Beleuchtung	4,71	0,63
belichten	3,92	1,10
belustigen	3,71	0,82

Lexem	Me	QuA/2
Bericht	5,54	0,79
Bescheidenheit	4,29	0,60
beschlagnahmen	3,25	0,58
Besichtigung	4,44	0,78
Besorgnis	3,95	0,71
betonen	4,04	0,45
betonen	4,44	0,55
Bilanz	4,57	0,96
Bild	5,35	0,60
billig	5,31	0,75
Binge	1,19	0,38
bitten	5,36	0,75
bitter	4,65	0,65
bleiern	2,81	0,78
Blödsinn	5,37	1,08
Blume	5,54	0,60
Blut	5,00	0,90
blutarm	3,00	0,82
bockig	3,78	0,76
Broschur	1,69	0,70
Brunnenkresse	1,96	0,48
Brut	2,71	0,55
Buch	6,67	0,53
Bunsenbrenner	2,81	0,78
buschig	2,69	0,72
Butter	6,11	0,47
Damm	3,25	0,67
Darstellung	4,89	0,52
Debatte	4,00	0,69
definitiv	2,75	0,65
Deich	2,55	0,64
deklarieren	2,92	0,96
Demokrat	4,89	0,83
demokratisch	5,28	0,80
Demonstrant	3,92	1,17
demonstrieren	4,82	0,67
dental	1,46	0,62
destruktiv	2,75	0,65
Diamant	3,04	0,58
Diener	3,20	1,21
Dienst	4,75	0,98
Dienstag	5,92	0,51
diktieren	4,12	0,56

Lexem	Me	QuA/2
Disput	3,31	0,75
Distel	3,00	0,63
Disziplin	5,25	0,67
Doktrin	3,43	0,84
dokumentarisch	3,80	0,99
dominant	2,46	0,60
drainieren	1,61	0,51
dramatisch	4,09	0,70
Drangsal	2,04	0,45
Dreieck	4,43	1,10
dreist	3,15	0,94
Drogerie	4,19	0,86
Druck	5,00	1,08
Druckerei	3,65	0,78
dublieren	2,25	0,52
Duft	4,44	0,66
dumm	5,18	0,62
dumpf	2,73	0,77
Dung	2,79	0,58
Dynamit	3,05	0,69
ehrerbietig	2,60	0,83
Eichhorn	2,44	0,79
Eid	3,06	0,78
Eintrag	3,44	0,90
empfangen	4,93	0,84
enthaltsam	3,29	0,63
entsinnen	3,69	0,82
Epigramm	2,57	0,76
ergiebig	4,06	0,78
erhaben	3,00	0,42
erkundigen	5,00	0,54
Erlas	2,78	0,69
erleichtern	4,27	0,61
ermuntern	4,15	0,64
Erziehung	5,36	0,55
Fabel	3,29	0,60
Fabrik	4,71	0,96
Fabrikant	2,89	0,73
Fabrikant	3,18	0,67
Fach	5,21	0,58
Falbe	1,85	0,52
Fall	4,86	0,89
Faschist	4,75	0,75

Lexem	Me	QuA/2
fatal	3,11	0,78
Feier	5,05	0,71
Feigheit	4,53	0,82
feist	2,13	0,46
feminin	2,45	0,78
Fieber	4,29	0,74
Film	5,27	0,67
Filtrat	2,65	0,62
filtrieren	3,00	0,87
Fisch	4,95	0,74
flach	4,65	0,73
Flasche	6,00	0,42
fleischig	2,85	0,64
fleißig	5,45	0,52
flink	3,78	0,71
Fluch	3,69	0,89
Flussbett	2,75	0,65
formal	4,12	1,00
Fossil	2,25	0,55
Frachter	3,25	0,67
fragen	6,13	0,57
Fraktur	2,18	0,43
frei	5,96	0,45
Freiheit	5,29	0,55
Fremdling	2,54	0,55
Frieden	6,67	0,56
friedfertig	3,18	0,51
frisieren	3,69	0,88
Frist	4,25	0,75
Fruchtsaft	4,14	0,91
Frühling	2,55	0,66
fuchsig	2,45	0,93
fundamental	4,11	0,72
funktionieren	5,38	0,54
Fuß	5,79	0,64
Gabel	5,92	0,55
Galaxis	2,56	0,66
Galerie	4,15	0,68
gallig	2,45	0,59
Gardine	4,75	0,91
Gardist	2,32	0,58
Garnele	1,69	0,70
Gas	4,91	0,71

Lexem	Me	QuA/2
gastieren	3,00	0,48
Gebaren	2,08	0,35
gebirgig	2,89	0,76
Geburt	4,77	0,53
Gedanke	5,18	0,64
Gedicht	5,09	0,60
geduldig	4,45	0,67
gefallen	5,55	0,93
Gefilde	2,56	0,66
Gegebenheit	3,71	0,82
Gehalt	4,12	0,57
Geheimnis	4,18	0,64
Geheimnis	4,62	0,62
Geheiß	2,23	0,71
gehorchen	4,15	0,95
Geige	4,33	0,82
Geist	5,00	0,93
Geiz	4,06	0,85
Gelatine	3,42	0,84
Geleier	4,12	1,13
geleiten	3,06	0,80
Gelübde	2,31	0,57
Gemeinde	4,00	1,14
gemeinsam	4,86	0,94
gemeinsam	5,85	0,72
Gemeinsamkeit	4,62	0,62
General	3,85	0,28
Genick	3,25	0,75
genug	5,71	0,89
Genuss	4,69	0,82
gerade	6,11	0,56
geradlinig	4,28	1,15
Gerassel	2,57	0,74
gering	4,71	0,96
Gerippe	3,15	0,43
Gesang	4,00	1,00
geschmacklos	4,79	0,60
geschmeidig	3,08	0,35
geschwind	2,85	0,66
Geschwulst	3,09	0,63
Gesinnung	3,79	0,64
Gestirn	2,55	0,60
Gesuch	2,56	0,64

Lexem	Me	QuA/2
gesunden	2,95	0,70
Gesundheit	5,44	0,74
Getier	2,58	0,84
Getreide	4,44	0,78
Getriebe	3,55	0,93
Gevatter	2,07	0,28
gewaltig	4,78	0,80
Geweih	2,71	0,55
gewinnen	5,00	0,74
Giebel	2,86	0,83
gießen	4,28	1,04
Gießerei	3,18	0,79
Gift	3,80	0,94
giftig	4,65	0,73
Gigant	2,78	0,69
gigantisch	3,06	1,17
Gips	3,75	0,67
Giraffe	3,06	0,67
Glanz	3,56	0,74
Glas	5,82	0,32
Glaserei	2,56	0,68
glatt	4,85	0,64
Glatteis	2,72	0,58
gleich	6,00	0,62
Gleichgültigkeit	4,36	0,65
Gleichnis	2,91	0,39
Gleichung	5,33	1,33
gleiten	3,88	0,62
Gliederung	4,08	1,16
Gliederung	4,75	0,91
glimmen	3,45	0,72
glitzern	3,67	0,68
gluckern	2,65	0,62
glucksen	2,64	0,62
Glut	3,86	0,85
Gnu	2,03	0,37
Golf	2,75	0,39
gönnen	4,29	0,67
Grab	3,71	0,75
graben	4,19	0,83
grafisch	3,63	1,16
Granit	3,25	0,70
Graphiker	3,00	0,62

Lexem	Me	QuA/2
Gratulant	3,18	0,67
gratulieren	4,58	0,91
grau	4,75	0,52
gravieren	2,67	0,69
Greisin	2,71	0,79
grimmig	3,22	0,80
Grippe	4,69	0,72
Grube	3,71	0,79
Grund	5,75	0,67
Gruppe	5,89	0,47
gruselig	3,32	0,62
Gruß	5,64	0,65
Gulasch	4,65	0,78
Gummi	4,44	0,78
gummieren	2,00	0,48
Gurgel	3,04	0,45
Gusseisen	2,46	0,74
gut	6,84	0,33
Haar	5,78	0,80
haben	6,93	0,28
habil	1,82	0,50
Habsucht	3,22	0,69
Hain	2,27	0,50
halbieren	4,25	0,81
halbseiden	1,85	0,52
Halunke	3,35	0,78
Hamsterei	3,44	1,07
Hand	5,88	0,59
handelsüblich	3,61	0,75
Harpune	2,46	0,64
hart	5,94	0,96
hauen	4,71	0,68
Hausierer	2,32	0,49
hegen	2,79	0,56
Heide	3,05	0,70
heikel	3,92	1,07
heilsam	3,61	0,54
heiraten	4,55	0,85
heiraten	5,00	0,89
Hieb	3,25	0,62
hierbleiben	5,19	0,83
hinderlich	3,33	0,66
Hindernis	4,55	0,66

Lexem	Me	QuA/2
hindurch	4,43	0,96
hinweisen	5,04	0,56
Hirn	3,69	0,75
hissen	3,05	0,70
Hornisse	2,27	0,60
Hubschrauber	3,73	1,12
Hudelei	3,75	1,15
Humanist	4,00	0,85
Hummel	3,27	0,67
Humorist	2,96	0,55
Humus	2,82	0,60
Hund	5,33	0,65
Ideal	4,85	0,66
illegal	4,43	1,03
Imbiss	4,19	0,94
Impuls	4,40	1,12
Industrie	5,65	0,65
Inhalt	5,88	0,37
inhuman	3,55	0,95
Inland	3,71	0,83
Inspirator	1,91	0,39
Institut	5,44	0,68
international	5,77	0,55
internieren	2,23	0,57
intim	3,75	0,78
Intoleranz	3,57	0,80
Intrige	3,19	0,82
Jahreszeit	4,25	0,94
Jahrhundert	4,92	0,88
jammern	3,75	0,87
Junge	5,69	0,65
Jupiter	2,56	0,66
Jurist	3,64	0,69
Kabale	1,95	0,33
Kadaver	2,38	0,54
kahl	4,31	0,88
Kaliber	3,09	0,77
Kalkulator	2,38	0,66
Kampf	5,25	0,68
Kanal	3,78	0,70
Kandidat	4,85	0,84
Kaninchen	4,00	0,90
Kapital	5,00	1,14
karg	2,91	0,39
Käse	5,73	0,67
Katapult	2,67	0,65
Kathedrale	3,31	0,82
Katholik	3,04	0,54
katzbuckeln	2,67	0,65
Keil	3,42	0,91
keimen	3,43	0,80
Kellerei	2,18	0,61
Kiefer	3,88	0,91
Kiesel	2,80	0,89
Kinderlied	4,00	0,86
Kindheit	4,77	0,82
Kino	5,79	0,68
kitschig	4,06	0,92
kitzlig	4,29	0,89
klagen	3,73	0,67
klassisch	4,38	0,61
Klausur	3,00	0,87
Klavier	3,75	0,91
Kleinigkeit	5,27	0,82
Klima	4,94	0,90
Klinik	4,38	0,66
Klub	5,04	0,58
klumpig	3,00	0,48
Knabe	3,23	0,72
knallen	3,94	0,80
kneifen	3,38	0,68
knicken	3,45	0,66
Knie	5,06	1,02
knirschen	3,56	0,90
knurren	3,81	0,88
Kollektiv	5,67	0,91
kommunal	3,75	0,71
Kommunist	5,77	0,72
Konditorei	3,78	0,72
konkurrieren	4,56	1,02
konservativ	4,40	1,01
kraus	2,62	0,56
kreuzen	3,44	0,94
Krone	3,33	0,77
Krug	3,82	0,62
krümmen	3,65	0,65

Lexem	Me	QuA/2
Kubismus	1,73	0,59
Kuli	5,77	0,60
Kultur	5,36	0,65
Lagune	2,12	0,31
lasieren	1,46	0,50
lateinisch	3,55	0,64
Leidenschaft	4,09	0,63
liberal	4,06	0,85
Lichtung	3,45	0,75
lieblich	3,56	0,64
Limit	3,25	0,80
linieren	2,27	0,77
linnig	1,07	0,28
Literat	2,69	0,72
lukullisch	2,45	0,63
lustig	5,82	0,64
mahnen	3,73	0,78
Mahnmal	3,73	0,71
Mais	3,22	0,73
Malerei	4,25	0,75
Mandarine	3,85	0,72
mannigfaltig	3,45	0,70
maritim	2,56	0,66
Marxist	5,55	0,72
Meile	3,69	0,75
Milch	6,06	0,42
mildern	3,21	0,60
Mineral	3,09	0,66
Minimum	4,75	1,12
Mittag	5,73	0,71
munter	5,05	0,70
Musikant	3,95	0,72
Muskel	4,18	0,72
mutig	4,45	0,82
nachahmen	3,89	0,80
nachweisen	4,96	0,57
nackt	4,12	0,57
nahrhaft	3,88	0,60
Offizier	4,71	0,63
organisch	4,00	0,92
Palast	4,56	0,83
panieren	3,09	0,61
Partei	6,07	0,42

Lexem	Me	QuA/2
peinigen	2,91	0,62
Pfannkuchen	3,82	0,72
Pfund	4,88	0,80
Phantasie	4,73	0,61
Pirat	2,82	0,49
Planung	5,35	0,62
plausibel	3,55	0,62
platzieren	3,54	0,67
plump	3,57	0,80
Politik	5,14	0,64
politisch	6,06	0,80
Polizei	4,67	0,73
praktisch	5,82	0,51
preinen	1,00	0,25
preisen	3,00	0,48
Prise	3,31	0,81
privat	5,22	0,75
probieren	5,62	0,62
Produkt	4,82	0,67
Profil	4,09	0,61
Programm	5,80	0,53
Pudel	3,25	0,66
Pult	2,89	0,76
putzen	5,14	0,87
Rache	3,56	0,76
Radierung	2,72	0,58
Radio	5,77	0,72
Rand	5,32	1,07
rascheln	3,63	0,92
Rasur	3,33	0,91
ratsam	4,09	0,62
Regal	4,58	0,96
Register	3,82	0,72
reiben	4,11	0,91
reich	4,85	0,59
reichen	4,75	1,00
Reichtum	4,25	0,94
reinigen	4,19	0,89
reinigen	4,58	0,91
reinlich	3,20	0,84
reisen	5,12	0,59
Reisig	2,46	0,67
reiten	3,44	0,66

Lexem	Me	QuA/2
Rennerei	4,80	0,98
Republik	5,87	0,49
Resultat	5,71	0,63
rieseln	3,44	0,74
Ring	4,65	0,62
rudern	3,55	0,66
Sack	4,15	0,74
salben	2,46	0,57
salzig	5,20	1,08
Samstag	5,08	1,17
Sandale	4,22	0,71
Schall	3,23	0,40
Schalmei	2,24	0,55
Scheidung	4,42	1,05
schematisch	4,14	1,05
Schicksal	4,33	0,73
schieben	4,25	0,95
Schiedsrichter	4,09	0,61
schief	5,43	0,77
schießen	4,78	0,75
Schiff	4,65	1,10
Schifffahrt	3,36	0,65
Schimpanse	3,19	0,79
schinden	3,25	0,71
schlaff	3,12	0,66
Schleier	3,15	0,66
schlicht	3,44	0,78
schlimm	5,56	0,49
schlitzäugig	2,81	0,47
Schluck	5,11	0,87
schlucken	4,65	0,68
Schlupfwinkel	3,15	0,64
schlurfen	3,44	0,73
schmal	4,45	0,66
Schmeichelei	3,36	0,75
schmeichlerisch	3,28	0,79
schmierig	3,46	0,60
Schmuckkasten	3,00	0,85
schmunzeln	3,81	0,78
schneiden	3,63	0,90
schneiden	5,43	0,84
Schnitt	4,31	0,88
schrill	2,85	0,64
schrumpfen	2,96	0,54
schuldig	4,25	0,65
schweigsam	4,19	0,81
Schwimmmeister	2,85	0,52
Schwimmerin	3,75	0,71
schwindlig	3,67	0,70
Schwindsucht	2,39	0,57
Schwur	3,38	0,62
siegen	5,00	0,64
Signal	4,00	0,70
silbern	3,33	0,69
simultan	2,56	0,71
Sinfonie	4,18	0,61
Sinn	5,44	0,61
sinnwidrig	2,56	0,66
Sintflut	2,56	0,61
sittsam	2,56	0,61
Sitz	4,65	0,65
sitzen	6,00	0,42
Sitzung	5,61	0,54
Skala	3,44	0,79
Skalp	2,19	0,39
skalpieren	2,39	0,57
Skandal	3,69	0,82
Sklaverei	3,33	0,82
Sklavin	2,95	0,70
Skrupel	3,18	0,67
Skulptur	3,43	0,80
Soldat	4,85	0,72
Solist	3,94	0,97
Span	3,09	0,71
spannen	4,00	0,89
sparsam	4,73	0,62
spazieren	5,09	0,39
speichern	4,22	0,75
speisen	4,00	1,13
spektral	2,18	0,53
Spekulant	3,15	0,59
Spelunke	3,00	0,88
spendieren	4,44	0,85
spicken	3,85	0,72
Spiegel	5,27	0,71
Spiel	5,25	0,86

Lexem	Me	QuA/2
spielen	5,61	0,69
Spinat	3,29	0,85
spionieren	3,88	0,57
Spirale	3,82	0,67
spitzfindig	3,82	0,67
splittern	2,59	0,34
sporadisch	2,18	0,64
Sprache	5,90	0,42
spreizen	3,00	0,63
springen	4,67	0,69
spritzig	3,45	0,59
Spruch	4,09	0,61
sprudeln	3,27	0,66
Spurt	4,00	0,75
spurten	3,46	0,76
Sputnik	3,75	0,91
Staatsrat	4,21	0,75
Staffelei	2,33	0,56
standhalten	4,11	0,92
stattfinden	5,75	0,52
Statur	3,18	0,64
steifen	1,93	0,42
steil	4,44	0,83
steinern	2,85	0,69
steril	3,12	0,59
Stickerin	2,32	0,52
stickig	3,79	0,58
Stil	4,14	0,86
Stimulus	3,05	0,71
stinken	4,44	1,02
streicheln	4,56	0,95
streichen	3,56	0,74
Streik	4,75	1,08
stricken	4,29	0,92
Strudel	2,58	0,83
Struktur	4,85	0,76
Strumpf	5,54	0,67
studieren	6,33	0,65
stumm	3,96	0,54
Tagesanbruch	2,94	0,87
Tal	4,00	0,42
tanken	4,14	0,94
Tanz	4,88	0,60

Lexem	Me	QuA/2
Teich	4,06	0,91
teilnehmen	5,85	0,54
Teilung	4,31	1,04
Temperatur	5,00	0,64
textil	3,56	0,78
Thematik	4,71	0,82
ticken	3,85	0,95
tief	5,54	0,62
Tier	5,62	0,72
Treiber	1,55	0,59
Tribunal	2,73	0,63
Trumpe	1,02	0,26
Tumult	3,45	0,66
turnen	4,06	0,81
Tusche	3,44	0,79
Uhr	6,36	0,50
Umgang	4,29	0,63
umkreisen	3,73	0,78
umstimmen	3,82	0,64
Umzug	3,64	0,83
undicht	3,82	0,64
unliebsam	3,56	0,74
Urkunde	4,09	0,63
Urteil	4,45	0,75
wach	5,27	0,86
wachrufen	2,67	0,69
Wachs	3,56	0,83
wandern	4,73	0,61
Wandzeitung	4,57	0,77
weichen	3,21	0,60
weichherzig	2,65	0,62
weigern	3,85	0,64
Wein	4,40	0,59
Weisheit	3,55	0,57
wiegen	4,29	0,83
wild	4,18	0,60
Wucherer	2,55	0,58
wurmstichig	2,54	0,52
Wut	4,82	0,64
zagen	2,27	0,50
Zank	3,05	0,77
zart	4,56	0,78
zeichnen	4,86	0,89

Lexem	Me	QuA/2
Zeit	6,67	0,53
zersplittern	3,79	0,60
ziehen	5,04	0,55
zivil	3,64	0,69
Zucker	5,85	0,76
Zufuhr	3,31	0,82

Tabelle 3. Frequenzliste 2, sortiert nach den Median-Werten

Nr.	Stimulus	Me	QuA/2
1	gut	6,88	0,31
2	können	6,81	0,40
3	wollen	6,72	0,58
4	geben	6,67	0,60
5	werden	6,67	0,60
6	mein	6,67	0,67
7	dein	6,61	0,62
8	alles	6,61	0,67
9	fragen	6,61	0,75
10	arbeiten	6,44	0,61
11	denken	6,44	0,61
12	mehr	6,44	0,66
13	verstehen	6,43	0,72
14	neu	6,40	0,91
15	Frieden	6,29	0,84
16	waschen	6,22	0,67
17	Kuli	6,17	0,61
18	Lehrer	6,15	0,64
19	schlafen	6,13	0,45
20	Uhr	6,11	0,84
21	Straße	6,09	0,60
22	halten	6,07	1,00
23	groß	6,05	0,67
24	wichtig	6,04	0,54
25	Licht	6,00	0,63
26	modern	6,00	0,82
27	Zimmer	6,00	0,88
28	lang	5,94	0,78
29	sich treffen	5,94	0,84
30	Teller	5,91	0,61
31	runter	5,90	1,31
32	Tee	5,89	0,49
33	drei	5,85	0,56
34	leben	5,82	0,67
35	Seite	5,82	0,67
36	interessant	5,80	0,99
37	zwei	5,75	0,75
38	Luft	5,75	1,03
39	Stuhl	5,69	0,76
40	antworten	5,69	0,88
41	sowjetisch	5,62	0,72
42	Erde	5,62	0,90
43	fünfzig	5,57	0,96
44	vier	5,56	0,90
45	sich interessieren	5,55	0,66
46	Welt	5,55	0,75
47	jung	5,54	0,79
48	anfangen	5,50	0,94
49	weggehen	5,44	0,79
50	gesund	5,43	0,89
51	beginnen	5,40	0,98
52	Hand	5,31	0,88
53	Kopf	5,22	0,75
54	Handtuch	5,19	0,82
55	Zucker	5,12	0,62
56	nächst	5,12	1,21
57	hundert	5,10	0,83
58	amerikanisch	5,00	0,63
59	Garten	5,00	0,73
60	rausgehen	5,00	1,23
61	Durcheinander	4,92	0,98
62	sich beschäftigen	4,92	1,04
63	rufen	4,86	1,02
64	zwanzig	4,86	1,05
65	Schere	4,85	0,59
66	national	4,85	0,83
67	neun	4,80	1,07
68	neunzehn	4,80	1,34
69	Bedingung	4,79	0,64
70	verändern	4,78	0,80
71	braun	4,71	1,05
72	schwach	4,67	0,83
73	elf	4,60	0,96
74	Schirm	4,58	0,87
75	spazieren gehen	4,56	0,78

Nr.	Stimulus	Me	QuA/2
76	Motorrad	4,55	0,79
77	dreißig	4,45	0,75
78	Nase	4,44	0,78
79	Million	4,43	0,96
80	fünfzig	4,43	1,20
81	Energie	4,42	1,04
82	eigen	4,42	1,19
83	gegenwärtig	4,37	1,20
84	Rücken	4,36	0,75
85	Knopf	4,22	0,75
86	folgen	4,22	0,80
87	schlagen	4,20	1,11
88	warnen	4,18	0,64
89	siebzehn	4,14	0,99
90	achtzig	4,11	1,05
91	schauen	4,08	1,04
92	Nebel	4,06	0,82
93	Tuch	4,06	0,92
94	einhundert	4,00	0,92
95	Kugelschreiber	4,00	1,21
96	künftig	3,89	0,92
97	nicken	3,81	0,87
98	Fotoapparat	3,81	0,94
99	Pionier	3,75	0,94
100	Affe	3,69	1,05
101	Denkmal	3,67	0,69
102	Kirsche	3,58	0,91
103	Aufnahme	3,58	0,92
104	Maus	3,58	0,95
105	Kombinat	3,58	1,04
106	Entwurf	3,43	0,71
107	Vorhang	3,36	0,75
108	Dicke (Stärke)	3,31	0,81
109	Pilz	3,31	0,82
110	hinausgehen	3,31	1,16
111	anblicken	3,29	1,01
112	Stütze	3,25	0,86
113	Kinn	3,25	1,08
114	(Park-)Bank	3,22	0,86
115	Oberfläche	3,21	0,59
116	blasen	3,19	0,82
117	hinunter	3,15	0,76
118	Käfer	3,11	0,71
119	Wirrwarr	3,09	1,18
120	Schal	3,08	0,58
121	Ausruf	3,04	0,45
122	Spaten	2,91	0,61
123	zustellen	2,71	0,87
124	Deckung	2,67	0,65
125	Kröte	2,58	0,91
126	Flieger	2,55	0,58
127	Kahn	2,55	0,60
128	Mühle	2,54	0,55
129	Getöse	2,45	0,70
130	Hamsterei	2,45	1,06
131	Känguru	2,33	0,53
132	Schildkröte	2,32	0,63
133	Eidechse	2,29	0,68
134	Walross	2,25	0,55
135	Kostenanschlag	2,18	0,62
136	Vergänglichkeit	2,07	0,42
137	Landzunge	2,03	0,39
138	Stab (milit.)	1,91	0,39
139	Fischfangsaison	1,84	0,52
140	Schirmwand	1,46	0,52
141	Buchweizen	1,33	0,53
142	Trumpe	1,02	0,26
143	preinen	1,00	0,25
144	linnig	1,00	0,25

Tabelle 4. Frequenzliste 2, alphabetisch sortiert

Stimulus	Me	QuA/2
achtzig	4,11	1,05
Affe	3,69	1,05
alles	6,61	0,67
amerikanisch	5,00	0,63
anblicken	3,29	1,01
anfangen	5,50	0,94
antworten	5,69	0,88
arbeiten	6,44	0,61
Aufnahme	3,58	0,92
Ausruf	3,04	0,45
(Park-)Bank	3,22	0,86
Bedingung	4,79	0,64

Stimulus	Me	QuA/2
beginnen	5,40	0,98
blasen	3,19	0,82
braun	4,71	1,05
Buchweizen	1,33	0,53
Deckung	2,67	0,65
dein	6,61	0,62
denken	6,44	0,61
Denkmal	3,67	0,69
Dicke (Stärke)	3,31	0,81
drei	5,85	0,56
dreißig	4,45	0,75
Durcheinander	4,92	0,98
Eidechse	2,29	0,68
eigen	4,42	1,19
einhundert	4,00	0,92
elf	4,60	0,96
Energie	4,42	1,04
Entwurf	3,43	0,71
Erde	5,62	0,90
Fischfangsaison	1,84	0,52
Flieger	2,55	0,58
folgen	4,22	0,80
Fotoapparat	3,81	0,94
fragen	6,61	0,75
Frieden	6,29	0,84
fünfzig	4,43	1,20
fünfzig	5,57	0,96
Garten	5,00	0,73
geben	6,67	0,60
gegenwärtig	4,37	1,20
gesund	5,43	0,89
Getöse	2,45	0,70
groß	6,05	0,67
gut	6,88	0,31
halten	6,07	1,00
Hamsterei	2,45	1,06
Hand	5,31	0,88
Handtuch	5,19	0,82
hinausgehen	3,31	1,16
hinunter	3,15	0,76
hundert	5,10	0,83
interessant	5,80	0,99
jung	5,54	0,79

Stimulus	Me	QuA/2
Käfer	3,11	0,71
Kahn	2,55	0,60
Känguru	2,33	0,53
Kinn	3,25	1,08
Kirsche	3,58	0,91
Knopf	4,22	0,75
Kombinat	3,58	1,04
können	6,81	0,40
Kopf	5,22	0,75
Kostenanschlag	2,18	0,62
Kröte	2,58	0,91
Kugelschreiber	4,00	1,21
Kuli	6,17	0,61
künftig	3,89	0,92
Landzunge	2,03	0,39
lang	5,94	0,78
leben	5,82	0,67
Lehrer	6,15	0,64
Licht	6,00	0,63
linnig	1,00	0,25
Luft	5,75	1,03
Maus	3,58	0,95
mehr	6,44	0,66
mein	6,67	0,67
Million	4,43	0,96
modern	6,00	0,82
Motorrad	4,55	0,79
Mühle	2,54	0,55
nächst	5,12	1,21
Nase	4,44	0,78
national	4,85	0,83
Nebel	4,06	0,82
neu	6,40	0,91
neun	4,80	1,07
neunzehn	4,80	1,34
nicken	3,81	0,87
Oberfläche	3,21	0,59
Pilz	3,31	0,82
Pionier	3,75	0,94
preinen	1,00	0,25
rausgehen	5,00	1,23
Rücken	4,36	0,75
rufen	4,86	1,02

Stimulus	Me	QuA/2
runter	5,90	1,31
Schal	3,08	0,58
schauen	4,08	1,04
Schere	4,85	0,59
Schildkröte	2,32	0,63
Schirm	4,58	0,87
Schirmwand	1,46	0,52
schlafen	6,13	0,45
schlagen	4,20	1,11
schwach	4,67	0,83
Seite	5,82	0,67
sich beschäftigen	4,92	1,04
sich interessieren	5,55	0,66
sich treffen	5,94	0,84
siebzehn	4,14	0,99
sowjetisch	5,62	0,72
Spaten	2,91	0,61
spazieren gehen	4,56	0,78
Welt	5,55	0,75
werden	6,67	0,60
wichtig	6,04	0,54
Wirrwarr	3,09	1,18
wollen	6,72	0,58
Zimmer	6,00	0,88
Zucker	5,12	0,62
zustellen	2,71	0,87
zwanzig	4,86	1,05
zwei	5,75	0,75

Stimulus	Me	QuA/2
Stab (milit.)	1,91	0,39
Straße	6,09	0,60
Stuhl	5,69	0,76
Stütze	3,25	0,86
Tee	5,89	0,49
Teller	5,91	0,61
Trumpe	1,02	0,26
Tuch	4,06	0,92
Uhr	6,11	0,84
verändern	4,78	0,80
Vergänglichkeit	2,07	0,42
verstehen	6,43	0,72
vier	5,56	0,90
Vorhang	3,36	0,75
Walross	2,25	0,55
warnen	4,18	0,64
waschen	6,22	0,67
weggehen	5,44	0,79

Glottometrics 2, 2002, 82-86

Der Zuwachs der Wörter auf *–ion* im Deutschen

Helle Körner, Göttingen[1]

Abstract. This paper shows that the number of German words having the ending *–ion* increases in accordance with the Piotrowski law.

Key words: Piotrowski law, language change

1. Einleitung

Ein wichtiger Bereich der Sprachwissenschaft beschäftigt sich mit dem Sprachwandel, also u.a. mit Fragen dazu, warum und wie sich Sprache verändert. „Sprachwandel" bezeichnet üblicherweise den Prozeß der Veränderung von Sprachelementen und Sprachsystemen in der Zeit. Die grundlegende Hypothese zum Verlauf des Sprachwandels beruht auf der Annahme, daß ein Mitglied einer Sprachgemeinschaft eine Neuerung verwendet, die von anderen Personen übernommen wird. Dieser Prozeß beginnt relativ langsam und breitet sich dann mit „wachsender Zahl der Kontaktpersonen" schneller aus, bis er „entweder einen tolerierten Sättigungsgrad erreicht oder sich völlig durchgesetzt hat" (Best 2001a: 102).

Ausgehend von dieser These lassen sich drei unterschiedliche Formen des Sprachwandels unterscheiden (Altmann 1983: 60-62):

1. vollständiger Sprachwandel, bei dem alte Formen vollständig aus der Sprache verdrängt und durch die neuen Formen ersetzt werden (z.B. *was* zu *war*),
2. unvollständiger Sprachwandel, bei dem sich neue Formen oder Wörter nur in einem begrenzten Maß durchsetzen (z.B. Fremdwörter) und
3. reversibler Sprachwandel, bei dem neue Formen oder Wörter aufkommen, sich ausbreiten und dann wieder verschwinden (z.B. bei der e-Epithese deutscher Verben).

Beispiele für diese Typen von Sprachwandel findet man u.a. in Best & Kohlhase (Hrsg. 1983) und Best (2002b).

Aus diesen Thesen entwickelte Altmann das verallgemeinerte Piotrowski-Gesetz: „Unter dem Piotrowski-Gesetz verstehen wir die hypothetische Aussage über den zeitlichen Verlauf einer beliebigen sprachlichen Entität" (Altmann, 1983: 59). Andere Autoren, die nur die ersten zwei Veränderungsarten berücksichtigten, haben bereits qualitative Vermutungen über einen „S-förmigen" zeitlichen Verlauf sprachlicher Veränderungen entwickelt (z.B. Osgood & Sebeok 1965: 155).

Für den unvollständigen Sprachwandel hat Altmann (1983) die Funktion

[1] Address correspondence to: Helle Körner, Lange-Geismar-Str. 51, 37073 Göttingen

$$(1) \quad p(t) = \frac{c}{1 + ae^{-kt}}$$

abgeleitet, wobei *c* die Asymptote darstellt.

Im Rahmen dieser Arbeit soll eine weitere Überprüfung des Piotrowski-Gesetzes stattfinden. Als Beispiel dafür wurden die Fremdwörter, die auf das Suffix *–ion* enden, ausgewählt. Bei der Übernahme bzw. Bildung der Wörter auf *–ion* handelt es sich ebenso wie bei anderen Fremdwörtern (Best 2001b) um einen Prozeß, der über Jahrhunderte hinweg stattgefunden hat. Anhand von (1) soll geprüft werden, ob auch dieser Sprachwandel den (für den unvollständigen Sprachwandel) typischen S-förmigen Verlauf einschlägt, wie ihn Best (2002a) bereits für das Suffix *–ical* sowie für Wörter auf *–ität* (Best 2001a: 107) und *-bar* (Best 2002b) als geeignetes Modell nachwies.

2. Methodik

Anhand dreier rückläufiger Wörterbücher (Theissen et al., 1992; Muthmann, 1988; Mater, 1965) wurde eine Liste der Wörter, die auf das Suffix *–ion* enden, zusammengestellt. Diese Wörterbücher wurden wegen ihres unterschiedlichen Erscheinungsdatums ausgewählt, so daß auch Wörter berücksichtigt werden konnten, die mittlerweile so vielleicht gar nicht mehr in der deutschen Sprache existieren (Bei Mater (1965) findet sich beispielsweise noch das Wort Akkord**ion**, das in den anderen benutzten Wörterbüchern bereits als Akkord**eon** lexikalisiert ist).

Im Gegensatz zu anderen Studien, die sich mit der Fremdwort-Thematik beschäftigen, konnten hier auch Wörter einbezogen werden, bei denen die Ursprungs- bzw. Gebersprache nicht eindeutig zugeordnet werden konnte. Berücksichtigt wurden dabei keine Determinativ-Komposita (wie etwa *Kultur-Region*) und keine Wörter, bei denen *Ion* als Lexem fungiert (wie *Wasserstoff-Ion*). Daraus ergab sich eine Liste mit 1566 Eintragungen. Von dieser Liste konnten jedoch nur 393 Wörter hinsichtlich ihrer Übernahme ins Deutsche ausreichend genau datiert werden. Für die Datierung wurden mehrere Wörterbücher herangezogen, um eine durch Übereinstimmung abgesicherte Basis zu gewinnen. Hauptsächlich beruht diese Arbeit auf den Angaben, die Pfeifer (2000) macht; zur weiteren Kontrolle wurden auch Carstensen (1996), Kluge (1995) und Drosdowski (1989) verwendet. Auf das Deutsche Fremdwörterbuch (DFWB) wurde weniger zurückgegriffen, da die anderen Wörterbücher es wohl zum größten Teil schon aufgrund der zeitlichen Differenz berücksichtigen und große Unterschiede bezüglich der Datierungen zwischen dem DFWB und den anderen drei Wörterbüchern bestehen. Bei Widersprüchen in der Datierung wurde die am häufigsten genannte Angabe berücksichtigt.

Um das Piotrowski-Gesetz testen zu können, ist es notwendig, daß ausreichende empirische Daten zur Verfügung stehen; dies ist leider nicht immer der Fall. Bei der Untersuchung der auf *–ion* endenden Wörter ist es gerade für den Zeitraum bis zum 14. Jahrhundert schwierig, überhaupt aussagekräftige Daten zu gewinnen. So lassen sich für das 9. Jahrhundert nur zwei Belege und für das 10. Jahrhundert gar kein Beleg finden.

Zum Verfahren der Auswertung muß an dieser Stelle noch gesagt werden, daß nur solche Entlehnungen berücksichtigt wurden, für die eine hinreichend genaue Datierung angegeben wurde, indem das Jahrhundert der Übernahme erkennbar war. Angaben wie „16./17. Jahrhundert“ wurden dabei dem erstgenannten Zeitraum zugeschlagen, Angaben wie „um 1700“ wiederum wurden etwa dem folgenden Jahrhundert zugerechnet; d.h. auch, daß undatierte Entlehnungen nicht berücksichtigt wurden (Vgl. Best 2001c: 8).

3. Ergebnis

Die Tabelle und die Abbildung zeigen, daß das Piotrowski-Gesetz sich bei der Entwicklung der Wörter auf *–ion* sehr gut bewährt. Dabei gibt *c* an, gegen welchen Wert der Sprachwandel strebt. *t* steht für die Zeit, und *a* und *k* sind Parameter.

Durch die Berechnung ergibt sich der sehr gute Determinationskoeffizient $D = 0.99$, was sich auch in der Graphik widerspiegelt (Die Abweichungen zwischen der berechneten Kurve und den als Punkte markierten gemessenen Daten sind sehr gering). Die eingangs aufgestellte Hypothese über den S-förmigen Verlauf dieses Sprachwandels konnte damit also bestätigt werden.

4. Fazit

Es wurden alle datierbaren *–ion*-Bildungen berücksichtigt; man kann wohl annehmen, daß sie den Trend, den die Zunahme der Wörter dieses Ableitungstyps im Deutschen aufweisen, wiedergeben.

An die Datei, die so gewonnen wurde, kann das Piotrowski-Gesetz in der unvollständigen Form mit sehr gutem Ergebnis angepaßt werden. Auch dieser Sprachwandel entspricht also den theoretischen Annahmen.

Der Parameter *c* steht für die obere Grenze dieser Entwicklung, so wie sie sich aus der berechneten Formel empirisch ergibt. Der tatsächliche Wert liegt entsprechend deutlich höher. Der Graphik ist zu entnehmen, daß dieser Sprachwandel voraussichtlich noch eine gewisse Entwicklung vor sich hat; er könnte sich aber bereits in seiner Endphase befinden.

Tabelle 1. Zuwachs der Wörter mit *-ion*

Jahrhundert	Zeit	Anzahl	Anzahl (kumuliert)	Anzahl (berechnet)
9	1	2	2	0.18
10	2	0	2	0.51
11	3	1	3	1.42
12	4	1	4	3.93
13	5	2	6	10.74
14	6	10	16	28.50
15	7	31	47	70.27
16	8	138	185	158.48
17	9	52	237	247.48
18	10	74	311	325.47
19	11	58	369	367.03
20	12	24	393	384.69
$a = 5975.07$	$k = 1.023$	$c = 395.36$	$D = 0.99$	

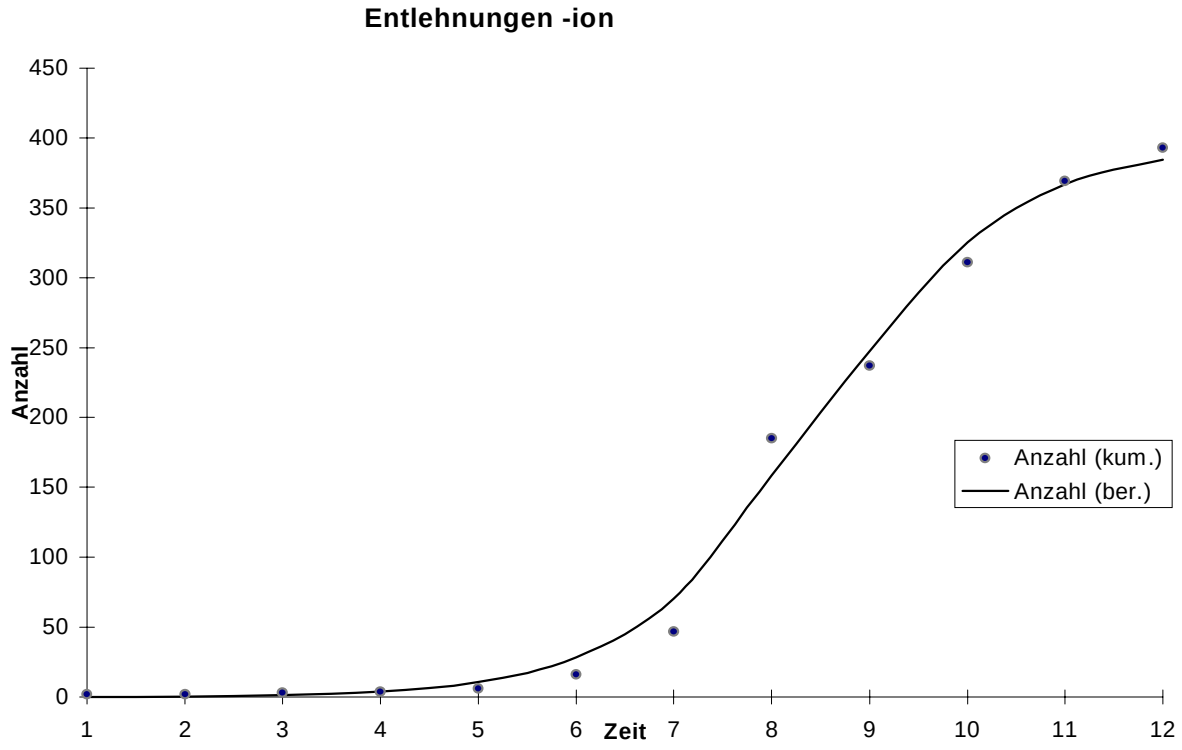

Abbildung 1. Graphische Darstellung der Daten aus Tabelle 1

Literatur

Altmann, Gabriel (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, Karl-Heinz, Kohlhase, Jörg (Hrsg.), *Exakte Sprachwandelforschung: 54-90.* Göttingen: edition herodot.

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

Best, Karl-Heinz (2001b). Ein Beitrag zur Fremdwortdiskussion. In: *Die deutsche Sprache in der Gegenwart. Festschrift f. Dieter Cherubim zum 60. Geburtstag: 263-270.* Hrsg. v. St. J. Schierholz in Zusammenarbeit mit E. Fobbe, St. Goes u. R. Knirsch. Frankfurt: Peter Lang Verlag.

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

Best, Karl-Heinz (2002a). Der Zuwachs der Wörter auf *–ical* im Deutschen. *Glottometrics 2, 11-16.*

Best, Karl-Heinz (2002b). Spracherwerb und Sprachwandel. Zur Reichweite des Piotrowski-Gesetzes. Ms.

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

Carstensen, Broder (1996). *Anglizismen-Wörterbuch: Der Einfluß des Englischen auf den deutschen Wortschatz nach 1945*. Berlin: de Gruyter.

Drosdowski, Günther (1989). *Duden. Etymologie: Herkunftswörterbuch der deutschen Sprache.* Mannheim: Duden-Verlag.

Kluge, Friedrich (1995). *Etymologisches Wörterbuch der deutschen Sprache.* 23. Auflage. Berlin: de Gruyter.

Mater, Erich (1965). *Rückläufiges Wörterbuch der deutschen Gegenwartssprache.* Leipzig: VEB Verlag Enzyklopädie.

Muthmann, Gustav (1988). *Rückläufiges deutsches Wörterbuch. Handbuch der Wortausgänge im Deutschen mit Beachtung der Wort- und Lautstruktur.* Tübingen: Niemeyer.

Osgood, Charles I., & Sebeok, Thomas (ed.) (1954/1965). *Psycholinguistics.* Bloomington, Indiana UP.

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

Theissen, S., Alexis, R., Kefer, M., Tewilt, G.-T. (1992). *Rückläufiges Wörterbuch des Deutschen.* Liège: C.I.P.L.

Glottometrics 2, 2002,87-94

Syllable Lengths in Russian, Bulgarian, Old Church Slavonic and Slovene

Otto A. Rottmann, Bochum/Hagen[1]

Abstract. This paper presents empirical syllable length data (with "syllable length" being based on the number of phonemes) from Slovene, modern Bulgarian, Old Church Slavonic (= Old Bulgarian) and modern Russian. The data analysed give evidence of the following adequate models: Hyperpoisson in Old Church Slavonic, modern Bulgarian and Slovene and Conway-Maxwell-Poisson or Morse in modern Russian.

Key words: syllable length; distributions: Hyperpoisson, Conway-Maxwell-Poisson, Morse; Russian, Bulgarian, Old Church Slavonic (= Old Bulgarian), Slovene

1. General

The present paper is meant as a first step towards a comprehensive analysis of syllable lengths in the four languages mentioned in the heading. The justification for this analysis is the evident fact that the structure of syllables is part of the nomological system of language. On the other hand, and this fact cannot be neglected, the definition of what a syllable is like is widely discussed in literature. We consider a "syllable" a chain of sounds centering around a vowel, i.e. the sound in the chain with maximum sound energy. However, the four Slavic languages analysed differ as to their syllable structures (= composition of vowels and consonants), and so isolating one syllable from another may be difficult and in cases of doubt require support by native speakers. At least, this applies to the modern languages Bulgarian, Slovene and Russian. Fortunately, this does not apply to the dead language Old Church Slavonic, because it had a fixed structure: all syllables were open, i.e. the last sound in a syllable was a vowel, which as a maximum was preceded by five consonants (five only if the fifth was /j/ directly preceding the vowel).

The theoretical basis for the present analysis is found in the ideas of synergetic linguistics, which, however, will not be detailed in this paper, but in a forthcoming monography on morphological and syntactical structures in Slavic languages; the present paper presents results on the one hand and a slight further development of the results in the monography on the other.

The syllable lengths analysed were determined by the number of phonemes. As different opinions can be found in literature as to the set of phonemes in a specific language, the works listed under "References" were taken as the basis for the sets of phonemes in the languages analysed. All texts were chosen at random.

The following texts were used for the examination (see Table 1):

[1] Address correspondence to: Otto Rottmann, Seminar für Slawistik, Universität GB3/8, Universitätstr. 150, D-44780 Bochum. E-mail: otto.rottmann@t-online.de

Table 1. Texts and their sources

Bulgarian:

No.	Text	Source of the text
1	Voennite spеclužbi pod partijna vlast	Kontinent 7-7-97
2	SŠA ni praščat doklad za skandala s “Armimeks”	Kontinent 8-7-97
3	Nezakonnite pari da vljazat v sociali fondove	Kontinent 6-7-97
4	Rakat: kade se namirame naistina	Kontinent 20-4-97
5	Zaščo se buntuva Albanija	Kontinent 19-4-97
6	Sture Alen: Pasim v tajna nominiranite nobelisti	Kontinent 17-4-97
7	Levicata triumfira sas silata na razuma	Kontinent 16-4-97
8	Da se sbližim platno s Južna Evropa	Kontinent 1-7-97
9	Na Balkanite ima tvarde mnogo istorii	Kontinent 14-7-97
10	Zaščo ne izlizaše igrata	Kontinent 4-7-97

Russian:

BG no.	Text	Source of the text
1	Tolsтoj L.N., Kavkazskij plennik, Gl. 1	Bradda Books. London 1962
2	Kaša iz topora	In: Kniga dlja čtenija po russkomy jazyku. Moskva 1970
3	Sud	see 2
4	Sem’ granatovych prut’ev	see 2
5	Mudrost’	see 2
6	Kakaja žena nužna	see 2
7	Delež gusja	see 2
8	Čudesnyj klad	see 2
9	Lentjajka	see 2
10	Verbljud i osel	see 2

Old Church Slavonic:

AKS no.	Text	Source of the text
1	Luke XIII	Codex Zographensis, acc. to Leskien, text in adapted form, i.e. after the dissolution of the language-specific abbreviations
2	Luke XII	see 1
3	Luke XI	see 1
4	Luke X	see 1
5	Luke V	see 1
6	Luke VI	see 1
7	Luke VII	see 1
8	Luke VIII	see 1
9	Luke IX	see 1
10	Kievskij Missal	see 1, excerpts acc. to Leskien, text in adapted form, i.e. after the dissolution of the language-specific abbreviations

Slovene:

Text no. SVE	Text	Source of the text
1	16. stoletje	Toporišič, Jože. Slovenski knjižni jezik. Bd. 2. 1966
2	17. stoletje	see 1

3	Se o 17. in 18. stoletje	see 1
4	Konec 18. in prva polovica 19. stoletja	see 1
5	Pohlin	see 1
6	Kopitar	see 1
7	Dnevnik	see 1
8	Pravopis	see 1
9	Raba	see 1
10	Vaje	see 1

Apart from the fact that phonemes were used as the basis for counting the following specific criteria were applied to the analysis:

(a) Numbers in the form of figures were counted as if being written in words at full length.
(b) Abbreviations were dissolved and the resulting words were examined as if not being abbreviated.
(c) The evaluation is based on $\alpha = 0.01$ as the threshold for acceptable to good results. In this paper we proceed in compliance with what was already discussed in Rottmann (forthcoming) on the decision in favour of the above limit value. In the case that there are no degrees of freedom for the chi-square test we use the contingency coefficient $C = X^2/N$

2. Results

2.1. Old Church Slavonic (= Old Bulgarian)

Syllable length in Old Church Slavonic is definitely controlled by the Hyperpoisson distribution. The relevant formula expressing the attractor is as follows (hereby the fact is taken into consideration that words consisting of 0 syllables do not exist, i.e. the distribution is 1-displaced):

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

with a and b being parameters and ${}_1F_1$ the hypergeometric function; the following quantities are specified:

x: syllable length, specified in phonemes
f_x: number of syllables having length x
NP_x: theoretical values according to the above formula
a, b: parameters
X_k^2 : chi-square with k degrees of freedom
P: probability of chi-square
C: contingency coefficient

Table 2. Fitting the Hyperpoisson d. to Old Church Slavonic data

	AKS 1		AKS 2		AKS 3	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	211	211.36	354	354.88	422	422.70
2	752	753.30	1644	1648.10	1563	1565.58
3	137	133.44	267	248.04┐	258	250.98

4	11	12.90	6	19.98	18	21.74
	a = 0.1864 b = 0.0523 X_1^2 = 0.38 P = 0.54		a = 0.1532 b = 0.8189 C = 0.00005		a = 0.1676 b = 0.0452 X_1^2 = 0.85 P = 0.36	

	AKS 4		AKS 5		AKS 6	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	362	362.30	315	314.64	415	413.82
2	1241	1242.02	1114	1118.38	1330	1338.39
3	190	187.44	191	177.72	271	252.56
4	14	15.24	6	15.26	15	26.23
	a = 0.1579 b = 0.0460 X_1^2 = 0.14 P = 0.71		a = 0.1663 b = 0.0468 C = 0.00006		a = 0.2003 b = 0.0620 X_1^2 = 6,21 P = 0.01	

	AKS 7		AKS 8		AKS 9	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	384	383.50	506	506.80	517	516.08
2	1339	1344.75	1591	1593.52	1709	1716.75
3	233	216.83	259	252.07	305	287.27
4	8	18.93	18	21.61	15	25.91
	a = 0.1690 b = 0.0482 C = 0.00007		a = 0.1666 b = 0.0530 X_1^2 = 0.80 P = 0.37		a = 0.1758 b = 0.0528 X_1^2 = 5.88 P = 0.02	

	AKS 10	
X	f_x	NP_x
1	133	132.74
2	435	437.88
3	94	86.86
4	5	9.53
	a = 0.2110 b = 0.0640 X_1^2 = 2.76 P = 0.10	

2.2. Modern Bulgarian

The same controlling mechanism for syllable length can be found in modern Bulgarian without exception. Obviously, there have not been any changes in the syllable controlling mechanism for over more than one thousand years despite changes in the phonological system like the modifications of the *jers* in weak and strong positions.

Table 3. Fitting the Hyperpoisson d. to Modern Bulgarian data

	BG 1		BG 2		BG 3	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	160	160.05	218	215.41	172	170.10
2	1083	1083.33	1361	1342.84	1164	1151.12
3	398	399.32	441	478.89	339	364.63
4	80	75.66	111	87.91	75	59.13
5	8	10.65	6	11.96	2	7.01
	a = 0.3898 b = 0.0576 X_2^2 = 0.91 P = 0.63		a = 0.3798 b = 0.0609 X_1^2 = 6.20 P = 0.01		a = 0.3322 b = 0.0491 X_1^2 = 3.75 P = 0.05	

	BG 4		BG 5		BG 6	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	383	378.30	533	533.14	357	357.51
2	2863	2827.84	3184	3184.84	2100	2069.62
3	879	952.13	1107	1107.44	650	700.89
4	202	163.61┐	216	198.31┐	155	122.26┐
5	16	20.53┘	10	26.26┘	4	15.73┘
	a = 0.3541 b = 0.0474 C = 0.0028		a = 0.3699 b = 0.0619 X_1^2 = 0.01 P = 0.94		a = 0.3593 b = 0.0621 C = 0.0022	

	BG 7		BG 8		BG 9	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	342	336.06	187	186.67	320	319.55
2	1738	1707.82	985	970.72	1530	1527.85
3	549	613.20┐	333	360.30	557	564.53
4	145	114.12┘	85	69.34	118	108.49
5	13	15.80	7	9.97	11	15.59
	a = 0.3864 b = 0.0760 X_2^2 = 2.80 P = 0.09		a = 0.3997 b = 0.0769 X_2^2 = 6.70 P = 0.09		a = 0.4004 b = 0.0838 X_2^2 = 2.29 P = 0.32	

	BG 10	
X	f_x	NP_x
1	191	192.79
2	1124	1134.52
3	442	419.32
4	76	80.00
5	5	11.37
	a = 0.3944 α = 0.0670 X_2^2 = 5.11 P = 0.08	

2.3. Slovene

The Slovene language knows the same controlling mechanism for syllable length, i.e. the Hyperpoisson distribution. The relevant data are as follows (see Table 4).

Table 4. Fitting the Hyperpoisson d. to Slovene data

	SVE 1		SVE 2		SVE 3	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	60	62.29	102	10181	64	65.12
2	256	265.78	957	963.58	550	559.67
3	154	125.48┐	357	345.81	227	203.31
4	21	37.45┘	61	63.25	29	37.72
5			3	7.76	1	5.18
6			3	0.77		
	a = 0.5529 b = 0.1296 C = 0.0009		a = 0.3730 b = 0.0394 X_2^2 = 1.24 P = 0.54		a = 0.3793 b = 0.0441 X_2^2 = 8.34 P = 0.02	

	SVE 4		SVE 5		SVE 6	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	61	61.82	70	70.10	117	113.90
2	489	495.59	421	421.59	656	638.60
3	183	167.15	154	153.48	125	157.00
4	22	28.79	30	28.81	31	19.73
5	2	3.64	3	4.02	2	1.78
	a = 0.3521 b = 0.0439		a = 0.3875 b = 0.0644		a = 0.2239 b = 0.0399	
	X_2^2 = 3.94 P = 0.14		X_2^2 = 0.31 P = 0.86		C = 0.0000	

	SVE 7		SVE 8		SVE 9	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	62	63.22	47	45.69	104	103.31
2	385	392.55	300	291.61	564	560.25
3	162	139.33	70	85.49	180	187.82
4	14	25.46	20	14.22	39	32.49
5	1	3.44			1	4.14
	a = 0.3847 b = 0.0619		a = 0.3073 b = 0.0481		a = 0.3573 b = 0.0659	
	C = 0.0004		X_1^2 = 5.43 P = 0.02		X_2^2 = 4.05 P = 0.13	

	SVE 10	
X	f_x	NP_x
1	119	119.84
2	854	860.01
3	337	322.28
4	62	62.00
5	1	8.87
	a = 0.3954 b = 0.0551	
	X_2^2 = 7.71 P = 0.02	

2.4. Modern Russian

Things, however, are different in modern Russian: as it is the case with word length (see Rottmann, forthcoming), where two applicable mechanisms were found (Extended Positive Binomial and Hyperpoisson with the latter occurring in almost as many cases as the former) the controlling mechanisms for syllable length seem to be the Morse and the Conway-Maxwell-Poisson distributions: the latter (*CMP* in the table) applies to the majority of cases; in two cases (RS 3 and RS 4), however, the Morse distribution (*Morse* in the table) yields better data, and in text RS 7 syllable length is only controlled by Morse.

The formula for the Conway-Maxwell-Poisson distribution is:

$$P_x = \frac{a^{x-1}}{((x-1)!)^b} P_1, \quad x = 1,2,3,... \tag{2}$$

whereas the formula for the Morse distribution is:

$$P_x = P_1 a^{x-1} b^{(x-1)(x-2)} \qquad x = 1,2,3,\ldots \tag{3}$$

This distribution can be obtained from the usual approach $P_x = g(x)P_{x-1}$ setting $g(x) = AB^x$, solving and substituting $A = ab^{-2}$ and $B = b^2$ in order to obtain the standard form.

Table 5. Fitting (2) and (3) to the data of Modern Bulgarian

	RS 1 - CMP		RS 2 - CMP		RS 3 - Morse	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	267	254.54	31	30.20	113	125.14
2	1185	1218.43	273	279.91	516	497.41
3	642	603.55	158	147.83	342	352.07
4	62	79.32	10	14.61	43	44.38
5	4	4.17	1	0.45	6	1.00
	a = 4.7868 b = 3.2725		a = 9.2674 b = 4.1331		a = 3.9748 b = 0.4220	
	X_2^2 = 7.76 P = 0.02		X_1^2 = 1.98 P = 0.16		X_1^2 = 2.45 P = 0.12	

	RS 4 - Morse		RS 5 - CMP		RS 6 - CMP	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	86	83.47	100	98.25	136	130.35
2	377	379.45	427	426.23	540	553.56
3	283	283.55	250	245.61	302	288.92
4	34	34.83	34	43.45	41	44.24
5	2	0.71	6	3.46	1	2.93
	a = 4.5461 b = 0.4054		a = 4.3383 b = 2.9124		a = 4.2467 b = 3.0244	
	X_1^2 = 0.10 P = 0.75		X_2^2 = 4.03 P = 0.14		X_2^2 = 2.68 P = 0.26	

	RS 7 – Morse		RS 8 - CMP		RS 9 - CMP	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	68	59.08	137	132.84	140	137.69
2	336	349.73	661	674.48	689	696.69
3	234	228.77	400	385.29	357	348.37
4	15	16.58	57	61.32	42	44.99
5	1	0.13	2	3.94	2	2.28
6			1	0.13		
	a = 5.9140 b = 0.3327		a = 5.0774 b = 3.1519		a = 5.0598 b = 3.3390	
	X_1^2 = 2.01 P = 0.16		X_2^2 = 1.55 P = 0.46		X_2^2 = 0.57 P = 0.75	

	RS 10 - CMP	
X	f_x	NP_x
1	43	40.71
2	184	189.62
3	127	121.21
4	24	26.47
	a = 4.6582 b = 2.8654	
	X_1^2 = 0.80 P = 0.37	

3. Conclusions

The length model outlined in Rottmann (forthcoming) for Slavic languages can be considered confirmed despite the fact that in each language analysed the sample only comprised ten texts. Nevertheless, such a statement is permitted thanks to the homogeneous character of the controlling mechanism found.

This also applies to modern Russian. Though two mechanisms (Extended Positive Binomial and Morse) turn out to be effective (which means this problem will have to be investigated in closer detail with a lot of further texts), both are within the "family of distributions" controlling lengths in Slavic languages. With respect to the Extended Positive Binomial distribution this fact was already stated in the above mentioned forthcoming publication.

References

Altmann, G. (1997). *Fitter. Software and handbook*. Lüdenscheid: RAM.

Balgarska akademija na nauikite (1982) (Hg.). *Gramatika na cavremennija balgarski knižoven ezik. Tom 1: Fonetika.* Sofija: Balgarskata Akademija Na Naukite

Leskien, A. (1962[8]). *Handbuch des Altbulgaris*chen. Heidelberg: Winter.

Rottmann, O.A. (forthcoming) Zu Form und Struktur von Wort und Satz in den slavischen Sprachen. Ein Beitrag zur quantitativ-typologischen Analyse.

Tauscher, E., Kirschbaum, E.-G. (1962). *Grammatik der russischen Sprache*. Berlin: Volk und Wissen.

Toporišič, J. (1978). *Glasovna in naglasna podoba slovenskega jezika*. Maribor: Založba Obzorja.

Book Reviews

Best, K.-H. (ed.), ***Häufigkeitsverteilungen in Texten*** (= Göttinger Linguistische Abhandlungen 4). Göttingen: Peust & Gutschmidt Verlag 2001. 310 pp. *By Simone Andersen*

"Everything abides by laws": Bunge's principle is obviously the motto of this book - it is cited repeatedly by the different contributors.

Looking for laws in linguistics means one has to face some difficulties. Linguistics as a science with an object of empirical character generates hypotheses that cannot be proven in the logical-mathematical way. Unfortunately, a hypothesis - even a good one - cannot be proven either by concordant observations. In fact, all you could do by observation is to falsify or to corroborate it. So, the best way of making a hypothesis a law is making it plausible by theory, i.e. to derive it from a theory, and corroborating it by empirical tests. Best's publication fulfils the second task for the law of lengths. Together with his co-researchers in his Göttingen project on Quantitative Linguistics, Best investigated frequency distributions of the lengths of linguistic units in texts.

The theory of word length distribution (Altmann 1988a, 1988b, 1991; Wimmer et al. 1994, Wimmer & Altmann, 1996) claims that word lengths - and probably any linguistic unit lengths, measured by the number of their direct constituents - in a text will occur in predictable proportions: e.g. the number of word lengths x in a text should be proportional to the number of word length *x*-1, with the proportions varying from one length class to the other. So the frequency of word length *x* can be considered as a function of the frequency of word length *x*-1: $P_x = g(x)P_{x-1}$

Depending on what is substituted for g(x) one gets a class of distributions mostly related to the Hyperpascal distribution of which the most important ones here seem to be the Poisson, the Hyperpoisson, the negative binomial and their modifications. Up to now about 50 different languages and more than 3000 texts have already been investigated by Best and his group with the data following at least one of these distributions. They tested the extension of the law by extreme conditions, concerning the variety of languages and sort of text, from Faroe letters to Chinese lexicon, and the wide range of units from morph and syllable lengths over word lengths to sen-tence lengths. They succeeded to fit

for word lengths:
- the Hyperpoisson distribution: 95 German letters of the last 500 years, 21 Faroe letters, 31 Gaelic texts, 34 of Luthers songs and tales, Dutch lexicon,
- the negative binomial distribution: 23 Chinese texts and 3 lexicons,
- the positive negative binomial distribution: 110 different types of Low German texts,
- the positive Cohen-Poisson distribution: 23 Chinese texts and 3 lexicons,
- the mixed Poisson distribution: 16 Rhineland palatine essays,
- the Conway-Maxwell-Poisson distribution: Dutch lexicon,
- the positive Singh-Poisson distribution: 20 Dutch letters,
- the Singh-Poisson or the Hirata-Poisson distribution: 24 Swiss German private letters.

Here the Conway-Maxwell-Poisson distribution belongs to a more general family (cf. Wimmer, Köhler, Altmann 2003) and the Hirata-Poisson is a generalized Poisson.

For syllable lengths:
- the Conway-Maxwell-Poisson distribution or the Hyperpoisson distribution: texts of German press.

For morph lengths:
- the Hyperpoisson distribution: German texts.

For sentence lengths:
- the Hyperpoisson distribution: 22 Chinese texts of different types, 22 Russian texts of different type,
- the negative binomial distribution: 25 German texts,
- the positive Poisson distribution: 80 German texts dating from the last 100 years.

The law of length can be said to have been well confirmed by repeated testing. And there is much evidence that it holds not only for word lengths but for any units. Best's studies can be considered to play the role of a test with a large power: They gave the law a considerable chance of being falsified.

Another important aspect of the book is that it provides a very fruitful base for further questions and future research; it makes a significant step towards a theory of texts. The last contribution is a study of the Zipf-Mandelbrot-law (by A. Knüppel), which can be understood as an outlook and a presentation of potentially important constructs, questions and proposals (closedness, balance of forces, role of the parameters etc.).

There are still several unsolved but important problems to be investigated: What determines the number and the values of the parameters? What do they mean? What are the forces deciding on their number: one (in the Poisson distribution), or two (as in the Hyper-poisson, the Conway-Maxwell-Poisson, the negative binomial etc. distributions) or even more (three in the mixed Poisson distribution for dialect essays) that have to be taken into account? What are the factors deciding on the choice of a simple, a mixed or a generalized distribution?

Another open question is the explanation of the universal validity of the law. Why do the unit lengths always follow the distributions found? Bunge's dictum that everything abides by laws leaves space for different interpretations which the different contributors seem to support in various ways. It could mean that speakers compose their texts unconsciously in accordance to the law - a concept of "following rules" that has a long tradition in linguistics. It could also mean a translation of the epistemic principle of "nihil sine causa" which holds in the natural sciences: "At random" and "according to statistical laws" has not necessarily to be a contradiction. Everything abides by law, even randomness. If speakers of different languages construct and plan texts for different purposes, and if the texts show unit lengths distributions that can be fitted by probability distributions, it could mean that unit length is something that hardly can be manipulated by the text producer. Under this interpretation, Altmann's law wouldn't look less attractive, perhaps even more revolutionary. The distribution could be due to universal text construction principles which reflect mental properties and communication laws, as well as due to universal principles of the evolution of language unit inventories which perhaps reflect information theoretical needs, categorization and psychological forces of memory (clustering, magical number 7 ± 2).

It seems to be necessary to investigate the degree of intention, the extent of free choice and the extent of constraints vs. changeability that affect a distribution of unit lengths in any text.

As a first step, one could think of looking at the "individual" frequencies and compare them with the class frequencies, in a kind of ANOVA (analysis of variance). For example, the

frequency of the monosyllabic words: is it the result of a few very frequently used words together with many words that are used only once in this text (which would yield a considerable variance)? Or does the class frequency result from a number of nearly equally used words (small variance)? If the variance in the classes is smaller than the variance between them, it could be a hint that a word is chosen "because of" its length, for what reasons ever.

One could try to vary the degree of intention as independent variable and watch the goodness-of-fit and the parameter as dependent variables. The degree of intention could be varied e.g. by differing time limits or the instruction of rewriting a text by making the words longer, or shorter.

Looking at the data it is striking that the means seem to be very similar for a given unit. Is it possible to construct a "super-distribution" for every unit showing the prototypical expected situation? We could perhaps suppose or estimate a "preference parameter" or "preferred length" parameter p (or μ) for any unit (μ near 1 for word lengths, for example) that would be determined by forces which are unchangeable for the individual text producer: universal properties of language and information. The degree of deviation from this super-distribution could signify the effects of other forces, like intention, purpose (text sort), special traits of the individual etc. so that bad fit would yield information, too.

When categorizing the forces influencing the unit lengths, it could be helpful to distinguish between influences that are manipulable by the text producer and those that are not, as well as between influences that are stable and those arising ad hoc in individual situations.

The book can be read as well as a textbook of Quantitative Linguistics: it shows the approach, the methods, the problems and the procedures exemplarily. Even beginners or specialists in one language will study it with profit, since the theoretical base is explained again by every contributor. One of the clearest explanations can be found in A.Ahlers contribution “The distribution of word length in different types of Low German texts” (43-58). A very clear and comprehensible description of the test procedure is in A. Knüppel’s contribution “Untersuchungen zum Zipf-Mandelbrot-Gesetz an deutschen Texten” (248-280).

A bibliographical note on S.G. Čebanov, the pioneer of word length research, written by K.-H. Best and Čebanov’s grandson, is a contribution to the history of Quantitative Linguistics. The annotated bibliography of the Göttingen project (284-310) shows the extent of this research; it is one of the greatest programs ever performed in quantitative linguistics.

References

Altmann, G. (1988a). Verteilungen der Satzlängen. *Glottometrika 9, 147-169*

Altmann, G. (1988b). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Altmann, G. (1991). Modelling diversification phenomena in language. In U. Rothe (ed), *Diversification processes in language: Grammar: 33-46.* Hagen: Margit Rottmann Medienverlag.

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

Wimmer, G., & Altmann, G. (1996). The theory of word length distribution. *Glottometrika 15, 112-133.*

Wimmer, G., Köhler, R., Altmann, G. (2003). Unified derivation of some language laws. (to appear in *Handbook of Quantitative Linguistics.* Berlin: de Gruyter)

Best, K.-H., ***Quantitative Linguistik. Eine Annäherung*** (= Göttinger Linguistische Abhandlungen 3). Göttingen: Peust & Gutschmidt Verlag 2001. 132 Seiten. *By Gabriel Altmann*

Es gibt wenige Einführungen in die Quantitative Linguistik. Die Ursache liegt darin, dass man nicht weiß, was alles hineingehört, weil die QL alle Disziplinen der Linguistik umfasst – und noch etwas mehr. Man schreibt daher lieber Einführungen in einzelne Disziplinen (Phonologie, Textanalyse, Typologie, Lexikologie u.a.) und zeigt geeignete Methoden, die man dann detaillierter darstellen kann, oder man schreibt Bücher, die mehrere Disziplinen umfassen, aber dafür nur über den status quo informieren. Beide haben ihre Vorteile.

Best wählt einen völlig anderen Weg. Er beschränkt sich auf einige wenige Probleme aus unterschiedlichen Bereichen (Länge, Komplexität, Rangverteilung, Diversifikation, einige andere Gesetze), die sich praktisch mit zwei Softwarepaketen lösen lassen, und widmet sich nur der Frage, ob bestimmte zu Gesetzen gewordene Hypothesen an umfangreichen Daten bekräftigt werden können. Es sind Bereiche, an deren Fortschritt er selbst sehr aktiv beteiligt war.

Die Philosophie, die sich hinter dieser Absicht verbirgt, kann mit einem Satz charakterisiert werden: „Teste auf induktive Weise deduktive Hypothesen!“ Der Unterschied zu üblichen Verfahren in der Linguistik – auch der quantitativen – ist deutlich: Es geht nicht um das induktive Heranpirschen an eine Erscheinung, die dann auf die übliche Weise charakterisiert wird (mit Indizes, Regelbeschreibungen, Klassifikation usw.), sondern um Prüfung von Theorien, die die Genese einer Erscheinung deduktiv erfassen. Die Befürchtung, man stöße hier an komplizierte Mathematik, ist unbegründet. Man findet im ganzen Buch nur 18 Formeln, die man nicht einmal beherzigen muss, weil die ganze Arbeit die entsprechende Software erledigt: für den Anfänger ein paradiesischer Zustand, der ihm auch ohne Kenntnis der Differential- und Differenzengleichungen erlaubt, sehr wertvolle Resultate zu erzielen, Theorien zu unterstützen oder umzuwerfen.

Die im Buch enthaltenen Problemkreise und Resultate kann man aus der Tabelle ablesen. In der ersten Spalte steht die Eigenschaft, der Prozess oder das Gesetz, in der zweiten Spalte die untersuchte Entität oder Messeinheit, in der dritten Spalte sind die überprüften Modelle.

Problem	(Mess)einheit	Modell
Wortlänge	Silbe	Hyperpoisson, Hyperpascal, Conway-Maxwell-Poisson
	Morph	Hyperpoisson
	Phonem	Hyperpascal
	Buchstabe	Hyperpascal
Kompositumlänge	Glied	Hyperpoisson
Morphlänge	Phonem	Hyperpoisson
Silbenlänge	Phonem	Conway-Maxwell-Poisson
Satzgliedlänge	Wort	Hyperpoisson
Satzgliedtiefe	Attribut	Hyperpoisson
Satzlänge	Wort, Clause	Negative Binomial
Rhythmische Einheit	Silbe	Hyperpoisson
Komplexität der Chinesischen Schriftzeichen	Strich	Binomial
Rangfrequenzverteilung	Buchstaben	Zipf-Mandelbrot
	Phoneme	Zipf-Mandelbrot, negative hypergeometrische
	Wörter	Zipf-Mandelbrot

Diversifikation	Wortarten	negative Binomial
	Satzglieder	Hyperpoisson
	Kasus	Hyperpoisson
	Allomorphe	Hyperpoisson
	Komposita	Hyperpoisson
	„pray“ (engl.)	gemischte Poisson
	Evidenzmarker	positive Poisson
	Eigennamen	Hirata-Poisson
	Wortbildung	Hyperpoisson
	„que“ (fr.)	Hyperpoisson
	Etymologien	negative Binomial-Poisson
Martins Gesetz		Differenzengleichung
Menzeraths Gesetz		Potenzkurve
Frumkinas Gesetz		negative hypergeometrische, Poisson
Piotrowskis Gesetz		logistische Kurve, Altmanns Modelle

Wie man sieht, hat der Autor die Zahl der Attraktoren (Modellen) für einzelne Erscheinungen vernünftigerweise stark eingeschränkt. Die Theorie der Länge erlaubt praktisch unendliche Variabilität, Bests Kollektiv, das etwa 40 Sprache und über 3000 Texte untersucht hat, fand zahlreiche andere Modelle; für den Anfänger ist es aber angenehm zu wissen, dass er auch mit bescheidenen Mitteln zum Ziel kommen kann. Im Bereich der Längen ist die komplexeste Verallgemeinerung die Hyperpascal-Verteilung (die die Hyperpoisson-, die Poisson- und die negative Binomialverteilung umfasst) und die Conway-Maxwell-Poisson-Verteilung, die mit dem Menzerathschen Gesetz assoziiert ist.

Im Bereich der Rangfrequenzverteilung ist die Zipf-Mandelbrot-Verteilung das Standardmodell. Die Forschung in diesem Bereich ist besonders stark entwickelt, weil sich hier viele Mathematiker und Physiker eingeschaltet haben.

Im Bereich der Diversifikation erwartet man besonders viele Modelle – Best führt nur 7 an – weil die Randbedingungen sehr unterschiedlich sind. Bests Verdienst liegt darin, dass er dieses Gebiet beträchtlich erweitert und implizite die folgende Hypothese testet: „Diversifiziert eine Spracherscheinung, dann folgen die Häufigkeiten einzelner Klassen einem „anständigen“ Rangverteilungsgesetz“. Die Anwesenheit dieses Gesetzes bzw. die Bestätigung seiner empirischen Realisation ist gleichzeitig das beste Kriterium für die Klassifikation sprachlicher Entitäten. Führt man eine Klassifikation durch (z.B. der Wörter in Wortarten), dann ist diejenige Klassifikation fruchtbar, die diesem Gesetz entspricht. Das Gesetz wird zu einem Kriterium.

Etwas sparsamer werden die anderen Gesetz mit Daten belegt, weil hier die Ermittlung der Daten recht beschwerlich ist: Martins Gesetz bezieht sich auf die Verteilung der Wörter nach ihrer Allgemeinheitsstufe; Menzeraths Gesetz bezieht sich auf das Verhältnis zwischen Konstrukt und Konstituente und bildet eine Analogie zum „allmächtigen“ Potenzgesetz, das bis in die Theorie der Fraktale reicht; Frumkinas Gesetz bezieht sich auf die Verteilung einer Entität auf Textpassagen bestimmter Länge; Piotrowskis Gesetz ist ein Sprachwandelgesetz, dessen drei Modelle alle Veränderungen (vollständige, partielle, reversible) von Sprachentitäten erfassen.

Alle Anpassungen der Modelle an Daten werden sowohl tabellarisch als auch graphisch dokumentiert, so dass sich der Leser ein Bild einzelner Verläufe leicht einprägen kann. Die

Literatur wurde so ausgewählt, dass der Leser auch auf die Theorie zugreifen kann, falls er geneigt ist an die Wurzel der Gesetze zu gehen. Pädagogisch ausgezeichnet zusammengestellt eignet sich dieses Buch genau für den Zweck, für den es bestimmt wurde: eine Einführung in die Problematik der quantitativen Linguistik.

Books received – Büchereingang

An dieser Stelle werden im Folgenden sowie in den künftigen Ausgaben der Glottometrics Neuerscheinungen aufgeführt, die der Zeitschrift zur Rezension angeboten wurden. Neben den hier aufgeführten Publikationen können auch Neuerscheinungen, die nicht verzeichnet sind, bei der Rezensionsredaktion der Glottometrics angefordert werden. Die Herausgeber möchten an dieser Stelle ausdrücklich dazu auffordern, aktuelle Neuerscheinungen, die für eine Rezension zur Verfügung stehen, anzuzeigen. Interessenten für Rezensionen oder für Neuerscheinungsangebote möchten sich bitte bei der Adresse Rezensionsredaktion melden.

Kontakt:
Dr. Arne Ziegler, Universität Münster, Institut für Deutsche Sprache und Literatur und ihre Didaktik, Leonardo Campus 11, D-48149 Münster. E-Mail: arneziegler@uni-muenster.de

Best, Karl-Heinz (Hrsg.) (2001). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt Verlag. XX + 310 pp.

Hřebíček, Luděk (2000). *Variation in Sequences. Contributions to General Text Theory*. Prague: Oriental Insitute. 132 pp.

Mayer, Felix (ed.) (2001). *Language for Special Purposes. Perspectives for the New Millenium. Vol 1.: Linguistics and Cognitive Aspects, Knowledge Representation and Computational Linguistics, Terminology, Lexicography and Didactics. Vol. 2: LSP in Academic Discourse and in the Fields of Law, Business and Medicine.* Tübingen: Narr.

Pawlowski, Adam (2001*). Metody kwantytatywne w sekwencyjnej analizie tekstu*. Warszawa: Uniwersitet Warszawski.

Naukovij Visnik Černiveckogo universitetu. Vipusk 114, 2001. Germans´ka filologija.

Naukovij Visnik Černiveckogo universitetu. Vipusk 115, 2001. Germans´ka filologija.