Glottometrics 12
2006

RAM-Verlag

ISSN 2625-8226

Glottometrics

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

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

Herausgeber – Editors

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

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

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

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 12 (2006), Lüdenscheid: RAM-Verlag, 2006. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet (Open Access) unter der Adresse
https://www.ram-verlag.eu/journals-e-journals/glottometrics/ verfügbar.
Bibliographische Deskription nach 12 (2006) **ISSN 2625-8226**

In Celebration of Yet Another Milestone:

The 80th Birthday

of the Founder of Quantitative Linguistics in Japan

Shizuo Mizutani

This issue of *Glottometrics* (Number 12) in 2006 is dedicated in celebration of the 80th birthday of the founder of quantitative linguistics in Japan, Sizuo Mizutani. The details of his academic work are introduced by Maruyama in *Glottometrics* (Number 10) in 2005.

Contents of Glottometrics 12

Abstract. This paper examines the inter-textual type-token relationship and tests some existing quantitative models describing the vocabulary size and text length relationship. 8,334 samples were randomly drawn from the British National Corpus, totaling 8,001,000 tokens. The result shows that the models by Herdan and Brunet can capture the dynamic inter-textual type-token relationship, and the latter is also robust in extrapolation.

Abstract: A simple method for quantifying the complexity of graphical signs is suggested. The complexity is defined as the number of crossing points which can maximally be achieved with an overlapping straight line. In signs composed of several disconnected elements, the complexity is to be computed for each element seperately. This method is compared with another complexity measure serving a similar purpose recently proposed by Altmann.

Abstract. The paper investigates Indian Language pattern with respect to word length distribution. The languages selected are Assamese, Marathi and Hindi belonging to Indo-Aryan family and Kannada and Tamil belonging to Dravidian Family. It examines data of different registers like Aesthetics, Social Science, Natural, Physical and Professional Sciences, Official and Media languages and Translated Material. It is observed that no single distribution fits across languages while across registers a single distribution fits majority of samples.

Abstract. Language acquisition abides by laws. These laws seem to be the same as in language change and text production. If one considers the stages of language acquisition separately then there are further models controlling distributions, rank orders and processes. This paper yields some evidence for these laws.

Abstract. Sentence length in German is studied in three formal length categories and three genres. We try to characterize the individual writer by means of a vector of properties which can be used both for classification and the study of development.

Abstract. The word *hinoki* or 'cypress' can be transcribed in two variant forms, (the so-called "traditional" variant) and (the "simplified" variant), in Japanese kanji. Such variant forms are called *kanji variants*. The present paper reviews a series of studies on Japanese kanji recognition (Yokoyama, 2006a, 2006b, 2006c), and proposes a logistic regression model which accounts for performance in a preference judgment task based on kanji frequency data. Yokoyama (2006a) administers preference and familiarity judgment tasks in which the participants were presented with 263 pairs of traditional and simplified variants and asked to choose the more preferable or familiar variant of each pair. The analyses indicate a positive contribution of frequency to variant preferences, supporting the so-called "mere exposure effect" theory of Zajonc (1968). This finding leads to a logistic regression model that describes preference behavior in kanji recognition, based on Fechner's law. Yokoyama (2006b) shows that the model is comparable to the so-called "generalized matching law" of Baum (1974) and to the "ideal free distribution theory" of Fagen (1987). Yokoyama (2006c) further examines the predictive validity of the model with empirical data obtained from a preference judgment task, administered in the Tokyo and Kyoto areas. Logistic regression analyses are performed with the ratio of preference for the given variants and the logit of the character frequencies, yielding significant correlations between the predicted probabilities and the observed responses (r = .804 for Asahi newspaper data). The present paper synthesizes these studies and proposes a logistic regression model that efficiently describes preference behavior in Japanese kanji recognition, integrating the theoretical perspectives of the mere exposure effect and the generalized matching law.

History of Quantitative Linguistics

Book reviews

Glottometrics 12, 2006, 1-10

Models for dynamic inter-textual type-token relationship

Fan Fengxiang[1]*, Dalian*

Abstract. This paper examines the inter-textual type-token relationship and tests some existing quantitative models describing the vocabulary size and text length relationship. 8,334 samples were randomly drawn from the British National Corpus, totaling 8,001,000 tokens. The result shows that the models by Herdan and Brunet can capture the dynamic inter-textual type-token relationship, and the latter is also robust in extrapolation.

Keywords: inter-textual type-token relationship, TTR-models, English

Introduction

The type-token relationship figures prominently in quantitative linguistics and language teaching. The type-token ratio (TTR) serves as one of the markers for genre classification (Biber, 1988, 1989), authorship attribution (Stamatatos et al., 1999), text categorization (Karlgren, Cutting, 1994), and as a measure for vocabulary diversity (Schmitt, 2002). In ESL/EFL (English as a Second Language/English as a Foreign Language) teaching, TTR is mainly used to measure lexical density of a text (Nation, 1990), and the learner's lexical variation in their compositions (Arnaud, 1984).

Linguists have noticed the sensitivity of the TTR to the number of tokens (Guiraud, 1954, Orlov, 1982, Sichel, 1986, Holmes, 1994), and methods have been devised to overcome the variability of the TTR, i.e. Yule's *K*, Guiraud's *R*, Sichel's *S*, Honoré's *H*, Scott's standardized TTR (Scott, 1996) and so on. Though the quest for a stabilized type-token relationship is both of theoretical and practical significance, the study on the variability itself is important, particularly for ESL/EFL teaching, where there is a lack of robust models describing the dynamic type-token relationship. Such models can be used for language course design and for the development of lexical acquisition theories.

The aim of this research is to study the quantitative behavior of types in relation to the increase of the number of cumulative tokens from different texts, and search for a model capturing such a dynamic relationship. To achieve such a purpose, 8,334 samples were randomly drawn from the British National Corpus (the BNC), totaling 8,001,000 tokens. The samples are of four different sizes: 500-word, 1000-word, 1500-word and 2000-word, which are the normal text length in intermediate and advanced ESL/EFL teaching. Considering the size of the well-known corpora, such as the Brown Corpus and the LOB Corpus, which respectively

[1] Address correspondence to: Fan Fengxiang, Foreign Language Department, Dalian Maritime University, Dalian 116026, China. E-mail: fanfengxiang@yahoo.com

contain 500 2000-word samples totaling 1,000,000 tokens each, the number of samples and the total number of the cumulative tokens are very large for a study of this nature. Such a number can test a model's descriptive power for a very large number of cumulative tokens since some models may have a threshold over which they will break down. A set of computer programs in Foxpro was used for sampling and calculating the type increase as the samples were drawn one by one.

There are two kinds of types, the string types and the lemma types. The former refers to different word-forms while the latter refers to the set of word-forms with the same sense, differing only in inflections. Under this definition, *give, gives, giving, gave, given* are five different string types but only one lemma type. In this paper *type* refers to *lemma type*. It excludes Arabic numerals, personal and place names, and non-word strings. Tokens include all the character clusters and single characters in a text except punctuations. In addition, *type* is used interchangeably with *vocabulary*. The difference is that *vocabulary* is used in relation to *text*.

Models for the vocabulary size and text length relationship

There have been a number of influential quantitative models describing the vocabulary and text length relationship. A selection of indices is listed below. In these models V stands for the size of vocabulary and N for text length. Other symbols are parameters.

(1) $V = a\sqrt{N}$ (Guiraud, 1954). Sánchez & Cantos (1997) proposed a similar model $Types = K\sqrt{Tokens}$.

(2) $V = \exp(\ln^{\alpha} N)$ (Somers, 1959)

(3) $V = \alpha N^{\beta}$ (Herdan, 1964). Heaps (1978) proposed a similar model in the form $D = KN^{\beta}$, which is known as the Heaps' Law.

(4) $V = \alpha(\ln N)^{\beta}$ (Brunet, 1978)

(5) $V = \dfrac{Z(\ln Z - \ln N)N}{(\ln Z + \alpha)(Z - N)}$ (Orlov, 1982). Z is the Zipf size.

(6) $V = \dfrac{aN}{\beta + N}$ (Tuldava, 1995)

(7) $V = \dfrac{aN}{1 - \beta + \beta N}$ (Köhler, Martináková, 1998)

These models will be tested on the observed type increase of the set of 8,334 random samples.

Analysis and results

The 8,334 random samples totaling 8,001,000 tokens produce 60,193 types. However, if Arabic numerals, personal and place names and non-word strings were included, the number of types would increase to 10,082. Figure 1 and Figure 2 are respectively the type growth curve and the TTR decrease curve.

We omitted to scrutinize the possible variation in sequencing of samples, though the problem of inhomogeneity with which we are confronted here plays an important role (cf. Altmann, 1992). Different order of samples may bring different (better or worse) results. We rely on the equal value of all possible sequencings of samples.

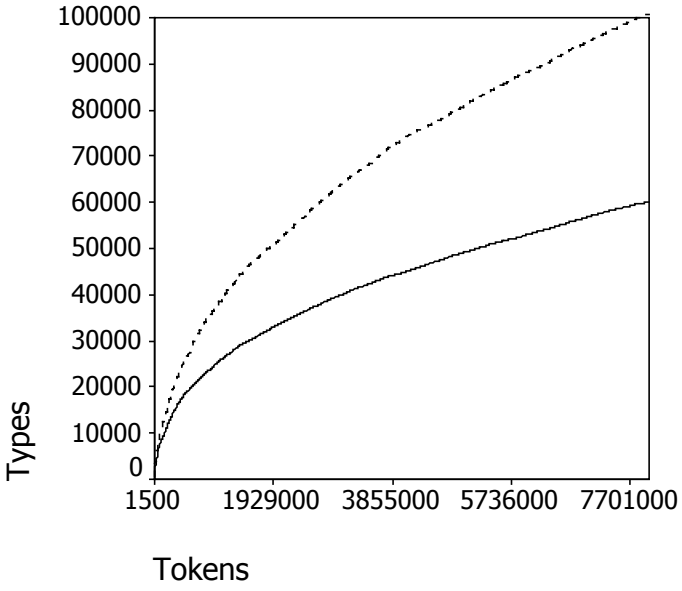


Fig. 1. Type growth curve. The solid line is the "clean" type growth curve, the dotted line the one with Arabic numerals, personal and place names, and non-word strings.

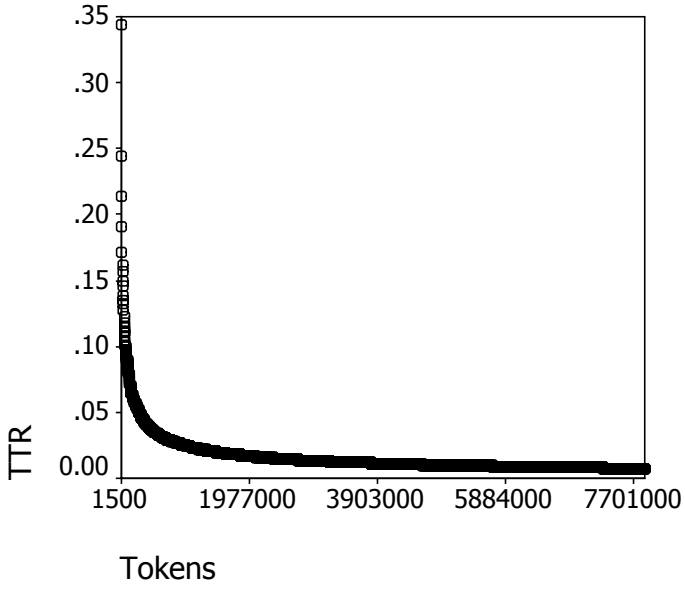


Fig. 2. TTR decrease curve

Despite the number of cumulative tokens, the type growth curve shows no sign of leveling out even when it reaches the end. It is still in the LNRE zone. The rate of the increase is very high initially; the first 50,000 tokens produce 5,564 types. But the rate of increase gradually slows down. For example, from 7,774,000 tokens to 7,986,000 tokens, the number of types increases from 59,474 to 60,009, a yield of only 536 new types out of 212,000 tokens. The initial TTR is 0.3433, dropping to 0.0075 at the end.

All the models were tested on the observed type growth of the set of 8,334 samples, and the results are shown in Figure 3, Figures 4, Figure 5 and Figure 6. The solid lines are the observed values and the dotted lines the model fits.

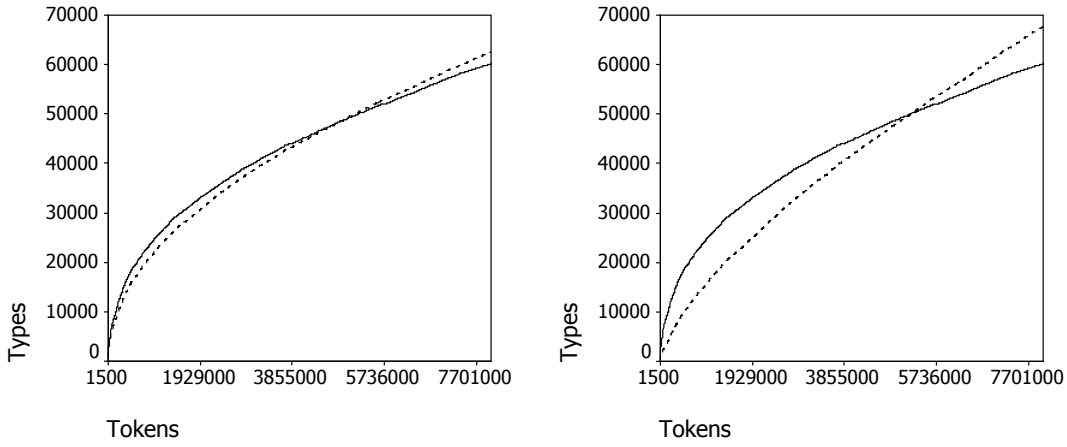


Fig 3. Fit of (1), (left), a = 65.7365677; and fit of (2), (right), a = 0.8698.

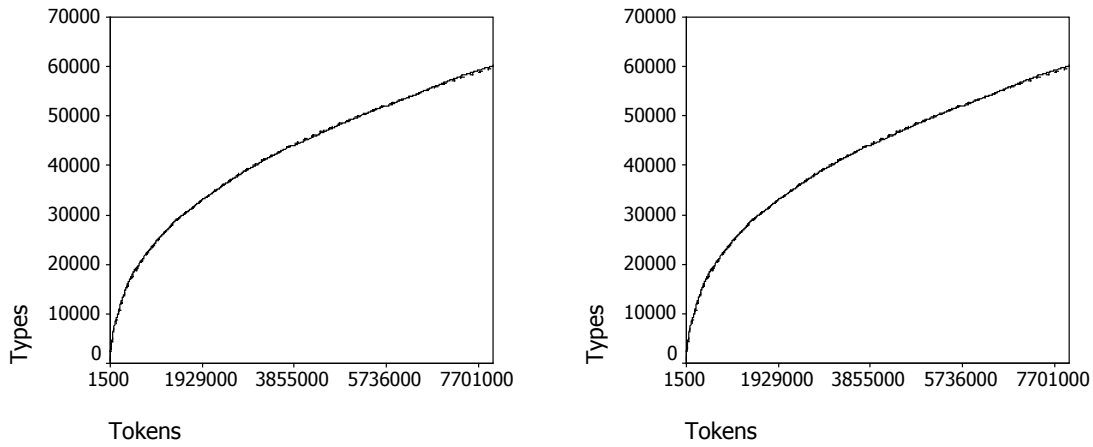


Fig. 4. Fit of (3), (left), a = 65.73656, β = 0.4291; and fit of (4), (right), a = 0.0014238, β = 6.345906.

Of these models, (3) and (4) give very good fits. Their determination coefficient (R^2) is respectively 0.99971 and 0.99968. Although (3) has a slightly but irrelevantly larger determination coefficient, it severely underestimates the observed values between 1,500 tokens and 150,000 tokens, while the initial deviation of (4) from the observed values is very mild. Table 1 reveals this fact. To save space, it only shows the number of tokens, the observed number of types, the model estimations and errors between 1,500 tokens and 42,000 tokens.

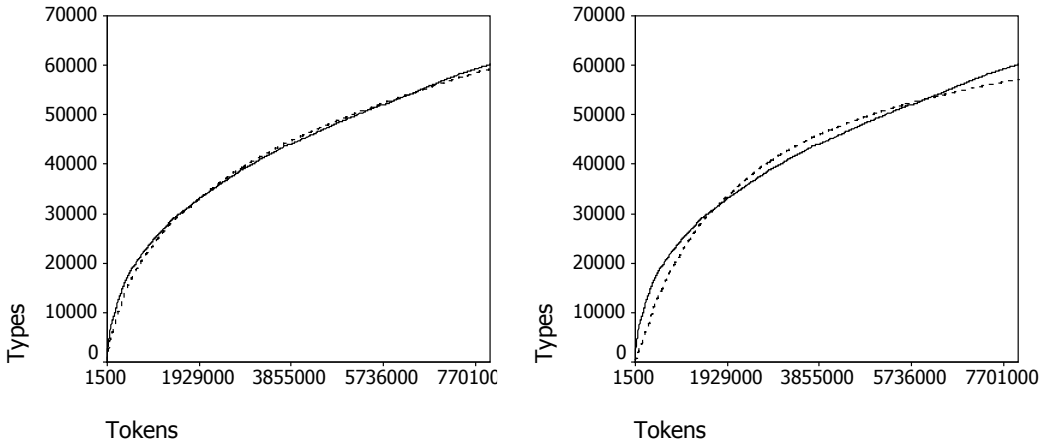


Fig. 5. Fit of (5), (left), Z = 132000, a = 33.9819; and fit of (6) and (7), (right). The latter two produce exactly the same fit that completely overlaps each other (the dotted line). For (6), a = 73573.696, β = 2308732.74; and for (7), a = 0.031868, β = 4.33138.

Table 1
The observed number of types, the model estimations and errors
between 1,500 and 42,000 tokens

TK	TP	ES4	E4	ES3	E3	TK	TP	ES4	E4	ES3	E3
1500	515	434	82	1516	-1001	23000	3516	3245	271	4893	-1377
3000	888	770	118	2042	-1154	24000	3594	3334	261	4983	-1389
4000	1088	964	125	2310	-1222	24500	3645	3377	268	5027	-1382
4500	1200	1054	146	2430	-1230	25000	3679	3420	259	5071	-1392
5000	1297	1140	157	2542	-1245	26000	3792	3505	287	5157	-1365
6500	1591	1382	209	2845	-1254	27500	3981	3630	351	5283	-1302
7000	1655	1458	197	2937	-1282	28500	4042	3711	331	5365	-1323
8500	1823	1673	150	3192	-1369	29500	4101	3791	310	5445	-1344
9000	1926	1742	184	3271	-1345	30000	4125	3830	295	5484	-1359
11000	2188	2000	188	3565	-1377	30500	4152	3869	283	5523	-1371
12500	2441	2181	260	3766	-1325	31000	4182	3908	274	5562	-1380
13000	2475	2239	236	3830	-1355	31500	4233	3947	286	5600	-1367
15000	2687	2463	225	4073	-1386	32000	4255	3985	270	5638	-1383
16000	2799	2569	230	4187	-1388	32500	4293	4023	270	5676	-1383
17000	2910	2673	237	4298	-1388	33000	4322	4061	262	5713	-1391
18500	3068	2824	244	4457	-1389	33500	4349	4098	251	5750	-1401
19500	3170	2921	249	4558	-1388	35500	4549	4245	304	5895	-1346
20500	3302	3016	286	4657	-1355	36000	4611	4281	330	5930	-1319
21000	3352	3063	289	4706	-1354	37000	4704	4352	352	6000	-1296
21500	3409	3109	300	4753	-1344	39000	4800	4492	308	6137	-1337
22000	3451	3155	296	4801	-1350	40000	4842	4561	281	6205	-1363
22500	3469	3201	269	4847	-1378	42000	4993	4696	297	6336	-1343

TK: inter-textual cumulative tokens
TP: the observed number of types
ES4: estimation of (4)
E4: errors of (4)
ES3: estimation of (3)
E3: errors of (3)

Figure 6 compares the errors of (3) and (4).

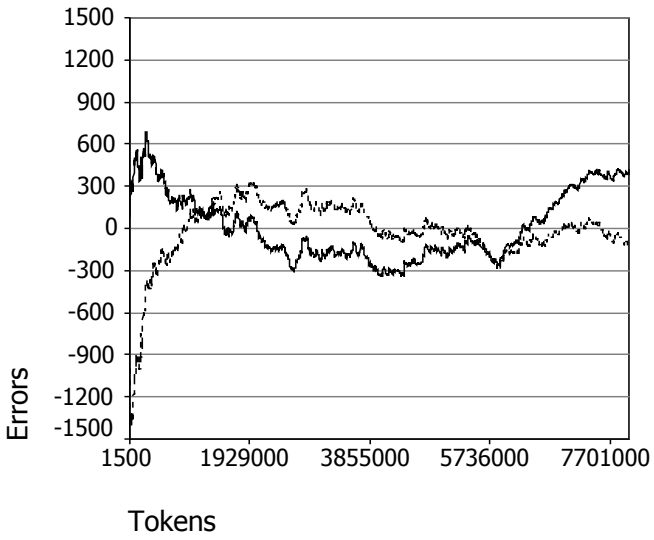


Fig. 6. The error curves of (3), (the broken line), and (4), (the solid line)

Figure 6 shows that from about 6,000,000 tokens upwards, (4) starts to have bigger deviations than (3). At the end of the curve the error of (4) is 414 while (3) is only –87. However, considering the corresponding observed number of types (60,193), 414 is almost negligible, while the errors of (3) at the initial stage are intolerable because the model fit departs substantially from the observed values.

The fit of (5) is acceptable, with a determination coefficient 0.99644; however, its initial deviation is too big. For example, at 60,000 tokens the observed number of types is 6,151, but the model's estimation is 4,042, an error of 2,110. The fit of (1), (6) and (7) is poor. The determination coefficient of (1) is 0.98473. (6) and (7) have the same determination coefficient 0.97361. The fit of (2) does not resample the observed curve at all. Its determination coefficient is 0.82007. Figure 7 and Figure 8 are the above mentioned model error curves.

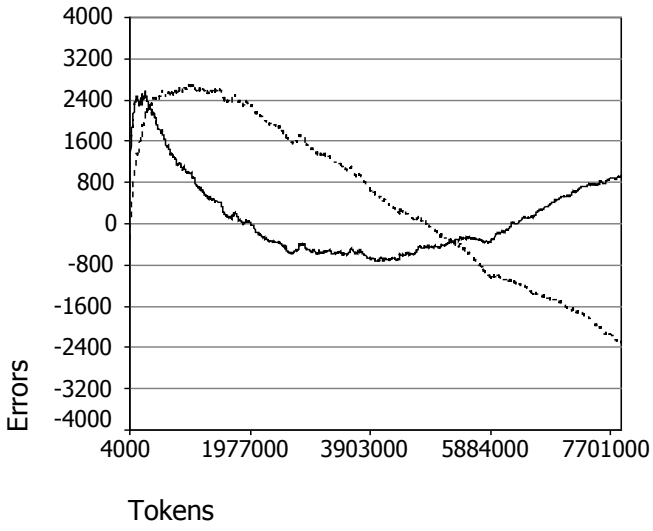


Fig. 7. Error curves of (5), (the solid line) and (1), (the broken line).

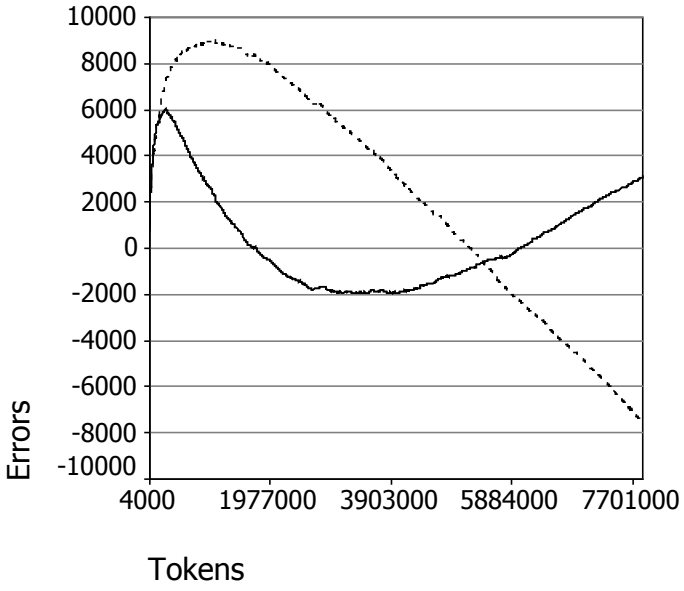


Fig. 8. Error curves of (6) and (7), (the solid line), and (2), (the broken line). The curves of (6) and (7) completely overlap each other.

Discussion

Curves (3) and (4) prove to be good for the description of the dynamic inter-textual type-token relationship under 8,001,000 inter-textual tokens. Since the fit of (4) has very mild deviation throughout the observed type growth curve, it is more appropriate for the study of lexical acquisition and ESL/EFL teaching.

It is estimated that on average educated native speakers have a vocabulary of 20,000 (Nation 1990, Nagy,1997). Radford et al. (1999) think it is around 30,000. If we put the figure at 25,000, the number of inter-textual tokens needed to produce this number of types can be estimated with (8), obtained from (4):

(8) $N = e^{\beta\sqrt{\frac{V}{\alpha}}}$.

With the parameters of (4) for the 8,001,000 inter-textual tokens, N is 1,040,078. This vocabulary size is not difficult for native students to acquire since generally they read about a million words of text a year (Nagy, 1997).

In ESL/EFL teaching, course designers must decide on the volume of input texts needed for the intended number of new words to be learned. If the learners already have a vocabulary of 1,000 words, and the course designer wishes to add 2,000 more, then the volume of texts in terms of cumulative tokens can be estimated with (8). V is set to 3,000; and N is 20,326, roughly the number of cumulative tokens contained in 20 1000-word texts.

Apart from mild deviations, (4) is also fairly robust for extrapolation while (3) is not. Both (3) and (4) were tested on a set of 1,000 samples randomly drawn from the BNC,

totaling 1,000,000 tokens. For (3) a = 30.481125, β = 0.4857; for (4) a = 0.002095567, β = 6.202270648. These parameters were used on the set of 8,001,000 cumulative tokens. Figure 9 and Figure 10 shows the results. The solid lines are the observed values, the dotted line the extrapolation.

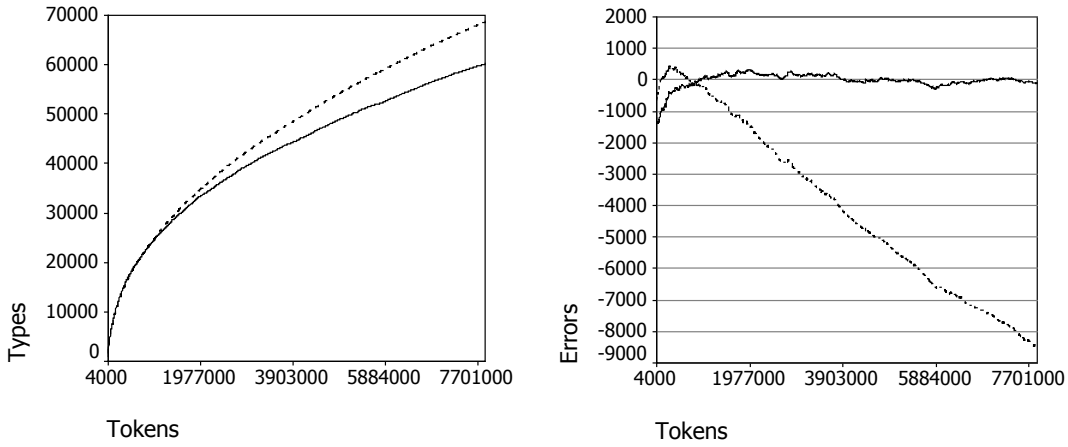


Fig. 9. Extrapolation of (3). Left: model fit with a = 30.481125, β = 0.4857. Right: error comparison. The solid line is the errors of the model fit with the original parameters, the broken line the errors of extrapolation.

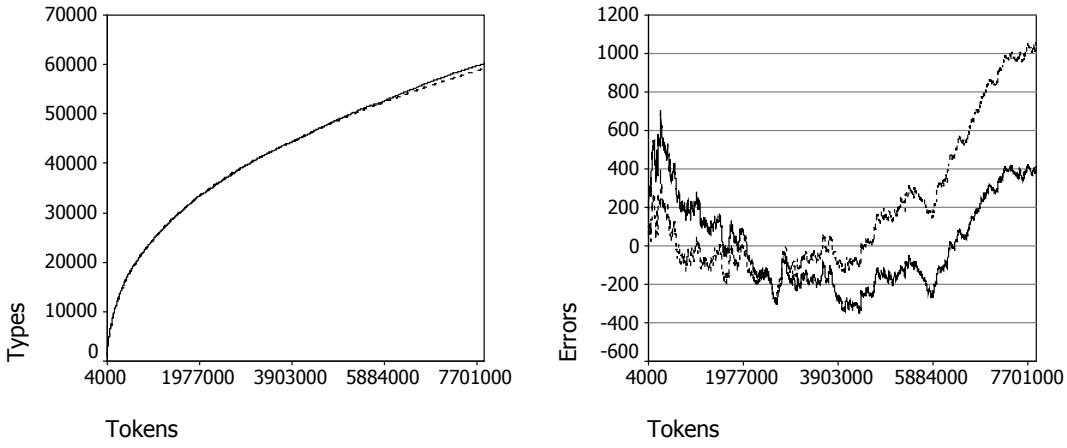


Fig. 10. Extrapolation of (4). Left: model fit with a = 0.002095567, β = 6.202270648. Right: error comparison. The solid line is the errors of the model fit with the original parameters, the broken line the errors of extrapolation.

The parameters of (3) for the set of 1,000,000 tokens can not be extrapolated to the set of 8,001,000 tokens while the parameters of (4) can. As Figure 10 shows, the fit of extrapolation of (4) is still good; its determination coefficient is 0.99903 while that of (3) is 0.86382. For the first half of the extrapolation curve of (4) the fit is even better than the fit with the original parameters. On the second half of the curve the deviation starts to increase. At the end of the curve the error of the extrapolation reaches its peak 1,056, but it is still acceptable because it is only 1.17% of the observed values.

We now stretch this robustness of (4) to the limit and use (8) to try to estimate the cumulative tokens needed to generate the entire set of the vocabulary in general English. For

the sake of simplicity, assuming this set of vocabulary is all contained in the Oxford English Dictionary of the second edition, which has 290,000 head words. With the parameters for the set of 8,001,000 tokens and V set to 290,000, N is 713,979,058. Tests reveal that N is over-estimated in extrapolation with (8) using parameters obtained from the smaller set of samples. Therefore it is highly possible that 713,979,058 inter-textual tokens would be enough to produce 290,000 types.

Summary and conclusion

In the inter-textual type-token relationship, the number of types is a function of the number of inter-textual cumulative tokens, with an ever decreasing TTR as the number of tokens increases. The number of types increases rapidly at the beginning then the rate of increase slows down. However, the type growth curve is still on the rise as it reaches the end. The starting value of TTR of the 8,001,000 tokens from 8,334 random samples is 0.3433, decreasing to 0.0075 at the end.

The inter-textual type-token relationship can be well captured with Herdan's model and Brunet's model, whose determination coefficient is respectively 0.99971 and 0.99968. Although Herdan's model has a negligibly larger determination coefficient than Brunet's it gravely underestimates the observed values under 150,000 tokens. In this respect Brunet's model is more suitable for use in lexical acquisition studies and EFL teaching. In addition, it is robust for extrapolation. Orlov's model fit is acceptable. Tuldava's model and Köhler-Martináková's model give poor fits. Since they are merely modifications of the same formula, the two produce exactly the same fit for the empirical data. Somers' model can not describe the inter-textual type-token relationship at all, though, perhaps, other sequencing of samples might have resulted in a better fit

Considering the vastness of English texts, the number of the samples and the number of the cumulative tokens are still too small. It would be of interest to study the behavior of Herdan's model and Brunet's model with much larger sets of empirical data.

References

Altmann, G. (1992). Das Problem der Datenhomogenität. *Glottometrika 13, 105-120.*

Arnaud, P. (1984). The lexical richness of L2 written productions and the validity of vocabulary tests. In T. Culhane, C. Klein Bradley, & D. K. Stevenson (Eds.), *Practice and problems in language testing: papers from the International Symposium on Language Testing: 14-28*. Colchester: University of Essex.

Biber, D. (1988) *Variation across Speech and Writing* Cambridge: Cambridge University Press

Biber, D. (1989). A Typology of English Texts *Linguistics 27, 3-43.*

Brunet, E. (1978). *Le vocabulaire de Jean Giraudoux. Structure et évolution*. Genève: Slatkine.

Guiraud, H. (1954). *Les Caractères Statistiques du Vocabulaire*, Paris: Presses Universitaires de France.

Heaps, H. (1978). *Information Retrieval, Computational and Theoretical Aspects*. New York: Academic Press.

Holmes, D. (1994). Authorship attribution. Computers and the Humanities *28(2), 87-106.*

Karlgren, J., Cutting, D. (1994). Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In: *Proceedings of the 15th International Conference on Computational Linguistics.*

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: 514-546.* Berlin: de Gruyter

Nagy, W. (1997). On the Role of Context in First- and Second-language Vocabulary Learning. In Schmitt, N. and McCarth, M. (eds), *Vocabulary: Description, Acquisition and Pedagogy: 64-83.* Cambridge: Cambridge University Press.

Nation, I. (1990). *Teaching and Learning Vocabulary*. New York: Newbury.

Orlov, J. (1982). Ein Modell der Häufigkeitsstruktur des Vokabulars. In: Orlov, J.K., Boroda, M.G, Nadarejšvili, I.Š. (1982), *Sprache, Text, Kunst. Quantitative Analysen*: *118-192.* Bochum: Brockmeyer.

Radford, A., Atkinson. M., Britain, D., Clahsen, H., Spencer, A. (1999). *Linguistics, An Introduction*. Cambridge: Cambridge University Press.

Sánchez, A., Cantos. P. (1997). Predictability of Word Forms (Types) and Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the CUMBRE Corpus: An 8-Million-Word Corpus of Contemporary Spanish. *International Journal of Corpus Linguistics 2(2), 259-280.*

Scott, M. (1996). *Wordsmith Tools*. Oxford: Oxford University Press.

Sichel, H. (1986). Word frequency distributions and type-token characteristics. *Mathematical Scientist 11, 45-72*

Schmitt, N. (2002). Using Corpora to Teach and Assess Vocabulary. In Tan, M. (Ed.), *Corpus Studies in Language Education: 31-44*. Thailand: IELE Press.

Somers, H. (1959). *Analyse mathématiquwe du langage: Lois générales et mesures statistiques*. Louvain: Nauwelaerts.

Stamatatos, E. (1999). Automatic Authorship Attribution In: *Proceedings of the Ninth Conference of the European Chapter of the Association for the Computational Linguistics*.

Tuldava, J. (1995). *Methods in quantitative linguistics*. Trier: WVT.

Glottometrics 12, 2006, 11-15

Script complexity revisited

Carsten Peust, Konstanz[1]

Abstract: A simple method for quantifying the complexity of graphical signs is suggested. The complexity is defined as the number of crossing points which can maximally be achieved with an overlapping straight line. In signs composed of several disconnected elements, the complexity is to be computed for each element seperately. This method is compared with another complexity measure serving a similar purpose recently proposed by Altmann.

It is intuitively clear that graphical signs can have different degrees of *complexity*: Q is obviously more complex than I, and A is more complex than A. Altmann made a proposal for numerically measuring this intuitive category recently in this journal (Altmann 2004). His proposal (in the following *composition method*) is to split a sign into basic elements, for which he defines specific costs (point: costs 1; straight line: costs 2; arch: costs 3). In addition, three types of contacts are to be considered, again involving costs (continuous: costs 1; crisp: costs 2; crossing: costs 3). The complexity of a sign is the sum of the costs of all of its basic elements as well as contacts. Altmann finds that the complexity counts achieved by his system agree well with our intuitive understanding of sign complexity.

Altmann's composition method indeed turns out to be a useful way of quantifiying sign complexity, but a slight feeling of dissatisfaction remains. Two quite heterogeneous principles are at work (elements and contacts), together with six cost values which have to be defined arbitrarily. In this paper, an alternative method is proposed which makes use only of a single principle and therefore involves fewer arbitrary assumptions, is applicable even more rapidly and unambiguously, and still reflects well our intuitive idea of sign complexity (in the following *intersection method*). I wish to thank Gabriel Altmann who has first drawn my attention to the topic and who generously offered me the opportunity to publish a counterproposal in this journal.

Assume that a straight line overlaps with the sign to be measured. This results in crossing points, whose number will vary for different positions and rotations of the line. What we will consider is the number of crossing points which can be maximally achieved with an optimal placement of the line over the graphical sign. I thus propose the following definition:

Rule 1: *The complexity of a sign is the maximal number of crossing points that can be achieved with a straight line.*

Instead of counting crossing points, it is an equivalent possibility to count the number of black/white-transitions and divide it by two. A few examples will suffice to illustrate the idea.

O has a complexity of 2 because there are, at maximum, two crossings or four black/white-transitions:

[1] Address correspondence to: Carsten Peust, Bücklestr. 68a, D-78467 Konstanz, Germany. E-mail: cpeust@gmx.de

A has a complexity of 3:

A has a complexity of 5:

In applying this method, the sign should be considered as being composed of lines with minimal width. The sign B arrives at a complexity of 4 because, assuming it as a shape of minimal thickness, a straight line can be placed so as to cross each of its two arches twice, which results in four crossings or eight black/white-transitions:

It seems desirable to achieve the same complexity also for the same sign in a bolder type (e.g. **B**), although eight black/white-transitions are not actually possible here because the outline is too thick.

The following table illustrates which complexity values are reached for the majuscles of the Latin alphabet in the fonts Arial and Courier New according to both algorithms:

Arial	*Composition Method*	*Intersection Method*	**Courier New**	*Composition method*	*Intersection Method*
A	12	3	A	22	5
B	16	4	B	16	4
C	7	2	C	11	3
D	9	2	D	9	3
E	14	3	E	26	5

F	10	3	F	22	5
G	15	3	G	15	4
H	10	3	H	26	5
I	2	1	I	10	3
J	3	2	J	7	3
K	10	3	K	26	6
L	6	2	L	14	3
M	14	4	M	26	6
N	10	3	N	20	5
O	8	2	O	8	2
P	9	3	P	13	4
Q	13	3	Q	21	4
R	14	4	R	22	6
S	15	3	S	23	5
T	6	2	T	18	3
U	3	2	U	11	3
V	6	2	V	14	3
W	14	4	W	22	5
X	7	2	X	23	4
Y	8	2	Y	20	4
Z	10	3	Z	18	3

It appears that a higher complexity in one system normally implies a higher complexity in the other: Both methods are roughly equivalent. Based on the given set of 52 test signs, the following correspondency table can be set up:

Intersection Method	1	2	3	4	5	6
Composition Method	2	3-9	7-18	13-23	20-26	22-26

On the other hand, there do exist sign couples for which the two methods lead to contradictory results. Some extreme cases are shown in the next table. They are arranged so that the left-hand sign is less complex and the right-hand sign more complex according to the composition algorithm, but vice versa according to the intersection algorithm:

Sign couple		Composition complexity		Intersection complexity	
J	D	7	9	3	2
P	T	13	18	4	3
N	X	20	23	5	4
R	H	22	26	6	5

The crucial question is whether the signs in the left-hand column are the "simpler" ones (which would favour the composition method) or the more "complex" ones (which would favour the intersection method). A spontaneous judgement may seem difficult, but it is my perception that the advantage lies slightly on the side of the intersection method.

Another difference is that Altmann's composition method leads to a much finer gradation of complexity than does the intersection method proposed here. This may be desirable. On the other hand, I believe that these gradations are on the whole not confirmed by our intuitive notions of sign complexity. Let us inspect all those signs which, with the intersection algorithm, are uniformly assigned the complexity of 3, sorted in increasing (1) and decreasing (2) order of their composition complexity, which varies as widely as from 7 to 18:

(1): J P D F H K N Z I C U A Q E L V G S T Z
(2): Z T S G V L E Q A U C I Z N K H F D P J

Can it be recognized that the complexity increases in (1) and decreases in (2), which is what Altmann would postulate? I do not think so. Let us examine two other sequences of signs, namely signs for which both methods unanimously predict an increase (3) or a decrease (4) of complexity:

(3): I U H B F M
(4): M F B H U I

There is no doubt in this case as to which row has increasing and which one has decreasing complexity.

There is, however, a considerable systematic advantage of Altmann's composition method compared with the intersection method as presented up to now. Altmann's system allows for the simple addition rule *C(a) + C(b) = C(a+b)*: The complexity of a sign group is the sum of the complexities of its individual elements. The composition complexity can thus be plausibly applied also to greater entities such as words or even texts. With the intersection method, however, the value for a word would normally fall drastically below the sum of its component values because most letters would not be crossed optimally or not be touched at all by a single line.

This theoretical problem leads to unintuitive results even in the computation of single letters as soon as they become more complex. Consider the Arabic letters س (*s*) and ش (*š*). س has an intersection complexity of 6 (cut twice each through the bottom of all three arches). Although the shape of ش is clearly derived from س by the addition of diacritical points, these points do not affect its intersection complexity, as defined up to now, because a straight line cannot be placed optimally so as to include any of them. It is clearly counterintuitive that both signs should have the same complexity.

As another example, consider the Korean group 음 (*ym*) (the readers are asked to take the bottom element as a simple rectangle with no serifes for the sake of this argument). The complexity will be 5 (cut through all three elements in vertical direction). The similar group 임 (*im*), which differs only in the arrangement of its elements and is likely to be regarded as identical in complexity by many, achieves, however, a complexity of only 4 (again with a vertical line).

In order to remedy such unintuitive results, I wish to posit the following additional rule:

Rule 2: *The complexity of a graphical cluster consisting of several disconnected components must not be computed with a single straight line. Instead, its complexity is defined as the sum of the complexities of its components.*

This ensures that the attractive additive behaviour of Altmann's composition complexity can be retained also in our system. The complexity of ش, a sign composed of four components separated by white space, will now be 6+1+1+1 = 9; the complexity of both 음 and 임 will be 2+1+2 = 5. Our additional rule does not change any of the values of the Latin capitals as discussed above since none of them is composed of disconnected elements.

What range of values will be achieved for sign complexity by applying the intersection method proposed here? As we saw, the values fall between 1 and 6 for Latin majuscles; the results are similar for the minuscles. In Arabic, values are largely the same, with an exceptionally high value of 9 for the already discussed letter ش (*š*) with three diacritics. In the Tamil alphabet we reach complexities up to 7: இ (*i*) and 8: ண (*ṇ*) (I am neglecting here the composite syllabic groups which can be even more complex). In Chinese, as might be expected, these values are outnumbered even by very familiar signs such as 瞧 (*qiáo* "to look", complexity 16), 餐 (*cān* "food", complexity 17), 罐 (*guàn* "tin, box", complexity 19), or 露 (*lù* "dew", complexity 20).

Reference

Altmann, G. (2004). Script complexity. *Glottometrics 8, 68-74.*

Glottometrics 12, 2006,16-38

Word Length Distribution in Indian Languages

B.D. Jayaram and M.N. Vidya[1]

Abstract. The paper investigates Indian Language pattern with respect to word length distribution. The languages selected are Assamese, Marathi and Hindi belonging to Indo-Aryan family and Kannada and Tamil belonging to Dravidian Family. It examines data of different registers like Aesthetics, Social Science, Natural, Physical and Professional Sciences, Official and Media languages and Translated Material. It is observed that no single distribution fits across languages while across registers a single distribution fits majority of samples.

Keywords: Word length, Assamese, Marathi, Hindi, Tamil, Kannada, Indian languages

Introduction

The Word occupies a central position in different linguistic disciplines. The study of Word length distribution is a topic of interest to corpus linguistics, quantitative linguistics, synergetic linguistics, and psycholinguists. The study serves as one of the basis for making language typology and searching for language laws. It is of interest from a historical perspective also because it provides a clue about the process of word formations. Such studies are also helpful to text processing for they provide insight into the laws that govern the occurrence of two-three-four syllabic words in a text. Word length studies also contribute to stylistic studies by distinctiveness of genre based on word length distributions (Milic, Slane, 1994). It is therefore necessary to conduct systematic studies in order to investigate how authorship, text type, temporal factors etc., may influence the frequency distribution of the word length in texts.

The word length distribution has been worked mostly on the languages of Europe and the majority of them fall under Indo-European language family. There are some languages outside Europe like Chinese (Best, Zhu, 2001), Japanese (Ejiri, Staeheli, Ooaku, 1994), Arabic (Abbe, 2000), Korean (Kim, Altmann, 1996), Old Hebrew (Balschun 1997), Eskimo (Meyer 1999), Ketchua (Best, Medrano 1997), Lakota (Pustet, Altmann 2005), Indonesian (Altmann et al. 2002) and Turkish (Best, Özmen 1996) on which scholars have worked on the phenomena of word length distribution. A complete bibliography can be found in http://www.gwdg.de/~kbest/litlist.htm. It is observed that in the majority of the cases the word length follows some modified, generalized or mixed Poisson distribution. The present paper investigates how Indian language pattern emerges vis a vis other languages of the world. The paper also presents the analysis of word length distribution for different registers like Aesthetics, Social Science, Natural Physical & Professional Sciences, Commerce, Official & Media Languages and Translated Materials across different languages spoken in India.

India is a multicultural and multilingual nation with 114 autonomous languages (Census 1991). These languages are classified linguistically into four genetic types or families: Indo-European, Dravidian, Austro-Asiatic and Tibeto-Burmese. One of the most important features of Indian languages is that they share many common vocabulary and some features of

[1] Address correspondence to: B.D. Jayaram, Central Institute of Indian Languages, Manasagangotri, Mysore - 570 006, India. E-mail: jayaram@ciil.stpmy.soft.net

grammar across language families owing to co-existence for centuries and also due to contact between them.

The distribution of these languages in terms of the families and the number of speakers are given in Table 1 (Jayaram et.al):

Table 1
Languages of India

Sl. No.	Language Family	Number of Languages	Number of Speakers	Percentage to the total populations
1.	Indo-European	20	631,351,789	75.299
	(a) Indo-Aryan	19	631,273,191	75.278
	(b) Germanic	1	178,598	0.021
2.	Dravidian	17	188,945,126	22.531
3.	Austro-Asiatic	14	9,490,157	1.132
4.	Tibeto-Burmese	63	8,092,940	0.003
	Total	114	838,001,987	99.930

Of these 114 languages, 22 major languages are listed in the VIII Schedule of the Constitution of India. The article 351 of the Indian constitution allows the government to compile a schedule (list) of languages recognized by the Government for use in State Legislature. So far, the VIII Schedule list 22 languages viz., Assamese, Bengali, Bodo, Dogri, Gujarathi, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu and Urdu. These languages represent all the four language families.

Data Source

The sample text analyzed for the present investigation was drawn from the Indian language written corpora developed by the Central Institute of Indian Languages, Mysore (Jayaram and Rajyashree, 2001). There are corpuses of 14 major Indian languages available. The size of the corpus varies from 1.5 million words to 3 million words depending on the availability of materials under different genres. The corpora were a general type meant to cater to multi-user covering all genres of the language. The criteria considered to categorize the language into different genres were informational, administrative, instructional & imaginative. Thus the data was collected under six main categories namely Aesthetics, Social Science, Natural Physical & Professional Sciences, Commerce, Official & Media Languages and Translated Materials which were further sub divided in to 76 text categories. The period of the corpora was restricted to one decade i.e. the text published between 1981 to 1990 were included and it represents the contemporary Indian language. This restriction ensured homogeneity of the text to some extent. The factors taken into consideration for selecting the period of one decade are the following. After independence the use of modern Indian languages was widened as they were used in domains like administration, judiciary, and higher education. Secondly, modern literary genera developed in most Indian languages around 1950 and the different languages have experienced various waves of modernity since then. After independence, when Indian languages assumed responsibilities as the official languages of the States, developmental programmes were taken to equip them with specialized vocabulary with coining words, translating or borrowing. About three decades saw hectic activity by both govern-

mental, semi and/or non-governmental agencies to develop different registers and to compile dictionaries, glossaries, thesaurus, encyclopedias and grammars. A considerable amount of literature was produced in different registers during this period. After about 30 years modernity got consolidated in terms of vocabulary and standardized use of language. Hence, the corpora were developed for the period mentioned above.

Method

The word length is measured in terms of syllables and the investigation is for the written languages. The syllables are counted in terms of number of vowels in the word. The word is defined orthographically representing between two spaces or separated by special characters like punctuation marks in the running text. Thus the words are categorized into monosyllabic, disyllabic, etc. The counting of words was realized based on the orthographic convention of each language. In some languages like Hindi, the case marker occurs as a postposition independently (eg. 'ghar me' – 'house in' as two words) while in some languages like Kannada, it occurs as a suffix of the word (eg. 'maneyalli' – 'house in' as one word). The compound words are counted as two different words. The foreign words given in different script is not counted for the present investigation.

Data for the present investigation

For the present investigation the languages analysed were 5 namely, Assamese, Bengali, Hindi, Marathi belonging to Indo-Aryan language family and Kannada & Tamil belonging to Dravidian language family. The text samples ranging from 6 to 38 were analysed drawing from each of the categories. The following table shows the sample text from different categories for the above-mentioned languages. In some languages, materials were not available for certain categories which are indicated by blank (-) in the table.

Table 2
Data and Categories

Sl. No.	Languages	Categories					
		Aesthetics	Commerce	Natural, Physical & Professional Sciences	Official and Media Languages	Social Sciences	Translated Material
1.	Assamese	30	-	30	30	30	30
2.	Hindi	30	18	30	28	-	24
3.	Kannada	30	38	30	-	30	-
4.	Marathi	30	20	38	11	30	14
5.	Tamil	30	6	30	30	30	16

Result and Discussion

The analysis of data was carried our using Altmann-Fitter software. Altmann-Fitter is an interactive program for fitting theoretical univariate discrete probability functions to empirical frequency distributions using methods of iterative optimization based on the simplex algorithm by Nelder-Mead. It is meant to be used by both, researchers as well as practical workers. Fitting starts with the common point estimates and is being optimized successively. Goodness of fit is determined by the chi-square test.

The following table (see Table 3) gives the overall pictures of the analysis of text samples from different languages drawn under various categories. It can be observed that out of the total samples analyzed (given in parenthesis), the majority of the samples follows a particular distribution. As was to be expected no single distribution fits across languages. While across categories there is a single distribution-fitting majority of the samples in Hindi (1-displaced extended positive binomial), Kannada (positive Cohen-Poisson) and Marathi (Dacey-negative binomial). In the case of Tamil, which belongs to Dravidian family, the samples from five categories fallow the Dacey-negative binomial. Though Marathi and Tamil belong to two different language families i.e. Indo-Arayan and Dravidian, the word length distribution follows the Dacey-negative binomial. The extreme case of variation is for the Assamese language where Dacey-Poisson fit samples from two categories. Dacey-negative binomial fit for two other categories and Extended positive binomial fit samples from the categories of official and Media languages. The reasons for these variations would be very interesting to probe further which is not the scope of the present paper.

Table 3
General results of fitting

Languages	**Categories**					
	Aesthetics	**Commerce**	**Natural Physical and Professional Sciences**	**Official and Media Languages**	**Social Sciences**	**Translated Material**
Assamese	28(30) Dacey Poisson	-	29(30) Dacey Negative Binomial	28(30) Extended Positive Binomial	29(30) Dacey Negative Binomial	29(30) Dacey Poisson
Hindi	30(30) Extended Positive Binomial	15(18) Extended Positive Binomial	29(30) Extended Positive Binomial	24(28) Extended Positive Binomial	-	21(23) Extended Positive Binomial
Kannada	30(30) Positive Cohen Poisson	25(38) Positive Cohen Poisson	29(30) Positive Cohen Poisson	-	28(30) Positive Cohen Poisson	-
Marathi	28(30) Dacey Negative Binomial	16(20) Dacey Negative Binomial	36(38) Dacey Negative Binomial	11(11) Dacey Negative Binomial	27(30) Dacey Negative Binomial	14(14) Dacey Negative Binomial
Tamil	29(30) Dacey Negative Binomial	5(6) Dacey Negative Binomial	30(30) Dacey Negative Binomial	24(30) Dacey Negative Binomial	30(30) Dacey Negative Binomial	14(16) Positive Cohen Poisson

In a multilingual country like India one automatically expects the existence of different substrates (e.g. Sanskrit), superstrates (e.g. Hindi and English) and both mutual areal as well as cultural influences. Hence for the distribution of word length in texts one cannot reckon with plain distribution models like binomial or Poisson, one must automatically take into acount the possibility of modifications. In the present data two tendencies can be found:

(i) Local modifications of a single frequency class giving rise to modified distributions, here the extended positive binomial (1) and the positive Cohen-Poisson (2). The formulas are:

$$(1) \qquad P_x = \begin{cases} 1-\alpha, & x=1 \\ \alpha \binom{n}{x-1} \dfrac{p^{x-1}q^{n-x+1}}{1-q^n}, & x = 2,3,...,n+1 \end{cases}$$

$$(2) \qquad P_x = \begin{cases} \dfrac{(1-\alpha)a}{e^a-1-\alpha a}, & x=1 \\ \dfrac{a^x}{x!(e^a-1-\alpha a)}, & x=2,3,4,... \end{cases}$$

In the first case (Formula (1)), the usual extended positive binomial has been displaced one step to the right. Only the first class is modified, the other classes are weighted in order to give a sum of 1. In the second case (Formula (2)), the first two classes (zero and one) of the Poisson distribution were added and weighted, the rest was weighted because of normalization. Cases like these appear in languages which earlier had zero-syllabic words which changed to clitics or prefixes (e.g. Slavic langauges)

(ii) If the influence of foreign strata is very strong, there are two word-length layers in the language which must be modeled by a superposition of distributions. In general we obtain a form

$$(3) \qquad P_x = \sum_{i=1}^{k} \alpha_i P^*_{xk}$$

where $\alpha_1 + \alpha_2 + \ldots + \alpha_k = 1$ and the k components of P^* can be either different distributions or equal distributions with different parameters. In our case the number of components is $k = 2$ in both cases, thus if $\alpha_1 = \alpha$, then $\alpha_2 = 1 - \alpha$. The resulting cases yield the so-called mixed Dacey-type distributions with two strata. The first stratum is a usual distribution weighted by 1- α, the second stratum is identical but shifted one step to the right and weighted by α. For the languages analyzed we obtained only two cases, namely the Dacey-Poisson distribution (4) and the Dacey-negative binomial distribution (5):

$$(4) \qquad P_x = \frac{(1-\alpha)a^{x-1}e^{-a}}{(x-1)!} + \frac{\alpha(x-1)a^{x-2}e^{-a}}{(x-1)!}, \; x=1,2,3,...$$

$$(5) \qquad P_x = (1-\alpha)\binom{k+x-2}{x-1}p^k q^{x-1} + \alpha\binom{k+x-3}{x-2}p^k q^{x-2}, \; x=1,2,...$$

Instead of presenting the fitting of all cases, we show merely one example, namely the fitting of the Dacey-Poisson distribution to Text 01 in aesthetics in Assamese as presented in Table 4.

Table 4
Fitting the Dacey-Poisson distribution to an Assamese text on aesthetics

x	f_x	NP_x (4)
1	303	295.60
2	855	904.47
3	879	820.60
4	417	421.04
5	141	149.29
6	40	40.32
7	6	8.78
8	1	1.89
N = 2642, a = 1.1078, α = 0.6612, DF = 5, X^2 = 8.85, P = 0.12, C = 0.0034		

The results for all texts in all languages are resumed in Tables 5 to 30.

Table 5
Fitting of the Dacey-Poisson distribution to word length data in Assamese, Aesthetics

Text No	N	a	α	DF	X^2	P	C
Text 01	2642	0.1078	0.6612	5	8.8515	0.1151	0.0034
Text 02	2417	1.1120	0.7214	5	8.7676	0.1187	0.0036
Text 03	1980	1.2676	0.6623	5	9.5402	0.0894	0.0048
Text 04	1964	1.3673	0.3540	4	25.1878	0.0000	0.0128
Text 05	2489	1.1048	0.5591	4	4.7097	0.3184	0.0019
Text 06	2051	0.8321	0.7132	3	22.8242	0.0000	0.0111
Text 07	2004	0.8758	0.7408	4	5.3055	0.2574	0.0026
Text 08	2297	0.9704	0.6021	4	12.7878	0.0124	0.0056
Text 09	2271	1.0365	0.6466	4	4.0572	0.3983	0.0018
Text 11	2852	1.1963	0.5223	4	15.6356	0.0035	0.0055
Text 12	1642	1.1316	0.5034	5	8.2123	0.1449	0.0050
Text 13	2856	1.0971	0.4523	4	3.8775	0.4228	0.0014
Text 14	2843	1.1542	0.5018	5	5.9764	0.3085	0.0021
Text 15	2887	1.0720	0.5114	5	14.2579	0.0141	0.0049
Text 16	3045	0.9741	0.5927	4	18.3884	0.0010	0.0060
Text 17	2783	1.1287	0.5457	5	15.9054	0.0071	0.0057
Text 19	2957	1.0203	0.6843	4	3.8442	0.4275	0.0013
Text 20	1819	1.0880	0.5866	5	4.2432	0.5150	0.0023
Text 21	2252	1.1665	0.4849	5	21.7032	0.0006	0.0096
Text 22	2381	1.1511	0.5416	5	20.0367	0.0012	0.0084
Text 23	2336	1.2136	0.5560	5	15.4616	0.0086	0.0066
Text 24	2981	1.0603	0.4675	4	10.4452	0.0336	0.0035
Text 25	2953	0.9864	0.4121	4	3.2372	0.5189	0.0011
Text 26	2404	1.1385	0.5948	4	20.9086	0.0003	0.0087
Text 27	2272	1.1502	0.5290	5	4.4661	0.4844	0.0020
Text 28	2485	0.9893	0.5761	4	1.8490	0.7635	0.0007
Text 29	2336	0.9935	0.5369	3	4.3652	0.2246	0.0019
Text 30	2214	1.0610	0.5215	5	16.0562	0.0067	0.0073

Table 6
Fitting of the Dacey-negative binomial distribution to word length data in Assamese, Natural Physical and Professional Sciences

Text No	N	k	p	α	DF	X^2	P	C
Text 01	2939	114.5813	0.9894	0.5555	4	11.9904	0.0174	0.0041
Text 02	2843	130.7329	0.9917	0.6284	3	16.6960	0.0008	0.0059
Text 03	2802	108.4727	0.9895	0.5543	3	7.9777	0.0465	0.0028
Text 04	2776	127.0135	0.9914	0.6201	3	1.9890	0.5747	0.0007
Text 05	3255	106.6741	0.9886	0.5343	4	4.2250	0.3764	0.0013
Text 06	3145	120.8904	0.9929	0.6431	3	18.1456	0.0004	0.0058
Text 07	2938	123.6812	0.9927	0.6451	3	34.0589	0.0000	0.0116
Text 08	2004	146.0929	0.9925	0.6694	4	20.3048	0.0004	0.0101
Text 09	2483	52.7465	0.9678	0.2574	5	17.1062	0.0043	0.0069
Text 10	2475	101.5781	0.9842	0.4500	5	8.5819	0.1269	0.0035
Text 11	2964	124.9948	0.9888	0.5552	5	14.9291	0.0107	0.0050
Text 12	2795	2.8728	0.7022	0.7022	5	21.2133	0.0007	0.0076
Text 13	2448	2.2719	0.6283	0.6283	7	26.1791	0.0005	0.0107
Text 14	2371	59.2513	0.9756	0.3106	5	16.8785	0.0047	0.0071
Text 15	2725	2.2559	0.6333	0.6743	6	9.2205	0.1616	0.0034
Text 16	2382	55.0231	0.9688	0.2553	5	20.0679	0.0012	0.0084
Text 17	2338	70.0857	0.9783	0.3446	5	21.7974	0.0006	0.0093
Text 18	2857	2.8805	0.6979	0.6979	6	38.2859	0.0000	0.0134
Text 19	3201	149.6734	0.9929	0.6797	4	10.8797	0.0279	0.0034
Text 20	2998	122.3918	0.9912	0.6054	4	15.4318	0.0039	0.0051
Text 21	3081	144.3942	0.9936	0.6925	1	19.2448	0.0000	0.0062
Text 22	1316	147.5703	0.9915	0.6481	4	4.0369	0.4010	0.0031
Text 24	2250	97.9029	0.9872	0.5068	3	26.8917	0.0000	0.0120
Text 25	1474	99.2491	0.9859	0.4857	4	11.6566	0.0201	0.0079
Text 26	1696	158.2766	0.9932	0.7021	3	11.0749	0.0113	0.0065
Text 27	1507	134.5813	0.9918	0.6366	4	19.9236	0.0005	0.0132
Text 28	1549	165.2941	0.9930	0.7019	3	2.9035	0.4067	0.0019
Text 29	1660	113.9750	0.9899	0.5640	3	6.3553	0.0955	0.0038
Text 30	1500	141.0789	0.9914	0.6374	4	11.5935	0.0206	0.0077

Table 7
Fitting the extended positive binomial distribution to word length data in Assamese: Official and Media Languages

Text No	N	n	p	α	DF	X^2	P	C
Text 01	2178	8	0.2633	0.9027	4	6.3644	0.1735	0.0029
Text 02	2385	6	0.2751	0.8264	2	2.2802	0.3198	0.0010
Text 03	2495	8	0.2101	0.8240	3	8.9355	0.0302	0.0036
Text 04	2301	8	0.2097	0.8101	3	16.5905	0.0009	0.0072
Text 05	1984	9	0.2155	0.8508	3	24.8456	0.0000	0.0125
Text 06	2029	9	0.2148	0.8590	3	17.9839	0.0004	0.0089
Text 08	2836	10	0.1894	0.8734	4	22.7110	0.0001	0.0080
Text 09	2432	9	0.1970	0.8507	3	19.1939	0.0002	0.0079

Text 11	2452	8	0.2161	0.8862	3	10.7168	0.0134	0.0044
Text 12	2238	6	0.2707	0.8699	2	24.9939	0.0000	0.0112
Text 13	2349	8	0.2103	0.8523	3	8.6131	0.0349	0.0037
Text 14	2703	11	0.1949	0.9142	4	29.8782	0.0000	0.0111
Text 15	2871	10	0.1936	0.9091	1	12.4967	0.0004	0.0044
Text 16	2267	9	0.2259	0.8937	4	18.9884	0.0008	0.0084
Text 17	2543	7	0.2476	0.8592	3	26.4066	0.0000	0.0104
Text 18	2090	9	0.2088	0.9249	3	17.0568	0.0007	0.0082
Text 19	3113	7	0.2336	0.8615	3	7.2302	0.0649	0.0023
Text 20	2813	8	0.2139	0.8685	3	19.0682	0.0003	0.0068
Text 21	1918	7	0.2496	0.8649	3	2.4298	0.4881	0.0013
Text 22	2238	7	0.2554	0.8731	3	27.4850	0.0000	0.0123
Text 23	2215	7	0.2646	0.9097	3	11.4720	0.0094	0.0052
Text 24	2316	6	0.2818	0.8541	2	2.4857	0.2886	0.0011
Text 25	2281	9	0.2131	0.8707	4	4.4431	0.3494	0.0019
Text 26	2263	8	0.2529	0.8754	3	22.4896	0.0001	0.0099
Text 27	2420	8	0.2383	0.8905	3	14.4909	0.0023	0.0060
Text 28	2269	9	0.2212	0.9074	4	17.0905	0.0019	0.0075
Text 29	2308	10	0.2127	0.9042	4	5.8875	0.2077	0.0026
Text 30	2311	9	0.2196	0.9156	2	33.2487	0.0000	0.0144

Table 8
Fitting the Dacey-negative binomial distribution to word length data in Assamese: Social Sciences

Text No	N	k	p	α	DF	X^2	P	C
Text 01	2537	32.5377	0.9442	0.1571	6	15.7110	0.0154	0.0062
Text 02	2029	166.4185	0.9902	0.6274	5	5.5254	0.3552	0.0027
Text 03	2403	117.1813	0.9847	0.4789	5	12.1539	0.0327	0.0051
Text 04	2501	4.0266	0.7193	0.7193	7	31.1376	0.0001	0.0125
Text 05	2503	165.0480	0.9902	0.6283	5	4.5860	0.4685	0.0018
Text 06	2483	61.6586	0.9689	0.2639	5	13.4252	0.0197	0.0054
Text 07	2698	138.5344	0.9882	0.5537	5	25.6117	0.0001	0.0095
Text 08	1712	144.7819	0.9905	0.6124	4	14.8803	0.0050	0.0087
Text 09	2876	125.0608	0.9887	0.5573	3	22.9891	0.0000	0.0080
Text 10	2813	127.7221	0.9901	0.5876	4	13.4742	0.0092	0.0048
Text 11	2913	120.3683	0.9892	0.5642	4	8.5108	0.0746	0.0029
Text 12	2718	131.4907	0.9902	0.5912	4	11.5301	0.0212	0.0042
Text 13	3078	133.7552	0.9897	0.5843	5	24.9357	0.0001	0.0081
Text 14	2534	123.2294	0.9881	0.5412	5	7.0196	0.2192	0.0028
Text 15	2863	146.1918	0.9911	0.6288	4	13.4546	0.0093	0.0047
Text 17	2198	160.2069	0.9922	0.6795	4	6.7602	0.1491	0.0031
Text 18	2528	87.3243	0.9856	0.4704	4	12.9765	0.0114	0.0051
Text 19	2561	99.5243	0.9892	0.5367	3	1.5135	0.6792	0.0006
Text 20	2596	109.6879	0.9894	0.5629	4	18.5864	0.0009	0.0072
Text 21	2636	90.7070	0.9855	0.4696	4	12.0861	0.0167	0.0046
Text 22	3088	126.5326	0.9904	0.5933	4	5.6530	0.2266	0.0018
Text 23	2980	2.4198	0.6777	0.6777	7	29.3354	0.0001	0.0098

Text 24	2640	59.1827	0.9748	0.3044	5	36.2620	0.0000	0.0137
Text 25	2637	99.1506	0.9868	0.4868	4	18.5162	0.0010	0.0070
Text 26	2873	137.7890	0.9917	0.6340	4	8.2322	0.0834	0.0029
Text 27	2715	151.5103	0.9928	0.6794	4	3.5893	0.4644	0.0013
Text 28	2847	130.9448	0.9909	0.6153	4	7.4973	0.1118	0.0026
Text 29	2829	147.7523	0.9925	0.6672	4	8.3303	0.0802	0.0029
Text 30	2671	131.0242	0.9890	0.5782	4	9.9477	0.0413	0.0037

Table 9
Fitting the Dacey-Poisson distribution to word length data in Assamese: Translated Material

Text No	N	a	α	DF	X^2	P	C
Text 01	2803	0.9131	0.6882	4	11.5601	0.0209	0.0041
Text 02	2797	0.9059	0.6788	3	12.8220	0.0050	0.0046
Text 03	2823	0.9782	0.6754	4	18.5737	0.0010	0.0066
Text 04	2614	0.9267	0.7073	4	14.6531	0.0055	0.0056
Text 05	2828	1.1634	0.4323	4	23.0698	0.0001	0.0082
Text 06	2740	1.2584	0.4080	5	39.6952	0.0000	0.0145
Text 07	2503	1.1902	0.3964	2	32.6492	0.0000	0.0130
Text 09	2154	1.0269	0.5862	4	20.7314	0.0004	0.0096
Text 10	1834	1.1611	0.4543	4	23.7835	0.0001	0.0130
Text 11	1585	1.3372	0.4599	5	14.9241	0.0107	0.0094
Text 12	2778	1.2490	0.4287	5	15.8397	0.0073	0.0057
Text 13	2773	1.2782	0.3999	5	5.3351	0.3764	0.0019
Text 14	2747	1.2313	0.4556	5	24.4279	0.0002	0.0089
Text 15	2748	1.0828	0.5096	5	16.2149	0.0063	0.0059
Text 16	2883	1.1077	0.5065	5	15.2448	0.0094	0.0053
Text 17	2593	1.1123	0.4637	4	30.3789	0.0000	0.0117
Text 18	1857	1.1459	0.4249	4	23.5836	0.0001	0.0127
Text 19	1987	1.2134	0.4116	4	9.5563	0.0486	0.0048
Text 20	2929	1.1738	0.4556	5	7.0982	0.2134	0.0024
Text 21	2978	1.1469	0.4511	5	8.0017	0.1561	0.0027
Text 22	2885	1.0452	0.5123	5	8.5982	0.1262	0.0030
Text 23	3005	1.0649	0.5141	5	14.0221	0.0155	0.0047
Text 24	3072	1.0972	0.4887	4	21.4913	0.0003	0.0070
Text 25	3030	1.1173	0.4714	5	22.0951	0.0005	0.0073
Text 26	3042	1.0783	0.4446	5	41.3566	0.0000	0.0136
Text 27	2238	1.2268	0.4952	5	28.8297	0.0000	0.0129
Text 28	2846	1.1592	0.5504	4	20.5107	0.0004	0.0072
Text 29	2814	1.0968	0.5603	4	20.3354	0.0004	0.0072
Text 30	2695	1.1478	0.5314	5	14.9429	0.0106	0.0055

Table 10
Fitting the extended positive binomial distribution to word length data in Hindi: Aesthetics

Text No	N	n	p	α	DF	X^2	P	C
Text 01	2520	5	0.1893	0.7365	1	25.0063	0.0000	0.0099
Text 02	2188	7	0.1605	0.7212	2	10.6481	0.0049	0.0049
Text 03	1617	4	0.2624	0.7520	1	1.7118	0.1907	0.0011
Text 04	2235	5	0.2328	0.7266	2	10.4181	0.0055	0.0047
Text 05	1882	4	0.1987	0.7539	1	6.4248	0.0113	0.0034
Text 06	1933	4	0.2442	0.7605	1	0.2423	0.6225	0.0001
Text 07	2356	5	0.1945	0.7097	1	11.5839	0.0007	0.0049
Text 08	2391	4	0.2648	0.7206	1	23.6205	0.0000	0.0099
Text 09	2317	6	0.1894	0.7583	2	19.2695	0.0001	0.0083
Text 10	1887	11	0.1319	0.7266	3	15.7436	0.0013	0.0083
Text 11	2775	8	0.1618	0.7239	2	4.4411	0.1086	0.0016
Text 12	2347	29	0.0484	0.7503	3	29.8625	0.0000	0.0127
Text 13	3293	4	0.2737	0.7883	1	7.1992	0.0073	0.0022
Text 14	3167	5	0.2290	0.7186	2	1.5717	0.4557	0.0005
Text 15	2588	6	0.2176	0.7159	2	6.7573	0.0341	0.0026
Text 16	2706	43	0.0315	0.6869	3	9.3652	0.0248	0.0035
Text 17	2773	6	0.1977	0.7364	2	3.2344	0.1985	0.0012
Text 18	2428	6	0.1914	0.7315	2	15.5099	0.0004	0.0064
Text 19	2685	5	0.3222	0.7695	2	10.8555	0.0044	0.0040
Text 20	2445	12	0.1196	0.7264	3	7.2206	0.0652	0.0030
Text 21	2509	7	0.1833	0.7329	2	28.2840	0.0000	0.0113
Text 22	2699	7	0.1828	0.7362	2	10.1072	0.0064	0.0037
Text 23	2500	5	0.1571	0.7400	1	0.0268	0.8700	0.0000
Text 24	2588	6	0.1934	0.7527	2	16.3350	0.0003	0.0063
Text 25	2412	7	0.1885	0.7276	2	1.0975	0.5777	0.0005
Text 26	2599	5	0.2616	0.7372	2	15.2119	0.0005	0.0059
Text 27	2237	7	0.2027	0.7488	2	8.8301	0.0121	0.0039
Text 28	2403	5	0.2125	0.7149	1	11.0401	0.0009	0.0046
Text 29	2203	9	0.1548	0.7249	3	5.7308	0.1255	0.0026
Text 30	2547	7	0.1839	0.7452	2	11.4456	0.0033	0.0045

Table 11
Fitting the extended positive binomial distribution to word length data in Hindi: Commerce

Text No	N	n	p	α	DF	X^2	P	C
Text 01	2421	10	0.1607	0.7600	3	3.0476	0.3844	0.0013
Text 02	2014	7	0.1866	0.7597	2	9.5698	0.0084	0.0048
Text 04	2344	7	0.1608	0.7124	1	26.8994	0.0000	0.0115
Text 05	2111	7	0.1746	0.7205	2	3.1688	0.2051	0.0015
Text 06	2419	7	0.1812	0.7325	2	2.3277	0.3123	0.0010
Text 08	2477	8	0.1808	0.7513	3	15.2442	0.0016	0.0062
Text 09	1723	6	0.3207	0.9611	2	24.0613	0.0000	0.0140

Text 10	2944	7	0.1446	0.7442	2	12.9365	0.0016	0.0044
Text 11	2138	7	0.2071	0.7558	2	13.8759	0.0010	0.0065
Text 12	2186	7	0.1979	0.7259	2	11.1707	0.0038	0.0051
Text 13	2803	7	0.1816	0.7367	2	17.8897	0.0001	0.0064
Text 14	4320	9	0.1506	0.7275	3	17.6389	0.0005	0.0041
Text 15	1849	7	0.2089	0.7355	2	2.9123	0.2331	0.0016
Text 17	2550	8	0.1681	0.7286	3	5.9367	0.1147	0.0023
Text 18	2796	9	0.1508	0.7332	3	9.2104	0.0266	0.0033

Table 12
Fitting the extended positive binomial distribution to word length data in Hindi: Natural Physical and Professional Sciences

Text No	N	n	p	α	DF	X^2	P	C
Text 01	2550	10	0.1406	0.7518	3	4.3417	0.2269	0.0017
Text 02	2416	7	0.1648	0.7285	2	13.2939	0.0013	0.0055
Text 03	2024	9	0.1371	0.7204	2	13.8094	0.0010	0.0068
Text 04	3501	7	0.2233	0.7615	3	1.4988	0.6826	0.0004
Text 05	2733	6	0.1971	0.7417	2	32.9802	0.0000	0.0121
Text 06	2344	7	0.1959	0.7031	2	6.3205	0.0424	0.0027
Text 07	3155	16	0.0816	0.7322	3	32.2342	0.0000	0.0102
Text 08	2420	6	0.2084	0.7343	2	5.7380	0.0568	0.0024
Text 09	2699	8	0.1579	0.7347	2	23.1270	0.0000	0.0086
Text 10	2569	6	0.1737	0.7213	2	7.9229	0.0190	0.0031
Text 11	2599	8	0.1555	0.7079	2	26.5879	0.0000	0.0102
Text 12	2732	7	0.1759	0.7130	2	25.1814	0.0000	0.0092
Text 13	2532	10	0.1407	0/7022	3	6.8171	0.0780	0.0027
Text 14	2663	8	0.1163	0.7191	2	28.8589	0.0000	0.0108
Text 15	2588	11	0.1167	0.7276	3	23.0360	0.0000	0.0089
Text 16	2550	10	0.1406	0.7518	3	4.3417	0.2269	0.0017
Text 17	2416	7	0.1648	0.7285	2	13.2939	0.0013	0.0055
Text 19	3139	6	0.1787	0.7104	2	3.7718	0.1517	0.0012
Text 20	2917	15	0.0925	0.7192	3	16.1798	0.0010	0.0055
Text 21	3292	15	0.0983	0.7163	3	21.6546	0.0001	0.0066
Text 22	5606	7	0.1875	0.7374	3	29.7198	0.0000	0.0053
Text 23	3536	5	0.2315	0.7248	2	13.8160	0.0010	0.0039
Text 24	2249	7	0.1500	0.7412	1	14.1527	0.0002	0.0063
Text 25	2233	6	0.1873	0.7376	2	24.3056	0.0000	0.0109
Text 26	2820	8	0.1547	0.7099	2	10.8375	0.0044	0.0038
Text 27	2338	8	0.1538	0.7318	2	25.5584	0.0000	0.0109
Text 28	2824	7	0.1868	0.7064	2	5.2724	0.0716	0.0019
Text 29	2781	9	0.1339	0.7098	2	6.5330	0.0381	0.0023
Text 30	3131	10	0.1422	0.7288	3	27.1460	0.0000	0.0087

Table 13
Fitting the extended positive binomial distribution to word length data in Hindi: Official and Media languages

Text No	N	n	p	α	DF	X^2	P	C
Text 02	1539	7	0.2056	0.7290	2	0.1780	0.9149	0.0001
Text 03	2287	9	0.1605	0.7188	2	1.1605	0.5598	0.0005
Text 04	1900	12	0.1148	0.7126	3	26.1716	0.0000	0.0138
Text 05	3082	14	0.1072	0.7323	3	8.2241	0.0416	0.0027
Text 07	2239	8	0.1900	0.7414	3	16.3897	0.0009	0.0073
Text 09	2185	8	0.1932	0.7483	1	32.8265	0.0000	0.0150
Text 10	1943	15	0.1179	0.7452	4	14.2418	0.0066	0.0073
Text 11	2668	7	0.2381	0.7586	3	33.7840	0.0000	0.0127
Text 12	2195	7	0.2403	0.7658	3	2.2604	0.5201	0.0010
Text 13	2100	11	0.1518	0.7595	3	14.7676	0.0020	0.0070
Text 14	2390	9	0.1986	0.7983	3	13.1520	0.0043	0.0055
Text 15	2407	8	0.1656	0.7461	2	2.8188	0.2443	0.0012
Text 16	2340	8	0.1572	0.7321	2	10.4338	0.0054	0.0045
Text 17	2367	10	0.1553	0.7520	3	25.0829	0.0000	0.0106
Text 18	2363	7	0.2439	0.7753	3	27.2178	0.0000	0.0115
Text 19	2491	11	0.1485	0.7916	2	37.3400	0.0000	0.0150
Text 20	2608	12	0.1316	0.7416	3	26.8905	0.0000	0.0103
Text 23	3706	7	0.1754	0.7591	2	38.1448	0.0000	0.0103
Text 24	1438	22	0.0557	0.7483	3	14.1728	0.0027	0.0099
Text 25	5905	12	0.0986	0.7145	3	53.4961	0.0000	0.0091
Text 26	1273	5	0.2632	0.7235	2	5.0702	0.0793	0.0040
Text 27	1539	7	0.2056	0.7290	2	0.1780	0.9149	0.0001
Text 28	2287	9	0.1605	0.7188	2	0.1605	0.5598	0.0005
Text 29	1988	6	0.2001	0.7042	2	9.4070	0.0091	0.0047

Table 14
Fitting the extended positive binomial distribution to word length data in Hindi: Translated Material

Text No	N	n	p	α	DF	X^2	P	C
Text 01	2144	6	0.1788	0.7351	2	20.5636	0.0000	0.0096
Text 02	2374	5	0.1591	0.7405	1	7.8681	0.0050	0.0033
Text 03	1976	6	0.1501	0.7055	1	16.6641	0.0000	0.0084
Text 04	2612	7	0.1400	0.7619	2	13.2374	0.0013	0.0051
Text 06	6419	7	0.1552	0.7378	2	8.2802	0.0159	0.0013
Text 07	1934	6	0.1795	0.7053	2	15.2883	0.0005	0.0079
Text 09	2745	6	0.1609	0.7042	2	4.7338	0.0938	0.0017
Text 10	2678	5	0.1796	0.7729	1	13.4340	0.0002	0.0050
Text 11	2001	7	0.1622	0.7196	2	7.0711	0.0291	0.0035
Text 12	2349	6	0.1876	0.7479	2	6.2681	0.0435	0.0027
Text 13	4735	7	0.1429	0.7337	2	47.8527	0.0000	0.0101
Text 14	1655	7	0.1815	0.7057	2	2.8170	0.2445	0.0017
Text 15	2856	7	0.1343	0.7510	2	34.2949	0.0000	0.0120

Text 16	2118	6	0.1792	0.7035	2	15.9399	0.0003	0.0075
Text 17	6869	6	0.1923	0.7409	2	76.5230	0.0000	0.0111
Text 18	2125	6	0.1495	0.7449	1	18.0849	0.0000	0.0085
Text 19	2745	6	0.1609	0.7042	2	4.7338	0.0938	0.0017
Text 21	9123	6	0.1864	0.7313	2	61.4007	0.0000	0.0067
Text 22	2381	9	0.1379	0.7329	2	21.2640	0.0000	0.0089
Text 23	2163	13	0.0767	0.7369	1	19.6025	0.0000	0.0091
Text 26	1939	5	0.2121	0.7607	1	1.7806	0.1821	0.0009

Table 15
Fitting the positive Cohen-Poisson distribution to word length data in Kannada: Aesthetics

Text No	N	a	α	DF	X^2	P	C
Text 01	5015	2.7595	0.8742	7	16.9011	0.0180	0.0034
Text 02	4910	3.3461	0.8326	9	21.9570	0.0090	0.0045
Text 03	4847	3.3350	0.8119	9	22.9443	0.0063	0.0047
Text 04	4886	3.4569	0.8472	9	15.3402	0.0820	0.0031
Text 05	5076	3.3956	0.8138	9	9.7418	0.3718	0.0019
Text 06	4978	3.2475	0.9022	8	5.1771	0.7385	0.0010
Text 07	4795	3.5406	0.5995	9	45.6798	0.0000	0.0095
Text 08	5433	3.2521	0.8104	8	22.8436	0.0036	0.0042
Text 09	6982	3.3229	0.6125	7	39.3169	0.0000	0.0056
Text 10	7085	3.3364	0.7897	9	23.6881	0.0048	0.0033
Text 11	5752	3.3139	0.7372	9	29.2232	0.0006	0.0051
Text 12	5117	2.5930	0.9432	7	20.4822	0.0046	0.0040
Text 13	1195	2.7514	0.9489	6	6.5121	0.3683	0.0054
Text 14	5351	2.7910	0.9035	7	26.6957	0.0004	0.0050
Text 15	5008	3.3119	0.8377	8	18.0836	0.0206	0.0036
Text 16	4926	3.3792	0.7891	9	18.4069	0.0307	0.0037
Text 17	4912	3.3298	0.7805	8	20.9926	0.0072	0.0043
Text 18	4952	2.5442	0.9498	7	3.8911	0.7922	0.0008
Text 19	5020	2.6225	0.9247	7	19.7394	0.0062	0.0039
Text 20	4691	2.8366	0.9236	7	44.2054	0.0000	0.0094
Text 21	4339	2.6694	0.9320	7	16.7488	0.0191	0.0039
Text 22	557	3.2197	0.6182	7	10.5154	0.1612	0.0189
Text 23	5076	3.1880	0.8562	8	31.9098	0.0001	0.0063
Text 24	5028	3.3409	0.7942	9	27.0055	0.0014	0.0054
Text 25	2693	3.2724	0.8622	8	14.9522	0.0601	0.0056
Text 26	1540	3.2308	0.8665	7	15.4861	0.0302	0.0101
Text 27	1666	3.1724	0.9173	7	7.6764	0.3620	0.0046
Text 28	1943	3.5572	0.7322	8	8.3907	0.3963	0.0043
Text 29	4939	3.3146	0.7091	8	4.2381	0.8350	0.0009
Text 30	5051	3.2133	0.7342	8	21.7259	0.0054	0.0043

Table 16
Fitting the positive Cohen-Poisson distribution to word length data in Kannada: Commerce

Text No	N	a	α	DF	X^2	P	C
Text 02	4733	3.8957	0.8694	10	53.3811	0.0000	0.0113
Text 03	4824	3.5648	0.8538	5	61.8179	0.0000	0.0128
Text 04	4728	3.4529	0.7977	9	45.9144	0.0000	0.0097
Text 05	4840	3.2990	0.7359	8	11.9249	0.1546	0.0025
Text 07	4937	3.4697	0.7979	9	58.3200	0.0000	0.0118
Text 09	5011	3.4036	0.6345	9	74.7270	0.0000	0.0149
Text 10	4100	3.2417	0.8172	8	24.4280	0.0019	0.0060
Text 11	2442	3.3556	0.7282	8	34.5436	0.0000	0.0141
Text 13	1602	3.4809	0.7957	8	12.6990	0.1226	0.0079
Text 14	4947	3.6921	0.7169	9	25.6221	0.0024	0.0052
Text 15	4715	3.3594	0.7921	5	69.0318	0.0000	0.0146
Text 20	4705	3.8075	0.7239	9	18.2285	0.0326	0.0039
Text 22	1550	3.6656	0.8901	8	16.7396	0.0329	0.0108
Text 23	4947	3.5402	0.8141	6	74.1148	0.0000	0.0150
Text 24	4902	3.5286	0.8315	9	14.3855	0.1093	0.0029
Text 26	5066	3.3535	0.7424	9	52.9363	0.0000	0.0104
Text 27	3191	3.4854	0.7948	7	46.6127	0.0000	0.0146
Text 28	2602	3.3450	0.8061	8	17.8427	0.0224	0.0069
Text 29	1482	3.1425	0.6373	7	9.4098	0.2246	0.0063
Text 30	5005	3.3858	0.4831	9	56.8116	0.0000	0.0114
Text 31	3243	3.3858	0.5490	8	31.3881	0.0001	0.0097
Text 32	3853	3.3800	0.7858	8	41.0335	0.0000	0.0106
Text 35	2855	3.3698	0.8411	8	27.3381	0.0006	0.0096
Text 36	3341	3.8162	0.6003	9	15.0607	0.0893	0.0045
Text 38	3903	3.6049	0.8333	6	58.3718	0.0000	0.0150

Table 17
Fitting the positive Cohen-Poisson distribution to word length data in Kannada:
Natural Physical and Professional Sciences

Text No	N	a	α	DF	X^2	P	C
Text 01	3032	3.2735	0.8292	8	23.7790	0.0025	0.0078
Text 02	3940	3.1504	0.9097	7	59.1357	0.0000	0.0150
Text 03	1649	2.9414	0.9628	5	23.8495	0.0002	0.0145
Text 04	603	3.3561	0.8286	7	7.5694	0.3721	0.0126
Text 05	3219	3.5126	0.8246	9	39.9246	0.0000	0.0124
Text 06	2745	3.4065	0.9082	8	17.4062	0.0261	0.0063
Text 07	2931	3.3201	0.9274	8	14.8409	0.0623	0.0051
Text 08	5919	3.2154	0.7759	8	39.2450	0.0000	0.0066
Text 09	4287	3.4049	0.7506	4	59.0380	0.0000	0.0138
Text 10	4729	3.5706	0.7947	9	52.3216	0.0000	0.0111
Text 11	5008	3.0452	0.8292	8	7.5128	0.4824	0.0015
Text 12	4985	3.1381	0.8694	8	21.6373	0.0057	0.0043
Text 13	4664	2.2430	0.9080	6	18.0838	0.0060	0.0039

Text 14	4211	3.2738	0.8842	8	19.8669	0.0109	0.0047
Text 15	4946	3.7618	0.3270	9	69.4515	0.0000	0.0140
Text 16	8838	3.4711	0.4147	9	127.0816	0.0000	0.0144
Text 17	4994	3.4650	0.5466	8	25.2465	0.0014	0.0051
Text 18	4575	3.4820	0.6185	9	17.4516	0.0421	0.0038
Text 19	6583	3.3934	0.7970	9	29.0337	0.0006	0.0044
Text 20	5159	3.4668	0.7306	9	61.1424	0.0000	0.0119
Text 21	5340	3.4496	0.7831	9	16.8214	0.0516	0.0032
Text 22	1730	3.3580	0.8207	8	21.0442	0.0070	0.0122
Text 23	3647	3.1128	0.8716	5	53.5528	0.0000	0.0147
Text 25	4966	3.7548	0.7840	9	21.3847	0.0110	0.0043
Text 26	4901	3.5146	0.7696	9	16.6347	0.0548	0.0034
Text 27	4715	3.7655	0.5443	9	36.4752	0.0000	0.0077
Text 28	5285	3.0681	0.8648	7	23.9991	0.0017	0.0044
Text 29	1482	3.1425	0.6373	7	9.4098	0.2246	0.0063
Text 30	5005	3.3858	0.4831	9	56.8116	0.0000	0.0114

Table 18
Fitting the positive Cohen-Poisson distribution to word length data in Kannada: Social Sciences

Text No	N	a	α	DF	X^2	P	C
Text 01	4938	3.4095	0.8751	9	10.4856	0.3126	0.0021
Text 02	4710	3.5504	0.7195	9	69.4340	0.0000	0.0147
Text 03	3009	3.2233	0.5153	3	45.0362	0.0000	0.0150
Text 04	4828	3.1992	0.8535	8	10.9697	0.2034	0.0023
Text 05	4482	3.5716	0.7565	9	16.5671	0.0559	0.0037
Text 07	4200	3.3009	0.8846	8	56.6272	0.0000	0.0135
Text 08	4402	3.1813	0.8778	8	49.5978	0.0000	0.0113
Text 09	3746	3.6673	0.7878	9	41.8780	0.0000	0.0112
Text 10	4999	3.0913	0.9129	8	53.7422	0.0000	0.0108
Text 11	2956	3.6626	0.7295	9	23.9607	0.0044	0.0081
Text 12	4938	3.4095	0.8751	9	10.4856	0.3126	0.0021
Text 13	611	3.8173	0.7280	8	12.7334	0.1214	0.0208
Text 14	4710	3.7070	0.7615	9	19.2570	0.0231	0.0041
Text 15	4488	3.2907	0.8569	8	33.7767	0.0000	0.0075
Text 16	4848	3.5809	0.8760	8	17.8613	0.0223	0.0037
Text 17	3642	3.2415	0.9029	8	12.6785	0.1234	0.0035
Text 18	5022	3.9357	0.8892	10	33.4145	0.0002	0.0067
Text 19	5087	3.9809	0.8196	9	49.7064	0.0000	0.0098
Text 20	5018	3.8869	0.9062	10	28.7256	0.0014	0.0057
Text 21	1369	3.2573	0.5465	7	20.3833	0.0048	0.0149
Text 22	4974	3.6117	0.7422	9	68.9132	0.0000	0.0139
Text 23	4876	3.6052	0.7829	9	28.4841	0.0008	0.0058
Text 25	2216	3.0636	0.8959	7	29.5148	0.0001	0.0133
Text 26	4918	3.6020	0.8532	9	29.3543	0.0006	0.0060
Text 27	4957	3.4224	0.8646	9	11.8164	0.2239	0.0024
Text 28	4689	3.3243	0.9040	8	19.8396	0.0110	0.0042

Text 29	3672	3.3851	0.9225	8	19.3230	0.0132	0.0053
Text 30	4960	2.8355	0.8883	7	36.9916	0.0000	0.0075

Table 19
Fitting the Dacey-negative binomial distribution to word length data in Marathi: Aesthetics

Text No	N	k	p	α	DF	X^2	P	C
Text 01	3575	199.3005	0.9935	0.7480	4	12.9883	0.0113	0.0036
Text 02	4238	3.5538	0.7709	0.7709	5	29.1700	0.0000	0.0069
Text 03	2971	4.4761	0.8048	0.8048	5	25.8010	0.0001	0.0087
Text 04	2763	194.9526	0.9939	0.7509	4	7.6821	0.1039	0.0028
Text 05	4693	196.8671	0.9940	0.7558	4	18.4166	0.0010	0.0039
Text 06	4864	3.7483	0.7693	0.7693	6	31.9182	0.0000	0.0066
Text 07	4652	6.3328	0.8439	0.8439	6	15.3734	0.0175	0.0033
Text 08	4282	182.7836	0.9936	0.7348	4	9.2785	0.0545	0.0022
Text 09	2695	3.6235	0.7787	0.7787	5	9.5713	0.0883	0.0036
Text 10	2791	5.5239	0.8263	0.8263	4	16.5591	0.0024	0.0059
Text 11	5203	173.3739	0.9938	0.7274	4	65.3083	0.0000	0.0126
Text 12	3058	164.4206	0.9930	0.7000	4	23.3199	0.0001	0.0076
Text 13	1975	198.9929	0.9938	0.7538	4	9.9462	0.0413	0.0050
Text 14	4329	195.7899	0.9941	0.7588	4	8.0831	0.0886	0.0019
Text 15	2821	4.6478	0.8178	0.8178	5	20.5310	0.0010	0.0073
Text 16	2731	4.1406	0.7675	0.7675	6	34.9128	0.0000	0.0128
Text 17	4484	157.0607	0.9920	0.6645	5	16.6355	0.0052	0.0037
Text 18	3718	4.3963	0.7793	0.7793	6	22.1552	0.0011	0.0060
Text 19	2725	5.4810	0.8401	0.8401	4	21.0313	0.0003	0.0077
Text 20	2781	192.4915	0.9935	0.7437	4	8.2672	0.0823	0.0030
Text 21	2880	5.7876	0.8306	0.8306	5	11.9334	0.0357	0.0041
Text 22	4062	160.1345	0.9934	0.7034	4	38.1616	0.0000	0.0094
Text 24	2302	238.2751	0.9952	0.8290	4	12.5998	0.0134	0.0055
Text 25	3495	233.8528	0.9950	0.8202	2	37.3558	0.0000	0.0107
Text 26	4433	4.0631	0.7630	0.7630	6	44.8298	0.0000	0.0101
Text 27	3593	189.6630	0.9934	0.7359	4	23.6153	0.0001	0.0066
Text 28	3240	55.1150	0.9806	0.3630	4	13.5629	0.0088	0.0042
Text 29	3437	7.0110	0.8501	0.8501	5	20.5551	0.0010	0.0060
Text 30	3854	179.9117	0.9929	0.7147	5	34.5014	0.0000	0.0090

Table 20
Fitting the Dacey-negative binomial distribution to word length data in Marathi: Commerce

Text No	N	k	p	α	DF	X^2	P	C
Text 01	3031	7.2511	0.8416	0.8416	6	19.5200	0.0034	0.0064
Text 02	3462	190.4193	0.9920	0.7056	5	40.4473	0.0000	0.0117
Text 03	4894	136.0278	0.9877	0.5528	5	35.0791	0.0000	0.0072
Text 04	563	151.3821	0.9901	0.6200	1	0.6771	0.4106	0.0012
Text 05	3078	6.6697	0.8300	0.8300	4	39.5820	0.0000	0.0129
Text 06	3733	181.9666	0.9920	0.6976	5	50.4328	0.0000	0.0135
Text 08	3035	5.2107	0.7732	0.7732	6	44.5443	0.0000	0.0147

Text 09	2810	6.1576	0.8208	0.8208	3	41.4295	0.0000	0.0147
Text 10	2874	5.3757	0.8056	0.8056	6	36.0919	0.0000	0.0126
Text 11	3573	144.5615	0.9895	0.6107	4	53.5497	0.0000	0.0150
Text 15	4214	186.2290	0.9925	0.7145	5	22.7026	0.0004	0.0054
Text 16	5264	188.0865	0.9943	0.7563	4	27.1116	0.0000	0.0052
Text 17	3552	4.4065	0.7866	0.7866	6	30.3870	0.0000	0.0086
Text 18	4183	164.2353	0.9921	0.6763	5	4.2483	0.5142	0.0010
Text 19	4246	176.0013	0.9928	0.7035	5	59.7549	0.0000	0.0141
Text 20	3635	166.2706	0.9913	0.6629	5	35.7103	0.0000	0.0098

Table 21
Fitting the Dacey-negative binomial distribution to word length data in Marathi: Natural Physical and Professional Sciences

Text No	N	k	p	α	DF	X^2	P	C
Text 01	3067	130.6534	0.9885	0.5636	4	8.7921	0.0665	0.0029
Text 02	5490	171.1501	0.9922	0.6986	5	80.8534	0.0000	0.0147
Text 03	3613	202.7398	0.9942	0.7737	4	24.5335	0.0001	0.0068
Text 04	3522	161.8243	0.9914	0.6640	5	42.6855	0.0000	0.0121
Text 05	3522	161.8411	0.9914	0.6641	5	43.0885	0.0000	0.0122
Text 06	3094	176.2587	0.9921	0.6860	5	41.9739	0.0000	0.0136
Text 07	4778	175.4391	0.9913	0.6748	4	67.2873	0.0000	0.0141
Text 08	4121	160.0737	0.9899	0.6309	5	46.0863	0.0000	0.0112
Text 09	3920	165.4624	0.9908	0.6446	5	28.9604	0.0000	0.0074
Text 10	5596	176.7545	0.9924	0.6975	5	10.6531	0.0587	0.0019
Text 11	3644	161.5556	0.9903	0.6426	5	46.4524	0.0000	0.0127
Text 13	3619	175.1490	0.9908	0.6602	2	54.0317	0.0000	0.0149
Text 14	3075	129.2178	0.9891	0.5726	5	7.9462	0.1592	0.0026
Text 15	4784	203.3965	0.9949	0.7904	4	43.0515	0.0000	0.0090
Text 16	4920	215.2932	0.9944	0.7860	5	8.2674	0.1421	0.0017
Text 17	4734	170.5713	0.9926	0.6916	5	27.5951	0.0000	0.0058
Text 18	3548	180.5306	0.9942	0.7444	4	19.2044	0.0007	0.0054
Text 19	3333	173.0849	0.9929	0.7108	4	12.1474	0.0163	0.0036
Text 20	955	180.4575	0.9936	0.7350	2	11.8272	0.0027	0.0124
Text 21	4161	154.2350	0.9907	0.6330	5	26.7610	0.0001	0.0064
Text 22	4830	184.0817	0.9942	0.7544	4	43.2714	0.0000	0.0090
Text 23	4429	182.9561	0.9927	0.7151	5	49.2314	0.0000	0.0111
Text 24	3247	169.6876	0.9926	0.6994	4	11.2244	0.0242	0.0035
Text 25	4215	200.5095	0.9935	0.7491	5	7.8978	0.1620	0.0019
Text 26	6402	87.3398	0.9787	0.3851	6	51.6325	0.0000	0.0081
Text 27	5081	232.0804	0.9942	0.7976	5	13.4040	0.0199	0.0026
Text 28	3781	143.3506	0.9899	0.6117	4	19.5318	0.0006	0.0052
Text 29	3234	155.5328	0.9913	0.6596	5	37.6799	0.0000	0.0117
Text 31	6288	179.6414	0.9921	0.6903	5	39.8220	0.0000	0.0063
Text 32	5266	121.8066	0.9869	0.5261	5	22.4340	0.0004	0.0043
Text 33	1341	161.7419	0.9913	0.6514	4	8.1341	0.0868	0.0061
Text 34	3492	184.0155	0.9929	0.7157	5	18.3563	0.0025	0.0053
Text 35	5405	172.3715	0.9929	0.7026	5	16.6273	0.0053	0.0031

Text 36	4693	87.4341	0.9803	0.3949	5	16.0496	0.0067	0.0034
Text 37	4842	187.2937	0.9931	0.7275	5	26.2636	0.0001	0.0054
Text 38	4059	167.1963	0.9920	0.6749	5	24.2841	0.0002	0.0060

Table 22
Fitting the Dacey-negative binomial distribution to word length data in Marathi: Official and Media Languages

Text No	N	k	p	α	DF	X^2	P	C
Text 01	4163	125.3526	0.9892	0.5684	5	18.1992	0.0027	0.0044
Text 02	4925	149.7798	0.9920	0.6582	4	43.7517	0.0000	0.0089
Text 03	4149	124.8937	0.9863	0.5033	6	16.3635	0.0119	0.0039
Text 04	4055	127.2282	0.9894	0.5746	4	11.5381	0.0211	0.0028
Text 05	4828	123.8570	0.9888	0.5557	5	24.0707	0.0002	0.0050
Text 06	3432	150.4144	0.9897	0.5941	5	17.6826	0.0034	0.0052
Text 07	3943	3.2937	0.7308	0.7308	6	31.7381	0.0000	0.0080
Text 08	4984	136.0585	0.9904	0.6009	5	28.2252	0.0000	0.0057
Text 09	3939	127.8267	0.9895	0.5753	5	12.8750	0.0246	0.0033
Text 10	4665	164.0989	0.9927	0.6862	4	20.0266	0.0005	0.0043
Text 11	3723	131.5840	0.9900	0.5917	5	6.9477	0.2246	0.0019

Table 23
Fitting the Dacey-negative binomial distribution to word length data in Marathi: Social Sciences

Text No	N	k	p	α	DF	X^2	P	C
Text 01	3429	173.1657	0.9926	0.7019	2	37.9758	0.0000	0.0111
Text 02	4762	214.1586	0.9938	0.7718	5	30.3643	0.0000	0.0064
Text 03	4709	120.8668	0.9846	0.4736	6	58.3113	0.0000	0.0124
Text 04	4016	151.1630	0.9909	0.6382	5	15.4929	0.0085	0.0039
Text 05	3445	156.6024	0.9895	0.6150	5	19.7484	0.0014	0.0057
Text 06	4336	183.1277	0.9913	0.6752	5	43.3188	0.0000	0.0100
Text 07	5454	4.4817	0.7710	0.7710	7	22.2718	0.0023	0.0041
Text 08	3908	174.0969	0.9924	0.7017	5	30.3736	0.0000	0.0078
Text 09	4365	159.7849	0.9916	0.6608	4	8.6537	0.0704	0.0020
Text 10	4327	5.2268	0.7748	0.7748	7	34.4138	0.0000	0.0080
Text 11	4167	205.5090	0.9936	0.7600	5	26.1632	0.0001	0.0063
Text 13	3854	184.5666	0.9915	0.6893	5	46.1552	0.0000	0.0120
Text 15	2067	235.0856	0.9957	0.8379	4	12.6204	0.0133	0.0061
Text 16	1389	217.9567	0.9951	0.8093	3	6.5219	0.0888	0.0047
Text 17	2827	234.3373	0.9955	0.8300	4	6.2722	0.1797	0.0022
Text 18	4964	173.7051	0.9915	0.6738	5	36.7160	0.0000	0.0074
Text 19	3873	4.3254	0.7722	0.7722	4	34.8027	0.0000	0.0090
Text 20	2899	184.9799	0.9916	0.6840	5	25.8453	0.0001	0.0089
Text 21	2824	177.1861	0.9907	0.6487	5	33.8792	0.0000	0.0120
Text 22	4258	149.9868	0.9902	0.6274	5	58.1804	0.0000	0.0137
Text 23	3338	188.0436	0.9927	0.7131	5	44.6703	0.0000	0.0134
Text 24	4115	166.8914	0.9907	0.6470	5	54.6610	0.0000	0.0133

Text 25	4661	157.0027	0.9917	0.6594	5	24.9513	0.0001	0.0054
Text 27	4190	4.3116	0.7742	0.7742	6	25.8938	0.0002	0.0062
Text 28	4797	146.2749	0.9903	0.6116	5	44.5332	0.0000	0.0093
Text 29	4032	133.6587	0.9895	0.5759	5	28.5929	0.0000	0.0071
Text 30	3751	128.6930	0.9901	0.5960	4	25.0251	0.0000	0.0067

Table 24
Fitting the Dacey-negative binomial distribution to word length data in Marathi: Translated Material

Text No	N	k	p	α	DF	X^2	P	C
Text 01	5603	5.7081	0.8456	0.8456	5	22.1229	0.0005	0.0039
Text 02	4044	141.7979	0.9911	0.6254	5	22.2969	0.0005	0.0055
Text 03	4056	205.1701	0.9949	0.7898	4	2.1199	0.7137	0.0005
Text 04	5680	4.4576	0.8205	0.8205	5	21.1575	0.0008	0.0037
Text 05	3965	4.4714	0.7910	0.7910	4	37.5090	0.0000	0.0095
Text 06	3476	4.2714	0.8093	0.8093	4	14.8473	0.0050	0.0043
Text 07	5336	220.8658	0.9952	0.8098	4	16.7105	0.0022	0.0031
Text 08	3712	5.1967	0.8172	0.8172	5	8.5598	0.1280	0.0023
Text 09	3553	191.4976	0.9941	0.7536	4	10.7373	0.0297	0.0030
Text 10	4067	167.0439	0.9930	0.6999	4	24.4051	0.0001	0.0060
Text 11	3691	184.9319	0.9943	0.7542	4	3.6675	0.4529	0.0010
Text 12	4188	4.6367	0.8038	0.8038	4	14.9625	0.0048	0.0036
Text 13	3961	5.3505	0.8254	0.8254	5	13.0046	0.0233	0.0033
Text 14	3521	5.4611	0.8252	0.8252	5	15.7275	0.0077	0.0045

Table 25
Fitting the Dacey-negative binomial distribution to word length data in Tamil: Aesthetics

Text No	N	k	p	α	DF	X^2	P	C
Text 01	5200	83.1434	0.9783	0.9783	5	77.5245	0.0000	0.0149
Text 02	5203	36.1178	0.9512	0.9512	6	49.9150	0.0000	0.0096
Text 03	4900	26.7689	0.9446	0.9446	6	39.1088	0.0000	0.0080
Text 04	5110	20.3283	0.9053	0.9053	7	34.6588	0.0000	0.0068
Text 05	1947	406.8742	0.9964	0.9438	5	12.3496	0.0303	0.0063
Text 06	5096	44.9856	0.9564	0.8894	7	43.5172	0.0000	0.0085
Text 07	5236	343.9627	0.9935	0.8226	7	44.4243	0.0000	0.0085
Text 08	5146	21.0507	0.9432	0.9432	5	39.2660	0.0000	0.0076
Text 09	5202	19.7661	0.9404	0.9597	5	18.3819	0.0025	0.0035
Text 10	5244	216.9893	0.9901	0.6871	7	45.0415	0.0000	0.0086
Text 11	5245	229.3696	0.9913	0.7232	7	38.4424	0.0000	0.0073
Text 12	5158	63.7081	0.9724	0.9724	6	17.6834	0.0071	0.0034
Text 13	5352	377.0271	0.9950	0.8996	6	15.2415	0.0185	0.0028
Text 14	2106	33.1137	0.9371	0.1527	6	19.0945	0.0040	0.0091
Text 15	5193	746.6686	0.9975	0.9939	6	6.5399	0.3655	0.0013
Text 16	5229	78.3240	0.9767	0.9767	6	15.7263	0.0153	0.0030
Text 17	4927	477.1826	0.9971	0.9877	5	24.2590	0.0002	0.0049
Text 19	5094	46.1802	0.9732	0.9732	5	35.0717	0.0000	0.0069

Text 20	5844	10.0477	0.8285	0.9646	8	38.9808	0.0000	0.0067
Text 21	5075	21.3598	0.9114	0.9114	3	75.0727	0.0000	0.0148
Text 22	6551	475.3605	0.9964	0.9613	6	13.8586	0.0313	0.0021
Text 23	2010	218.1164	0.9933	0.8706	5	29.1320	0.0000	0.0145
Text 24	5300	55.1576	0.9586	0.9586	8	14.2961	0.0744	0.0027
Text 25	3536	162.0942	0.9912	0.9912	5	14.3126	0.0137	0.0040
Text 26	2496	17.6071	0.8897	0.8897	7	37.1469	0.0000	0.0149
Text 27	5480	47.8643	0.9716	0.9716	5	14.1562	0.0146	0.0026
Text 28	3909	114.2721	0.9833	0.8223	6	34.3737	0.0000	0.0088
Text 29	4029	19.2780	0.9105	0.9105	7	33.7295	0.0000	0.0084
Text 30	3084	18.5787	0.9189	0.9189	6	18.8103	0.0045	0.0061

Table 26
Fitting the Dacey-negative binomial distribution to word length data in Tamil: Natural Physical and Professional Sciences

Text No	N	k	p	α	DF	X^2	P	C
Text 01	4875	16.1498	0.9019	0.9019	7	32.3535	0.0000	0.0066
Text 02	4982	10.0003	0.8441	0.8441	3	74.2110	0.0000	0.0149
Text 03	4877	328.4434	0.9939	0.8203	2	62.7873	0.0000	0.0129
Text 04	1962	27.5578	0.9444	0.9444	5	5.4186	0.3670	0.0028
Text 05	4876	10.16791	0.8605	0.9661	6	4.8157	0.5677	0.0010
Text 06	5005	14.9372	0.8986	0.8986	6	56.6510	0.0000	0.0113
Text 07	4995	16.3117	0.8978	0.8978	2	74.9965	0.0000	0.0150
Text 08	4756	173.9403	0.9898	0.8259	6	8.9253	0.1778	0.0019
Text 09	4968	26.8019	0.9381	0.9381	6	15.4388	0.0171	0.0031
Text 10	4994	53.0328	0.9667	0.9667	6	19.5883	0.0033	0.0039
Text 11	4807	35.9980	0.9535	0.9172	6	7.2084	0.302	0.0015
Text 12	5014	312.0059	0.9952	0.9676	5	12.4919	0.0286	0.0025
Text 13	4886	71.1330	0.9800	0.9800	5	17.2941	0.0040	0.0035
Text 14	5304	43.8097	0.9690	0.9690	5	16.5647	0.0054	0.0031
Text 15	5092	23.2133	0.9312	0.9312	6	29.0230	0.0001	0.0057
Text 16	5135	262.6838	0.9930	0.7786	6	22.1058	0.0012	0.0043
Text 17	5038	31.9043	0.9466	0.9466	6	18.3582	0.0054	0.0036
Text 18	5081	23.3495	0.9202	0.9202	5	74.3942	0.0000	0.0146
Text 19	5015	47.0082	0.9648	0.9648	5	65.3095	0.0000	0.0130
Text 20	4946	62.5717	0.9699	0.9699	7	7.5470	0.3742	0.0015
Text 21	5108	24.5264	0.9328	0.9328	7	17.5860	0.014	0.0034
Text 22	5033	579.2906	0.9968	0.9744	6	23.2487	0.0007	0.0046
Text 23	5100	36.8263	0.9501	0.9501	7	40.6311	0.0000	0.0080
Text 24	4980	16.8887	0.8895	0.9626	8	12.4246	0.1332	0.0025
Text 25	5023	39.1597	0.9554	0.9554	6	71.2481	0.0000	0.0142
Text 26	4931	92.3290	0.9810	0.9810	3	60.7595	0.0000	0.0123
Text 27	5508	9.4754	0.8171	0.8171	7	80.1596	0.0000	0.0146
Text 28	5180	156.6311	0.9866	0.8314	7	65.8145	0.0000	0.0127
Text 29	4786	26.7418	0.9291	0.9584	7	57.5044	0.0000	0.0120
Text 30	5134	51.5393	0.9665	0.9665	6	43.3029	0.0000	0.0084

Table 27
Fitting the Dacey-negative binomial distribution to word length data in Tamil: Commerce

Text No	N	k	p	α	DF	X^2	P	C
Text 01	4756	314.6970	0.9934	0.8232	7	49.3315	0.0000	0.0104
Text 02	4975	81.3059	0.9800	0.9800	6	61.7491	0.0000	0.0124
Text 03	2936	28.0239	0.9368	0.9368	6	31.7207	0.0000	0.0108
Text 04	4990	12.0092	0.8633	0.8633	7	17.9239	0.0123	0.0036
Text 05	4752	54.7885	0.9736	0.9736	5	34.0021	0.0000	0.0072

Table 28
Fitting the positive Cohen-Poissondistribution to word length data in Tamil: Translated Material

Text No	N	a	α	DF	X^2	P	C
Text 01	2214	2.9160	0.9999	4	28.8579	0.0000	0.0130
Text 03	4943	3.3809	0.9039	9	58.9534	0.0000	0.0119
Text 04	5094	3.5521	0.9698	8	37.9777	0.0000	0.0075
Text 05	4975	3.2039	0.9238	8	45.0100	0.0000	0.0090
Text 06	4926	3.3885	0.9377	8	36.1859	0.0000	0.0073
Text 07	5229	3.0920	0.9789	8	61.9705	0.0000	0.0119
Text 08	4807	3.6443	0.9463	8	66.8213	0.0000	0.0139
Text 09	3127	3.5249	0.9681	9	37.0071	0.0000	0.0118
Text 11	4867	3.3526	0.9465	8	54.5215	0.0000	0.0112
Text 12	5129	3.4059	0.9793	8	28.0802	0.0005	0.0055
Text 13	3844	3.5267	0.9999	8	35.0269	0.0000	0.0091
Text 14	3337	3.5563	0.9822	8	45.2675	0.0000	0.0136
Text 15	3436	3.5145	0.9832	7	45.7987	0.0000	0.0133
Text 16	3571	2.9450	0.9323	8	47.4601	0.0000	0.0133

Table 29
Fitting the Dacey-negative binomial distribution to word length data in Tamil: Official and Media Languages

Text No	N	k	p	α	DF	X^2	P	C
Text 01	1046	39.8950	0.9508	0.9508	5	11.0404	0.0506	0.0106
Text 02	954	387.2934	0.9949	0.8801	5	8.1724	0.147	0.0086
Text 04	954	71.8277	0.9800	0.9800	4	13.7051	0.0083	0.0144
Text 05	993	94.4394	0.9815	0.8602	2	14.5053	0.0007	0.0146
Text 06	1022	31.7282	0.9426	0.9426	3	15.1221	0.0017	0.0148
Text 07	1006	13.5551	0.8844	0.8844	2	14.0512	0.0009	0.0140
Text 08	949	354.2594	0.9947	0.7781	1	12.3703	0.0004	0.0130
Text 09	977	597.6133	0.9966	0.8435	2	14.5061	0.0007	0.0148
Text 11	976	280.8498	0.9937	0.8065	5	9.5188	0.0901	0.0098
Text 12	1059	283.3817	0.9924	0.7644	6	10.4849	0.1057	0.0099
Text 14	945	303.8665	0.9937	0.8272	2	12.8092	0.0017	0.0136
Text 15	994	168.8275	0.9878	0.5570	5	10.5477	0.0611	0.0106
Text 17	1022	8.5959	0.8225	0.8225	6	6.0748	0.4149	0.0059

Text 19	983	29.2094	0.9423	0.9423	3	14.7826	0.0020	0.0150
Text 20	1001	13.0102	0.8619	0.8619	3	13.8932	0.0031	0.0139
Text 21	967	528.4906	0.9964	0.8550	2	11.1696	0.0038	0.0116
Text 22	1010	20.5560	0.9161	0.9161	3	15.1520	0.0017	0.0150
Text 24	1038	332.1895	0.9944	0.7447	1	13.8967	0.0002	0.0134
Text 25	998	9.1077	0.8333	0.8333	5	13.6828	0.0178	0.0137
Text 26	1030	95.1574	0.9780	0.6478	2	15.4856	0.0004	0.0150
Text 27	1016	265.8643	0.9922	0.6306	1	12.6925	0.0004	0.0125
Text 28	1019	13.3028	0.8845	0.8845	2	14.3395	0.0008	0.0141
Text 29	987	23.3862	0.9337	0.9337	3	14.0040	0.0029	0.0142
Text 30	955	336.8037	0.9939	0.8360	6	8.3185	0.2157	0.0087

Table 30
Fitting the Dacey-negative binomial distribution to word length data in Tamil: Social Sciences

Text No	N	k	p	α	DF	X^2	P	C
Text 01	2406	683.6750	0.9980	0.9999	3	7.6890	0.0529	0.0032
Text 02	5142	246.7137	0.9912	0.7148	7	10.8809	0.1439	0.0021
Text 03	6175	112.4412	0.9820	0.8409	7	25.5867	0.0006	0.0041
Text 04	5131	347.7708	0.9938	0.8553	7	24.3910	0.0010	0.0048
Text 05	5124	445.4117	0.9954	0.9317	6	13.1681	0.0404	0.0026
Text 06	6205	16.1234	0.8770	0.8770	8	62.0140	0.0000	0.0100
Text 07	3374	564.5479	0.9971	0.9739	5	22.0915	0.0005	0.0065
Text 08	5381	259.7408	0.9914	0.7375	7	15.1859	0.0337	0.0028
Text 09	3836	305.3744	0.9925	0.7621	3	57.1195	0.0000	0.0149
Text 10	5052	47.6926	0.9655	0.9655	6	15.1308	0.0193	0.0030
Text 11	5799	233.5780	0.9903	0.6853	6	14.7447	0.0223	0.0025
Text 12	4454	287.6103	0.9913	0.7286	7	31.2472	0.0001	0.0070
Text 13	4998	24.1529	0.9160	0.9160	7	42.8964	0.0000	0.0086
Text 14	3651	182.3368	0.9881	0.5829	7	22.5271	0.0021	0.0062
Text 15	4980	81.4072	0.9800	0.9800	6	24.9911	0.0003	0.0050
Text 16	4891	146.4599	0.9881	0.8780	4	72.5445	0.0000	0.0148
Text 17	5169	68.7798	0.9755	0.9755	3	76.7513	0.0000	0.0148
Text 18	5078	19.2543	0.9133	0.9133	7	59.0615	0.0000	0.0116
Text 19	5117	24.3124	0.9267	0.9267	3	50.9334	0.0000	0.0100
Text 20	5103	16.5389	0.8628	0.8628	4	76.2726	0.0000	0.0149
Text 21	5187	22.3424	0.9006	0.9006	3	75.6449	0.0000	0.0146
Text 22	5062	44.6656	0.9597	0.9597	7	31.0009	0.0001	0.0061
Text 23	5159	33.0798	0.9465	0.9465	7	58.1115	0.0000	0.0113
Text 24	2400	82.7955	0.9749	0.8335	2	35.0297	0.0000	0.0146
Text 25	6053	13.7969	0.8527	0.8527	4	89.2290	0.0000	0.0147
Text 26	4950	53.9447	0.9647	0.8818	6	70.0692	0.0000	0.0142
Text 27	1789	349.8144	0.9950	0.8341	1	25.6031	0.0000	0.0143
Text 28	5015	69.3434	0.9779	0.9779	6	32.4958	0.0000	0.0065
Text 29	4940	24.5280	0.9398	0.9398	6	33.4275	0.0000	0.0068
Text 30	5146	92.9959	0.9777	0.8386	7	19.3825	0.0071	0.0038

References

Abbe, S. (2000). Word length distribution in Arabic letters. *J. of Quantitative Linguistics 7, 121-127.*

Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie.* Götingen: Peust & Gutschmidt.

Balschun, C. (1997). Wortlängenhäufigkeiten in althebräischen Texten. In: Best, K.-H. (ed.), *Glottometrika 16, 174-179.* Trier: WVT.

Best, K.-H., Medrano, P. (1997). Wortlängen in Ketschua-Texten. In: Best, K.-H. (ed.), *Glottometrika 16: 204-212.* Trier: WVT.

Best, K.-H., Özmen, E. (1996). Wortlängenhäufigkeiten in türkischen Texten und ihre linguistischen Implikationen. *Archív Orientální 64, 19-30.*

Best, K.-H., Zhu, J. (2001). Wortlängenverteilungen in chinesischen Texten und Wörterbüchern. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 101-114.* Göttingen: Peust & Gutschmidt.

Bhattacharya, S.S., Jayaram, B.D., Itagi, N.H. (2005). Documenting Indian Languages: LIS India Experience. *Paper presented in Prof. M.B.Emeneau centenary International conference on South Asian Linguistics, June 1-4, 2005, Mysore, India*

Ejiri, K., Staeheli, N., Ooaku, S. (1994). Word frequency distribution in Japanese text. *J. of Quantitative Linguistics 1, 212-223.*

Jayaram, B.D., Rajyashree, K.S. (2001). Indian Language corpora Development. *Paper presented in Language Engineering in South Asian Languages, April 23-25, 2001, NCST, Mumbai, India.*

Kim, I., Altmann, G. (1996). Zur Wortlänge in koreanischen Texten. In: Schmidt, P. (Hg.), *Glottometrika 15, 205-213.* Trier: WVT.

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

Milic, L.T., Slane, S. (1994). Style. Quantitative aspects of genre in the century of prose corpus, *Spring94, Vol. 28, Issue 1.*

Glottometrics 12, 2006, 39-54

Gesetzmäßigkeiten im Erstspracherwerb

Karl-Heinz Best[1]*, Göttingen*

Abstract. Language acquisition abides by laws. These laws seem to be the same as in language change and text production. If one considers the stages of language acquisition separately then there are further models controlling distributions, rank orders and processes. This paper yields some evidence for these laws.

Keywords: First language, language acquisition, German

Gesetzmäßigkeiten im Erstspracherwerb

In diesem Beitrag soll noch einmal der Frage nachgegangen werden, ob die Prozesse und die im Ergebnis eintretenden Zustände im Erstspracherwerb sich gesetzmäßig verhalten oder nicht. Es wäre ja immerhin denkbar, dass die gesetzmäßigen Prozesse oder Zustände, die in Sprachsystem und -wandel immer wieder zu beobachten sind, eine gewisse Zeit benötigen, um sich einstellen zu können. In Best (2003a,b) konnte aber bereits gezeigt werden, dass zumindest einige Erwerbsprozesse den gleichen Gesetzen folgen, die auch für den Sprachwandel vorgeschlagen wurden (Altmann 1983) und sich vielfach bewährt haben (Best 2003c). Dass der Verlauf des Spracherwerbs dem logistischen Gesetz folgt, deutet bereits Lenneberg (1972: 166) graphisch an; Wagner, Altmann & Köhler (1987: 139, Formel 8) schlagen das entsprechende Modell dazu vor. Dies hat sich mehrfach bestätigen lassen; einige weitere Überprüfungen werden folgen. Es wurde jedoch anscheinend noch nicht untersucht, ob auch die Zustände, die sich schon bei Kindern in verschiedenen Altersstufen einstellen, gesetzmäßige Züge tragen. Auch dies wird hier an einigen Beispielen demonstriert.

I. Zur Gesetzmäßigkeit von Erwerbsprozessen

I.1. Lauterwerb

Nachdem in den bereits erwähnten Untersuchungen hinreichend Belege dafür zu finden sind, dass verschiedenste Spracherwerbsprozesse dem logistischen Gesetz (auch: Piotrowski-Gesetz) folgen, soll hier nun ein Fall vorgeführt werden, der sich etwas anders darstellt. Es handelt sich um die Lauterwerbsprozesse, so wie sie von Chen und Irwin erforscht wurden. Die bekannteste Untersuchung hierzu dürfte Chen & Irwin (1946) sein. Die Autoren beobachteten den Lauterwerb von insgesamt 95 Kindern von den ersten Lebensmonaten an bis zum Alter von 30 Monaten und schlugen als Modell für diesen Prozess

$$(1) \qquad y = ax^b$$

[1] Address correspondence to: Karl-Heinz Best: kbest@gwdg.de

vor. Dieses Modell entspricht dem sog. Menzerath-Altmann-Gesetz in seiner einfachsten Form (Altmann 1980: 3). Tabelle 1 gibt die Beobachtungen von Chen & Irwin (1946) getrennt für Konsonanten und Vokale wieder (MW: Mittelwerte der Altersgruppen); die Berechnung wurde mit *NLREG* erneut durchgeführt und bestätigt ebenso wie die Graphik mit nur geringfügigen Abweichungen die Angaben der Autoren.

Tabelle 1
Lauterwerb (mittlere Anzahl der Lauttypen nach Alter)

Alter (in Monaten)	*t*	Konsonanten		Vokale	
		$MW_{beobachtet}$	$MW_{berechnet}$	$MW_{beobachtet}$	$MW_{berechnet}$
1-2	1	2.7	2.83	4.5	5.22
3-4	2	4.5	4.04	6.6	6.39
5-6	3	5.2	5.70	7.1	7.20
7-8	4	7.0	6.84	8.1	7.83
9-10	5	7.7	7.88	8.3	8.36
11-12	6	9.7	8.85	8.9	8.82
13-14	7	9.8	9.76	10.0	9.22
15-16	8	10.8	10.63	9.7	9.59
17-18	9	10.9	11.46	10.1	9.93
19-20	10	12.4	12.25	10.2	10.24
21-22	11	12.7	13.01	10.3	10.53
23-24	12	13.7	13.75	10.7	10.80
26-26	13	14.8	14.47	11.0	11.05
27-28	14	15.1	15.17	11.0	11.30
29-30	15	15.8	15.85	11.4	11.53
		$a = 2.8349$ $b = 0.6356$	$D = 0.99$	$a = 5.2166$ $b = 0.2928$	$D = 0.97$

Anm.: *a* und *b* sind die Parameter des Modells; *D* ist der Determinationskoeffizient, der mit $D \geq 0.90$ eine sehr gute Anpassung des Modells anzeigt, wie auch die Graphik (Abb. 1) bestätigt:

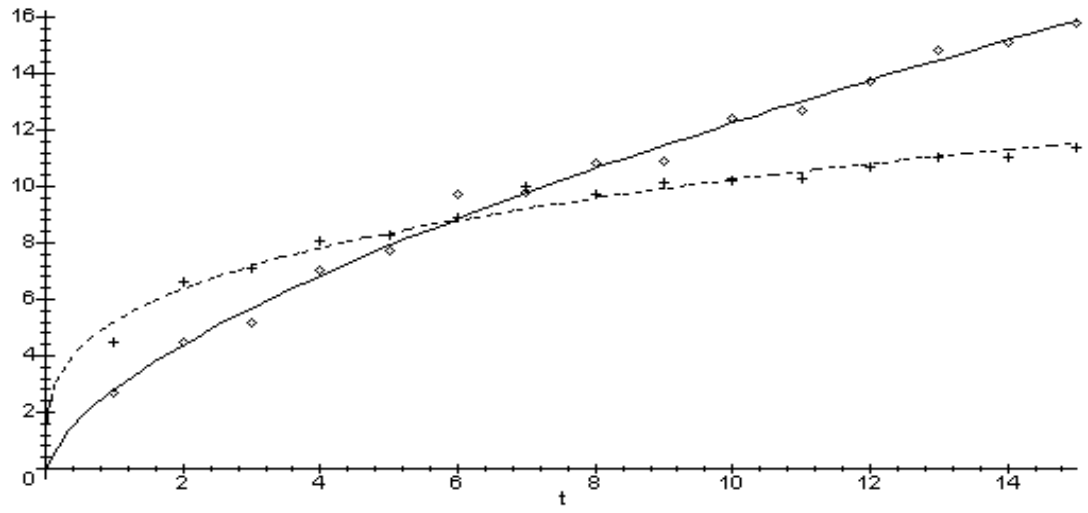


Abb. 1. Lauterwerb (mittlere Anzahl der Lauttypen nach Alter)

Chen & Irwin (1946: 28) ziehen daraus das Resumee: „It is suggested that the equations reported in this article may be considered the expressions of two laws of phoneme type development during the period of infancy." Das gleiche Modell bewährt sich in Irwin & Chen (1946), worin die Autoren den Lauterwerb der Kinder u.a. für Jungen und Mädchen getrennt bearbeiten; in diesem Fall wurden Konsonanten und Vokale nicht je für sich ausgewertet.

Irwin (1948) untersucht den Lauterwerb von Kindern aus zwei unterschiedlichen sozialen Milieus und zeigt, dass die Erwerbsprozesse dem abgewandelten Modell

$$(2) \quad y = a\,e^{bx}$$

folgen, was einer weiteren Variante des Menzerath-Altmann-Gesetzes entspricht (Altmann 1980: 3). Weitere Modifikationen des gleichen Modells nutzt Irwin (1947).

An diesen Untersuchungen ist eines besonders interessant: Der Lauterwerb setzt offensichtlich mit hohen Zuwachsraten ein und flacht mit der Zeit allmählich ab; es gibt keine Hinweise darauf, dass es sich um einen typischen logistisch verlaufenden Prozess handelt, der ja allmählich einsetzt, sich beschleunigt, einen Wendepunkt erreicht und dann sich zunehmend verlangsamt. Bei den beobachteten Lauterwerbsprozessen fehlt offenbar die erste Hälfte der Entwicklung bis zum Wendepunkt; es ist ja kaum vorstellbar, dass diese erste Hälfte sich innerhalb der ersten beiden Lebensmonate abspielt und nur dadurch verloren geht, dass für diese Zeitspanne ein Durchschnittswert gebildet wird. Wenn diese Schlussfolgerung richtig ist, dann bedeutet das, dass man mit zwei unterschiedlichen Formen von Spracherwerbsprozessen zu rechnen hat: einer, die dem Menzerath-Altmann-Gesetz entspricht, und einer weiteren, die dem logistischen Gesetz folgt. Für den erstgenannten Typ gibt es bisher nur die Beobachtungen zum Lauterwerb, während alle anderen Erwerbsprozesse anscheinend nach dem Piotrowski-Gesetz verlaufen. Daten, die Grohnfeldt (1980: 170) und Templin (1957: 30) mitteilen, setzen erst mit dem Alter von 2 bzw. 3 Jahren ein und geben daher keinen Aufschluss über die ersten Anfänge des Lauterwerbs und damit auch nicht über die Wahl eines geeigneten Modells; beide Modelle können in allen erwähnten Fällen mit guten Ergebnissen angepasst werden.

I.2. Die Entwicklung des Wortschatzes im Schulalter

Der Aufbau des kindlichen Wortschatzes folgt dem logistischen Gesetz (Best 2003a). Weitere Daten zum Wortschatz in der Schulzeit findet man zum Deutschen bei Rest u.a. (1977). Eine gewisse Vorsicht ist jedoch geboten; so stellt Augst (1984: V) fest, dass die Angaben zum Wortschatz der Kinder „im Laufe der Forschungsjahrzehnte ständig anwächst." Auch Oksaar (1977: 188) meldet Vorbehalte gegenüber älteren Untersuchungen an: „Es ist anzunehmen, daß die Entwicklung des Wortschatzes heute schneller verläuft."

Die Untersuchung von Rest u.a. (1977: III) galt dem „Verhältnis von Umgangssprache (Sprechsprache) und Schulsprache (Schriftsprache)." Da nur Schulbücher des 1. und 2. Schuljahres und Sprechsprache des 1. – 3. Schuljahres erhoben wurden, sind die Daten für eine Untersuchung dieses Spracherwerbsprozesses nicht ausreichend.

I.3. Das erste Auftreten von Gegenstandsbenennungen

Bühler ([4]1967: 95) gibt in Form einer Graphik wieder, in welchem Alter (in Monaten) bei insgesamt 49 Kindern zum ersten Mal Gegenstandsbenennungen auftreten. Die Zusammen-

stellung beruht auf den Angaben mehrerer Forscher; sie setzt mit 8 Monaten ein und reicht bis zum Alter von 20 Monaten. Die drei Kinder, die erst mit 20 Monaten erste Gegenstandsbenennungen entwickeln, werden als „pathologisch“ aufgefasst (Bühler [4]1967: 95; Fußnote 1). Sie werden bei den Berechnungen mit berücksichtigt. In diesem Fall ist das logistische Gesetz

$$(3) \quad p = \frac{c}{1 + ae^{-bt}}$$

ein sehr gutes Modell, da das Testkriterium mit $D = 0.97$ erfüllt ist (vgl. Tabelle 2 u. Abb. 2).

Tabelle 2
Erstes Auftreten von Gegenstandsbenennungen

t Alter (in Monaten)	Kinder mit ersten Gegenstandsbenennungen (absolut)	Kinder mit ersten Gegenstandsbenennungen (kumuliert)	Kinder mit ersten Gegenstandsbenennungen (berechnet)
8	3	3	7.49
9	7	10	12.43
10	13	23	19.12
11	6	29	26.78
12	5	34	34.01
13	5	39	39.71
14	4	43	43.58
15	1	44	45.97
16	1	45	47.34
17	1	46	48.10
18	0	46	48.52
19	0	46	48.74
20	3	49	48.86
$a = 876.3509$	$b = 0.6329$	$c = 49$	$D = 0.97$

a, *b*, *c*: Parameter des Modells.

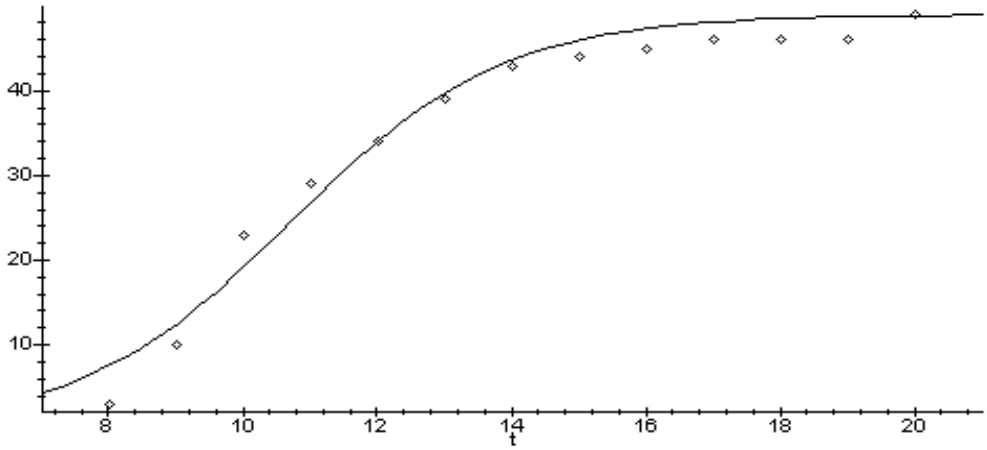


Abb. 2. Erstes Auftreten von Gegenstandsbenennungen

I.4. Zur Entwicklung der Wortlängen

Ergebnisse zur Zunahme der Wortlänge bei Kindern finden sich schon bei Busemann (1925: 95). Er stellt auf der Grundlage von 163 Aufsätzen die Entwicklung bei 10-15jährigen Mädchen dar. Es ist eindeutig, dass Wortlängen mit dem Alter zunehmen; die Schwankungen sind jedoch so groß, dass Modell (3) mit $D = 0.45$ nur ein völlig unakzeptables Ergebnis erbringt. Belässt man es bei den wenigen von Busemann mitgeteilten Daten, erscheinen die beobachteten Schwankungen der Wortlängen überdimensioniert.

Als Ausweg bietet sich an, einen realistischen Wert für das Erwachsenenalter zu ergänzen. Einige Hinweise dazu finden sich in Best (2001b: 31), Deußing (1927/1969: 120). Setzt man in Busemanns Datei zusätzlich $t = 26$ für ein Alter von 35 Jahren ein und eine Wortlänge von 1.910 (Best 2001b), so erhält man mit Modell (3) eine hervorragende Anpassung mit $D = 0.98$; mit einer angenommenen Wortlänge von 1.830 (Deußing 1927/1969) kommt man bei gleichem Alter ebenfalls auf $D = 0.98$. Es ist also ganz klar, dass die Wertestreuung bei Busemann (1925) an Bedeutung verliert, wenn man die Altersspanne erweitert. Für die folgende Tabelle und Berechnung wurde ein Wert von 1.700 Silben pro Wort für ein Erwachsenenalter von 35 Jahren als Durchschnittswert angenommen. Dieser Wert ist laut Fucks (1968: 80) repräsentativ für „mittlere Autoren"; er entspricht genau demjenigen, der im Durchschnitt für weit über 100 Briefe mehrerer Autoren des 20. Jahrhunderts aus verschiedenen Untersuchungen des *Göttinger Projekts zur Quantitativen Linguistik* (Best 2001) ermittelt werden kann. Der Gesamtprozess folgt dem logistischen Gesetz (Modell 3) (vgl. Tabelle 3 u. Abb. 3).

Tabelle 3
Entwicklung der Wortlänge bei Mädchen (Aufsätze), (ergänzt um $t = 26$ für 35 J.)

Alter	t	n_x	$\hat{n}_x$	Alter	t	n_x	$\hat{n}_x$
10	1	1.528	1.521	14	5	1.563	1.552
11	2	1.544	1.529	15	6	1.567	1.560
12	3	1.506	1.537	35	26	1.700	1.700
13	4	1.537	1.545				
$a = 0.4401$ $b = 0.0171$ $c = 2.1795$ $D = 0.94$							

n_x = Wortlänge, beobachtet; $\hat{n}_x$ = Wortlänge, aufgrund des logistischen Gesetzes berechnet.

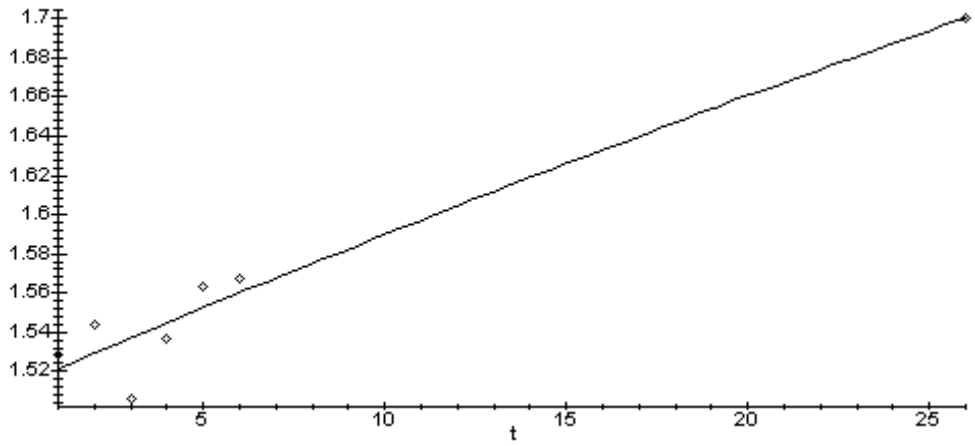


Abb. 3. Entwicklung der Wortlänge bei 10-15jährigen Mädchen (ergänzt um $t = 26$ für 35 J.)

Weitere Daten zur Entwicklung der Wortlänge findet man bei Deußing (1927/1969: 117-122). Bei der Untersuchung der Wortlänge in Silben gemessen differenziert er zwischen der gesprochenen und geschriebenen Sprache. Leider gibt Deußing für vier Altersstufen die Daten für nur je ein Kind an. Dieser Datenbestand ist also noch schwächer als der bei Busemann (1925) und alles andere als repräsentativ. Man muss jedenfalls mit Unregelmäßigkeiten aufgrund der individuellen Unterschiede der berücksichtigten Kinder rechnen. So nimmt es nicht Wunder, dass das logistische Gesetz in der Form von Formel (3) hier als Modell versagt. Um zu zeigen, dass der Prozess dennoch nicht chaotisch verläuft, kann jedoch Modell (2) mit guten Ergebnissen verwendet werden (vgl. Tabelle 4 und Abb. 4):

Tabelle 4
Entwicklung der Wortlänge bei Kindern (mündlich und schriftlich)

		gesprochene Sprache		geschriebene Sprache	
Alter	t	n_x	$\hat{n}_x$	n_x	$\hat{n}_x$
8;4	1	1.43	1.43	1.46	1.47
10;5	26	1.47	1.48	1.53	1.51
12;4	49	1.54	1.53	1.55	1.55
14;3	72	1.57	1.58	1.58	1.59
		$a = 1.4267$ $b = 0.0014$ $D = 0.98$		$a = 1.4709$ $b = 0.0011$ $D = 0.93$	

t: Alter der Kinder in Monaten, beginnend mit $t = 1$ für ein Alter von 8;4 Jahren.

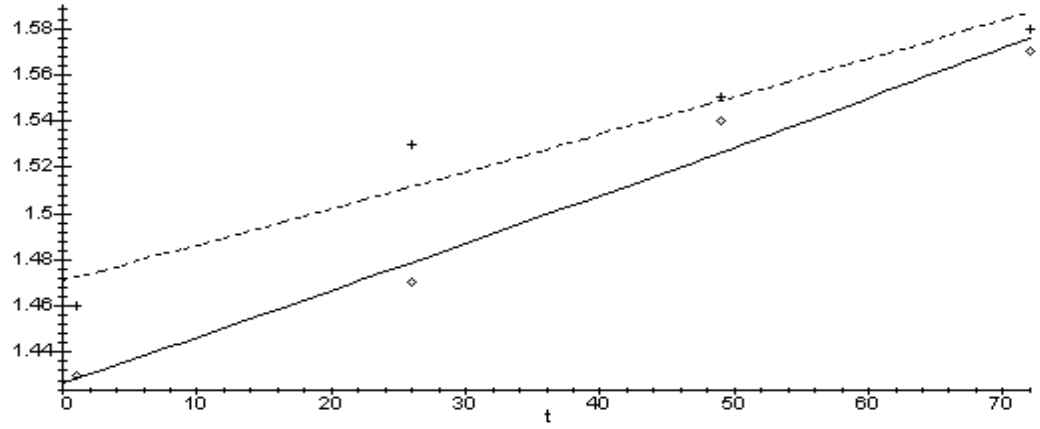


Abb. 4. Entwicklung der Wortlänge bei Kindern (mündlich: beobachtet: Raute, berechnet: durchgezogene Linie; schriftlich: beobachtet: Kreuze, berechnet: gestrichelte Linie)

Deutlich höhere Durchschnittswerte als Busemann und Deußing teilt Fischer ([4]1971: 173) für Aufsätze von Mittelschülern zwischen 13 und 17 Jahren mit. Dies zeigt nur, dass Wortlängen je nach schriftlichem oder mündlichem Sprachgebrauch, Textsorte, Alter der Probanden, Aufgabenstellung bei der Textproduktion und weiteren Randbedingungen erheblich schwanken können. Immer stellt sich das Problem der Repräsentativität und Vergleichbarkeit der Untersuchungsstrategie und -ergebnisse.

I.5. Entwicklung der Satzlänge

An zwei Beispielen soll die gesetzmäßige Entwicklung der Satzlänge (in Wörtern gemessen) demonstriert werden. Es handelt sich zunächst um die Untersuchung von Smith (1935: 186) zu insgesamt 305 amerikanischen Kindern im Alter von 1 ½ - 5 ½ Jahren. Der Prozess folgt dem logistischen Gesetz (Modell 3) (vgl. Tabelle 5 und Abb. 5):

Tabelle 5
Entwicklung der Satzlänge (Wörter pro Satz) bei Kindern

Alter	n_x	$\hat{n}_x$
1 ½	1.2	1.10
2	1.8	1.80
2 ½	2.5	2.65
3	3.5	3.49
3 ½	4.3	4.17
4	4.6	4.62
4 ½	4.9	4.89
5	5.0	5.04
5 ½	5.1	5.12
a = 28.4423	b = 1.3545	c = 5.2011 D = 0.997

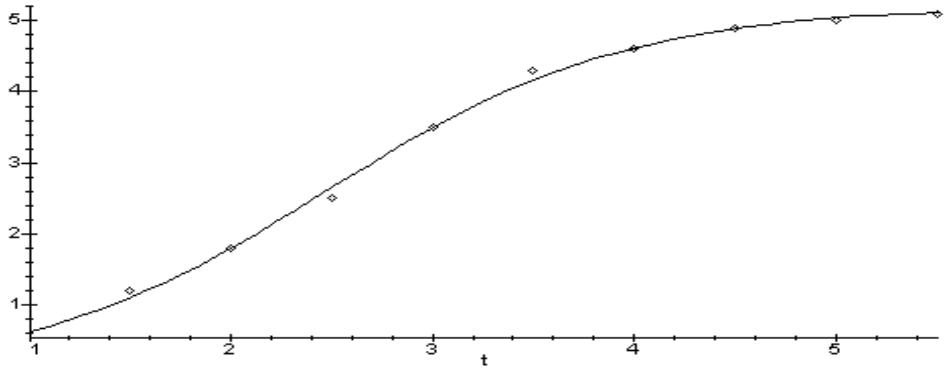


Abb. 5. Entwicklung der Satzlänge (Wörter pro Satz) bei Kindern

Satzlängen in geschriebener Sprache ab 9 Jahren bis zum Erwachsenenalter haben Stormzand & O'Shea (1924: 19) auf der Basis von 10000 Sätzen erhoben (s. auch McCarthy 1954: 550). Für „College freshmen" wurde 18 Jahre angesetzt, für „College Upper Classmen" 21 Jahre und für „Adults" 30 Jahre. Der Prozess folgt dem logistischen Gesetz (Modell 3) (vgl. Tabelle 6 und Abb. 6).

Tabelle 6
Entwicklung der Satzlänge (Wörter pro Satz) bei Kindern und Erwachsenen

Alter	n_x	$\hat{n}_x$	Alter	n_x	$\hat{n}_x$
9	11.1	9.99	16	18.0	18.48
11	12.0	12.90	17	19.8	19.16
12	13.5	14.27	18	19.9	19.73
13	15.2	15.54	21	21.5	20.82
14	17.3	16.67	30	20.9	21.71
15	17.8	17.65			
$a = 13.3259$ $b = 0.2692$ $c = 21.7950$ $D = 0.96$					

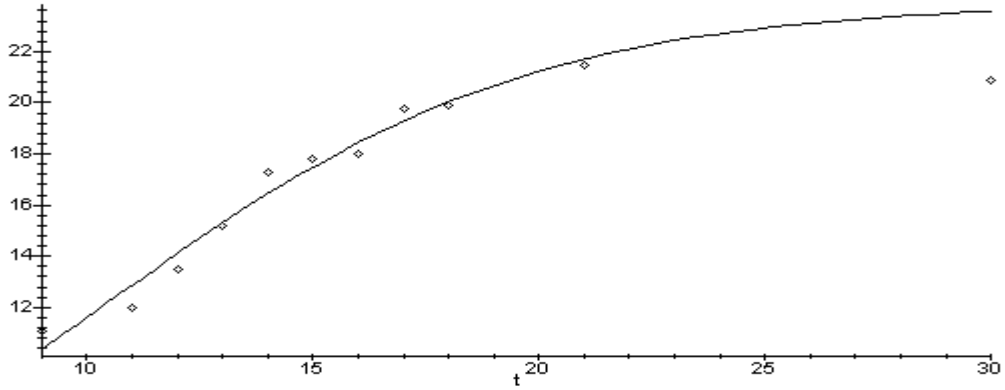


Abb. 6. Entwicklung der Satzlänge (Wörter pro Satz) bei Kindern und Erwachsenen

Stormzand & O'Shea (1924) untersuchen die Entwicklung syntaktischer Leistungen noch nach weiteren Kriterien: Der Anteil einfacher Sätze nimmt zunächst zu, dann aber mit der verbesserten Beherrschung komplexer Sätze wieder ab. Für komplexe Sätze werden unterschiedliche Auswertungen durchgeführt: durch Semikolon getrennte Teilsätze, Anzahl der Teilsätze (clauses), Anzahl abhängiger Teilsätze je 100 Sätzen. Außerdem wird auch die Veränderung der Fehlerquote pro Satz und pro Wort mit zunehmendem Alter betrachtet. Fast alle diese Fälle folgen mit guten bis sehr guten Ergebnissen dem logistischen Gesetz in Form des Modells (3); der reversible Prozess der Zu- und dann Abnahme einfacher Sätze unterliegt einem entsprechend veränderten Modell, das Altmann (1983: 78-86) ableitet und überprüft. Nur einer der genannten Fälle macht Probleme: die Abnahme des Anteils von Sätzen mit 2 und mehr clauses zeigt erhebliche Schwankungen, die eine akzeptable Anpassung von Modell (2) verhindern. Vermutlich ist diese Untersuchungskategorie lediglich zu undifferenziert und führt deshalb zu diesem Misserfolg.

McCarthy (1954: 546-549) stellt tabellarisch Ergebnisse zur Satzlängenentwicklung in gesprochener und in geschriebener Sprache (McCarthy 1954: 550) zusammen. Bildet man aus Ergebnissen von Smith und Templin zur gesprochenen und Heider & Heider sowie Stormzand & O'Shea zur geschriebenen Sprache eine gemeinsame Übersicht, so lässt sich auch in diesem Fall zeigen, dass der gesamte Prozess trotz der Störungen, die die verschiedenen Register bedeuten sollten, mit sehr gutem Ergebnis ($D = 0.99$) dem logistischen Gesetz folgt.

Weitere Ergebnisse zum Wachstum von clause- und Satzlänge und zur Länge sog. „T-units", einer Einheit zwischen clause und Satz, mit zunehmendem Alter stellt Hunt (1965) vor. An diese Daten kann das gleiche Modell erfolgreich angepasst werden. Auf eine Wiedergabe der Ergebnisse wird hier verzichtet, weil der Datenbestand doch vergleichsweise unbe-

friedigend ist. Hunt stellt dar, wie lang Sätze von Kindern im 4., 8. und 12. Schuljahr (grades) sind und gibt als Vergleich Satzlängen für „superior adults“ an, ohne dazu eine Altersangabe zu liefern. Man erfährt nur, dass die Daten durch Auswertung von *Harper's* und *Atlantic* gewonnen wurden. Nimmt man für Erwachsene ein Durchschnittsalter von 40 Jahren an, kann man das logistische Gesetz mit Erfolg anwenden.

I.6. Entwicklung der Textlänge

Pregel & Rickheit (1975: 37) haben die Zunahme der Wortzahl mündlicher Texte bei Kindern im Alter von 6;0 bis 9;11 Jahren zusammengestellt und betonen die auffälligen Schwankungen, die dabei zu beobachten sind. Dennoch lässt sich ein klarer Trend feststellen, der auch dem logistischen Gesetz (Modell 3) folgt (vgl Tabelle 7 und Abb. 7).

Tabelle 7
Entwicklung der Länge mündlicher Texte (Wörter pro Text)
bei Kindern im Schulalter

Alter	t	n_x	$\hat{n}_x$
6;0 – 6;3	1	62.8	55.33
6;4 – 6;7	2	57.3	62.22
6;8 – 6;11	3	59.6	68.48
7;0 – 7;3	4	70.7	73.96
7;4 – 7;7	5	89.0	78.57
7;8 – 7;11	6	84.3	82.35
8;0 – 8;3	7	87.6	85.36
8;4 – 8;7	8	82.9	87.72
8;8 – 8;11	9	90.1	89.54
9;0 – 9;3	10	91.8	90.93
9;4 – 9;7	11	94.9	91.97
9;8 – 9;11	12	88.7	92.76
$a = 0.9757$ $b = 0.3081$ $D = 0.83$			

Das Modell (3) lässt sich mit zufriedenstellendem $D = 0.83$ an die Beobachtungwerte anpassen, wie auch Abb. 7 zeigt:

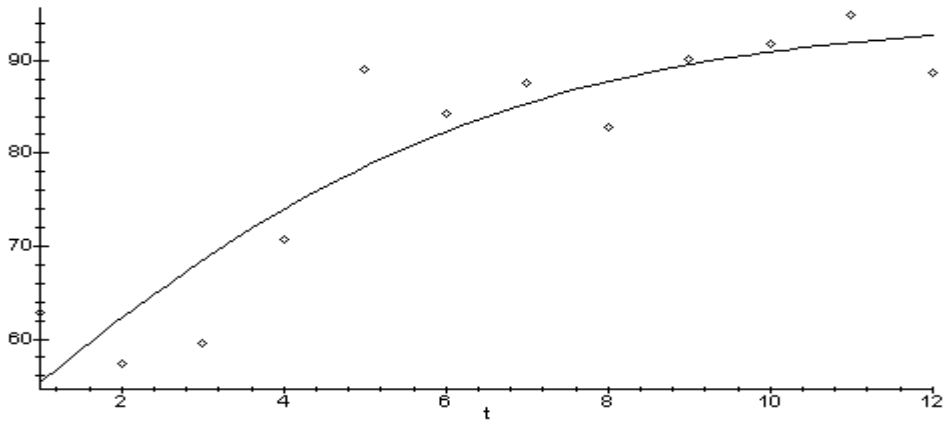


Abb. 7. Entwicklung der Länge mündlicher Texte bei Kindern im Schulalter

Dem gleichen Modell entsprechen auch die Daten zur Entwicklung der Textlänge, die für schriftliche Texte von Kindern im Alter von 8;8 – 8;11 ($D = 0.97$) und für mündliche Texte vom 1. – 4. Schuljahr ($D = 0.99$) vorgestellt werden (Pregel & Rickheit 1975: 39, 41).

II. Gesetzmäßigkeiten als Ergebnis von Erwerbsprozessen

In diesem Abschnitt wird nun die Perspektive geändert: Es geht nicht mehr um Modelle für den Erwerbsprozess selbst, sondern darum, ob sich die Zustände, die bei Kindern eines bestimmten Alters beobachtet werden, ebenfalls gesetzmäßig verhalten. Es gibt ja eine ganze Reihe von Gesetzeshypothesen, die sich bei der Untersuchung des Sprachsystems und seiner Verwendung vielfach bewährt haben. Unklar ist bisher aber weitgehend, ob dies auch schon für die Phasen des Spracherwerbs gilt oder sich erst später einstellt.

II.1. Zur Wortschatzdynamik

Untersucht man das Auftreten neuer Wörter im Text, so kann man nachweisen, dass mit zunehmender Textlänge immer weniger neue Wörter hinzukommen. Entsprechende Daten zu einem Kind (Christiane, Alter 12;2) haben Wagner, Altmann & Köhler (1987: 136) präsentiert; es kann gezeigt werden, dass sie dem bereits vorgestellten Menzerath-Altmann-Gesetz (1) folgen (Best ²2003b: 20). Das gleiche Modell ist auch für die von Tuldava (1995: 135) wiedergegebenen Daten für 7jährige russische Kinder anwendbar (Best ²2003b: 20f.). Die Testergebnisse sind dokumentiert und werden daher hier nicht wiederholt. Sie bestätigen, dass Gesetzmäßigkeiten sich auch schon relativ früh im Spracherwerb einstellen können.

II.2. Wortartenverteilungen

Zählt man in einzelnen Texten aus, welche Wortart wie oft vorkommt, und bringt diese Wortarten dann in eine Rangordnung, so lässt sich zeigen, dass diese Rangordnungen ebenfalls bestimmten Modellen folgen (Best 2000, 2001a). Templin (1957: 101) stellt nun dar, wie die Verteilung von Wortarten bei Kindern im Alter von 3 bis 8 Jahren ist, und zwar einmal getrennt für alle verwendeten Wörter und weiter nur für die verschiedenen Wörter. Anders als in Best (2000, 2001a) liegen hier nur relative Werte vor, weshalb nicht die dort verwendeten Modelle benutzt werden können. Stattdessen erweist sich wiederum das Menzerath-Altmann-Gesetz in Form von Modell (2) für sämtliche Daten als geeignet. Zur Demonstration werden nur die Ergebnisse für die 3- und 8jährigen Kinder vorgestellt (Tabelle 8):

Tabelle 8
Wortartverteilung bei 3jährigen Kindern

	alle Wörter			nur verschiedene Wörter		
Rang	Wortart	n_x	$\hat{n}_x$	Wortart	n_x	$\hat{n}_x$
1	Verb	22.6	23.79	Substantiv	25.5	26.71
2	Pronomen	19.4	18.66	Verb	23.4	20.04
3	Substantiv	17.7	14.64	Pronomen	12.1	15.04
4	Adverb	10.0	11.49	Adverb	11.5	11.29
5	Vermischt	7.1	9.01	Adjektiv	8.8	8.47

6	Artikel	6.8	7.07	Vermischtes	7.6	6.36
7	Präposition	6.5	5.55	Präposition	5.8	4.77
8	Adjektiv	6.3	4.35	Artikel	2.2	3.58
9	Interjektion	2.1	3.41	Interjektion	2.0	2.69
10	Konjunktion	1.5	2.68	Konjunktion	1.1	2.02
	a = 30.3233	b = -0.2426	D = 0.95	a = 35.5931	b = -0.2871	D = 0.96

Tabelle 9
Wortartverteilung bei 8jährigen Kindern

	alle Wörter			nur verschiedene Wörter		
Rang	Wortart	n_x	$\hat{n}_x$	Wortart	n_x	$\hat{n}_x$
1	Verb	24.3	24.16	Substantiv	27.4	28.68
2	Pronomen	17.8	18.78	Verb	24.2	20.88
3	Substantiv	17.0	14.60	Adverb	12.4	15.19
4	Adverb	9.1	11.35	Adjektiv	11.9	11.06
5	Artikel	8.1	8.83	Pronomen	8.7	8.05
6	Präposition	7.9	6.86	Präposition	5.6	5.86
7	Adjektiv	7.4	5.34	Vermischtes	4.5	4.26
8	Konjunktion	3.7	4.15	Konjunktion	2.5	3.10
9	Vermischtes	2.9	3.23	Artikel	1.5	2.26
10	Interjektion	1.2	2.51	Interjektion	1.3	1.64
	a = 31.0708	b = -0.2517	D = 0.96	a = 39.4114	b = -0.3177	D = 0.97

Die folgende Tabelle zeigt, dass in allen Altersstufen das Modell (2) mit sehr guten Ergebnissen an die Beobachtungswerte für alle Wörter und für verschiedene Wörter angepasst werden kann (vgl. Tabelle 10):

Tabelle 10
Übersicht über die gesamten Testergebnisse

	alle Wörter	nur verschiedene Wörter
Alter	Ergebnis: D	Ergebnis: D
3	0.95	0.96
3.5	0.95	0.94
4	0.96	0.93
4.5	0.97	0.95
5	0.97	0.95
6	0.97	0.96
7	0.97	0.97
8	0.96	0.97

Zur Veranschaulichung wird in der folgenden Graphik (vgl. Abb. 8) dargestellt, wie gut beobachtete und berechnete Werte für alle Wörter bei 8jährigen Kindern übereinstimmen. (Die schwarzen Balken stehen für die beobachteten Werte.)

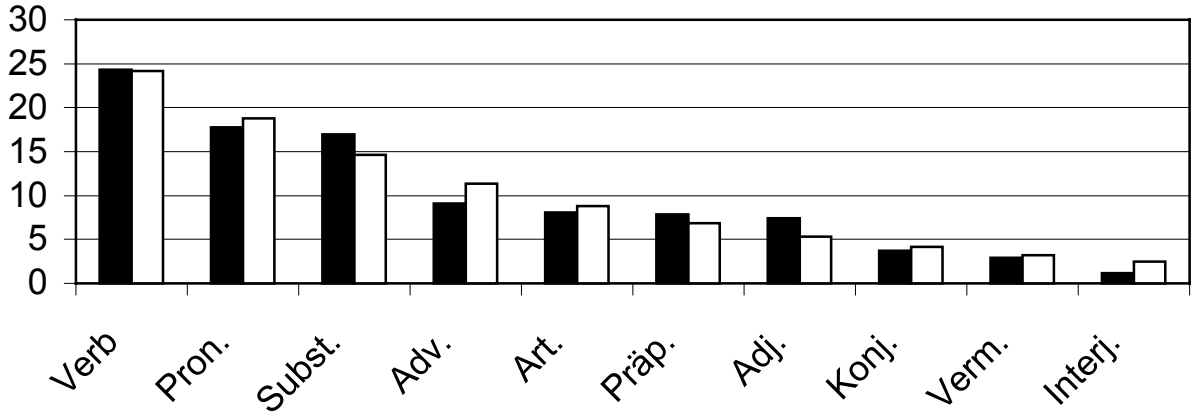


Abb. 8. beobachtete und berechnete Werte für die Wortartanteile bei 8jährigen Kindern

II.3. Wortartverteilung eines 3jährigen Mädchens

Bühler ([4]1967: 162) präsentiert die Verteilung des Gesamtwortschatzes von 2282 Wörtern eines 3jährigen Mädchens auf die Wortarten; die Verteilung entspricht wieder Modell (2) (vg. Tabelle 11 und Abb. 9):

Tabelle 11
Wortartverteilung eines 3jährigen Mädchens

Wortart	Häufigkeit (beobachtet)	Proportion	Proportion (berechnet)
Substantiv	1171	51.31	53.03
Verb	732	32.08	25.94
Adjektiv	198	8.68	12.69
Adverb	98	4.29	6.21
Pronomen	36	1.58	3.03
Präposition	20	0.88	1.48
Interjektion	15	0.66	0.73
Konjunktion	12	0.53	0.36
a = 108.4253 b = -0.7152 D = 0.97			

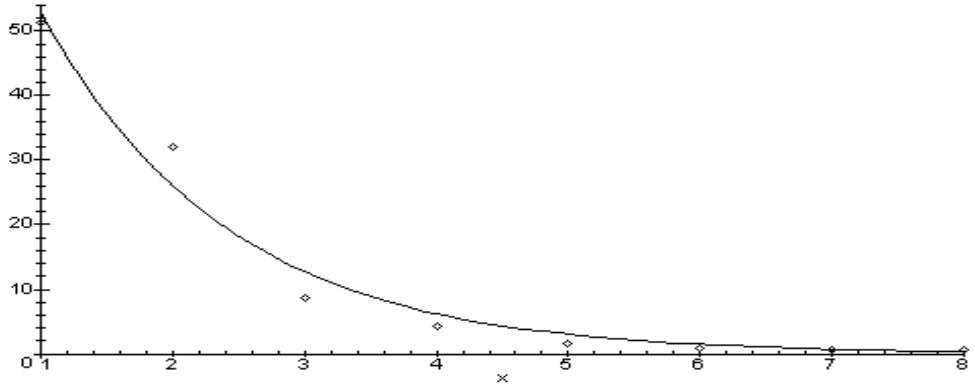


Abb. 9. Beobachtete und berechnete Werte für die Wortarten eines 3jährigen Mädchens

II.4. Wortschatzbeherrschung im Schuleintrittsalter

Weiterhin soll eine Querschnittstudie zur Beherrschung des Wortschatzes in einem bestimmten Lebensabschnitt vorgestellt werden. Auch in diesem Fall lässt sich nachweisen, dass die Ergebnisse sich als gesetzmäßig erweisen. Dabei geht es um die Frage, wie viele Wörter (aktiver Wortschatz) ein Kind bei seiner Einschulung beherrscht? Dazu vermittelt Augst (Hrsg., 1984: XV) einen ersten Eindruck; an seine Daten wurde mit Hilfe des Altmann-Fitters (1997) die für Rangordnungen bewährte 1-verschobene negative hypergeometrische Verteilung (vgl. Grzybek, Kelih & Altmann 2004) angepasst (vgl. Tabelle 12 und Abb. 10):

$$(4)\quad P_x = \frac{\binom{-M}{x-1}\binom{-K+M}{n-x+1}}{\binom{-K}{n}}, \qquad x = 1, 2, \ldots, n+1$$

Tabelle 12
Wortschatzbeherrschung im Schuleintrittsalter

Rang	Kind; Beruf der Eltern (Vater, Mutter)	n_x	NP_x
1	Christian; Hochschullehrer, Realschulleherin	5300	5328.06
2	Wolfgang; Hochschullehrer, Realschullehrerin	4380	4473.53
3	Carsten; Chemiker, Germanistikstudentin	4215	4104.76
4	Charlotte; Bauingenieur, Kindergärtnerin	4015	3871.29
5	Kristina; Hauptschullehrer, Bibliothekarin/Germanistikstudentin	3630	3698.72
6	Philipp; Hochschullehrer; Psychologiestudentin	3525	3558.89
7	Gero; Dipl. Wirtschaftsingenieur, Sozialpädagogin	3460	3437.25
8	Mirko; Schreiner, Hausfrau/ Lehrerin	3215	3323.48
9	Eva; Pastor, Hausfrau/ Lehrerin	3200	3205.28
10	Alexandra; Landwirt, Hausfrau (Marokkanerin)	3110	3048.74
$K = 1.8757$ $M = 0.8427$ $n = 9$ $X^2 = 16.929$ $FG = 6$ $C = 0.0004$			

K, M und *n* sind die Parameter des Modells, *C* ist der Diskrepanzkoeffizient, der mit $C \leq 0.01$ ein sehr gutes Ergebnis anzeigt. Die Untersuchung galt 6jährigen Kindern kurz vor Schuleintritt; es wurden in natürlichen Gesprächssituationen alle aktiv verwendeten Wörter einschließlich der Neubildungen erhoben.

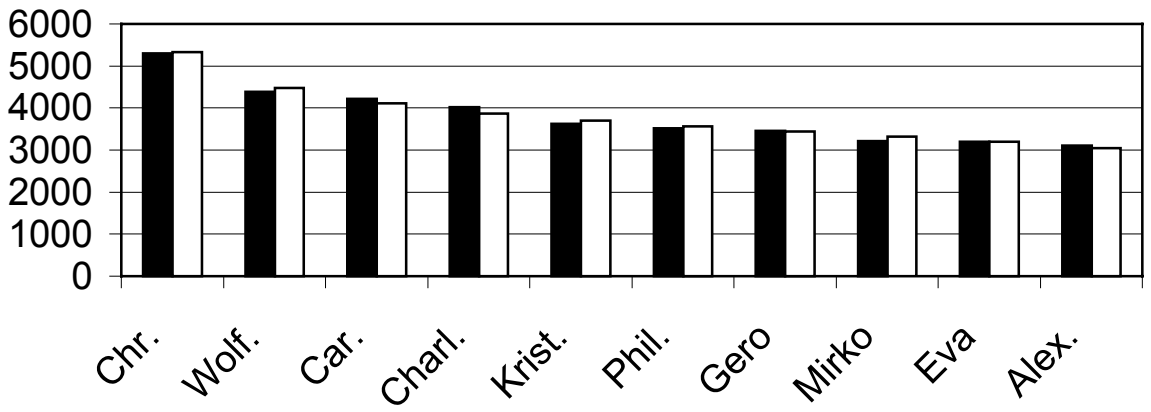


Abb. 10. Wortschatzbeherrschung im Schuleintrittsalter

II.5. Textumfang bei Kindern verschiedener sozialer Schichten

Pregel & Rickheit (1975: 45) geben eine Übersicht über Textlängen (nach der Zahl der Wörter pro Text) bei Kindern aus fünf Sozialschichten, die von der Oberschicht bis zur unteren Unterschicht als Extremen reichen. Es handelt sich um Durchschnittswerte für das Gesamtkorpus mit einer Altersstreuung von 6;0 – 10 Jahren. Mit Modell (2) erreicht man eine überzeugende Anpassung (vgl. Tabelle 13 und Abb. 11):

Tabelle 13
Textumfang nach Sozialschichten

Sozialschicht	$\text{Textlänge}_{\text{beob.}}$	$\text{Textlänge}_{\text{ber.}}$
1	103.8	107.54
2	97.4	93.39
3	82.0	81.10
4	75.2	70.42
5	54.9	61.16
a = 123.8335 b = -0.1411 D = 0.94		

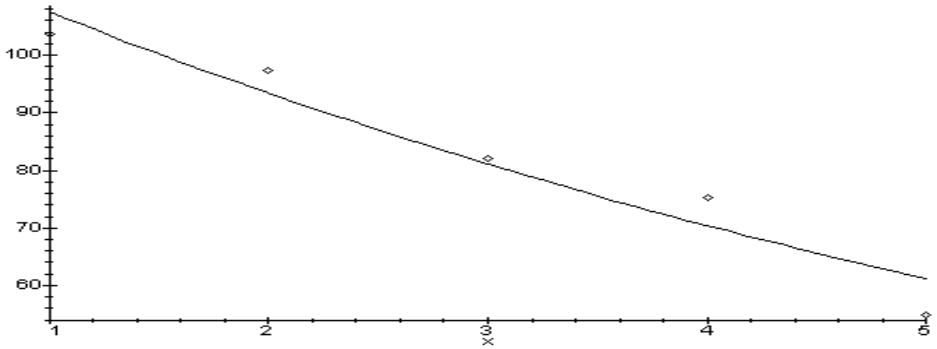


Abb. 11. Textumfang nach Sozialschichten

Zusammenfassung und Perspektive

Die Untersuchung zeigt, dass sowohl im Prozess des Spracherwerbs als auch in den Zuständen, die mit einem bestimmten Alter erreicht werden, in aller Regel gesetzmäßige Verläufe bzw. Entwicklungen nachweisbar sind. Es sieht jedoch so aus, als ob die Spracherwerbsprozesse zwei recht unterschiedlichen Gesetzmäßigkeiten folgten. Man kann zwar auch für den Lauterwerb das logistische Modell anpassen; es zeigen sich jedoch für die ersten Anfänge des Lauterwerb erkennbare systematische Abweichungen, die bei Verwendung von Modell (1) nicht vorkommen.

Ein Problem tritt bei der Modellierung der Zunahme der Wortlänge auf. Aufgrund der vergleichsweise ungünstigen Datenlage darf vermutet werden, dass dieses Problem sich mit verbesserten Erhebungen überwinden lassen sollte. Vor allem eine Ausdehnung der Daten auf weitere Altersklassen bei weniger künstlicher Kommunikationssituation und höherer Anzahl von Probanden könnte zu Verbesserungen führen. Die Einbeziehung eines Wertes von 1.700 Silben pro Wort für Erwachsene im Alter von 35 Jahren ist dagegen nur eine Notlösung, die mit c = 2.1795 einen für das Deutsche wohl etwas zu hohen Zielwert erbringt.

Wechselt man die Perspektive und untersucht, wie sich in einem bestimmten Alter die sprachlichen Leistungen verteilen, so kommt man offenbar immer wieder zu der Erkenntnis, dass diese sich gesetzmäßig verteilen.

Prinzipiell unterliegt also auch der Spracherwerb den Gesetzen, die bereits in anderen Zusammenhängen für das Sprachsystem und seine Verwendung entwickelt wurden.

Literatur

Altmann, Gabriel (1980). Prolegomena to Menzerath's Law. In: Grotjahn, Rüdiger (ed.), *Glottometrika 2* (S. 1-10). Bochum: Brockmeyer.

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

Augst, Gerhard (Hrsg.) (1984). *Kinderwort. Der aktive Wortschatz (kurz vor der Einschulung) nach Sachgebieten geordnet mit einem alphabetischen Register.* Frankfurt u.a.: Peter Lang.

Best, Karl-Heinz (2000). Verteilungen der Wortarten in Anzeigen. *Göttinger Beiträge zur Sprachwissenschaft 4, 37-51.*

Best, Karl-Heinz (2001). Kommentierte Bibliographie zum Göttinger Projekt. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten* (S. 284-310). Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (2001a). Zur Gesetzmäßigkeit der Wortartenverteilungen in deutschen Pressetexten. *Glottometrics 1, 1-26.*

Best, Karl-Heinz (2001b). Wortlängen in Texten gesprochener Sprache. *Göttinger Beiträge zur Sprachwissenschaft 6, 31-42.*

Best, Karl-Heinz (2003a). Zur Entwicklung von Wortschatz und Redefähigkeit bei Kindern. *Göttinger Beiträge zur Sprachwissenschaft 9, 7-20.*

Best, Karl-Heinz (²2003b). *Quantitative Linguistik. Eine Annäherung*. 2., überarbeitete und erweiterte Auflage. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (2003c). Spracherwerb, Sprachwandel und Wortschatzwachstum in Texten. Zur Reichweite des Piotrowski-Gesetzes. *Glottometrics 6, 18-43.*

Bühler, Charlotte ([4]1967). *Kindheit und Jugend. Genese des Bewusstseins.* Darmstadt: Wissenschaftliche Buchgesellschaft.

Busemann, Adolf (1925). *Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik. Sprachstatistische Untersuchungen.* Jena: Verlag von Gustav Fischer. Teildruck in: Helmers, Hermann (Hrsg.) (1969), *Zur Sprache des Kindes* (S. 1-59). Darmstadt: Wissenschaftliche Buchgesellschaft.

Chen, Han Piao, & Irwin, Orvis C. (1946). Infant Speech Vowel And Consonant Types. *Journal of Speech Disorders 11, 27-29.*

Deußing, Hans (1927/ 1969). Der sprachliche Ausdruck des Schulkindes. In: Helmers, Hermann (Hrsg.), *Zur Sprache des Kindes* (S. 60-131). Darmstadt: Wissenschaftliche Buchgesellschaft.

Fischer, Hardi ([4]1971). Entwicklung und Beurteilung des Stils. In: *Mathematik und Dichtung. Versuche zur Frage einer exakten Literaturwissenschaft* (S. 171-183). Zusammen mit Rul Gunzenhäuser hrsg. von Helmut Kreuzer. 4. durchgesehene Auflage. München: Nymphenburger.

Fucks, Wilhelm (1968). *Nach allen Regeln der Kunst.* Stuttgart: Deutsche Verlags-Anstalt.

Grohnfeldt, Manfred (1980). Erhebungen zum altersspezifischen Lautbestand bei drei- bis sechsjährigen Kindern. *Die Sprachheilarbeit 25, 169-177.*

Grzybek, Peter, Kelih, Emmerich, & Altmann, Gabriel (2004). Graphemhäufigkeiten (am Beispiel des Russischen). Teil II: Modelle der Häufigkeitsverteilung. *Anzeiger für Slavische Philologie XXXII, 25-54.*

Hunt, Kellogg W. (1965). A Synopsis of Clause-to-Sentence Length Factors. *English Journal 54, 300-309.*

Irwin, Orvis C., & Chen, Han Piao (1946).Development of Speech During Infancy: Curve of Phonemic Types. *Journal of Experimental Psychology 36, 431-436.*

Irwin, Orvis C. (1947).Development of Speech During Infancy: Curve of Phonemic Frequnecies. *Journal of Experimental Psychology 37, 187-193.*

Irwin, Orvis C. (1948). Infant Speech: The Effect of Family Occupational Status and of Age on Sound Frequency. *Journal of Speech and Hearing Disorders 13, 320-323.*

Lenneberg, Eric H. (1972). *Biologische Grundlagen der Sprache.* Frankfurt: Suhrkamp.

McCarthy, Dorothea (1954). Language Development in Children. In: Carmichael, Leonard (ed.), *Manual of Child Psychology. 2nd Ed.* (S. 492-630). New York: Wiley & Sons/ London: Chapman & Hall.

Oksaar, Els (1977). *Spracherwerb im Vorschulalter. Einführung in die Pädolinguistik.* Stuttgart/ Berlin/ Köln/ Mainz: Kohlhammer.

Pregel, Dietrich, & Rickheit, Gert (1975). *Kindliche Redetexte. Variablentypische Auswahl aus einem Korpus zur Sprache des Grundschulkindes.* Düsseldorf: Schwann.

Rest, Walter, Brose, Karl, Heitkämper, Peter, & Neumann, Siegfried (1977). *Wortschatzuntersuchung: Das normale Kind.* Opladen: Westdeutscher Verlag.

Smith, Madorah E. (1935). A Study of Some Factors Influencing the Development of the Sentence in Preschool Children. *Journal of Genetic Psychology 46, 182-212.*

Stormzand, Martin J., & O'Shea, M.V. (1924). *How much English Grammar?* Baltimore: Warwick & York.

Templin, Mildred C. (1957). *Certain Language Skills in Children. Their Development and Interrelationships.* Minneapolis: The University of Minnesota Press.

Tuldava, Juhan (1995). *Methods in Quantitative Linguistics.* Trier: Wissenschaftlicher Verlag Trier.

Wagner, Klaus R., Altmann, Gabriel, & Köhler, Reinhard (1987). Zum Gesamtwortschatz der Kinder. In: Wagner, Klaus R. (Hrsg.), *Wortschatz-Erwerb* (S. 128-142). Bern u.a.: Peter Lang.

Software

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

NLREG*. Nonlinear Regression Analysis Program*. Ph. H. Sherrod. Copyright (c) 1991 - 2001.

Glottometrics 12, 2006, 55-62

Sentence length as a feature of style
(applied to works of German writers)

Tetiana Dzhuriuk[1], Chernivtsi

Abstract. Sentence length in German is studied in three formal length categories and three genres. We try to characterize the individual writer by means of a vector of properties which can be used both for classification and the study of development.

Keywords: German, style, sentence length

1. Introduction

Since the work of Sherman (1888), the study of style has continuously attracted the attention of linguists. Several scholars (cf. Vinogradov 1961; Grigorev 1983; Domašnev 1983; Kucharenko 2002, Chrapčenko 1976; Elsberg 1965; to mention only some eastern European writers) consider an author's (or individual artistic) style to be a matter of the writer's skill. They suggest that, by means of style, a writer expresses their creative individuality and their point of view. The character of that individual style is distinctly expressed by the means and methods of composition and the architectonics of the text (Chrapčenko 1976: 118). Unfortunately, even this delimitation of possibilities permits a practically infinite choice of description means. Consequently, researchers must restrict their study to a few features or concentrate on just one. Even this restriction does not yield unique results, because text features represent conceptual constructs created by us, and concept formation may be different for every researcher.

Though there have been a large number of studies concerned with the external aspects of text, a statistical study of the frequency of different linguistic elements could help to specify the concept of functional style and to identify some further specific features of the author's idiolect.

In fact, although the individual style of an author depends to a great extent on their talent (not relevant from the linguistic point of view), it depends still more on their writing habits – which can be scrutinized analytically. A writer uses the national language but they choose and combine the elements of the word stock and grammar in accordance with their literary aim and their habits – if there are any. The dialect of an author is a selection of a set of means which enable them to construct specific schemes. The study of these schemes sometimes necessitates different research methods. Since individual elements of language and their co-occurrences are characterized by certain frequencies of occurrence, written and spoken texts can be investigated and described using the methods of mathematical statistics. Using quantitative methods, one can discover some aspects of the "syntactic feather" of a writer.

This study investigates sentence length distribution in German prose at the beginning of the 20th century. Since we are studying the individual style, we can refrain from an analysis of the probability distribution of sentence length and concentrate on a raw class formation

[1] Address correspondence to: Tetjana Dzurjuk: uatanya@ukr.net

concerning sentence length. For our study we have chosen novels and stories by K.Tucholsky, T.Mann, F.Kafka, I.Keun. The study is based on the following works:

Kafka, F. (1968). Das Schloß. Frankfurt am Main und Hamburg: Fischer Bücherei GmbH.

Kafka, F. (1969). Das Urteil und andere Erzählungen. Frankfurt am Main und Hamburg: Fischer Bücherei GmbH.

Keun, I. (1994). Das Kunstseidene Mädchen. – München, Deutscher Taschenbuch Verlag GmbH & Co. KG.

Keun, I. (1982). Das Mädchen, mit dem die Kinder nicht verkehren durften. Moskau: Verlag Progress.

Mann, T. (1973). Buddenbrooks. Verfall einer Familie. Bukarest: Kriterion Verlag.

Mann, T. (1975). Das Wunderkind. In: Erzählungen. Leipzig: Verlag Philipp Reclam jun., 81-92.

Mann, T. (1975). Schwere Stunde. In: Erzählungen. Leipzig: Verlag Philipp Reclam jun., 214-224.

Tucholsky, Kurt (1962). Der kranke Zeisig. In: Tucholsky. Ein Lesebuch für unsere Zeit. Weimar: Volksverlag, 34-41.

Tucholsky, Kurt (1962). Die Erdolchten. In: Tucholsky. Ein Lesebuch für unsere Zeit. – Weimar: Volksverlag, 279-289.

Tucholsky, Kurt (1962). Paris 1924–1927, I– IV. In: Tucholsky. Ein Lesebuch für unsere Zeit. – Weimar: Volksverlag, 105 -116.

Tucholsky, Kurt (1991). Schloß Gripsholm: Eine Sommergeschichte. Reinbek bei Hamburg: Rowohlt.

All these works were written in the first half of the 20th century (up to 1940).

The sentence is the major syntactic unit of text. There are basic differences between spoken and written language in terms of the structure, completeness and length of sentences. However, in fiction the characteristics of the norm are superimposed by an individual author's features; hence, the hypothesized standards are distorted by some boundary conditions.

As sentences are important elements of text structure, the following hypothesis can be set up: *sentence length distribution is invariant in different texts by the same author and represents a characteristic of that author's style*. Here, it is not the mean sentence length but the form of the distribution that is key.

Most researchers assert that it is necessary to partition text into three components, namely author's speech (AS), language of characters (LC) and reported speech (RS). These components are part of the boundary conditions. If they are *a priori* isolated, it is easier to treat them as the ceteris-paribus condition.

Sentence size can be considered a function of its structure. It becomes evident when we compare the mean sentence length and the number of simple and complex sentences in texts. There is a direct dependence between the mean length of the sentences in a text and the number of complex sentences in it (Lesskis 1963: 106; Šubik 1999: 76).

Among many other factors which determine the length of the sentence in prose, the main factor is the author's perception of the world, their attitude toward the represented characters and events, and aesthetic principles (Sil´man 1967: 7). These are the factors that can account for the uniformity of sentence length, which is often observed in different works by one author or in works representative of a certain literary trend. It must be remembered that identical sentences lengths can result from different, sometimes contradictory conditions: for instance, a short sentence can represent both laconic brevity of wisdom and narrowness of outlook.

Considerable variations in the mean sentence length in different works by one and the same author can be signs of important changes, in author's style and in the subject; this can serve as an index of the creative evolution of the writer. Thus, sentence length is embedded in

a net of thematic, compositional, stylistic and structural characteristics of a text. A change of structure leads to increase or decrease in sentence length; but there is no one-to-one correspondence between simple structure and short length or complex structure and large length. That is, the dependence is not functional but stochastic.

Following some linguists (e.g. Kucharenko) we divide all sentences into short (up to 10 words), middle (up to 30 words), and long (up to 60 words) and super-long ones (anything greater, with no upper limit) (Kucharenko 2002: 66). The German researcher Braun (1993: 105) suggests another partitioning: short sentences (1-4, 5-8, 9-12); middle sentences (13-16, 17-20); long sentences (21-24, more than 24). The lower limit of sentence length is one word; there is no upper limit, as theoretically a written sentence of any length could exist (cf. Andersen 2005). We are aware that any such subdivision is arbitrary. Those researchers who fit distributions to sentence length data use the intervals 1-5, 6-10, 11-15,…

Long and very long sentences are used less often than middle and short sentences; however, the appearance of even one such sentence influences the text's mean length of sentence. By *mean length of sentence*, we refer to the relationship of the number of words in the text under study to the number of sentences in that text. It is specific for every text; obviously it is a characteristic of the author's style or genre. Mean length of sentence is not the same in texts by different authors, but it can also vary in different texts by one author. This is caused by the fact that sentence length has a very great variance, it always represents a long-tail distribution; so the arithmetic mean does not yield a sufficient, stable characteristic. For stylistic purposes, the geometric mean should be used instead; this reduces the weight of any outliers, or the parameters of the theoretical distribution.

The more dialogues are in a text, the shorter the mean length of sentence appears to be, and vice versa. Philosophical works which express the internal world of a character generally have a higher mean length of sentence. Mean length of sentence in literary texts can be used to characterize the author's style; but since it is a very unstable quantity, in every case at least the variance of length should be taken into account.

In the texts investigated here, certain types were selected: description, narration and argumentation, which contained author's speech, speech of characters, and reported speech. These types yield a sufficiently comprehensive picture of the language of artistic texts of each author in relation to the length of sentence.

Not only individual genres, works or authors, but also different historical epochs may display differences in sentence length (cf. Wittek 2001).

Some researchers assert that the dominance of short sentences in modern German prose is determined by the influence of the spoken language (Schneider 1969: 434; Altmann 1988: 96). Of course, the importance of spoken language is great, but it is not a unique or even basic factor. Short sentences are widely used in modern literature to show those unnoticeable everyday sides of reality, which play an important role in the integration of events. A second important factor is a trend toward artistic laconism; instead of a verbose description of an object's characteristic features, conspicuous or expressive details are given. The shortness of sentences can even result from contrary factors like emotional saturation and factual dryness which are components of modern prose (Šubik 1969: 81).

Evidently, the above mentioned factors work in different texts to different degrees and combine with other factors, which must be considered as *ceteris paribus*.

2. Analysis

For our purposes we took a sample of 10513 sentences from texts of four German authors. The sentences of the sample were divided into four types:

short (to 10 words), /**S**/
middle (to 30 words), /**M**/
long (to 60 words), /**L**/
and super-long (more than 60 words) /**SL**/.

The absolute frequencies of these types are presented in Table 1.

Table 1
Frequency of sentences in prose (to 1940-s)

	Kafka	Keun	Tucholsky	T.Mann	All authors
short	665	1331	2842	872	5710
middle	616	1478	1124	669	3887
long	192	239	180	189	800
super-long	64	31	14	7	116
Number of sentences	1537	3079	4160	1737	10513

We are aware that any partitioning of the variable is artificial; but even keeping the original values would yield a variable with many zero frequency classes. Hence any partitioning is as good as any other. To compare the writers we have two possibilities: (a) we evaluate the homogeneity of the distributions using the common chi-square test for independence, or (b) if we want to know what is characteristic for individual writers, we perform a series of tests for individual cells of the contingency tables.

In case (a) we compute the criterion

$$(1) \qquad X^2 = \sum_{i=1}^{4}\sum_{j=1}^{4}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

where O_{ij} are the observed and E_{ij} the expected values computed in the usual way. The result, $X^2 = 1027.5$ considerably exceeds the critical value $\chi^2_9 = 16.92$ at $\alpha = 0.05$; $\chi^2_9 = 21.67$ at $\alpha = 0.01$. Thus, there are considerable differences between the frequencies shown in Table 1.

In case (b), in order to localize the cells which diverge from the expectation and constitute the basis of the individual style, we perform a test for each cell separately ignoring Bonferroni´s rule. For each cell we use the criterion

$$(2) \qquad z = \frac{O_{ij}-E_{ij}}{\sqrt{n_i.n._j(n-n_i.)(n-n._j)/[n^2(n-1)]}}$$

where $n_i.$ is the marginal sum of the row i in Table 1, $n._j$ is the marginal sum of the column j and n is the total sum. Following Altmann et al. (2002: 30ff) we make the following decisions:

(i) if $z \geq 1.645$, we consider the class as *preferred* (*P*)
(ii) if $-1.645 < z < 1.645$, we consider the class as *neutral* (N)
(iii) if $z \leq -1.645$, we consider the class as *avoided* (A)

For example, for the cell "Kafka-short" we get (using $n^2(n\text{-}1) = n^3$)

$$z = \frac{665 - 5710(1537)/10513}{\sqrt{5710(1537)(10513-5710)(10513-1537)/10513^3}} = \text{-9.41.}$$

Since –9.41 < -1.645 we consider this cell as avoided, i.e. Kafka displays a significant tendency to avoid short sentences. The results for the other cells are shown symbolically in Table 2a,b.

Table 2a
Individual sentence-length tendencies with 4 German authors

	Kafka	Keun	Tucholsky	Mann
short	-9.41	-14.68	23.32	-3.77
middle	2.73	15.08	17.11	1.46
long	7.81	0.38	-10.27	5.63
super-long	12.43	-0.61	-6.09	-3.06

Table 2b
Qualitative sentence length tendencies

	Kafka	Keun	Tucholsky	Mann
Short	A	A	P	A
Middle	P	P	A	N
Long	P	N	A	P
Super-long	P	N	A	A

Note that these tendencies can only be claimed to exist within the class of these four authors. If we compared more German authors the results could be quite different. In any case a characterization is possible: Tucholsky prefers short sentences, Keun prefers sentences of middle length, Mann prefers long sentences and Kafka avoids short sentences. Though the partitioning of the contingency table is not correct (the sum of squares of z-values is not equal to the chi-square for independence), the evaluation of individual trends is more conspicuous than can be achieved by any other method.

3. The impact of text type

A question arises: is there any system in the use of certain types of sentences by the author? Obviously, the desire to bring the language of their works nearer to spoken language and at the same time the pressure of literary tradition leads to a considerable increase of number of short sentences in Tucholsky's works, especially in dialogue. These unexpected results can be explained by more frequent use of dialogue with comparatively short remarks by characters and short authorial sentences introducing direct speech. To investigate this point, we partitioned the texts into three components:

author's speech (AS),
speech of characters (SC)
and reported speech (RS).

The distribution of these three components in the texts under investigation is shown in Table 3.

Table 3
Frequency of the use of sentences in different components of the text

Speech	KAFKA	KEUN	TUCHOLSKY	T. MANN	Prose
AS	837	2504	2540	973	6854
SC	681	426	1603	748	3458
RS	19	149	17	16	201
Sum	1537	3079	4160	1737	10513

An overall chi-square test shows that the authors differ in using the three speech types (X^2_6 = 999.2). Again, we can test the individual cells to ascertain the preference of the authors for a special type of speech. Using the above test we obtain the symbolic results in Table 4.

Table 4
Qualitative speech-type tendencies

	Kafka	Keun	Tucholsky	T. Mann
AS	A	P	A	A
SC	P	A	P	P
RS	A	P	A	A

The result is, again, valid only if these four authors are compared, not generally. In any case, the result shows that there exist differences among these authors.

The interaction of speech type and sentence length can be tested in many different ways. But even if we set up three-dimensional contingency tables and use appropriate software, the result remains relative to the group of writers under examination. Adding further authors would require, each time, new computations. Hence, we decided to test each author separately and characterize that author with a vector of symbols. In Tables 5a to 5d the sentence-length classes are shown in relation to speech types for each author separately.

Table 5a
Speech types and sentence lengths with Kafka

	short	middle	long	super-long
AS	326	367	108	36
SC	338	244	76	23
RS	1	7	6	5
Sum	665	618	190	64

Table 5b
Speech types and sentence lengths with Keun

	short	middle	long	super-long
AS	1099	1185	204	16
SC	223	173	20	10
RS	10	19	5	5
Sum	1332	1477	239	31

Table 5c
Speech types and sentence lengths with Tucholsky

	short	middle	long	super-long
AS	1555	821	152	12
SC	1285	295	23	-
RS	1	9	5	2
Sum	2841	1125	180	14

Table 5d
Speech types and sentence lengths with Kafka

	short	middle	long	super-long
AS	338	466	162	7
SC	526	197	25	-
RS	8	6	2	-
Sum	872	669	189	7

For testing individual cells we use formula (2). It is not necessary to show the whole computation. The results can again be interpreted qualitatively (as P, N, A) and for each author we can set up a vector in the form

$$\text{Author} = \langle AS_{short}, AS_{middle}, AS_{long}, AS_{super\text{-}long}, SC_{short}, \ldots, RS_{super\text{-}long} \rangle$$

where the elements can be replaced either by the resulting *z*-values or simply by qualitative decisions (P,A,N). We choose the second alternative and obtain:

Kafka = <A,P,N,N,P,A,N,N,A,N, P,P>
Keun = <A,P,P, A,P,A,A,P, A,N,N,P>
Tucholsky = <A,P,P, N,P,A,A,A,A, P, P,P>
T. Mann = <A,P,P, P, P,A,A,A,N,N,N,N>.

As can be seen, the authors differ in these two dimension (length, speech); no two authors have identical vectors. We might add further dimensions to these vectors, if necessary; it is possible to perform any kind of classification. A similarity measure between the vectors can easily be set up. Restricted to the specified properties, we can distinguish genres and within a genre we can follow the development of texts in time. An extended study of German literature is in preparation.

References

Admoni, W. (1966). *Razvitie struktury predloženija v period formirovanija nemeckogo nacionalnoho jazyka.* Leningrad: Nauka.

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

Altmann, G. , Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie.* Göttingen: Peust & Gutschmidt.

Altmann, G., Schwibbe, M.H. (1989). *Das Menzeratsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim, Zürich, New York: Georg Olvs Verlag.

Andersen, S. (2005). Word length balance in texts. *Glottometrics 11, 32-50.*

Chrapčenko, M. (1976). Tvorča indyvidualnist pysmennyka i rozvytok literatury. Kyjiv: Dnipro.

Fleischer W., Michel G. (1979). *Stilistik der deutschen Gegenwartssprache*. Leipzig: VEB Bibliographisches Institut.

Kucharenko,V. (2002) *Interpretacija teksta: Učebnik dlja studentov filologičeskich specialnostej.* Odessa: Latstar.

Levickij,V. (2004). *Kvantitativnye metody v lingvistike*. Chernovzy: Ruta.

Lesskis, G. (1963). O zavisimosti meždu razmerom predloženija i charakterom teksta. In: *Voprosy jazykoznanija 12, 3, 92-112.*

Schneider W. (1969). *Stilistische deutsche Grammatik*. Freiburg, Basel, Wien: Herder.

Sherman, L.A. (1888). Some observation upon the sentence-length in English prose. *University of Nebraska Studies 1, 119-130.*

Sil´man T. (1967). Problemy sintaksičeskoi stilistiki. Leningrad**:** Prosvesčenie.

Šubyk S. (1969). Razmer predlozheniya v nemezkoi khudozhestvennoi proze. In: *Sbornik statej po metodike prepodovanija inostrannych jazykov i filologii 4*, *77-79* (Leningrad).

Vašak, P. (1974). Dlina slova i dlina predloženija v tekstach odnoho avtora In: *Voprosy statističeskoi stilistiki 314-329*. Kiev: Naukova dumka.

Vinogradov, V. (1961). *Problema avtorstva i teorija stilji*. Moskva: Nauka.

Wittek, M. (2001). Zur Entwicklung der Satzlänge im gegenwärtigen Deutschen. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 219-247*. Göttingen: Peust & Gutschmidt.

Glottometrics 12, 2006, 63-74

A logistic regression model of variant preference in Japanese kanji: an integration of mere exposure effect and the generalized matching law

Shoichi Yokoyama, Yukiko Wada[1]
The National Institute for Japanese Language, Tokyo

Abstract: The word *hinoki* or 'cypress' can be transcribed in two variant forms, 檜 (the so-called "traditional" variant) and 桧 (the "simplified" variant), in Japanese kanji. Such variant forms are called *kanji variants*. The present paper reviews a series of studies on Japanese kanji recognition (Yokoyama, 2006a, 2006b, 2006c), and proposes a model which accounts for performance in a preference judgment task based on kanji frequency data. Yokoyama (2006a) administers preference judgment task in which the participants were presented with 263 pairs of traditional and simplified variants and asked to choose the more preferable variant of each pair. The analyses indicate a positive contribution of frequency to variant preferences, supporting the so-called "mere exposure effect" theory of Zajonc (1968). This finding leads to a logistic regression model that describes preference behavior in kanji recognition, based on Fechner's law. Yokoyama (2006b) shows that the model is comparable to the so-called "the generalized matching law" of Baum (1974) and to "the ideal free distribution theory" of Fagen (1987). Yokoyama (2006c) further examines the predictive validity of the model with empirical data obtained from a preference judgment task, administered in the Tokyo and Kyoto areas. Logistic regression analyses are performed with the ratio of preference for the given variants and the logit of the character frequencies, yielding significant correlations between the predicted probabilities and the observed responses (r = .804 for Asahi newspaper data). The present paper synthesizes these studies and proposes a logistic regression model that efficiently describes preference behavior in Japanese kanji recognition, integrating the theoretical perspectives of mere exposure effect and the generalized matching law.

Key words: mere exposure effect, Fechner's law, generalized matching law, logistic regression analysis, variation theory, kanji, variants, preference, familiarity

1. Variant Forms in Japanese Kanji Characters

1.1. Kanji Variants in Japanese Language

Pairs of kanji characters that share the same meaning and pronunciation but exhibit varieties in their visual forms are called *variants*. Variants are commonly found in Japanese kanji characters. For example, the pair of kanji "桧" and "檜", both representing "cypress", has two variants: the traditional "檜" and the simplified "桧". Very few studies have provided insights into familiarity of and preference for orthographic variants with inter-disciplinary perspectives (Sasahara & Yokoyama, 2000). Yokoyama (2006a) employs a two-alternative

[1] Address correspondence to: Shoichi Yokoyama or Yukiko Wada, The National Institute for Japanese Language, Midori-cho, Tachikawa-shi, Tokyo, 190-8561, Japan.
E-mail: yokoyama@kokken.go.jp or ywada@kokken.go.jp.

forced-choice task, in which the participants are presented with pairs of kanji variants and asked to choose the item in each pair that is more preferable. The participants were instructed to perform the tasks assuming that they were word-processing with digital tools, such as computers or cell phones. The number of participants was approximately 200; they were college students in the Tokyo area. The data indicates that the participants' preference for variants are not attributable to graphic complexity or historical reasons, but to the frequencies of the given variants. Yokoyama (2006b) further analyzes the contribution of the frequency data, introducing a logistic regression model that accounts for performance in the preference judgment task based on frequency data obtained from a newspaper corpus.

1.2. Scope of this paper

The purpose of the present paper is to review related studies and to propose a logistic regression model that accounts for preference performance in kanji character recognition by integrating mere exposure effect and the generalized matching law. Yokoyama (2006a) demonstrates that character frequency data from Asahi and Yomiuri Newspaper corpora can account for performance in a preference judgment task in which 263 pairs of variants were presented to native speakers of Japanese. A subsequent study by Yokoyama (2006b) reveals the corresponding relationship between the logistic regression model, introduced by Labov (1972) in sociolinguistics, and the generalized matching law, commonly applied in animal behavioral research. Yokoyama (2006c) further proposes a quantitative model that accounts for performance in the preference judgment task, which is reviewed in this paper. Such a model is expected to provide an innovative framework to investigate recognition and use of orthographic variation, because linguistic activities involve choices among multiple alternatives, and because such choices are often based on preferences for certain forms.

1.3. Mere Exposure Effect in Social Psychology

A number of studies have shown that the preference mechanisms in human psychology involve memory and cognitive factors. Zajonc (1968), for example, has proposed the so-called "mere exposure effect" in social psychology. Based on the positive correlation between frequency and preference, in which high-frequency words are more preferred than low- frequency words, he asserts that repeated exposure to unfamiliar items increases preference frequency is a contributing factor for preference development (Zajonc, 1968).

Following the framework of mere exposure effect, various studies have shown that the mere exposure effect is observed regardless of the participants' awareness of the exposure. A method used by studies in brain research, among others, is to present kanji characters as primes and then examine whether the primed stimuli are preferred, in comparison to un-primed items. In Elliot & Dolan (1998), each of the twenty primes was presented for .05 second, followed by a masking stimulus for .45 second, ten times in all. The total time used to present the primes was 1 minute and 40 seconds. The participants were native speakers of English, completely unfamiliar with kanji. Thus, it was assumed, in the given condition, that the participants were not able to recognize the prime characters, and that they were even unable to perceive the visual representations of the kanji characters. However, the results showed that they preferred the primed stimuli to the un-primed counterparts, indicating the contribution of primes, i.e. exposure, to preference judgment performance.

1.4. Frequency, Preference, and Familiarity

Monin (2003) has shown that preference for certain items increases perceived familiarity, whether the exposure was conscious or unconscious. In his experiments, the participants judged their familiarity to stimuli that exhibited varying degrees of positive attributes. For example, the stimuli included photographs of beautiful-looking persons and real words that were semantically positive. The participants' task was to judge their perceived familiarity to such stimuli. Monin's analyses have shown that positive attributes of the stimuli, such as physical and semantic characteristics, contributed to performance in the familiarity inference task. In other words, items with positively preferred attributes were perceived as being more familiar.

The findings of these studies indicate that exposure frequency increases familiarity and that familiarity and preference are closely related to each other. Therefore, the present study assumes that exposure to kanji, which is operationally defined as frequency, contributes to preference, which is defined as performance in the preference judgment task in the study. It is also assumed that preference and familiarity are closely related, reflecting each other.

2. Fechner's Law and Psychophysical Model

2.1. A Model to Predict Familiarity Judgment Performance Based on Frequency

Yokoyama (2006a) proposes a model to account for performance in a variant preference task based on Fechner's law, which is often applied in perception studies in psychophysics. Fechner's law defines the perception degree as S, strength of stimulus as I, common logarithm as *log*, slope as constant K, and the intercept on the S axis as constant C as follows:

$$S = K \log I + C, \tag{1}$$

where Equation (1) is a linear combination equation. By analogy with Equation (1) above, kanji familiarity is expressed as follows:

$$Familiarity = K \log (frequency) + C, \tag{2}$$

Based on Equation (2), the familiarity of traditional variants, referred to as *FamTrad* hereafter, is expressed as follows:

$$FamTrad = K \log (FreqTrad) + C, \tag{3.1}$$

where the frequency of the given traditional variants in the newspaper corpora is defined as the *frequency of traditional variants* (*FreqTrad* hereafter), slope as K, and intercept on the S axis as C.

In the same way, the familiarity of simplified variants, referred to as *FamSimp* hereafter, is expressed as follows:

$$FamSimp = K \log (FreqSimp) + C, \tag{3.2}$$

where *FreqSimp* stands for the frequency of the simplified variants.

A value of 1 is added to the frequency data in advance to avoid zero values in logarithm computation. Namely, the frequency of a traditional variant is 1 plus the frequency of that

traditional variant, the frequency that of a simplified variant is 1 plus the frequency of that simplified variant, respectively, in the subsequent analyses. This conversion method is commonly employed in various studies.

2.2. A Model to Predict Preference Performance Based on Character Frequency

As discussed in Yokoyama (2006a, 2006b), it can be assumed that preference for variants is attributable to frequency differences between the given pair of variants. Preference for traditional variants, referred to as *PrefTrad*, can be described with Equation (4.1) as follows:

$$PrefTrad = a\ (FamTrad - FamSimp) + b, \tag{4.1}$$

where the explanatory variable is defined as the difference in familiarities between the traditional and simplified variants, the slope as a, and intercept as b.

When Equation (3.1) is substituted for the familiarity of traditional variants in Equation (4.1), and Equation (3.2) for the familiarity of simplified variants in Equation (4.1), Equation (4.2) as the following is available:

$$PrefTrad = a\ K\ \{\log (FreqTrad) - \log (FreqSimp)\} + b, \tag{4.2}$$

Equation (4.3) below expresses preference for traditional variants as follows:

$$PrefTrad = a\ K \log (FreqTrad\ /\ FreqSimp) + b, \tag{4.3}$$

where the frequencies of the variants are transformed to a logarithm of the frequency ratio. This model may be extended to multiple regression models with other plausible variables, such as stroke numbers. However, this study will mainly examine a simple regression model, namely Equation (4.3).

3. Empirical Validation of the Simple Regression Model

3.1. The Simple Regression Model and Empirical Evidence

This section empirically tests Equation (4.3), which is a simple regression model introduced in the previous section. The model, in theory, describes performance in a preference judgment task on character recognition in Japanese kanji.

3.2. Variant Preference Task

3.2.1. Materials

Yokoyama (2006a) selected stimuli from 263 pairs of traditional and simplified variants in the CD-ROM data of Sasahara & Yokoyama (2000). The original 263 pairs of variants were selected based on the number of variants that each character exhibits. Technical issues were also considered, so that the stimuli could be displayed and printed with the 83JIS standard, which is the 1983 version of the Japanese Industrial Standards. Of the original 263 pairs, 86 pairs of variants were selected for the purpose of the study. Figure 1 shows some sample stimuli used in the task. Presentation order was randomized in the task.

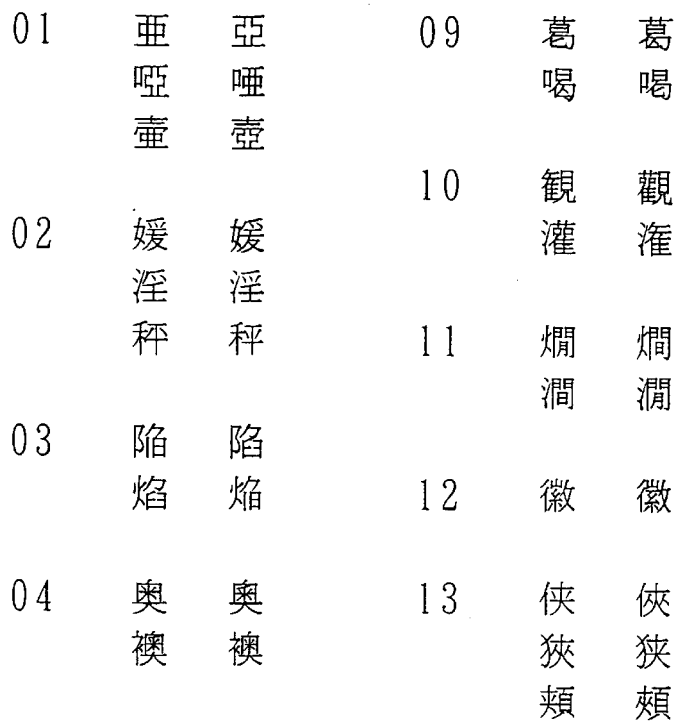


Figure 1. Sample stimuli in the preference judgment task

3.2.2. Task, instruction, and procedure

The task was a two-alternative forced-choice task in a paper-and-pencil format. The cover page of the task asked about demographic information and experience in word processing with digital tools. The task was administered as a part of instruction in classes at colleges.

The participants were asked to suppose that they were word-processing and to choose the variant of each pair, i.e. either traditional or simplified, that they would prefer (Yokoyama, 2006a, 2006b). Word processing was chosen as the context of the task in order to minimize the effects of non-target variables. Effects of economy, such as efficiency in hand-writing, for example, which may lead the participants to prefer graphically simpler forms, were presumably minimized in the task.

3.2.3. Participants

The data were obtained from two groups of participants, the Tokyo and Kyoto groups, both of which consisted of college students. Data used in the analyses were obtained from the participants who met the following criteria: (1) native speakers of Japanese; (2) relatively younger generation, i.e. 25 years old or younger; and (3) have experience in word processing and typing Japanese characters. The Tokyo group consisted of eighty-five female college students in the Tokyo area, who participated in the study in the academic year 1996-1997. The Kyoto group was male and female college students in the Kyoto area, twenty males and fifty-two females, seventy-two in total, who participated in the study in 1998.

3.3. Frequency Data from Newspaper Corpus

Frequency data from the Asahi Newspaper were obtained from two sources: (1) Chikamatsu, Yokoyama, Nozaki, Long, & Fukuda (2000); and (2) Yokoyama, Sasahara, Nozaki, & Long (1998). The frequency data were based on a corpus of the Asahi Newspaper (Asahi--Shinbun-Sha, 1994), which included the text of the morning and evening papers from January 1 through December 31, 1993. Based on this corpus, the researchers manually corrected the inconsistencies between the hard-copy representations and the digital data. The total number of the kanji characters included in the data, i.e. the number of tokens, was 17,117,320 and the number of types was 4,546.

3.4. Analyses and Results

Preference for traditional variants, in percentage form, was computed for the 86 pairs of variants according to Equation (4.3) (discussed above), which is a simple regression model with logarithms of frequencies. Table 1 summarizes some of the computation results based on the Tokyo group data along with the frequency data based on the Asahi Newspaper corpus.

Table 1
Preference for traditional variants (PrefTrad) and frequencies in Asahi Newspaper

Pair	PrefTrad %	Frequency		Pair	PrefTrad %	Frequency	
		Simplified	Traditional			Simplified	Traditional
亜亞	2.4	1035	5	蝿蠅	10.6	11	1
壷壺	74.1	59	20	竈竈	12.9	1	0
陥陷	8.2	1285	0	条條	24.1	9948	116
奥奧	1.2	2590	0	嬢孃	12.9	108	0
蛍螢	54.1	98	2	飲飮	1.2	2274	0
学學	7.1	54725	7	真眞	15.3	12248	149
誉譽	3.5	2198	0	慎愼	15.3	2724	2
鴬鶯	65.9	16	4	槙槇	44.7	230	2
鴬鷺	25.9	16	0	靭靱	32.9	17	2
会會	4.7	161051	7	尽盡	2.4	1175	2
桧檜	71.8	230	15	侭儘	30.6	0	0
覚覺	1.2	4990	3	数數	1.2	29439	2
撹攪	10.7	12	2	薮藪	40.0	81	27
観觀	0.0	7794	0	銭錢	9.4	2503	1
潅灌	84.7	11	2	賎賤	65.9	2	9
狭狹	17.6	948	0	曽曾	20.0	926	13
尭堯	31.8	72	45	騒騷	3.5	1619	0
焼燒	3.5	2553	1	捜搜	5.9	5404	0
区區	1.2	28396	0	沢澤	38.8	14489	643
欧歐	24.7	8001	0	駅驛	10.6	3315	1
経經	5.9	38698	9	訳譯	3.5	3161	1
頚頸	81.2	5	40	鈬鐸	45.9	0	38
倹儉	7.1	36	0	単單	1.2	6886	0

A regression analysis was performed with Equation (4.3) with the frequency data from Asahi Newspaper as follows:

$$PrefTrad = 10.30 \log (FreqTrad / FreqSimp) + 41.88, \quad (4.3a)$$

where the explanatory variable was defined as the logarithms of the frequency ratio of the two variants.

The results show a significant correlation between the logarithm of frequency ratio and the preference for traditional variants with $r = .73$ ($p < .01$, $df = 84$), accounting for 52.90% of the variance. However, this analytic method may produce values for the estimate probabilities that are negative, 100, or greater than 100, suggesting that it is statistically invalid. A solution for this problem is introduced in section 4.2, in which logistic regression analysis is discussed.

The same procedure was applied to Equation (5.1) in order to compare the predictive power of Equations (4.3) and (5.1) as follows:

$$PrefTrad = a\ K\ (FreqTrad - FreqSimp) + b, \quad (5.1)$$

where the frequencies are not transformed to logarithms. Equation (5.1) can be expanded to Equation (5.2) as follows:

$$PrefTrad = 22.31\ (FreqTrad - FreqSimp) + 0.00, \quad (5.2)$$

which yielded little correlation, with $r = .21$ ($p < .05$, $df = 84$) in fact, accounting only for 4.61% of the variance. As Equation (5.2) differs from Equation (4.3) in that the frequency data are not transformed into logarithms, this result indicates that the logarithm of frequencies has stronger predictive power than the frequency data *per se*.

4. Application of the Matching Law in Mathematical Linguistics

4.1. The Generalized Matching Law in Animal Psychology

Yokoyama (2006b) demonstrated that kanji frequency data from the Asahi newspaper corpus explained the performance in the preference task by applying the generalized matching law. A constructive extension of this line of research is to evaluate the theoretical contribution of the generalized matching law and the role of frequency in written language.

The basic principle of the matching law was proposed by Herrnstein (1961), which was initially applied to behavioral studies of animals, and was expressed as follows:

$$R1 / (R1 + R2) = r1 / (r1 + r2), \quad (6)$$

where R refers to responses and r refers to frequencies of reinforcers. The mathematical model of the generalized matching law developed by Baum (1974), which expresses the relationship between the reinforcers and the response allocation as follows:

$$(R1/R2) = B\ (r1/r2)^{S}. \quad (7)$$

Equation (7) can be expanded to Equation (8) as follows:

$$\log (R1/R2) = S \log (r1/r2) + \log B, \quad (8)$$

where *log* refers to natural logarithms with base *e*, parameter *S* to the slope of the line representing sensitivity of reaction, and *log B* to the intercept representing response bias. Equation (8) is a logit equation, predicting the logit of response based on stimulus logit.

Ratio of reinforcement, i.e. *r1/r2* in Equation (8), and that of response allocation, i.e. *R1/R2*, may be explained by reference to an example experiment. Suppose that pigeons or rats in cages are rewarded with food by pushing levers called Levers 1 and 2, for example. The frequency of reward obtained by pushing Lever 1 is represented by *r1*, and that by pushing Lever 2 is represented by *r2*. The ratio of the frequencies of these two reward opportunities is referred as the ratio of reinforcers, i.e. *r1/r2*. The frequencies of lever-pushing behavior by the animals are represented as *R1* and *R2*, with *R1* referring to the frequency of the subjects' pushing Lever 1 and *R2* to that of pushing Lever 2. The ratio of these two frequencies, i.e. *R1/R2*, is referred to as the *ratio of response allocation*. Previous research in animal behavior has shown that response allocation is well expressed by Equation (8).

The generalized matching law seems to exhibit a wide range of applicability. In fact, it is comparable to the ideal free distribution theory in ecological studies, which describes the distribution of wild animal communities across multiple food sites (Fagen, 1987; Yamaguchi & Ito, 2006). The model of the ideal free distribution theory is identical to Equation 8, when the distribution ratio of individuals is replaced with *R1/R2* and the amount of food with *r1/r2*. More generally, it should be noted that the generalized matching law exhibits generalizability with empirical evidence across different fields of study.

Quantitative models, which are comparable to the generalized matching law, are applied in linguistic investigation as well. Well-known examples are sociolinguistic studies by Labov, in which linguistic variation and change are described by a logistic regression model (Wardhaugh, 1986). Labov's (1972) quantitative perspective is prominent among his various contributions to linguistics, in that quantitatively-represented variables quite efficiently account for, and possibly predict, actual language use.

4.2. Correspondence between the Generalized Matching Law and the Logistic Regression Model

Logistic regression analysis conceivably contributes to an investigation of the relationship between mere exposure effect and the generalized matching law. Logistic regression is a method often used in medical statistics, biology, and sociolinguistic studies (Matsuda, 1993), and is expressed as follows:

$$\log\{p1 / (1 - p1)\} = Z, \tag{9.1}$$

where *Z* is a linear function, the element *p1* refers to the probability of choosing Alternative 1, and *1–p1* describes the probability of choosing Alternative 2 in two-alternative forced-choice tasks. The ratio of the difference in the probabilities between the two options, i.e. *p1/(1–p1)* in Equation (9.1), is called *odds*. Equation (9.1) can be expanded to Equation (9.2) as follows:

$$p1 = 1 / \{1 + \exp(- Z)\}, \tag{9.2}$$

which yields probability values between 0 and 1, never allowing values less than 0% or equal to / greater than 100%.

When the response frequencies of Alternatives 1 and 2 are replaced with *R1* and *R2*, the sum of response frequencies *N* is expressed as *R1+R2*, as in *N=R1+R2*. Since *p1* refers to the

probability of choosing Alternative 1, *p1* can be defined as *R1/N* and *p2* as *R2/N*. Thus, the ratio of probabilities is expressed by the odds represented by Equation (10).

$$p1 / (1 - p1) = (R1/N) / (R2/N) = R1 / R2, \tag{10}$$

Yokoyama (2006b) shows that Equation (9.1) becomes identical to Equation (8), i.e. the generalized matching law, when *R1/R2* in Equation (10) substitutes *p1/(1–p1)* in Equation (9.1) and *S log (r1/r2)+log B* in Equation (8) replaces *Z* in Equation (9.1) as follows:

$$\log \{p1 / (1 - p1)\} = \log (R1/R2) \quad \text{and} \quad Z = S \log (r1/r2) + \log B,$$
$$\log (R1/R2) = S \log (r1/r2) + \log B, \tag{8}$$

where Equation (8), i.e. the generalized matching law, is a form of logistic regression model. It may be applicable to a wide range of phenomena across various fields of science, gathering evidence from animal behavior, economic, and ecological studies. However, no previous research seems to have pointed out the comparability between logistic regression models and the generalized matching law.

4.3. The Generalized Matching Law and Mathematical Linguistics in Japanese

Yokoyama (2006b) examines the applicability of the generalized matching law in mathematical linguistics, using a preference judgment task, in which the participants choose which of a pair of kanji variants they prefer. The stimuli were pairs of kanji variants, such as “桧” vs. “檜”. The generalized matching law as in Equation (8) is represented by Equation (8a) as follows:

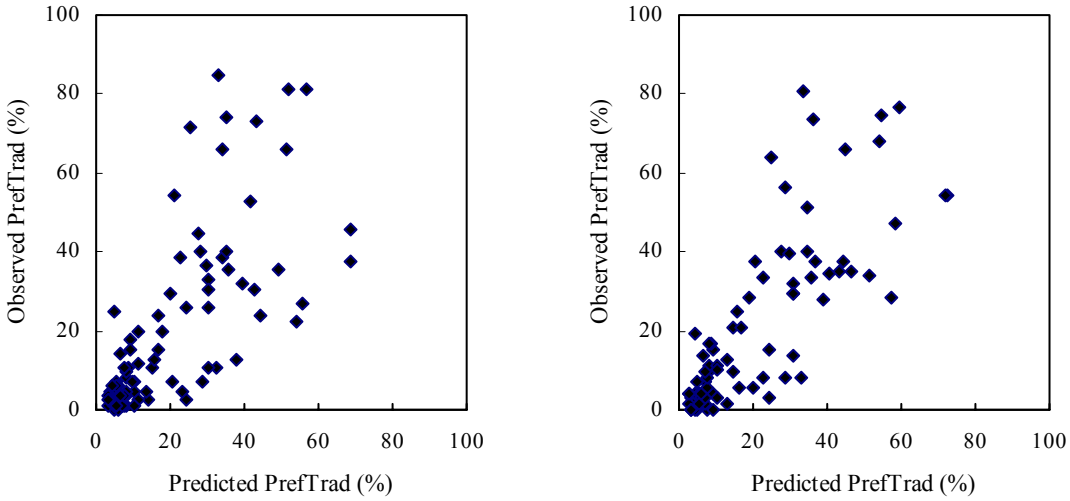
$$\log (R1/R2) = S \log (r1/r2) + \log B = a (FamTrad - FamSimp) + b, \tag{8a}$$

where the numbers of participants who chose the traditional variant and the simplified variant are represented as *R1* and *R2* respectively. Comparability between Equations (8) and (8a) is explained as follows:

$$\begin{aligned} a (FamTrad - FamSimp) + b &= a \{[K \log (FreqTrad) + C] - [K \log (FreqSimp) + C]\} + b \\ &= a K \{\log (FreqTrad) - \log (FreqSimp)\} + b \\ &= a K \log (FreqTrad / FreqSimp) + b \\ &= S \log (r1/r2) + \log B, \end{aligned}$$

where the frequency of the traditional variants, i.e. *FreqTrad*, is defined as *r1*, and frequency of the simplified variants, i.e. *FreqSimp*, as *r2*. The logarithm of the frequency ratio of the characters, i.e. log*(r1/r2)*, is referred to as *exposure relativity*.

Yokoyama (2006b) computes the logarithm of ratio of reinforcer frequencies, i.e. log*(r1/r2)* in Equation (8), based on the frequency data from a newspaper corpus, and estimates the values of the parameter *S* and *log B* by the least square method. It should be noted that the model of Yokoyama (2006b) is an innovative contribution, for studies on mere exposure effect, originally proposed by Zajonc (1968), have not previously presented any specific models. As discussed, the model describes preference behavior in natural language quite reliably, although it still awaits further empirical validation with applicable data.



(a) Tokyo [$r^2 = .517$] (b) Kyoto [$r^2 = .646$]

Figure 2. Predicted and observed preference for traditional variants

Yokoyama (2006c) estimates the parameters in the Tokyo group by applying the method of maximum likelihood estimation to the model as follows ($p <. 01$, $df = 1$):

$$\log (R1/R2) = 0.294 \log (r1/r2) - 0.301, \quad (11)$$

$$p1 = 1 / \{ 1 + \exp[- 0.294 \log (r1/r2) + 0.301] \}, \quad (12)$$

which expresses the probability of choosing the traditional variant of the pair. Although the maximum likelihood estimation is commonly applied in various fields, its application to a study on variant preference behavior in Japanese kanji recognition is a novel exploration.

The results reveal that the model accounted for 51.7% ($r = .719$, $p < .01$, $df = 84$) and 64.6% ($r = .804$, $p < .01$, $df = 84$) of the variance of the observed data in the Tokyo and Kyoto groups respectively. This predictive power is considered significantly strong in studies on natural language. Figure 2 shows the correlation between the predicted probabilities and the observed responses of *PrefTrad*.

An inter-group difference of 15% was observed in the accountability between the Tokyo and Kyoto groups (Yokoyama, 2006c). This difference may be attributable to the types and frequencies of kanji characters to which the participants were exposed in their daily lives in the two different locations. Another factor which may have contributed to the inter-group differences is the genders of the participants. The Tokyo group only consisted of females, while the Kyoto group included twenty males (30% of the group). Although these two factors may be responsible for the inter-group differences, further investigation of this point is necessary.

5. Conclusion

The present paper has reviewed a series of related studies and proposed a model which accounts for performance in the preference judgment task in Japanese kanji recognition. The analysis has indicated the strong predictive power of the model of Yokoyama (2006b) with empirical data from the preference judgment task of Yokoyama (2006a) and Yokoyama

(2006b). For example, the model accounts for 64.6% of the variance of the observed responses in the Kyoto group, based on the frequency data from the Asahi Newspaper corpus. The models discussed in the paper are conceptually straightforward, providing opportunities for immediate applications. It should also be noted that they are quite efficient, in that they provide reliable accountability using only frequency data from newspaper text, though newspapers only constitute a small portion of those activities which involve written language.

The basic principle of the mere exposure effect theory mentioned earlier explains this phenomenon quite well. Studies including but not limited to Zajonc (1968) and Kunst-Wilson & Zajonc (1980) assert that repeated exposure to unfamiliar stimuli increases familiarity, and that, as a consequence, preference for such items increases. Provided that frequency data from newspapers represents the actual use of the given orthography, highly frequent items in newspaper corpora are likely to be high-exposure items in actual written communication. Given that repeated exposure to certain items increases preference, high- frequency characters, i.e. highly exposed characters, are more likely to be preferred than less-exposed characters. In addition, preferred items are likely to be used more frequently, further increasing their exposure. In short, the three variables, i.e. actual language use, frequency of exposure, and elevated preference, are mutually dependent, constantly affecting one another. The results of the present study as well as the empirical evidence from Yokoyama (2006a) and Yokoyama (2006b) support such a theoretical framework, describing the role of mere exposure effect and the applicability of the generalized matching law.

The model with strong predictive power included the familiarity difference between the traditional and simplified variants as the explanatory variable. This is probably because the participants in the study compared their degrees of familiarity with the traditional and simplified variants and chose the more familiar items as they performed the task, for the task *per se* was to compare and select one of the two alternatives at a participant-controlled speed. This account leaves room for replication studies for methodological issues to be examined. Another issue is the validity of newspaper corpora as representative of actual language use in general. Further research on this issue must await corpora that represent the full diversity of orthographic characteristics and types of communication in written language.

Acknowledgement: The authors would like express gratitude to Dr. Katsuo Tamaoka for his efforts and encouragement while exchanging views during the various stages of manuscript preparation. We also appreciate the meaningful feedback provided by anonymous reviewers. The paper is based on studies published in *Mathematical Linguistics* Volume 25, Numbers 4 and 5. The authors acknowledge the contributions of the Mathematical Linguistic Society of Japan, which celebrates its 50th anniversary this year. We are especially grateful for the continuous contributions to mathematical linguistics in Japan of Dr. Sizuo Mizutani, the founder of the Mathematical Linguistic Society of Japan and a former member of the National Institute for Japanese Language, who celebrates his 80th birthday this year.

References

Asahi Shinbun-sha (1994). *CD-HIASK 1993 Asahi Shinbun Database*. Tokyo: Kinokuniya & Nichigai Associates.

Baum, W. M. (1974). On two types of deviation from the matching law: Bias and undermatching. *Journal of the Experimental Analysis of Behavior, 22, 231-242*.

Chikamatsu, N., Yokoyama, S., Nozaki, H., Long, E., & Fukuda, S. (2000). A Japanese logographic character frequency list for cognitive science research. *Behavior*

Research Methods, Instruments, and Computers, No. 32, Vol. 3, *482-500.*

Elliot, R., & Dolan, R. (1998). Neural response during preference and memory judgments for subliminally presented stimuli: A functional neuroimaging study. *The Journal of Neuroscience*, *1998, 18*, *4697-4704.*

Fagen, R. (1987). A generalized habitat matching rule. *Evolutionary Ecology*, *1, 5-10.*

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, *4, 267-272*

Kunst-Wilson, W.R., & Zajonc, R.B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science*, *207*, *557-558*.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Matsuda, K. (1993). Dissecting analogical leveling quantitatively : The case of the innovative potential suffix in Tokyo Japanese. *Language Variation and Change*, 5, 1-34.

Monin, B. (2003). The warm glow heuristic: When liking leads to familiarity. *Journal of Personality and Social Psychology*, *85*, *1035-1048.*

Sasahara, H. & Yokoyama, S. (2000) Familiarity with kanji variants and user preference. *Japanese Linguistics*, *8, 110-125*.

Wardhaugh, R. (1986) *An introduction to sociolinguistics*. Oxford: Blackwell Publishers.

Yamaguchi, T., & Ito, M. (2006). An experimental test of the ideal free distribution in humans: The effects of reinforcer magnitude and group size. *The Japanese Journal of Psychology*, *76, 547-553.*

Yokoyama, S. (2006a). Can we predict preference for kanji form from newspaper data on character frequency? *Mathematical Linguistics*, *25, 181-194.*

Yokoyama, S. (2006b). Mere exposure effect and general matching law for preference of kanji form. *Mathematical Linguistics*, *25, 199-214*.

Yokoyama, S. (2006c, July). Corpus data and prediction of language use by the logistic regression analysis. Paper presented at the research meeting of the National Research for Japanese Institute, Tokyo.

Yokoyama, S., Sasahara, H., Nozaki, H., & Long, E. (1998) *Study on the use of kanji in electronic-media newspapers (Shinbun denshi media no kanji)*. Tokyo: Sanseido.

Zajonc, R.B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9, 1-27.*

Glottometrics 12, 2006, 75-96

History of Quantitative Linguistics

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

XV. Jean Paul (1763-1825)

Jean Paul (vollständiger Name: Jean Paul Friedrich Richter) ist einer der bedeutendsten deutschen Autoren des ausgehenden 18. und beginnenden 19. Jahrhunderts. Er wurde am 21.3.1763 in Wunsiedel (Fichtelgebirge) geboren, besuchte das Gymnasium in Hof, studierte 1781-84 in Leipzig Theologie und Philosophie, musste sein Studium aber aus Armut abbrechen und lebte dann wieder in Hof. 1790-94 arbeitete er als Lehrer an einer von ihm selbst gegründeten Elementarschule in Schwarzenbach, konnte dann aber vom Erfolg seiner Bücher leben. Weitere Lebensabschnitte verbrachte er 1798-1800 in Weimar und 1800 in Berlin, war bis 1803 Legationsrat in Meiningen und kam schließlich über Coburg 1804 nach Bayreuth, wo er am 14.11.1825 verstarb. Er hatte Kontakt zu vielen Zelebritäten seiner Zeit und war Mitglied der Berliner und der Frankfurter *Gesellschaft für deutsche Sprache* (Jean Paul 1820: 76). Neben seinem literarischen Werk befasste er sich u.a. mit linguistischen Themen (Fugen, Wortbildung, etc.). Für die generelle Vorstellung und Würdigung seiner ästhetisch-linguistischen Bemühungen sei auf Faust (1983) verwiesen.

Jean Paul taucht in den Annalen der Quantitativen Linguistik praktisch nicht auf; der einzige mir bekannte Verweis auf ihn findet sich in Best (1997: V), wo es um die Verwendung von Eigennamen geht. Die folgenden kurzen Hinweise sollen auch nicht dafür argumentieren, in Jean Paul einen der frühen Vertreter der Quantitativen Linguistik zu sehen, sondern vielmehr zeigen, dass er auf einige Themen einging, die uns nach wie vor beschäftigen, und sich dabei der Statistik bediente. Damit trägt er zu einem gedanklichen Klima bei, das der Quantitativen Linguistik förderlich sein kann. Vielleicht können diese Hinweise dazu anregen, auch anderweitig nach Autoren oder Strömungen zu suchen, die eine ähnliche Leistung erbrachten oder noch erbringen. Es ist ja doch auffällig zu sehen, dass die Quantitative Linguistik mancherorts gedeiht und anderswo keinerlei Resonanz erfährt.

Themen, die für die Quantitative Linguistik einschlägig sind, werden zuerst in seiner Poetik *Vorschule der Ästhetik* (Jean Paul 1804, ²1813; vgl. Faust 1983) in verschiedenen Paragraphen angeschnitten. Dabei geht es um Wort- und Satzlänge, zwei in der Quantitativen Linguistik (Köhler 1986; 1999) ebenso wie in der Verständlichkeitsforschung zentrale Größen (Mikk 2000; Best 2005), und um den „Reichtum" einer Sprache:

1. *Wortlänge*: Hierzu werden mehrere Aspekte behandelt: a) Jean Paul spricht sich gegen zu lange Wörter aus: „Je länger aber ein Wort, desto *unanschaulicher*; daher geht schon durch die Wurzel-Einsilbigkeit der ‚Lenz' dem ‚Frühling' mit seinen Ableitern vor, ebenso ‚glomm' dem ‚glimmte'" (Jean Paul ²1813: 307), ein Zitat, auf dessen ersten Teil sich Schneider ([13]2004: 41) bei seinem Plädoyer für möglichst verständliches Schreiben der Journalisten beruft. Die schiere Länge erscheint also schon Jean Paul als ein beachtenswertes Merkmal. Die

zitierte Behauptung ist so beschaffen, dass man daraus eine testbare Hypothese gewinnen kann, wenn man ein Kriterium für „Anschaulichkeit“ bestimmt. b) Die Länge von Eigennamen hat ihre eigene Bedeutung, wie Jean Paul (²1813: 270) unter Bezug auf Wieland ausführt: „So hat z.B. der uns bekannte Autor nicht ohne wahren Verstand unbedeutende Menschen einsilbig: Wutz, Stuß getauft, andere schlimme oder scheinbar wichtige mit der Iterativ-Silbe *er*: Lederer, Fraischdörfer...“ Fischer & Roth (1996: 62) folgern daher: „Mit der Gewichtigkeit und psychostrukturellen Komplexität des Personals wächst demzufolge die Anzahl der Silben.“ c) Wie sehr Jean Paul die bloße Länge von Wörtern, speziell von Komposita, beschäftigt, erweist sich später noch einmal in dem selbstgebildeten „Wortbandwurmstockabtreibmittellehrbuchstempelkostenersatzberechnung“ (1820: 67). d) Bei der Untersuchung der Bedingungen für das Vorkommen des Fugen-{s} kommt Jean Paul (²1813: 321) zu der Behauptung: „Je länger das Bestimmwort [= Determinans, Verf.] ist, desto gewisser verzerren wir es noch durch eine Verlängerung mit S.“ Ähnlich heißt es später (Jean Paul 1820: 41): „Je länger das Bestimmwort ist, das mit einem *s* verzischt, und je länger folglich das Ohr darauf warten müssen, desto heißer fodert es sein *s*. Z.B. Wahrheitliebe statt Wahrheitsliebe läßt sich das gedachte Glied noch gefallen, aber Wahrhaftigkeitliebe, wo es um zwei Sylben länger auf den Schlangen-Mitlauter vergeblich gepaßt, oder gar Wissenschaftlichkeitliebe will ihm durchaus nicht ein.“ Dass diese Behauptung sich als Hypothese für entsprechende Untersuchungen geradezu aufdrängt, dürfte klar sein.

2. *Satzlänge*: Zu diesem Kriterium stellt Jean Paul ein Prinzip auf: „Sprachkürze muß dem Leser nicht längere Zeit kosten, sondern ersparen“ (²1813: 318) und erläutert dies (²1813: 319): „Zur Achtung gegen den Leser gehört ferner weit mehr *ein* langer Periode als zwanzig kurze. Die letzten muß er zuletzt doch selber zu *einem* umschaffen, durch Wiederlesen und Wiederholen.“ Jean Paul erkannte offenbar bereits, dass sprachliche Ökonomie viele Facetten hat (Moser 1971; 1980), wobei u.a. gilt, dass das, was für den Sprecher/ Schreiber weniger Aufwand bedeutet, den des Hörers/ Lesers erhöht (Köhler 1986: 20ff.).

3. *Reichtum der Sprache*: Hierzu heißt es (²1813: 306): „Wenn man den Reichtum unserer Sprache ... am vollständigsten ausgelegt sehen will: so überzähle man den deutschen Schatz an sinnlichen Wurzel-Zeitwörtern.“ Als Beleg dafür gibt Jean Paul an, wie viele verschiedene Verben in einem von ihm selbst angelegten Wortregister in unterschiedlichen Verbklassen vorkommen. Hier wird also eine Eigenschaft des Deutschen, die Differenziertheit des Verbwortschatzes, mit Hilfe einer Statistik charakterisiert und als Maßstab für sprachlichen „Reichtum“ gewertet.

4. *Wortbildung*: Sprachstatistisch fundierte Argumentation findet sich auch bei der Behandlung eines speziellen Problems der deutschen Wortbildung, und zwar dann, wenn er sich mit der von ihm abgelehnten -{s}-Fuge beschäftigt (Jean Paul 1820): „it became clear that the linking -*s*- appeared less frequently than any other linking elements... This provided Jean Paul with a quantitative argument for the normative proposal“ (Faust 1983: 240). Jean Paul verweist im Zusammenhang mit der Wortbildung auf seine „Wörtervolkzählungen“ (1820: 39), die er durchführt, um dem Sprachgebrauch hinsichtlich der -{s}-Fuge auf die Spur zu kommen. Mit „Wörtervolk“ sind Deklinationsklassen gemeint.

5. *Buchstabenhäufigkeit*: In (Jean Paul 1820: 37) beklagt er sich über das „deutsche Schwa“ und führt als Beleg an: „Kaufen Sie von einem Schriftgießer vier Zentner klein Cicero, so bekommen sie nur 4900 Fraktur-a, dagegen aber 11000 Fraktur-e.“

Nimmt man alles zusammen, so kann man feststellen, dass Jean Paul immer wieder statistische Erhebungen zu sprachlichen Phänomenen durchführt oder wenigstens fordert, um seine Argumentation zu unterstützen. In Einzelfällen kommt er zu Formulierungen, die Zusammenhänge zwischen verschiedenen sprachlichen Eigenschaften behaupten. In beiden Aspekten wirkt er durchaus modern, auch wenn man möglicherweise seinen Ideen, etwa zur Wortbildung, nicht unbedingt folgen mag.

Literatur

Best, Karl-Heinz (1997). Warum nur: Wortlänge? Nicht nur ein Vorwort. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: V-XII.* Trier: Wissenschaftlicher Verlag Trier.

Best, Karl-Heinz (2005). Sind Wort- und Satzlänge brauchbare Kriterien der Lesbarkeit von Texten? In: Wichter, Sigurd, & Busch, Albert (Hrsg.), *Wissenstransfer - Erfolgskontrolle und Rückmeldungen aus der Praxis.* Frankfurt/ M. u.a.: Lang (erscheint).

Faust, Manfred (1983). Jean Paul's essay on word formation. In: Faust, Manfred, Harweg, Roland, Lehfeldt, Werner, & Wienold, Götz (Hrsg.); *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik. Festschrift für Peter Hartmann: 237-248*. Tübingen: Narr.

Fischer, Susanne, & Roth, Jürgen (1996). Faust. Wuz. Flok. *Spiegel Special Nr. 10: 62-65.*

Jean Paul (1804, ²1813). Vorschule der Ästhetik. In: Jean Paul, *Sämtliche Werke. Abt. I, Bd. 5: 7-456*. Hrsg. v. Norbert Miller. Frankfurt: Zweitausendeins 1996 (Nachdruck der Ausgabe des Hanser-Verlags 1963).

Jean Paul (1820). Über die deutschen Doppelwörter; eine grammatische Untersuchung in zwölf alten Briefen und zwölf neuen Postskripten. In: Jean Paul, *Sämtliche Werke. Abt. II, Bd. 3: 9-108.* Hrsg. v. Norbert Miller. Frankfurt: Zweitausendeins 1996 (Nachdruck der Ausgabe des Hanser-Verlags 1963).

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

Köhler, Reinhard (1999). Syntactic Structures: Properties and Interrelations. *Journal of Quantitative Linguistics 6*, *46-57.*

Mikk, Jaan (2000). *Textbook: Research and Writing*. Frankfurt: P. Lang.

Moser, Hugo (1971). Typen sprachlicher Ökonomie im heutigen Deutsch. In: *Sprache und Gesellschaft. Jahrbuch 1970: 89-117.* Düsseldorf: Schwann.

Moser, Hugo (1980). Zum Problem der sprachlichen Ökonomie. *Zeitschrift für deutsche Philologie 99, 98-100.*

Schneider, Wolfgang ([13]2004). *Deutsch fürs Leben. Was die Schule zu lehren vergaß.* Reinbek: Rowohlt.

Karl-Heinz Best, Göttingen

XVI. Ernst Wilhelm Förstemann (1822-1906)

0. Biographisches

Förstemann wurde am 18.9.1822 in Danzig geboren, studierte nach dem Schulabschluss Vergleichende Sprachwissenschaft in Berlin und Halle (u.a. bei A.F. Pott); Promotion 1844 in Halle; danach Arbeit als Lehrer in verschiedenen Stellungen. Ab 1851 Bibliothekar und Lehrer in Wernigerode, ab 1865 Bibliothekar in verschiedenen Stellungen in Dresden. 1899 Ruhestand. 1900 Übersiedlung nach Berlin, dann Charlottenburg. Er verstarb am 4.11.1906.

I. Förstemanns Plädoyer für Sprachstatistik

Förstemann (1853a: 339) erklärt zur „numerischen methode“, er habe bereits „in frueher jugend“ eine „neigung fuer diese richtung“ gefasst.

Programmatisch sind schon die ersten Sätze in Förstemann (1846: 83):
„Wer es gesehn hat, wie die neuere Statistik aus der Betrachtung bloßer Zahlenangaben die überraschendsten Resultate für das Leben und die Fortbildung der Völker erlangt, wird nicht darüber spotten, wenn auch in der Sprachwissenschaft der Versuch gemacht wird, durch Zählung der Sprachindividuen, der Buchstaben, zu einigen Resultaten zu gelangen oder wenigstens schon gefundene Resultate von einer neuen Seite her zu bestätigen. Man weiß z.B., daß gewisse Laute in den Sprachen allmälig entweder häufiger oder seltener werden; sollte es nun nicht von Interesse sein, dieses Steigen oder Sinken mit mathematischer Genauigkeit zu messen und dadurch solche Erscheinungen gegen verwandte ins rechte Licht zu setzen? Sollte man nicht ferner dazu kommen können, von den Veränderungen der Sprachen zwischen je zwei gegebenen Zeitpunkten und den inzwischen verflossenen Zeiträumen Proportionen zu bilden und aus diesen für die größere oder geringere Vitalität einer Sprache zu einer gewissen Zeit ein annähernd sicheres Urteil zu erhalten? So lange man sich wenigstens vor dem Ueberschreiten der vernunftgemäßen Grenze hütet, dürfte diese Methode nicht unergiebig sein.“[1]

In der Fußnote zu dieser Seite heißt es: „Es ist merkwürdig, daß man bisher nie in den Grammatiken an eine solche Lautstatistik gedacht hat, während Buchdrucker und Schriftgießer doch von der Nothwendigkeit eines Theils derselben von jeher überzeugt sind.“

Mit diesen Zitaten soll auf einige Aspekte hingewiesen werden, die für Förstemann und teils auch für die Linguistik seiner Zeit wichtig sind:

- Es sind Einflüsse von außen, die wieder einmal einen Anstoß für Neuerungen in der Linguistik geben. Eines dieser Vorbilder findet Förstemann (1852a: 166) in der Geographie, wo die Proportion von Küstenlänge und Flächeninhalt der Länder thematisiert wird.
- Förstemann verspricht sich durch Verwendung der Statistik teils neue Ergebnisse, teils Bestätigung bereits vorhandenen Wissens durch eine zusätzliche Methode, mit der „mathematische[...] Genauigkeit“ erreicht werden kann.
- Er spricht davon, „Proportionen“ zwischen unterschiedlichen Sprachentwicklungsstadien bilden zu können.
- Charakteristisch für Förstemann ist auch der Hinweis, man müsse sich „vor dem Ueberschreiten der vernunftgemäßen Grenze hüte(n).“ Ähnlich z.B. in Förstemann (1853c: 44).

Die letzten drei Aspekte sind in vielen Arbeiten Förstemanns zu finden, auch die Warnung, keine vorschnellen Schlüsse zu ziehen.
Im Folgenden sollen einige Themen aus Förstemanns Arbeiten vorgestellt werden, soweit sie in einem weiten Sinne als Beiträge zur Quantitativen Linguistik verstanden werden können.

II. Laute und Lautgruppen

In Förstemann (1846) geht es um zwei Aspekte:

1. Es werden „Proportionen“ zwischen den Häufigkeiten von Lauten oder Lautklassen innerhalb des Gotischen, Althochdeutschen, Mittelhochdeutschen und Neuhochdeutschen in Form von Mittelwerten (arithmetisches Mittel) aufgrund mehrerer Zählungen vorgestellt. Man findet u.a. Angaben zum Verhältnis von Konsonanten und Vokalen, zum relativen Anteil

[1] Zitate folgen so gut wie möglich dem Original; bei der Wiedergabe der Umlaute und der Schreibung von <ss> wird mangels entsprechender Zeichen eine modernisierte Form verwendet.

der Grundvokale und der Konsonanten, zu den Quotienten zwischen hellen und dunklen Vokalen, zum Anteil verschiedener Konsonantengruppen sowie zur Verteilung von Konsonanten und Vokalen auf Anlaut, Mitte und Auslaut.

2. Die Veränderungen zwischen den genannten Sprachen werden schon dadurch deutlich, dass deren Lautrelationen in der richtigen zeitlichen Anordnung untereinander aufgeführt sind. Förstemann stellt Berechnungen dazu an, bei welchen Übergängen mehr Veränderungen pro Zeiteinheit stattfinden und benutzt diese Befunde, um die Vitalität der Sprache in den einzelnen Zeitabschnitten zu beurteilen.

Panconcelli-Calzia (1941: 47) kommentiert die Untersuchung von 1846: „Förstemann veröffentlicht die erste vollständige Statistik über die Laute im Gotischen, sowie im Alt-, Mittel- und Hochdeutschen. Es ist die erste Arbeit dieser Art, die verdient, als Statistik bezeichnet zu werden."

Diese Arbeit wird in Förstemann (1852a) fortgesetzt, indem das Deutsche (bzw. das Gotische als ein früher Vertreter der germanischen Sprachen) unter ähnlichen Gesichtspunkten mit dem Griechischen und Lateinischen verglichen wird. Auch hier geht es wieder um die Proportion der Häufigkeit von Konsonanten und Vokalen, um die Proportionen verschiedener Laute und Lautklassen innerhalb der Sprachen und dann auch zwischen ihnen. Er charakterisiert noch einmal seinen Versuch von 1846: Dort habe er versucht „darzuthun, dass durch statistische angaben ueber das vorkommen der einzelnen laute sich resultate ueber die entwickelung der sprachen und ueber das verhältnis der einzelnen idiome zu einander erzielen lassen" und plädiert dafür, „dass dieser weg der erkenntnis des sprachgeistes und sprachlebens naeher zu kommen, ein erlaubter und förderlicher sei. Denn für manches auf andern wegen erkannte finden wir hier schärfe und genauigkeit, irrthuemer werden hier leicht und schlagend berichtigt, und, täuscht mich nicht alles, so lässt sich sogar von diesem wege aus mehrfach bahn brechen in dunkle und sonst unzugängliche parthien der wissenschaft. Darf man sonst neue bahnen nur mit einer gewissen schüchternheit und in der furcht betreten, festen boden zu verlieren, so giebt uns dagegen hier das mathematische element, als die sicherste sphaere des menschlichen erkennens, vielfach die bürgschaft, dass wir uns aus dem sicher erkannten nicht zu weit in das luftige reich unhaltbarer hypothesen verlieren werden" (Förstemann 1852a: 164).

Während diese Passage einige Motive der Arbeit von 1846 wiederholt und verstärkt, kommt an späterer Stelle eine neue Idee zur Sprache. Förstemann vergleicht dort die Häufigkeiten, mit denen einzelne Laute in den drei behandelten Sprachen verwendet werden und berechnet daraus Distanzen zwischen ihnen. Dabei kommt heraus, das Griechisch und Gotisch sich stärker voneinander unterscheiden als die beiden anderen Paarungen. Bemerkenswert ist hier der Gedanke, dass man den Abstand zwischen Sprachen berechnen kann. Diesen Gedanken spinnt Förstemann (1852a: 175) weiter aus:

„Bei aufstellung dieser zahlen muss ich mich ausdrücklich gegen den vorwurf verwahren, als masste ich mir an, mit ihnen im allgemeinen den abstand der sprachen von einander auszudrücken. Dazu würden noch andere elemente berücksichtigt werden müssen, wie der abstand in der flexion, der abstand des genus, der abstand des sprachschatzes u.s.w., elemente, bei denen ich die anwendung der mathematischen methode gleichfalls nicht fuer unmoeglich halte. Genau genommen erschöpfe ich durch die mitgetheilten zahlen nicht einmal den l a u t l i c h e n unterschied der sprachen, denn dazu müsste ich auch in anschlag bringen, wie (nach euphonischen gesetzen) die laute in jeder der drei sprachen vereint werden."

Förstemann nutzt diese Methoden, um der historischen Erforschung der Sprachen zu dienen; seine Untersuchungen sollen die statistische Absicherung der historischen Klassifikation der Sprachen fördern. Er stellt aber nicht nur Vergleiche zwischen Vorgänger- und Nachfolgersprachen an, sondern charakterisiert auch die griechischen Dialekte je für sich und im Vergleich untereinander statistisch (Förstemann 1853d). Knauer (1955: 143) würdigt diese

Überlegungen: „Und es dauerte noch bis in die Mitte des 19. Jahrhunderts, bis E. FÖRSTEMANN in mehreren Aufsätzen als erster nicht nur zählte, sondern darauf aufbauend Möglichkeiten der Sprachcharakterisierung durch Feststellung von Lautmengen-Proportionen darlegte."

Ein weiterer Gedanke wirkt recht modern: So überlegt Förstemann (1852a: 176f.) zunächst auf rein theoretischer, mathematischer Grundlage, wie groß die Distanzen zwischen Sprachen minimal und wie groß sie maximal sein könnten, und schließt dann eine Überlegung dazu an, wie groß diese Differenz bei den Sprachen der Welt tatsächlich sein könnte. Hätte man diese Grenze, dann könnte man alle erforschten Sprachen auf einer Distanzskala zwischen dem minimalen und dem maximalen Distanzgrad anordnen. Er vermutet, dass diese Skala das Verhältnis der Sprachen untereinander als Dialekte, verwandte oder nicht verwandte Sprachen explizieren könnte.

In Förstemann (1853c) bezieht er das Sanskrit in seine Lautuntersuchungen ein. Die Untersuchungsaspekte stimmen mit den bereits erwähnten überein: Es geht wieder um Proportionen von Lauten und Lautklassen innerhalb der betrachteten Sprachen und zwischen ihnen. Diese münden in vorsichtigen Verallgemeinerungen, etwa wenn er feststellt, dass bei den vier behandelten Sprachen da, wo der Vokalismus sich stärker verändert, dies auch für den Konsonantismus gilt (Förstemann 1853c: 42f.).

In Förstemann (1853d) geht es um die Lautproportionen griechischer Dialekte. Am bemerkenswertesten ist aus der Sicht der Quantitativen Linguistik die Anfangspassage, in der Förstemann Korrekturen an seinem statistischen Vorgehen darstellt, das bisher auf einer inzwischen als zu wenig repräsentativ erkannten Datenbasis beruhte. Entsprechend erhöht er nun die Textbasis für seine Zählungen erheblich, um auch für die selteneren Laute ein hinreichend sicheres Ergebnis zu gewinnen.

Später findet man eine Statistik zu Lautverschiebungen im Konsonantismus, getrennt nach Anlaut und Inlaut, die er wie folgt kommentiert: „Man ersieht aus diesen statistischen Angaben die Stärke der Erscheinung im Allgemeinen so wie die verhältnismässige Stärke der einzelnen Richtungen, in die sie auseinander geht. Und zur Schätzung der r e l a t i v e n Stärke sind sie völlig brauchbar, wenn auch die Zahlen a b s o l u t keineswegs feststehn (Förstemann 1874: 366).

III. Lexikalische Untersuchungen

Ganz analog zu den Lautuntersuchungen behandelt Förstemann (1852b, 1854) Bezeichnungen für Tiere im Deutschen, Griechischen, Lateinischen und Sanskrit daraufhin, inwieweit diese Sprachen einen gemeinsamen Wortschatz aufweisen, und zwar paarweise ebenso wie insgesamt, denn „...dann ist es zeit, aus den numerischen angaben ueber die zahl der verwandten wörter folgerungen ueber den gegenseitigen l e x i c a l i s c h e n abstand der sprachen zu machen, so wie sie jetzt schon ueber ihren l a u t l i c h e n abstand gemacht werden können. Nur darf man nie erwarten, daß beide arten der sprachdistancen unter einander uebereinstimmen, denn der leblose laut folgt zum theil ganz anderen einflüssen als das beseelte wort" (Förstemann 1854: 62).

In Förstemann (1874: 100, 280, 452) stellt der Autor dar, wie sich der Erbwortschatz vom Indogermanischen bis zu den Anfängen des Deutschen entwickelt, in dem er aufschlüsselt, wie viele Wörter aus welcher der aufeinanderfolgenden Sprachperioden stammen. Es handelt sich um einen Gesamtwortschatz von 2417 (Förstemann verrechnet sich und gibt 2413 an.), die auf die Wortarten aufgeschlüsselt werden.

IV. Namen

Ein weiterer thematischer Schwerpunkt Förstemanns ist die Namenforschung. Die Quantitative Linguistik hat in diesem Zusammenhang besonders zwei Aufsätze zu entdecken (Förstemann 1852a,d), in denen er Daten und auch Schätzungen zum ererbten Namensbestand vorstellt. In einigen Fällen sind diese Daten geeignet, an ihnen Gesetzeshypothesen zu testen, die die Quantitative Linguistik erst in den letzten Jahrzehnten entwickelt hat. Dazu sollen zwei Beispiele gegeben werden. Als erstes folgt ein Beispiel aus dem sog. *Verbrüderungsbuch von St. Peter zu Salzburg*, in dem 32 Schreiber hinreichend genau datiert werden können. Förstemann (1852a: 338f.) hat nun Daten zusammengestellt, die zeigen, wie <ai> allmählich in <ei> übergeht. (Förstemann unterscheidet nicht strikt zwischen Buchstaben und Lauten.) Aus diesen Angaben lässt sich die folgende Tabelle 1 erstellen, wobei das logistische Gesetz in der Form

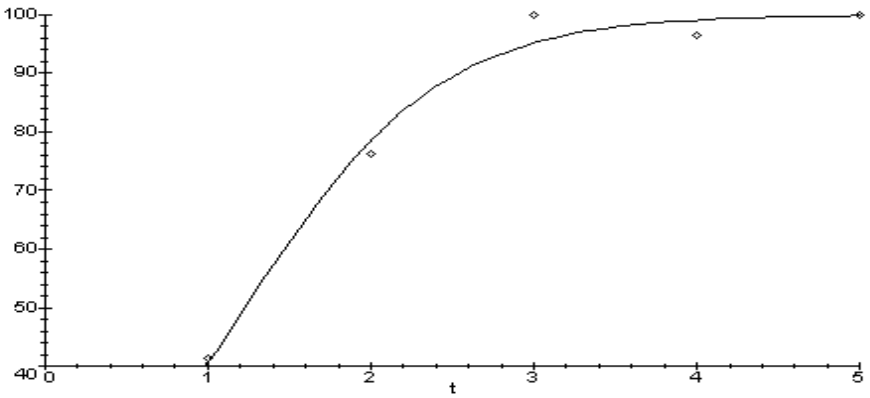
$$p = \frac{1}{1 + a e^{-bt}}$$

(Altmann 1983: 60) angewendet wird:

Tabelle 1
Der Übergang von <ai> zu <ei> bei Namen

t	Zeit bis zum Jahr	Anteil <ei> beobachtet	Anteil <ei> berechnet
1	800	41.41	40.67
2	900	76.27	78.54
3	1000	100.00	95.13
4	1100	96.55	99.05
5	1200	100.00	99.82
a = 7.7894		*b* = 1.6750	*D* = 0.9859

Legende:
a und *b* sind die Parameter des Modells; Der Determinationskoeffizient *D* = 0.9859 zeigt eine sehr gute Übereinstimmung zwischen dem Modell und den Daten an, wie auch die folgende Graphik bestätigt:



Graphik 1. Der Übergang von <ei> zu <ai> bei Namen

Ein zweiter Aspekt, der sich aus heutiger Perspektive aufgreifen lässt, ist die Diversifikation der „etwas ueber 6000 Personennamen“ nach Wortbildungsstrukturen (Förstemann 1852d: 102, 103). Orientiert man sich dazu an Altmann (1991), kann man als Modell für solche Fälle die erweiterte positive negative Binomialverteilung

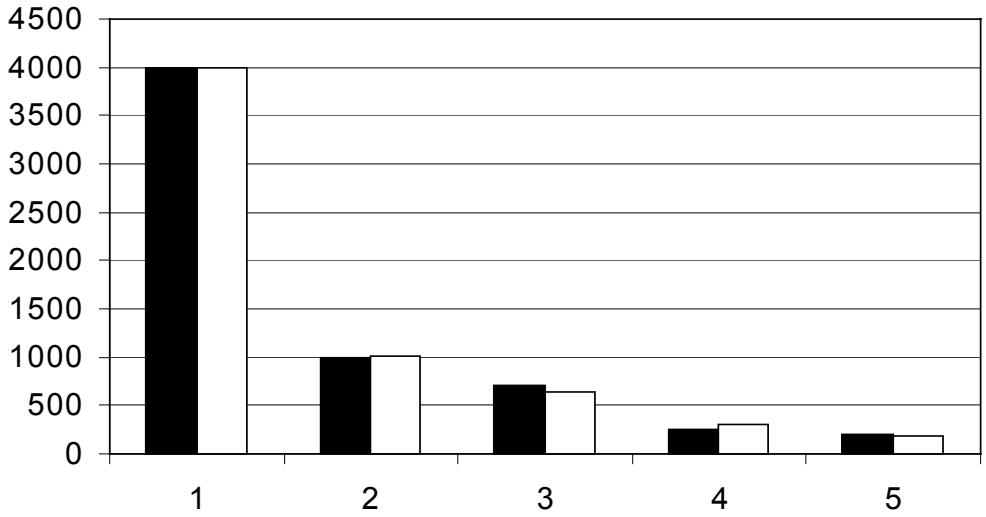
$$P_x = \frac{\alpha \binom{k+x-1}{x} p^k q^x}{1-p^k}, \quad x=1,2,3,...$$

verwenden:

Tabelle 2
Diversifikation der Wortstruktur von Namen

Rang	Wortstruktur	Namen (beobachtet)	Namen (berechnet)
1	Stamm, Stamm	4000	4000.00
2	Stamm, Endung	1000	1016.50
3	Stamm	700	640.38
4	Stamm, Endung, Endung	250	306.03
5	Stamm, Endung, Stamm	200	187.10
k = 6.2540 p = 0.8263 α = 0.3496 FG = 1 X^2 = 16.966 C = 0.0028			

Legende:
k, p, α sind die Parameter des Modells; FG sind die Freiheitsgrade; X^2 ist das Chiquadrat; C ist der Diskrepanzkoeffizient, der hier mit C = 0.0028 eine sehr gute Übereinstimmung zwischen Beobachtung und Modell anzeigt, wie auch die Graphik bestätigt:



Graphik 2. Diversifikation der Wortstruktur von Namen

V. Grammatische Themen

Auch grammatische Verhältnisse werden von Förstemann statistisch erhoben, so die Anzahl starker Verben in den von ihm angenommenen Verbklassen und ihre Einteilung nach Stammauslauten (Förstemann (1874: 577) sowie die Bedeutung der Genitive auf *-an* und *-on* (Förstemann 1867).

VI. Weitere Themen Förstemanns

Nur der Vollständigkeit halber sei darauf verwiesen, dass auch die Entzifferung der Maya-Handschriften und die Mundartforschung (vgl. Cherubim 2003: 502) sowie die Volksetymologie (Förstemann 1852c, 1877) zu Förstemanns Arbeitsfeldern zählen. Paalzow (1906) würdigt seine Verdienste als Bibliothekar, geht aber auch auf einige seiner sprachwissenschaftlichen Themen ein.

VII. Nachwirkung

Leo Meyer (1869), den Förstemann (1875: 78) als „meinen Freund" bezeichnet, nennt – ohne sich an dieser Stelle ausdrücklich auf Förstemann zu beziehen – Zahlenangaben zu Lauthäufigkeiten in einigen indogermanischen Sprachen. Wie Pott (1884: 24) verweisen Zwirner & Zwirner (1935/ 1969: 56) auf Förstemann, der wohl als erster Häufigkeitszählungen von Buchstaben durchgeführt habe. Noch beeindruckender gerät die bereits zitierte Würdigung durch Panconcelli-Calzia (1941). Mehrfach geht Herdan (1966) auf Förstemanns lautstatistische Verdienste ein. Meier (1967: 7, 349, 379) würdigt Förstemann (mit falschem Vornamen!) als den ersten, der Zählungen mit sprachwissenschaftlicher Zielsetzung durchgeführt habe. Dazu ist anzumerken: sprachstatistische Zählungen finden sich schon in Jean Pauls *Vorschule der Ästhetik* (Jean Paul 1804, ²1813; Best 2005). Knauer (1955: 143) macht aber klar, dass Förstemann nicht nur Zählungen durchführt, sondern diese als Mittel zu weiterreichenden Zwecken, eben der „Sprachcharakterisierung", einsetzt. Altmann & Lehfeldt (1980: 115) verweisen darauf, dass er auch einer Universalie auf der Spur war, wenn er meinte, „daß das Vorherrschen der Zungenlaute eine gemeinsame Eigenschaft aller menschlichen Sprachen sei." Es ist aber auffällig, dass Förstemann trotz dieser Verdienste in etlichen Darstellungen der Geschichte der Sprachwissenschaft keine Erwähnung findet. In Köhlers Bibliographie (1995) ist er immerhin mit drei Arbeiten vertreten, allerdings nicht mit seiner bahnbrechenden Untersuchung von 1846. Hier ist offensichtlich eine Lücke zu schließen.

Auf den indirekten Einfluss Förstemanns - über August Schleicher - auf die russische Linguistik weisen Grzybek & Kelih (2003: 136; 2004: 95) hin.

Förstemanns Idee einer Skala, auf der man alle Sprachen aufgrund ihrer Distanzen einordnen könnte, lässt sich als erstes Konzept zu einer quantitativen Sprachtypologie interpretieren, in der später euklidische Distanzen dazu genutzt werden, um eine Taxonomie von Sprachen zu erarbeiten (Altmann & Lehfeldt 1973).

Während Förstemanns lautstatistische Untersuchungen immerhin eine gewisse Resonanz in der Quantitativen Linguistik gefunden haben, sind seine anderen quantitativen Ansätze offenbar bisher ihrer Aufmerksamkeit entgangen. Dass u.a. auch im Bereich der Namenforschung Entdeckungen zu machen sind, wurde oben bereits an zwei Beispielen demonstriert.

VIII. Abschließende Bemerkung

Nimmt man alles bisher Gesagte zusammen, kann man Ernst Wilhelm Förstemann wohl als einen der ersten Quantitativen Linguisten überhaupt ansehen. Zwar haben andere schon vor ihm einzelne Themen der Quantitativen Linguistik behandelt (vgl. die einschlägigen Beiträge in *Glottometrics 6/ 2003ff.)*; es scheint aber niemanden zu geben, der gleichzeitig mit Förstemann oder gar vor ihm die Statistik in thematisch derart vielfältiger Weise immer wieder eingesetzt hat, um Zustände oder Veränderungen der Sprache darzustellen.

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 (1991). Modelling diversification phenomena in language. In: Rothe, Ursula (Hrsg.), *Diversification Processes in Language: Grammar* (S. 33-46). Hagen: Margit Rottmann Medienverlag.

Altmann, Gabriel, & Lehfeldt, Werner (1973). *Allgemeine Sprachtypologie.* München: Fink.

Altmann, Gabriel, & Lehfeldt, Werner (1980). *Einführung in die Quantitative Phonologie.* Bochum: Brockmeier.

Best, Karl-Heinz (2005). Jean Paul (1763-1825). (in diesem Band).

Cherubim, Dieter (2003). Förstemann, Ernst Wilhelm. In: *Internationales Germanistenlexikon 1800-1950.* Herausgegeben und eingeleitet von Christoph König. (S. 502-503). Berlin/ New York: de Gruyter.

Ernst Wilhelm Förstemann. In: Förstemann, Ernst (³1913). *Altdeutsches Namenbuch. Zweiter Band. Orts- und sonstige geographische Namen. Erste Hälfte A-K.* 3., völlig neu bearbeitete, um 100 jahre (1100-1200) erweiterte auflage, hrsg. v. Hermann Jellinghaus. (S. III-XXVIII). Bonn: Peter Hanstein Verlagsbuchhandlung.

Förstemann, Ernst (1843). Noch etwas über Idisi. *Germania 5, 219-221.*

Förstemann, Ernst (1843). Zur Bedeutungslehre der deutschen Adverbia. *Germania 6, 44-51.*

Förstemann, Ernst (1846). Ueber die numerischen Lautverhältnisse im Deutschen. *Germania 7, 83-90.*

Förstemann, Ernst (1850). Ueber ein künftiges Wörterbuch altdeutscher Eigennamen. *Germania 9, 36-62.*

Förstemann, Ernst (1852a). Numerische lautverhältnisse im Griechischen, Lateinischen und Deutschen. *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen [= Kuhns Zeitschrift] 1, 163-179.*

Förstemann, Ernst (1852b). Sprachlich-naturhistorisches. *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen [= Kuhns Zeitschrift] 1, 491-506.*

Förstemann, Ernst (1852c). Ueber deutsche Volksetymologie. *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen [= Kuhns Zeitschrift] 1, 1-25.*

Förstemann, Ernst (1852d). Die zusammensetzung altdeutscher personennamen. *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen [= Kuhns Zeitschrift] 1, 97-116.*

Förstemann, Ernst (1853a). Die diphthonge im verbruederungsbuch von St. Peter zu Salzburg. *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen [= Kuhns Zeitschrift] 2, 337-350*.

Förstemann, Ernst (1853b). Nicht vorhandene Eigennamen. *Germania 10, 26-36.*

Förstemann, Ernst (1853c). Numerische lautbeziehungen des griech., latein. und deutschen zum sanskrit. *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen [= Kuhns Zeitschrift] 2, 35-44.*

Förstemann, Ernst (1853d). Numerische lautverhältnisse in griechischen dialecten. *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen [= Kuhns Zeitschrift] 2, 401-414.*

Förstemann, Ernst (1853e). Unorganisch anlautendes H in altdeutschen Personennamen.. *Germania 10, 37-55.*

Förstemann, Ernst (1854). Sprachlich-naturhistorisches. *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen [= Kuhns Zeitschrift] 3, 43-62.*

Förstemann, Ernst (1867). Zur geschichte altdeutscher declination. IV. Der genitivus singularis. *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen [= Kuhns Zeitschrift] 16, 321-343.*

Förstemann, Ernst (1869, 1870, 1871). Der urdeutsche Sprachschatz. *Germania 14 (= NF 2), 337-372; Germania 15 (= NF 3), 385-410; Germania 16 (=NF 4), 414-438.*

Förstemann, Ernst (1869, 1870, 1871). Straßennamen von Gewerben. *Germania 14 (= NF 2), 1-26; Germania 15 (= NF 3), 261-284; Germania 16 (=NF 4), 265-286.*

Förstemann, Ernst (1872). Assimilation im deutschen. *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen [= Kuhns Zeitschrift] 20, 401-430.*

Förstemann, Ernst (1874/75). *Geschichte des deutschen Sprachstammes. 1. u. 2. Band.* Nordhausen: Verlag von Ferd. Förstemann.

Förstemann, Ernst (1877). Ueber deutsche Volksetymologie. *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen [= Kuhns Zeitschrift] 23, 375-384.*

Förstemann, Ernst (1883). Thumelicus. *Germania 28, 188-190.*

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

Grzybek, Peter, & Kelih, Emmerich (2004). Anton Semënovič Budilovič (1848-1908) - A Forerunner of Quantitative Linguistics in Russia? *Glottometrics 7, 94-96.*

Herdan, Gustav (1966). *The Advanced Theory of Language as Choice and Chance.* Berlin/ Heidelberg/ New York: Springer.

Knauer, Karl (1955). Grundfragen einer mathematischen Stilistik. *Forschungen und Fortschritte 29, 140-149.*

Köhler, Reinhard (1995). *Bibliography of Quantitative Linguistics*. With the Assistance of Christiane Hoffmann. Amsterdam: J. Benjamins.

Meier, Helmut (1967). *Deutsche Sprachstatistik.* Zweite erweiterte und verbesserte Aufl. Hildesheim: Olms.

Meyer, Leo (1869). *Die gothische Sprache. Ihre Lautgestaltung insbesondere im Verhältniss zum Altindischen, Griechischen und Lateinischen.* Berlin: Weidmannsche Buchhandlung.

Paalzow, Hans (1906). Ernst Förstemann. *Zentralblatt für das Bibliothekswesen 23, 552-563.*

Panconcelli-Calzia, Giulio (1941). *Geschichtszahlen der Phonetik. 3000 Jahre Phonetik.* Hamburg: Hansischer Gildenverlag. Reprint in: Panconcelli-Calzia, Geschichtszahlen der Phonetik. Quellenatlas der Phonetik. New edition with an English Introduction by Konrad Koerner. Amsterdam/ Philadelphia: John Benjamins.

Pott, August Friedrich (1884). Einleitung in die allgemeine Sprachwissenschaft. *Internationale Zeitschrift für allgemeine Sprachwissenschaft 1 (= Techmers Zeitschrift), 1-68.* Neudruck in: AUGUST FRIEDRICH POTT, EINLEITUNG IN DIE ALLGEMEINE SPRACHWISSENSCHAFT preceded by the same author's ZUR LITERATUR DER SPRACHENKUNDE EUROPAS. Newly edited together with a bio-bibliographical sketch of Pott by Paul Horn by E.F.K. KOERNER. With a preface and a new index of names: 201-268. Amsterdam: John Benjamins 1974.

Zwirner, Eberhard, & Ezawa, Kennosuke (Hrsg.) (1969). *Phonometrie. Dritter Teil: Spezielle Anwendungen.* Basel/ New York: Karger.

Zwirner, Eberhard, & Zwirner, Kurt (1935). Lauthäufigkeit und Zufallsgesetz. *Forschungen und Fortschritte 11, Nr. 4*: 43-45. (Auch in: Zwirner & Ezawa (Hrsg.), Dritter Teil: 55-59.)

Karl-Heinz Best, Göttingen

XVII. Karl Knauer (1906-1966)

Karl Knauer (vollständig: Karl August Friedrich Knauer) wurde am 16.8.1906 in Hamburg geboren, besuchte die Oberrealschule in Augsburg und studierte ab SS 1925 neuere Sprachen in München. Promotion Juni 1929, 1929-1932 Lektor für deutsche Sprache in Lille, Frankreich. Am 15.3.35 Habilitation in Münster mit der Venia legendi für romanische Philologie und der Ernennung zum Dozenten. Ab 24.4.42 apl. Professor. Knauer wurde in seinem Entnazifizierungsverfahren im Mai 1948 als nicht belastet eingestuft. 1960 Wissenschaftlicher Rat und Prof. für Romanische Philologie, Univ. Münster. Er verstarb am 22.5.1966 in Münster (Hausmann 2000; Lausberg 1980; Schuder 1966; Untiedt 2003).

Linguistische Arbeitsschwerpunkte: Farbbezeichnungen, Sprachkurs (Spanisch), Stilistik, Wörterbuch (Bertelsmann Wörterbuch; Sachs-Vilatte), Wortschatz.

Knauers Auffassung von philologischer Forschung äußert sich darin, dass er generell von „einer Wissenschaft vom Wort" spricht, und „damit nicht Linguistik *oder* Literaturwissenschaft...[meint], sondern ihr *Ganzes*" (Knauer 1950, 1080).

Karl Knauer ist für die Quantitative Linguistik bedeutsam, weil er sich mit der Klangästhetik der romanischen Sprachen auf exakter Grundlage, d.h. mit Methoden wissenschaftlicher Stilistik, befasste und eindringlich für eine stärkere Berücksichtigung statistischer Methoden in Linguistik und Literaturwissenschaft plädierte. Er beklagte „ein Übergewicht an qualitativen Urteilen", ja „eine unleugbare Mathematikfeindlichkeit", die im Widerspruch zu „der von Platon geforderten zentralen Bedeutung der Mathematik" stehe:

„Bei der Beobachtung wissenschaftlicher Objekte machen wir, nicht anders als im täglichen Leben, teils qualitative, teils quantitative Feststellungen, wobei diese Erfahrungsarten durch vielfache gegenseitige Bedingtheiten miteinander verflochten erscheinen"; er fordert „eine grundsätzliche Aufrollung quantitativer Problemstellungen" (alle Zitate: Knauer 1955, 141f.). Zur Demonstration seiner Ansichten befasst sich Knauer (1955, 145ff.) u.a. mit der These, das Italienische weise spezifische Klangeigenschaften auf, indem er die Rangordnung der Lauthäufigkeiten des Italienischen mit der des Französischen vergleicht. In

Knauer (1958, 174) meint er, die Einsicht nehme zu, „daß die quantifizierende Arbeitsweise dazu geeignet ist, das Verständnis des Forschers über die Findung und Erklärung des neuen Einzelfalles in größere, unifizierende Zusammenhänge zu leiten."

Ein gutes Beispiel für Knauers Bemühungen um eine statistisch fundierte Stilistik ist seine Habilitationsschrift zu Marmontel. Er vertritt die These, dass Marmontel „seine stilistischen Bestrebungen so gut wie ausschließlich auf das Gebiet der lautlichen Formgebung beschränkt" (Knauer 1936: 144). Um seine Auffassung zu stützen vergleicht er u.a. die Verteilungen von rhythmischen Gruppen (Kola) in drei Abschnitten von Marmontels Roman *Les Incas, ou La destruction de l'empire du Pérou* (Paris 1777) mit denjenigen, die er in Werken von Rousseau und Voltaire vorfindet, welche in dieser Hinsicht weniger Auffälligkeiten zeigen. Zuvor hatte schon Servien (1930: 103, passim) entsprechende Erhebungen durchgeführt. Betrachtet man die ausgezählten Abschnitte aus heutiger Perspektive, dass Einheiten beliebiger Art, die in Texten in verschiedener Länge auftreten, bestimmten Gesetzmäßigkeiten folgen sollten (Wimmer, Köhler, Grotjahn, Altmann 1994; Wimmer & Altmann 1996, Wimmer, Witkovsky & Altmann 1999, u.a.), so kann man Knauers Annahme, dass Marmontel zumindest in den betreffenden Textabschnitten deutlich andere stilistische Mittel einsetzt als Rousseau und Voltaire, nur bestätigen.

Die von Knauer dargebotenen Textabschnitte wurden daraufhin geprüft, ob sie der verschobenen Hyperpoisson-Verteilung (hier in 1-verschobener Form)

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

folgen. Die Besonderheit der rhythmischen Gestaltung durch Marmontel wird u.a. dadurch deutlich, dass die Tests bei den Texten von Rousseau und Voltaire erfolgreich waren, bei den Textabschnitten von Marmontel aber nicht, wie die folgen Tabellen zeigen:

Tabelle 1
Verteilung rhythmischer Gruppen in *Chant de mort* des greisen Indianers

Kl	*x*	n_x	NP_x	*Kl*	*x*	n_x	NP_x
1	3	1	0.22|	7	9	3	8.99
2	4	1	1.23|	8	10	10	7.14
3	5	1	3.40	9	11	1	4.96
4	6	4	6.29	10	12	6	3.07
5	7	1	8.73	11	13	2	3.25
6	8	27	9.71				
$a = 5.5703$ $b = 1.0129$ $X^2 = 51.990$ $FG = 7$ $P = 0.00$							

Daten: Knauer (1936: 34); Text: *Les Incas*, Kap. XVII, T. I, S. 226ff. (Knauer 1936: 27-33).

Legende zu den Tabellen:

Kl: Längenklasse;
x: Zahl der Silben pro rhythmischer Gruppe;
n_x: beobachtete Häufigkeit der rhythmischen Gruppen mit *x* Silben;

NP_x: aufgrund der Hyperpoisson-Verteilung berechnete Häufigkeit der rhythmischen Gruppen mit x Silben;
a, b: Parameter der Hyperpoisson-Verteilung;
X^2: Wert des Chiquadrats;
FG: Freiheitsgrade;
P: Überschreitungswahrscheinlichkeit des Chiquadrats.

Die senkrechten Striche zeigen eine Zusammenfassung der betreffenden Längenklassen an. Die Anpassung der Hyperpoisson-Verteilung an die beobachteten Daten wird als erfolgreich betrachtet, wenn $P \geq 0.05$; $0.01 \leq P < 0.05$ signalisiert ein schwaches, gerade noch akzeptables Ergebnis.

Tabelle 1 zeigt, dass die Verteilung rhythmischer Gruppen in diesem Text bei den mittleren Längen einen mehrfachen Wechsel zwischen häufigen und seltenen Längen aufweist, und nicht, wie zu erwarten wäre, zunächst eine kontinuierliche Zunahme und dann eine ebenso kontinuierliche Abnahme der Häufigkeiten. Die Werte schwanken so stark, dass die Hyperpoisson-Verteilung ebenso wenig wie andere sonst verwendete Modelle angepasst werden kann. Dies gilt auch für die Cohen-C-Verteilung und die Pandey-Poisson-Verteilung, die einen Verschiebeparameter aufweisen und damit lokale Störungen in den Daten ausgleichen können. (Die von Knauer 1936: 34, Fußn. 1a angedeutete geringfügig andere Analyse des Textes ändert an diesem Befund nichts Wesentliches.)

Ein ähnlicher Fall wurde bereits bei Wortlängen in chinesischen Texten verschiedener Funktionalstile beobachtet (Best & Zhu 2001); auch da traten mehrfache Wechsel zwischen hohen und niedrigen Häufigkeiten im Verlauf der Daten auf. Durch Zusammenfassung von jeweils zwei Nachbarklassen konnte gezeigt werden, dass trotz der Schwankungen gesetzmäßige Verteilungen vorliegen. Wendet man dieses Verfahren auf die Verteilung rhythmischer Gruppen an, so lässt sich zeigen, dass im Fall *Chant de Mort* die Pandey-Poisson-Verteilung mit $P = 0.07$ erfolgreich angepasst werden kann. (Zur Auswirkung unterschiedlicher Zusammenfassungen vgl. Kelih & Grzybek 2004).

Die Hyperpoisson-Verteilung wird hier und in den folgenden Fällen trotz der schlechten Ergebnisse präsentiert, um den Vergleich mit den Texten von Rousseau und Voltaire zu ermöglichen. So wird die Besonderheit der rhythmischen Gestaltung durch Marmontel deutlich. Die Testergebnisse, die nach Zusammenfassung benachbarter Längenklassen erzielt werden, sollen aber genannt werden, um zu zeigen, dass auch in Marmontels Textgestaltung Gesetzmäßigkeiten nachweisbar sind.

Tabelle 2
Verteilung rhythmischer Gruppen im Dialog zwischen Alonzo und Cora

Kl	x	n_x	NP_x	Kl	x	n_x	NP_x
1	4	3	3.19	8	11	1	3.92
2	5	3	6.21	9	12	4	2.27
3	6	10	9.04	10	13	1	1.19|
4	7	0	10.50	11	14	0	0.58|
5	8	31	10.16	12	15	1	0.26|
6	9	3	8.41	13	16	0	0.11|
7	10	4	6.09	14	17	1	0.06|
$a = 5.7681$ $b = 2.9649$ $X^2 = 63.038$ $FG = 7$ $P = 0.00$							

Daten: Knauer (1936: 68f.); Text: *Les Incas*, Kap. XXVIII, T. II, S. 34ff. (Knauer 1936: 61-67). Bei Zusammenfassung von Nachbarklassen kann die Pandey-Poisson-Verteilung mit $P = 0.51$ angepasst werden.

Tabelle 3
Verteilung rhythmischer Gruppen im Abschnitt *Der Sonnenkult der Inkas*

Kl	*x*	n_x	NP_x	*Kl*	*x*	n_x	NP_x
1	1	1	1.24	12	12	9	5.00
2	2	1	1.95	13	13	2	4.12
3	3	2	2.85	14	14	1	3.25
4	4	0	3.86	15	15	1	2.47
5	5	5	4.88	16	16	4	1.80
6	6	10	5.79	17	17	3	1.26
7	7	0	6.45	18	18	2	0.85
8	8	26	6.80	19	19	0	0.56
9	9	1	6.78	20	20	1	0.35
10	10	3	6.43	21	21	1	0.50
11	11	0	5.80				
$a = 18.9293$ $b = 11.9727$ $X^2 = 94.556$ $FG = 15$ $P = 0.00$							

Daten: Knauer (1936: 91); Text: *Les Incas*, Kap. I, T. II, S. 30ff. (Knauer 1936: 81-89). Bei Zusammenfassung von Nachbarklassen ergibt die Cohen-C-Poisson-Verteilung $P = 0.07$.

Ein ganz anderes Bild ergibt sich, wenn man die rhythmischen Gruppen bei Rousseau und Voltaire untersucht. Knauer (1936) hat dazu je einen Textauszug verwendet und auf zweierlei Weise ausgewertet: einmal in „prosaischer Zählung“ und dann in „poetischer Zählung (Knauer 1936: 127). Die Auswertung unterscheidet sich dadurch, dass das „e muet“ und die metrischen Diphthonge bei der poetischen Zählung einen anderen Silbenwert erhalten als bei der prosaischen.

Tabelle 4
Verteilung rhythmischer Gruppen in Rousseau, *Nouvelle Heloïse*

		prosaische Auswertung		poetische Auswertung	
Kl	*x*	n_x	NP_x	n_x	NP_x
1	3	1	0.29	1	0.23
2	4	0	0.82	0	0.66
3	5	3	1.82	2	1.49
4	6	2	3.32	3	2.79
5	7	7	5.13	5	4.44
6	8	6	6.87	5	6.15
7	9	8	8.12	10	7.53
8	10	9	8.59	6	8.27
9	11	6	8.22	6	8.24
10	12	9	7.19	11	7.50
11	13	5	5.78	6	6.29
12	14	5	4.31	4	4.89

13	15	2	2.99	2	3.55
14	16	4	1.94	5	2.41
15	17	1	2.60	2	3.55
		$a = 10.0860$ $b = 3.5344$ $X^2 = 6.875$ $FG = 11$ $P = 0.81$		$a = 10.6514$ $b = 3.6972$ $X^2 = 8.455$ $FG = 10$ $P = 0.58$	

Daten: Knauer (1936: 140); Text: *Nouvelle Heloïse*, IV² partie, aus Brief XVII (Knauer 1936: 132-134).

Tabelle 5
Verteilung rhythmischer Gruppen in Voltaire, *L'homme aux quarante écus*

		prosaische Auswertung		poetische Auswertung	
Kl	*x*	n_x	NP_x	n_x	NP_x
1	3	2	1.24	1	1.11
2	4	2	3.19	2	2.62
3	5	4	6.30	3	5.01
4	6	12	10.13	10	8.09
5	7	12	13.71	12	11.27
6	8	19	16.04	13	13.81
7	9	20	16.53	20	15.12
8	10	12	15.20	13	14.95
9	11	15	12.63	14	13.47
10	12	7	9.58	6	11.16
11	13	6	6.67	10	8.55
12	14	3	4.30	6	6.09
13	15	3	2.58	5	4.06
14	16	2	1.44	3	2.54
15	17	2	1.45	2	1.49
16	18			1	1.68
		$a = 8.6074$ $b = 3.3569$ $X^2 = 6.344$ $FG = 12$ $P = 0.90$		$a = 10.2099$ $b = 4.3280$ $X^2 = 6.746$ $FG = 13$ $P = 0.91$	

Daten: Knauer (1936: 130); Text: *L'homme aux quarante écus, Des Proportions, Abschn. 1-13* (Knauer 1936: 123-126).

Die Texte von Rousseau und Voltaire unterscheiden sich bei der Verteilung der rhythmischen Gruppen von denen Marmontels dadurch, dass sie keine mehrfachen, auffälligen Wechsel zwischen häufigen und seltenen benachbarten Längenklassen aufweisen. Dies schlägt sich auch in den unterschiedlichen Anpassungen der Hyperpoisson-Verteilung nieder, womit eine weitere Bestätigung der Ideen von Knauer gewonnen ist. (Die gleiche Verteilung lässt sich übrigens mit $P = 0.68$ auch an die von Servien 1930: 103 mitgeteilte Datei zu Chateaubriand, *Atala* anpassen.)

Gegen Knauers Einteilung der Texte in rhythmische Gruppen wurden deutliche Vorbehalte erhoben (Cuénot 1938, 391). Solche Einwände sind m.E. nicht sehr erheblich, da derartige Auswertungen immer beträchtlich von der Interpretation der Texte abhängig sind, die sich bei verschiedenen Bearbeitern, ja sogar bei ein und demselben Bearbeiter zu verschiedenen Zeiten unterscheiden mag. Das gleiche Problem begegnete bereits im Zusam-

menhang mit der Untersuchung von Texten im Hinblick auf die Verteilung rhythmischer Einheiten (Best 2002, 138). Cuénot hat aber natürlich Recht mit dem Hinweis, dass man die Datenerhebung in höherem Maß intersubjektiv gestalten kann.

Ein anderes Thema taucht in Knauer (1965, [4]1971, 198) auf: hier widmet er sich dem Problem der Rekurrenz von [k] in 75 Sonetten Baudelaires, indem er u.a. die Zahl der Gedichte erfasst, in denen [k] an Versanfängen keinmal, einmal, zweimal usw. bis zu fünfmal vorkommt. Man kann dies m.E. als ein Diversifikationsphänomen (Altmann 1991) ansehen und entsprechend durch Anpassung der Poisson-Verteilung

$$P_x = \frac{e^{-a}\, a^x}{x!}, \quad x = 0,1,2,...$$

modellieren:

Tabelle 6
Sonette mit Versanfängen auf [k] bei Baudelaire

x	n_x	NP_x
0	15	15.46
1	26	24.41
2	17	19.28
3	13	10.15
4	3	4.01
5	1	1.69
a = 1.5794 X^2 = 1.723 FG = 4 P = 0.79		

Legende:

x	Zahl der Versanfänge auf [k];
n_x	beobachtete Zahl der Gedichte mit 0, 1, 2...5 Versanfängen auf [k];
NP_x	aufgrund der Poisson-Verteilung berechnete Zahl der Gedichte mit 0, 1, 2...5 Versanfängen auf [k];
a	Parameter der Poisson-Verteilung.

Knauer nutzt diese Beobachtungen, um eine stärkere Affinität Baudelaires zu [k] als zu [p] und [t] zu konstatieren. Aus heutiger Sicht ist aber auch bedeutsam, dass hier wieder eine gesetzmäßige Verteilung nachweisbar ist.

Die dargebotenen Beispiele von Textauswertungen vermitteln einen ersten Eindruck von Knauers Vorgehensweise bei der klangästhetischen Analyse von literarischen Kunstwerken. Die beiden folgenden Zitate (Knauer 1955, 148) verdeutlichen nochmals Knauers Vorstellungen von der Bedeutung der Statistik:

„Die mathematische Statistik sucht ... die der direkten Anschauung oft verborgenen, aber für das Verständnis wesenhafter Zusammenhänge im unbelebten wie im belebten Reich der Dinge wichtigen Ordnungen auf und ist deshalb in den jüngsten Jahrzehnten ... in den verschiedensten Zweigen der Forschung zu einem bedeutenden Erkenntnismittel geworden." Und: „Die wissenschaftliche Statistik stellt sich die Aufgabe, durch möglichst wenige und möglichst einfache Ausdrücke in mathematischer Prägung die in solchen Häufigkeitsverteilungen angelegten Gesetzmäßigkeiten kenntlich zu machen."

Knauer hat gewiss seinen Beitrag „zur Schaffung einer Forschungsatmosphäre, die in unseren sprach- und literaturwissenschaftlichen Disziplinen wünschenswert und notwendig erscheint“ (Knauer 1955, 149), geleistet. Außerhalb der Romanistik dürfte sein Beitrag in der Sammlung von Gunzenhäuser & Kreuzer (1965, ^{4}1971) noch am ehesten bekannt sein. Mehr als ein schwaches, zeitweise versickerndes und gelegentlich mit großem Unverständnis aufgenommenes Rinnsal in der Gesamtentwicklung der philologischen Forschung ist aus derartigen Bemühungen und Plädoyers aber für lange Zeit nicht geworden.

Es ist hier nicht der Ort, Knauers klangästhetische Untersuchungen im Detail zu entwickeln; die gegebenen Andeutungen sollen genügen. Die folgenden Literaturhinweise mögen ggfs. den Weg zu einer intensiveren Befassung mit ihnen weisen. Es ging vielmehr darum, zu zeigen, dass Knauer wissenschaftliche Positionen bezogen hat, die die Position der Quantitativen Linguistik stärken, und dass er außerdem zu sonst kaum beachteten Themen Daten erarbeitete, die man heute aus neu begründeter theoretischer Perspektive aufgreifen kann. Er gehört zu den Autoren, die die Quantitative Linguistik wiederentdecken sollte.

Literatur

Diese Liste enthält, abgesehen von Rezensionen, die meisten wissenschaftlichen Arbeiten von Knauer, da eine solche Übersicht nirgends in publizierter Form zu existieren scheint. Einige Arbeiten, die in der Bearbeitungszeit nicht zu verifizieren waren, fehlen allerdings.

Altmann, Gabriel (1991). Modelling diversification phenomena in language. In: Rothe, Ursula (Hrsg.), *Diversification Processes in Language: Grammar* (S. 33-46). Hagen: Margit Rottmann Medienverlag.

Best, Karl-Heinz (2002). The distribution of rhythmic units in German short prose. *Glottometrics 3, 136-142.*

Best, Karl-Heinz, & Zhu, Jinyang (2001). Wortlängenverteilungen in chinesischen Texten und Wörterbüchern. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten* (S. 101-114). Göttingen: Peust & Gutschmidt.

Cuénot, C. (1938). Rez. zu Knauer (1936). *Zeitschrift für romanische Philologie 58, 389-393.*

Gunzenhäuser, Rul, & Kreuzer, Helmut (1965, ^{4}1971). *Mathematik und Dichtung. Versuche zur Frage einer exakten Literaturwissenschaft.* 4. durchgesehene Auflage. München: Nymphenburger.

Hausmann, Frank-Rutger (2000). *„Vom Strudel der Ereignisse verschlungen.“ Deutsche Romanistik im „Dritten Reich“.* Frankfurt/M.: Vittorio Klostermann.

Kelih, Emmerich, & Grzybek, Peter (2004). Häufigkeiten von Satzlängen: Zum Faktor der Intervallgröße als Einflussvariable (am Beispiel slovenischer Texte). *Glottometrics 8, 23-41.*

Knauer, Karl (1930). Beiträge zum Ausdruck von Abstraktem im Französischen. *Romanische Forschungen XLIV, 185-254.* (Diss. München)

Knauer, Karl (1935). Lodovico Ariosto. Zum Wesen und Wirken seiner Kunst. *Germanisch-romanische Monatsschrift XXIII, 368-389.*

Knauer, Karl (1936). *Ein Künstler poetischer Prosa in der französischen Vorromantik: Jean-François Marmontel. Habilitationsschrift.* Bochum-Langendreer: Druck: Heinrich Pöppinghaus.

Knauer, Karl (1936). Voltaire- und Rousseau-Stil. *Germanisch-romanische Monatsschrift XXIV, 282-299.*

Knauer, Karl (1937). Die klangästhetische Kritik des Wortkunstwerks am Beispiel französischer Dichtung. *Deutsche Vierteljahresschrift für Literaturwissenschaft und Geistesgeschichte 15, 69-91.*

Knauer, Karl (1937). Musik in der Sprache – Wohlklang und Mißklang. *Sprachkunde Nr. 5, 11-13.*

Knauer, Karl (1938). Französische Sprache als Klangkunst und als Gegenstand ästhetischer Forschung. *Zeitschrift für französische Sprache und Literatur LXI, 257-272.*

Knauer, Karl (1942). Einheit und Vielgestalt im Klangbau der europäischen Sprachen. *Sprachkunde, Nr. 2, 4-5.*

Knauer, Karl (1943). Exaktheit in der Wissenschaft vom Wort. *Sprachkunde, Nr. 3/ 4, 1-3.*

Knauer, Karl (1943). Sprachwissenschaftliche Klangästhetik auf exakter Grundlage. *Helicon IV, 147-160.*

Knauer, Karl (1943/44). Die Proportionen des sprachlichen Lautgefüges. Eine methodische Studie über den Klangcharakter der französischen Umgangssprache. *Zeitschrift für französische Sprache und Literatur LXV, 47-52.*

Knauer, Karl (1950). Grenzen der Wissenschaft vom Wort. *Akademie der Wissenschaften und der Literatur in Mainz, Abhandlungen der Geistes- und Sozialwissenschaftlichen Klasse, Jahrgang 1950, Nr. 13, 1077-1093.* Verlag der Akademie der Wissenschaften und der Literatur in Mainz; Wiesbaden: in Komm. Franz Steiner Verlag.

Knauer, Karl (1953). Quelques aspects de l'exigence d'exactitude en critique littéraire. *Essais de philologie moderne (1951): Bibliothèque de la Faculté de Philosophie et Lettres de L'Université de Liège, Fasc. CXXIX, 201-208.* Paris: Société d'Édition „Les belles Lettres".

Knauer, Karl (1954). *Vulgärfranzösisch. Charakterzüge und Tendenzen des gegenwärtigen französischen Wortschatzes.* München: Hueber.

Knauer, Karl (1955). Grundfragen einer mathematischen Stilistik. *Forschungen und Fortschritte 29, 140-149.*

Knauer, Karl (1956). Der Anteil der Sprachwissenschaft an der Erforschung menschlicher Verhaltensweisen mit Hilfe quantifizierender Methoden. *Actes du IV^e^ congrès international des sciences anthropologiques et ethnologiques, Vienne, 1-8 Septembre 1952. Tome III: Ethnologica. Seconde Partie et Rapport Général, 213-223.* Wien: Holzhausen.

Knauer, Karl (1956). Die Einbeziehung des Mengenfaktors in die Interpretation sprachästhetischer Ordnungen dargestellt an einem Problem aus Théophile Gautiers Reimstilistik. *Forschungen und Fortschritte 30, 13-17.*

Knauer, Karl (Bearb.) (1956). *Nachtrag 1956 zu Sachs-Villatte, Enzyklopädisches Wörterbuch der französischen und deutschen Sprache. Teil II: Deutsch – Französisch.* Berlin: Langenscheidt.

Knauer, Karl (1958). Ein Gestaltproblem der altfranzösischen Laissenkunst. Als Beitrag zur Erforschung meßbarer Ordnungsgebilde in Werken der Dichtung. *Forschungen und Fortschritte 32, 174-179.*

Knauer, Karl (1964). Elektronische Filter bei der Untersuchung sprachlicher Klangstrukturen. *Forschungen und Fortschritte 38, 332-337.*

Knauer, Karl (1965; [4]1971). Die Analyse von Feinstrukturen im sprachlichen Zeitkunstwerk. Untersuchungen an den Sonetten Baudelaires. In: *Mathematik und Dichtung. Versuche zur Frage einer exakten Literaturwissenschaft: 193-210.* Zusammen mit Rul Gunzenhäuser hrsg. von Helmut Kreuzer. 4. durchgesehene Auflage. München: Nymphenburger.

Knauer, Karl (1965). Über Klangfarbenstrahlung bei Baudelaire. *Zeitschrift für französische Sprache und Literatur LXXV, 228-246.*

Knauer, Karl (1967). Die Klangfarbendriften in Baudelaires Alexandrinersonetten. *Archiv für das Studium der neueren Sprachen und Literaturen 118. Jg./ 203. Bd., 272-277.*

Knauer, Karl, & Knauer, Elisabeth (1960). *Bertelsmann Wörterbuch: Französisch – Deutsch/ Deutsch – Französisch.* Gütersloh: Bertelsmann.

Knauer, Karl, & Pizzaro, José ([15]1957). *30 Stunden Spanisch für Anfänger.* Berlin: Langenscheidt.

Lausberg, Heinrich (1980). Die Romanistik an der Universität Münster. In: Dollinger, Heinz (Hrsg.), *Die Universität Münster 1780-1980* (S. 401-410). Münster: Aschendorff.

Schuder, Werner (Hrsg.) (1966). *Kürschners Deutscher Gelehrten-Kalender 1966. Zehnte Ausgabe. A-M.* Berlin: de Gruyter.

Servien, Pius (1930). *Les rhythmes comme introduction physique à l'estétique. Nouvelles méthodes d'analyse et leur application notamment à la musique, aux rhythmes du français et aux mètres doriens.* Paris: Boivin.

Untiedt, Frank (2003). *Die Fächer Anglistik und Romanistik an der Westfälischen Wilhelms-Universität in der NS-Zeit.* Staatsexamensarbeit, Münster.

Wimmer, Gejza, & Altmann, Gabriel (1996). The Theory of Word Length Distribution: Some Results and Generalizations. In: Schmidt, Peter (Hg.), *Glottometrika 15: 112-133.* Trier: Wissenschaftlicher Verlag Trier.

Wimmer, Gejza, Köhler, Reinhard, Grotjahn, Rüdiger, & Altmann, Gabriel (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics 1, 98-106.*

Wimmer, Gejza, Witkovský, Viktor, & Altmann, Gabriel (1999). Modification of Probability Distributions Applied to Word Length Research. *Journal of Quantitative Linguistics 6, 257-268.*

Ich danke Karl-Heinz Kieselbach, Elisabeth Knauer-Romani, Marie-Rose Wörner und dem Universitätsarchiv, Westfälische Wilhelms-Universität Münster, für persönliche und fachliche Informationen.

Karl-Heinz Best, Göttingen

XVIII. August Friedrich Pott (1802-1887)

Pott wurde am 14.11.1802 in Nettelrede (bei Bad Münder am Deister) geboren und verstarb am 5.7.1887 in Halle. Er begann sein Studium 1821 in Göttingen mit Theologie, fühlte sich aber vor allem zur Philologie hingezogen und hörte u.a. bei dem Germanisten Georg Friedrich Benecke. Lehrer am Gymnasium in Celle. 1827 Promotion in Göttingen; nach Habilitation in Berlin 1833 Professor für allgemeine Sprachwissenschaft in Halle (vgl. Bense 1979). Horn (1888: 317) bezeichnet Pott in seinem Nekrolog als „nestor der sprachforscher, der letzte der noch lebenden begründer der vergleichenden sprachforschung." Seine Themen waren u.a.: Etymologie, Zigeunersprache, Zahlwörter, Personennamen, Doppelungen; seine von Koerner (1973: 14, 21, 26) besonders gewürdigten Werke *Zur Litteratur der Sprachenkunde Europas* (1887) und *Einleitung in die allgemeine Sprachwissenschaft* (1884-1890) geben eine Übersicht über das linguistische Wissen seiner Zeit und sind zugleich kommentierte Führer zur entsprechenden Literatur.

Für die Quantitative Linguistik ist auf Potts *Einleitung in die allgemeine Sprachwissenschaft* (1884-1890) hinzuweisen, deren 1. Teil 1884 erschien (Best ²2003: 8). Hier werden mehrere einschlägige Themen angeschnitten:

1. Pott (1884: 19) bezieht sich auf Leibniz, *de arte combinatoria*, wo ja die Frage behandelt wird, wie viele Wörter man bilden kann, wenn ein Alphabet eines bestimmten Umfangs zur Verfügung steht. Dass Leibniz hiermit in einer langen Tradition steht, kommt allerdings nicht zum Ausdruck (Best 2005a,b). In diesem Zusammenhang schneidet Pott (1884: 19f.)

2. Ein weiteres bedeutsames Thema an: Die Zahl der denkbaren Wörter ist ja davon abhängig, wie lang Wörter in einer Sprache sein können. Hierzu gibt Pott einen groben Überblick, in dem er vor allem auf die langen Wörter in mittel- und nordamerikanischen Sprachen sowie im Grönländischen (Potts Terminus) hinweist und einige Wörter in Silben und Buchstaben beziffert. Auch ein althochdeutsches Beispiel findet sich hier.

3. Pott relativiert Leibniz' Berechnung mit dem Hinweis, dass ja nicht jeder Buchstabe (Laut) mit jedem anderen kombiniert werden könne und selbst in allen Sprachen der Welt zusammen die berechnete Zahl der Wörter nicht vorkomme. Er führt diese Überlegungen mit Hinweisen auf die Zahl der Verbalwurzeln und Formelemente im Sanskrit fort, das „nicht mehr als ... 2000 V e r b a l w u r z e l n und ... höchstens 200 f o r m a l e [...] E l e m e n t e [...]" (Pott 1884: 20) enthalte und weist auch darauf hin, dass ja nicht jede Wurzel mit jedem Formelement kombiniert werden könne.

4. Ein weiteres Thema, das er in diesem Zusammenhang anschneidet, ist die Frage danach, wie viele Wörter denn ein Individuum verwenden könne (Pott 1884: 21). Ohne sich diese Zahlen zu eigen zu machen, berichtet er, dass man Angaben finden könne, die zwischen weniger als 300 Wörtern bei vielen Mitgliedern „der arbeitenden Klasse" (Pott 1884: 21) und „ungefähr 15000 Wörtern" (Pott 1884: 21) in den Dramen Shakespeares schwanken.

5. Auf den Klangeindruck von Sprachen bezogen meint er: „Von besonderer Wichtigkeit betreffs der Gesamtwirkung eines Sprachidioms auf das Ohr und sonst ist aber das **statistische** Verhalten der Lautklassen und Einzellaute..." Er verweist in diesem Zusammenhang auf Arbeiten seines Schülers Förstemann, der als erster Untersuchungen zur Lautstatistik durchgeführt habe, und auf Whitney. Man muss ergänzen: Es ging Förstemann damit vor allem um sprachvergleichende Untersuchungen, die Rückschlüsse auf die verwandtschaftlichen Verhältnisse zwischen den Sprachen zulassen sollten (Best 2006).

6. Dass für Pott die Statistik ein wichtiges Mittel der Erkenntnis ist, wird später noch einmal deutlich, wenn er im Zusammenhang mit Wortbildungsaffixen „eine s t a t i s t i s c h begründete Einsicht in die M i t t e l" fordert, „worüber eine gegebene Sprache im ganzen oder einzelnen zu verfügen hat. An sich ist es doch auch wissenswert, zu erfahren, wie sich dieses oder jenes Idiom desfalles gegen andere im Vorteil oder Nachteil befindet" (Pott 1884: 46).

7. Ein letztes sprachstatistisches Thema ist zu erwähnen: die Frage nach der Zahl der Sprachen der Erde. Im Zusammenhang mit der Frage nach der Klassifikation wird gefragt, wie viele Sprachen es denn gebe (Pott 1884: 51f.), und er meint, man könne „vielleicht die Zahl 1000" (Pott 1884: 68) erreichen.

Man kann also konstatieren, dass Pott etliche Themen ansprach, die für die Weiterentwicklung der Quantitativen Linguistik bedeutsam waren. Allerdings muss man mit Koerner (1974: VII) feststellen: „Pott's contribution to the study of language ... was already beginning to be largely ignored during the 1870's and 1880's." Dies gilt anscheinend in besonderem Maße auch für seine sprachstatistischen Hinweise.

Literatur

Bense, Gertrud (1976). Bemerkungen zu theoretischen Positionen im Werk von A.F. Pott. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 29, 519-522.*

Bense, Gertrud (1979). August Friedrich Pott 1802-1887. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 32, 19-23.*

Best, Karl-Heinz (²2003). *Quantitative Linguistik: Eine Annäherung.* 2., überarb. u. erw. Auflage. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (2005a). Georg Philipp Harsdörffer (1607-1658). *Glottometrics 9, 86-88.*

Best, Karl-Heinz (2005b). Gottfried Wilhelm Leibniz (1646-1716). *Glottometrics 9, 79-82.*

Best, Karl-Heinz (2006). Ernst Wilhelm Förstemann (1822-1906). *Glottometrics.* In Arbeit.

Gabelentz, Georg von der (1888). Pott. *Allgemeine deutsche Enzyklopädie 26, 478-485.* Leipzig: Duncker & Humblot. Wieder abgedruckt in: Sebeok, Thomas A. (ed.) (1966). *Portraits of Linguistis. A Biographical Source Book for the History of Western Linguistics, 1746-1963. Vol. 1: From Sir William Jones to Karl Brugmann* (S. 251-261). Bloomington/ London: Indiana University Press.

Horn, Paul (1888). August Friedrich Pott. *Beiträge zur Kunde der indogermanischen Sprachen 13: 317-341.* Neudruck in: August Friedrich Pott, EINLEITUNG IN DIE ALLGEMEINE SPRACHWISSENSCHAFT preceded by the same author's ZUR LITERATUR DER SPRACHENKUNDE EUROPAS. Newly edited together with a bio-bibliographical sketch of Pott by Paul Horn by E.F.K. Koerner. With a preface and a new index of names: XVII-XLI. Amsterdam: John Benjamins 1974.

http://www.catalogus-professorum-halensis.de/indexb1871.html

Koerner, E.F.K. (1973). *THE IMPORTANCE OF F. TECHMER'S „INTERNATIONALE ZEITSCHRIFT FÜR ALLGEMEINE SPRACHWISSENSCHAFT" IN THE DEVELOPMENT OF GENERAL LINGUISTICS.* Amsterdam: Benjamins.

Koerner, E.F.K. (1974). Preface to the new edition. In: August Friedrich Pott, *EINLEITUNG IN DIE ALLGEMEINE SPRACHWISSENSCHAFT preceded by the same author's ZUR LITERATUR DER SPRACHENKUNDE EUROPAS.* Newly edited together with a bio-bibliographical sketch of Pott by Paul Horn by E.F.K. Koerner. With a preface and a new index of names: VII-XVI. Amsterdam: John Benjamins 1974.

Leopold, Joan (1983). *The letter liveth. The life, work and library of August Friedrich Pott (1802-1887).* Amsterdam: Benjamins.

Plank, Frans (1993). Professor Pott und die Lehre der Allgemeinen Sprachwissenschaft. *Beiträge zur Geschichte der Sprachwissenschaft 3, 95-128.*

Pott, August Friedrich (1884). Einleitung in die allgemeine Sprachwissenschaft. *Internationale Zeitschrift für allgemeine Sprachwissenschaft 1 (= Techmers Zeitschrift), 1-68.* Neudruck in: August Friedrich Pott, EINLEITUNG IN DIE ALLGEMEINE SPRACHWISSENSCHAFT preceded by the same author's ZUR LITERATUR DER SPRACHENKUNDE EUROPAS. Newly edited together with a bio-bibliographical sketch of Pott by Paul Horn by E.F.K. Koerner. With a preface and a new index of names: 201-268. Amsterdam: John Benjamins 1974.

Für Hinweise zu A.F. Pott danke ich Frau Gertrud Bense.

Karl-Heinz Best, Göttingen

Book review. *Kvantitativnaja Lingvistika: Issledovanija i modeli* (Klim-2005). Materialy Vserossijskoj naučnoj konferencii (6-10 ijunja 2005 g.). Novosibirsk: Novosibirskij Gosudarstvennyj Pedagogičeskij Universitet. (Redakcionnaja kollegija A.A. Polikarpov, G.G. Sil'nickij, V.V. Poddubnyj). Reviewed by **Emmerich Kelih.**

0.

The book under review, "Quantitative linguistics: Analyses and models (Klim 2005). Proceedings of the All-Russian Scientific Conference of June 6-10, 2005" (hereafter Klim 2005), is an omnibus volume containing the proceedings of the conference on quantitative linguistics held in Novosibirsk (Russia).

The conference was organized by the National Novosibirsk Pedagogical University – the publisher of the omnibus volume – under the chairmanship of the scientific secretary of the conference Victor V. Kromer. The lasting importance of the conference is attested to by the fact that 45 scientists[1] took part in it. The volume, which appeared in 2005, provides an ideal starting point for reviewing individual contributions.

The 370-page volume (Klim 2005) contains 35 contributions, written in Russian by 39 authors. The editors ("redakcionnaja kollegija") were A.A. Polikarpov, G.G. Sil'nickij and V.V. Poddubnyj. Contrary to usual practice, Klim (2005) does not – unfortunately – contain either a preface or an introduction by the editors. This would provide at least an initial survey of the main topics of the conference or the published papers. Also somewhat striking is that the book is not subdivided into thematic sections. It would, of course, be possible to recognize a thematic level-subdivision on the basis of the order of the papers (from phonosemantics through style studies up to general theoretical questions); but in this review we shall, rather, consider the papers as being classified into three domains: (1) quantitative text linguistics, (2) quantitative linguistics, and (3) computational linguistics/data mining. The review of particular analyses will be followed by (4) a general summary of the most important research areas in Russian quantitative text and language analysis resulting from Klim (2005).

1.

The first thematically extensive division in Klim (2005) is quantitative text linguistics. The defining factor of this systematisation is an explicit interest in individual texts (distinguishing style and verse analysis respectively), authorship attribution, and general considerations on "text homogeneity".

Let us start with the domain of quantitative style analysis, which rests quite on a long-standing tradition in quantitative research in text science. A consistent interest in the study of sentence length and the frequency of syntactic constructions may be observed in three of the contributions to Klim (2005).

T.V. Džurjuk ("Sentence length as a parameter of individual author style of German writers at the beginning of the 20th century: based on data from F. Kafka, I. Keun, Th. Mann and K. Tucholsky", pp. 182-194)[2] reports on studies of length and frequency of punctuation in sentences. The frequency of sentence lengths has been examined statistically in more detail by means of the chi-square contingency test and the contingency coefficient. According to the author, sentence lengths pooled to groups ("short", "mean" and "long" sentences) give important information about the stylistic individuality of the writers mentioned above. The high proportion of short sentences found in the texts motivates the examination of the hitherto neglected text-internal heterogeneity: we may distinguish direct speech, speech insertions and narrative text parts. Even if the examination of these text components is restricted to the state-

[1] The number of participants has been calculated from the conference's homepage (http://klim.nightmail.ru/).
[2] The translations of individual titles were made by the reviewer. In some cases the titles have been shortened on stylistic grounds without substantially changing the meaning.

ment of absolute frequencies for each writer, it would clearly be possible to derive from these facts a broader relationship between sentence length and its affiliation to a certain text component.

Quite similar, with regard to the methods applied, is the contribution of *Ju. P. Bojko* ("Selectivity of the author's style at the sentence level", pp. 303-315). The author is not directly interested in sentence length, but rather in the frequency of syntactic constructions (subject, predicative clauses, etc.) in English writers. Furthermore, in this case the ascertained "range" of frequency of occurrence is interpreted as originating from the individuality of the author. Because of the fulfilled syntactic segmentation of the sentence, its length should be examined as a subsequent step.

Apart from this descriptive capturing of authorial style based on subtly distinguished units of investigation (parts of speech, clause) and tested by means of the chi-square criterion, *V.V. Poddubnyj* and *O.G. Ševelev* ("Comparison and cluster analysis of text features (frequency) on the basis of the hypergeometric criterion", pp. 205-217) address the use of the clustering procedure in quantitative text analysis and text typology. To this end, the authors use the sentence length in works by several Russian authors. The text size is given in the somewhat unusual specification of megabytes (MB). The authors examine in greater detail whether the combination of the assumed "syntactic" parameters, e.g. frequency of sentences (measured in quite arbitrary fixed lengths of 5, 10 etc. words) or the proportion of chapters, can influence the "effectivity" of text classification when using the clustering method. The partially ascertained differences between individual clustering results are, however, not interpreted further; the authors are content with the resulting classification. Hence, on this point we must agree with the authors' statement that further experimental examinations and a theoretical foundation of the results are required (cf. Klim 2005: 216).

Three further articles are dedicated to the quantitative study of disputed authorship. The well-known St Petersburg specialist in this domain, *M.A. Marusenko*, discusses jointly with *E.E. Mel'nikova* and *E.S. Rodionova* ("Attribution of anonymous and pseudoanonymous articles published in the journals 'Vremja' and 'Epocha' in 1861-1865", pp. 283-293) the problem of certain articles of disputed authorship that have, up until now, generally been ascribed to Dostoevsky. The quantitative criteria for the solution of this problem are the length of simple sentences, the number of sentences without a subject, and some further morphosyntactic considerations. With regard to the categorization of the applied properties, Marusenko (1990) is still very helpful. The previous attribution of these articles to Dostoevsky is challenged on the basis of computed Euclidean distances. Instead, it is supposed that the disputed articles could have been written by several authors.

The same problem is followed up in the article by *N.V. Semjannikova* ("Verification of attribution methods with translated texts", pp. 294-302). However, the author is less interested in content problems than in testing the significance of different methods. She postulates the inadequacy of (1) the Q-sum method (cf. Tweedie 2005: 393f.) for Russian, although no empirical evidence is presented. On the other hand, (2) N.A. Morozov's well-known method of "authors' invariant" (frequency of prepositions) clearly allows, in the opinion of the author, statements about the author's individuality to be made. This interpretation arises solely from graphical representations and is not – as should be expected – substantiated by statistical testing. The greatest success rate is ascribed to (3) the method of frequency of letter bigrams, based on Khmelev's (2000) considerations. This contribution should be considered an approximation to the problem of finding an adequate method of attribution. The stated intention, namely comparing translated texts by different writers, may be considered quite a prolific perspective.

A further methodologically oriented contribution, which nevertheless can be placed in the domain of authorship attribution, is that of *A.P. Kovalevskij* ("Application of the invariance

principle in the analysis of text homogeneity", pp. 195-204).[3] The author is interested in the problem of whether a text is homogeneous, i.e. not written by two or more authors. The concrete motivation is the discovery of an objective criterion for identifying plagiarism, for example in schoolwork, which do not necessarily originate from one author. The approach taken here to this interesting problem is to some extent original. In ascertaining text homogeneity or individual author style, the starting assumption is that the authors use the same words with different frequency. In order to obtain an adequate basis for investigation, the author sets up frequency lists of words for several Russian novels (by Dostoevsky, Tolstoy, and Bulgakov). From these frequency lists, an intersection set of words occurring in all the novels is constructed. This intersection set is called "standard language" *Z* and it contains 27,000 word forms. Using a special stochastic process (Brownian motion), the author tries to show that a sudden increase in the occurrence of frequent words of the "standard language" in a text block (which is not clearly defined) is a break of homogeneity and signals plagiarism. The relevance of the method is illustrated by a comparison of a literary text, its retelling by a pupil, and a combination of these two texts. In this paper, two things seem to be lacking. On the one hand, the reader feels the need for a discussion of the complex and multi-layered problem of text plagiarism; on the other hand, we might ask for the linguistic reasons for the relevance of high-frequency words in plagiarism detection. Further, it is quite conspicuous that the rich Russian tradition of systematic compilation of frequency dictionaries is totally ignored. These could surely serve as a starting point for this approach. Last but not least, the law of Frumkina must be consulted as an initial step.

The contribution of *O.N. Grinbaum* ("Quantitative analysis of the verse: from linguistics to 'art-metrics'", pp. 94-107) is concerned with quantitative verse analysis. The theoretical interest is concentrated on rhythm, which together with meter represents an important research field both in literary and linguistic domains. Grinbaum is, however, not interested in the empirical study of rhythm, but rather in an adequate definition of this phenomenon. In this connection, he provides a thorough survey of one hundred years of the Russian discussion of the concept of "rhythm". From this study, it is evident that he does not considers rhythm in the structuralist sense as a sequence of accents, but defines it as a "harmony of relationships". In addition, such a conception, in sense of the Golden Mean, can be quantified. The author follows this up intensively elsewhere (Grinbaum 2000)

A.S. Gumenjuk and *A.S. Kostyšin* work directly in the domain of quantitative verse analysis. In their investigation ("Acoustic elements of verses and the formal procedures of their recognition", pp. 34-43), the authors present a new procedure of text segmentation, whose fundamentals have been shown in detail already in Gumenyuk, Kostyshin and Simonova (2002) and Gumenjuk et al. (2004). A text is segmented into acoustic text units called 'consonances' which can be set up in different length forms on the basis of neighbouring phonemes with the aid of an algorithm.

Actual empirical examination of these acoustic text units is given in the next contribution by *A.S. Kostyšin* ("Investigations into a segmentation algorithm", pp. 44-60). Here texts are segmented by means of the algorithm mentioned above in acoustic text units of different length, and a frequency dictionary is compiled. In this connection, the well known hypothesis of Ju.K. Orlov concerning the validity of Zipf-Mandelbrot law for full and closed texts (Zipf size) is applied as a criterion or test procedure. It should corroborate the correctness of the chosen segmentation of poetic texts in its psycholinguistic relevance.

Some further articles in Klim (2005) can be classified as selected problems in quantitative text analysis and empirical literary science. *N.L. Zeljanskaja* ("Philological reception of a literary fact: experimental analysis and modelling", pp. 61-72) presents an empirical inquiry

[3] A reference from this contribution could not be stated correctly, namely "Herdan, E. Calculus of legomena / E. Herdan. – N.-Y. 1964" (cf. Klim 2005: 204).

into the assessment of the literary and cultural-historical relevance of classical Russian authors. To this end, she interviews test persons considered by her to be professional readers (philologists, literature scientists). Unfortunately, it is not possible to reconstruct from this article what is to be examined. Even if there are cursory hints to the statistical evaluation of the questionnaires – up to now 7 persons have been interviewed – the aim and the method of this contribution remains generally unclear.

An innovative and original contribution to experimental methods in the domain of linguistics and text science goes back to *K.I. Belousov* ("Modelling the interplay of intratextual spaces", pp. 73-93): 21 test persons segmented a given text into "micro-themes" and on lexico-semantic basis ascribed the word occurring therein to another word. The frequency of this ascription performed by test persons serves as a starting point for a graphic representation of the so-called semantic-intratextual relationships. In the next step, these subjective text relationships are supplemented by further prosodic properties: the test persons must expressively read aloud the text they have segmented into semantic relations. The reading is recorded. In the opinion of the author, it is possible to set up a relationship of "semantic" and "prosodic" contour between the semantic-lexical intratextual relations of a text and the loudness of the declaimed part of the text. Apart from the unclear meaning of "intratextal" semantic text relationships, one might surely question whether loudness, as a property of the prosodic contour taken into account by the author, should not be supplemented by further experimental phonetic factors (intonation, speech pause, etc.)

The investigation of *Ju.N. Kovšova* and *V.A. Seleznev* ("Measurement of the activity of the supporting and the logical language functions for the ascertainment of text quality", pp. 137-145) treats the problem of a quantifiable evaluation of the quality of mathematics textbooks. The starting point is text parts with illustrating ("supporting") and explanative ("logical") functions. Somewhat strange is the quantitative determination of these text parts: it is not the frequency or the length of the passages having this function in textbooks but their *area*, measured in square centimetres. The areas yield some ratios but these are not standardized. Their values in individual chapters are simply juxtaposed with those of other textbooks. This method, which according to the authors should enable a statement about the "didactic" quality of textbooks, must be critically scrutinized because it represents a very raw, exclusively optical approximation to a very interesting problem.

A slightly different, but in any case acceptable, view of text homogeneity can be found in *V.A. Seleznëv* and *E.V. Isaeva* ("The Hurst parameter in word sequences", pp. 146-151). Considering the text as a time series, the sequential order of word lengths measured in terms of letter numbers is analysed: first in full, closed Russian novels, then in texts which were "shuttled" by means of a special algorithm. As a matter of fact, Hurst's parameter is different in differently "processed" texts. But also, in this case, the ascertained differences are interpreted exclusively on the background of graphical presentations. Nevertheless, the method raises the possibility of quantitatively scrutinizing the semantic "unity" of texts.[4]

Similar to this contribution is a further examination of text homogeneity by means of the Hurst parameter and Brownian motion by *N.S. Zakrevskaja* ("Study of text homogeneity using the moving average", pp. 26-33). The author studies the sequence of word lengths, measured in terms of syllable numbers, in two literary texts and in a "quasi-text" representing a cumulative mixture of the two texts. In contrast to the previous article, the author is slightly more cautious in the interpretation of her results, and understands her contribution explicitly as a test of a method using texts as illustrations (cf. Klim 2005: 33).

[4] For this investigation one would expect at least a reference to the first discussion of Hurst´s parameter by Krylov (1994: 113f) or the detailed discussion accompanied by a series of empirical investigations concerning sentence length by Hřebíček (1997: 132ff.).

2.

Another large thematic block in Klim (2005) consists of works belonging to the "core domain" of quantitative linguistics. Here, the characterisation of the text/author is of secondary importance. Rather general language properties and language phenomena that are scrutinized by means of quantitative methods become the focus of attention.

Four contributions of the "Smolenks Group" led by Prof. V.V. Silnickij belong in this block. The starting point of this group is the use of different correlation methods for stating the mutual relationships between phonetic, morphosyntactic and semantic features of language. Until now, the English verb has stood especially in the focus of attention (cf. Sil'nickij et al 1990). On the basis of the articles published in Klim (2005), some new meaningful problems and methodological innovations can be envisioned.

In the contribution of *A.G. Sil'nickij* ("Quantitative semantic classification of "economic" situations of modeling English verbs", pp. 167-181) 500 English verbs (especially those describing economic relations between subjects) are analyzed as to their syntactic-semantic properties. On the basis of a maximal number of different criteria using correlation analysis, verb subclasses are identified that should be put in mutual relations. A further step using cluster analysis corroborates the result.

A classification of German verbs can be found in *E.A. Il'jušina und D.S. Goršenin* ("Testing the feature system of German verbs using multivariate procedures", pp. 364-368). Here a classification of 4892 verbs is performed using 33 properties (phonetic to semantic) and the clustering method. Because of the shortness of the paper, and the omission of tabular and graphical presentations of individual steps in the analysis, no direct conclusions may be drawn for the time being.

In contrast to this, the aim of *L.A. Kuz'min's* contribution ("Correlations between different levels of adjectives in modern English", pp. 108-121) is much more clearly evident: he is concerned with internal mutual relations within the class of English adjectives and connects phonetic, morphological, morphosyntactic, etymological, chronological and semantic properties, testing statistically the significance of interrelations.

The contribution by *G.G. Sil'nickij* ("Correlation and discriminance analysis of languages and language properties", pp. 152-166) does not concern language internal correlations, but rather a multivariate discriminance analysis of typological properties. In 78 areally and genetically different languages, first all correlated features (out of 47 phonetic and grammatical ones) are excluded. The non-correlated features are the starting point of discriminance analysis, yielding no corroboration of the known genetic classification, but five new typological groups. All in all, this is an interesting paper which should be further discussed. It must be remarked that correlation and discriminance analysis as well as factor analysis are merely inductive explorative means not directly yielding theoretical results. Classification itself is a very shaky ground, easily manipulable. It is theoretically prolific if it can be derived from a theory. Hence its direct relevance is not evident. Nevertheless, this direction at least gives rise to many hypotheses which can later on be founded theoretically.

Semantic structures and associated partial domains are explicitly studied in three contributions. *A.A. Vengrenovič* and *V.V. Levickij* ("Quantitative parameters of synonymy in German", pp. 228-231) study the noun in German quantitatively. An analysis of three synonymy dictionaries (containing 64,076 nouns) displays among other things a significant negative correlation between the synonyms of a lemma and the frequency of stylistic markers. No less interesting is the discovery of a relation between the gender category of nouns and the number of synonyms.

The contribution of *L.V. Gikov* and *G.V. Gikova* ("Adverb-verb connections in German", pp. 232-243) studies the frequency of word class combinations on the basis of random

samples from prose texts by German writers. The complete sample contains 3,000 adverb-verb connections. In this corpus, the adverbs are divided into six subclasses (temporal local, modal, causal, conditional and consecutive) and the verbs are classified according to three groups (state, process and activity verbs), consisting of a further 26 subcategories. In this set of data, the authors try to ascertain whether certain adverb-verb combinations, having given semantic-syntactic properties, are significantly frequent. The authors succeed in filtering out about 20 very frequent combinations using the chi-square test, and give relevant information about the strength of adverb-verb combinability.

The investigation of *N.L. L'vova* ("Study of phonosemantic interrelations between consonant clusters at the beginning of words in different texts", pp. 3-10) concerns the question of whether consonantal bigrams at word beginnings have some relation to the functional style of the text. A similar approach can be found in Lvova (2005). The author shows that in English poetic, literary and journalistic texts, the frequency of some consonant clusters possesses a functional stylistic property. This result, tested by means of the chi-square test, should also be tested in larger samples (so far the results hold for 12 poems, 8 literary texts and 10 press texts, where the sample size is never greater than 500 words). In that case it could, perhaps, be corroborated that the frequency of consonant clusters at the beginning of words is not a general language-specific phenomenon but, as a matter of fact, a genre-specific feature. Here, perhaps, more references concerning the study of iconism might be expected.

A.A. Polikarpov dedicates his paper ("Evolutionary foundations of Menzerath's law and the search for the dependence of morpheme length on its positional features", pp. 351-363) to Menzerath's law, which is well established in quantitative linguistics. This article (see also Polikarpov 2006) tries to integrate Menzerath's law into the theory of "life cycle" of words developed by the author. While (traditionally) Menzerath's law describes the relation between the size of linguistic constructs and the length of their constituents, the author tries to supplement it with a positional aspect. He concentrates on the length of suffixes and prefixes in dependence on the distance (position) from their stem morphemes. As a matter of fact, it can be observed that prefixes which are more distant from the stem morpheme tend to be longer, while suffixes get smaller with increasing distance from the "centre". So far the oinly data at our disposal concerns Russian. In a dictionary of Russian word forms segmented morphologically compiled by the author (50,747 items), morpheme length (measured in the number of letters) is considered separately for prefixes, roots and suffixes. Now, if we consider the length of prefixes in dependence on their distance from the stem (i.e. whether a word form contains one, two or three prefixes) we see that with increasing distance the prefixes get longer (an example from the published data: prefixes in the third position to the left of the stem morpheme have a mean length of 2.597 letters, those in the second position 2.249 and those in the first left position 2.08 letters). On the other hand, with increasing distance from the stem morpheme the length of suffixes decreases. This tendency is not as markedly expressed as the dependence of prefix length on its position; a kind of length oscillation can be observed. For the time being there are only data from the morphologically very rich Russian, but they are interesting solely because of their quantitative ratios: in the given corpus, word forms have maximally 3 prefixes and 7 affixes. A comparison with other languages would show that this length tendency is a special feature of Russian or even Indo-European languages; it is not present in strongly agglutinating languages.

In this contribution, relationships between morpheme length and word age, frequency, and semantic class are postulated. These interesting cross-connections represent the core part of the "life cycle" theory which will surely be discussed intensively in (quantitative) linguistics. Above all the integration of the positional dependence of morpheme length into the "traditional" Menzerath's law could be of great interest.

There is only one contribution in Klim (2005) to glottochronology, which is very intensively studied in Russian linguistics, namely that by *L.A. Selezneva-Eleckaja* ("Semantic factors of different susceptibility levels against decay of units in the glottochronological list: using data from Indo-European languages", pp. 322-335), in which the author tries to classify the lexical items of the Swadesh list and their modifications according to semantic fields. The aim of such an investigation is to ascertain whether there is a relation between the semantic content and the degree of survival of lexical units. According to the author, this holds for certain groups (e.g. words expressing feelings), but this can in turn be understood as a hint at the internal non-homogeneity of the word forms in the Swadesh lists. Here we should recall V.V. Levickij's objection that the classification of semantic units e.g. according to the degree of abstractness etc. is notoriously ambiguous (cf. Levickij 2005: 462). In addition to this global problem there is another global flaw in this work: the author operates exclusively with percentages and graphical presentations. Such a procedure in no case meets the methodological expectation of quantitative linguistics which approaches data of this kind with at least a statistical test.

The contribution by *M.V. Usmanova* ("Linguistic and psychological peculiarities of gender: quantitative aspects", pp. 316-321) concerns gender study. The author performs an empirical enquiry (48 test persons) into gender-specific language customs in Russian. On the basis of the evaluation of conversations about certain themes, she could ascertain differences in the use of modal verbs between men and women.

The article of *M.K. Timofeeva* ("Measurement and modelling techniques of the naturalness level of systems with natural language interaction", pp. 336-350) is of a rather theoretical nature. The author shows the possibilities and limits of a quantitative differencing of natural and artificial languages which, according to the author, will grow in importance in the man-machine dialogue.

3.

The third and last block of papers in Klim (2005) address general aspects of computer-based processing of language texts, corpus linguistics and "data mining". Statistical evaluations are rare here but we report on them briefly, in order to round off the depiction of Klim (2005).

Surprisingly there are only two articles in the volume that can be attributed to corpus linguistics. While *I.V. Arzamasceva* ("Statistical investigation of German terminology to the theme 'Fuzzy logic' by means of the program Fuzzy-Base", pp. 244-255) is concerned with the establishment of a data bank (frequency dictionary) of the German lexicon in the domain of fuzzy sets, *A.I. Izotov* ("Quantitative aspects of the description of functional-semantic categories of the imperative in present-day Czech", pp. 122-136) investigates the frequencies of Czech imperative constructions on the basis of the Czech National Corpus. The data are contrasted with those of Russian.

V. G. Klimov ("The linguistic component of the computer based processing of data from natural languages: problems and perspectives of information-communicative technologies in school education", pp. 11-25) dedicates his article rather to general problems of computational linguistics than to the use of quantitative methods. The point is a basic discussion of automatic semantic text analysis. Also the contribution by *A.M. Naletov* ("A computer-based system for the analysis of natural language texts", pp. 218-227) is oriented to basic research and concerns the analysis of semantic networks, a theme that belongs rather to the domain of "small worlds".

An attempt at the automatic recognition of "similar" texts can be found in the article by *E.N. Benderskaja* and *S.V. Žukova* ("Processing of text information by means of chaotic neuronal networks", pp. 271-282) who discuss some concepts of fuzzy logic. Finally, *V.D.*

Gusev and *N.V. Salomatina* ("L-gram analysis of natural language texts and its possibilities", pp. 256-270) present an algorithm for the identification of steady collocations (called here L-grams) and supply the first frequencies. Even if the results are not further evaluated statistically, it is at least an important step toward quantitative analysis on the syntactic level.

4. Summary

Having reported on 31 contributions in Klim (2005), encapsulating a wide and heterogeneous spectrum of applications of quantitative methods in the Russian linguistics and text science, it is possible to detect some general tendencies. From the great number and wide scope of the articles, which as far as content and methodology are concerned display a high level of mathematical-statistical competence, we may draw the conclusion that this omnibus volume can be considered representative[5] of the present state of the art in Russia.

The individual contributions provide an excellent insight into some special domains developed in recent years: worth mentioning are the correlation work from Smolensk, the St Petersburg analyses of authorship attribution and verse analysis, the ideas on the "word life cycle" and an explicit interest in quantitative investigations of semantic structures (synonymy, polysemy, etc.). This is the profile of the Russian quantitative linguistics and text science.

On one hand this profile is a conscious continuation of the experience of Russian quantitative linguistics. Worth mentioning is especially Ju. K. Orlov's et al. hypothesis of "Zipf size" which provides even today a starting point for other innovative investigations. On the other hand, we may recognize a break in tradition. In Klim (2005) there are neither references to the works of Ju.A. Tuldava and Ju.A. Krylov, which are still current, nor to the group "Statistika reči« which dominated the field of statistical investigations in Russia for decades.

The omission of these references may be ascribed to the dynamics or to the sense of a new era in Russian scientific enterprise. At least from the point of view of the application of quantitative methods, this process brings forth a series of new, partially innovative and original approaches. Though innovation and originality is welcome in every scientific discipline, there must be an equal emphasis on the maintenance of certain "standards". "Standards" here are meant in the sense of the "working method of quantitative linguistics" (cf. Altmann 1972: 3ff.; Köhler 2005: 8ff.)[6] which controls the course of linguistic or text-analytical investigations using quantitative methods, and which is not prescriptive but can, nevertheless, be considered a raw and acceptable guideline.

---

[5] Klim (2005) is surely comparable with the 1991 conference in Smolensk called "Ėvrističeskie vozmožnosti kvantitativnych metodov issledovanija jazyka//Heuristic possibilities of quantitative methods in language analysis (cf. Sil'nickij, Tuldava and Polikarpov 1991) and with "2-aja Meždunarodnaja konferencija po kvantitativnoj lingvistike//Second international conference on quantitative linguistics" (cf. Polikarpov 1994) in Moscow, September 1994. The proceedings of these conferences contain only longer abstracts.

[6] In this connection we expressly refer to the recently published handbook of quantitative linguistics (cf. Köhler, Altmann and Piotrowski 2005). This book represents not only the actual state of the arts in quantitative linguistics concerning its theoretical and methodological bases but also a compact source of information allowing to embed one own research in a wider framework. In other words, Klim (2005) displays a partly small readiness to discuss and absorb work from the non-Russian area. This is a phenomenon already criticized by the well-known Russian verse theoretician B.V. Tomaševskij who laid the central foundation stones of the statistical and probabilistic verse analysis and discovered this feature in the Russian verse statistics in the second and third decade of the 20th century. Especially the "domoroščennaja statistika/homemade statistics" was a thorn in his side. Tomaševskij emphasized as an alternative the »philological statistics« which should explicitly take note of and discuss works beyond the Russian scientific area (cf. Tomaševskij 1923: 139).

The formulation of a linguistic/text-analytical hypothesis, the operationalization of the units to be quantified, the translation of the hypothesis in the language of statistics, and above all the (not always easy) choice of an adequate statistical method, should as a rule precede an interpretation. It is not our aim to evaluate the individual contributions according to this criterion. Rather we wish to say that linguistic and text-analytic investigations claiming to work "quantitatively" should not in general be based only on graphical presentations of characteristics or ratios. They should, in agreement with the required intersubjectivity, follow the above-mentioned course of work in order to present a lasting and relevant contribution.

Apart from this partially noticeable methodological weakness, this extensive volume is, nevertheless, a successful and exciting mixture of tradition, innovation and originality. It is to be hoped that this remains characteristic of quantitative procedures in linguistics and text analysis in the future, not only in Russia.

References

Altmann, G. (1972): Status und Ziele der quantitativen Linguistik. In: Jäger, S. (ed.), *Linguistik und Statistik*. Braunschweig: Vieweg, 1-9.

Dshurjuk, T.V.; Levickij, V.V. (2003): "Satztypen und Satzlängen im Funktional- und Autorenstil", in: *Glottometrics* 6, 2003, 40-51.

Grinbaum, O.N. (2000): *Garmonija strofičeskogo ritma v ėstetiko-formal'nom izmerenii.* Sankt Peterburg: Izdatel'stvo Sankt-Peterburgskogo Universiteta.

Gumenjuk, A.; Kostyshin, A.; Borisov, K.; Salnikova, O. (2004): "On the acoustic elements of a poem and the formal procedures of their segmentation", in: *Glottometrics* 8, 42-67.

Gumenyuk, A.; Kostyshin, A.; Simonova, S. (2002): "An approach to the analysis of text structure," in: *Glottometrics 3,* 61-89.

Hřebíček, L. (1997): *Lectures on Text Theory.* Prague: Oriental Institute, Academy of Sciences of the Czech Republic.

Khmelev, D. (2000): "Disputed Authorship Resolution through Using Relative Empirical Entropy for Markov Chains of Letters in Human Language Text", in: *Journal of Quantitative Linguistics*, 7, 3; 201-207.

Köhler, R. (2005): Gegenstand und Arbeitsweise der Quantitativen Linguistik. In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.): *Quantitative Linguistik/Quantitative Linguistics.* Berlin u.a.: de Gruyter, 1-16. [= Handbücher zur Sprach- und Kommunikationswissenschaft, 27].

Krylov, Y.K. (1994): Hurst's law as a Universal Law of Quantitative Linguistics of a coherent text. In: Polikarpov, A.A. (ed.) (1994), 113-114.

Levickij, V. (2005): Polysemie. In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.): *Quantitative Linguistik/Quantitative Linguistics.* Berlin u.a.: de Gruyter, 458-464. [= Handbücher zur Sprach- und Kommunikationswissenschaft, 27]

Lvova, N.L. (2005): "Semantic functions of English initial consonant clusters", in: *Glottometrics* 9, 21-28.

Marusenko, M.A. (1990): *Atribucija anonimnych i psevdoanonimnych literarturnych proizvedenij metodami teorii raspoznavanija obrazov.* Leningrad: Izdatel'stvo Leningradskogo Universiteta.

Polikarpov, A.A. (ed.) (1994a): *Qualico-94. 2nd International Conference of Quantitative Linguistics. September 20-24 1994//2-aja Meždunarodnaja konferencija po kvantitativnoj lingvistike 20-24 sentjabrja 1994 goda.* Moskva: MGU im. M.V. Lomonosova, Filologičeskij fakul'tet.

Polikarpov, A.A. (2006): Towards the Foundations of Menzerath's Law. In: Grzybek, P. (ed.): *Contributions to the Science of Language.* New York u.a.: Springer, 255-272.

Sil'nickij G.G.; Andreev, S.N; Kuz'min, L.A.; Kuskov, M.I. (1990): *Sootnešenie glagol'nych priznakov različnych urovnej v anglijskom jazyke.* Minsk: Navuka i Tėchnika.

Sil'nickij, G.G.; Tuldava, Ju.A.; Polikarpov, A.A. (eds.) (1991): *Ėvrističeskie vozmožnosti kvantitativnych metodov issledovanija jazyka. Tezisy dokladov Vsesojuznogo seminara v gor. Smolensk 11-13 sentjabrja 1991 g.* Smolensk: Smolenskij SGPI.

Tomaševskij, B.V. (1923): "Problema stichotvornogo ritma" in: *Literaturnaja mysl'* 2, 124-140.

Tweedie, F.J. (2005): Statistical models in stylistics and forensic linguistics. In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.): *Quantitative Linguistik/Quantitative Linguistics.* Berlin u.a.: de Gruyter, 387-397. [= Handbücher zur Sprach- und Kommunikationswissenschaft, 27].

Rezension: Gabriel Altmann/Viktor Levickij/Valentina Perebyinis (Hg.): *Problemy kvantytatyvnoji lingvistyky: zbirnyk naukovych pracj* (Probleme der Quantitativen Linguistik: Sammelband). Černivci: Ruta, 2005. 352 Seiten. Von **Juri Kijko.**

Der vorliegende Band mit internationalen Beiträgen präsentiert die neuesten Ergebnisse der Quantitativen Linguistik als einer wissenschaftlichen Disziplin und verfolgt das Ziel, ihren heutigen Stand und ihre Perspektiven darzustellen. Der Sammelband besteht aus fünf Teilen: 1. Quantitative Linguistik als eine wissenschaftliche Disziplin, 2. Zur Anwendung quantitativer Methoden in der Linguistik, 3. Quantitative Untersuchungen der Sprach- und Texteinheiten, 4. Korpuslinguistik, 5. Quantitative Gesetze, Verteilungen und Programme. Der Band vereint 22 Beträge von Sprachwissenschaftlern aus acht Ländern.

Am Anfang des ersten Teils steht als Vorwort ein Beitrag eines der Herausgeber Gabriel Altmann (Deutschland), der seine Überlegungen zu Mode und Wahrheit in der Wissenschaft anstellt und für die Quantitative Linguistik plädiert.

Auf die Ziele und Methoden der Quantitativen Linguistik gehen Reinhard Köhler und Gabriel Altmann (Deutschland) in ihrem Beitrag (S. 12-41) ein. Die Autoren begründen die Einführung von quantitativen Konzepten, Modellen und Methoden in die Linguistik sowie in die Textwissenschaft, wie es in den betreffenden Disziplinen der Fall ist.

Im einem weiteren Beitrag (S. 42-59) von Gabriel Altmann und Peter Meyer (Deutschland) setzen die Autoren Überlegungen über die Quantitative Linguistik und ihre Rolle in der Sprachwissenschaft fort. Sie sind der Meinung, dass „the intervention of physicists in linguistics can help us to open our science for the theory of general systems".

Ramon Ferrer i Cancho (Italien/Spanien) wendet sich (S. 60-75) der Erforschung der Struktur des syntaktischen Netzes zu und entdeckt aufgrund der neuesten Ergebnisse mögliche Wege zum Verstehen der universalen Eigenschaften der menschlichen Sprache.

Karl-Heinz Best (Deutschland) stellt in seinem Beitrag (S. 76-88) fest, „dass die Quantitative Linguistik in den deutschsprachigen Ländern zur Zeit einigermaßen prosperiert, verdankt sie ganz wesentlich den vielfältigen Anregungen der osteuropäischen Forscher und den verbesserten Kontaktmöglichkeiten, die sich seit etwa 1990 entwickelt haben".

Der zweite Teil des Bandes thematisiert die Erfahrungen der Wissenschaftler bei der Anwendung der quantitativen Methoden in der Linguistik.

Valentina Perebeynos (S. 89-99) stellt die Geschichte der sprachstatistischen Forschungen in der Ukraine vor und betont, dass die Anwendung der statistischen Methoden bei linguistischen Untersuchungen in der Ukraine eine lange Traditionen aufweise und zu vielen Errungenschaften geführt habe.

Nataliya Darchuk (S. 100-110) macht den Leser mit der parametrisierten Datenbank der modernen ukrainischen Sprache aufgrund der modernen Häufigkeitswörterbücher bekannt.

Viktor Levickij (S. 111-133) stellt Ergebnisse der quantitativen Studien vor, die am Lehrstuhl für Germanische, Allgemeine und Vergleichende Sprachwissenschaft (Universität Tscherniwzi, Ukraine) auf der phonetischen, morphologischen, lexikalischen und syntaktischen Ebene durchgeführt worden sind. Der Autor plädiert für die Verbindung von quantitativen und qualitativen Ansätzen in den linguistischen Untersuchungen.

Olexandr Oguy als Vertreter der Tschernowizer quantitativen Schule (S.134-148) schließt sich den oben erwähnten Überlegungen an und erörtert Perspektiven der Anwendung von approximativen Methoden in den semasiologischen Studien.

Der dritte Teil des Sammelbandes ist den quantitativen Untersuchungen von Sprach- und Texteinheiten gewidmet.

Viktor Kromer und Anatoliy Polikarpov (Rußland) stellen eine Methode der Zusammenstellung eines zweidimensionalen sprachgenetischen Dendrogramms (S. 149-158) vor, die aufgrund von 10 indogermanischen Sprachen überprüft worden ist.

Peter Grzybek und Emmerich Kelih (Österreich) wenden sich der Frage nach den Graphemhäufigkeiten im Ukrainischen (Teil 1: ohne Apostroph) zu. Ihre Untersuchung (S. 149-179) gehe weit über das „einfache Zählen“ von Buchstaben hinaus und beinhalte weitreichende Perspektiven für Empirie und Theorie.

Mykhaylo Bilynskyi (Ukraine) befasst sich (S. 180-193) mit der semantischen Nähe in der rückläufigen Synonymie bei der deverbalen Rekategorisierung am Beispiel der englischen Sprache. Er stellt fest, dass die aufgrund des Koeffizienten der semantischen Nähe vorgeschlagene Methode eine Modellierung der virtuellen lexikographischen Gruppenbildung ermöglicht.

Emilia Nemcová und Kvetoslava Serdelová (Slowakische Republik) gehen in ihrem Beitrag der Frage der Synonymie in der slowakischen Sprache (S. 194-209) nach. Es werden die Relationen zwischen Synonymie und anderen Eigenschaften wie Polysemie, Worthäufigkeit und Wortlänge untersucht. Die Ergebnisse der Studie unterstützen die synergetische Theorie von R. Köhler.

Svitlana Kijko und Viktor Levickij (Ukraine) präsentieren die Ergebnisse einer komplexen Studie der verbalen Polysemie im Deutschen (S. 210-244): die Distribution der polysemen Verben, den Zusammenhang von Polysemie und Semantik sowie die Abhängigkeit zwischen Polysemie und Häufigkeit.

Sergij Kantemir / Mykola Luchak / Nadiya Lvova (Ukraine) behandeln Möglichkeiten der Wortfeldbildung mithilfe quantitativer Methoden (S. 245-272). Nach der kritischen Analyse der Wortfeldforschung stellen die Autoren verschiedene Feldtypen vor. M. Luchak befasst sich mit dem grammatischen Feld der Kategorie Zeit in verschiedenen Stilen der heutigen englischen Sprache. Sergij Kantemir behandelt das lexikalisch-semantische Feld der Farbenbezeichnung im heutigen Deutsch. N. Lvova untersucht das phonetisch-semantische Feld von 30 Gruppen von Anfangskonsonanten wie *St-, Sp-, Br-, Fr-* u.a. im heutigen Englisch.

Marina Kovtanyuk (Ukraine) vergleicht das semantische Feld der Stärke im Englischen und Französischen mithilfe statistischer Methoden (S. 273-283). Dazu werden Adjektive mit der genannten Semantik auf paradigmatischer und syntagmatischer Ebene analysiert. Die Autorin kommt zu interessanten Schlussfolgerungen im Geiste von Humboldt.

Leonid Hikov (Ukraine) setzt sich zum Ziel (S. 284-291), den Zusammenhang zwischen dem Gebrauch der lexikalisch-semantischen Subklassen und dem Stil des Schriftstellers mit-

hilfe quantitativer Methoden festzustellen. Die Untersuchung wird aufgrund der Werke von sechs deutschen Schriftstellern durchgeführt. Die erhaltenen Ergebnisse sind von Bedeutung für die Stilostatistik.

Yuliya Boyko (Ukraine) versucht in einem weiteren Beitrag (S. 292-305), den Zusammenhang zwischen der Frequenz der Satzgefüge, der Anzahl der Nebensätze im Satzgefüge und dem Autorenstil aufgrund der englischen und amerikanischen Literatur des 20. Jahrhunderts festzustellen. Nach den Ergebnissen der Studie hängen die Frequenz der Satzgefüge und die Anzahl der Nebensätze im Satzgefüge vom Autorenstil ab.

Im vierten Teil des Sammelbandes werden Probleme der Korpuslinguistik, die im engen Kontakt zur Quantitativen Linguistik steht, behandelt. Adam Pawlowski (Polen) stellt Überlegungen zu Perspektiven und Gefahren der Korpuslinguistik (S. 306-322) an. B.D. Jayaram (Indien) stellt Korpora indischer Sprachen vor, die den Forschern frei zur Verfügung stehen (S. 323-329).

Der fünfte Teil des Sammelbandes ist den Fragen der quantitativen Gesetze, Verteilungen und Programme gewidmet. Ján Mačutek (Slowakische Republik) analysiert die Eigenschaften der in der Linguistik oft benutzten Naranan-Balasubrahmanyan-Verteilung im Bezug auf andere Verteilungen (S. 330-334). Fan Fengxiang (China) macht auf die Anwendung des Computerprogramms Visual FoxPro in der Quantitativen Linguistik (S. 335-348) aufmerksam, das sich als ein ergebniswirksames und universales Mittel für die Quantitative Linguistik erwiesen hat.

Insgesamt vermittelt der internationale Sammelband Anregungen zu weiterer Forschung und zeigt, dass die quantitativen Untersuchungen nach wie vor von Bedeutung sind.