Glottometrics 11
2005

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. 11 (2005), Lüdenscheid: RAM-Verlag, 2005. 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 11 (2005) **ISSN 2625-8226**

Contents

Glottometrics 11, 2005, 1-8

Can simple models explain Zipf's law for all exponents?

Ramon Ferrer i Cancho[*]*, Rome*
Vito D. P. Servedio, Rome

Abstract. H. Simon proposed a simple stochastic process for explaining Zipf's law for word frequencies. Here we introduce two similar generalizations of Simon's model that cover the same range of exponents as the standard Simon model. The mathematical approach followed minimizes the amount of mathematical background needed for deriving the exponent, compared to previous approaches to the standard Simon's model. Reviewing what is known from other simple explanations of Zipf's law, we conclude there is no single radically simple explanation covering the whole range of variation of the exponent of Zipf's law in humans. The meaningfulness of Zipf's law for word frequencies remains an open question.

Keywords: Zipf's law, Simon model, intermittent silence

INTRODUCTION

Zipf's law for word frequencies is one of the most striking statistical regularities found in human language. If f is the frequency of a word, the proportion of words having frequency f follows

$$P(f) \sim f^{-\beta}, \tag{1}$$

where $\beta \approx 2$ in normal adult speakers (Zipf, 1932; Zipf, 1949; Ferrer i Cancho, 2005a). β is the exponent of Zipf's law. For simplicity, here we assume Eq. 1 for word frequencies, although other functional forms have been proposed (Chitashvili & Baayen, 1993; Tuldava, 1996; Naranan & Balasubrahmanyan, 1998). H. A. Simon proposed a process constructing a random text for explaining Zipf's law (Simon, 1955; Simon, 1957). At each iteration step, the text grows by one word. The (t+1)-th word will be either a new one (with probability ψ) or an old word (with probability 1-ψ). An old word means a word that has already appeared in the text. The old word is obtained choosing one word occurrence of the existent text sequence at random. All occurrences of words in the sequence have the same probability of being chosen. Equivalently, the old word can be chosen in the following way: the probability of choosing the word i is proportional to f_i, the normalized frequency of word i in the text sequence. The asymptotic distribution of the process follows Eq. 1 with

$$\beta = 1 + \frac{1}{1-\psi}, \tag{2}$$

[*] Address correspondence to Ramon Ferrer i Cancho, Dip. di Fisica, Università 'La Sapienza', Piazzale A. Moro 5, ROMA 00185, ITALY. E-mail: ramon@pil.phys.uniroma1.it.

(Simon, 1955; Simon, 1957; Rapoport, 1982; Manrubia & Zanette, 2002; Zanette & Montemurro, 2005).

Simon's model reproduces Zipf's law with $\beta > 2$. $\beta \approx 2$ is obtained for small values of ψ. Simon's model has been applied to many contexts. Some examples are the scale-free degree distribution of complex networks (Bornhodlt & Ebel, 2001) and the distribution of family names (Zanette & Manrubia, 2001; Manrubia & Zanette, 2002; Manrubia *et al*., 2003).

THE MODELS

Our models are simple generalizations of Simon's and follow essentially the same idea.

> **MODEL A.** At each iteration step, the text grows by m copies of a word ($m > 0$). The m copies will be either from a new one (with probability ψ) or from an old one (with probability $1-\psi$). The old word is obtained choosing one word of the text sequence at random, as in Simon's standard model. The probability of choosing the i-th word is proportional to f_i, the frequency of the word.

As it is formulated, model A means that the text sequence consists of blocks of m copies of the same word. That is not realistic when $m > 1$. A slight variation makes the generalized model more realistic, while giving the same word frequency distribution (see below for the mathematical details about the equivalence):

> **MODEL B**. At each iteration step, the text grows by one word. The $(t+1)$-th word will be either a new one (with probability ψ) or an old word (with probability $1-\psi$). The old word is obtained choosing one member of the text sequence at random. In this case, the probability of choosing the i-th word is proportional to k_i, the weight of the word. Every time a word is chosen, m is added to its weight. New words have zero weight. f_i , the frequency of the i-th word of the text is $f_i = k_i / m$.

When $m = 1$, we have $f_i = k_i$. In that case, Simon's model and model A and B are identical.
The Appendix shows that the frequency distribution of models A and B follows Zipf's law (Eq. 1) and the exponent is given by Eq. 2 (as in the standard Simon model). Interestingly, the exponent does not depend on m ($m \geq 1$). The Appendix provides a calculation of $P(f)$ for Simon's model ($m = 1$) that is more detailed and requires less mathematical background than existing calculations (Simon, 1955; Simon, 1957; Rapoport, 1982; Manrubia & Zanette, 2002; Zanette & Montemurro, 2005).

DISCUSSION

Our extensions of Simon's model account for the same range of exponents as the standard Simon model. Neither the standard Simon model nor our generalizations cover the full interval of real exponents in word frequencies. As far as we know, real exponents lie within the interval [1.6,2.42] in single author text samples (Ferrer i Cancho, 2005a). $1 < \beta < 2$ has been found in some schizophrenics (Whitehorn & Zipf, 1943; Zipf, 1949; Piotrowski et al., 1995). $\beta = 1.6$ was reported by Piotrowski et al. (1995). Young children have been shown to follow $\beta = 1.6$ (Brilluen, 1960; Piotrowski et al., 1995). $\beta = 1.7$ was found in military combat texts

(Piotrowski et al., 1995). The standard Simon model and our extensions exclude the exponents of children and some schizophrenics.

Simon's model is a *birth stochastic process*. A 'birth' means choosing a word that has not yet been used. ψ is the birth rate. Simon's model has been extended to consider also 'deaths' (Manrubia & Zanette, 2002), which here means the possibility that a word occurrence disappears from the text. One may think that the fact that a word occurrence disappears means that the occurrence has been 'forgotten'. If ω is the death rate (i.e. the probability that a 'word' disappears at any iteration step), the extended Simon model by Manrubia and Zanette obeys Zipf's law (Eq. 1) with

$$\beta = 1 + \frac{1-\omega}{1-\omega-\psi}. \tag{3}$$

When words never 'disappear', i.e. $\omega = 0$, Simon's standard model is recovered (Simon, 1955). It is easy to show that we have $\beta < 1$ or $\beta > 2$ for the birth-death model. When $1 - \omega - \psi < 0$, it follows that

$$\frac{1-\omega}{1-\omega-\psi} < 0, \tag{4}$$

which leads to $\beta<1$ using Eq. 3. $\beta<1$ is problematic since $P(f)$ can only be a probability function if time is finite. When $1-\omega-\psi>0$ it follows that $1-\omega>\psi$. So, assuming $\psi>0$ we obtain $1-\omega>1-\omega-\psi$. Dividing by $1-\omega-\psi$ on both sides of the previous equation we obtain

$$\frac{1-\omega}{1-\omega-\psi} > 1, \tag{5}$$

which leads to $\beta > 2$ using Eq. 3. The interval $\beta \in [1,2]$ is covered neither by the standard Simon's model, nor by our extension, nor by Manrubia and Zanette's extension.

While Manrubia & Zanette's extensions were not motivated by Zipf's law for word frequencies, Zanette & Montemurro extended Simon's model to make it more realistic for word frequencies (Zanette & Montemurro, 2005; Montemurro & Zanette, 2002). Recall N_t is the vocabulary size at time t. Their first extension (ZM1) takes into account that vocabulary growth in Simon's model is linear ($N_t \sim \psi t$) while real vocabulary growth is sublinear. ZM1 gets $N_t \sim t^{\nu}$ by replacing the constant rate ψ at which new words are added by a time dependent rate

$$\psi = \psi_0 t^{\nu-1}, \tag{6}$$

where $0 < \nu < 1$ and ψ_0 is the initial rate. ZM1 gives $\beta = 1 + \nu$ (Zanette & Montemurro, 2005). It is easy to see that $1 < \beta < 2$ for ZM1. The second extension (ZM2) tries to give recently introduced words more chance of being used again. After a series of simplifications, the final analytical model extends the standard Simon model with a probability γ. Any time an old word must be added, one word is chosen among all the words in the text regardless of its frequency (all existent words are equally likely) with probability γ and with probability $1-\gamma$, an old word is chosen as in the standard Simon model, i.e. with a probability proportional its frequency of occurrence. ZM2 gives

$$\beta = 1 + \frac{1}{(1-\psi)(1-\gamma)}. \quad (7)$$

Simon's original model is recovered for $\gamma = 0$. Since $\gamma \geq 0$, it is easy to see that

$$\beta \geq 1 + \frac{1}{1-\psi}. \quad (8)$$

Knowing $\psi > 0$ we have $\beta > 2$. The third extension combines the extensions of ZM1 and ZM2 leading to

$$\beta = 1 + \frac{\nu}{1-\gamma}. \quad (9)$$

Knowing that $\gamma \geq 0$, we obtain $\beta > 1 + \nu$. Knowing $\nu > 0$ we get $\beta > 1$.

Simon's model and intermittent silence are among the simplest explanations for Zipf's law in word frequencies. Intermittent silence models consist of concatenating characters from a set including letters (or phonemes) and blanks (or silences). Every time in a sequence a blank is produced, a new word starts (Miller, 1957; Miller & Chomsky, 1963; Mandelbrot, 1966; Li, 1992, Suzuki *et al.* 2005). For simplicity, it is assumed that all letters are equally likely. If σ is the probability of blank (or silence) and L is the number of letters, intermittent silence reproduces Zipf's law with

$$\beta = \frac{1}{1 - \frac{\log(1-\sigma)}{\log L}} + 1. \quad (10)$$

The model in (Li, 1992; Suzuki *et al.*, 2005) is recovered when blanks have the same probability as any of the letters, that is, when $\sigma = 1/(L + 1)$. Assuming $L > 1$ and $\sigma > 0$, $L \to \infty$ and $\sigma \to 0$ give

$$-\infty < \frac{\log(1-\sigma)}{\log L} < 0. \quad (11)$$

Using the bounds in Eq. 11 on Eq. 10 we get $\beta \in (1,2)$. Therefore, intermittent silence accounts for a fraction of the interval of variation of real exponents. Interestingly, the widespread skepticism about the meaning of Zipf's law among scientists is mostly based on intermittent silence (Miller & Chomsky, 1963; Nowak *et al.*, 2000; Wolfram, 2002), which turns out to be incomplete.

The main argument against Zipf's law meaningfulness is that simple models can reproduce the law (Miller & Chomsky, 1963; Rapoport, 1982; Suzuki *et al*. 2005). It is worth to mention that Suzuki *et al.* take a special position, acknowledging the possible meaningfulness of Zipf's law for word frequencies but denying the relevance of Zipf's law for units (e.g. symbols) from unknown sources. Suzuki *et al.*'s main argument is that the presence of Zipf's law is not, in general, a sufficient for communication of any kind. If Zipf's law alone is considered a sufficient condition, false positives may be obtained. The criticisms mentioned above consider a very narrow interval of real exponents and use essentially two simple mod-

els, i.e. intermittent silence and the standard Simon's model. Those models are actually simple but do not cover, indeed, the whole range of real exponents.

We have seen that it is not easy to find a model covering the whole range of exponents. ZM3 is, as far as we know, the only modification of a simple model that covers the whole range. ZM3 is less simple than Simon's original simple model. In fact, arguing that ZM3 is a simple model is problematic, since it incorporates two particular extensions at the same time. Therefore, one can safely conclude that, so far, there is no radically simple explanation of Zipf's law covering the whole range of exponents.

The classic criticisms against Zipf's law meaningfulness need to be reviewed in the light of the range of real exponents. If the key problem against Zipf's law meaningfulness is finding a simple explanation, two different models are actually needed depending on the value of β: intermittent silence when $\beta < 2$ and Simon's model when $\beta > 2$. That is not very elegant since human language could be essentially the same system when considering the variations of β below or above 2 in normal adult speakers (Ferrer i Cancho, 2005a). Whether variations of β below or above 2 are the outcome of essentially the same communication system in normal adult speakers is a matter of current debate (Ferrer i Cancho, 2005a; Ferrer i Cancho, 2005b). We believe that ZM3 will be a crucial model in future discussions about the relevance of Zipf's law. In the absence of a radically simple model for Zipf's law covering the whole interval of real exponents, it seems wiser to give Zipf's law meaningfulness a chance in human language.

ACKNOWLEDGMENTS

This work was supported by the ECAgents project, funded by the Future and Emerging Technologies program (IST-FET) of the European Commission under the EU RD contract IST-1940. The information provided is the sole responsibility of the authors and does not reflect the Commission's opinion. The Commission is not responsible for any use that may be made of the data appearing in this publication.

APPENDIX

We start with the derivation of the word frequency distribution for model A. Thus, we consider an extended Simon where m identical words are added to the text at every time step t ($t > 0$). The identical words added are new with probability ψ or they are a copy of an existent word (which is chosen at random from the text) with probability 1 - ψ. The text sequences has N_0 different words and m_0 word occurrences at the 0-th step ($N_0 \leq m_0$). $P(f)$ can be easily derived with the mean field schema used for the Barabási-Albert scale-free network model (Barabási & Albert, 1999a, Barabási & Albert, 1999b). If k_i, the number of occurrences of the i-th word in Model A, and also t, are treated as continuous variables, then the expected variation of k_i is

$$\frac{dk_i}{dt} = (1-\psi)m\pi_i \,, \tag{12}$$

where

$$\pi_i = \frac{k_i}{\sum_{j=1}^{N_t} k_j} \tag{13}$$

and N_t is the number of different words at time t, π_i can be seen as a continuous rate of change of k_i.

Replacing Eq. 13 with $\sum_{j=1}^{N_t} k_j = m_0 + mt$ into Eq. 12 we obtain

$$\frac{dk_i}{dt} = \frac{(1-\psi)mk_i}{m_0 + mt} . \tag{14}$$

We define t_i as the time at which word i arrived. Integrating Eq. 14 for a word that appeared for the first time at $t = t_i$ with m copies we may write

$$\int_m^{k_i} \frac{dk_i}{k_i} = (1-\psi)m \int_{t_i}^{t} \frac{dt}{m_0 + mt} . \tag{15}$$

The previous Eq. leads to

$$k_i = m\left(\frac{m_0 + mt}{m_0 + mt_i}\right)^{1-\psi} . \tag{16}$$

According to Eq. 16., $\tau(k, t)$, the expected time of arrival of a word with k occurrences when the process is at time t (the time at which a word with k_i occurrences was added for the first time when the process is in the t-th iteration) becomes

$$\tau(k,t) = \frac{1}{m}\left((m_0 + mt)\left(\frac{k}{m}\right)^{-\frac{1}{1-\psi}} - m_0\right) . \tag{17}$$

We define $P(k_i < k)$ as the probability that word i has less than k occurrences and $P(t_i > t)$ as the probability that word i arrives at time t or later. We have that

$$P(k_i < k) = P(t_i > \tau(k,t)) , \tag{18}$$

so

$$P(k_i < k) = 1 - P(t_i \leq \tau(k,t)) . \tag{19}$$

The number of words with $t_i < \tau(k,t)$ is $\psi\tau(k,t)$. Thus, the expected proportion of words with $t_i < \tau(k,t)$ is

$$P(t_i \leq \tau(k,t)) = \frac{\psi\tau(k,t)}{m_0 + mt} , \tag{20}$$

where m_0+mt is the total amount of words at time t.

Replacing Eq. 19 with Eq. 20 into

$$P(k) = \frac{\partial P(k_i < k)}{\partial k} \tag{21}$$

we obtain

$$P(k) = \frac{\psi m^{\frac{1}{1-\psi}-1}}{1-\psi} k^{-1-\frac{1}{1-\psi}}. \tag{22}$$

Interestingly, the distribution does not depend on either t, m_0 or N_0 and the exponent depends only on ψ. For $m = 1$ as in the standard Simon model, Eq. 22 gives

$$P(k) = \frac{\psi}{1-\psi} k^{-1-\frac{1}{1-\psi}}. \tag{23}$$

Hence,

$$P(k) \sim k^{-\beta} \tag{24}$$

with

$$\beta = 1 + \frac{1}{1-\psi} \tag{25}$$

as in the standard Simon process. For deriving the word frequency distribution in model B, we need to take into account that $f = k/m$, where f is the random variable for the frequency of words and k is the random variable for the weight of words. We define $P(f)$ as the probability that a word has frequency f in model B. We have $\widetilde{P}(f) = P(k(f))$ with $k(f) = mf$ so we get

$$\widetilde{P}(f) \sim f^{-\beta} \tag{26}$$

using Eq. 24. Therefore, model B follows Zipf's law with the same exponent as model A.

REFERENCES

Barabási, A.-L. and Albert, R. (1999a). Emergence of scaling in random networks. *Science 286, 509-511.*

Barabási, A.-L., Albert, R., Hawoong, J. (1999b). Mean-field theory for scale-free random networks. *Physica A 272, 173-187.*

Bornholdt, S., Ebel, H. (2001). World wide web scaling exponent from Simon's 1955 model. *Physical Review E 64, 035104 (R).*

Brilluen, L. (1960). *Science and theory of information* (Russian translation). Moscow: Gosudarstvennoe Izdatel´stvo Fiz.-Mat. Literatury.

Chitashvili, R.J., Baayen, R. H. (1993). Word frequency distributions. In: G. Altmann and L. Hřebíček (eds.), *Quantitative text analysis: 54-135.* Trier: Wissenschaftlicher Verlag Trier.

Ferrer i Cancho, R. (2005a). The variation of Zipf's law in human language. *European Physical Journal B 44, 249-257.*

Ferrer i Cancho, R. (2005b). Zipf's law from a communicative phase transition. *Submitted to European Physical Journal B.*

Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution, *IEEE Transactions on Information Theory 38, 1842-1845.*

Mandelbrot, B. (1966). Information theory and psycholinguistics: A theory of word frequencies. In: P. F. Lazarsfield and N. W. Henry (eds.). *Readings in mathematical social sciences: 151-168.* Cambridge: MIT Press.

Manrubia, S.C, Derrida, B., Zanette, D.H. (2003). Genealogy in the era of genomics. *American Scientist 91, 158-165.*

Manrubia, S.C., Zanette, D.H. (2002). At the boundary between biological and cultural evolution: the origin of surname distributions. *Journal of Theoretical Biology 216, 461-477.*

Miller, G.A. (1957). Some effects of intermittent silence. *American Journal of Psychology 70, 311-314.*

Miller, G.A., Chomsky, N. (1963). Finitary models of language users. In Luce, R.D., Bush, R., & Galanter, E. (Eds.), *Handbook of Mathematical Psychology*. Vol 2. New York: Wiley.

Montemurro, M. A., Zanette, D. (2002). New perspectives on Zipf's law in linguistics: from single texts to large corpora. *Glottometrics 4, 87-99.*

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

Nowak, M.A., Plotkin, J.B., Jansen, V.A.A. (2000). The evolution of syntactic communication. *Nature 404, 495-498.*

Piotrowski, R.G., Pashkovskii, V.E., Piotrowski, V.R. (1995). Psychiatric linguistics and automatic text processing. In: *Automatic Documentation and Mathematical Linguistics, 28(5), 28-35.* First published in *Naučno-Techničeskaja Informacija, Serija 2, Vol. 28, No. 11. pp. 21-25, 1994.*

Rapoport, A. (1982). Zipf's law re-visited. *Quantitative Linguistics 16, 1-28.*

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika 42, 425-440.*

Simon, H. A. (1957). *Models of Man*. Chapter 6: On a class of skew distributions functions. New York: John Wiley & Sons.

Suzuki, R., Tyack, P.L., Buch, J. (2005). The use of Zipf's law in animal communication analysis. *Animal Behavior 69, F9-F17.*

Tuldava, J. (1996). The frequency spectrum of text and vocabulary. *Journal of Quantitative Linguistics 3, 38-50.*

Whitehorn, J.C., Zipf, G.K. (1943). Schizophrenic language. *Archive of Neurology and Psychiatry 49, 831-851.*

Wolfram, S. (2002). *A new kind of science*. Champaign: Wolfram Media.

Zanette, D.H., Manrubia S.C. (2001). Vertical transmission of culture and the distribution of family names. *Physica A 295, 1-8.*

Zanette, D.H., Montemurro, M.A. (2005). Dynamics of text generation with realistic Zipf distribution. *Journal of Quantitative Linguistics. In press.*

Zipf, G.K. (1932). *Selected studies of the principle of relative frequency in language.* Cambridge, MA: Harvard University Press.

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

Glottometrics 11, 2005, 9-31

Zur Häufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten

Karl-Heinz Best, Göttingen[1]

Abstract. This paper deals with frequency distributions of letters in German. As in other natural languages, letters of the alphabet do not appear equally often. On the contrary, each letter has its own characteristic frequency. Moreover, the frequencies of the letters differ from text to text. The purpose of this paper is to present some further frequency distributions of letters in German texts and text corpora and to show that they all follow the negative hypergeometric distribution. In some well-chosen cases, spaces and punctuation marks are considered, too**.**

Keywords: letter frequency, German

1. Ziele der Untersuchung

Die vorliegende Arbeit behandelt die Häufigkeit von Buchstaben bzw. von Schriftzeichen generell (also einschließlich der Interpunktions-, Leer- und sonstigen Schriftzeichen) in deutschen Texten und Textkorpora. Zwei Aspekte sollen dabei berücksichtigt werden:

Mehrere Anfragen zeigen, dass es ein Interesse daran gibt, mit welchen Häufigkeiten Buchstaben im Deutschen auftreten. Es scheint nur relativ wenige Daten dazu zu geben, die z.T. recht alt sind und daher die Frage provozieren, ob sie für die heutigen Verhältnisse noch repräsentativ sind (Hoffmann ²1985: 83, mit Angaben auch zur Häufigkeit aller anderen Schriftzeichen und der Lücken; Meier [2]1967: 334, Graphik dazu: Müller 2003: 209; Nasvytis 1953: 80; Schönpflug 1969 unter Berücksichtigung der Position der Buchstaben im Wort; für Fachsprachen: Hoffmann ²1985: 90, ebenfalls mit Angaben auch zur Häufigkeit aller anderen Schriftzeichen und der Lücken). Recht aktuelle und umfangreiche Daten findet man in der Kryptologie (Bauer 2000; Beutelspacher [4]1994: 18): Bei Bauer ([3]2000: 303-304) ist die prozentuale Verteilung von 681972 Zeichen in Artikeln der Süddeutschen Zeitung vom März 1992 erfasst. Andere Erhebungen gelten der Häufigkeit von Anfangsbuchstaben im Lexikon (Finkenstaedt & Wolff 1973: 42; Mater 1966: Anfangskapitel, ohne Seitenzählung; Muthmann 1996; Schmidt 1985: 92, 94; 1986: 100-101) oder auch dem Druckraum („Alphabethstrecken"), den diese in 13 Wörterbüchern verschiedener Art einnehmen (Schmidt 1985: 91; Schmidt 1986: 101; Haß-Zumkehr 2001: 383f.), bzw. der Häufigkeit von Endbuchstaben (Muthmann 1988: 65). Soweit es sich um Daten aus Texten handelt, sind − meist heterogene − Textkorpora, aber keine Einzeltexte ausgewertet worden. Es gibt auch Zusammenstellungen, in denen die Umlaute und/ oder <ß> fehlen (Küpfmüller 1954: 267; Zemanek 1959: 34, 51). In vielen Fällen werden nur die relativen Anteile der Schriftzeichen genannt, mehrfach sogar ohne Nennung der Grundgesamtheit, wodurch solche Angaben genau genommen unbrauchbar sind. Die relativen Anteile der Schriftzeichen unterscheiden sich z.T. erheblich. Ordnet man sie in einer Rangliste nach ihrer Häufigkeit, so stehen im Deutschen anscheinend

[1] Address correspondence to: K.-H. Best, e-mail: kbest@gwdg.de

<e> und <n> in dieser Reihenfolge immer am Anfang; schon vom dritten Buchstaben an ergeben sich aber Unterschiede: meist handelt es sich dabei um <i>, aber auch <r> kann als dritthäufigster vorkommen (Zemanek 1959: 34). Erhebungen zu Einzeltexten und weiteren Textkorpora sollen in dieser Arbeit die Datenbasis erweitern.

Ein zweiter Aspekt der Untersuchung besteht darin, der Frage nachzugehen, ob es ein theoretisch begründetes Modell gibt, das geeignet ist, die empirischen Daten von Rang-Häufigkeitsverteilungen der Buchstaben und anderer Schriftzeichen angemessen zu repräsentieren. Solche Versuche sind nicht neu; es sei etwa auf Belevitch (1956) verwiesen, der Rang-Frequenz-Untersuchungen an Wörtern, Phonemen und Buchstaben durchführt. Ein neues Beispiel findet sich in Best (²2003: 79) für einen kurzen literarischen Text aus Lichtenbergs Sudelbüchern. Alle Dateien in dieser Arbeit werden daraufhin geprüft, ob sie ebenfalls einem solchen Modell folgen.

2. Modelle von Rang-Frequenz-Verteilungen der Buchstaben und anderen Schriftzeichen

Bei Rang-Frequenz-Verteilungen spielen im Deutschen besonders drei Modelle eine hervorragende Rolle: die geometrische Verteilung (Altmann & Lehfeldt 1980: 144), die Zipf-Mandelbrot-Verteilung (Zörnig & Altmann 1995) und die negative hypergeometrische Verteilung (Altmann 1988: 69ff.; Best ²2003: 78-85; Knüppel 2001). In dieser Arbeit wird gezeigt, dass die negative hypergeometrische Verteilung − anders als die beiden anderen − für die Rang-Frequenz-Verteilung von Schriftzeichen generell bzw. nur von Buchstaben im Deutschen ein sehr gutes Modell darstellt, deren Formel in 1-verschobener Form

$$P_x = \frac{\binom{-M}{x-1}\binom{-K+M}{n-x+1}}{\binom{-K}{n}}, \qquad x = 1, 2, ..., \; n+1 \tag{1}$$

lautet. Das Modell hat den Vorteil, dass es sich aus der allgemeinen Theorie von Wimmer und Altmann (2005) einfach durch Reparametrisierung gewinnen lässt. Setzt man in der Rekursionsformel

$$P_x = \left(1 + a_0 + \frac{a_1}{x + b_1} + \frac{a_2}{x + b_2}\right) P_{x-1}, \tag{2}$$

die den Ausgangspunkt ihres theoretischen Ansatzes bildet und zu zahlreichen Sprachgesetzen führt, $a_0 = b_2 = 0$, $a_1 = K\text{-}2 - (M\text{-}1)(n\text{-}1)/(K\text{-}M\text{+}n)$, $a_2 = (n + 1)(M - 1)/(K\text{-}M\text{+}n)$, $b_1 = K\text{–}M\text{+}n$ dann bekommt man die Rekursionsformel der negativen hypergeometrischen Verteilung mit Lösung (1), die hier mit Verschiebung um einen Schritt nach rechts präsentiert wurde.

Dieses Modell soll nun mit Hilfe des Altmann-Fitters (1997) am Beispiel einer Reihe von literarischen und sonstigen Texten und Korpora erprobt werden. Zur Veranschaulichung werden die Testergebnisse meist auch in Form von Graphiken dargestellt.

Erläuterungen zu den folgenden Tabellen:

n_x: beobachtete Anzahl der Schriftzeichen im jeweiligen Text (absolute Häufigkeiten, ohne Rücksicht auf Groß- und Kleinschreibung);
NP_x: Anzahl der Schriftzeichen nach der negativen hypergeometrischen Verteilung;
K, M, n: Parameter der Verteilung;
X^2: Werte des Chiquadrates;
FG: Freiheitsgrade;
P: Überschreitungswahrscheinlichkeit für das entsprechende Chiquadrat;
C: Diskrepanzkoeffizient $C = X^2/N$.

Die übrigen Angaben verstehen sich wohl von selbst. Die Anpassungen der negativen hypergeometrischen Verteilung an die Textdateien werden als erfolgreich betrachtet, wenn $P \geq 0.05$ bzw. $C \leq 0.01$. Diese Bedingungen sind bei allen Texten erfüllt, wenn nur die Buchstaben erhoben wurden. Bei den Dateien, die alle Schriftzeichen erfassen, erhält man mit $0.01 < C \leq 0.02$ mehrfach ein Ergebnis, das die genannte Bedingung verfehlt, aber doch immer in einem noch tolerablen Bereich $0.01 < C \leq 0.02$ liegt. Bei umfangreichen Dateien wie den meisten hier vorgelegten ist der Diskrepanzkoeffizient *C* das zu bevorzugende Testkriterium, auch wenn seine Stichprobenverteilung nicht gegeben ist; nur bei der Fabel Pestalozzis als einer Datei mit nur wenigen Buchstaben (675) ist *P* geeignet.

Die Ergebnisse der Untersuchungen zu Texten und Textkorpora im Deutschen sind in den folgenden Tabellen und Graphiken zusammengestellt.

3. Verteilungen von Buchstaben und sonstigen Schriftzeichen in deutschen Texten

Tabelle 1
Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Johann Heinrich Pestalozzi, *Hühner, Adler und Mäuse* (675 Buchstaben)

Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	e	18.96	128	116.87	15	g	2.07	14	14.63
2	n	10.37	70	76.15	16	o	2.07	14	12.90
3	i	7.26	49	60.92	17	k	1.04	7	11.29
4	r	6.37	43	51.67	18	w	1.04	7	9.79
5	s	6.07	41	45.01	19	f	0.89	6	8.38
6	l	6.07	41	39.78	20	z	0.89	6	7.07
7	d	5.78	39	35.46	21	ü	0.89	6	5.86
8	t	5.33	36	31.76	22	v	0.89	6	4.73
9	h	5.19	35	28.52	23	ä	0.59	4	3.68
10	u	4.89	33	25.64	24	j	0.30	2	2.73
11	a	4.59	31	23.04	25	p	0.15	1	1.87
12	m	2.96	20	20.67	26	ö	0.15	1	1.11
13	c	2.81	19	18.50	27	ß	0.15	1	0.47
14	b	2.22	15	16.49					
$K = 3.0071$, $M = 0.6847$, $n = 26$, $X^2 = 17.63$, $FG = 22$, $P = 0.73$									

(Der Text wurde ohne Verfasser und Überschrift ausgewertet; er findet sich in Johann Heinrich Pestalozzi. 1992. *Fabeln*. Ausgewählt und mit einem Nachwort von Heinz Weder. Zürich: Manesse. S. 48. Die Buchstaben <q, x, y> kommen in diesem Text nicht vor. Das Testergebnis ist mit $P = 0.73$ sehr gut.)

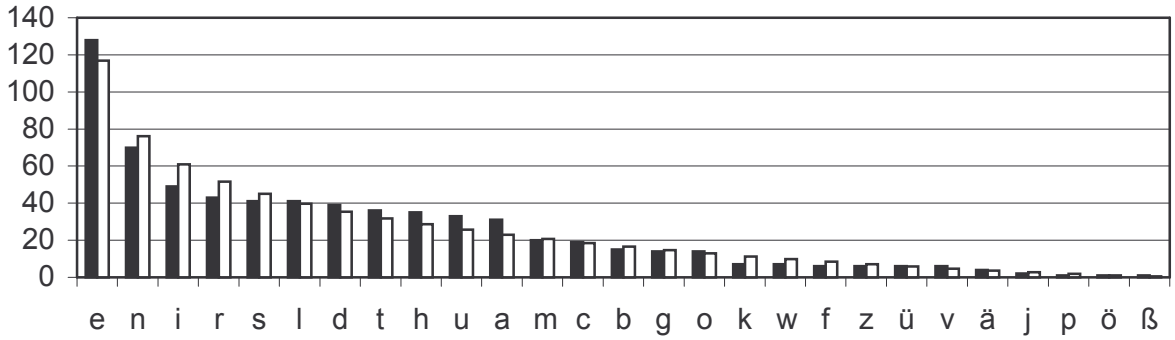


Abbildung 1. Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Pestalozzi, *Hühner, Adler und Mäuse*

Tabelle 2
Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Gottfried August Bürger, *Münchhausen* (137476 Buchstaben)

Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	e	17.26	23729	21439.76	16	w	1.91	2624	2703.45
2	n	10.68	14684	14970.98	17	f	1.74	2397	2375.70
3	i	8.03	11043	12280.74	18	b	1.56	2140	2072.73
4	r	6.74	9270	10556.97	19	k	1.25	1717	1792.94
5	h	5.64	7759	9271.00	20	z	1.19	1631	1535.07
6	s	5.62	7723	8235.67	21	v	0.77	1059	1298.09
7	a	5.54	7622	7364.02	22	ü	0.64	885	1081.18
8	t	5.13	7052	6608.79	23	ß	0.53	725	883.73
9	d	4.65	6389	5941.48	24	ä	0.51	705	705.31
10	u	3.97	5463	5343.65	25	p	0.50	692	545.66
11	l	3.70	5080	4802.65	26	ö	0.26	362	404.70
12	c	3.65	5014	4309.47	27	j	0.13	175	282.58
13	g	3.12	4286	3857.46	28	y	0.02	26	179.73
14	m	2.92	4010	3441.57	29	x	0.02	25	97.00
15	o	2.31	3172	3057.92	30	q	0.01	17	35.99
$K = 3.4096$, $M = 0.7385$, $n = 29$, $X^2 = 1311.63$, $C = 0.0095$									

(Textgrundlage: elektronische Textversion. Damit wird hier und später darauf verwiesen, dass der Text entweder aus dem Internet oder von einer CD ROM genommen wurde. In solchen Fällen ist nicht immer klar, in welcher Version der betreffende Text vorliegt. Der Text wurde vollständig, einschließlich Titelseite und Fußnoten ausgewertet; er unterscheidet sich von der Erstausgabe durch einige zusätzliche Zwischentitel.)

Zwecks Anpassung der negativen hypergeometrischen Verteilung wurden die letzten sieben Buchstaben zu einer Klasse zusammengefasst.

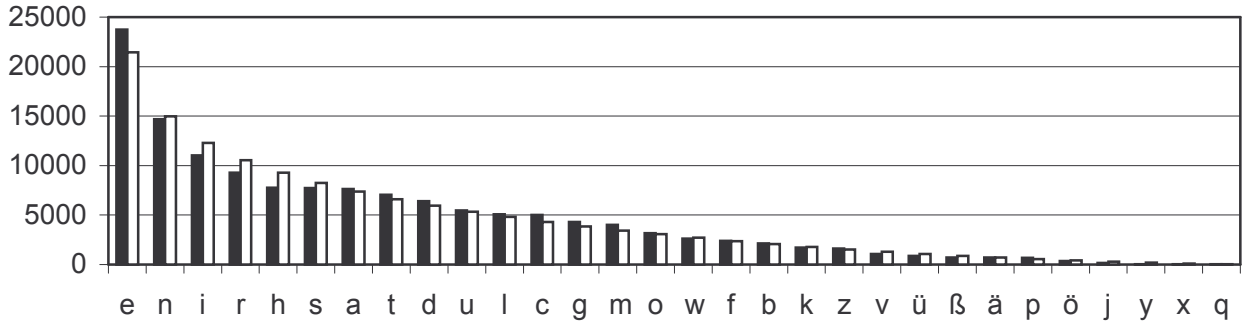


Abbildung 2. Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Gottfried August Bürger, *Münchhausen*

Es folgt die Ballade *Lenore* von G.A. Bürger, die doppelt ausgewertet wurde: Einmal nur hinsichtlich ihrer Buchstaben, dann aber unter Berücksichtigung aller Schriftzeichen einschließlich der Leerzeichen. Solche Erhebungen hat auch schon Hoffmann (²1985: 83; 90) vorgestellt.

Tabelle 3
Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Gottfried August Bürger, *Lenore* (6215 Buchstaben)

Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	e	13.64	848	809.16	15	m	3.06	190	156.50
2	n	10.20	634	623.15	16	b	2.03	126	138.62
3	i	7.85	488	534.77	17	w	1.91	119	121.73
4	r	7.14	444	473.87	18	f	1.59	99	105.80
5	h	6.36	395	425.95	19	k	1.50	93	90.79
6	t	6.31	392	385.69	20	p	1.19	74	76.68
7	s	5.23	325	350.54	21	z	0.87	54	63.47
8	a	5.05	314	319.09	22	ü	0.50	31	51.15
9	d	5.05	314	319.09	23	v	0.50	31	51.15
10	u	4.59	285	264.17	24	ß	0.42	26	29.33
11	l	4.23	263	239.75	25	ä	0.39	24	19.94
12	g	3.44	214	216.95	26	ö	0.16	10	11.72
13	o	3.43	213	195.57	27	j	0.10	6	4.89
14	c	3.27	203	175.46					
$K = 3.1886$, $M = 0.8109$, $n = 26$, $X^2 = 45.992$, $C = 0.0074$									

(Textgrundlage: elektronische Textversion. Der Text wurde vollständig, einschließlich Titel ausgewertet. Die Buchstaben <q>, <x> und <y> kommen im Text nicht vor.)

Tabelle 4
Anpassung der negativen hypergeometrischen Verteilung an die Schriftzeichen in Gottfried August Bürger, *Lenore* (insgesamt 7962 Schriftzeichen einschließlich Leerzeichen)

Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	LZ	15.10	1202	1244.95	20	!	1.47	117	118.00
2	e	10.65	848	787.04	21	f	1.24	99	106.94
3	n	7.96	634	626.52	22	k	1.17	93	96.54
4	i	6.13	488	532.48	23	p	0.93	74	86.73
5	r	5.58	444	466.57	24	GST	0.92	73	77.49
6	h	4.96	395	415.93	25	„	0.82	65	68.78
7	t	4.92	392	374.82	26	z	0.68	54	60.59
8	s	4.08	325	340.19	27	.	0.68	54	52.89
9	a	3.94	314	310.26	28	AP	0.45	36	45.66
10	d	3.94	314	283.89	29	ü	0.39	31	38.90
11	u	3.58	285	260.32	30	v	0.39	31	32.59
12	l	3.30	263	239.01	31	ß	0.33	26	26.73
13	g	2.69	214	219.57	32	?	0.33	26	21.33
14	o	2.68	213	201.72	33	ä	0.30	24	16.39
15	c	2.55	203	185.22	34	;	0.18	14	11.94
16	m	2.39	190	169.90	35	ö	0.13	10	7.99
17	,	1.93	154	155.62	36	j	0.08	6	4.61
18	b	1.58	126	142.27	37	:	0.08	6	1.87
19	w	1.49	119	129.75					
$K = 3.0934$, $M = 0.6574$, $n = 36$, $X^2 = 51.709$, $C = 0.0065$									

(Der gesamte Text wurde einschließlich Überschrift, aber ohne Autor ausgewertet. LZ: Leerzeichen; GST: Gedankenstrich; AP: Apostroph.)

Tabelle 5
Anpassung der negativen hypergeometrischen Verteilung
an die Buchstaben in Georg Büchner, *Lenz* (42608 Buchstaben)

Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	e	17.05	7263	6814.29	16	o	1.88	801	823.49
2	n	10.69	4555	4695.07	17	b	1.84	783	722.83
3	i	8.21	3496	3829.01	18	f	1.61	686	629.92
4	r	6.75	2874	3279.04	19	z	1.19	508	544.27
5	s	6.03	2571	2871.34	20	k	1.07	457	465.44
6	a	5.82	2480	2544.72	21	ü	0.61	259	393.11
7	t	5.60	2386	2270.84	22	ä	0.60	256	327.02
8	h	5.38	2293	2034.35	23	v	0.52	223	266.95
9	d	4.75	2023	1826.02	24	p	0.49	210	212.76
10	l	4.07	1736	1639.88	25	ß	0.39	166	164.35
11	u	3.56	1516	1471.83	26	ö	0.33	141	121.70
12	c	3.13	1333	1318.97	27	j	0.16	69	84.82
13	g	3.12	1330	1179.16	28	y	0.03	14	53.83
14	m	3.04	1296	1050.77	29	q	0.03	12	28.98
15	w	2.03	867	932.54	30	x	0.01	4	10.71
$K = 3.4083$, $M = 0.7289$, $n = 29$, $X^2 = 425.28$, $C = 0.0100$									

(Textgrundlage: elektronische Textversion. Der Text wurde vollständig mit Autor und Titel ausgewertet.)

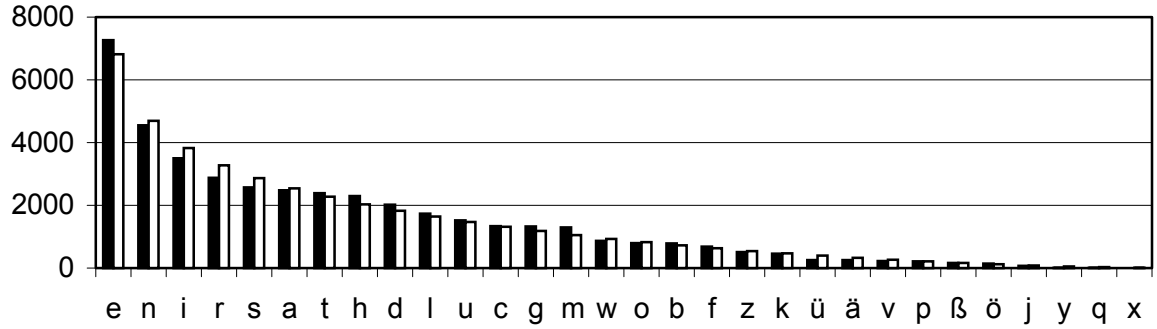


Abbildung zu Tab. 5. Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Georg Büchner, *Lenz*

Die folgende Tabelle enthält wiederum alle Buchstaben, aber auch alle sonstigen Schriftzeichen einschließlich der Leerzeichen:

Tabelle 6
Anpassung der negativen hypergeometrischen Verteilung an die Schriftzeichen in Georg Büchner, *Lenz* (insgesamt 53443 Zeichen einschließlich der Leerzeichen)

Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	LZ	16.91	9039	9218.26	25	ä	0.48	256	297.56
2	e	13.59	7263	6025.47	26	v	0.42	223	256.69
3	n	8.52	4555	4779.36	27	p	0.39	210	220.20
4	i	6.54	3496	4008.73	28	;	0.34	182	187.74
5	r	5.38	2874	3450.64	29	ß	0.31	166	158.98
6	s	4.81	2571	3013.62	30	ö	0.26	141	133.61
7	a	4.64	2480	2655.45	31	j	0.13	69	111.33
8	t	4.46	2386	2353.25	32	„	0.12	64	91.90
9	h	4.29	2293	2093.18	33	:	0.11	60	75.04
10	d	3.79	2023	1866.22	34	GST	0.10	51	60.54
11	l	3.25	1736	1666.13	35	!	0.07	37	48.16
12	u	2.84	1516	1488.40	36	?	0.04	21	37.70
13	c	2.49	1333	1329.65	37	y	0.03	14	28.97
14	g	2.49	1330	1187.26	38	q	0.02	12	21.77
15	m	2.43	1296	1059.15	39	AP	0.02	11	15.94
16	,	1.76	943	943.65	40	BST	0.02	9	11.30
17	w	1.62	867	839.36	41	ZST	0.01	6	7.70
18	o	1.50	801	745.11	42	x	0.01	4	4.99
19	b	1.47	783	659.92	43	0	0.00	1	3.02
20	f	1.28	686	582.93	44	2	0.00	1	1.67
21	z	0.95	508	513.38	45	3	0.00	1	0.80
22	k	0.86	457	450.63	46	4	0.00	1	0.31
23	.	0.76	407	394.08	47	8	0.00	1	0.07
24	ü	0.48	259	343.22					
$K = 4.8953$, $M = 0.6991$, $n = 46$, $X^2 = 750.30$, $C = 0.0140$									

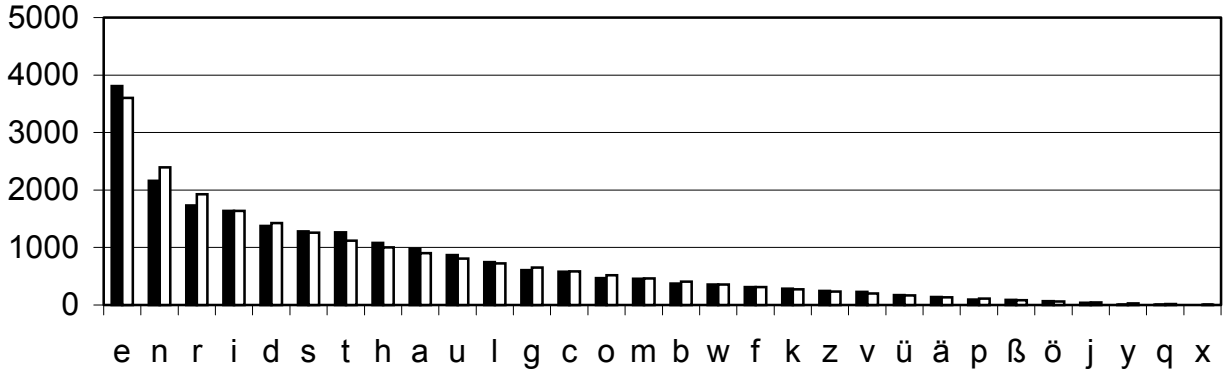
(Der Text wurde vollständig mit Autor und Titel ausgewertet. LZ: Leerzeichen; BST: Binde- und Auslassungsstrich; AP: Apostroph; GST: Gedankenstrich; ZST: Zitierstrich für Titel von Werken. Der Text enthält keinen Trennungsstrich.)

Tabelle 7
Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Georg Büchner, *Hessischer Landbote* (21452 Buchstaben)

Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	e	17.76	3810	3602.92	16	b	1.73	372	408.25
2	n	10.07	2161	2394.38	17	w	1.67	358	359.20
3	r	8.07	1731	1926.01	18	f	1.45	310	313.89
4	i	7.63	1637	1636.52	19	k	1.32	283	272.07
5	d	6.42	1378	1425.85	20	z	1.15	246	233.51
6	s	5.98	1283	1259.39	21	v	1.06	227	198.04
7	t	5.89	1264	1121.29	22	ü	0.81	173	165.52
8	h	5.03	1080	1003.05	23	ä	0.65	139	135.85
9	a	4.57	981	899.59	24	p	0.43	92	108.95
10	u	4.05	868	807.64	25	ß	0.42	90	84.77
11	l	3.49	748	724.99	26	ö	0.32	68	63.31
12	g	2.82	605	650.06	27	j	0.18	38	44.58
13	c	2.69	577	581.70	28	y	0.03	7	28.66
14	o	2.18	468	519.03	29	q	0.01	3	15.69
15	m	2.12	455	461.39	30	x	0.00	0	5.94
$K = 3.3167$, $M = 0.7016$, $n = 29$, $X^2 = 147.25$, $C = 0.0069$									

(Textgrundlage: elektronische Textversion. Der Text wurde vollständig, einschließlich Verfasser, Titel und Zwischentitel ausgewertet. Die Zahlen wurden nicht berücksichtigt.)

Zwecks Anpassung der negativen hypergeometrischen Verteilung wurden die letzten vier Buchstaben zu einer Klasse zusammengefasst.



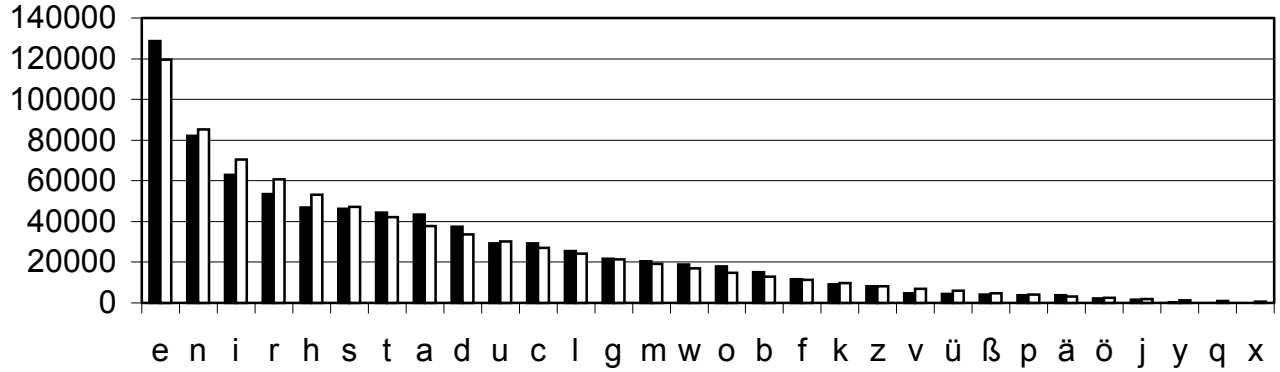
Graphik zu Tab. 7. Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Georg Büchner, *Hessischer Landbote*

Tabelle 8
Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Karl May, *Winnetou I* (777368 Buchstaben)

Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	e	16.56	128724	119449.27	17	b	1.93	15041	12980.96
2	n	10.55	82034	85379.17	18	f	1.49	11575	11289.51
3	i	8.09	62891	70467.08	19	k	1.17	9103	9744.18
4	r	6.88	53495	60659.56	20	z	1.04	8074	8336.00
5	h	6.05	47011	53224.13	21	v	0.62	4798	7057.29
6	s	5.94	46214	47176.88	22	ü	0.59	4553	5901.37
7	t	5.70	44285	42054.77	23	ß	0.51	3953	4862.40
8	a	5.59	43427	37603.28	24	p	0.49	3834	3935.23
9	d	4.82	37492	33667.40	25	ä	0.49	3777	3115.31
10	u	3.78	29406	30145.57	26	ö	0.27	2132	2398.55
11	c	3.75	29172	26967.50	27	j	0.21	1598	1781.29
12	l	3.29	25539	24082.39	28	y	0.03	226	1260.20
13	g	2.80	21729	21452.27	29	q	0.01	106	832.28
14	m	2.62	20390	19047.92	30	x	0.00	34	494.77
15	w	2.42	18837	16846.35	31	é	0.00	6	245.11
16	o	2.30	17911	14829.06	32	ñ	0.00	1	80.94
$K = 3.7659$, $M = 0.7610$, $n = 31$, $X^2 = 9973.40$, $C = 0.0128$									

(Textgrundlage: elektronische Textversion. Der Text wurde vollständig mit Titel ausgewertet. Angegebene Druckfehler wurden beseitigt. <Ae>, <Oe> und <Ue> am Wortanfang wurden durch <Ä>, <Ö> und <Ü> ersetzt. Bildunterschriften wurden entfernt.)

Das Graphem <I> für die römische Zahl I, das im Titel vorkommt, ist bei den Berechnungen nicht berücksichtigt worden.



Graphik zu Tab. 8: Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Karl May, *Winnetou I*

(In der Graphik wurden die Buchstaben <é> und <ñ> nicht berücksichtigt.)

Tabelle 9
Anpassung der negativen hypergeometrischen Verteilung an die Schriftzeichen in Karl May, Winnetou I (insgesamt 974506 Zeichen einschließlich der Leerzeichen)

Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	LZ	16.33	159093	162931.14	25	v	0.49	4798	5834.30
2	e	13.21	128724	107424.01	26	ü	0.47	4553	5059.61
3	n	8.42	82034	85648.76	27	ß	0.41	3953	4364.38
4	i	6.45	62891	72148.04	28	p	0.39	3834	3742.45
5	r	5.49	53495	62348.61	29	ä	0.39	3777	3188.10
6	h	4.82	47011	54656.92	30	ö	0.22	2132	2696.04
7	s	4.74	46214	48337.67	31	!	0.20	1969	2261.35
8	t	4.54	44285	42991.93	32	?	0.19	1815	1879.41
9	a	4.46	43427	38378.86	33	j	0.16	1598	1545.86
10	d	3.85	37492	34341.51	34	;	0.15	1478	1256.64
11	u	3.02	29406	30771.52	35	:	0.06	548	1007.86
12	c	2.99	29172	27590.63	36	BST	0.05	441	795.88
13	l	2.62	25539	24740.20	37	y	0.02	226	617.21
14	g	2.23	21729	22174.97	38	AST	0.02	188	468.54
15	m	2.09	20390	19859.07	39	q	0.01	106	346.71
16	w	1.93	18837	17763.51	40	AP	0.01	54	248.71
17	o	1.84	17911	15864.35	41	x	0.00	34	171.65
18	b	1.54	15041	14141.54	42	KL	0.00	30	112.76
19	,	1.49	14482	12578.03	43	*	0.00	28	69.39
20	f	1.19	11575	11159.13	44	é	0.00	6	39.00
21	k	0.93	9103	9872.04	45	1	0.00	1	19.16
22	.	0.88	8577	8705.51	46	8	0.00	1	7.51
23	„	0.87	8434	7649.54	47	ñ	0.00	1	1.83
24	z	0.83	8074	6695.16	48	I[1]	-	1	-
$K = 4.7707$, $M = 0.7033$, $n = 46$, $X^2 = 12316.23$, $C = 0.0126$									

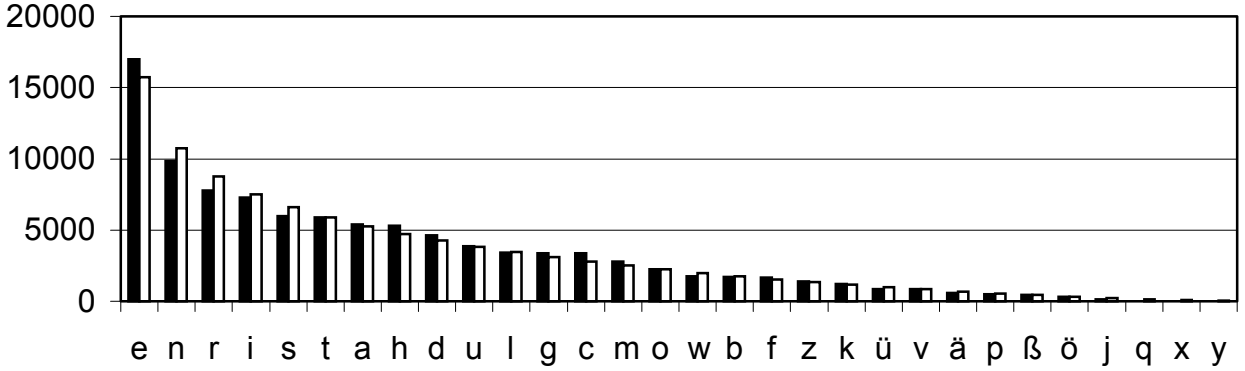
(Der Text wurde vollständig mit Titel ausgewertet. LZ: Leerzeichen; BST: Binde- und Auslassungsstrich; AP: Apostroph; AST: Anführungsstriche zur Kennzeichnung von Begriffen, Wörtern; GST: Gedankenstrich; KL: rechte und linke Klammer; ZST: Zitierstrich für Titel von Werken.)

[1] Das Graphem <I> für die römische Zahl I, das im Titel vorkommt, ist bei den Berechnungen nicht berücksichtigt worden.

Tabelle 10
Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Franz Kafka, *Die Verwandlung* (99559 Buchstaben)

Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	e	17.05	16976	15711.54	16	w	1.77	1765	1990.78
2	n	9.88	9837	10760.37	17	b	1.73	1718	1759.39
3	r	7.82	7783	8773.25	18	f	1.68	1669	1544.57
4	i	7.31	7275	7523.51	19	z	1.41	1401	1345.17
5	s	5.99	5963	6602.70	20	k	1.22	1218	1160.31
6	t	5.91	5888	5867.87	21	ü	0.85	844	989.27
7	a	5.42	5392	5253.12	22	v	0.84	833	831.50

8	h	5.31	5288	4722.85	23	ä	0.60	594	686.60
9	d	4.66	4642	4255.75	24	p	0.49	492	554.31
10	u	3.86	3847	3838.04	25	ß	0.47	463	434.49
11	l	3.42	3406	3460.37	26	ö	0.33	332	327.19
12	g	3.39	3371	3116.07	27	j	0.15	147	232.63
13	c	3.37	3356	2800.25	28	q	0.01	10	151.28
14	m	2.82	2808	2509.22	29	x	0.00	2	84.00
15	o	2.25	2239	2240.14	30	y	0.00	0	32.43
$K = 3.2877$, $M = 0.7218$, $n = 29$, $X^2 = 954.41$, $C = 0.0096$									



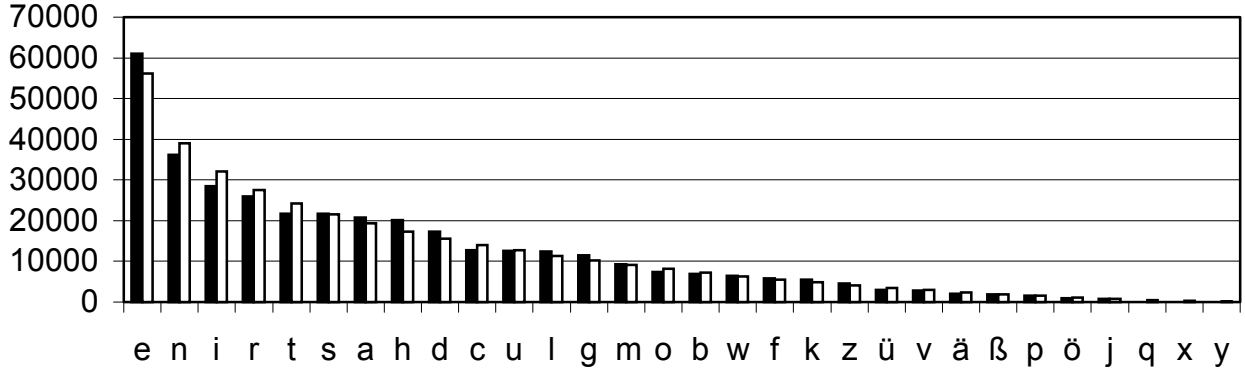
Graphik zu Tab. 10. Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Franz Kafka, *Die Verwandlung*

Tabelle 11
Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Franz Kafka, *Der Prozeß* (361848 Buchstaben)

Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	e	16.86	60999	56145.32	16	b	1.92	6931	7220.13
2	n	10.01	36239	39081.95	17	w	1.77	6396	6363.94
3	i	7.86	28436	32048.49	18	f	1.62	5878	5570.39
4	r	7.15	25883	27563.11	19	k	1.54	5583	4835.42
5	t	6.01	21760	24226.99	20	z	1.26	4551	4155.78
6	s	6.00	21699	21546.29	21	ü	0.82	2982	3528.88
7	a	5.76	20827	19292.03	22	v	0.79	2862	2952.71
8	h	5.56	20101	17339.97	23	ä	0.59	2120	2425.79
9	d	4.77	17249	15615.38	24	ß	0.52	1896	1947.10
10	c	3.51	12698	14069.88	25	p	0.45	1645	1516.13
11	u	3.49	12634	12670.42	26	ö	0.28	1008	1132.90
12	l	3.45	12467	11393.45	27	j	0.20	734	798.06
13	g	3.19	11561	10221.64	28	q	0.01	26	513.08
14	m	2.56	9250	9141.86	29	x	0.00	5	280.75
15	o	2.05	7423	8144.00	30	y	0.00	5	106.16
$K = 3.3554$, $M = 0.7350$, $n = 29$, $X^2 = 3437.86$, $C = 0.0095$									

(Textgrundlage: elektronische Textversion. Der Text wurde vollständig, einschließlich Titelseite ausgewertet.)

Zwecks Anpassung der negativen hypergeometrischen Verteilung wurden die letzten vier Buchstaben zu einer Klasse zusammengefasst.



Graphik zu Tab. 11: Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Franz Kafka, *Der Prozeß*

Tabelle 12
Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Guntram Vesper, *Fugen* (6259 Buchstaben)

Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	e	15.61	977	954.87	15	o	2.17	136	149.14
2	n	10.51	658	661.08	16	b	2.14	134	132.80
3	r	7.75	485	543.48	17	f	2.01	126	117.40
4	i	7.61	476	469.62	18	w	1.66	104	102.86
5	a	6.26	392	415.16	19	k	1.45	91	89.12
6	t	6.15	385	371.57	20	z	1.05	66	76.14
7	d	5.93	371	334.94	21	ü	0.67	42	63.87
8	h	5.43	340	303.16	22	v	0.64	40	52.32
9	s	5.06	317	274.95	23	ä	0.58	36	41.47
10	u	3.80	238	249.51	24	p	0.53	33	31.35
11	c	3.28	205	226.29	25	ß	0.42	26	21.98
12	g	3.20	200	204.90	26	ö	0.32	20	13.46
13	m	2.88	180	185.04	27	j	0.14	9	5.98
14	l	2.75	172	166.51					
$K = 2.9541$, $M = 0.7251$, $n = 26$, $X^2 = 46.28$, $C = 0.0074$									

(Guntram Vesper, *Fugen*. In: Guntram Vesper. 1984. *Landeinwärts. Prosa und Gedichte.* Stuttgart: Reclam. S. 6-9). Die Buchstaben <q, x, y> kommen in der Erzählung nicht vor.

Tabelle 13
Anpassung der negativen hypergeometrischen Verteilung an die Schriftzeichen in Guntram Vesper, *Fugen* (7555 Schriftzeichen einschließlich der Leerzeichen)

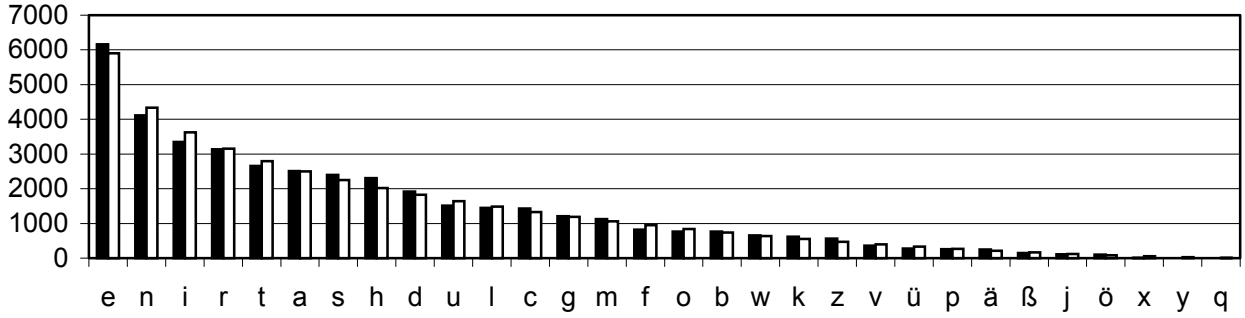
Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	LZ	14.57	1101	1215.51	17	b	1.77	134	140.36
2	e	12.93	977	796.70	18	f	1.67	126	125.29
3	n	8.71	658	640.78	19	.	1.42	107	111.18
4	r	6.42	485	546.48	20	w	1.38	104	97.96

5	i	6.30	476	478.85	21	k	1.20	91	85.58
6	a	5.19	392	425.93	22	,	1.10	83	73.98
7	t	5.10	385	382.31	23	z	0.87	66	63.15
8	d	4.91	371	345.11	24	ü	0.56	42	53.04
9	h	4.50	340	312.61	25	v	0.53	40	43.65
10	s	4.20	317	283.73	26	ä	0.48	36	34.98
11	u	3.15	238	257.71	27	p	0.44	33	27.02
12	c	2.71	205	234.04	28	ß	0.34	26	19.81
13	g	2.65	200	212.33	29	ö	0.26	20	13.37
14	m	2.38	180	192.31	30	j	0.12	9	7.79
15	l	2.28	172	173.75	31	:	0.07	5	3.22
16	o	1.80	136	156.47					
$K = 3.0799$, $M = 0.6859$, $n = 30$, $X^2 = 92.00$, $C = 0.0122$									

Tabelle 14
Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben
in Otto Jägersberg, *Dazugehören* (40977 Buchstaben)

Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	e	15.04	6161	5899.79	16	o	1.87	768	835.94
2	n	10.03	4109	4330.64	17	b	1.86	761	734.18
3	i	8.18	3351	3626.33	18	w	1.60	657	639.96
4	r	7.66	3137	3156.48	19	k	1.52	622	552.86
5	t	6.49	2660	2796.09	20	z	1.37	561	472.55
6	a	6.11	2504	2499.81	21	v	0.88	360	398.75
7	s	5.85	2398	2246.25	22	ü	0.67	274	331.26
8	h	5.63	2309	2023.64	23	p	0.63	260	269.92
9	d	4.67	1914	1824.79	24	ä	0.61	251	214.61
10	u	3.68	1510	1645.04	25	ß	0.36	146	165.27
11	l	3.54	1452	1481.15	26	j	0.28	113	121.90
12	c	3.48	1425	1330.81	27	ö	0.25	104	84.53
13	g	2.94	1206	1192.28	28	x	0.01	6	53.30
14	m	2.75	1126	1064.27	29	y	0.01	4	28.44
15	f	2.01	824	945.76	30	q	0.01	4	10.38
$K = 3.4964$, $M = 0.7775$, $n = 29$, $X^2 = 269.79$, $C = 0.0066$									

(Die Erzählung wurde vollständig, aber ohne Überschrift und Autor ausgewertet. Text in: Otto Jägersberg, *Der letzte Biß*. Zürich: Diogenes 1977, S. 113-164)



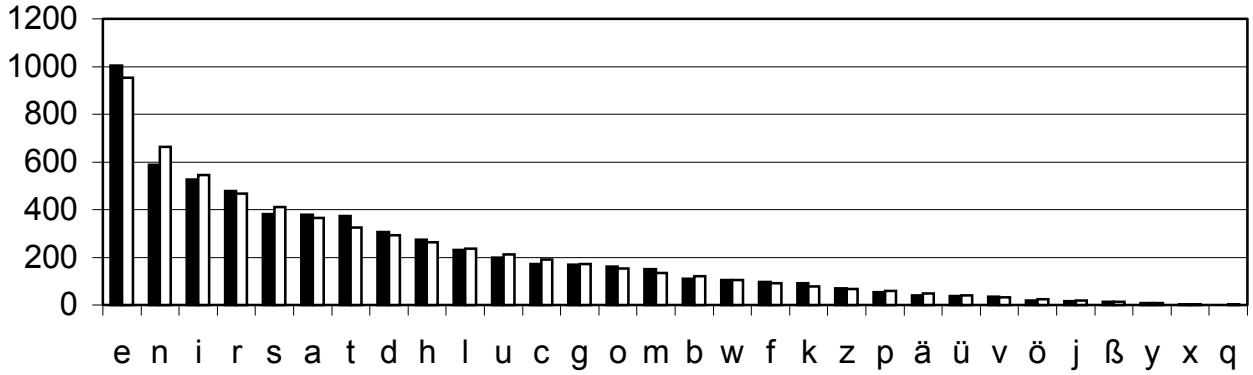
Graphik zu Tab. 14: Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Otto Jägersberg, *Dazugehören*

Bei den nun folgenden Tabellen und Graphiken ist darauf zu achten, dass sie fast alle erst nach Einführung der neuen Rechtschreibung verfasst wurden, die zum 1.8.1998 in Kraft getreten ist. Gewisse Verschiebungen gegenüber den alten Rechtschreibregeln betreffen vor allem die <ss>- und <ß>-Schreibung sowie die Wiedergabe etlicher Fremdwörter (<Photograph> oder <Fotograf>, aber nur: <Philosophie>).

Tabelle 15
Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Josef Joffe, *Nach dem Bruderkrieg* (6091 Buchstaben)

Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	e	16.47	1003	951.77	16	b	1.81	110	119.76
2	n	9.65	588	663.52	17	w	1.74	106	105.26
3	i	8.64	526	543.96	18	f	1.59	97	91.86
4	r	7.86	479	467.45	19	k	1.48	90	79.48
5	s	6.24	380	410.43	20	z	1.13	69	68.07
6	a	6.22	379	364.56	21	p	0.90	55	57.58
7	t	6.14	374	325.96	22	ä	0.66	40	47.97
8	d	5.02	306	292.52	23	ü	0.62	38	39.23
9	h	4.50	274	262.99	24	v	0.56	34	31.32
10	l	3.78	230	236.54	25	ö	0.33	20	24.24
11	u	3.27	199	212.61	26	j	0.25	15	17.99
12	c	2.84	173	190.79	27	ß	0.23	14	12.57
13	g	2.79	170	170.80	28	y	0.11	7	8.00
14	o	2.64	161	152.41	29	x	0.03	2	4.32
15	m	2.48	151	135.44	30	q	0.02	1	1.61
$K = 3.4038$, $M = 0.7372$, $n = 29$, $X^2 = 35.42$, $C = 0.0058$									

(Es wurde nur der laufende Text ausgewertet. Text In: Josef Joffe, *Nach dem Bruderkrieg. Bush und Schröder: Ernüchterung als schöpferische Chance.* DIE ZEIT 40/2003. Internetversion.)



Graphik zu Tab. 15: Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Joffe, *Nach dem Bruderkrieg*

Tabelle 16
Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Ralf Hoppe, *Das gierige Gehirn* (20075 Buchstaben)

Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	e	15.87	3186	3078.24	16	b	1.94	389	406.05
2	n	10.01	2010	2149.45	17	k	1.84	370	358.61
3	i	8.30	1666	1765.99	18	f	1.65	332	314.55
4	r	7.51	1508	1521.28	19	w	1.56	313	273.66
5	s	6.88	1382	1339.14	20	p	1.10	220	235.76
6	t	6.71	1348	1192.68	21	z	1.07	214	200.72
7	a	5.80	1164	1069.42	22	v	0.99	199	168.42
8	h	4.51	906	962.58	23	ü	0.55	111	138.80
9	d	4.14	832	868.10	24	ä	0.48	96	111.80
10	l	3.68	739	783.34	25	ö	0.36	73	87.39
11	u	3.23	648	706.50	26	y	0.28	56	65.60
12	c	2.95	593	636.29	27	j	0.23	46	46.45
13	o	2.79	560	571.78	28	ß	0.15	30	30.06
14	m	2.69	541	512.25	29	x	0.06	13	16.58
15	g	2.62	525	457.15	30	q	0.02	5	6.34
$K = 3.3278$, $M = 0.7366$, $n = 29$, $X^2 = 103.44$, $C = 0.0052$									

(Der Text wurde vollständig mit Überschrift und Autor, aber ohne Quellenangabe ausgewertet. Text In: Ralf Hoppe, *Das gierige Gehirn*. DER SPIEGEL 39/2003. Internetversion.)

Die folgende Tabelle enthält wiederum alle Buchstaben, aber auch alle sonstigen Schriftzeichen einschließlich der Leerzeichen:

Tabelle 17
Anpassung der negativen hypergeometrischen Verteilung an die Schriftzeichen in Ralf Hoppe, *Das gierige Gehirn* (insgesamt 24977 Zeichen einschließlich der Leerzeichen)

Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	LZ	14.72	3676	3830.31	28	ä	0.38	96	117.16
2	e	12.76	3186	2646.52	29	ö	0.29	73	103.07
3	n	8.05	2010	2145.25	30	y	0.22	56	90.42
4	i	6.67	1666	1822.98	31	BST	0.20	50	79.09
5	r	6.04	1508	1583.85	32	j	0.18	46	68.95
6	s	5.53	1382	1393.45	33	:	0.18	45	59.92
7	t	5.40	1348	1235.58	34	0	0.16	39	51.88
8	a	4.66	1164	1101.25	35	1	0.15	38	44.74
9	h	3.63	906	984.95	36	ß	0.12	30	38.43
10	d	3.33	832	883.00	37	GST	0.12	29	32.86
11	l	2.96	739	792.84	38	2	0.07	18	27.97
12	u	2.59	648	712.57	39	?	0.07	18	23.68
13	c	2.37	593	640.75	40	9	0.06	15	19.94
14	o	2.24	560	576.25	41	5	0.06	14	16.69
15	m	2.17	541	518.15	42	4	0.05	13	13.89
16	g	2.10	525	465.72	43	x	0.05	13	11.47
17	,	1.89	472	418.32	44	6	0.05	13	9.41
18	b	1.56	389	375.43	45	;	0.05	12	7.65
19	k	1.48	370	336.58	46	3	0.05	12	6.17
20	f	1.33	332	301.40	47	7	0.04	9	4.92
21	w	1.25	313	269.53	48	8	0.04	9	3.89
22	.	1.01	252	240.67	49	q	0.02	5	3.03
23	p	0.88	220	214.54	50	AP	0.01	3	2.33
24	z	0.86	214	190.90	51	!	0.01	3	1.77
25	v	0.80	199	169.53	52	(	0.00	1	1.32
26	„	0.64	160	150.23	53	)	0.00	1	2.99
27	ü	0.44	111	132.83					
$K = 6.4117$, $M = 0.7414$, $n = 64$, $X^2 = 305.04$, $C = 0.0122$									

(Der Text wurde vollständig mit Autor und Titel, aber ohne Quellenangabe ausgewertet. LZ: Leerzeichen; BST: Binde- und Auslassungsstrich; AP: Apostroph; GST: Gedankenstrich.)

Es folgen Tabellen zu einigen Textkorpora. Auf die Angaben von Zemanek (1959: 34) wird hier nur verwiesen. Sie umfasst 103500 Buchstaben, jedoch ohne Umlaute und <ß>. Außerdem ist nicht klar, woher die Daten stammen bzw. auf welcher Textgrundlage sie gewonnen wurden. Meier (^{2}1967: 334) gibt für 40120 Buchstaben nur die relativen Anteile wieder, so dass kein Test möglich ist; aus dem gleichen Grund kommt auch die Darstellung von Bauer (^{3}2000: 303-304) mit 681972 Buchstaben (ohne Umlaute und <ß>) nicht infrage. Daher wird als erstes eine Erhebung von Schönpflug (1969) berücksichtigt, bei der es dem Autor darum ging, eine repräsentative Auswahl von Texten zugrunde zu legen. Sie ist repräsentativ für eine studentische Versuchspersonengruppe, deren Aufnahme von öffentlich zugänglichem Sprachmaterial in der Auswahl von 1000 Textabschnitten mit einer Länge von jeweils 100 Wörtern angemessen abgebildet sein sollte. Privatgespräche wurden nicht berücksichtigt.

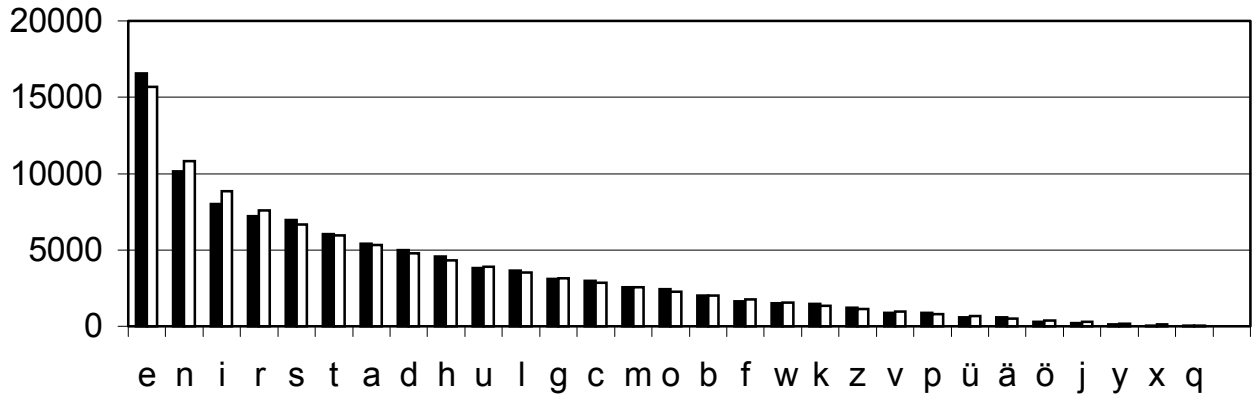
Tabelle 18
Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Schönpflugs Textkorpus (99984 Buchstaben)

Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	e	16.57	16568	15685.67	16	b	2.01	2014	2007.11
2	n	10.14	10139	10815.77	17	f	1.63	1633	1768.68
3	i	8.00	7999	8845.09	18	w	1.52	1524	1546.82
4	r	7.23	7227	7600.09	19	k	1.49	1485	1340.48
5	s	6.97	6966	6679.53	20	z	1.23	1232	1148.82
6	t	6.03	6028	5942.63	21	v	0.89	892	971.21
7	a	5.39	5392	5324.40	22	p	0.87	873	807.18
8	d	5.00	5002	4789.71	23	ü	0.61	606	656.44
9	h	4.57	4567	4317.48	24	ä	0.60	601	518.88
10	u	3.83	3826	3894.14	25	ö	0.29	288	394.57
11	l	3.67	3665	3510.43	26	j	0.19	192	283.80
12	g	3.09	3090	3159.76	27	y	0.11	107	187.20
13	c	2.98	2982	2837.33	28	x	0.04	43	105.88
14	m	2.58	2575	2539.50	29	q	0.02	22	41.93
15	o	2.45	2446	2263.48					
$K = 3.2268$, $M = 0.7265$, $n = 28$, $X^2 = 462.28$, $C = 0.0046$									

Die Tabelle enthält nur die Gesamthäufigkeit der Buchstaben (Schönpflug 1969: 162f.); es wird darauf verzichtet, auch noch die Verteilung der Buchstaben auf die Wortpositionen 1 - 4 und ≥ 5 einzeln zu untersuchen.

In Schönpflugs Korpus wird ohne jeden Kommentar der Buchstabe <ß> ausgelassen; ob er durch <s> und <z> ersetzt oder einfach nur ausgelassen wird, ist nicht erkennbar.

Schönpflug (1969: 181) lässt auch eine gewisse Skepsis hinsichtlich der Repräsentativität seiner Daten erkennen: „Die Frage muß offen bleiben, wie weit die untersuchten Texte auch repräsentativ für dasjenige Material sind, das Probanden begegnet, welche sich nicht gerade im Studium befinden. Wahrscheinlich wird die hier getroffene Auswahl gegenüber anderen möglichen seltenere Wörter und Wortformen bevorzugt haben, da sie einen hohen Anteil wissenschaftlicher Fachliteratur enthielt."

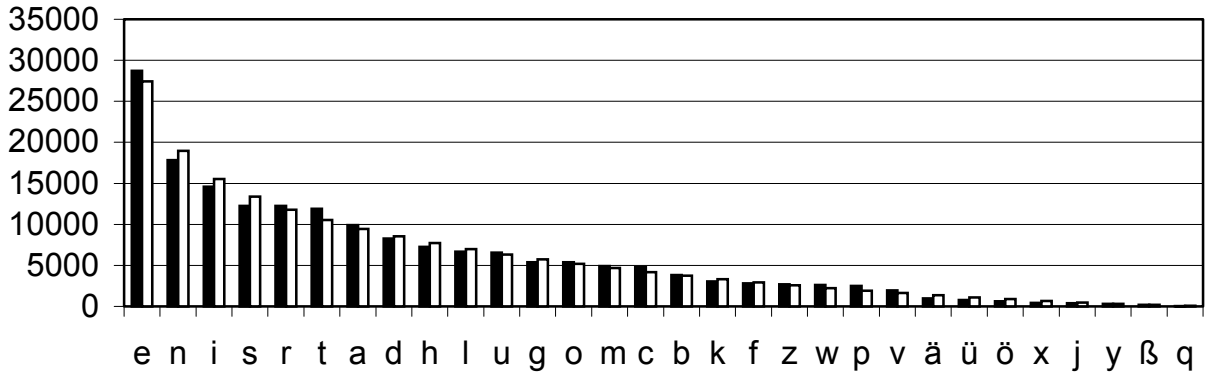


Graphik zu Tab. 18: Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in Schönpflugs Textkorpus

Tabelle 19
Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in einem linguistischen Skript (179922 Buchstaben)

Rang	<...>	%	n_x	NP_x	Rang	<...>	%	n_x	NP_x
1	e	15.96	28711	27444.96	16	b	2.13	3825	3743.91
2	n	9.90	17814	18944.47	17	k	1.68	3028	3328.97
3	i	8.10	14575	15525.42	18	f	1.56	2804	2941.49
4	s	6.81	12258	13373.70	19	z	1.49	2686	2579.59
5	r	6.81	12251	11786.93	20	w	1.46	2626	2241.75
6	t	6.62	11908	10519.00	21	p	1.39	2496	1926.80
7	a	5.50	9898	9456.44	22	v	1.08	1948	1633.87
8	d	4.60	8282	8537.98	23	ä	0.54	974	1362.33
9	h	4.04	7270	7726.91	24	ü	0.44	785	1111.82
10	l	3.72	6691	6999.57	25	ö	0.35	633	882.24
11	u	3.65	6559	6339.84	26	x	0.25	446	673.81
12	g	3.00	5390	5736.31	27	j	0.22	388	487.10
13	o	2.99	5386	5180.57	28	y	0.17	312	323.24
14	m	2.74	4923	4666.32	29	ß	0.12	212	184.22
15	c	2.67	4808	4188.69	30	q	0.02	35	73.74
$K = 3.2002$, $M = 0.7254$, $n = 29$, $X^2 = 1423.51$, $C = 0.0079$									

(Es handelt sich um ein Skript des Verf. in der noch nicht ganz abgeschlossenen Neubearbeitung von 2003, das als Begleitlektüre zu linguistischen Seminaren und als Repetitorium eingesetzt wird. Es enthält die üblichen Themen, die in einer Einführung in die Linguistik vorkommen müssen: Phonetik, Phonologie, Morphologie, Syntax, usw. Der Text wurde für die hier vorgestellte Auswertung gekürzt um: mathematische Formeln, fremdsprachige Zitate, Literaturverzeichnisse, die meisten Tabellen und Transkriptionen. Im Fall von <ss>- und <ß>-Schreibung wurde die neue Rechtschreibung befolgt. Die Fremdwortschreibung ist konservativ; sie folgt in der Regel nicht der neuen Rechtschreibung.)



Graphik zu Tab. 19: Anpassung der negativen hypergeometrischen Verteilung an die Buchstaben in einem linguistischen Skript

4. Ergebnisse und Perspektiven

Die Untersuchung hat folgende Ergebnisse erbracht:

Die Tabellen und Graphiken zeigen, mit welchen Häufigkeiten Buchstaben und in einigen Fällen auch andere Schriftzeichen in deutschen Texten vertreten sind. Die relativen Häufigkeiten der Schriftzeichen schwanken, ebenso ihre Abfolge in der Häufigkeitsrangfolge. Ob es sich dabei um autor- oder textsortenspezifische Abweichungen handelt, lässt sich erst dann sagen, wenn wesentlich mehr Vergleichsdaten zur Verfügung stehen. Immerhin ist klar, dass unter bestimmten Umständen mit einem erhöhten Anteil einzelner Buchstaben zu rechnen ist. In mathematischen Texten etwa wird <x> eine andere Rolle spielen als in vielen anderen Textsorten. Nasvytis (1953: 86) wiederum führt <c> als extrem seltenen Buchstaben an, der in den hier vorgelegten Tabellen dagegen immer eine mittlere Häufigkeit aufweist. In schweizerischen deutschen Texten wird <ss> auf Kosten von <ß> bevorzugt. Eine Verschiebung zwischen <ss> und <ß> ist auch nach der neuen Rechtschreibung zu erwarten (Duden ^{22}2000: 1120, § 25). Die Vorschläge zur vorsichtigen Anpassung von Fremdwörtern, die zum allgemein gebräuchlichen Wortschatz gehören (z.B. <Exposé> → <Exposee> und <Typographie> → <Typografie> als Hauptform; Kürschner 2001: 24f.), dürften ebenfalls zu Verschiebungen in der Häufigkeit von Buchstaben führen.

Bei den in dieser Arbeit ausgewerteten Texten fällt auf, dass in allen Fällen <e> und <n> − in dieser Reihenfolge − die häufigsten Buchstaben sind, wenn man einmal von den anderen Schriftzeichen absieht. Schon beim dritthäufigsten Buchstaben treten jedoch Unterschiede auf: meist ist es <i>, in Büchners *Landbote*, Kafkas *Verwandlung* und Vespers *Fugen* aber <r>. Solche Verschiebungen sind dann über die gesamte Rang-Häufigkeitsskala hinweg zu beobachten.

In allen hier vorgelegten Fällen wurde nachgewiesen, dass die Buchstaben, in eine Rang-Häufigkeitstabelle gebracht, entsprechend der negativen hypergeometrischen Verteilung vertreten sind. Das gilt sowohl für die Einzeltexte als auch für die hier betrachteten Textkorpora, und es gilt auch dann, wenn alle Schriftzeichen berücksichtigt werden und nicht nur die Buchstaben. Es wäre nach Erarbeitung weiterer Daten zu klären, ob dies für deutsche Texte generell gilt oder ob Einflüsse von Autoren, Entstehungszeit, Funktionalstil oder Textsorten möglicherweise dazu zwingen, unterschiedliche Verteilungen anzuwenden. In Altmann (1988: 69ff.) und Best (^{2}2003: 78-85) wird gezeigt, dass Buchstaben, Phoneme und Wörter in etlichen Fällen der Zipf-Mandelbrot-Verteilung folgen; in anderen Fällen entsprechen sie der geometrischen Verteilung (Altmann & Lehfeldt 1980: 144). Zörnig & Altmann (1983: 205) stellten jedoch bereits fest, dass die geometrische Verteilung kein immer geeignetes Modell ist. Beide Verteilungen wurden auch für die hier bearbeiteten Texte und Korpora in Betracht gezogen; die Ergebnisse waren jedoch wesentlich schlechter als die in dieser Arbeit vorgestellten. Für die Rang-Häufigkeitsordnung von Lauten/ Phonemen in anderen deutschen Textkorpora musste das Modell von Altmann (1993) verwendet werden (Best 2004/05), das sich dann auch bei weiteren Untersuchungen zum Deutschen und Englischen bewährt (Best 2005).

Erst nach Abschluss der vorliegenden Untersuchungen sind mir die Arbeiten von Grzybek u.a. zum gleichen Thema in Anwendung auf slavische Sprachen bekannt geworden. In Grzybek, Kelih & Altmann (2004) werden unterschiedliche Modelle für die Verteilung von Buchstaben entwickelt und anhand russischer Texte geprüft. Als Ergebnis dieser Untersuchung erwies sich nur die negative hypergeometrische Verteilung als ein geeignetes Modell, während die anderen in Betracht gezogenen Verteilungen zu keinen akzeptablen Anpassungen führten. Grzybek & Kelih (2003) bestätigen dieses Ergebnis auch für Slowenisch. Die Unter-

suchungen zum Deutschen einerseits und zwei slavischen Sprachen andererseits kommen also unabhängig voneinander zu den gleichen Ergebnissen.

Ein vorläufig eindeutiges Bild ergibt sich, wenn man die Häufigkeitsränge der Anfangs- (Muthmann 1996: 36) und der Endbuchstaben in einem Lexikon (Muthmann 1988: 65) betrachtet; in beiden Fällen kann die negative hypergeometrische Verteilung nicht als Modell der Wahl betrachtet werden. Gute Ergebnisse erzielt man dagegen wiederum mit Altmanns Modell (Altmann 1993: 62), womit die Befunde in Best (2004/05; 2005) unterstützt werden. Dieses Modell bewährt sich auch bei Muthmanns eigenwilliger Darstellung der Anfangsphoneme im Lexikon (Muthmann 1996: 36); in diesem Fall kann aber auch die negative hypergeometrische Verteilung angepasst werden.

In diesem Zusammenhang sei noch auf eine interessante Verwendungsperspektive hingewiesen: Wheeler (2003) entwickelt ein Verfahren, Textprofile auf der Grundlage der Häufigkeit von ASCII-Zeichen zu erstellen. Er nimmt an, dass man auf dieser Grundlage einmal in der Lage sein könnte, unerwünschte von erwünschten E-Mails zu unterscheiden, da sich diese Textarten anscheinend danach unterscheiden, ob sie mehr oder weniger Leerzeichen enthalten und teilweise auch danach, welche Zeichen in ihnen mit welchen Häufigkeiten vorkommen oder auch nicht.

Es muss also festgestellt werden, dass unter den beiden Gesichtspunkten, die dieser Untersuchung zugrunde liegen (Erweiterung der Datenbasis; Überprüfung, welchen Verteilungen die Schriftzeichen entsprechen), weiterer Forschungsbedarf besteht und dass mit solchen Untersuchungen durchaus vielversprechende Perspektiven (Wheeler 2003) verbunden sind. Zu letzteren gehört m.E. auch eine Idee von Belevitch, auf die hier abschließend hingewiesen sei. Er sieht einen engen Zusammenhang zwischen der sehr unterschiedlichen Verwendungshäufigkeit von Buchstaben und Phonemen und dem Aufwand, sie zu produzieren. Also, so schließt er, muss man den Produktionsaufwand der Einheiten messen können, was bei Phonemen auf der Grundlage der phonetischen oder phonologischen Merkmale relativ leicht ist. Auch bei chinesischen (Yu 2001) und japanischen Schriftzeichen (katakana: Belevitch 1956: 435f.) ist das wegen ihrer besonderen Gestaltung gut möglich. „De toutes façons, le coût d'un symbole peut s'interpréter comme étant le nombre de dichotomies nécessaires pour son décodage" (Belevitch 1956: 431). Also: Wenn man die Komplexität messen kann, welche jedem einzelnen Schriftzeichen eigen ist, dann kann man auf dieser Basis ein Maß für ihren unterschiedlichen Schwierigkeitsgrad gewinnen und damit wiederum die empirisch gewonnenen Rang-Häufigkeitsverteilungen begründen. Einen neuen Vorschlag zur Messung der Komplexität von Schriftzeichen findet man bei Altmann (2004).

5. Wissenschaftshistorische Nachbemerkung

Buchstabenzählungen gehören zu den frühesten sprachstatistischen Untersuchungen zum Deutschen; Meier (^{2}1967: 349ff.) führt in seiner „Zeittafel sprachstatistischer Arbeiten" als zweitältestes Werk Förstemann (1846) an, eine Untersuchung, bei der der Titel auf eine Lautstatistik hindeutet; im Text betont Förstemann (1846: 83) aber ausdrücklich, dass er Buchstaben ausgezählt hat. Anhand einer Untersuchung der relativen und wechselnden Häufigkeit von Buchstaben bzw. Buchstabengruppen im Gotischen, Althochdeutschen, Mittelhochdeutschen und Neuhochdeutschen möchte er u.a. Aufschluss über Phasen unterschiedlicher Vitalität der Sprachentwicklung gewinnen. So stellt er fest, dass die Proportion zwischen Vokalen und Konsonanten sich in der Phase vom Althochdeutschen zum Mittelhochdeutschen wesentlich stärker verändert hat als in der nachfolgenden vom Mittelhochdeutschen zum Neuhochdeutschen. Entsprechend hält er das Deutsche in der früheren Phase für vitaler als in der späteren. In einer nachfolgenden Untersuchung widmete sich Förstemann (1852) auch dem

Griechischen und Lateinischen. Im Anschluss an Förstemann hat dann Schleicher (1852) entsprechende Daten zum Altkirchenslawischen erhoben (Grzybek & Kelih 2003: 136f.).

Zur Bedeutung von Buchstabenzählungen betonen Grzybek & Kelih (2003: 159) mit Blick auf das Russische, dass dies „niemals nur Selbstzweck war": „Immer ging es um weiterführende Fragen, angefangen von mathematischen und methodologischen Problemen, über Fragen der Optimierung technischer Einrichtungen oder der Strukturierung von Codes und Prozessen der Informationsübertragung, bis hin zu Fragen der Textstilistik und Texttypologie." Mit Förstemann kommen sprachhistorische und typologische Aspekte hinzu, mit Wheeler aktuelle Probleme der Informatik. Auch wenn das Thema vielleicht auf den ersten Blick trivial zu sein scheint, dürfte damit einleuchten, dass es unter vielen Gesichtspunkten linguistischer und außerlinguistischer Art bedeutsam, und, wie die neuen Bemühungen zeigen, auch nach wie vor aktuell ist.

Literatur

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

Altmann, Gabriel (1993). Phoneme Counts. In: Altmann, Gabriel (ed.), *Glottometrika 14, 54-68*. Trier: Wissenschaftlicher Verlag Trier.

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

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

Bauer, Friedrich L. (³2000). *Entzifferte Geheimnisse.* 3., überarbeitete und erweiterte Auflage. Berlin/ Heidelberg: Springer.

Belevitch, V. (1956). Théorie de l'information et statistique linguistique. *Académie royale de Belgique, Bulletin de la classe des sciences, 5^{e} Série, Tome XLII/ Koninklijke Academie van Belgiën, Mededelingen van de klasse der wetenschappen, 5de Reeks, Boek XLII*: 419-436.

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

Best, Karl-Heinz (2004/05). Laut- und Phonemhäufigkeiten im Deutschen. *Göttinger Beiträge zur Sprachwissenschaft 10/11, 21-32.*

Best, Karl-Heinz (2005). Buchstabenhäufigkeiten im Deutschen und Englischen. *Naukovyj Visnyk Černivec'koho Universytetu. Vypusk 231, 119-127.*

Beutelspacher, Albrecht (^{4}1994). *Kryptologie.* 4., abermals leicht verbesserte Auflage. Braunschweig/ Wiesbaden: Vieweg.

Duden. Die deutsche Rechtschreibung. 22., völlig neu bearbeitete und erweiterte Auflage. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag 2000.

Finkenstaedt, Thomas & Wolff, Dieter (1973). *Ordered Profusion*. Studies in Dictionaries and the English Lexicon with contributions by H. Joachim Neuhaus and Winfried Herget. Heidelberg: Winter.

Förstemann, Eduard (1846). Ueber die numerischen Lautverhältnisse im Deutschen. *Germania. Neues Jahrbuch der Berlinischen Gesellschaft für Deutsche Sprache und Alterthumskunde, Bd. 7, 82-90.*

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

Grzybek, Peter & Kelih, Emmerich (2003). Grapheme Frequencies in Slovene. In: Benko, Vladimir (ed.), *Slovko 2003.* Bratislava (erscheint).

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

Haß-Zumkehr, Ulrike (2001). *Deutsche Wörterbücher.* Berlin/ New York: de Gruyter.

Hoffmann, Lothar (²1985). *Kommunikationsmittel Fachsprache. Eine Einführung.* Zweite, völlig neu bearb. Aufl. Tübingen: Narr.

Knüppel, Anke (2001). Untersuchungen zum Zipf-Mandelbrot-Gesetz an deutschen Texten (S. 248-280). In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten.* Göttingen: Peust & Gutschmidt.

Küpfmüller, K. (1954). Die Entropie der deutschen Sprache. *Fernmeldetechnische Zeitschrift 7, 265-272*.

Kürschner, Wilfried (2001). *Neue Rechtschreibung kompakt.* Vechta: Plaggenborg Verlag.

Mater, Erich (1966). *Deutsche Verben. 1. Alphabetisches Verzeichnis*. Leipzig: Bibliographisches Institut.

Meier, Helmut (²1967). *Deutsche Sprachstatistik*. Hildesheim: Olms.

Müller, Gerhard (2003). Beantwortung einer Anfrage. *Der Sprachdienst 47, 209.*

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

Muthmann, Gustav (1996). *Phonologisches Wörterbuch der deutschen Sprache.* Tübingen: Niemeyer.

Nasvytis, A. (1953). *Die Gesetzmäßigkeiten kombinatorischer Technik*. Berlin u.a.: Springer.

Schleicher, August (1852). *Die Formenlehre der kirchenslawischen Sprache, erklärend und vergleichend dargestellt*. Bonn/ Wien/ Prag: H.B. König.

Schmidt, Hartmut (1977). *Untersuchungen zu konzeptionellen Problemen der historischen Lexikographie.* Diss. phil., Berlin.

Schmidt, Hartmut (1985). *Untersuchungen zu konzeptionellen Problemen der historischen Lexikographie (Bedeutungen, Definitionen, Stichwortlisten, Aussagebereich).* Akademie der Wissenschaften der DDR, Zentralinstitut für Sprachwissenschaft, Linguistische Studien, Reihe A: Arbeitsberichte 134. (Auszug aus Schmidt 1977)

Schmidt, Hartmut (1986). *Wörterbuchprobleme. Untersuchungen zu konzeptionellen Fragen der historischen Lexikographie.* Tübingen: Niemeyer. (Fast unveränderter Nachdruck von Schmidt 1977)

Schönpflug, Wolfgang (1969). n-Gramm-Häufigkeiten in der deutschen Sprache. I. Monogramme und Digramme. *Zeitschrift für experimentelle und angewandte Psychologie XVI*: *157-183.*

Wheeler, Eric S. (2003). Multidimensional scaling to visualize text peparation. *Glottometrics 6, 65-69.*

Wimmer, Gejza & Altmann, Gabriel (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Handbook of Quantitative Linguistics*. Berlin: de Gruyter (im Druck).

Yu, Xiaoli (2001). Zur Komplexität chinesischer Schriftzeichen. *Göttinger Beiträge zur Sprachwissenschaft 5, 121-129.*

Zemanek, Heinz (1959). *Elementare Informationstheorie.* Wien/ München: Oldenbourg.

Zörnig, Peter, & Altmann, Gabriel (1983). The repeat rate of phoneme frequencies and the Zipf-Mandelbrot law. In: *Glottometrika 5, 205-211.* Ed. by Reinhard Köhler & Joachim Boy. (S.). Bochum: Brockmeyer.

Zörnig, Peter, & Altmann, Gabriel (1995). Unified representation of Zipf distributions. *Computational Statistics & Data Analysis 19*, *461-473.*

Software

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

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

Adresse des „Göttinger Projekts“ im Internet (mit ausführlicher Bibliographie): http://wwwuser.gwdg.de/~kbest

Glottometrics 11, 2005, 32-50

Word length balance in texts: Proportion constancy and word-chain-lengths in Proust's longest sentence

Simone Andersen[1], *Düsseldorf*

Abstract. Constancy phenomena in word length distributions of texts are demonstrated. The regularity of proportions is shown by intercorrelation of parts under differing kinds of partitioning. Length homogeneity r_A as a measure for the stability of the values of the distribution is developed. Balance number B refers to word-chains in line: Every B words the total number of syllables tends to be equal, indicated by decreased variance.

Keywords: word length, homogeneity, intercorrelation, length balance, word-chain-length, constancy

1. The problem of length proportions

The overall shape of the distribution of word lengths in a given text is well predictable by the word length laws (Zipf 1949; Altmann 1988; Wimmer, Köhler, Grotjahn & Altmann 1994; Wimmer & Altmann 1996, Altmann & Best 1996; Best 2001; Grzybek 2005; http//:www.gwdg.de/~kbest/litlist.htm).

How are these lengths scattered over the text? Obviously there is no fixed order of short and longer words, as long as we refer to prose. But if their patterns were completely arbitrary, it would be conceivable to find a possibly uneven scattering with heterogeneous components, e.g. a text where all the short words occur at the beginning, so that the longer ones have to crowd together at the remainder to compensate for it at the end. There are concepts in linguistics proposing that only the entire text reveals the true frequency proportions. Orlov presumes (Orlov 1982; Best 2003) that the author of a text organizes the frequency structure of the elements only for the text as a whole. So the proportions do not hold for parts of the text.

Opposed to this, we believe that length balance in speaking and writing becomes visible from the beginning. Whatever the reason for the individual distribution, we suppose that its effects on word length proportions of the text will work in a homogeneous way and presumably should be rather independent of the text producer's talent for organization. Consequently the distribution should be found in the parts as in the whole. We tried to find evidence for this by investigating a text and detecting the degree of homogenity of the word length proportions characterizing it. Additionally we were looking for hints indicating length balance in very small text segments as well. Tendency towards balance could also be revealed by another constancy phen-

[1] Address correspondence to: AndersenSC@aol.com

omenon: We were looking for units in spoken or written text that could be considered as recurring patterns within which the lengths are balanced out so that the pattern units tend to be equal.

2. Method

We partitioned one text in varying ways and observed the properties of the resulting parts. Word length was measured in number of syllables. In order to find indications of length balance we used two kinds of investigation.

The first step was comparing distributions and lengths under differing kinds of text partitioning. In the second step we partitioned the text into a finer grid and examined the length (number of syllables) of word chains - sequences of words occurring after one another in line in the text - regardless of their grammatical or semantic relations, i.e. without taking into account grammatical constituents, clauses or phrases. We studied the variability of the chain lengths depending on the number of words in the chain.

In order to eliminate influences by sentence limits, we looked for a sentence as long as possible. We chose Marcel Proust's longest sentence, detected by the writer Alain de Botton (2001), from Proust's work *A la recherche du temps perdu*.

Using the German translation of the original text makes sure that the word lengths cannot result from the poet's intention or individual taste (sense of rhythm etc.) but are to a greater extent determined by constraints coming from the language system and its properties.

3. Results

The total number of words in this longest sentence is $n = 519$.

The shape of their length distribution is as to be expected (see Table 1), with a slight overrepresentation of two-syllable words – when compared to one of the Altmann-Fitter distributions (1994; 1997) – probably due to Proust' s very detailed descriptions which typically request the use of very specific words.

Table 1
Lengths of single words in the entire sentence

Length x (number of syllables)	Frequency f_x (number of tokens with length x)	Proportion
1	203	0.391
2	194	0.374
3	63	0.120
4	40	0.077
5	16	0.031
6	1	0.002
7	2	0.004
	$n = 519$, $\overline{x} = 1.99$ $s^2 = 1.2256$	

In the first step we want to know what happens to the shape of the distribution under varied kinds of text partitioning:

Disregarding the last 19 words, we divide the remaining 500 words of the text into two parts of 250 words and get two distributions of lengths:

Table 2
Split half: Word lengths in the first and second half of text
(without the last 19 words)

Text part	1-syll	2-syll	3-syll	4-syll	5-syll	6-syll	7-syll
(I) 1-250	98	96	27	19	8	1	1
(II) 251-500	99	90	34	19	8	0	0

In the next step we partitioned the whole text into five parts containing 100 words each:

Table 3
Word lengths under text partitioning into 5 parts

Parts	1-s	2-s	3-s	4-s	5-s	6-s	7-s
first: the first 100 words	34	37	15	6	6	1	1
second: words 101-200	45	37	9	8	1	0	0
third: words 201-300	40	38	8	11	3	0	0
fourth: words 301-400	41	38	15	5	1	0	0
fifth: words 401-500	37	36	14	8	5	0	0
last 19 words	6	8	2	2	0	0	1
n = 519	203	194	63	40	16	1	2
Proportion	0.391	0.374	0.120	0.077	0.031	0.002	0.004

We observe a remarkable constancy of proportions as illustrated in Fig. 1.

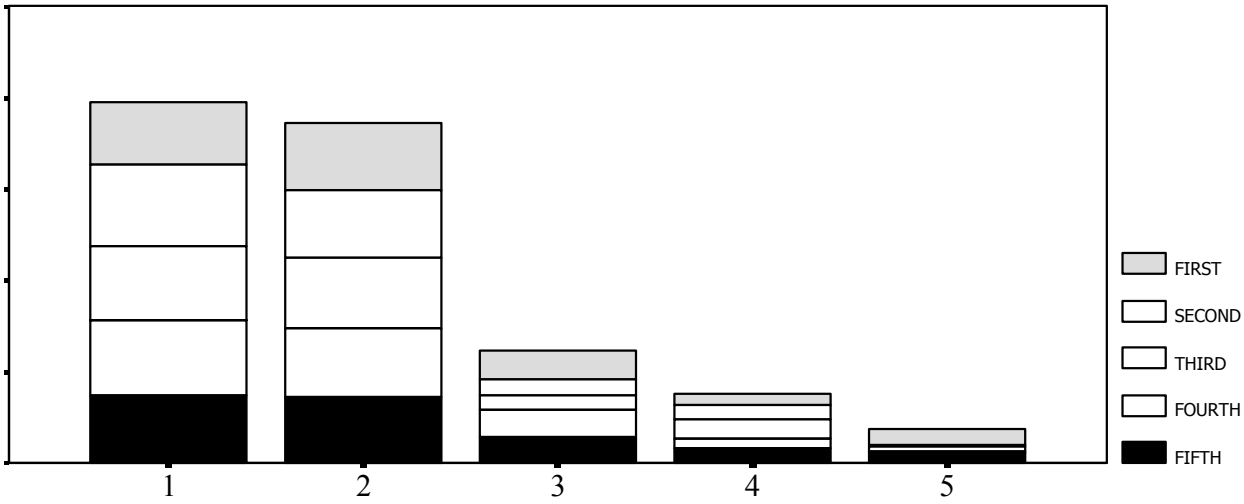


Fig. 1. Proportions of word lengths in the five parts of the text
(Length 5 = 5, 6 or 7 syllables)

The distributions of word lengths in the different parts and in the entire text are shown in Fig.

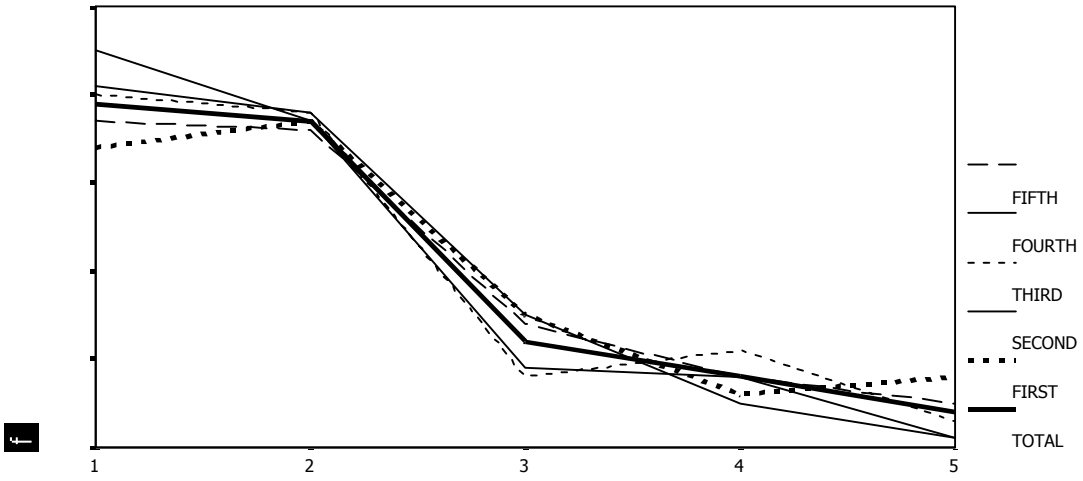


Fig. 2. Length distributions in the five parts and in the entire text (total)

Table 4 shows the proportions for word lengths in the entire text (column 1) in the first and second half (column 2 resp. 3) and in the five parts of 100 words (columns 4 – 8):

Table 4
Proportions of lengths for different parts and entire text

word length (syllables)	total	1.half	2.half	1-100	101-200	201-300	301-400	401-500
1	0.39	0.39	0.40	0.34	0.45	0.40	0.41	0.37
2	**0.37**	**0.38**	**0.36**	**0.37**	**0.37**	**0.38**	**0.38**	**0.36**
3	0.12	0.10	0.14	0.15	0.09	0.08	0.15	0.14
4	0.08	0.08	0.08	0.06	0.08	0.11	0.05	0.08
5	0.03	0.03	0.03	0.06	0.01	0.03	0.01	0.05
6	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
	519	250	250	100	100	100	100	100

Something which is very striking is the nearly constant proportion of the two syllable words (see in the second row above): It is nearly constantly $p = 0.37$ which is the overall proportion for the entire text and it can be found in every part, in the first and in the second half as in every 100 words of the text. We will go back to this later.

4. Length homogeneity of texts

Now we are able to calculate the length correlations between the different parts.

The following table (Table 5) shows the correlations r_{pp} between the parts consisting of 100 words and their intercorrelations r_{pt} as the mean of each row (without the diagonal) which is the mean of the correlations between one part and the remaining text (= the four remaining parts, without the last 19 words as explained above). The word lengths of 6 or 7 syllables could be of confounding influence: Because of their extremely low proportions (rounded values of 0.00 almost everywhere) they increase the correlations improperly. So we grouped them together with the 5-syllable-words and counted them as length class no. five (already visible in Fig.1 and Fig.2).

Table 5
Homogeneity: Intercorrelations of the parts of the text

Parts	Parts					
	1-100	101-200	201-300	301-400	401-500	r_{pt}
1-100	1.00	0.9527	0.9514	0.9829	0.9882	0.9688
101-200	0.9527	1.00	0.9915	0.9803	0.9855	0.9775
201-300	0.9514	0.9915	1.00	0.9669	0.9802	0.9725
301-400	0.9829	0.9803	0.9669	1.00	0.9971	0.9818
401-500	0.9882	0.9855	0.9802	0.9971	1.00	0.9878
						r_{int} = 0.9777

The mean of the last column is

$$r_{int} = (0.9688 + 0.9775 + 0.9725 + 0.9818 + 0.9878)/5 = 0.9777$$

which indicates the degree of intercorrelation of all parts. In analogy to test theoretical scale analysis in psychological diagnostics we could try to interprete the degree of intercorrelation to be a measure of homogeneity related to length proportions.

We will call this length proportion homogeneity or just length homogeneity r_Λ with $r_\Lambda = r_{int}$ and propose that it may be a useful measure of a given text to characterize the stability of its length distribution. In Proust's sentence length homogenity r_Λ is extremely high (0.9777), but we suppose that any text written or produced by a single author within a narrow time span will yield a considerable length homogeneity. In classical test theory, the concepts of homogeneity or stability converge towards a measure for reliability.

If we look at the proportion of an individual word length in an entire text (for example, the proportion of 0.39 as a score for one-syllable-words), we can put the question of how reliable this value is for each part of the text. Is it the resulting average of very inhomogeneous parts? Or is it a typical proportion value, valid for many text parts?

Thus the length homogeneity is a measure for the precision of assessment in determining the characteristic length distribution in a given text.

An additional step for improving the results could be eliminating those lengths that come to less than 0.01 per cent of the entire text: word lengths of 6 or 7 syllables and more are too rare to be a useful measure. Nearly always producing the same value (here: zero-proportion) means that

they have poor discrimination power: they provide no information, they are levelling the results and will be disregarded – not only in this example but generally. As we observe in Table 5.a, removal of the word lengths of 6 and 7 would not change a lot of the intercorrelation: it would increase by a very small amount.

Table 5a
Homogeneity: Intercorrelations of the parts of the text without lengths of 6 or more

Parts	Parts					
	1-100	101-200	201-300	301-400	401-500	r_{pt}
1-100	1.00	0.9575	0.9562	0.9887	0.9921	0.9736
101-200	0.9575	1.00	0.9915	0.9803	0.9855	0.9787
201-300	0.9562	0.9915	1.00	0.9669	0.9802	0.9737
301-400	0.9887	0.9803	0.9669	1.00	0.9971	0.9834
401-500	0.9821	0.9855	0.9802	0.9971	1.00	0.9887
						r_{int} = 0.9796

5. Length homogeneity r_{Λ} as a text characteristic

Now we can use r_{Λ} as a text characteristic if we observe and determine the decrease of the intercorrelation in dependence of the number of parts *t*.

Unlike in psychological scale analysis, we have no "natural" units, like the items of a test. Instead, we are able to divide a text into parts of any size, and we can make use of this fact by measuring how far a text can be partitioned into equal parts without losing a considerable amount of homogenity.

The number *N* of words within the parts that are to be compared and intercorrelated ranges from the upper limit of $N = n/2$ (with n = the number of words in the entire text) down to N = 100, because proportions of less than 10 percent (for words with 4 or more syllables) can be compared meaningfully only if it makes a difference between occurring and non-occurring in the text.

From that it follows that there is the minimum size of n = 100 words in a text to determine its word length homogenity r_{Λ}.

N = number of words in the parts; t = number of parts of equal size

$100 \geq N = n/t$

Number of parts *t*	*N*	r_{Λ}	
1	519	1.00	
2	250	0.9912	
3	170	*	
4	125	*	(* = has not been calculated here)
5	100	0.9777	

In this text, the limit for *t* is 5, because 100(6) > 519.

During partitioning up to the individual limit, the intercorrelation does not fall below 0.9. We interprete this as a very high degree of homogeneity of the length distribution in text. Only correlations of less than 0.9 should be considered as loss of homogenity.

6. Visibility of the frequency distribution and best sample

Additionally we could try to search out those parts that show the highest correlation with the entire text (although this is a rather subtle question, because of the extremely high correlations of all parts). The values are shown in Table 6.

Table 6
Correlations of text parts with the entire text (= total)

	FIRST	SECOND	THIRD	FOURTH	FIFTH
TOTAL	.9814	.9925	.9895	.9927	.9983

We can see that the correlation between part 5 (last part) with the entire text is nearly perfect (0.9983), followed by part 4. Perhaps it could mean that the proportions are being settled best towards the end, so the last part of a text reveals the frequency distribution most apparently and can be considered as the best sample of a given text. Here we approach the point at which we must discuss the question of what causes the length distribution of words in a given text. In view of the fact that the investigated sample is a translated text, and the translater has little choice in respect of determining the lengths of the words even unconsciously, we are minded to propose that the considerable homogenity provides additional evidence for what we already found and claimed in another context (Andersen 2002): other than in musical composition, the frequencies in texts are to a great extent out of reach for the individual text producer. Probably the even higher correlations of the last parts indicate a small amount of controlled or intentional production in the beginning.

7. Word-chains and their lengths

Let us now look at the text from another perspective. As we could observe already in Table 3 (6th row), the last 19 words in the text do not reveal the typical distribution. Of course, we do not expect that in every text part of any length we will find the proportions above; this would be a very straight pattern. And as we said above, we don't look for proportion constancy in parts smaller than 100 words.

Apart from proportion constancy in larger parts, we are looking for constant or at least similar patterns of length in order to find hints indicating length balance within smaller units of a text. Are we able to find a certain number *B* (balance) of words where the total number of syllables tends to be equal, regardless of the lengths of the component words?

So our investigation objects will be *word chains*. Word chains are sequences of words occurring one after another in line in the text, regardless of their grammatical or semantic relations.

The idea of length balance arising when investigating a number of units in line follows from the Menzerath-Altmann-law (Altmann 1980; Altmann & Schwibbe 1989; Hřebíček 2000): The components of the shorter units are longer (per average) than those of the longer units.

We divided the text into single words ("one-word-chains"), two-word-chains (2-w), three-word-chains (3-w) etc. up to 12-word-chains. Instead of considering the distributions of lengths within them we recorded their length, measured by number of syllables in the chain. We are interested in the variability of the chain lengths depending on the number of words in the chain.

To keep the number of tokens constant, in the beginning we considered the first 90 words (tokens) of the text. Values are shown in Table 7.

Table 7
Lengths (number of syllables) of single words, two-word-chains, three-word-chains, 10-word-chains, 11-word-chains with their frequencies f (the first 90 tokens)

Length (= number of syllables)	f(1-w)	f(2-w)	f(3-w)	f(10-w)	f(11-w)
1	30	-	-		
2	33	6	-		
3	13	11	2		
4	7	8	4		
5	5	7	4		
6	1	8	5		
7	1	1	4		
8	-	1	4		
9	-	1	4		
10	-	1	2	-	
11	-	-	-	-	-
12	-	-	-	-	-
13	-	-	1	-	-
…				1	-
18				-	-
19				1	1
20				1	2
21				1	-
22				4	1
23				–	-
24				–	2
25				–	2
26				–	-
27				1	-
28				-	1
Number of chains	90	45	30	9	9

To remember the limitations:

Because of combinatorics, the distribution of chain lengths with more than two words per chain cannot show the same shape as the distribution of single words. The shortest chains can hardly be the most frequent, because with an increasing number of words in the chain there are still fewer possibilities of realization. For example, to get the shortest three-word-chain with a length of 3 syllables, there has to be a coincidence of 1 + 1 + 1 syllables which is only one out of 7^3 possible states - taking the length of 7 syllables as a kind of an upper limit for a word. This is

a rare case, regardless of the greater probability of short words, as long as we refer to European languages.

Instead of showing the typical word length shape, the chain lengths should converge towards a favourable length as soon as length balance can be found, or as soon as some typical patterns occur where short and long words are combining in a favoured proportion.

As shown in Tables 8.1 and 8.2, we observe that the variance changes, it is fluctuating. For ten-word-chains (when standardized, see Table 8.2) it is at minimum, and then increases again:

Table 8.1
Number of syllables in word chains (the first 90 tokens)

	chains in the sample	mean	var	s	range
One-word-chains	90	2.233	1.709	1.307	6
Two-word-chains	45	4.467	3.618	1.902	8
Three-word-chains	30	6.700	5.528	2.351	10
Nine-word-chains	10	20.10	8.989	2.998	9
Ten-word-chains	9	22.33	7.500	2.739	10
11-word-chains	9**	24.00	8.750	2.958	9
12- word-chains	8*	26.25	14.21	3.770	13

*in this case: 96 words (= 8 x 12); ** 99 words (= 9 x 11)

Table 8.2
Values of Table 8.1, standardized by chain length

	chains in the sample	mean	var	s	range	mean ± s	2s
One-word-chains	90	2.233	1.709	1.307	6	0.926 – 3.540	2.614
Two-word-chains	45	2.234	1.809	0.951	4	1.282 – 3.184	1.902
Three-word-chains	30	2.233	1.842	0.784	3.33	1.450 – 3.017	1.567
Nine-word-chains	10	2.233	0.999	0.333	1	1.900 – 2.566	0.666
Ten-word-chains	9	2.233	0.750	0.274	1	1.959 – 2.507	0.548
11-word-chains	9**	2.182	0.795	0.269	0.82	1.911 – 2.449	0.538
12- word-chains	8*	2.188	1.185	0.314	1.1	1.873 – 2.502	0.629

*in this case: 96 words (= 8 x 12); ** 99 words (= 9 x 11)

The fluctuating variance is striking because it is related to the same 90 words and differs depending on the kind of partitioning. The range (standardized) decreases and increases.

The amount of mean ± s (the "2/3 area" of all values) decreases and increases again with increasing chain length. For ten-word-chains and eleven-word-chains it is at minimum, for 12-word-chains it increases again.

In the first 90 tokens, we considered a sample with constant size but differing number of chains.

In order to eliminate the varying number of cases, we now consider for every chain length the first 10 chains and record their lengths, and their variance (see Tables 9.1 and 9.2):

Table 9.1
The first 10 chains: average number of syllables

	Words in the sample	mean	var	s	range
Two-word-chains	20	4.4	1.822	1.35	3
Three-word-chains	30	6.2	4.844	2.2	7
Nine-word-chains	90	20.1	8.989	2.998	9
Ten-word-chains	100	21.9	8.544	2.923	10
11-word-chains	110	23.6	9.378	3.062	9
12- word-chains	120	25.2	15.96	3.994	13

Table 9.2
Values of Table 9.1, standardized by chain length

	mean	var	s	mean ± s
two-word-chains	2.200	0.911	0.675	1.520 – 2.875
Three-word-chains	2.066	1.615	0.733	1.330 – 2.799
Nine-word-chains	2.233	0.999	0.333	1.900 – 2.566
Ten-word-chains	2.190	0.854	0.292	1.898 – 2.480
11-word-chains	2.145	0.853	0.278	1.867 – 2.423
12- word-chains	2.100	1.330	0.333	1.767 – 2.433

Again, we find that the standardized variance decreases and increases with increasing chain length.

8. Comparing the variances

Let us now look at the variability of the data values. We are not allowed to calculate a statistical Analysis of variance, because the assumptions required are not fulfilled (independent groups, normal distribution). But we do not need it either, because we are not really interested in the question of whether the chain means are equal or not. Of course, they are not. Rather, we are interested in finding evidence for greater homogenity of the first ten 10-word chains compared to the first ten 3-word-chains.

We want to know if the variability of each can be attributed to the variability among the chains, or rather to some characteristics of the individual chains. So we must compare the variance within each chain to the variance between the chains. For this purpose we make use of the starting procedure of an analysis of variance.

10-word-chains:

In Table 9.1 we find the variance between the chains: $s^2(\text{betw})_{10} = 8.544$.
If we divide this by the number of words per chain (here: 10 words), we get the average variance *between* the chains: $s_w^2(\text{betw})_{10} = 0.854$ (see Table 9.2).

To get the variance *within* the chains, we have to sum up the ten single variances (see Table 10, last row): $s^2(\text{in})_{10} = 16.85$ and divide it by the number of chains, so we get the mean variance in a chain as: $s_w^2(\text{in})_{10} = 16.85 : 10 = 1.685$

Table 10
Distributions of word lengths (number of syllables) in 10-word-chains

Syll	Chain no.									
	1	2	3	4	5	6	7	8	9	10
1	3	2	7	4	2	4	4	3	1	4
2	5	4	1	3	4	1	5	4	6	4
3	0	3	0	1	2	4	0	1	2	2
4	2	1	1	0	0	1	0	1	1	0
5	0	0	1	2	1	0	0	1	0	0
6	0	0	0	0	0	0	1	0	0	0
7	0	0	0	0	1	0	0	0	0	0
Total	10	10	10	10	10	10	10	10	10	10
Mean	2.1	2.3	1.8	2.3	2.8	2.2	2.0	2.3	2.3	1.8
Var	1.21	0.9	2.18	2.46	3.51	1.29	2.22	1.79	0.68	0.62

This is certainly not the same value as: $s_w^2(\text{betw})_{10} = 0.854$, but as we said before, we do not want to calculate an ANOVA or an F-test.

We only want to compare the variance that is due to the length variation of the words within a chain to the variance due to the length variation between the chains.

variance for 10-word-chains
between within chains
0,854 1,685

If we do the same for the 3-word-chains, there is a noticeable difference:

3-word-chains:

Calculating the variance within the chains, we sum up the last row of Table 11 (the single variances of the 3-word-chains) and get $s^2(\text{in})_3 = 12.664$. If we standardize, we get 1.266 as the average variance within the chains.

We compare to the variance between the chains which is $s^2(\text{betw})_3 = 4.844$ (see Table 9.1) then standardized as average variance between: $s_w^2(\text{betw})_3 = 1.615$ (see Table 9.2)

Variance for 3-word-chains
between within chains
1.615 1.266

Table 11
Distributions of word lengths (number of syllables) in 3-word-chains:

Syll	Chain no.									
	1	2	3	4	5	6	7	8	9	10
1	1	1	1	0	0	1	2	3	1	2
2	1	1	2	2	1	1	1	0	0	1
3	0	0	0	0	2	1	0	0	0	0
4	1	1	0	1	0	0	0	0	1	0

5	0	0	0	0	0	0	0	0	1	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
Total	3	3	3	3	3	3	3	3	3	3
Mean	2.33	2.33	1.67	2.67	2.67	2	1.33	1	3.33	1.33
Var	2.33	2.33	0.33	1.33	0.33	1	0.33	0	4.33	0.33

We observe that the variance between the 3-word-chains exceeds the variance within them, which means that the 3-word-chains are more "individual": there are typical longer chains and typical shorter ones, and for the length of a word it is more important in which 3-word-chain it occurs than the fact that it occurs in a 3-word-chain. And conversely: 10-word-chains determine the lengths of their words more than 3-word-chains do. The 10-word-chains are more similar to one another than the 3-word-chains are. In 10-word-chains there seems to be a balancing influence that effects the lengths of their words.

Because of that we take a look at the variance of all ten-word-chains in the entire text. We divided the complete text (519 words) into 52 ten-word-chains. Their lengths are given in Table 12.

Table 12
Lengths of all 52 ten-word-chains

Length (number of syllables) x	Frequency f_x	Proportion observed f_x/n
14	1	0.019
15	4	0.077
16	**6**	**0.115**
17	2	0.038
18	**6**	**0.115**
19	**6**	**0.115**
20	4	0.077
21	3	0.058
22	3	0.058
23	**9**	**0.173**
24	5	0.096
25	1	0.019
26	1	0.019
27	1	0.019
$\Sigma x f_x = 1040$	52	1.00
Mean = 20.00, s = 3.331, var = 11.098		

As it became already visible in Table 7, the length of 23 syllables is a preferred length for a 10-word-chain. Another typical length seems to be the number of 16, 18 or 19 syllables. Half of the observed chains are showing one of these lengths.

The observed length of a 10-word-chain ranges between 14 and 27 syllables, as illustrated in Fig. 3.

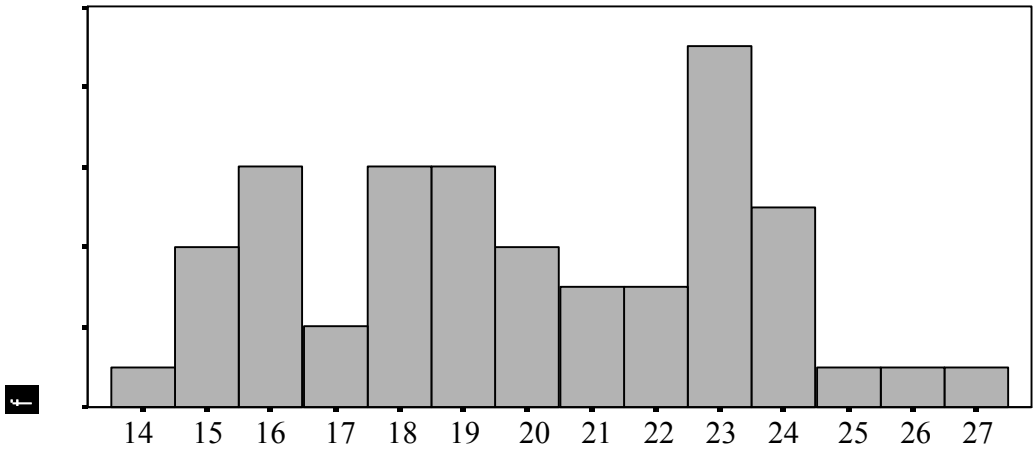


Fig. 3. Syllable numbers of all 52 ten-word-chains

We want to compare the variance between the chains and the variance within them for the entire text: The variance between all 52 chains is given in Table 12: $s^2(\text{betw})_{10} = 11.098$

If we divide it by the number of words per chain (here: 10 words), we get the variance between the chains per word on average: $s_w^2(\text{betw})_{10} = 1.1098$

To get the variance within the chains, we have to sum up the 52 single variances (data see appendix): $s^2(\text{in})_{10} = 63.556$ and divide it by the number of chains, so we get the mean variance within a chain as: $s_w^2(\text{in})_{10} = 63.556 : 52 = 1.222$.

variance for all 52 ten-word-chains	
between	within chains
1.1098	1.222

The variance within the 10-word chains exceeds the variance between the 10-word-chains. In a classical test of analysis of variance, one usually would try to corroborate the hypothesis that the units (here for example: the chains) are differing significantly from one another by showing that the variance between them is significantly greater than the variance within them. Here it is the reverse. Not only are we unable to find the variance between the chains significantly greater, it is in fact even smaller than the variance within the chains. Furthermore, it is also smaller than the total variance of all the words in the text that equals the variance within (see Table 1).

So we can state that the chains are more similar to one another than the words within the chains, and more similar than could be expected by the total variance of word lengths in the text as a whole.

9. Constancy of the proportion of two-syllable-words

In search of possible causes for this constancy we want to take a second look at the phenomenon of Table 4: the constant proportion of two-syllable-words.

We determined the occurrence of them in all of the 52 ten-word-chains and compared their proportion to the values of the binomial distribution with $n = 10$ and $p = 0.374$ (the proportion of two-syllable-words in the text, see Table 1 above). Values are given in Table 13:

Table 13
Occurrence (number x) of two-syllable-words in a 10-word-chain

Occurrence of 2-syllabic words in a 10-word-chain x	Cases (number of chains) f_x	$P_{observed}$ f_x/N	P_{exp}	Expected number (binomial) NP_{exp}
0	0	0.0	0.009	0.48
1	3	0.057	0.055	2.87
2	8	0.154	0.148	7.72
3	8	0.154	0.236	12.30
4	**20**	**0.385**	**0.247**	**12.86**
5	8	0.154	0.177	9.22
6	4	0.078	0.088	4.59
7	1	0.019	0.030	1.57
8	0	0.0	0.0067	0.35
9	0	0.0	0.0009	0.05
10	0	0.0	≈ 0	0.003
Sum	52	1.00	1.00	52

Our aim was not to fit the distribution or to determine the goodness of fit. But we can see that everything happens as to be expected – with the exception of the case of four 2-syllable words. Again we find (see row no. 4) that this event is far more frequent than expected by chance. Nearly half of the ten-word-chains contain exactly four two-syllable-words. This is deviating clearly from the expectable proportion.

To be sure that this result is not due to the fact of a too small sample:

Taking the theoretical probability from the binomial distribution as $P = 0.247$ and using it as the parameter p we calculate the probability for such a result to be created at random:

In two out of five chains ($n = 5$ and $x = 2$) a probability would result which is $P = 0.260$, so this could be possible.

Even in 5 out of 13 chains (a quarter of the sample) ($n = 13$ and $x = 5$) the probability would yield $P = 0.1222$ which is still conceivable.

But in 20 out of 52 chains we get $P = 0.0103$ which has to be considered as very improbable to be produced by chance.

Ten-word-chains with just 4 two-syllable-words are definitely preferred compared to the random situation ($P(X = 4) = 0.247$).

For every 10 words, a rather constant pattern can be found that cannot be explained by chance.

10. Conclusion

Length balance in a text can be proven and characterized by the measures of r_Λ and B.

The index of length homogeneity r_Λ indicates the degree of intercorrelation of the proportions of word lengths in parts of the text. In Proust's sentence, length homogeneity $r_\Lambda = 0.9777$ with t = 5 (individual limit).

We suppose that for every author (or text, sort of text, time, genre, style etc.) there is a style characteristic rhythm number B, so that in every B words the lengths of the resulting word-chains are balanced out and tend to be constant. In Proust's longest sentence, $B = 10$.

Further research should be done to corroborate the observation of B in any texts and to investigate the exceptional regularity of the two-syllable-words.

Length homogeneity r_A as a measure for the reliability of assessment in recording word length proportions should be determined, at least at split half level (t = 2), when investigating frequency distributions of word lengths in texts.

.

References

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

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

Altmann, G. & Best, K.-H. (1996). Zur Länge der Wörter in deutschen Texten. In: Schmidt, P. (ed.) *Glottometrika 15, 166-180*. Trier: Wissenschaftlicher Verlag Trier.

Altmann, G. & Schwibbe, M.H. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim: Olms.

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

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

Andersen, S. (2002). Freedom of choice and the psychological interpretation of word frequencies in texts. *Glottometrics 2, 45-52.*

Best, K.-H. (ed.) (1997). *The Distribution of Word and Sentence Length (= Glottometrika 16).* Trier: Wissenschaftlicher Verlag Trier.

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

Best, K.-H. (2003). *Quantitative Linguistik. Eine Annäherung.* Göttingen: Peust & Gutschmidt.

Botton, A. de (1997). *How Proust Can Change Your Life.* London: Picador Macmillan. (1998). *Wie Proust Ihr Leben verändern kann.* Frankfurt: S. Fischer.

Grzybek, P. (ed.) (2005). *Word length studies and related issues*. Boston/Dordrecht: Kluwer.

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

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

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

Wimmer, G. & Altmann, G. (1996). The theory of word length distribution: some results and generalizations. In: Schmidt, P. (ed.), *Glottometrika 15, 112-133*. Trier: Wissenschaftlicher Verlag Trier.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort.* Cambridge, Mass.: Addison-Wesley.

1. Appendix

I.

Proust's longest sentence from his work "A la recherche du temps perdu" in German translation
(detected by Alain de Botton 1997; 1998)

Diejenigen der alten Verdurinschen Möbel, die hier, manchmal sogar unter Beibehaltung einer bestimmten Anordnung, erneut Platz gefunden hatten und denen ich selbst in La Raspelière wiederbegegnet war, fügten in den gegenwärtigen Salon Teile des alten ein, die augenblicksweise mit nahezu halluzinatorischer Deutlichkeit jenen früheren noch einmal heraufbeschworen, gleich darauf aber fast unwirklich schienen, weil sie inmitten der umgebenden Wirklichkeit Bruchstücke einer untergegangenen Welt, die man an einem andern Orte wähnte, wiedererstehen ließen: ein aus Träumen entstiegenes Kanapee zwischen neuen, sehr wirklichen Sesseln, kleine, mit rosa Seide bezogene Stühle, eine durchwirkte Tischdecke auf dem Spieltisch, die zur Würde einer Person erhoben schien, denn wie eine Person besaß sie eine Vergangenheit, ein Gedächtnis, behielt sie doch im kalten Dunkel des Salons am Quai Conti jene Bräunung bei, welche die durch die Fenster der Rue Montalivet einfallende Sonnenstrahlung (deren genaue Stunde die Decke ebenso gut kannte wie Madame Verdurin selbst) bewirkt hatte, sowie die, die durch die Glasfenster der Gegend bei Deauville sich ergoß – wohin man jenes Requisit mitgenommen und wo es den ganzen Tag über den Blumengarten hinweg das tiefe Tal überschaut hatte in Erwartung der Stunden, da Cottard und der Geiger ihre Kartenspiele absolvieren würden – oder auch ein Strauß aus Veilchen und Stiefmütterchen in Pastell, Geschenk eines befreundeten großen Künstlers, der seither verstorben war, einziges hinterbliebenes Fragment eines Lebens, das sonst keine Spuren hinterlassen hatte; jetzt sprach nur dieses Bild noch – in ganz summarischen Zügen – von einem großen Talent und von einer langen Freundschaft, als einziges Überbleibsel erinnerte es noch an Elstirs sanften Blick, an die schöne, füllige und traurige Hand, mit der er immer gemalt hatte; ein gefälliges Durcheinander, eine Wirrnis aus Geschenken der Getreuen, die der Hausherrin überallhin gefolgt waren und schließlich die feste Prägung eines Charakterzuges, einer Schicksalslinie angenommen hatten, eine Fülle von Blumensträußen und Pralinenschachteln, die hier wie dort in einer ganz gleichen Art von üppigem Wachstum wuchernd sich entfalteten; eine merkwürdige Einsprengung aus sonderbaren und überflüssigen Objekten, jenen Dingen, die noch aussehen, als kommen sie eben erst aus der Verpackung hervor, in der sie als Geschenk überreicht worden sind, und die das ganze Leben hindurch bleiben, was sie zunächst gewesen sind, nämlich Geschenke zum 1. Januar, alle jene Gegenstände endlich, die man von den anderen nicht hätte trennen können, die aber für Brichot, den alten Besucher der Verdurinschen Feste, eine Patina und Weichheit bekommen hatten, wie sie Dingen eigen sind, denen ein geistiges Abbild ihrer selbst in unserem Innern eine Art von Tiefe hinzuzufügen scheint – alles dies ließ perlend in ihm jeweils Töne erwachen, welche in seinem Herzen geliebte Anklänge weckten: verworrene Erinnerungen, die gerade hier in diesem ganz und gar die Gegenwart verkörpernden Salon, indem sie vereinzelte Lichtflecke schufen – so wie an einem schönen Tage die Sonne im Viereck geradezu in die Atmosphäre eines Raumes hineingezeichnet – die Möbel und Teppiche gleichsam ausschnitten und mit einer Rahmenlinie umzogen, wobei sie von einem Kissen zu einer Blumenvase, einem Hokker zu einem noch lose anhaltenden Duft, einer Beleuchtungsart zu einem Vorherrschen bestimmter Farben hinübereilten und in plastischer und gleichzeitig beseelter Gestalt eine Form vor

Augen rückten, welche gleichsam die ideale, allen aufeinanderfolgenden Heimen anhaftende Urgestalt des Salons der Verdurins war.

Counting modalities:
French proper names were counted:
- syllable number by sound (Deauville = 2 syllables, Madame = 2 syllables)
- as entire word, if German translation would yield one (Rue Montalivet as "Montalivetstrasse" = 1 word, 5 syllables; Quai Conti as "Contiquai" oder "Contiufer" = 1 word, 3 syllables; La Raspeliere = 1 word, 5 syllables; but Madame Verdurin = 2 words, 2 and 3 syllables: "Frau Verdurin")

II.

52 ten-word-chains c with lengths of the words (number of syllables)

c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
4	4	1	2	3	3	6	2	2	1
1	2	1	2	2	2	1	1	2	1
2	3	1	1	3	1	1	1	1	2
4	3	5	2	1	1	1	2	2	1
2	2	5	1	2	3	1	4	2	1
1	1	1	1	5	1	2	3	4	2
1	3	2	5	1	4	2	2	2	2
2	2	1	1	2	3	2	2	2	2
2	1	1	3	2	3	2	1	3	3
2	2	5	7	1	3	5	3	3	1

c11	c12	c13	c14	c15	c16	c17	c18	c19	c20
1	2	2	4	1	1	2	1	1	2
1	1	2	4	2	3	3	4	2	2
2	1	1	2	3	1	4	2	1	1
2	1	2	3	1	2	1	1	2	1
2	2	1	2	2	1	1	2	1	1
1	2	1	1	2	2	1	1	1	1
2	1	1	2	2	1	1	3	2	2
4	2	2	3	1	2	2	2	2	1
1	1	1	1	1	2	1	1	4	4
3	3	5	2	1	1	2	3	4	1

c21	c22	c23	c24	c25	c26	c27	c28	c29	c30
2	3	2	2	1	1	2	1	1	2
2	5	1	1	3	1	2	3	2	1
2	2	1	2	4	2	2	1	2	4

4	2	1	2	4	3	1	1	2	1
2	2	2	2	1	1	4	3	5	5
2	1	1	1	1	3	4	4	2	1
1	1	1	1	1	1	2	2	4	1
2	2	1	2	2	1	2	2	4	1
3	2	1	2	2	1	1	1	2	1
1	4	4	2	1	1	3	2	2	1

c31	c32	c33	c34	c35	c36	c37	c38	c39	c40
2	2	1	3	1	3	2	1	2	1
1	4	1	2	1	1	1	2	3	2
2	3	3	1	1	2	1	1	1	1
1	1	1	1	2	3	1	2	2	3
1	4	2	1	2	1	1	1	3	2
3	1	1	1	2	2	3	2	2	2
2	5	2	2	2	3	1	3	1	1
2	3	1	3	1	2	2	1	1	1
1	2	1	2	1	2	2	4	2	3
4	2	1	1	2	4	2	2	2	2

c41	c42	c43	c44	c45	c46	c47	c48	c49	c50	c51	c52
2	1	3	1	2	2	1	1	2	3	2	7
1	1	2	1	1	4	3	1	1	2	1	2
1	2	4	1	1	1	2	2	2	5	1	4
2	2	5	3	1	1	3	2	4	1	2	3
5	3	1	4	2	4	1	1	1	1	2	1
1	2	3	2	2	2	1	2	2	3	2	2
2	1	1	2	2	2	2	4	4	1	2	1
1	2	1	1	1	5	4	2	1	3	1	3
1	2	2	4	2	1	3	2	2	3	3	1
2	3	1	3	1	2	2	1	3	2	2	1

III.

Variances of 52 ten-word-chains:

C1	1.211	C9	0.678	C17	1.067
C2	0.900	C10	0.489	C18	1.111
C3	3.567	C11	0.989	C19	1.333
C4	4.056	C12	0.489	C20	0.933
C5	1.511	C13	1.511	C21	0.767
C6	1.156	C14	1.156	C22	1.600
C7	3.122	C15	0.489	C23	0.944
C8	0.989	C16	0.489	C24	0.233

C25	1.556	C35	0.278	C45	0.278
C26	0.722	C36	0.900	C46	2.044
C27	1.122	C37	0.489	C47	1.067
C28	1.111	C38	0.989	C48	0.844
C29	1.600	C39	0.544	C49	1.289
C30	2.178	C40	0.622	C50	1.600
C31	0.989	C41	1.511	C51	0.400
C32	1.789	C42	0.544	C52	3.611
C33	0.489	C43	2.011		
C34	0.678	C44	1.511		

$\sum$ = 63.556 63.556 / 52 = 1.222

Glottometrics 11, 2005, 51 -55

Discrete distributions connected by partial summations[1]

Ján Mačutek, Bratislava[2]

Abstract. In this article it will be shown that (1) usual discrete probability (mass) distributions can be transformed in (almost) any partial sums distribution, (2) discrete probability distributions can be presented as partial sums distributions if an adequate transformations can be found.

Keywords: partial sums distributions

1. Introduction

In linguistic modeling one mostly starts from the assumption that self-regulation is based on the principle of proportionality between neighbouring frequency classes x and x+1. If a class with a greater value of x arises, then its "volume" depends on all existing classes. This fact can easily be expressed by the proportionality

(1) $P_x \sim P_{x\text{-}1}$

which is, as a matter of fact, a recurrence formula showing that class x depends on all classes 0,1,…,x-1. This relationship can be considered not only actual for modeling but also based on genetic assumptions: if X represents any type of complexity then genetically a simpler class existed before a more complex class arose. Usually, the proportionality is not constant but changes with x. Thus we obtain the familiar expression

(2) $P_x = g(x)P_{x\text{-}1}$

used both in synergetic linguistics and in general linguistic theory building (Köhler 1986, 2002; Altmann, Köhler 1996; Wimmer, Altmann 2005). However, for the zero-class the feedback is given only indirectly by the fact that the sum of the probabilities must be 1. A fact having no sense in empirical frequency distributions where sample size N is merely a sum of all frequencies. The proportionality function $g(x)$ is interpreted in most cases as a rational fraction $g(x) = s(x)/h(x)$, $s(x)$ representing the effect of the speaker (creativity, diversification), $h(x)$ that of the hearer or of the speech community and its task is the control the speakers creativity. The majority of language laws can be subsumed under (2) .

But further considerations have shown that if a frequency distribution of a linguistic property should represent a system, then a given class x need not depend only on the classes smaller than x but also on some combination of classes of another, parent distribution, that are greater and abide in turn by (2). This can be arrived at by a special summing of the classes of the parent distribution $\{P^*\}$, thus obtaining

---

[1] Acknowledgement: Supported by VEGA grant no. 1/0264/03.

[2] Address correspondence to: Ján Mačutek, e-mail: Jan.Macutek@fmph.uniba.sk

(3) $$P_x = g(x) \sum_{j=x+k}^{\infty} h(j) P_j^* .$$

According to the choice of k and $g(x)$ we obtain different types of partial sums distributions (in linguistics cf. Martindale et al. 1996; Wimmer, Šidlík, Altmann 1999; Wimmer, Altmann 2001a,b). In all cases they turned out to be adequate, especially in connection with ranking problems which are very common in linguistics.

Nevertheless, they present some problems because they are not easily systematized, i.e. embedded in a general theory. In the present article we try to show some of their properties and their relationship to usual simple distributions.

2. Connections between distributions

Let us consider the partial summation

(4) $$P_x = \sum_{j=x}^{\infty} g(j) P_j^* .$$

It is easy to see that the summation (4) is a special case of (3). This is not only a way how to create distributions for fitting data (see, e.g., Wimmer, Altmann 2001a). In fact, almost every pair of discrete distributions (the meaning of „almost every“ will be explained below) is connected by a proper partial summation. Let the distributions $\{P_x\}$ and $\{P_x^*\}$ be given and let

(5) $$P_x^* \neq 0 \text{ if } P_{x+1} - P_x \neq 0, \; x = 0,1,2,\ldots .$$

We put

$$P_x = \sum_{j=x}^{\infty} g(j) P_j^* ,$$

$$P_{x+1} = \sum_{j=x+1}^{\infty} g(j) P_j^*$$

and we obtain

$$P_x - P_{x+1} = g(x) P_x^* ,$$

which implies (if $P_x^* \neq 0$; in the opposite case also the left side of the previous equation is equal to zero and the value of the function $g(x)$ is irrelevant)

(6) $$g(x) = \frac{P_x - P_{x+1}}{P_x^*} .$$

Hence every distribution satisfying (5) is a parent distribution of every discrete distribution, which is true for both distributions different from its parent and identical with it (in other words, every distribution is a result of a partial summation). The function connecting them is

given by (6). The problem of finding invariant distributions (i.e., distributions which remain unchanged under (4)) is solved in general in Mačutek (2003).

3. Examples

1. In searching for distributions capturing the rank-frequency relation of phonemes/ graphemes (P. Grzybek, personal communication) it has been ascertained that the negative hypergeometric and the Whitworth distribution represent in the most cases the best fitting to data. Since they do not directly belong to common classes allowing us to embed them in the same higher model, we show at least the possibility of transformation.

The Whitworth distribution is defined by the summation

$$(7) \quad P_x = \frac{1}{n}\sum_{j=x}^{n}\frac{1}{j}, \quad x = 1,2,\ldots,n\,.$$

One can see that the summation above is a special case of (4) if $g(x) = \frac{1}{x}$ and the parent distribution $\left\{P_x^*\right\}$ is the discrete uniform distribution. However, the discrete uniform distribution defined on the set $\{1,2,\ldots,n\}$ can be considered as a special case of the zero truncated negative hypergeometric distribution. The question whether the Whitworth distribution can be obtained by a partial summation also from the zero truncated negative hypergeometric distribution can be easily answered using (6). The zero truncated negative hypergeometric distribution has the probability mass function

$$(8) \quad P_x^* = \frac{\binom{M+x-1}{x}\binom{K-M+n-x-1}{n-x}}{\binom{K+n-1}{n}-\binom{K-M+n-1}{n}}, \quad x = 1,2,\ldots,n.$$

We obtain

$$(9) \quad g(x) = \frac{P_x - P_{x+1}}{P_x^*} = \frac{\frac{1}{n}\sum_{j=x}^{n}\frac{1}{j} - \frac{1}{n}\sum_{j=x+1}^{n}\frac{1}{j}}{\frac{\binom{M+x-1}{x}\binom{K-M+n-x-1}{n-x}}{\binom{K+n-1}{n}-\binom{K-M+n-1}{n}}} = \frac{\binom{K+n-1}{n}-\binom{K-M+n-1}{n}}{nx\binom{M+x-1}{x}\binom{K-M+n-x-1}{n-x}}.$$

Hence the Whitworth distribution (7) is a result of the partial summation (4) if we take $g(x)$ from (9) and $\left\{P_x^*\right\}$ from (8) and insert them in (4). As a special case (for $K = 2$ and $M = 1$) we have $g(x) = \frac{1}{x}$ and $P_x^* = \frac{1}{n}$, $x = 1,2,\ldots,n.$

In the same way it can be shown that the relation holds also for the 1-displaced negative hypergeometric distribution.

2. As already mentioned, some distributions remain unchanged under (4), i.e. the result of partial summing does not change them (cf. Mačutek 2003). Let

(10) $P_{x+j} = 0, j = 1,2,\ldots$ if $P_x = 0$

and let $P_x^* = P_x, x = 0,1,\ldots,$ (i.e., the parent distribution and the result of a partial summation are identical), according to Mačutek (2003) we obtain

(11) $g(x) = 1 - \frac{P_{x+1}}{P_x}, x = 0,1,\ldots.$

If $\{P_x\}$ is, e.g., the Poisson distribution (i.e., $P_x = \frac{e^{-\lambda}\lambda^x}{x!}, \lambda > 0$), we have

$$g(x) = 1 - \frac{\frac{e^{-\lambda}\lambda^{x+1}}{(x+1)!}}{\frac{e^{-\lambda}\lambda^x}{x!}} = 1 - \frac{\lambda}{x+1} = \frac{x - \lambda + 1}{x+1},$$

which means that the Poisson distribution (with the parameter λ) is invariant with respect to the summation

(12) $P_x = \sum_{j=x}^{\infty} \frac{x - \lambda + 1}{x+1} P_x.$

3. Let $\{P_x\}$ be the geometric distribution. Inserting $P_x = p(1-p)^x$ in (11) we obtain

$$g(x) = 1 - \frac{p(1-p)^{x+1}}{p(1-p)^x} = 1 - (1-p) = p.$$

The geometric distribution remains unchanged under the partial summation

$$P_x = p\sum_{j=x}^{\infty} P_x.$$

This property of the geometric distribution was proved in Wimmer, Kalas (1999).

For several more examples see Mačutek (2003) (the function $g(x)$ is written in a slightly different form there and it can be derived without the constraint (10), i.e., for all discrete distributions). The function $g(x)$ from (11) is uniquely determined. Hence the invariances with respect to partial summations (4) are characteristic properties of discrete distributions (e.g., the Poisson distribution is the only distribution which remains unchanged under (12)).

References

Altmann, G., Köhler, R. (1996). „Language Forces“ and synergetic modeling of language phenomena. *Glottometrika 15, 62-76*

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

Köhler R. (ed.) (2002*). Korpuslinguistische Untersuchungen zur quantitativen und system-theoretischen Linguistik.* http://ubt.opus.hbz-nrw.de/volltexte/2004/279/

Mačutek J. (2003). On two types of partial summations. *Tatra Mountains Mathematical Publications 26, 403-410.*

Martindale, C., McKenzie, D., Gusein-Zade, S.M., Borodovsky, M.Y. (1996). Comparison of equations describing the frequency distribution of graphemes and phonemes. *J. of Quantitative Linguistics 3, 106-112.*

Wimmer, G., Altmann, G. (2001a). Models of rank-frequency distributions in language and music. In: Uhlířová, L. et. al (eds.) *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček: 10-20.* Trier: WVT.

Wimmer, G., Altmann, G. (2001b). A new type of partial-sums distributions. *Statistics and Probability Letters 52, 359-364.*

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In Altmann, G., Köhler, R., Piotrowski, R.G. (eds.), *Handbook of quantitative linguistics: 791-807*. Berlin: de Gruyter.

Wimmer, G., **Kalas, J.** (1999). A characterization of the geometric distribution. *Tatra Mountains Mathematical Publications 17, 325-329.*

Wimmer, G., Šidlík, P., Altmann, G. (1999). A new model of rank-frequency distribution. *J. of Quantitative Linguistics 6, 188-193.*

Glottometrics 11, 2005, 56-63

Turzismen im Deutschen

Karl-Heinz Best, Göttingen[1]

Abstract. Many authors examined the influence of loanwords in German (cf. Best 2001; Körner 2004). This above all applies to borrowings from French, English, and Latin, and for some other languages as well (for ex. Greek, Italian, Spanish). But there are many languages like Turkish the influence of which on the German lexicon is nearly unknown. The present paper presents the development of Turkish borrowings in German and demonstrates that this process abides by the logistic law which in linguistics is known as Piotrowski Law.

Keywords: borrowings, Turkish, German, Piotrowski law

Das logistische Gesetz als Modell für Entlehnungsprozesse

Altmann (1983), Altmann u.a. (1983) sowie Best & Altmann (1986) haben die Hypothese entwickelt, dass Sprachwandel generell und damit auch Entlehnungsprozesse einem Sprachgesetz des logistischen Typs entsprechend verlaufen sollten. Hinzu kam der sog. „reversible" Sprachwandel, der auf die gleiche Weise einsetzt, dann aber irgendwann wieder zurückgenommen wird. Eine ganze Reihe von Untersuchungen haben gezeigt, dass dieses Modell in seinen unterschiedlichen Formen sich bewährt, wenn nur ausreichend Daten verfügbar sind (Best 2003b; Körner 2004; und viele andere mehr). Der Verlauf der Entlehnung von Turzismen konnte mangels geeigneter Daten noch nicht getestet werden.

Bisher wurden solche Prozesse daraufhin untersucht, wie viele Wörter aus einer bestimmten Sprache oder Sprachgruppe direkt ins Deutsche gelangten. Dies betrifft eine ganze Reihe türkischer Wörter, darunter Joghurt, Kaftan und Pascha. Solche unmittelbar aus dem Türkischen stammenden Wörter erreichen in Wörterbüchern, die den Zeitpunkt der Übernahme ins Deutsche datieren, bei nur selten mehr als 20000 Stichwörtern lediglich einen Anteil von 0.15% (Duden. Herkunftswörterbuch ³2001) bzw. 0.06% (Duden. Etymologie 1963) und in Deutsches Fremdwörterbuch (1913-1988) von 0.18% aller Übernahmen (Best 2001; Körner 2004); die damit verbundene Zahl der Belege ist zu gering, um die Gesetzmäßigkeit des Übernahmeprozesses zu testen. Zu diesem Zweck muss die Datenbasis also erweitert werden. Um das zu erreichen, wurden für diese Untersuchung alle Wörter als Turzismen betrachtet, die entweder aus dem Türkischen stammen oder doch wenigstens über das Türkische und oft weitere Sprachen (bes. romanische und slawische) ins Deutsche gelangten. So erhält man genügend Belege, um das Piotrowski-Gesetz noch einmal zu prüfen. Das Vorgehen ist damit weitgehend dasselbe wie im Falle der Arabismen (Best 2004).

[1] Address correspondence to: K.-H. Best, e-mail: kbest@gwdg.de

Wörter türkischer Herkunft im Deutschen

Grundlage der Datengewinnung sind Wörterbücher des Deutschen, die auf CD ROM vorliegen und damit eine Computerrecherche ermöglichen: *Duden. Die deutsche Rechtschreibung* ([22]2000), *Duden. Das Fremdwörterbuch* ([8]2005), *Duden. Das große Wörterbuch der deutschen Sprache in 10 Bänden* ([3]1999) und *Kluge* ([24]2002). Alle Wörter, die nach Auskunft eines dieser Wörterbücher als Turzismen zu betrachten sind, wurden in eine Liste aufgenommen, möglichst mit Angabe zum Zeitpunkt der Übernahme ins Deutsche. Auf diese Wiese kam ein Bestand von genau 158 Wörtern türkischer Herkunft zustande, von denen 39 datierbar sind. Die Zeitangaben stützen sich in fast immer auf *Kluge* ([24]2002). In allen Fällen wurden auch *Duden. Herkunftswörterbuch* ([3]2001) und *Pfeifer* ([Ltg.] [2]1993) verglichen, sofern das möglich war. Zu zwei Wörtern findet man in *Duden. Das große Wörterbuch der deutschen Sprache in 10 Bänden* ([3]1999) Hinweise auf Kunstwerke, in denen der entsprechende Ausdruck vorkommt. Damit ist wahrscheinlich nicht der Zeitpunkt gewonnen, zu dem das entsprechende Wort ins Deutsche übernommen wurde; man kann aber immerhin sagen, dass es seitdem bekannt ist. Alle ermittelten Turzismen sind in Tabelle 1 aufgelistet. Diese enthält in der ersten Spalte das Stichwort, in der nächsten Spalte - soweit möglich - das Jahrhundert der Übernahme ins Deutsche, ferner eine sehr grobe Bedeutungsangabe und zuletzt die Entlehnungsgeschichte des Wortes. Speziell zur Entlehnungsgeschichte wurde in einigen Fällen auch *Duden. Das große Fremdwörterbuch* ([3]2003) konsultiert.

Zur Aufnahme in die Liste der Turzismen: Es wurden auch Eigennamen aufgenommen, wenn sie als Bezeichnungen für Gegenstände (Kopfbedeckungen und Teppiche) dienten. Fragezeichen in der Tabelle weisen auf Angaben hin, die in dem jeweils berücksichtigten Wörterbuch als unsicher charakterisiert sind. Weitere Widersprüche sind nicht markiert. So nimmt *Kluge* ([24]2002) für *Kaviar* griechische Herkunft an, während andere Wörterbücher es als „türkisch" charakterisieren. *Kaviar* wurde in die Liste aufgenommen, da es sich dabei möglicherweise um einen Turzismus handelt. Manche Angaben werden sicherlich zu korrigieren sein, wenn die Wortforschung zu verbesserten Einsichten gekommen sein wird.

Duden. Das Große Wörterbuch der deutschen Sprache in 10 Bänden enthält laut Werbung über 200000 Stichwörter; darauf bezogen machen 158 Wörter türkischer Herkunft knapp 0.08% aus. Als datierbar erwiesen sich gar nur 0.02%.

Tabelle 1
Turzismen im Deutschen

Entlehnung	Jhd.	Bedeutungshinweis	Entlehnungsweg
Aga		Herr; früher: Titel	türk.
Altin		alte russische Kupfermünze	türk.
Anatol		Teppich (n. Region)	türk.
Angora-		nach altem Namen von Ankara	türk.
Atabeg		ehemaliger Titel für Emire	türk.
Baba		Vater; Ehrentitel	türk.
Bairam		Name islamischer Feste	türk.
Baklava		Strudelgebäck	türk.
balkanisieren		staatlich zersplittern	lat. - türk.
Balkanologe		Wissenschaftler	griech./ türk.
Ban		Statthalter	serbo-kroat. - türk.
Beg		höherer Titel	türk.
Beglerbeg		Provinzstatthalter	türk.
Begum		Titel ind. Fürstinnen	engl. - urdu - türk.

Bei		wie Beg	türk.
Bergamotte	18.	eine Birnenart	frz. - it. - türk.
Bektaschi		Angehöriger eines Derwischordens	türk.
Besemer/ Desem(er)	13.	Handschnellwaage	ndd. - russ. - türk.?
Bursa		ein Teppich (n. Stadt)	türk.
Busuki		Lauteninstrument	griech. - türk.
Café	19.	Kaffeehaus	frz. - it. - türk. - arab.
Cafeteria	20.	Selbstbedienungsrestaurant	amerik.-span. - it. - türk. - arab.
Cafetier		Kaffeehausbesitzer	frz. - it. - türk. - arab.
Chagrin		Leder mit Narbenmuster	frz. - türk.
chagrinieren		Narbenmuster aufpressen	frz. - türk.
Defterdar[2]	18.	Schatzmeister	türk. - pers.
Derwisch	16.	Bettelmönch	türk. - pers.
Diwan	17.	Liegesofa	frz. - türk. - pers.
Dolma		türkisches Nationalgericht	türk.
Dolman		Überrock, Jacke	ung. - türk.
Dolmetscher	13.	Übersetzer	ung.?/russ.? - türk.
Döner	20.	Kurzwort zu Dönerkebab	türk.
Dönerkebab	20.	Kebab vom Drehspieß	türk.
Dudelsack	17.	„Dudel“-	poln./tschech. - türk.
Efendi/ Effendi		Titel und Anrede	türk. - griech.
Entari		Gewand	türk.
Ferman		Erlass	türk. - pers.
Fes/ Fez	19.	Kopfbedeckung (n. Stadt)	türk.
Gajda		türkische Sackpfeife	türk. - span.
Ghasi/ Gazi		Ehrentitel türkischer Herrscher	türk. - arab.
Giaur		Ungläubiger	türk. - pers. - arab.?
Gilet		Weste	frz. - span. - arab. -türk.
Hadschi/ Haddschi		Pilger	türk. - arab.
Hamam		türk. Bad	türk. - arab.
Hanum		Höflichkeitsanrede zu Frauen	türk./ pers.
Harem	18.	Frauenräume, Frauen	türk. - arab.
Hereke		Teppich (n. Ort)	türk.
Hodscha		geistlicher Lehrer	türk.
Horde	15.	Menge, Schar	poln. - türk. - tat.
Ilchan		Titel mongolischer Herrscher	türk. - mongol.
Irade		Erlass	türk. - arab.
Janitschar		Soldat	türk.
Jaschmak		Schleier	türk.
Jastik/ Yastik		kleiner Teppich	türk.
Jirmilik		türk. Silbermünze	türk.
Jatagan		Säbel	türk.
Joghurt/ Jogurt	20.	Dickmilch	türk.
Jurte	17.	Nomadenzelt	russ. - türk.
Jürük/ Yürük		Teppich	türk.
Kaffee	17.	Getränk	frz. - it. - türk. - arab.
Kaftan	16.	Obergewand	türk. - arab. - pers.

[2] In Lessing, Nathan der Weise.

Kalpak/ Kolpak		Mütze	türk. - tat.
Kantschu		Riemenpeitsche	slaw. - türk.
Kapu		Amtsgebäude	türk.
Karaburan		Sommersandsturm	türk.
Karagös		Figur in Schauspiel	türk.
Karakal		Wüstenluchs	türk.
Karaman		Teppich (n. Stadt)	türk.
Karbatsche		Riemenpeitsche	tschech. - ung. - türk.
Kaviar	16.	gesalzener Rogen	it. - türk. - pers.
Kawass/ Kawasse		Wächter, Bote	türk. - arab.
Kayseri		kleiner Teppich (n. Stadt)	türk.
Kebab/ Kebap	20.	Spießgebratenes	türk. - arab.
Kelek/ Kelik		Floß aus Häuten	türk. - pers.
Kelim/ Kilim		Teppich	türk. - pers.
Khan		Herrschertitel	türk. - mongol.
Khanat		Amt, Land eines Khans	lat. - türk.
Khedive		Titel des Vizekönigs v. Ägypt.	türk. - pers.
Kiosk	18.	Verkaufsstelle	frz. - türk. - pers.
Kismet	19.	Schicksal	türk. - arab.
Koffein/ Kaffein	19.	Inhaltsstoff des Kaffes	lat. - engl. - türk. arab.
Köfte		Hackfleischbällchen	türk.
Konak		Amtsgebäude	türk.
Konya		Teppich (n. Stadt)	türk.
Kula		Gebetsteppich (n. Ort)	türk.
Kurgan		Hügelgrab	russ. - türk.
Kuruş		Geldeinheit (Groschen)	türk. - it./dt.
Kuvasz		Hirtenhund	ung. - türk.
Ladik		Gebetsteppich (n. Ort)	türk.
Liman		lagunenartiges Gewässer	russ. - türk. - griech.
Lira		türk. Währungseinheit	türk. - it. - lat.
Makramee		Knüpftechnik, -arbeit	it. - türk. - arab.
Medrese/ Medresse		islamische Hochschule	türk. - arab.
Minarett	18.	Turm einer Moschee	frz. - türk. - arab.
Muchtar		Ortsvorsteher	türk. - arab.
Mudir		Verwaltungsleiter	türk. - arab.
Mulla/ Mullah		Rechts-/ Religionslehrer	türk. - pers. - arab.
Muschir/ Müschir		hoher Beamter	türk. - arab.
Muselman(n)	17.	Moslem	it./frz. - türk. - pers. - arab.
Nahie/ Nahije		Verwaltungsbezirk	türk. - arab.
Namas/ Namaz		Stundengebet	türk. - pers. - sanskr.
Odaliske		weiße türkische Haremssklavin	frz. - türk.
Okka		Gewicht	türk.-arab.- aram. - griech. - lat.
Ottoman		Ripsgewebe	frz. - türk.
Pallasch		Säbel	slaw. - ung. - türk.
Panderma		Gebetsteppich (n. Ort)	türk.
Para		jugoslawische Währungseinheit	serb. - türk.
Pascha	18.	Titel	türk. - pers.
Paschalik		Amtsbezirk	türk.

Perkal		Baumwollstoff	frz. - türk. pers.
Perkalin		Baumwollgewebe	lat. - frz. - türk. - pers.
Pilau/ Pilaw		Reisgericht	türk. - pers.
Rajah		Untertan, Verwaltungsbezirk	türk. - arab.
Raki		Branntwein	türk. - arab.
Ramasan		Ramadan	türk./pers.
Saffian	17.	Ziegenleder	russ. - türk. - pers.
Saki		Mundschenkfigur in Dichtung	türk./ pers. - arab.
Sandal		türk. Boot	türk. - arab. -pers.
Sandschak		Standarte, Regierungsbezirk	türk.
Schabracke	17.	verzierte Decke	ung.? - türk.
Schakal	17.	Raubtier	frz. - türk. - pers. - altind.
Schalwar		weite Hose	türk./ pers.
Scheck	19.	Zahlungsmittel	engl. - arab.? - türk.? - pers.
Selamlik		Empfangsraum	türk.
Serail[3]	18.	Palast/ Schloss	frz. - it./türk. - pers.
Sinopie		Vorzeichnung (n. Stadt)	türk.
Sivas		Teppich (n. Stadt)	türk.
Sofa	17.	gepolstertes Sitzmöbel	frz./it. - türk. - arab.
Softa		Student	türk. - pers.
Sorbet(t)/ Scherbett	17.	Halbgefrorenes	frz./it./span. - türk. - arab.
Spahi		Reitersoldat	frz. - türk. - pers.
Taft	15.	Seidengewebe, Futterstoff	it. - türk. - pers.
Tarbusch		Fes	frz. - arab. - türk./pers.
Tefsir		Koranauslegung	türk. - arab.
Tsatsiki/ Zaziki		dickflüssige Soße	türk./ griech.
Tschausch		Polizist	türk.
Tschibuk		Tabakpfeife	türk.
Tschiftlik		Landgut	türk.
Tugh		Art von Ehrenzeichen	türk.
Tughra		Namenszug des Sultans	türk.
Tulipan	16.	Tulpe	lat./it. - frz. - türk./pers.
Tulpe	17.	s. Tulipan	ndl. - türk./pers.
Turban	17.	Kopfbedeckung	it. - türk. - pers.
Türbe		Grabbau	türk. - arab.
türkis	15.	blaugrün	frz. - türk.
turkisieren		türkisch machen	lat. - türk.
Turko		farbiger frz. Fußsoldat	frz. - it. - türk.
Turkologie		Wissenschaft	türk./ griech.
Turzismus		türk. Spracheigentümlichkeit	lat. - türk.
Ulan	18.	Lanzenreiter	poln. - türk.
Ulema		Rechts- und Religionsgelehrter	türk. - arab.
Uschak		Teppich (n. Stadt)	türk.
Wali/ Weli		Statthalter	türk. - arab.
Walide		Mutter des Sultans	türk. - arab.?
Wekil		Minister	türk. - arab.
Wesir		Großwesir, Minister	türk. - arab.

[3] Im Titel eines Singspiels von W.A. Mozart

Wilajet		Verwaltungsbezirk	türk. - arab.
Zurna		eine Art von Oboe	türk. - pers.

In der folgenden Tabelle 2 sind die 39 datierbaren Turzismen als „beobachtet“ aufgeführt, getrennt nach den Jahrhunderten ihrer Übernahme. Die beobachteten Werte wurden für die Untersuchung zusätzlich in kumulierte Werte überführt. Tabelle 2 beginnt mit dem 13. Jahrhundert; aus dieser Zeit stammen die ältesten datierten Belege.

Als Nächstes wurde mit einem entsprechenden Programm in NLREG geprüft, ob auch der Zuwachs von Wörtern türkischer Herkunft dem logistischen Gesetz

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

entspricht. Das Ergebnis findet sich in der Tabelle 2 unter „berechnet“. Es handelt sich um die Werte, die man erhält, wenn man die Formel (1) an die kumulierten Werte anpasst. a, b und c sind die Parameter des Modells; c gibt den Zielwert an, auf den nach der Berechnung der Prozess hinausläuft. D ist der Determinationskoeffizient, der höchstens den Wert 1 erreichen kann. Das Ergebnis ist hervorragend, wie der Testwert $D = 0.9950$ und die folgende Graphik (Abb. 1) zeigen.

Tabelle 2
Entwicklung der Turzismen im Deutschen

Jhd.	t	beobachtet	kumuliert	berechnet
13.	1	2	2	0.74
14.	2	0	2	1.91
15.	3	3	5	4.73
16.	4	4	9	10.57
17.	5	12	21	19.65
18.	6	8	29	28.95
19.	7	5	34	35.18
20.	8	5	39	38.27
$a = 143.6958$ $b = 0.9828$ $c = 40.3824$ $D = 0.9950$				

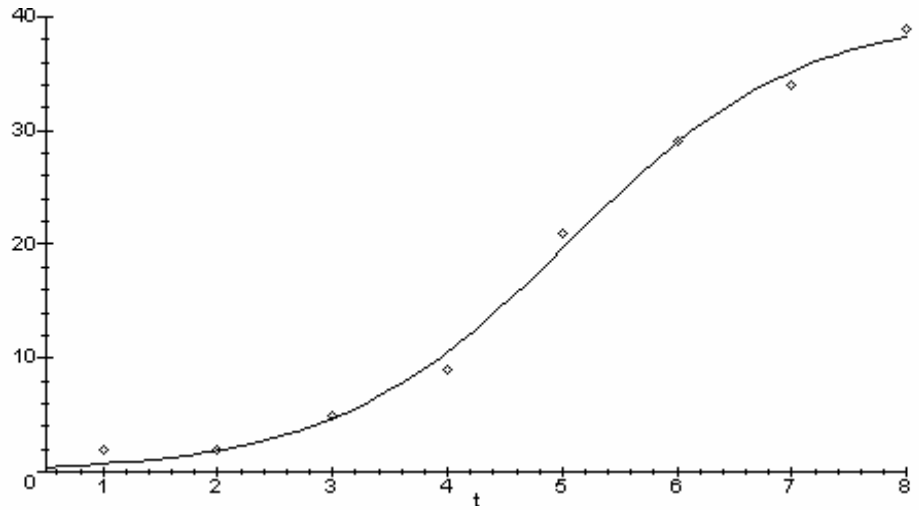


Abb.1 Die Entwicklung der Turzismen im Deutschen (In dieser Graphik steht $t = 1$ für die Entlehnungen des 13. Jahrhunderts, $t = 2$ für die des 14. Jahrhunderts; etc.)

Probleme

Die hier vorgelegte Darstellung ist in verschiedener Hinsicht problematisch: Als Turzismen wurden Wörter aufgefasst, die aus dem Türkischen stammen oder über das Türkische und ggfs. über weitere Sprachen ins Deutsche gelangten. Nur die Turzismen wurden aufgenommen, die in einem der genannten Wörterbücher als solche aufgeführt sind. Eigenmächtige Entscheidungen über die Zugehörigkeit eines Wortes zum deutschen Wortschatz und seine Entlehnungsgeschichte sollten durch diese Vorsichtsmaßnahme vermieden werden. Damit wird in Kauf genommen, dass womöglich das eine oder andere Wort zwar bereits weite Vebreitung gefunden hat, aber noch nicht in den Wörterbüchern dokumentiert ist. Eine ähnliche Vorsichtsmaßnahme gilt für die chronologische Einordnung der Wörter: Es wurden Datierungen nur dann vorgenommen, wenn sie in wenigstens einer der konsultierten Quellen angegeben waren.

Die Angaben der benutzten Handbücher weichen in einigen Fällen erheblich voneinander ab (vgl. *Kaviar*). Die Entscheidung für eine dieser Angaben ist in gewissem Sinne willkürlich. Ziel war dabei, möglichst vollständige Angaben zur Entlehnungsgeschichte und Datierung zu geben.

Um diese Bemerkungen zu konkretisieren: Obwohl Gastronomie und Handel der Türken in Deutschland in den letzten Jahrzehnten erheblich an Bedeutung gewonnen haben, findet man kaum etwas von diesem Wortschatz in den Wörterbüchern. Die oben angeführte Tabelle lässt sich mit jeder Speisekarte eines türkischen Restaurants ergänzen. Ähnliches mag für andere Lebensbereiche gelten. Dieser Wortschatz fehlt entsprechend auch in dieser Untersuchung.

Schlussbemerkung

Nach Auswertung von *Deutsches Fremdwörterbuch* (Kirkness [Hrsg.] 1988) ist das Türkische unter insgesamt 35 Sprachen, aus denen das Deutsche Wörter entlehnt hat, an 10. Stelle platziert (Best 2001: 14); nach *Duden. Herkunftswörterbuch* (^{3}2001) an 14. Stelle unter 32 „Geber"-Sprachen (Körner 2004: 30). Bei diesen Angaben handelt es sich jedoch nur um die Direktentlehnungen aus dem Türkischen.

Eine deutliche Veränderung des langfristigen Trends der Übernahme türkischer Wörter ins Deutsche, die ja aufgrund der Zuwanderung von türkischen Arbeitnehmern nach dem letzten Weltkrieg denkbar, wenn nicht erwartbar wäre, ist derzeit noch nicht erkennbar. Vielleicht ändert sich dies aber schon etwas mit der nächsten Wörterbuch-Generation. Diese einstweilige Ungewissheit führt jedoch zu der Frage, ob man den bisher entdeckten Trend mit Aussicht auf Erfolg für Prognosezwecke nutzen kann. Probleme von Prognosemöglichkeiten wurden in ähnlichen Zusammenhängen in Best (2003a: 20; 2003c; 2005) thematisiert.

Die Untersuchung hat ergeben, dass die bisher erfassten Entlehnungen aus dem Türkischen ins Deutsche dem sog. Piotrowski-Gesetz in der Form des unvollständigen Sprachwandels (Formel 1) folgen. Dieses Gesetz wird damit einmal mehr als ein sehr valides Modell für jegliche Art von Sprachwandel bestätigt.

Literatur

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

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

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

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

Best, Karl-Heinz (2003b). Spracherwerb, Sprachwandel und Wortschatzwachstum in Texten. Zur Reichweite des Piotrowski-Gesetzes. *Glottometrics 6, 9-34.*

Best, Karl-Heinz (2003c). Slawische Entlehnungen im Deutschen. In: Kempgen, Sebastian, Schweier, Ulrich, & Berger, Tilman (Hrsg.), *Rusistika · Slavistika · Lingvistika. Festschrift für Werner Lehfeldt: 464-473.* München: Vlg. Otto Sagner.

Best, Karl-Heinz (2004). Zur Ausbreitung von Wörtern arabischer Herkunft im Deutschen. *Glottometrics 8, 75-78.*

Best, Karl-Heinz (2005). Sind Prognosen in der Linguistik möglich? *Typen von Wissen. Beiträge zum 4. Kolloquium Transferwissenschaften in Halle, 4.-6- Oktober 2004.* In: Antos, Gerd, & Weber, Tilo (Hrsg.). Frankfurt/M.: Lang (erscheint).

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

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

Duden. Die deutsche Rechtschreibung. (^{22}2000). 22., völlig neu bearbeitete und erweiterte Auflage. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag.

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

Duden. Fremdwörterbuch. (^{8}2005). 8., neu bearbeitete und erweiterte Auflage. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag.

Duden. Das große Fremdwörterbuch. Herkunft und Bedeutung der Fremdwörter. (^{3}2003). 3., überarbeitete Auflage. Mannheim/ Wien/ Zürich: Dudenverlag.

Duden. Herkunftswörterbuch (^{3}2001). 3., völlig neu bearbeitete und erweiterte Auflage. Mannheim/ Wien/ Zürich: Dudenverlag.

Duden. Das große Wörterbuch der deutschen Sprache in 10 Bänden. (^{3}1999). 3., völlig neu bearbeitete und erweiterte Auflage. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag.

Kluge. Etymologisches Wörterbuch der deutschen Sprache. (^{24}2002). Bearb. v. Elmar Seebold. 24., durchgesehene und erweiterte Auflage. Berlin/ New York: de Gruyter.

Körner, Helle (2004). Zur Entwicklung des deutschen (Lehn-)Wortschatzes. *Glottometrics 7, 25-49.*

Pfeifer, Wolfgang [Ltg.] (^{2}1993/1995). *Etymologisches Wörterbuch des Deutschen.* München: dtv.

Verwendete Software

MAPLE V Release 4. 1996. Berlin u.a.: Springer.
NLREG. Nonlinear Regression Analysis Program. Ph.H. Sherrod. Copyright (c) 1991-2001.

Information im Internet zum „Göttinger Projekt zur Quantitativen Linguistik":
http://wwwuser.gwdg.de/~kbest.

Glottometrics 11, 2005, 64-97

Die statistische Analyse des semantischen Feldes der Farbbezeichnungen im Deutschen

Sergej Kantemir, Viktor Levickij[1]

Abstract. The syntagmatic and paradigmatic characteristics of colour terms in German were investigated. The use of statistical methods enabled by means of formal criteria to objectively reveal similarities and dissimilarities of syntagmatic and paradigmatic peculiarities of colour terms. The findings of the research are:

1. In modern German the colour characteristics of different things and phenomena are defined via diverse system of colour terms that is comprised of a number of frequent groups of words, and where the most frequently used colour terms such as *weiß, schwarz, rot, blau, grün, grau, braun* and *gelb* (67,77%) make up the nucleus of the semantic field, the rest of the words (32,23%) belong to its periphery.
2. By means of the basic colour terms modern German can structure colour perception of an individual and "verbalise" the semantic structure of colour phenomenon in national *linguistic image of the world.*

Keywords: colour terms, lexical semantics, compatibility, paradigmatic relations, lexical microsystem, semantic fields, statistic methods, corelation analysis

1. Ziele

Der Untersuchung der Farbbezeichnungen in verschiedenen Sprachen sind viele Arbeiten gewidmet. Die Forschungen werden mit unterschiedlicher Zielsetzung, von vielfältigen Blickwinkeln aus sowie anhand verschiedenster methodischer Ansätze und Arbeitsweisen durchgeführt. So wurden die Farbwörter von A. Wierzbicka (Wierzbicka 1996) im Rahmen der kognitiven Linguistik und von B. Berlin & P. Kay (1969), R. Frumkina (1984) etc. aus psycholinguistischer Sicht behandelt. Es wurden auch Methoden der strukturellen Linguistik (Wortfeldtheorie von J. Trier) in Arbeiten von L. Weisgerber (1962) und S. Wyler (1992) verwendet und ihre Häufigkeit wurde von A. Pawlowski (1999) untersucht.

Als eine der ersten Arbeiten, in denen statistische Ansätze zur Farbwortuntersuchung verfolgt worden sind, gilt das Buch von V. Moskovič „Semantika i statistika" (1969). Unter Hinzuziehung der Korrelationsanalyse und einiger anderen statistischen Kriterien und Koeffizienten untersuchte V. Moskovič die Gebrauchshäufigkeit der Farbadjektive in Texten, die Gebrauchshäufigkeit verschiedener Bedeutungen dieser Adjektive sowie den Grad ihrer semantischen Nähe anhand der Ähnlichkeit ihrer Kombinierbarkeit mit den Nomen im Russischen, Englischen und Französischen.

Aufgabe der vorliegenden Arbeit wird es sein, die syntagmatischen und paradigmatischen Eigenschaften und Beziehungen der Farbadjektive im Deutschen mithilfe statistischer Ansätze zu erschließen.

[1] Address correspondence to: Viktor Levickij, Radiščeva Str. 6/5, UA-58000 Tscherniwzi; Sergej Kantemir, Entusiastiv Str. 5/360, UA-58032 Tscherniwzi. E-mail: kantuzzi@web.de

2. Materialien

Als Forschungsgegenstand diente eine Stichprobe, die anhand von 33 modernen deutschen Prosawerken von W. Borchert, H. Fallada, L. Feuchtwanger, G. Grass, E. Heller, S. Lenz, H. Lind, T. Mann etc. (insgesamt drei Millionen lexikalischen Einheiten) durchgeführt worden ist. Der relative Fehler der Stichprobe δ beträgt 3,3%. Darüber hinaus wurden einige Wörterbücher der deutschen Sprache (Duden, Knaur, Wahrig) zur Bestandsaufnahme der Farbbezeichnungen verwendet.

3. Bestandsaufnahme der Farbadjektive

Bevor die Gebrauchshäufigkeit von Farbwörtern in Texten untersucht werden konnte, musste eine Prozedur erarbeitet werden, die in der modernen Semasiologie unter Bezeichnung *Inventarisierung der Gruppe* (s. Levickij 1989) bekannt ist. Dieses Verfahren besteht darin, dass der Forscher mithilfe einer bestimmten Methodik ein Register von Wörtern erstellen soll, das einer weiteren statistischen Bearbeitung unterzogen wird.

Die Bestandsaufnahme der zu untersuchenden Gruppe ist von uns in einigen Etappen durchgeführt worden.

In der ersten Etappe wurde aus sechs Wörterbüchern das Anfangsregister der Farbbezeichnungen (s. Anlage 1) durch eine allgemeine ununterbrochene Auswahl und Komponentenanalyse zusammengestellt.

In der zweiten Etappe wurde die Gebrauchshäufigkeit der ins Anfangsregister eingetragenen Adjektive in Texten analysiert (s. Anlage 2). Mit Hilfe der erwähnten Verfahrensweise wurde die Gruppe der hochfrequenten Farbadjektive (s. Tab. 1) bestimmt. Gerade diese Adjektive kann man wohl für die Grundfarbwörter der deutschen Gegenwartssprache halten.

Tabelle 1
Gebrauchshäufigkeit der Farbadjektive

Ausgewählte Farbwörter	Gebrauchshäufigkeit im Text	Prozentual
weiß	569	15,75%
schwarz	551	15,25%
rot	349	9,66%
blau	288	7,97%
grün	206	5,7%
grau	201	5,56%
braun	148	4,09%
gelb	136	3,77%
sonstige	1164	32,23%
Insgesamt	3612	100%

4. Kombinierbarkeit des Wortes

Als syntagmatische Eigenschaft des Wortes wird seine Kombinierbarkeit bezeichnet; darunter versteht man in weitem Sinne seine Fähigkeit, eine Verbindung mit anderen Wörtern nach entsprechenden Modellen im Text einzugehen.

Moderne linguistische Studien zur Wortverbindung (s. Amosova 1963: 36-41, Apresjan 1969: 81) zeigen, dass die Kombinierbarkeit auf drei Ebenen behandelt werden kann:

1) *auf der Ebene der Wortklasse (morphologische Klasse)* – tritt eine syntaktische Kombinierbarkeit auf als eine syntaktische Verbindung mit anderen Wörtern durch nähere Bestimmung von nominalen Satzgliedern (dies ist aufgrund der Distributionsformeln leicht nachzuweisen);

2) *auf der Ebene der Wortsubklasse*: Eine semantische Kombinierbarkeit als eine Verbindung von Lexemen aufgrund einer gemeinsamen semantischen Komponente;

3) *auf der Ebene eines Einzelwortes*: Eine lexikalische Kombinierbarkeit als (im engen Sinne) die Fähigkeit des Wortes, mit dem zum begrenzten Register gehörenden Einzelwort eine Kollokation einzugehen.

Als semantisch realisierte Einheit tritt in jedem Fall natürlich ein Wort auf (Levickij 1989: 38).

Es sollte auch hervorgehoben werden, dass die Kombinierbarkeit des Wortes mit dessen Bedeutung eng verbunden ist. So sind einige Linguisten der Auffassung, dass die Bedeutung des Wortes seine potenzielle Kombinierbarkeit sei (Zvegincev 1957: 132). Dabei ist auch zu vermuten, dass die Semantik eines Adjektivs in attributiver (manchmal prädikativer) Verwendung im Satz durch die Semantik eines entsprechenden Substantivs wesentlich beeinflusst wird. Kommen mehrere Adjektive als Attribute zu demselben Nomen in einem attributiven Syntagma vor, so existiert zwischen diesen Adjektiven eine bestimmte semantische Beziehung. Der Grad einer solchen Beziehung kann durch spezielle statistische Verfahren festgestellt werden, falls quantitative Charakteristiken der Kombinierbarkeit der zu erforschenden Adjektive im Text bekannt sind.

5. Kombinierbarkeitsweite und Vorkommenshäufigkeit des Wortes

Wie oben (4) gezeigt wurde, kann die Kombinierbarkeit nicht nur im Modell *[Wort + Wort]* (lexikalische Kombinierbarkeit), sondern auch im Modell *[Wort + Wortsubklasse]* (semantische Kombinierbarkeit) untersucht werden.

Zur Erschließung der Kombinierbarkeit von Adjektiven im Modell *[Wort + Wortsubklasse]* muss vorher ein Inventar von Subklassen der Substantive aufgestellt werden, die unter einem bestimmten semantischen Merkmal zusammengefasst sind. Obwohl derzeit keine objektiven Klassifizierungstechniken lexikalischer Einheiten aufgrund formaler Kriterien bestehen, gibt es doch in der modernen Sprachwissenschaft gewisse Empfehlungen bezüglich der Klassifikation gerade solcher Einheiten (s. Merten 1983).

Unter Hinzuziehung von Erfahrungen einiger Forscher (s. Levickij 1989: 135-136; Levickij/Ogoui u.a. 2001: 52-53) haben wir die eigene semantische Klassifikation der Substantive erarbeitet, die entsprechend den Ergebnissen unserer Stichprobe den Charakter der kontextuellen Umgebung der Farbadjektive im Deutschen in vollem Umfang darstellt.

Folglich sind alle Substantive des aufgestellten Registers, die in Wortverbindungen mit den Farbwörtern fixiert sind, zu folgenden Subklassen (Clustern) zusammengefasst wurden:

1) **das Äußere des Menschen, innere Organe und Körperteile:** *die Augen*; *das Gesicht*; *die Galle*; *die Haare*; *die Zähne*;
2) **Eigennamen, Vornamen und Pronomen:** *Amanda, Rufus, Minna, ich, er*.
3) **Personen:** *der Mann*; *die Brüder*; *der Kerl*; *der Junge*; *das Mütterchen*;
4) **einen Sozialstatus des Menschen oder eine Organisation:** *der Hauptmann*; *der Doktor*; *der Verkehrsschutzmann*; *die Regierung; die Reichswehr*;
5) **Alltagsgegenstände und Einrichtungen:** *der Tisch*; *der Kasten*; *die Brille*; *der Sportwagen*; *die Straßenbahn*;

6) **Materialien und Stoffe:** *die Seide*; *das Pulver*; *die Plastik*; *das Leder*; *der Marmorstein*;
7) **Fauna (Tiere, Vögel) und ihre Konstitution:** *der Schwan*; *die Pfote*; *die Kiemen*; *der Löwe*; *der Hund*;
8) **Flora (Pflanzen) und ihre Morphologie, landwirtschaftliche Kulturen:** *die Rose*; *die Blüte*; *die Gurke*; *das Gewächs*; *der Baum*;
9) **Nahrungsmittel (Nahrung und Getränke):** *der Hering*; *das Brot*; *die Milch*; *der Tee*; *die Grütze*;
10) **Naturerscheinungen und Objekte, Raumbegriffe:** *der Himmel*; *der Regen*; *die Luft*; *das Meer*; *die Wiese*;
11) **Kleidung und Schuhe:** *das Hemd*; *das Kleid*; *der Kragen*; *die Schuhe*; *der Hut*;
12) **Bauten und ihre Elemente:** *das Haus*; *die Windmühle*; *die Wand*; *die Tür*; *das Dach*;
13) **Zeitbegriffe:** *der Tag*; *die Nacht*; *der Monat*; *die Sprühregenperiode*; *die Woche*;
14) **Formen und Figuren:** *die Linie*; *der Streifen*; *die Kante*; *die Silhouette*;
15) **Hell-Dunkel-Begriffe, Lichtquellen:** *das Licht*; *das Feuer*; *die Flamme*; *die Blinklichter*; *der Schatten*;
16) **Färbung, Schattierungen:** *die Farbe*; *der Grund*; *der Anstrich*; *das Rot*; *der Hauch*;
17) **sonstige Begriffe:** *das Leben*; *die Antwort*; *die Gefahr*; *die Autoschlangen*; *die Punkte*.

Die Gebrauchshäufigkeit aller 17 Subklassen der Substantive mit den zu untersuchenden Farbbezeichnungen ist in Tabelle 2 aufgeführt.

Tabelle 2
Gebrauchshäufigkeit der Farbbezeichnungen mit den Subklassen der Substantive

Die Subklassen der Substantive	Farbbezeichnungen									
	weiß	schwarz	rot	blau	grün	grau	braun	gelb	sonstige	insgesamt
Äußeres des Menschen	81	86	76	39	11	44	54	15	213	619
Eigennamen etc.	22	32	32	11	10	5	11	2	46	171
Personen	4	9	7	1	4	4	2	1	19	51
Organisationen	7	5	6	2	1	3	1	0	17	42
Alltagsgegenstände	61	61	45	50	37	18	28	27	157	484
Materialien	40	34	14	10	19	13	5	7	71	213
Fauna	42	36	12	9	6	15	0	6	46	172
Flora	14	2	24	33	28	2	3	15	46	167
Nahrungsmittel	27	30	15	1	4	0	9	9	44	137
Natur, Raum	33	50	8	48	19	36	5	6	65	270
Kleidung, Schuhe	143	124	52	48	23	20	12	17	214	652
Bauten	54	12	5	8	7	10	5	2	78	181
Zeitbegriffe	1	7	2	0	0	10	1	0	3	24
Formen, Figuren	15	32	36	11	3	3	8	7	52	167
Hell-Dunkel-Begriffe	10	5	10	0	20	3	1	6	21	76
Färbung	6	14	1	7	0	2	3	3	27	63
Sonstige Begriffe	9	12	4	10	14	13	0	13	48	123
Insgesamt	569	551	349	288	206	201	148	136	1071	3612

Anhand dieser Tabelle wird deutlich, dass sich die Frequenzen der Kombinierbarkeit der untersuchten Adjektive mit Substantiven voneinander merklich unterscheiden:

1. Solche Wörter wie *weiß, schwarz, rot* zeichnen sich durch die breiteste Palette der Kombinierbarkeit aus; andererseits haben die syntagmatischen Relationen des Wortes *gelb* einen eingeschränkten Charakter.

2. Auch muss darauf aufmerksam gemacht werden, dass die Frequenzen der Kombinierbarkeit ausgewählter Adjektive, die sich mit allen Subklassen der Substantive verbinden, ungleichmäßig verteilt werden: So verhält sich die Gebrauchshäufigkeit von *weiß* mit der Subklasse *Flora* zu der mit der Subklasse *Bauten* wie 14:54; und dieselben Werte bei *schwarz* verhalten sich zueinander wie 2:12; bei *rot* wie 24:2.

Die Adjektive *blau* und *grün* kommen mit 15 Subklassen der Substantive vor, aber mit den Substantiven zur Bezeichnung der *Naturerscheinungen, Objekte und Raumbegriffe* verbindet sich *blau* 48-mal, *grün* 19-mal; und die syntagmatischen Relationen dieser Wörter mit den Substantiven, die *Hell-Dunkel-Begriffe* bezeichnen, verhalten sich zueinander wie 0:20.

Im Allgemeinen liegen der Fähigkeit des Wortes, syntagmatische Verbindungen mit anderen Wörtern einzugehen, sowie der ungleichmäßigen Verteilung der Frequenzen dieser Verbindungen inner- und außersprachliche Einflüsse zugrunde. Einerseits werden die Wechselbeziehungen von zwei Wörtern in der Rede durch bestimmte „gegenständliche“ Beziehungen zwischen den Denotaten dieser Wörter in der außersprachlichen Realität bedingt. Andererseits geschehen die Auswahl und die Einprägung verschiedener Wortverbindungen innerhalb der Sprache infolge der Wirkung eines solchen selektiven Faktors wie sprachlicher Usus. In der Linguistik werden der außersprachliche Faktor als *Vereinbarkeit von Denotaten* und der innensprachliche Faktor als *Vereinbarkeit von Lexemen* bezeichnet (Levickij 1989: 44).

Nun zum Grund der quantitativen Variabilität syntagmatischer Verbindungen zwischen den Wörtern. Man kann vermuten, dass die Frequenz des gemeinsamen Auftretens von zwei Wörtern im Text von der Gebrauchshäufigkeit jedes einzelnen Wortes abhänge. So kann die Kombinierbarkeit im Modell *weiß + Subklasse der Substantive zur Bezeichnung der Kleidung* sowohl durch die relativ hohe Frequenz von *weiß* als auch durch die Frequenz der angegeben Substantive bedingt sein. Das Wort *weiß* nimmt nach der Gebrauchshäufigkeit (569 Belege) tatsächlich den ersten Platz unter den untersuchten Adjektiven (genauso wie die Subklasse *Kleidung und Schuhe* (652 Belege) unter den Subklassen der Substantive) ein (s. Tab. 3). Wird dabei die Kombinierbarkeit anderer Adjektive mit den Subklassen der Substantive verglichen, so wird eine ungleichmäßige Verteilung der Frequenzen des gemeinsamen Wortgebrauchs in Verbindungen deutlicher.

Die Tabelle 3 verdeutlicht u.a., dass wir für eine vollständige Häufigkeitscharakteristik der Substantive, d.h. der kontextuellen Partner der Farbadjektive, auch andere quantitative Parameter der Kombinierbarkeit verwendet haben wie etwa Kombinierbarkeitsweite, Umfang und Gebräuchlichkeit der Subklasse.

Unter *der Kombinierbarkeitsweite* versteht man die Zahl der Partner, mit denen eine bestimmte Subklasse in Kollokation tritt. In unserer Arbeit werden mithilfe dieses Parameters Unterschiede in der Kombinierbarkeit jeder Subklasse der Substantive mit den zu untersuchenden Farbadjektiven veranschaulicht. Die Kombinierbarkeitsweite (als relativer Wert) zeigt auf, dass die Zahl der Subklassen, mit denen das Untersuchungsobjekt syntagmatische Beziehungen eingehen kann, für jeden konkreten Fall der Forschung anders ist. Eine solche relative Größe lässt sich bestimmen, nachdem die Zahl der Kontextpartner durch die Zahl der zu untersuchenden Einheiten geteilt ist.

Der Terminus *Umfang*, der mit der *Gebräuchlichlichkeit* eng verbunden ist (s. auch S. 70), repräsentiert hier das Verzeichnis der Substantive, die in jeder Subklasse bei allen Partnern (*weiß, schwarz, rot* etc.) je nach der Häufigkeit auftreten.

Wir haben festgestellt (s. Tab. 3), dass die Subklassen *Kleidung und Schuhe, das Äußere des Menschen*, *Alltagsgegenstände und Einrichtungen*, *Naturerscheinungen und Objekte*, *Stoffe* sowie *Bauten und ihre Elemente* zu den höchstfrequenten Subklassen syntagmatischer Partner gehören. In diesen Subklassen ist auch die größte Kombinierbarkeitsweite zu beobachten. Die Subklasse der Substantive *Zeitbegriffe*, in der die Gebrauchshäufigkeit am niedrigsten ist, ist durch die kleinste Kombinierbarkeitsweite gekennzeichnet.

Tabelle 3
Die quantitativen Charakteristiken der Subklassen der Substantive

Subklassen	Gebrauchs-häufigkeit	Kombinierbar-keitsweite	Gebräuchlichkeit	Umfang
Äußeres des Menschen	619	1	3,94	112
Eigennamen etc.	171	1	2,05	62
Personen	51	1	1,34	25
Organisationen	42	0,88	0,99	21
Alltagsgegenstände	484	1	1,35	220
Materialien	213	1	1,47	95
Fauna	172	0,88	1,41	78
Flora	167	1	2,52	57
Nahrungsmittel	137	0,88	1,74	44
Natur, Raum	270	1	1,85	99
Kleidung, Schuhe	652	1	1,84	203
Bauten	181	1	1,25	69
Zeitbegriffe	24	0,63	0,82	14
Formen, Figuren	167	1	1,38	73
Hell-Dunkel-Begriffe	76	0,88	1,63	30
Färbung	63	0,88	1,44	22
Sonstige Begriffe	123	0,88	1,65	42

Es fragt sich nun, ob die festgestellten Tatsachen in einem direkten Verhältnis stehen? Allerdings wird die erwähnte Korrelation nicht überall beobachtet: Solche Subklassen der Substantive wie *Flora*, *Formen und Figuren*, *sonstige Begriffe* lassen sich auch durch die größte Kombinierbarkeitsweite charakterisieren, wobei die Frequenz ihres Gebrauchs nur mittlere Werte erreicht. Die Subklassen *Fauna* und *Eigennamen*, die eine relativ hohe Gebrauchshäufigkeit haben, sind durch eine kleinere Kombinierbarkeitsweite gekennzeichnet als die Subklassen der Substantive zur Bezeichnung *der Flora, der Formen und Figuren* sowie *sonstiger Begriffe*. Andererseits ist die größte Kombinierbarkeitsweite mit den Farbadjektiven bei den Substantiven zur Bezeichnung von *Personen* bestimmt worden, obwohl die Gebrauchshäufigkeit der Bestandteile dieser Subklasse ganz gering ist, d.h., es besteht hier kein direktes Verhältnis zueinander.

Um festzustellen, inwieweit die erwähnten quantitativen Parameter miteinander verbunden sind, haben wir aufgrund der Tabelle 3 eine Korrelationsanalyse durchgeführt (s. Tab. 4).

Wie in Tabelle 4 deutlich ist, lassen sich die Charakteristiken der Gebrauchshäufigkeit der Substantive und der Kombinierbarkeitsweite, indem sie statistisch signifikante Werte ($r = 0{,}56$) erreichen, im Allgemeinen nicht durch eine relativ hohe Korrelationsabhängigkeit

voneinander kennzeichnen. (In diesem Fall ergibt sich bei *df =17-2 = 15* Freiheitsgraden (FG) und dem Signifikanzniveau 0,05 der minimale Korrelationskoeffizient $R_{0,05;15} = 0,48$; beim Signifikanzniveau 0,01 (1%) ist der minimale Korrelationskoeffizient gleich $R_{0,01;15} = 0,61$). Es muss aber auch darauf hingewiesen werden, dass eine sehr starke Korrelation ($r = 0,80$) zwischen der Kombinierbarkeitsweite und der Gebräuchlichkeit der Subklasse von Substantiven besteht, d.h. je höher die Gebräuchlichkeit der Subklasse von Substantiven ist, umso größer wird die Kombinierbarkeitsweite dieser Substantive mit den zu untersuchenden Farbadjektiven. Unter Berücksichtigung des berechneten Wertes des Korrelationskoeffizienten ($r = 0,91$) kann die Schlussfolgerung vollzogen werden, dass zwischen der Gebrauchshäufigkeit und dem Umfang der Subklasse eine volle Übereinstimmung bei den angegebenen Wortverbindungen beobachtet wird.

Tabelle 4
Das Verhältnis zwischen den quantitativen Charakteristiken (Korrelationskoeffizienten)

	Kombinierbarkeitsweite	Gebräuchlichkeit	Umfang
Häufigkeit	0,56	0,70	0,91
Kombinierbarkeitsweite		0,80	0,60
Gebräuchlichkeit			0,53

Andererseits hängen die signifikanten Werte des Gebrauchs von Subklassen der Substantive mit den Farbadjektiven von der Gebrauchshäufigkeit der ganzen Subklasse ($r = 0,70$) oder deren einzelnen Elemente ab – so zeugt diese Tatsache allein von einem großen Potenzial der syntagmatischen Eigenschaften einzelner Kontextpartner. Wir sehen, dass die Größe des Gebrauchs der Subklasse in gewissem Maß ($r = 0,53$) von der Anzahl ihrer Elemente abhängt. Anders gesagt, ist der Gebrauch durch das Verhältnis zweier Merkmale der Subklasse – Umfang und Frequenz – zueinander bedingt: Die größten Gebrauchswerte sind da zu ermitteln, wo eine geringe Anzahl von Wörtern vielmals gebraucht worden ist.

Darüber hinaus haben wir die Kombinierbarkeitsweite der Farbadjektive mit allen Subklassen bestimmt. So wurde festgestellt, dass die Farbwörter *weiß, schwarz, rot* in Texten mit 17 (von 17) Subklassen der Substantive und *blau, grün, braun, gelb* nur mit 15 Subklassen kollokieren (s. Tab. 2). Im ersten Fall beträgt die Kombinierbarkeitsweite 17:17 = 1 und im zweiten Fall 15:17 = 0,88.

Die Kombinierbarkeitsweite der Farbadjektive ist in Tabelle 5 angegeben.

Tabelle 5
Die Werte der Kombinierbarkeitsweite von Farbadjektiven

Farbadjektive	Kombinierbarkeitsweite
weiß	1
schwarz	1
rot	1
blau	0,88
grün	0,88
grau	0,94
braun	0,88
gelb	0,88

Man sieht, dass sich die von uns untersuchten Farbadjektive durch unterschiedliche Kombinierbarkeitsweiten charakterisieren lassen: Die höchstfrequenten Farbwörter *weiß, schwarz, rot* zeichnen sich durch den größten Bereich der Kombinierbarkeit aus. Es wurde auch ein relativ hoher Wert der Kombinierbarkeitsweite bei *grau* (0,94) festgestellt; dabei erscheint dessen Gebrauchshäufigkeit als nicht so groß (nur 201 Belege).

Wie die Korrelationsanalyse (s. Kap. 8) aufgezeigt hat, hängt die Weite der lexikalischen Kombinierbarkeit der Adjektive in den durchgeführten Forschungen hauptsächlich von der Zahl syntagmatischer Verbindungen ($r = 0{,}93$), der Zahl paradigmatischer Verbindungen ($r = 0{,}95$) und der Gebräuchlichkeit ($r = 0{,}92$) ab, und in etwas kleinerem Maß von der Frequenz ihres Gebrauchs ($r = 0{,}65$); dabei ergibt sich bei $df = 8 - 2 = 6$ FG und dem Signifikanzniveau 0,05 der minimale Korrelationskoeffizient $R_{0,05;6} = 0{,}71$; beim Signifikanzniveau 0,01 (1%) ist der minimale Korrelationskoeffizient gleich $R_{0,01;\,6} = 0{,}83$.

Trotz der ziemlich hohen, fast homogen repräsentierten Kombinierbarkeitsweite der Farbadjektive mit ihren kontextuellen Partnern wird der innere Charakter syntagmatischer Relationen einer Subklasse zu jedem untersuchten Wort immer anders sein. Das lässt sich durch einen solchen relativen Wert wie *die Gebräuchlichkeit* der Subklasse von Substantiven feststellen. Zur Gewinnung dieses Parameters muss die Zahl des Wortgebrauchs durch den Umfang der Subklasse geteilt werden: So wurden beispielsweise mit *grün* 25 Substantive, die Alltagsgegenstände bezeichnen, von den Autoren 37-mal gebraucht. Dabei beträgt die Größe der Gebräuchlichkeit dieser Subklasse mit dem Farbwort *grün* 37:25 = 1,48. In der Tabelle (s. Anlage 3) führen wir die Forschungsergebnisse zur Gebräuchlichkeit aller Subklassen der Substantive mit den Farbadjektiven in der deutschen schöngeistigen Literatur an.

Unserer Überzeugung nach verweisen solche Abweichungen in der Größe, die für die Kombinierbarkeitsweite charakteristisch sind, auf die Verschiedenartigkeit der semantischen Komponenten, aus denen die Bedeutung der untersuchten Adjektive besteht, sowie sicherlich darauf, dass die Breite syntagmatischer Beziehungen einer lexikalischen Einheit durch inner- und außersprachliche Faktoren beeinflusst wird.

Tabelle 6
Vergleichende Werte der quantitativen Parameter der Substantive und der Farbadjektive (Korrelationskoeffizienten)

	Farbadjektive	Subklasse der Substantive
Kombinierbarkeitsweite / Gebräuchlichkeit	0,92	0,80
Häufigkeit / Gebräuchlichkeit	0,79	0,70
Häufigkeit / Kombinierbarkeitsweite	0,65	0,56

Nachdem die Ergebnisse der Korrelationsanalyse der Subklasse von Nomen und der Farbadjektive im Deutschen (s. Tab. 4 und 8) vergleichend betrachtet worden sind, können wir schließlich behaupten, dass zwischen solchen Charakteristiken syntagmatischer Partner wie *Gebrauchshäufigkeit, Kombinierbarkeitsweite* und *Gebräuchlichkeit* ein direktes Verhältnis zueinander besteht (s. Tab. 6). Mit anderen Worten werden die syntagmatischen Eigenschaften kontextualer Partner durch die deutlich ausgeprägten Relationen der Übereinstimmung charakterisiert, und zwar:

1) durch eine außerordentlich starke Korrelation zwischen der Kombinierbarkeitsweite und der Gebräuchlichkeit (vgl. 0,92 : 0,80);

2) durch einen enormen Einfluss der Gebrauchshäufigkeit auf die Gebräuchlichkeit (vgl. 0,79 : 0,70);

3) durch eine ganz geringe Abhängigkeit der Kombinierbarkeitsweite von der Gebrauchsfrequenz (vgl. 0,65 : 0,56).

6. Die Kombinierbarkeit der Adjektive im Modell [Wort + Subklasse]

Die Stufe der semantischen Verbindung zwischen Wörtern kann, wie V. Levickij (1989) zeigte, mit Hilfe des χ^2-Tests und des Bernoullischen Koeffizienten Φ gemessen werden.

$$\chi^2 = \sum_i \frac{\left(O_i - E_i\right)^2}{E_i}, \tag{1}$$

wobei χ^2 = Kriterium der Homogenität;
Σ = Summe;
O_i = empirische Werte;
E_i = theoretisch mögliche Werte.

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, \tag{2}$$

wobei Φ = Bernoullischer Kontigenzkoefizient;
a, b, c, d = empirische Werte in der Vierfeldertafel.

Anstelle dieses statistischen Ansatzes könnte auch die Formel (3) benutzt werden.

$$\mathrm{u} = \frac{n_{ij} - E_{ij}}{\sqrt{\frac{n_i \cdot n_j \cdot (N - n_i) \cdot (N - n_j)}{N^2 \cdot (N-1)}}}, \tag{3}$$

wobei n_i = die Randsummen auf der rechten Seiten der Tabelle;
n_j = die Randsummen unterhalb der Tabelle;
N = Summe aller Häufigkeiten in der Tabelle;
n_{ij} = Häufigkeit in der Zelle (i,j) der Tabelle;
E_{ij} = erwartete Häufigkeit in Zelle (i,j) = $n_i n_j / N$

Die Praxis statistischer Forschungen zeigt, dass es zweckmäßig ist, die Angaben für eine Analyse als sog. alternative Verteilung insbesondere in Form von Vierfelder-Tabellen darzustellen. Solche Tabellen kann man aufgrund einer größeren Tabelle zusammenstellen, indem man die Spalten- und Zeilenzahl in einer aus mehreren Feldern bestehenden Tabelle verringert oder vereinigt.

So lässt sich eine alternative Tabelle, mit deren Hilfe Relationen zwischen den zu untersuchenden Merkmalen paarweise bestimmt werden könnten, folgendermaßen gestalten (s. Tab. 7).

Nach der durchgeführten statistischen Analyse wurde festgestellt, dass die syntagmatischen Verbindungen des Farbadjektivs *weiß* mit folgenden Subklassen der Substantive statistisch signifikant sind: 1) *weiß* + die Substantive zur Bezeichnung der Bauten und ihrer Elemente (χ^2 = 28,47; Φ = 0,09); 2) *weiß* + Substantive zur Bezeichnung der Kleidung und Schuhen (χ^2 = 22,89; Φ = 0,08); 3) *weiß* + Substantive zur Bezeichnung der Fauna (χ^2 = 10,22; Φ = 0,05); 4) *weiß* + Substantive zur Bezeichnung der Flora (χ^2 = 7,16; Φ = 0,04); 5)

weiß + Substantive zur Bezeichnung sonstiger Begriffe (χ^2 = 6,83; Φ = 0,04); 6) *weiß* + Substantive, die Formen und Figuren bezeichnen (χ^2 = 6,02; Φ = 0,04); 7) *weiß* + Substantive zur Bezeichnung der Alltagsgegenstände (χ^2 = 4,18; Φ = 0,03); 8) *weiß* + Substantive, die das Äußere des Menschen bezeichnen (χ^2 = 4,01; Φ = 0,03).

Tabelle 7
Die alternative Verteilung der Frequenzen eines Wortes

a	b	(a+b)
c	d	(c+d)
(a+c)	(b+d)	N

wobei *a, b, c, d* – empirische Größen in der Vierfelder-Tabelle;
N – die Gesamtmenge der Beobachtungen.

Darüber hinaus wurde festgestellt, dass die syntagmatischen Wortverbindungen mit den Substantiven, die Eigennamen (22 Belege), Fauna (18), den Sozialstatus und Organisationen (7), Materialien und Stoffe (40), Nahrungsmittel (27), Naturerscheinungen, Objekte und Raumbegriffe (33 Belege), Hell-Dunkel-Begriffe (10) sowie Farbe und Färbung (6) bezeichnen, keine statistische Signifikanz erreichen. So kann man behaupten, dass die Beziehungen zwischen den Elementen dieser Verbindungen einen zufälligen Charakter haben; deshalb sind sie nicht als stabile Wortverbindungen zu betrachten. Wegen niedriger Häufigkeitswerte haben wir keine Tabellen für die Verbindungen des Adjektivs *weiß* mit den Subklassen der Substantive zur Bezeichnung der Personen (4 Belege) und Zeitbegriffe (1) zusammengestellt.

Es wurde auch festgestellt, dass für das Farbadjektiv *schwarz* folgende Verbindungen statistisch signifikant sind: 1) *schwarz* + Substantive, die Bauten und ihre Elemente bezeichnen (χ^2 = 10,97; Φ = 0,06); 2) *schwarz* + Substantive, die Kleidung und Schuhe bezeichnen (χ^2 = 8,72; Φ = 0,05); 3) *schwarz* + Substantive, die Alltagsgegenstände bezeichnen (χ^2 = 8,22; Φ = 0,05); 4) *schwarz* + Substantive zur Bezeichnung der Nahrungsmittel und Getränke (χ^2 = 4,86; Φ = 0,04); 5) *schwarz* + Substantive, die Hell-Dunkel-Begriffe, Lichtquellen bezeichnen (χ^2 = 4,52; Φ = 0,04); 6) *schwarz* + Substantive zur Bezeichnung der Fauna (χ^2 = 4,50; Φ = 0,04).

Die Verbindungen *schwarz* mit den Substantiven, die a) das Äußere des Menschen (86 Belege); b) Eigennamen (32); c) Materialien (34); d) Naturerscheinungen und Objekte (50 Belege) bezeichnen, lassen sich nicht statistisch signifikant nennen, obwohl die Kombinierbarkeit von *schwarz* mit diesen Subklassen der Substantive durch eine relativ hohe Gebrauchshäufigkeit charakterisiert wird. Es wurde auch erkannt, dass die syntagmatischen Beziehungen von *schwarz* mit solchen lexikalischen Einheiten, die a) Leute (9 Belege); b) den Sozialstatus des Menschen (5); c) Zeitbegriffe (7); d) Farbe und Färbung (14); e) sonstige Begriffe (12) bezeichnen, keine statistische Signifikanz haben. Wir haben keine Vierfelder-Tabelle zur Ermittlung der Kombinierbarkeit von *schwarz* mit Substantiven zur Bezeichnung der Flora (2 Belege) zusammengestellt, weil die Größen ihrer Gebrauchshäufigkeit zu niedrig sind.

Das Farbadjektiv *rot* kommt in Verbindungen mit allen Subklassen der Substantive vor. Statistisch signifikant sind die syntagmatischen Beziehungen des Farbwortes *rot* mit folgenden Subklassen der Substantive: 1) *rot* + Substantive zur Bezeichnung der Formen und Figuren (χ^2 = 28,38; Φ = 0,09); 2) *rot* + Substantive zur Bezeichnung der Eigennamen (χ^2 = 16,85; Φ = 0,07); 3) *rot* + Substantive zur Bezeichnung der Naturerscheinungen, Objekte und Raumbegriffe (χ^2 = 15,00; Φ = 0,06); 4) *rot* + Substantive, die Bauten und ihre Elemente bezeich-

nen (χ^2 = 10,39; Φ = 0,05); 5) *rot* + Substantive, die das Äußere des Menschen bezeichnen (χ^2 = 5,86; Φ = 0,04); 6) *rot* + Substantive zur Bezeichnung der Flora (χ^2 = 4,45; Φ = 0,04).

Eine sorgfältige quantitative Analyse der 12 Vierfelder-Tabellen lässt bestimmen, dass solche Verbindungen des Adjektivs *blau* mit hochfrequenten Subklassen der Substantive statistisch signifikant sind: 1) *blau* + Substantive, die Naturerscheinungen und Objekte, Raumbegriffe bezeichnen (χ^2 = 38,23; Φ = 0,10); 2) *blau* + Substantive, die Flora bezeichnen (χ^2 = 33,15; Φ = 0,09); 3) *blau* + Substantive, die Alltagsgegenstände und Geräte bezeichnen (χ^2 = 4,23; Φ = 0,03).

Keine statistische Signifikanz haben die Verbindungen von *blau* mit den Substantiven, die a) das Äußere des Menschen (39 Belege); b) Eigennamen (11); c) Materialien und Stoffe (10); d) Fauna (9); e) Kleidung und Schuhe (48 Belege); f) Bauten und ihre Elemente (8); g) Formen und Figuren (11); h) Farben und Färbung (7); i) sonstige Begriffe (10) bezeichnen. Das heißt, solche Syntagmen gehören nach dem Charakter der Kombinierbarkeit nicht zu stabilen, sondern freien Verbindungen. Und mit den Subklassen der Substantive, die Personen (1 Beleg), einen Sozialstatus und Organisationen (2) sowie Nahrungsmittel und Getränke (1) bezeichnen, haben wir niederfrequente Wortverbindungen mit dem untersuchten Adjektiv *blau* fixiert.

Die syntagmatischen Verbindungen des Farbadjektivs *grün* sind mit folgenden Subklassen der Substantive statistisch signifikant: 1) *grün* + Substantive zur Bezeichnung der Hell-Dunkel-Begriffe und Lichtquelle (χ^2 = 61,33; Φ = 0,13); 2) *grün* + Substantive zur Bezeichnung der Flora (χ^2 = 39,85; Φ = 0,11); 3) *grün* + Substantive zur Bezeichnung des Äußeren des Menschen (χ^2 = 21,41; Φ = 0,08); 4) *grün* + Substantive, die Kleidung und Schuhe bezeichnen (χ^2 = 7,00; Φ = 0,04); 5) *grün* + Substantive, die Formen und Figuren bezeichnen (χ^2 = 4,97; Φ = 0,04); 6) *grün* + Substantive, die Materialien und Stoffe bezeichnen (χ^2 = 4,36; Φ = 0,03); 7) *grün* + Substantive, die Alltagsgegenstände und Geräte bezeichnen (χ^2 = 3,92; Φ = 0,03).

Es sollte auch hervorgehoben werden, dass sich die Kombinierbarkeit des Adjektivs *grün* mit den Substantiven, die Fauna (6 Belege), Naturerscheinungen und Objekte, Raumbegriffe (19), Bauten und ihre Elemente (7), Eigennamen (10) bezeichnen, wegen eines zufälligen Charakters der Beziehungen zwischen den Elementen der Verbindungen als niedriger als die theoretisch erwarteten Größen erwiesen hat.

Wegen niedriger Häufigkeitswerte haben wir keine Vierfelder-Tabellen für die Verbindungen des Adjektivs *grün* mit den Subklassen der Substantive zur Bezeichnung der Personen (4 Belege), des Sozialstatus und der Organisationen (1 Beleg), Nahrungsmittel und Getränke (4 Belege) sowie Formen und Figuren (3 Belege) zusammengestellt.

Die Kombinierbarkeit des Farbwortes *grau* mit den Substantiven, die Personen (4 Belege), den Sozialstatus des Menschen und Organisationen (3), Flora (2), Formen und Figuren (3), Hell-Dunkel-Begriffe, Lichtquellen (3), Farbe und Färbung (2) bezeichnen, ist niederfrequent.

Nachdem 10 alternative Vierfelder-Tabellen zusammengestellt und untersucht worden sind, lässt sich bestimmen, dass nur die folgenden Verbindungen des Adjektivs *grau* mit Subklassen der Substantive statistisch signifikant ist: 1) *grau* + Substantive, die Zeitbegriffe bezeichnen (χ^2 = 59,92; Φ = 0,13); 2) *grau* + Substantive, die Naturerscheinungen und Objekte oder Raumbegriffe bezeichnen (χ^2 = 33,51; Φ = 0,10); 3) *grau* + Substantive, die Kleidung und Schuhe bezeichnen (χ^2 = 9,44; Φ = 0,05); 4) *grau* + Substantive zur Bezeichnung sonstiger Begriffe (χ^2 = 6,07; Φ = 0,04).

Wie die Ergebnisse einer statistischen Analyse der syntagmatischen Verbindungen von *braun* (9 alternative Tabellen) bezeugen, sind sie mit folgenden Subklassen der Substantive statistisch signifikant: 1) *braun* + Substantive, die das Äußere des Menschen bezeichnen (χ^2 =

40,69; Φ = 0,11); 2) *braun* + Substantive, die Kleidung und Schuhe bezeichnen (χ^2 = 10,31; Φ = 0,05); 3) *braun* + Substantive, die Eigennamen bezeichnen (χ^2 = 7,68; Φ = 0,05); 4) *braun* + Substantive, die Alltagsgegenstände und Geräte bezeichnen (χ^2 = 4,05; Φ = 0,03).

Das Farbadjektiv *gelb* kommt in Verbindungen mit fast allen Subklassen der Substantive vor. So wurden 11 Tabellen für die Durchführung einer quantitativen Untersuchung und die Feststellung der statistisch signifikanten Verbindungen dieses Adjektivs mit allen Kontextpartnern außer den niederfrequenten (Substantive, die Eigennamen (2 Belege), Personen (1 Beleg), Bauten und ihre Elemente (2 Belege) sowie Farbe und Färbung (3 Belege) bezeichnen) zusammengestellt.

Statistisch signifikant sind die syntagmatischen Beziehungen des Farbwortes *gelb* mit folgenden Subklassen der Substantive: 1) *gelb* + Substantive zur Bezeichnung sonstiger Begriffe (χ^2 = 16,27; Φ = 0,07); 2) *gelb* + Substantive zur Bezeichnung der Flora (χ^2 = 13,75; Φ = 0,06); 3) *gelb* + Substantive zur Bezeichnung der Alltagsgegenstände und Geräte (χ^2 = 5,07; Φ = 0,04).

Wie die Forschungsergebnisse bezeugen, gehen die Adjektive *weiß, schwarz, rot, grün* die meisten statistisch signifikanten syntagmatischen Verbindungen ein. Übrigens sind die Farbwörter *weiß, schwarz, rot* durch den breitesten Umfang der Kombinierbarkeit und die höchste Gebrauchsfrequenz gekennzeichnet. Die anderen Farbadjektive haben durchschnittlich drei oder vier statistisch signifikante Beziehungen mit den zusammengefassten Subklassen der Substantive. Bemerkenswert ist, dass für das hochfrequente Adjektiv *blau* (288 Belege) nur drei statistisch signifikante Verbindungen mit den Subklassen der Substantive festgestellt wurde, während die Farbbezeichnung *grün* bei gleichem Umfang der Kombinierbarkeit und viel kleineren Werten der Gebrauchshäufigkeit statistisch signifikante Beziehungen immerhin mit sieben Subklassen der Substantive hat.

Die meisten statistisch signifikanten syntagmatischen Relationen der Farbadjektive wurden in Verbindungen mit den Subklassen der Substantiven festgestellt, die Alltagsgegenstände und Geräte, das Äußere des Menschen, Flora, Kleidung und Schuhe sowie sonstige Begriffe bezeichnen. Denn gerade mit diesen kontextuellen Partnern bilden die Farbadjektive stabile Wortverbindungen im Deutschen. Dabei haben die Verbindungen der zu untersuchenden Adjektive mit den Substantiven zur Bezeichnung der Personen, des Sozialstatus des Menschen und der Organisationen sowie der Farben und Färbung keine statistische Signifikanz, was von einem zufälligen Charakter der Beziehungen zwischen den Elementen dieser Verbindungen an und für sich zeugt.

7. Die Kombinierbarkeit im Modell [Wort + Wort]

Die Praxis der statistischen Erforschung der Lexik an der Universität Tscherniwzi hat die Zweckmäßigkeit und die Möglichkeit der Untersuchung syntagmatischer Verbindungen der Adjektive nicht nur im Modell *[Wort + Subklasse]*, sondern auch im Modell *[Wort + Wort]* aufgezeigt.

Die von uns durchgeführte quantitative Analyse der Kombinierbarkeit der Farbadjektive im erwähnten Modell hat gezeigt, dass die syntagmatischen Verbindungen von *weiß* mit folgenden Substantiven statistisch signifikant sind: *weiß* + Zähne (χ^2 = 62,34; Φ = 0,13); *weiß* + Wolke (χ^2 = 24,92; Φ = 0,08); *weiß* + Bluse (χ^2 = 23,87; Φ = 0,08); *weiß* + Hemd (χ^2 = 23,56; Φ = 0,08); *weiß* + Kleid (χ^2 = 14,67; Φ = 0,06); *weiß* + Gesicht (χ^2 = 13,61; Φ = 0,06); *weiß* + Schuhe (χ^2 = 9,98; Φ = 0,05); *weiß* + Bart (χ^2 = 8,47; Φ = 0,05).

Das Adjektiv *schwarz* hat statistisch signifikante syntagmatische Verbindungen mit folgenden Substantiven: *schwarz* + Haar (χ^2 = 31,12; Φ = 0,09); *schwarz* + Katze (χ^2 = 29,43;

Φ = 0,09); *schwarz* + Wasser (χ^2 = 24,67; Φ = 0,08); *schwarz* + Bart (χ^2 = 21,21; Φ = 0,07); *schwarz* + Schuhe (χ^2 = 12,79; Φ = 0,06); *schwarz* + Kleid (χ^2 = 9,19; Φ = 0,05).

Bei der Untersuchung syntagmatischer Relationen von *rot* mit konkreten Wörtern wurde festgestellt, dass das Farbadjektiv *rot* statistisch signifikante syntagmatische Verbindungen mit folgenden Substantiven hat: *rot* + Rosen (χ^2 = 102,36; Φ = 0,17); *rot* + Grütze (χ^2 = 93,13; Φ = 0,16); *rot* + Lippen (χ^2 = 67,62; Φ = 0,14); *rot* + Fleck (χ^2 = 37,70; Φ = 0,10); *rot* + Gesicht (χ^2 = 30,58; Φ = 0,09).

Bezüglich des Adjektivs *blau* wurde festgestellt, dass syntagmatische Beziehungen von *blau* mit allen ausgewählten Substantiven statistisch signifikant sind: *blau* + Blume (χ^2 = 202,75; Φ = 0,24); *blau* + Himmel (χ^2 = 183,63; Φ = 0,23); *blau* + Augen (χ^2 = 69,52; Φ = 0,14).

Aufgrund der Ergebnisse der durchgeführten statistischen Analysen wurde festgestellt, dass nur ein Substantiv in den syntagmatischen Verbindungen mit dem Farbadjektiv *grün* statistische Signifikanz hat: *grün* + Licht (χ^2 = 144,92; Φ = 0,20).

Statistisch signifikant sind auch die syntagmatischen Beziehungen von *grau* mit allen drei ausgewählten Substantiven: *grau* + Straße (χ^2 = 155,15; Φ = 0,21); *grau* + Haar (χ^2 = 25,66; Φ = 0,08); *grau* + Augen (χ^2 = 12,64; Φ = 0,06).

Wie aus den Ergebnissen einer quantitativen Untersuchung folgt, hat nur ein Substantiv eine statistisch signifikante syntagmatische Verbindung mit der Farbbezeichnung *braun*: *braun* + Augen (χ^2 = 146,16; Φ = 0,20).

Hiermit haben wir viele statistisch signifikante stabile Wortverbindungen erhalten, die aus einem minimalen Quantum der Glieder (zwei Einheiten) bestehen.

Die Entwicklungswege und Funktion von Wortverbindungen eines stabilen Kontextes sind sowohl durch außersprachliche als auch innersprachliche Faktoren bedingt. Die Lebensbedingungen der Gesellschaft, Sitten und Bräuche, alltäglicher Sachverhalt, ästhetische und sprachliche Normen etc. – alle erwähnten Faktoren führen zur Bildung stereotyper lexikalischer Redewendungen (Klischees), die als Bindeglied zwischen Wahrnehmung, Denken und Sprache einer menschlichen Gemeinschaft im Folgenden dienen.

In der letzten Zeit werden von einigen Wissenschaftlern psycholinguistische Experimente zur Bestimmung des Universalen und Spezifischen in semantischen Assoziationen bei verschiedenen Sprachträgern sowie zur Entdeckung der Besonderheiten der Wahrnehmung der Umwelt und des Menschen als seinen Komponenten im sprachlichen Weltbild durchgeführt (s. Terechova 1997). Es wurde herausgefunden, dass die Farbbezeichnungen z.B. im Ukrainischen als hochfrequente, stereotype Assoziationen zur Charakteristik der Somatismen sehr häufig auftreten; Beispiele hierfür sind: *čorni brovy* ‘schwarze Augenbrauen’, *velyki kari abo holubi oči* ‘große braune oder blaue Augen’, *dovhi čorni viji* ‘lange schwarze Wimpern’, *bila abo smuhlava škira* ‘helle (weiße) oder dunkle Haut’, *červoni ščoky* ‘rote Wangen’, *červoni povni huby* ‘rote volle Lippen’, *bili rivni zuby* ‘weiße regelmäßige Zähne’ (Terechova 1997: 45). Begründet wird dies damit, dass das Äußere des Menschen nach solchen Parametern wie Farbe, Größe, Form, Qualität etc. wahrgenommen wird und dann adäquat in der Sprache zum Ausdruck kommt.

Aufgrund der Ergebnisse solcher Experimente lassen sich stereotype Assoziationen in der Form typischer syntagmatischer Wortverbindungen bestimmen. Sie gelten als Konstante des sprachlichen Weltbildes, denn durch diese Wörter werden jene üblichen Vorstellungen von der Welt in das konzeptuelle Weltbild eingebunden, die in einer konkreten Sprache festgehalten sind.

Typische Wortverbindungen werden auch in lexikographischen Quellen fixiert (z.B. Wörter und Wendungen. Wörterbuch zum deutschen Sprachgebrauch, herausgegeben von E. Agricola). In diesem Wörterbuch werden solche Verbindungen gesammelt, die veranschaulichen, in welchen Wendungen das definierte Wort angewendet werden kann. Allerdings

sind die Auswahlmerkmale solcher Verbindungen nicht eindeutig ausgewiesen. Man kann vermuten, dass sie zu typischen oder hochfrequenten Modellen gehören. Darunter sind Verbindungen mit unterschiedlichem Grad der Festigkeit (von freien bis festen Fügungen) zu finden. Dennoch können wir manchmal, wenn die objektiven Abgrenzungskriterien für variable und stabile, stabile und feste Kontexte in der lexikographischen Praxis fehlen, eine Inkonsequenz der Prinzipien bei der Auswahl und dem Eintragen ins Wörterbuch verschiedener Wortverbindungen beobachten. So z.B. wird die Verbindung *weiße Wolke* in Agricolas Wörterbuch nicht fixiert. Im durchgeführten Experiment ist die Summe des Chi-Quadrats für diese Kombination 24,92 gleich und der Koeffizient $\Phi = 0{,}08$, d.h. die Verbindung *weiße Wolke* wird durch einen relativ hohen Grad der Gebundenheit ihrer Komponenten gekennzeichnet, aber sie besitzt nach Meinung des Verfassers des Wörterbuches nicht den Status einer typischen und hochfrequenten Verbindung. In einem anderen einsprachigen Wörterbuch (Duden Deutsches Universalwörterbuch) hingegen ist die Wortverbindung *weiße Wolke* als eine der ersten (nach den Ausdrücken *weißer Schnee* und *weiße Schwäne*) angeführt (Duden 1996: 1725).

Wie die Ergebnisse der statistischen Experimente zeigen, ergibt sich der hohe Grad der Gebundenheit zwischen Komponenten zweigliedriger Verbindungen niemals durch Zufall: Stabile Wortverbindungen werden in der Lexikografie entweder als „typisch“ oder als „fest“ oder als „phraseologisch verbunden“ interpretiert (Bystrova 1977: 9).

Von 27 statistisch signifikanten Verbindungen, die nach der statistischen Analyse syntagmatischer Beziehungen im Modell *Wort + Wort* festgestellt worden sind, haben wir 17 Wendungen unter den Stichwörtern des o.g. Wörterbuches entdeckt. In diesem Zusammenhang ist nun jede Wortverbindung genauer zu erwähnen.

Das Farbadjektiv *weiß* bildet mit den abgesonderten Substantiven 8 statistisch signifikante Verbindungen, aber im Wörterbuch „Wörter und Wendungen“ sind nur vier davon aufgeführt. So wird die Wendung *ein weißes Kleid*, das nach der Verbindungskraft die fünfte Position einnimmt, im Wörterbuch als eine der ersten festgelegt. Eine solche Wendung müsste nach unserer Auffassung das Syntagma *weiße Zähne* sein, weil es durch die größten Werte von χ^2 und Φ ($\chi^2 = 62{,}34$; $\Phi = 0{,}13$) gekennzeichnet ist. Im Wörterbuch fehlt der o.g. Ausdruck *weiße Wolke* mit der zweitgrößten Verbindungskraft, obwohl diese Wendung im Deutschen sicherlich zu den stereotypen Sprachmodellen zur Beschreibung der Natur und Umwelt gehört. Wir finden in Agricolas Werk auch nicht die Wortverbindungen *eine weiße Bluse* und *ein weißes Hemd*, die durch relativ große Werte des Chi-Quadrats und des Koeffizienten Φ charakterisiert werden. Im Wörterbuch gibt es jedoch zwei Wendungen *weiße Strümpfe* und *weiße Handschuhe*, die semantisch am nächsten zu den erwähnten Verbindungen liegen. Die Wendungen *weißes Gesicht* und *weiße Schuhe*, die von E. Agricola auch am Anfang der Stichwortdefinition (nach den Wortverbindungen *ein weißes Kleid* und *weiße Zähne*) fixiert werden, könnten wegen der geringeren Kraft ihrer syntagmatischen Relationen eher am Ende der Stichwortdefinition aufgeführt werden. Unter den typischen Wendungen fehlt im Wörterbuch der Ausdruck *weißer Bart*, der sich auch nach Angaben einer quantitativen Analyse durch statistisch signifikante Verbindungen auszeichnet. Das Wörterbuch enthält nur die Wortverbindung *weißes Haar*, obwohl das Adjektiv *weiß* mit dem Substantiv *Haar* (11 Belege), wie die Ergebnisse unserer Experimente zeigen, keine statistische Signifikanz erreicht.

Von sechs statistisch signifikanten Verbindungen mit dem Farbadjektiv *schwarz* erscheinen im Wörterbuch von E. Agricola nur drei Wendungen, nämlich *schwarzes Haar, die schwarze Katze* und *schwarzes Kleid*. Übrigens müsste der Ausdruck *schwarzes Kleid* erst am Ende der Stichwortdefinition (wenigstens nach fünf vorhergehenden Wörtern) folgen, weil die Werte des Chi-Quadrats und des Bernoullischen Koeffizienten Φ für dieses Syntagma am niedrigsten sind. Im Wörterbuch finden wir die Wendungen *schwarzes Wasser, schwarzer Bart* und *schwarze Schuhe* nicht, die sich, wie die Forschung aufgezeigt hat, nicht nur durch

eine hohe Gebrauchsfrequenz, sondern auch durch hohe statistische Signifikanz ihrer syntagmatischen Beziehungen charakterisieren lassen. Der Verfasser, dessen Auswahl sich offensichtlich nur auf eigene Erfahrungen bzw. Intuition im Bereich der Natur und Naturerscheinungen stützt, führt in seinem Werk die Fügung *schwarze Wolken* auf, die ihrem Wortinhalt nach wohl am nächsten zur Wortverbindung *schwarzes Wasser* liegt; bei der Charakterisierung der Kleidung sind das die Wendungen *ein schwarzer Rock, ein schwarzer Schleier, eine schwarze Krawatte,* allerdings kommt im Wörterbuch zum deutschen Sprachgebrauch kein Syntagma *schwarz + Schuhe* vor, obwohl das Substantiv *Schuhe* in der deutschen Gegenwartssprache unumstritten häufiger angewendet wird als die Substantive *Rock, Krawatte* und *Schleier* – das hat auch unsere Forschung gezeigt. Außerdem wird der Gebrauch des Wortes *Schuhe* durch solche Merkmale wie Geschlecht, Alter oder Sozialstatus nicht begrenzt. Die Wortverbindung *schwarzer Bart*, die sich durch hohe χ^2- und Φ-Werte auszeichnet, wird vom Verfasser des Wörterbuches auch nicht festgelegt verzeichnet. Im Bereich des Äußeren des Menschen beschränkt sich E. Agricola nur auf die Wendungen *schwarzes Haar* und *schwarze Augen*. Übrigens haben wir festgestellt, dass das Farbwort *schwarz* mit dem Substantiv *Augen* (22 Belege) keine statistisch signifikante Verbindung bildet.

Nach der genauen Untersuchung der Wortverbindungen mit dem Farbadjektiv *rot* haben wir entdeckt, dass es zwischen dem Ergebnis unserer Forschung und den Angaben des Wörterbuches zum deutschen Sprachgebrauch viel Gemeinsames gibt. Die beiden Listen der stabilen Kontexte fangen mit der Verbindung *rote Rosen* an und enthalten die Wendungen *rote Grütze* und *rote Lippen*. Offensichtlich könnte eine solche Kongruenz im lexikalischen Stoff von der Unvoreingenommenheit und der Objektivität der Ergebnisse unserer Analyse zeugen. Allerdings erscheint in Agricolas Wörterbuch zuerst der Ausdruck *rote Lippen* und dann erst die Wortverbindung *rote Grütze*.

Die Ergebnisse der Widerspiegelung der typischen Verbindungen mit dem Farbwort *rot* im Wörterbuch von E. Agricola unterscheiden sich natürlich ein wenig von den Ergebnissen unseres statistischen Experimentes. Im Wörterbuch finden wir die Ausdrücke *rote Farbe* und *rote Tinte*, deren Sinnbereich durch die Wortverbindung *ein roter Fleck* mit syntagmatischen Beziehungen von großer statistischen Signifikanz (χ^2 = 37,70; Φ = 0,10) ergänzt werden kann, nichtsdestoweniger ist sie in lexikographischen Quellen immer noch nicht zu finden. Andererseits führt der Verfasser viele Ausdrücke mit Somatismen an (*rote Backen, rote Wangen, ein roter Kopf, rote Augen, eine rote Nase, ein roter Mund* etc.); dazu könnte man noch die Verbindung *ein rotes Gesicht* ergänzen, weil sie auch durch einen hohen Grad der Gebundenheit ihrer Komponenten gekennzeichnet ist.

Das Farbadjektiv *blau* hat drei statistisch signifikante Verbindungen im Modell *[Wort + Wort]*. Alle Ausdrücke (*eine blaue Blume, blauer Himmel* und *blaue Augen*) erscheinen auch (allerdings in einer anderen Reihenfolge) im Wörterbuch von E. Agricola: Auf der ersten Position steht die Wortverbindung *blauer Himmel*, und dann folgt die Fügung *eine blaue Blume*. Offensichtlich können für eine solche Reihenfolge keine gewichtigen Gegenargumente vorgebracht oder zur Diskussion gestellt werden, obwohl die größten Werte des Chi-Quadrats und des Koeffizienten Φ, wie unsere Ergebnisse aufzeigen, für das Modell *blau + Blume* (χ^2 = 202,75; Φ = 0,24) festgestellt wurden.

Man kann vermuten, dass der Grund für diese Befunde darin besteht, dass die lexikalisch-semantische Varianten des Wortes *blau* während der statistischen Analyse nicht berücksichtigt bzw. abgegrenzt seien. So z.B. kommt das erwähnte Modell *blau + Blume* (23 Belege) im Deutschen in zwei Ausdrücken vor: *eine blaue Blume* (direkte Bedeutung) und *die blaue Blume [der Romantik]* (übertragene Bedeutung). Nichtsdestoweniger widersprechen solche Angaben den allgemeinen Erkenntnissen über die Kombinierbarkeit des Adjektivs *blau* nicht. Die Wortverbindung *blaue Augen* nimmt den dritten Rang unter den typischen Kontexten (sowohl im Wörterbuch zum deutschen Sprachgebrauch als auch in unserer Forschung) zu

Recht ein.

Die Adjektive *grün* und *braun* haben je eine statistisch signifikante Verbindung mit sehr starken syntagmatischen Relationen zwischen den Komponenten, z.B.: *grünes Licht* und *braune Augen*. Beide Ausdrücke werden auch von E. Agricola verzeichnet.

Die Farbbezeichnung *grau* geht in drei Ausdrücke mit statistisch signifikanten Verbindungen ein; zwei davon (*graues Haar* und *graue Augen*) kann man unter typischen Wortverbindungen im Wörterbuch von E. Agricola finden: Der Verfasser führt auf der ersten Position den Ausdruck *graue Augen* auf, und dann kommt die Verbindung *graues Haar*, obwohl es nach den χ^2-Werten und dem Bernoullischen Koeffizienten Φ, die den erwähnten Mikrostrukturen entsprechen, umgekehrt sein müsste. Eine solche Reihenfolge im Wörterbuch kann eigentlich durch die Struktur der Stichwortdefinition erklärt werden.

Nach der Bezeichnung von E. Agricola hat das Adjektiv *grau* vier Bedeutungen: 1. ⟨*von grauer Farbe*⟩ (und zwar die angeführte Verbindung *graue Augen*); 2. ⟨*grauhaarig*⟩ (entsprechend: *graues Haar*); 3. ⟨*trübe, eintönig, unbestimmt*⟩ und 4. ⟨*längst vergangen*⟩ (s. Agricola 1977: 280-281).

Als eine der ungewöhnlichsten tritt die Wortverbindung *graue Straße* auf, die nach Ergebnissen der statistischen Forschung den höchsten Grad der Gebundenheit ihrer Komponenten hat (vgl. $\chi^2 = 155{,}15$; $\Phi = 0{,}21$) und als Erste ins Verzeichnis typischer Umgebungen eingetragen werden sollte. Allerdings ist es sehr schwer, diese Verbindung stereotypen Wendungen hinzuzufügen, weil wir das hochfrequente Modell *grau* + *Straße* (11 Belege) nur bei einem Autor – W. Borchert – festgestellt haben. Die okkasionelle Wortverbindung *graue Straße* hat sich durch die in diesem Fall verwendete Variante der Prozedur der distributiv-statistischen Analyse und die hohe „Gebräuchlichkeitsschwelle“ nicht abgrenzen lassen.

Eine statistische Analyse der Verbindungen des Farbadjektivs *gelb* mit einzelnen Substantiven wurde wegen ungenügender Häufigkeitscharakteristiken im Rahmen dieser Forschung nicht durchgeführt. Es heißt natürlich gar nicht, dass es im Deutschen keine stereotypen syntagmatischen Verbindungen mit *gelb* bzw. andere Wendungen mit den o.g. Farbwörtern gibt.

Die Ergebnisse unserer Analyse bestätigen im Grunde eine Gesetzmäßigkeit, laut der zwischen der Breite syntagmatischer Beziehungen des Adjektivs und seiner Gebrauchshäufigkeit ein direktes Verhältnis besteht. So haben die hochfrequenten Farbadjektive (*weiß, schwarz, rot*) auch den breitesten Umfang kontextueller Einheiten. Die festgestellten Divergenzen zwischen den Daten der Analyse des Wörterbuches und den Daten, die wir aufgrund der lexikalischen Kombinierbarkeit der Farbbezeichnungen mit zwanzig Substantiven bekommen haben, zeigen auf, dass *intuitive Auswahlkriterien beim Verfassen von Wörterbüchern nicht die ganze Tiefe syntagmatischer Verbindungen zwischen den Wörtern widerspiegeln*. Deshalb lässt die angebotene Methodik der Erforschung der lexikalischen Kombinierbarkeit bestimmte Berichtigungen zu den entsprechenden Wörterbüchern beitragen.

8. Paradigmatische Relationen zwischen den Adjektiven

Die paradigmatischen Beziehungen zwischen den Adjektiven können, wie V. Moskovič, A. Suprun und andere Forscher aufgezeigt haben, aufgrund der Eigenschaften ihrer Kombinierbarkeit behandelt werden. Als Ausgangsthese gilt dabei die folgende Hypothese: Die Ähnlichkeit der lexikalischen Kombinierbarkeit müsste von der semantischen Ähnlichkeit der Wörter zeugen, weil die lexikalische Bedeutung des Wortes von seiner potentiellen Kombinierbarkeit in größtem Maße abhängt (Zvegincev 1957). Folglich: Je mehr die Kombinierbarkeit zweier Wörter übereinstimmt, umso enger ist ihre paradigmatische Relation. Die Kom-

binierbarkeitseigenschaften zweier Wörter können mit Hilfe der Korrelationsanalyse paarweise berücksichtigt und gemessen werden. Eine solche Analyse haben wir aufgrund der Angaben in der Tab. 2 durchgeführt. Dabei wurden nur die Fälle berücksichtigt, bei denen die gesamte Gebrauchshäufigkeit des Adjektivs mit der Subklasse der Substantive mindestens 5 Belege betragen. Die Ergebnisse der Korrelationsanalyse sind in Tab. 8 angeführt.

Tabelle 8
Die paradigmatischen Relationen der Farbadjektive im Deutschen
(Korrelationskoeffizienten)

	schwarz	rot	blau	grün	grau	braun	gelb
weiß	0,93	0,72	0,71	0,49	0,61	0,56	0,61
schwarz		0,81	0,77	0,45	0,70	0,65	0,63
rot			0,69	0,46	0,58	0,88	0,69
blau				0,76	0,73	0,59	0,78
grün					0,38	0,32	0,83
grau						0,69	0,45
braun							0,59

In der vorliegenden Tabelle ist die Tatsache bemerkenswert, dass alle Koeffizienten das Zeichen "+" haben. Das bedeutet, dass alle untersuchten Adjektive im Großen und Ganzen mit allen ausgewählten Subklassen der Substantive kollokieren.

Der durchschnittliche Wert des Korrelationskoeffizienten dieser Klasse lässt sich als ziemlich hoch (r = 0,64) anerkennen. Eine solche positive Korrelation kann von einer bestimmten Tendenz zeugen, die in der Linguistik als *Tendenz der funktionalen Anziehung* (Golovin 1971: 165-166) bekannt ist; in unserem Fall ergibt sie sich, wenn die Aktivitätsverstärkung eines Farbadjektivs eine entsprechende Aktivitätsverstärkung bei einem anderen Adjektiv verursacht.

Ein hoher Grad der Korrelation (von 0,61 bis zu 0,93) ist bezeichnend für 18 Paare (oder 64,28 %) der Adjektive. Die größte Verbindungsstärke – also die Ähnlichkeit der Distribution – wurde zwischen vier Wortpaaren: *weiß - schwarz* (0,93); *rot - braun* (0,88); *grün - gelb* (0,83); *schwarz - rot* (0,81) festgestellt.

Durch mittlere und schwache Verbindungen lassen sich jeweils 5 (oder je 17,86 %) Paare der Adjektive charakterisieren. Ganz niedrige Korrelationskoeffizienten ergeben sich zwischen den Oppositionen *grün - grau* (0,38) und *grün - braun* (0,32). Übrigens haben die Adjektive *grau* und *grün* bei durchschnittlichen Werten der Gebräuchlichkeit die geringste Anzahl statistisch signifikanter paradigmatischer Verbindungen; d.h. dass einer größeren Menge der Fälle des gemeinsamen Gebrauchs der Wörter nicht immer eine größere Kraft der paradigmatischen Verbindung entspricht.

Eine solche Verteilung der Korrelationskoeffizienten lässt sich mit der Kombinierbarkeit der untersuchten Adjektive mit den Substantiven erklären. Wenn wir annehmen, dass das Adjektiv *weiß* mit dem Adjektiv *schwarz* am häufigsten korreliert, so bedeutet das, dass sich beide Adjektive mit identischen Subklassen der Substantive gleich intensiv verbinden. Als gemeinsam für beide Farbadjektive treten in unserer Forschung die Substantive auf, die das Äußere des Menschen; Eigennamen; den Sozialstatus des Menschen und Organisationen; Alltagsgegenstände; Materialien und Stoffe; Fauna; Nahrungsmittel; Naturerscheinungen und Objekte, Raumbegriffe; Kleidung und Schuhe; Bauten und ihre Elemente; Formen und Figuren; Hell-Dunkel-Begriffe, Lichtquellen und sonstige Begriffe bezeichnen.

Und umgekehrt, wenn die Adjektive untereinander schwach korrelieren, so heißt das, dass sie nicht mit gemeinsamen, sondern mit unterschiedlichen Gruppen der Substantive öfter

kollokieren. So wird das Adjektiv *grün* beispielsweise oft mit den Substantiven zur Bezeichnung der Materialien und Stoffe, der Flora, der Naturerscheinungen, Objekte und der Lichtquellen gebraucht; das Adjektiv *braun* seinerseits kommt mit den Substantiven vor, die das Äußere des Menschen, Nahrungsmittel sowie Formen und Figuren bezeichnen. Deshalb besteht zwischen diesen Adjektiven keine statistisch signifikante Korrelation.

Folglich zeigt ein hoher Korrelationskoeffizient zweier Adjektive darauf, dass sie semantisch miteinander verbunden sind, und ein niedriger Koeffizient zeugt davon, dass eine solche Verbindung nicht stark genug ist.

Darüber hinaus wird anhand dieser Tabelle deutlich, dass die Zahl paradigmatischer Verbindungen nicht immer von der Gebrauchshäufigkeit des Adjektivs abhängt. So haben wir die meisten paradigmatischen Verbindungen (jeweils sieben) bei den höchstfrequenten Farbadjektiven *weiß* und *schwarz* (entsprechend 569 und 551 Belege), bei der Farbbezeichnung *rot* (349 Belege) und dem Adjektiv mit der mittleren Häufigkeit *blau* (288 Belege) festgestellt. Andererseits haben die Farbwörter *braun* (148 Belege) und *gelb* (136 Belege) – als die niedrigstfrequenten Adjektive im gewählten Mikrosystem – je sechs statistisch signifikante paradigmatische Verbindungen, und je fünf paradigmatische Verbindungen wurden bei den Adjektiven mit der mittleren Häufigkeit *grün* (208 Belege) und *grau* (201 Belege) festgehalten. Im Allgemeinen hat sich der Korrelationskoeffizient zwischen solchen Charakteristiken wie den Gebrauchshäufigkeiten und der Zahl der paradigmatischen Verbindungen der untersuchten Farbadjektive als relativ hoch ($r = 0{,}71$) erwiesen.

Tabelle 9
Die quantitativen Charakteristiken der paradigmatischen und syntagmatischen Relationen der Farbadjektive

Farbadjektive	Zahl der syntagmatischen Beziehungen	Zahl der paradigmatischen Beziehungen	Kombinierbarkeitsweite	Gebräuchlichkeit	Gebrauchshäufigkeit
weiß	15	7	1	1,82	569
schwarz	16	7	1	2,16	551
rot	14	7	1	2,09	349
blau	12	7	0,88	1,87	288
grün	11	5	0,88	1,64	206
grau	10	5	0,94	1,47	201
braun	9	6	0,88	1,28	148
gelb	11	6	0,88	1,29	136

Das analysierte Mikrosystem zeigt sich wie ein zusammenhängender lexikalisch-semantischer Komplex, dessen Glieder mit nach der Kraft und dem Charakter verschiedenen Verbindungen untereinander vereinigt sind. Alle Komponenten demonstrieren freilich vollkommen die Eigenschaften der ganzen Gruppe, aber am vollkommensten (wie die Ergebnisse der Analyse bezeugen) ist dies in den Farbwörtern *weiß, schwarz, rot, blau* und *gelb* repräsentiert. Übrigens fixieren die Wissenschaftler, die die Psychologie des Denkens, insbesondere den Übergang von der Empfindung bis zu dem Gedanken, erforschen, die Bezeichnungen gerade für diese und einige andere Farben in Form eines festen Kernes, der in verschiedenen Sprachen deutlich sichtbar ist. Offensichtlich handelt es sich hier um Bezeichnungen für sog. primäre, abstrakte Grundfarben, deren Mischung eine bestimmte Lexikalisierung im Bereich

anderer Farben determiniert, z.B.: *rot + blau = violett, rot + gelb = orange, blau + gelb = grün, schwarz + weiß = grau* etc.

Wir haben festgestellt, inwieweit die syntagmatischen und die paradigmatischen Relationen voneinander und von der Gebrauchshäufigkeit der untersuchten Adjektive abhängen. Dabei wurden die Angaben über die Zahl der paradigmatischen (starke und mittlere Beziehungen) und der syntagmatischen Verbindungen (mindestens 5 Belege) sowie ihre Gebrauchshäufigkeit in eine Tabelle (s. Tab. 9) eingetragen und dann wurde eine Korrelationsanalyse durchgeführt.

Wie unsere Analyse aufgezeigt hat (s. Tab. 10), lassen sich solche Parameter wie die Zahl der syntagmatischen Beziehungen und die Gebräuchlichkeit der Adjektive durch eine sehr große positive Abhängigkeit (r = 0,97) charakterisieren. Zwischen der Gebräuchlichkeit und der Kombinierbarkeitsweite existiert auch eine deutliche Korrelation (r = 0,92): je mehr feste syntagmatische Verbindungen das Adjektiv hat, desto öfter wird es in der Rede gebraucht. Andererseits zeugt die Verminderung des aktualisierten Kombinierbarkeitspotenzials eines Wortes von seinem Ausgang aus dem aktiven Wortschatz.

Tabelle 10
Das Verhältnis zwischen den quantitativen Charakteristiken der Farbadjektive

	Paradigmatik	Kombinierbarkeitsweite	Gebräuchlichkeit	Frequenz
Syntagmatik	0,94	0,93	0,97	0,86
Paradigmatik		0,95	0,94	0,71
Kombinierbarkeitsweite			0,92	0,65
Gebräuchlichkeit				0,79

Die Zahl der paradigmatischen Verbindungen steht hauptsächlich in einem direkten starken Verhältnis zur Zahl der syntagmatischen Verbindungen (r = 0,94), der Kombinierbarkeitsweite (r = 0,95) und der Gebräuchlichkeit der Farbadjektive (r = 0,94); dabei ergibt sich bei df = 8 – 2 = 6 FG und dem Signifikanzniveau 0,05 der minimale Korrelationskoeffizient $R_{0,05;\ 6}$ = 0,71; beim Signifikanzniveau 0,01 (1%) ist der minimale Korrelationskoeffizient gleich $R_{0,01;\ 6}$ = 0,83. Zwischen der Gebrauchshäufigkeit und der Zahl der paradigmatischen Beziehungen besteht eine ganz geringe gegenseitige Abhängigkeit von 0,65; davon zeugt auch die Verteilung paradigmatischer Beziehungen unter den Farbadjektiven mit verschiedener Frequenz.

Die Zahl der paradigmatischen Relationen kann also als Ergebnis der komplizierten Wechselwirkung der Gebrauchshäufigkeit des Wortes, seiner Kombinierbarkeitsweite, der Zahl seiner syntagmatischen Verbindungen (d.h. der Zahl seiner semantischen Komponenten) und der Menge gemeinsamer semantischer Übereinstimmungen betrachtet werden.

9. Der Aufbau eines Modells des semantischen Feldes

Das semantische Feld der Farbbezeichnungen kann aufgrund der Werte der Korrelationskoeffizienten aufgebaut werden. Dabei lassen sich nur „starke“ und „mittlere“ paradigmatische Beziehungen berücksichtigen, wie V. Moskovič aufgezeigt hat.

Da bei df = 17-2 = 15 FG und dem Signifikanzniveau 0,05 der minimale Korrelationskoeffizient $R_{0,05;15}$ = 0,48 und beim Signifikanzniveau 0,01 der minimale Korrelationskoeffizient $R_{0,01;15}$ = 0,61 sind, haben wir sie in drei Gruppen eingeteilt: starke 0,93-0,61, mittlere 0,60-0,48 und schwache Werte 0,47-0,34.

Das Modell des semantischen Feldes der Farbbezeichnungen in der deutschen Sprache ist aufgrund der starken und mittleren Werte der Korrelationskoeffizienten aufgebaut worden.

Es wurde festgestellt, dass die Korrelationspaare der Farbadjektive eine große Menge starker Verbindungen bilden, sodass in der Figur nur die starken Verbindungen zwischen den Wörtern mit den Koeffizienten von 0,61 bis 0,93 (eine ununterbrochene Linie) und die mittleren Verbindungen - von 0,48 bis zu 0,60 (eine unterbrochene Linie) (s. Abb. 1) - dargestellt sind.

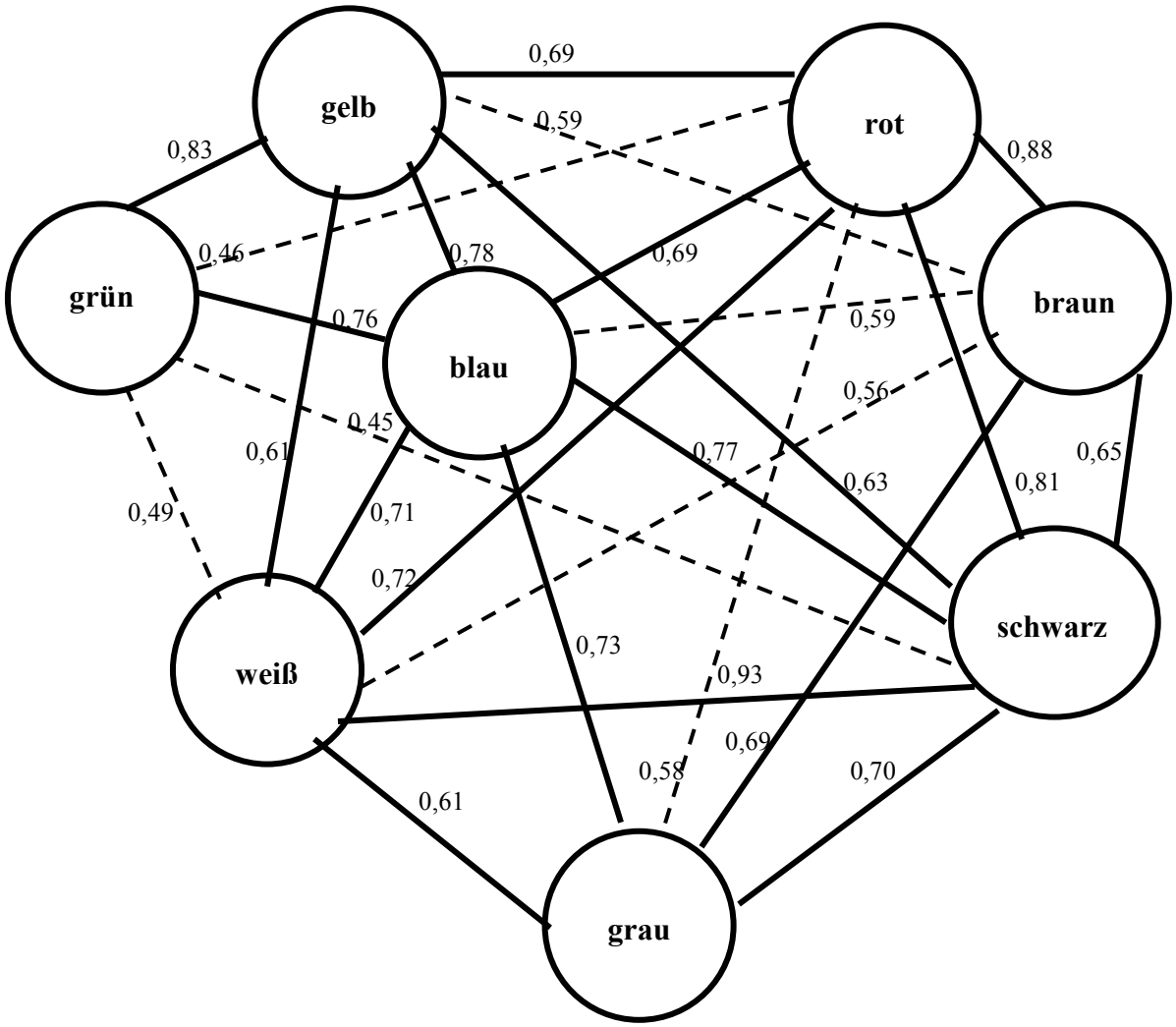


Abbildung 1. Das Schema der paradigmatischen Verbindungen der untersuchten Farbadjektive.

Schlussfolgerungen

Beim Blick auf die graphische Darstellung des untersuchten Mikrosystems, die die Kraft semantischer Beziehungen zwischen den Farbadjektiven auf syntagmatischer und paradigmatischer Ebene der Sprache veranschaulicht, können folgende Schlussfolgerungen gemacht werden:

Die lexikalischen Einheiten des semantischen Feldes der Farben werden nach „kalten“ und „warmen“ Farbtönen gruppiert.

Einerseits bilden die dunklen und kalten Farben (*schwarz, grau, braun, blau*) eine Gruppe, und andererseits gehören die hellen und warmen Farben zu einer anderen Gesamtheit (*weiß, rot, gelb, grün*).

Das Adjektiv *blau*, dass zusammen mit *schwarz* die meisten starken paradigmatischen Relationen hat, „vereinigt" beide Gruppen.

Die Adjektive, die bunte Farben bezeichnen, schließen sich (im Gegensatz zu den Adjektiven zur Bezeichnung unbunter Farben) nicht zu einem abgesonderten Komplex zusammen.

Das Lexem *schwarz* hat starke Relationen sowohl zu den Adjektiven *weiß* und *grau* als auch zu *gelb, rot* etc.

Andererseits wurden in der Gruppe unbunter Farbwörter in allen Paaren (*weiß - grau, weiß - schwarz* und *grau - schwarz*) auch starke Beziehungen festgestellt.

Wenn man die Farbbezeichnungen in einer Reihenfolge dem Farbspektrum gemäß (*rot – gelb – grün – blau...*) anordnet, wäre daraus noch eine Besonderheit zu ersehen, dass die Verbindungsstärke (Korrelationskoeffizienten) zwischen den benachbarten Farbbegriffen im Allgemeinen viel größer ist, als die zwischen den im Farbenband entfernten Bezeichnungen: z.B. *rot – gelb* 0,69 gegenüber *rot – grün* 0,46; *gelb – grün* 0,83 gegenüber *gelb – blau* 0,78; *grün – blau* 0,76 gegenüber *grün – rot* 0,46 oder *blau – rot* 0,69.

Darüber hinaus bilden diese Adjektive eine Menge relativ starker Verbindungen mit *braun*, das als Bezeichnung der Mischfarbe nicht zum Spektrum gehört, aber im Farbton wohl zwischen *rot* und *gelb* liegt: *braun – rot* 0,88; *braun – gelb* 0,59 gegenüber *braun – grün* 0,32. Es geht hier wohl auch um die Platzbestimmtheit durch die Begriffsnachbarn.

Zusammenfassend kann man behaupten, dass sich das untersuchte Mikrosystem durch das Vorhandensein von festen, statistisch signifikanten Verbindungen zwischen fast allen seinen Elementen charakterisieren lässt.

Literaturverzeichnis

Agricola, E. (1977). *Wörter und Wendungen. Wörterbuch zum deutschen Sprachgebrauch* / Hrsg. von E. Agricola unter Mitwirkung von H. Görner und R. Küfner. Leipzig: Bibliographisches Institut.

Amosova, N.N. (1963). *Osnovy anglijskoj frazeologii*. Leningrad: Leningradskij universitet.

Apressjan, J.D. (1969). Sinonimia i sinonimy. *Voprosy jazykoznania 4, 1969, 75-91.*

Berlin B., Kay P. (1969). *Basic color terms. Their universality and evolution*. Berkeley and Los Angeles: Univ. of California Press.

Bystrowa, L.V. (1977). *Prilagatelnye so značenijami "silnyj" i "slabyj" v sovremennom anglijskom jazyke. (Semantiko-statističeskoe issledovanie):* Autoreferat dissertacii. Kiev.

Duden. (1996*). Deutsches Universalwörterbuch* / Hrsg. von G. Drosdowski. Mannheim / Wien / Zürich: Duden-Verlag.

Frumkina, R.M. (1984). *Cvet, smysl, schodstvo: Aspekty psicholingvističeskogo analiza*. Moskva: Nauka.

Golovin, B.N. (1971). *Jazyk i statistika.* Moskva: Prosvesčenie.

Levickij, V., Hikow, L. (2004). Zum Gebrauch der Wortarten im Autorenstil. *Glottometrics 8, 2004, 12-22.*

Levickij, V.V. (1989). *Statističeskoe izučenie leksičeskoj semantiki*. Kiev: Minvuz.

Levickij, V.V., Ogoui, O.D., Kijko S.V., Kijko J.E. (2000). *Aproximatyvni metody vyvčennja lexyčnoho skladu.* Tscherniwzi: Ruta.

Merten, K. (1983). *Inhaltsanalyse. Einführung in Theorie, Methode und Praxis.* Opladen: Westdeutscher Verlag.

Moskovič, V.A. (1969*). Semantika i statistika*. Moskva: Nauka.

Pawlowski, A. (1999). The quantitative approach in cultural anthropology: application of linguistic corpora in the analysis of basic color terms. *J. of Quantitative Linguistics 6, 222-234.*

Zvegincev, V.A. (1957*). Semasiologia.* Moskva: Moskovskij universitet.

Terechova, D.I. (1997). Associatyvnyj portret ukrajinc'a ta rossijanyna. *Movoznavstvo 6, 43-50.*

Weisgerber, L. (1962*). Sprachliche Gestaltung der Welt*. Düsseldorf: Schwann.

Wierzbicka, A. (1996*). Jazyk. Kultura. Poznanie*. Moskva: Progress.

Wyler, S. (1992). *Colour and language. Colour Terms in English*. Tübingen: Gunter Narr Verlag.

Anlagen

Anlage 1

Bestandsaufnahme der Farbbezeichnungen im Deutschen
(Anfangsregister nach ein-/zweisprachigen Wörterbüchern)

№	Farbwörter	LEXIKOGRAFISCHE QUELLEN					
		I*	II*	III*	IV*	V*	VI*
1.	altgolden	+	+				
2.	altrosa		+	+		+	
3.	amarant	+	+	+			+
4.	apfelgrün		+	+			
5.	aschbleich		+				+
6.	aschblond	+	+	+			+
7.	aschfahl	+	+			+	+
8.	aschfarben		+			+	+
9.	aschgrau	+	+	+	+	+	+
10.	azurblau		+	+	+	+	+
11.	azurn	+	+	+	+	+	+
12.	beige	+	+	+	+	+	+
13.	beigefarben			+	+	+	+
14.	bernsteinfarben		+	+		+	+
15.	bernsteingelb					+	
16.	blaßblau			+			+
17.	blau	+	+	+	+	+	+
18.	blauäugig	+	+	+	+	+	+
19.	blaugrau	+	+	+		+	+
20.	blaugrün	+	+	+		+	
21.	bläulich	+	+	+	+	+	+
22.	blaurot	+	+	+		+	+
23.	blauschwarz	+	+	+		+	
24.	blaustichig		+	+			+
25.	blauviolett					+	
26.	bleich	+	+	+	+	+	+
27.	bleiern	+	+	+	+	+	+

28.	bleifarben		+	+		+	+
29.	bleifarbig		+			+	+
30.	bleigrau		+				
31.	bleu	+	+	+		+	+
32.	blitzblau		+	+			+
33.	blond	+	+	+	+	+	+
34.	blütenweiß	+	+	+			+
35.	blutrot	+	+	+		+	+
36.	bordeaux		+	+			+
37.	bordeauxrot		+	+			+
38.	brandrot	+	+	+			+
39.	braun	+	+	+	+	+	+
40.	braunäugig		+	+		+	+
41.	braungebrannt		+	+	+	+	+
42.	braunhaarig		+	+		+	
43.	bräunlich	+	+	+		+	+
44.	braunrot					+	+
45.	bronzefarben		+	+			+
46.	bronzenfarbig			+			+
47.	bronzen		+	+	+	+	+
48.	brünett	+	+	+	+	+	+
49.	champagner			+			+
50.	champagnerfarben			+			
51.	champagnerfarbig			+			
52.	cerise		+	+		+	+
53.	chamois		+	+			+
54.	creme	+	+	+		+	+
55.	cremefarben	+	+	+	+	+	+
56.	cremefarbig		+	+			+
57.	dämmergrau		+				+
58.	dunkel	+	+	+	+	+	+
59.	dunkelblau		+	+		+	+
60.	dunkelbraun			+			+
61.	dunkelgelb		+	+			
62.	dunkelgrün		+	+			+
63.	dunkelfarbig		+	+			+
64.	dunkelrot		+	+	+	+	+
65.	dunkelviolett		+				
66.	dunkelweiß					+	+
67.	eisblau	+	+	+		+	+
68.	eisgrau	+	+	+		+	+
69.	elfenbeinfarben					+	+
70.	elfenbeinfarbig					+	
71.	feldgrau	+	+	+			+
72.	feuerrot	+	+	+		+	+
73.	flachsblond	+	+	+		+	+
74.	flachsfarben					+	
75.	fleischfarben	+	+	+		+	+

76.	fleischfarbig	+	+	+		+	+
77.	fliederfarben	+		+	+	+	+
78.	fliederfarbig	+	+	+		+	+
79.	fuchsig	+	+	+			+
80.	fuchsrot				+	+	+
81.	gelb	+	+	+	+	+	+
82.	gelbbraun	+	+	+	+		+
83.	gelbgrün	+	+	+		+	
84.	gelbicht						+
85.	gelblich	+	+	+	+	+	+
86.	gelblich grün			+			
87.	gelblich weiß			+			+
88.	gelbrot	+	+	+			
89.	gämsfarben	+	+	+			+
90.	gämsfarbig		+	+			+
91.	giftgrün	+	+	+	+	+	+
92.	glutrot		+	+	+	+	+
93.	goldblond	+	+	+			
94.	goldbraun		+	+			+
95.	golden	+	+	+	+	+	+
96.	goldfarben		+	+		+	+
97.	goldfarbig		+	+		+	+
98.	goldgelb	+	+	+	+		+
99.	goldgrün	+	+	+			
100.	goldrot	+	+	+			
101.	grasgrün	+	+			+	+
102.	grau	+	+	+	+	+	+
103.	graubärtig		+	+		+	+
104.	graublau	+	+	+	+	+	+
105.	graubraun	+	+	+	+		+
106.	graugrün	+	+				+
107.	grauhaarig	+	+	+	+	+	+
108.	graulich	+	+	+		+	+
109.	gräulich	+	+	+		+	+
110.	graumeliert	+	+	+	+		+
111.	grauschwarz		+			+	
112.	grauweiß					+	+
113.	grün	+	+	+	+	+	+
114.	grünblau	+	+	+	+	+	+
115.	grüngelb		+	+		+	+
116.	grünlich	+	+	+	+	+	+
117.	haselnußfarben		+				
118.	hell	+	+	+	+	+	+
119.	hellblau		+	+	+	+	
120.	hellblond		+	+	+	+	
121.	hellbraun		+	+		+	+
123.	hellgelb		+	+			+
124.	hellgrau		+	+			

125.	hellgrün		+	+			
126.	hellrot		+	+		+	
127.	himmelblau	+	+	+	+	+	+
128.	hochrot		+	+	+		+
129.	immergrün	+	+	+		+	+
130.	indigoblau	+		+		+	+
131.	infrarot	+	+	+	+	+	+
132.	jade		+				+
133.	jadegrün	+	+	+		+	+
134.	kackbraun			+			
135.	kaffeebraun	+	+	+		+	+
136.	kalkweiß			+			+
137.	karmesinrot		+	+		+	+
138.	karminrot		+	+		+	+
139.	käseweiß		+	+		+	+
140.	kastanienbraun	+	+	+	+	+	+
141.	khakibraun		+	+			
142.	khakifarben		+	+		+	+
143.	khakifarbig		+	+			+
144.	kirschrot		+	+	+	+	+
145.	knallbunt		+	+			+
146.	knallgelb					+	
147.	knallrot		+	+	+	+	+
148.	kobaltblau	+	+	+		+	+
149.	kohlpechrabenschwarz	+	+			+	
150.	kohlrabenschwarz	+	+	+	+	+	+
151.	kohlschwarz	+	+	+	+	+	+
152.	königsblau	+	+	+		+	+
153.	kornblumenblau	+	+	+	+	+	+
154.	krebsrot	+	+	+		+	+
155.	kupferfarben		+	+	+		+
156.	kupferfarbig		+				+
157.	kupferrot	+	+	+		+	+
158.	lachsrosa			+			
159.	lachsfarben	+	+	+	+	+	+
160.	lachsfarbig			+		+	+
161.	lederbraun			+			+
162.	lederfarben			+			+
163.	lederfarbig			+			+
164.	leichenblaß	+	+	+	+	+	+
165.	lichtblau	+	+	+			+
166.	lila	+	+	+	+	+	+
167.	lilafarben			+			+
168.	lilafarbig			+			+
169.	lindgrün	+	+			+	
170.	malvenfarben	+	+	+		+	
171.	malvenfarbig	+	+	+		+	+
172.	marengo	+	+				

173.	marineblau						+
174.	mausfarben		+				+
175.	mausfarbig		+				+
176.	mausgrau		+			+	+
177.	mauve	+	+	+			+
178.	mehrfarbig	+	+	+	+	+	+
179.	milchig	+	+	+	+	+	+
180.	nachtblau	+	+	+			
181.	naturfarben	+	+	+		+	
182.	nebelgrau		+	+			+
183.	negroid	+	+	+		+	+
184.	noir		+				
185.	nußbraun		+	+			+
186.	oliv	+	+	+		+	+
187.	olivgrün	+	+	+	+	+	+
188.	orange	+	+	+		+	+
189.	orangenfarben	+	+	+	+	+	+
190.	orangenfarbig		+	+		+	+
191.	purpurfarben					+	+
192.	purpurfarbig					+	+
193.	purpurn		+	+		+	+
194.	purpurrot					+	+
195.	puterrot	+		+	+	+	+
196.	rabenschwarz	+	+	+	+	+	+
197.	rosa	+	+	+	+	+	+
198.	rosafarben			+		+	+
199.	rosafarbig			+		+	+
200.	rosarot		+			+	+
201.	rosé	+	+	+		+	
202.	rosenrot		+	+		+	+
203.	rosig	+	+	+	+	+	+
204.	rostbraun		+	+	+	+	
205.	rostfarben	+	+	+	+	+	+
206.	rostfarbig		+	+		+	+
207.	rostrot	+	+	+		+	+
208.	rot	+	+	+	+	+	+
209.	rotbärtig	+	+	+		+	+
210.	rotblond	+	+	+	+	+	+
211.	rotbraun	+	+	+	+	+	+
212.	rotglühend	+	+	+	+	+	+
213.	rothaarig	+	+	+	+	+	+
214.	rotwangig	+	+	+	+	+	+
215.	rötlich	+	+	+	+	+	+
216.	rotnasig		+	+			
217.	rouge		+				+
218.	rubinrot	+	+	+	+	+	+
219.	rubinfarben			+			
220.	rubinfarbig			+			

221.	safflorgelb		+	+			
222.	safrangelb		+	+			+
223.	sandfarben	+	+	+		+	+
224.	sandfarbig	+	+	+		+	+
225.	sattblau		+	+			
226.	sattgrün		+	+			
227.	schamrot	+	+	+	+	+	+
228.	schneeweiß		+	+	+	+	+
229.	schokoladenbraun			+	+	+	
230.	schokoladenfarben			+			+
231.	schokoladenfarbig			+			+
232.	schwarz	+	+	+	+	+	+
233.	schwarzblau	+	+			+	
234.	schwarzbraun	+	+	+	+	+	+
235.	schwarzbunt		+	+			
236.	schwarzgallig		+				+
237.	schwarzgrau	+	+				
238.	schwarzgrün	+	+				
239.	schwarzhaarig	+	+	+	+	+	+
240.	schwärzlich	+	+	+		+	+
241.	schwarz-weiß	+	+	+		+	+
242.	schwefelgelb	+	+	+		+	+
243.	schwefelfarben			+			
244.	schwefelfarbig			+			
245.	semmelblond	+	+	+		+	+
246.	senffarben	+	+	+		+	
247.	senffarbig			+			
248.	silberfarben			+		+	+
249.	silberfarbig			+			
250.	silbergrau	+	+	+		+	
251.	silberig	+	+	+			+
252.	silbern	+	+	+	+	+	+
253.	silberweiß	+	+	+		+	+
254.	silbrig	+	+	+	+	+	+
255.	stockdunkel		+	+		+	+
256.	stockfinster		+	+		+	+
257.	strohgelb	+	+	+	+	+	+
258.	strohfarben			+			
259.	strohfarbig			+			
260.	taubenblau	+	+	+		+	
261.	taubengrau	+	+	+		+	+
262.	tiefblau	+	+	+		+	
263.	tiefrot	+	+				
264.	tiefschwarz	+	+	+			
265.	tomatenrot		+	+			
266.	türkis	+	+	+		+	+
267.	türkisfarben	+	+	+		+	+
268.	türkisfarbig		+	+			+

269.	türkisgrün	+	+			+	
270.	ultramarin	+				+	+
271.	ultrarot	+	+	+			+
272.	ultraviolett	+	+	+		+	+
273.	veilchenblau	+	+	+	+		+
274.	veilchenfarben						+
275.	veilchenfarbig						+
276.	violett	+	+	+	+	+	+
277.	wächsern	+		+	+	+	+
278.	wachsbleich		+	+			+
279.	wasserblau	+	+	+			
280.	weinrot	+	+	+		+	+
281.	weiß	+	+	+	+	+	+
282.	weißblond	+	+	+		+	+
283.	weißgelb	+	+				+
284.	weißgrau	+	+	+			+
285.	weißlich	+	+	+	+	+	+
286.	ziegelrot		+	+	+	+	+
287.	zimtfarben			+	+	+	+
288.	zimtfarbig			+		+	+
289.	zinnoberrot			+			+
290.	zitronengelb		+	+		+	+
291.	zitronenfarben						+
292.	zitronenfarbig						+

Anm.:

I – KNAUR. Das deutsche Wörterbuch;
II – Wahrig. Deutsches Wörterbuch;
III – DUDEN. Deutsches Universalwörterbuch;
IV – Deutsch-russisches Wörterbuch von Daum/Schenk;
V – PONS-Globalwörterbuch Deutsch-Englisch;
VI – Das große deutsch-russisches Wörterbuch von O.I. Moskalskaja

Anlage 2
Die Gebrauchshäufigkeit der Farbbezeichnungen
(nach einer Stichprobe aus deutschen Prosawerken)

№	Farbbezeichnungen	Frequenz
1.	weiß	569
2.	schwarz	551
3.	rot	349
4.	blau	288
5.	grün	206
6.	grau	201
7.	braun	148
8.	gelb	136
9.	rosa	93
10.	blond	33
11.	lila	28
12.	golden	27
13.	dunkelblau	22
14.	gelblich	20
15.	hellblau	19
16.	knallrot	18
17.	schneeweiß	18
18.	dunkelbraun	17
19.	rötlich	17
20.	kinderkackegelb	16
21.	bläulich	15
22.	dunkelrot	14
23.	orange	14
24.	rosig	14
25.	beige	13
26.	himmelblau	13
27.	bräunlich	12
28.	dunkel	12
29.	rosarot	12
30.	schwarzweiß	12
31.	violett	12
32.	goldbraun	11
33.	rotbraun	10
34.	blaßblau	9
35.	blutrot	9
36.	grünlich	8
37.	purpurn	8
38.	dunkelgrün	7
39.	graublau	7
40.	orangenfarben	7
41.	tiefschwarz	7
42.	blaßbraun	6
43.	blauschwarz	6
44.	blauweiß	6
45.	bordeauxrot	6
46.	hellbraun	6
47.	hochrot	6
48.	orangerot	6
49.	pechschwarz	6
50.	rothaarig	6
51.	braungebrannt	5
52.	goldgelb	5
53.	grün-weiß	5
54.	hellgrün	5
55.	schwärzlich	5
56.	silbern	5
57.	tiefblau	5
58.	wasserblau	5
59.	ziegelrot	5
60.	blaurot	4
61.	graubraun	4
62.	graugelb	4
63.	grauschwarz	4
64.	grauweiß	4
65.	hellgelb	4
66.	hellgrau	4
67.	kalkweiß	4
68.	kupferrot	4
69.	mausgrau	4
70.	pinkfarben	4
71.	rosafarben	4
72.	rostrot	4
73.	staubgrün	4
74.	türkisblau	4
75.	blauäugig	3
76.	brandrot	3
77.	braunschwarz	3
78.	brünett	3
79.	champagnerfarben	3
80.	dunkelgrau	3
81.	feldgrau	3
82.	gelbbraun	3
83.	giftgrün	3
84.	grün-weiß-gestreift	3
85.	hellblond	3
86.	hellrot	3
87.	knallgelb	3

88.	kohlschwarz	3
89.	marineblau	3
90.	mattgelb	3
91.	meerfarben	3
92.	ockergelb	3
93.	orange-blau	3
94.	purpurrot	3
95.	rotblond	3
96.	schwarzgrün	3
97.	schwarzseiden	3
98.	silberweiß	3
99.	tiefrot	3
1.	weißblond	3
2.	weißlich	3
3.	anthrazitgrau	2
4.	aschblond	2
5.	azurblau	2
6.	beigebraun	2
7.	bierflaschengrün	2
8.	blaßgrün	2
9.	blaßrosa	2
10.	blauseiden	2
11.	blau-weiß	2
12.	blütenweiß	2
13.	braungelb	2
14.	braungolden	2
15.	braun-orange-weiß-kariert	2
16.	cremefarben	2
17.	dunkelrosa	2
18.	eidottergelb	2
19.	fahlgelb	2
20.	flaschengrün	2
21.	fleischfarben	2
22.	giftrot	2
23.	glutrot	2
24.	goldblond	2
25.	goldfarben	2
26.	grasgrün	2
27.	graugrün	2
28.	grüngolden	2
29.	grünweiß	2
30.	herzrot	2
31.	hoffnungsgrün	2
32.	kackbraun	2
33.	kastanienbraun	2
34.	kaviarschwarz	2
35.	lachsrosa	2

36.	milchweiß	2
37.	mittelblau	2
38.	nachtblau	2
39.	naturfarben	2
40.	porzellanweiß	2
41.	purpurblau	2
42.	rabenschwarz	2
43.	rosagrau	2
44.	rostbraun	2
45.	safrangelb	2
46.	scharlachrot	2
47.	schmutzig-gelb	2
48.	schwarzblau	2
49.	schwarzgrau	2
50.	schwarz-rot-golden	2
51.	schwarzweiß gewürfelt	2
52.	schwefelgelb	2
53.	senfgelb	2
54.	silbergrau	2
55.	tomatenrot	2
56.	türkis	2
57.	wachsfahl	2
58.	wasserstoffblond	2
59.	weinrot	2
60.	weißblau	2
61.	weiß-blau gestreift	2
62.	weißgelb	2
63.	weißgelblich	2
64.	weißgrau	2
65.	zinkgrün	2
66.	zitronengelb	2
67.	zornrot	2
68.	altgolden	1
69.	altrosa	1
70.	aschgrau	1
71.	backsteinrot	1
72.	basaltgrau	1
73.	beige-braunrot gesprenkelt	1
74.	beigefarben	1
75.	beige-grau	1
76.	beigegrünlich	1
77.	bernsteinfarben	1
78.	blaßgolden	1
79.	blaßkotzgrün	1
80.	blaublütig	1
81.	blaudunkel	1
82.	blau-durchsichtig	1

83.	blauflimmernd	1
84.	blaugeblümt	1
85.	blauglänzend	1
86.	blau-gold-marmoriert	1
87.	blaugrün	1
88.	bläulichweiß	1
89.	blau-orange-gestreift	1
90.	blauschimmernd	1
91.	blauweißgestreift	1
92.	blau-weiß-rot	1
93.	bleichgrün	1
94.	bleifarben	1
95.	blendendweiß	1
96.	blühweiß	1
97.	blumenblau	1
98.	blutig-rot	1
99.	blutrosa	1
100.	blütweiß	1
101.	braunäugig	1
102.	braungeschminkt	1
103.	braun-golden	1
104.	braungrau	1
105.	braungrün	1
106.	bräunlich-grünlich-gräulich	1
107.	braunorangekariert	1
108.	braun-weiß	1
109.	brennendrot	1
110.	butterblumengelb	1
111.	butterbraun	1
112.	chinablau	1
113.	dreckgrün	1
114.	druckerschwarz	1
115.	dunkelgelb	1
116.	dunkelgraugrün	1
117.	düsterblau	1
118.	eierschalengelblich	1
119.	eisblau	1
120.	eisengrau	1
121.	elfenbeinfarben	1
122.	elfenbeingelb	1
123.	erdbraun	1
124.	ergrauend	1
125.	fahlgrau	1
126.	fahlweiß	1
127.	februargrau	1
128.	feuchtgrün	1
129.	feuerrot	1
130.	finster	1
131.	flachsblond	1
132.	fleischig-dunkelrot	1
133.	gammelbraun	1
134.	gelb-braun	1
135.	gelb-durchscheinend	1
136.	gelbgraubraun gestreift	1
137.	gelb-grün-weiß gesprenkelt	1
138.	gelblich-blaß	1
139.	gelblich-siffig	1
140.	gelblich-verkrümelt	1
141.	gelblichverraucht	1
142.	gelblichweiß	1
143.	gelbweiß	1
144.	goldgelb-weiß-schwarz	1
145.	goldig	1
146.	goldschwarz	1
147.	granatrot	1
148.	graubleich	1
149.	grau-braun gemustert	1
150.	graubraungrün	1
151.	graugestuft	1
152.	grau-grün gemustert	1
153.	graugrünbraun	1
154.	graulackiert	1
155.	gräulich-milchig	1
156.	gräulich-rosa	1
157.	graurot	1
158.	grau-schwarz patiniert	1
159.	grautrüb	1
160.	grell-gelb	1
161.	grellrot	1
162.	grobbraun	1
163.	grün-bräunlich	1
164.	grün-grau gestreift	1
165.	grünlich-grau	1
166.	hellbeige	1
167.	hellgelblich	1
168.	hellrosa	1
169.	himbeerfarben	1
170.	himbeerrot	1
171.	indischrot	1
172.	intensiv-blau	1
173.	kaffeebraun	1
174.	kanariengelb	1
175.	karamelfarben	1
176.	kardinalsrot	1

177.	karmesinrot	1
178.	khakibraun	1
179.	kirschrot	1
180.	knallgrün	1
181.	knattergelb	1
182.	königsblau	1
183.	kornblumenblau	1
184.	krebsrot	1
185.	kreideweiß	1
186.	kreidigweiß	1
187.	krötenfarben	1
188.	lackschwarz	1
189.	leuchtendblau	1
190.	leuchtendgelb	1
191.	leuchtendweiß	1
192.	lilafarben	1
193.	lila-rot geschminkt	1
194.	matschgrün	1
195.	mattbraun	1
196.	mattgrau	1
197.	mattgrün	1
198.	mattschwarz	1
199.	mattweiß	1
200.	mausbraun	1
201.	mauseschwarz	1
202.	mehlweiß	1
203.	messinggelb	1
204.	metallicblau	1
205.	metallic-blau	1
206.	mittelanthrazitgrau	1
207.	mittelbraun	1
208.	nachtschwarz	1
209.	naßgrau	1
210.	nikotingelb	1
211.	normalbraun	1
212.	nußbraun	1
213.	ochsenblutrot	1
214.	ockerbraun	1
215.	onyx-schwarz	1
216.	orange-beige gestreifelt	1
217.	orange-blau gewürfelt	1
218.	orange-braun	1
219.	originalblau	1
220.	pastellgrün	1
221.	pechrabenschwarz	1
222.	perlweiß	1
223.	pfirsichfarben	1
224.	pinkrot	1
225.	puterrot	1
226.	rasengrün	1
227.	rauchblau	1
228.	rehbraun	1
229.	rokokorosa	1
230.	rosablühend	1
231.	rosa-braun-gelb	1
232.	rosa-bräunlich	1
233.	rosaorangeschwarz	1
234.	rosenrot	1
235.	rotäugig	1
236.	rot-blau gestreift	1
237.	rotgelb	1
238.	rot-gelb	1
239.	rotgrün	1
240.	rötlichgelb	1
241.	rot-qualmig	1
242.	rotweiß gesprenkelt	1
243.	rotweiß gestreift	1
244.	rotweiß getüpfelt	1
245.	rot-weiß-rot	1
246.	samtschwarz	1
247.	sandbraun	1
248.	sandfarben	1
249.	sanftgraurosa	1
250.	sattgrün	1
251.	scharlachfarben	1
252.	schiefergrau	1
253.	schimmelgrün	1
254.	schlappgrün	1
255.	schlohweiß	1
256.	schmutzgelb	1
257.	schmutziggrau	1
258.	schmutzig-grün	1
259.	schmutzigrot	1
260.	schmutzig-weiß	1
261.	schnapsgrün	1
262.	schwachsilbern	1
263.	schwarzbraun	1
264.	schwarzrot	1
265.	schwarzsilbern	1
266.	schwarzviolett	1
267.	schwarz-weiß	1
268.	schwarzweiß gestreift	1
269.	schwarzweiß-lappig	1
270.	schwarzweißrot	1
271.	schweinchenrosa	1
272.	seidigblau	1

273.	siegellackrot	1
274.	silbermetallicfarben	1
275.	silbrig	1
276.	smaragden	1
277.	sonnenblumengelb	1
278.	stahlblau	1
279.	stahlgrau	1
280.	staubgelb	1
281.	staubgrünlich	1
282.	strohgelb	1
283.	stumpfweiß	1
284.	tabakgelb	1
285.	taubenblau	1
286.	tellerweiß	1
287.	teufelsbraun	1
288.	türkisgrün	1
289.	türkis-weiß	1
290.	vanillegelb	1
291.	veilchenfarben	1
292.	wachsgelb	1
293.	weißbläulich	1
294.	weiß-dunkelblau-golden	1
295.	weiß-gelb	1
296.	weiß-grün	1
297.	weißlich-gelb	1
298.	weiß-silbrig	1
299.	weizenblond	1
300.	wollweiß	1
301.	zartgrün	1
302.	zartrosa	1
303.	zartrosig	1
304.	zigarrenbraun	1
305.	zinnoberrot	1
306.	zornesrot	1
307.	zuckergußfarben	1

Anlage 3

Die Werte der Gebräuchlichkeit der Subklassen der Substantive mit den Farbadjektiven

№	Subklassen der Substantive	Farbadjektive							
		weiß	**schwarz**	**rot**	**blau**	**grün**	**grau**	**braun**	**gelb**
1.	Äußeres des Menschen	3,24	3,44	3,8	5,57	5,5	3,38	3,6	3
2.	Eigennamen etc.	1,38	2,46	2,67	3,67	1,43	1	2,75	1
3.	Personen	1	2,25	1,17	1	1,33	1	1	1
4.	Organisationen	1,17	1,25	1,5	1	1	1	1	-
5.	Alltagsgegenstände	1,27	1,45	1,29	1,56	1,48	1,13	1,33	1,29
6.	Materialien	1,38	1,62	1,4	1,43	1,9	1,63	1,25	1,17
7.	Fauna	2,33	1,38	1,09	1,8	1,2	1,5	-	2
8.	Flora	2	1	8	4,13	1,4	1	1,5	1,14
9.	Nahrungsmittel	2,45	3	3,75	1	1	-	1,29	1,5
10.	Natur, Raum	2,2	2,27	1,33	3	1,27	2,25	1,25	1,2
11.	Kleidung, Schuhe	3,04	2,58	1,93	1,41	1,92	1,11	1,2	1,55
12.	Bauten	2,08	1	1	1,33	1,17	1,43	1	1
13.	Zeitbegriffe	1	2,33	1	-	-	1,25	1	-
14.	Formen, Figuren	1,5	1,6	2,57	1,22	1	1	1,13	1
15.	Hell-Dunkel-Begriffe	1,43	2,5	1,11	-	4	1	1	2
16.	Färbung	1,2	4,67	1	1,17	-	1	1,5	1
17.	Sonstige Begriffe	2,25	1,09	1	2,5	2,33	1,86	-	2,17

Glottometrics 11, 2005, 98-119

Hidden communication aspects in the exponent of Zipf's law

Ramon Ferrer i Cancho[1]

Abstract. This article focuses on communication systems following Zipf's law, in a study of the relationship between the properties of those communication systems and the exponent of the law. The properties of communication systems are described using quantitative measures of semantic vagueness and the cost of word use. The precision and the economy of a communication system is reduced to a function of the exponent of Zipf's law and the size of the communication system. Taking the exponent of the frequency spectrum, it is demonstrated that semantic precision grows with the exponent, whereas the cost of word use reaches a global minimum between 1.5 and 2, if the size of the communication system remains constant. The exponent of Zipf's law is shown to be a key aspect for knowing about the number of stimuli handled by a communication system, and determining which of two systems is less vague or less expensive. The ideal exponent of Zipf's law, it is therefore argued, should be very slightly above 2.

Keywords: Zipf's law, frequency spectrum, exponent, precision, economy

INTRODUCTION

Word frequencies in human language arrange themselves according to what is known as Zipf's law. If $P(f)$ is the proportion of words whose frequency is f in a given sample (e.g. a text), we say that a sample follows Zipf's law (Zipf, 1932, 1935, 1949) if

$$P(f) \sim f^{-\beta}, \tag{1}$$

where β is the exponent of the law. We assume that $\beta > 1$.

The previous equation appears as a straight line when $P(f)$ is plotted on a logarithmic scale. Although different functions have been proposed for modelling $P(f)$ (Chitashvili & Baayen, 1993; Tuldava 1996; Naranan & Basubrahmanyan, 1998), the basic trend described in simplified form by Eq. 1 appears to hold without exceptions in word frequencies. This article uses the functional form in Eq. 1 because its simplicity is extremely helpful for the analytical calculations discussed here.

Typically, $\beta \approx 2$ is found (Zipf, 1932, 1935, 1949) but significant deviations from that value have been reported in single author samples:

- $\beta > 2$ in fragmented discourse schizophrenia. This type of speech is characterized by multiple topics and the absence of a consistent subject. The lexicon of such a text may be varied and chaotic (Piotrowski *et al.* 1995, Piotrowska *et al.,* to appear). $\beta \in [2.11,2.42]$ is found. Schizophrenic patients of this kind tend to be in the acute phase of the disease.

[1] Address for correspondence: Ramon Ferrer i Cancho, Dip. di Fisica, Università 'La Sapienza', Piazzale A. Moro 5, ROMA 00185, ITALY. E-mail: ramon@pil.phys.uniroma1.it

- Values suspiciously above the ideal $\beta = 2$ have been found in nouns from single author samples. More precisely, $\beta \in [2.15,2,32]$ (Balasubrahmanyan & Naranan, 1996).
- $1 < \beta < 2$ in advanced forms of schizophrenia (Whitehorn & Zipf, 1943; Zipf, 1949; Piotrowski *et al.*, 1995; Piotrowska *et al.*, to appear). Texts are filled mainly with words and word combinations related to the patient's obsessional topic. The variety of lexical units employed here is restricted and repetitions are many. $\beta = 1.66$ is reported in (Piotrowski *et al.* 1995; Piotrowska *et al.*, to appear).
- $\beta = 1.6$ in very young children (Brillouin, 1960; Piotrowski *et al.*, 1995). Older children conform to the typical $\beta \approx 2$ (Zipf, 1942).
- Exponents larger than $\beta \approx 2$ can be obtained as a result of deficient sampling from a text with the typical $\beta \approx 2$ (Piotrowski *et al.*, 1995; Piotrowska *et al.*, to appear).

Therefore, the exponents that are of interest here seem to be constrained to a very narrow domain, i.e. $\beta \in [1.66,2.42]$ (Ferrer i Cancho, 2005b). Whether Zipf's law can distinguish between acute and chronic schizophrenic patients is a matter of current research. The main message concerning schizophrenia here is that the disease shows exponents on both sides of the interval of variation in humans and that the value of the exponent may be related to the stage of the disease. Significant variations of β have also been found in multi-author samples (Piotrowski *et al*., 1994; Ferrer i Cancho, 2005d, Ferrer i Cancho & Solé, 2001; Montemurro, 2001; Montemurro & Zanette, 2002), particular word classes (Balasubrahmanyan & Naranan, 1996) and both (Ferrer i Cancho, 2005a).

The focus of the present paper is communicative aspects of single individuals. Significant deviations in multi-author texts will not be considered. The aim of the present paper is to show the connection between the exponents and various types of quantitative measures suggesting that the variation of the exponent may be due to tuning the vagueness and the cost of word use. Most of the measures of vagueness and the measure of cost of word use that are employed here are defined using Shannon's information theory (Ash, 1990). Support for the hypothesis of the strong association between Zipf's law and communication, comes from recent models where Zipf's law and/or the value of the exponent can be explained as the outcome of minimizing or constraining various standard information theory measures (Ferrer i Cancho, 2005a, 2005d; Ferrer i Cancho & Solé, 2003).

THE MODEL

We assume a general communication system mapping words to stimuli. We have a set of n words $S = \{s_1, ..., s_i, ..., s_n\}$ and a set of m stimuli $R = \{r_1, ..., r_j, ..., r_m\}$. We assume that words connect to stimuli to build their meaning. Word-stimuli associations are defined by a binary matrix $A = \{a_{ij}\}$ where $a_{ij} = 1$ if s_i and r_j are linked and $a_{ij} = 0$ otherwise. Let us consider in greater detail what is meant here by "stimuli". Various experiments have shown that words are associated with the activation of different brain areas (Pulvermüller, 2003). Generally speaking, nouns tend to activate visual areas. Verbs tend to activate motor areas if the corresponding action can be performed by the individual and visual areas otherwise. The activated areas are associated to different types of stimuli experienced with the word. Let us take one of the definitions of the Webster's Revised Unabridged Dictionary (1913)[2] for the word *write*: "to inscribe on any material by a suitable instrument". In our view, the verb *write* is associated to the motor stimuli of the action of writing and the visual (tactile, olfactory,...) stimuli of the instruments used for writing. The construction of a complex meaning would

[2] www.dict.org.

involve a structure combining diverse stimuli. From that point of view, a word in *S* does not refer to stimuli in *R*, but it is merely associated to them. We do not claim that words in *S* refer to stimuli in *R* via *A* although they may. We do not use the term reference because it is stronger than association. In our example, *write* can only refer to the motor stimuli of the action of writing. *write* cannot refer to the instrument used for writing, although it is associated with it. The action and the instrument are both stimuli involved in the construction of the complex meaning of the verb *write*. Defining "word meaning" is an open problem in various fields ranging from cognitive science to philosophy. In our view, complex meaning would emerge from the interaction between different stimuli. Referential associations may be a subset or a higher order structure of the associations defined by *A*. It makes sense to assume that the more stimuli a word is associated with, the higher the probability of using that word.

It is important to note that when we say that a word has no meaning we usually mean that it has no referential power. Nonetheless, if a word lacks referential power it does not imply that it has no associations with stimuli. Our framework is not inconsistent with the existence of words with no apparent meaning, such as prepositions, conjunctions or articles. Real words with no apparent meaning are the words with the highest frequencies. The five most frequent word in the British National Corpus[3], a large collection of text samples, are *the*, *of*, *and*, *to* and *a*. The framework here predicts that the most frequent words would have the largest number of connections with stimuli in *R*. Since those connections are merely associative (and *not* always referential) there is no inconsistency here. Furthermore, that high number of associations may underlie those words' lack of referential power or "meaning". The uncertainty associated with the interpretation of highly connected words is so large (Ferrer i Cancho, 2005c,e) that reference cannot be effectively attributed. Words with no meaning may have two different origins: words that have no links, and words having too many links. It makes sense to suppose that words with no meaning may have an excess of connections rather than a lack thereof, although those connections could be very weak given the high frequency of the words involved (Ferrer i Cancho & Reina, 2002).

A first approach to the semantic vagueness of the set of words could be the average number of links per word, that is, $\langle k \rangle$. The number of links of a vertex (e.g. a word) is called "degree" in standard graph theory (Bollobás, 1998), so $\langle k \rangle$ is the mean signal degree. The idea behind the relationship between $\langle k \rangle$ and vagueness is very simple: the more links a word has, the higher the number of possible interpretations in the context where it appears. The higher the value of $\langle k \rangle$, the lower the precision of the communication system. Hereafter we assume that 'precision' and 'vagueness' have opposite meaning. $H(R|S)$, that is, the average uncertainty (or entropy) associated with the interpretation of every stimulus once the corresponding word is known, is a more precise measure, from the information theory point of view. That measure is defined as

$$H(R \mid S) = \sum_{i=1}^{n} p(s_i) H(R \mid s_i), \tag{2}$$

where $H(R|s_i)$ is the uncertainty (or entropy) associated with the interpretation of s_i , and $p(s_i)$ is the probability of using s_i. $H(R|S)$ is the average uncertainty associated with the interpretation of the words in *S*. The higher the value of $H(R|S)$, the lower the precision of the communication system. Since $H(R|S)$ is mathematically a hard function to manipulate, a simpler version has been considered (Ferrer i Cancho, 2005a):

[3] www.natcorp.ox.ac.uk

$$G(R \mid S) = \frac{1}{n}\sum_{i=1}^{n} H(R \mid s_i)\,. \tag{3}$$

$G(R|S)$ is the amount of uncertainty per word associated with the interpretation of the words in S. $G(R|S)$ and $H(R|S)$ have similar properties. The upper and lower bounds are the same, i.e. $0 \leq G(R \mid S), H(R \mid S) \leq \log m$. $G(R|S)$ has the virtue of allowing Zipf's law (Eq. 1) to be derived using the maximum entropy principle (Ferrer i Cancho, 2005a).

A possible approach to the cost of word use is $H(S)$, the entropy of the set of words (Ferrer i Cancho, 2005a,d; Ferrer i Cancho & Solé, 2003). This is defined as

$$H(S) = -\sum_{i=1}^{n} p(s_i) \log p(s_i)\,. \tag{4}$$

Support for $H(S)$ as a measure of the cost of word use comes from two different sources. Firstly, it is known in psycholinguistics that the availability of a word in various linguistic tasks is correlated with the frequency of that word. The availability of a word obeys the so-called *word frequency effect*, i.e. the more frequent the word, the higher its availability (Akmajian *et al.*, 1995; Carroll, 1994). The best availability is achieved when a word has probability one, which means that the rest of the words have probability 0. In that case, $H(S) = 0$. The worst case is when all words are equally likely, that is when $p(s_i) = 1/n$ for each word. In that case, $H(S) = \log n$. That is, $H(S)$ is a good measure of cost of word use. Second, the use of $H(S)$ as a measure of cost is justified by models leading to Zipf's law when the information transfer is maximized while $H(S)$ is minimized (Ferrer i Cancho, 2005d; Ferrer i Cancho & Solé, 2003). Those models explain Zipf's law as the outcome of maximizing the communicative efficiency, but saving as much cost as possible. Interestingly, if those models replace $H(S)$ with the effective vocabulary size (i.e. the proportion of words with at least one link) as a measure of cost of word use, Zipf's law is not reproduced. Vocabulary size is an important ingredient for the cost of a communication system (Köhler, 1986, 1987) but it does not seem to be essential for Zipf's law.

We may assume that the $p(s_i)$, the probability of occurrence of word s_i, is proportional to k_i, the number of connections of s_i, that is

$$p(s_i) = \frac{k_i}{M}\,, \tag{5}$$

where

$$k_i = \sum_{j=1}^{m} a_{ij} \tag{6}$$

and

$$M = \sum_{i=1}^{n} k_i \tag{7}$$

(as in Ferrer i Cancho, 2005a,b,d). Eq. 5 contains the basic assumption that words are used according to the number of semantic associations they have. Eq. 5 states that a word is used with a probability proportional to the number of stimuli it is associated with. Eq. 5 is chosen

for simplicity and its predictive power: it can explain the interval of variation of β in human language (Ferrer i Cancho, 2005b).

We may also assume that $P(k)$, the proportion of words with k links obeys

$$P(k) \sim k^{-\beta}. \tag{8}$$

Zipf's law (Eq. 1) is recovered from Eqs. 5 and 8 (Ferrer i Cancho, 2005a,b). We assume a fixed $P(k)$ or $P(f)$, given the surprising tendency of human language to arrange according to Zipf's law even in atypical cases. Although there is variation in β for (human) words, the basic trend described by Eq. 1 has essentially no exceptions, as far as we know. From Eq. 5 and

$$p(s_i) = \sum_{j=1}^{m} p(s_i, r_j) \tag{9}$$

it follows that the probability that s_i and r_j are associated by the communication system is

$$p(s_i, r_j) = \frac{a_{ij}}{M}. \tag{10}$$

We may write Eq. 5 as

$$p(s_i \mid k_i = k) = \frac{k}{M} = \frac{k}{n < k >}, \tag{11}$$

where <...> is the expectation operator over $P = \{P(1),...,P(k),...P(m)\}$ and $P(k)$ is the proportion of words having k connections.

Assuming Eq. 8, the uncertainty (or entropy) associated to the interpretation of s_i becomes $H(R|s_i) = \log k$ if $k_i = k$ (Ferrer i Cancho, 2005a). Thus, $H(R|S)$ in Eq. 2 becomes (Ferrer i Cancho, 2005b)

$$H(R \mid S) = \frac{< k \log k >}{< k >} \tag{12}$$

and $G(R|S)$ becomes (Ferrer i Cancho, 2005a)

$$G(R \mid S) = < \log k >. \tag{13}$$

Figs. 1-3 show that $<k>$, $G(R|S)$ and $H(R|S)$ are decreasing functions of β for different values of m. The three functions grow with m for a given value of β.

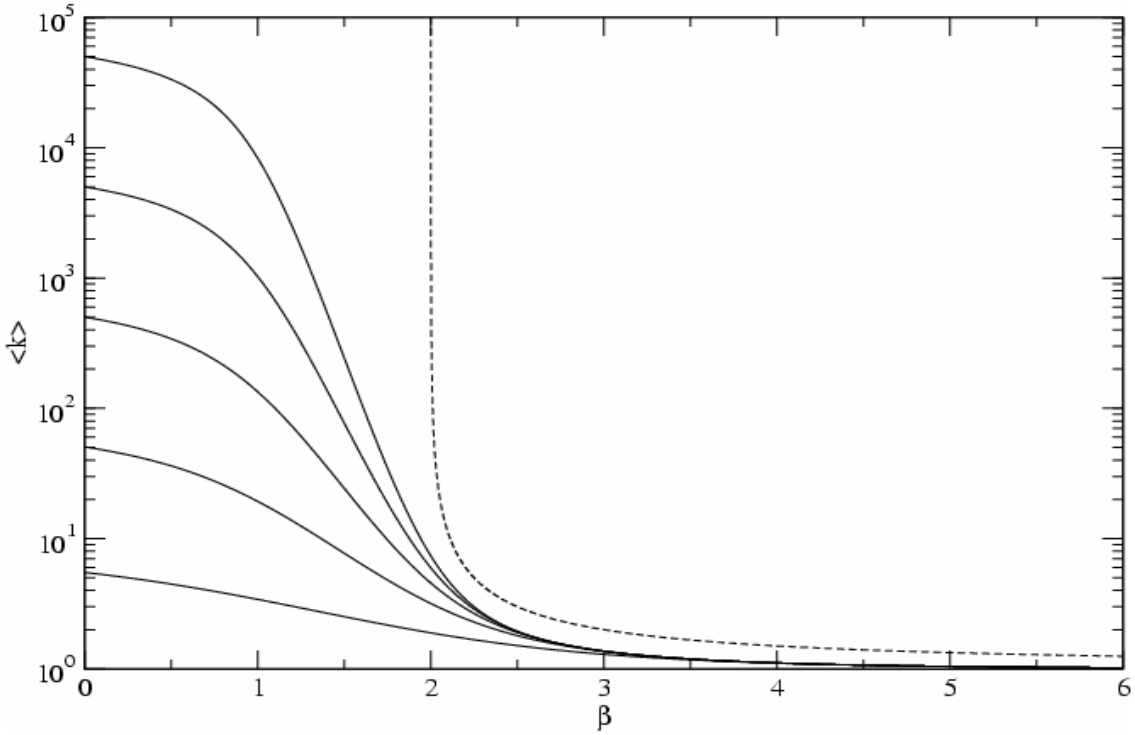


Fig 1. <*k*>, the mean word degree, versus β, the exponent of the frequency spectrum of Zipf's law. Series from the bottom to the top are for $m = 10$, $m = 10^2$, $m = 10^3$, $m = 10^4$ and $m = 10^5$ (solid lines). The approximated expected curve for $m \to \infty$ is also shown (dashed line).

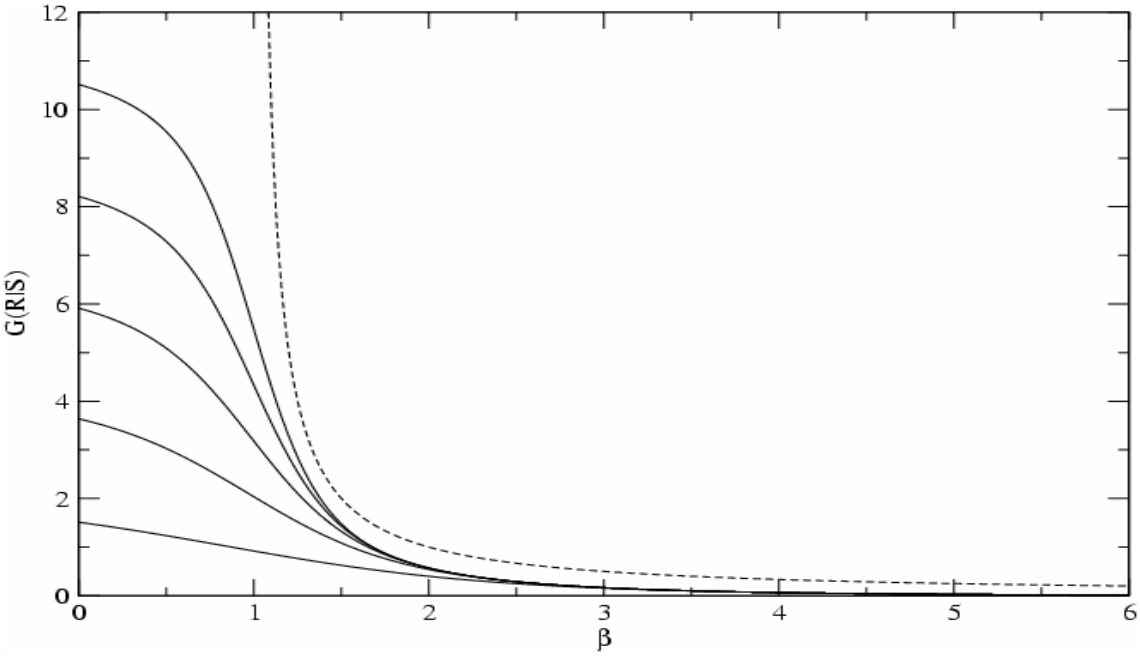


Fig 2. $G(R|S)$, the uncertainty per word associated with the interpretation of every word, versus β, the exponent of the frequency spectrum of Zipf's law. Series from the bottom to the top are for $m = 10$, $m = 10^2$, $m = 10^3$, $m = 10^4$ and $m = 10^5$ (solid lines). The approximated expected curve for $m \to \infty$ is also shown (dashed line). Natural logarithms were used.

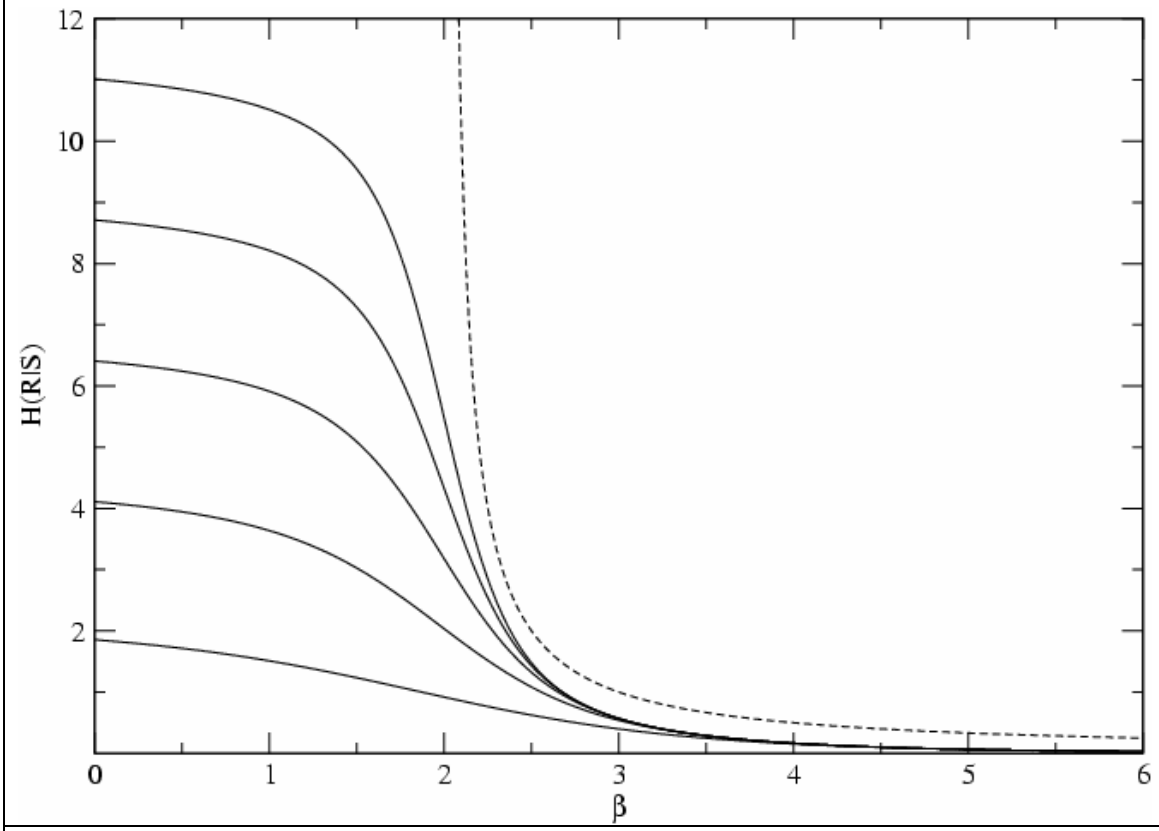


Fig 3. $H(R \mid S)$, the average uncertainty associated with the interpretation of every word, versus β, the exponent of the frequency spectrum of Zipf's law. Series from the bottom to the top are for $m = 10$, $m = 10^2$, $m = 10^3$, $m = 10^4$ and $m = 10^5$ (solid lines). The approximated expected curve for $m \to \infty$ is also shown (dashed line). Natural logarithms were used.

Here we will define vagueness as the opposite of precision. $<k>$, $G(R|S)$ and $H(R|S)$ are inverse measures of precision and direct measures of vagueness.

As for cost of word use, substituting Eq. 11 into Eq. 4 we get,

$$H(S) = \log(n < k >) - \frac{< k \log k >}{< k >}, \tag{14}$$

where $M = n<k>$ is the total amount of links. Knowing Eq. 12, Eq. 14 can be written as

$$H(S) = \log(n < k >) - H(R \mid S). \tag{15}$$

Fig. 4 shows $H(S)$ for $n = 10^3$ and different values of m. $H(S)$ decreases as m grows for a fixed value of β whereas the vagueness measures behave inversely. $H(S)$ has a minimum at $\beta = \beta^*$, a critical value of β, such that $1 < \beta^* < 2$ for the values of m that we used here. Notice that although the exact value of $H(S)$ depends on n, β^* depends only on m (recall Eq. 15). Fig. 5 shows β^* versus m.

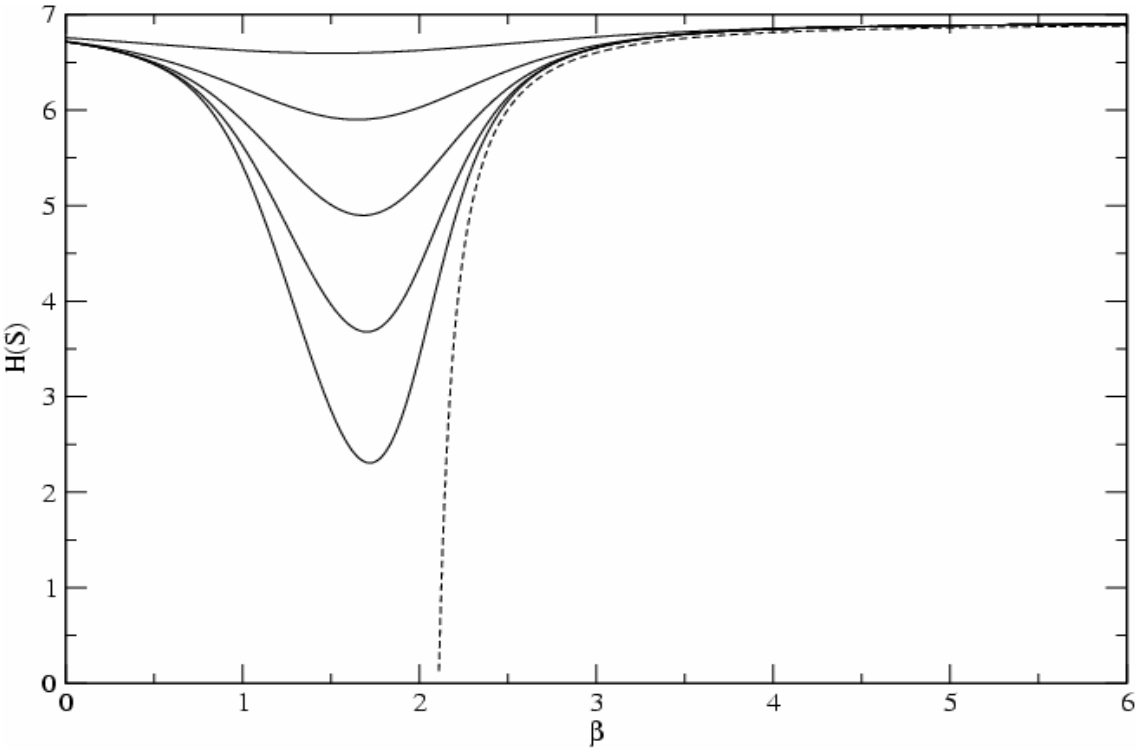


Fig. 4. $H(S)$ versus β where $H(S)$ is the entropy of the set of words S and β is the exponent of the frequency spectrum of Zipf's law. Series from top to the bottom are for $m = 10$, $m = 10^2$, $m = 10^3$, $m = 10^4$ and $m = 10^5$ (solid lines). $n = 10^3$ is used in all cases, although the point where the minimum $H(S)$ is reached is independent of n. The approximated expected curve for $m \to \infty$ is also shown (dashed line). Natural logarithms were used.

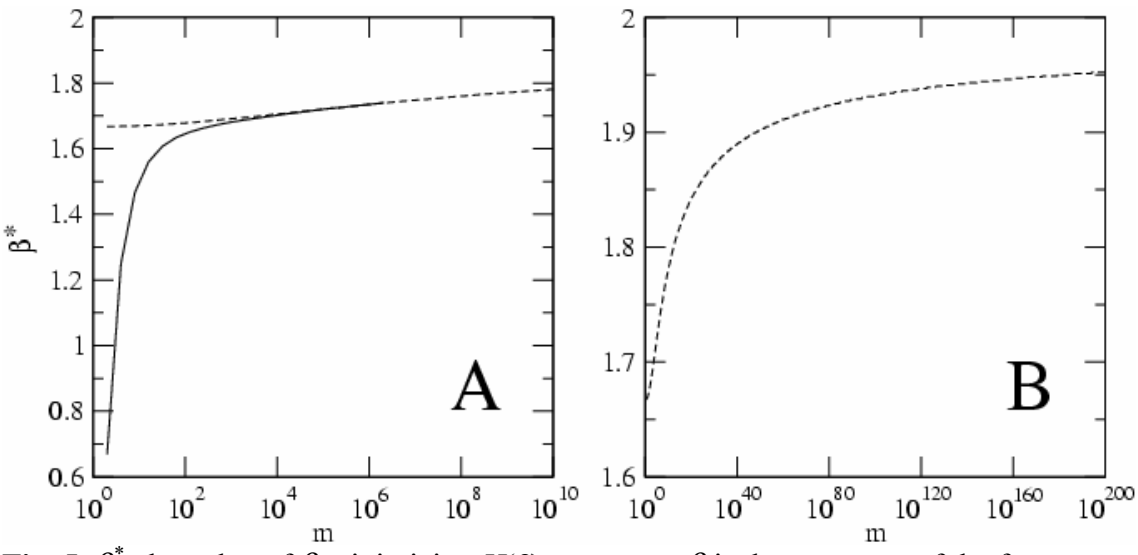


Fig. 5. β^*, the value of β minimizing $H(S)$, versus m. β is the exponent of the frequency spectrum representation of Zipf's law, $H(S)$ is the entropy of the set of words S and m is the number of stimuli. A. β^* versus m calculated without integrals using Eq. 15 (solid line) and with integrals using Table 1 (dashed line) B. β^* versus m calculated with integrals till very large values of m.

RESULTS

We can obtain formulae for the measures of vagueness and cost using approximation with integrals (see Appendix A). Results are summarized in Table 1. The measures of vagueness are functions of β and m whereas $H(S)$ is a function of β, m and n. When $m \to \infty$, we can obtain simple mathematical expressions in particular domains of β (Table 2).

Table 1
Summary of the relationship between Zipf's law and communication measures

Function	Information theory	Approximation	
$< k >$	-	$\beta \neq 1$ and $\beta \neq 2$	$\frac{(1-\beta)(m^{2-\beta}-1)}{(2-\beta)(m^{1-\beta}-1)}$
		$\beta = 1$	$\frac{m-1}{\log m}$.
		$\beta = 2$	$\frac{\log m}{1-\frac{1}{m}}$
$< \log k >$	$G(R\|S)$	$\beta \neq 1$	$\frac{1}{m^{1-\beta}-1}\left[m^{1-\beta}\left(\log m - \frac{1}{1-\beta}\right)+\frac{1}{1-\beta}\right]$
		$\beta = 1$	$\frac{\log m}{2}$
$\frac{< k \log k >}{< k >}$	$H(R\|S)$	$\beta \neq 1$ and $\beta \neq 2$	$\frac{1}{m^{2-\beta}-1}\left[m^{2-\beta}\left(\log m - \frac{1}{2-\beta}\right)+\frac{1}{2-\beta}\right]$
		$\beta = 1$	$\frac{m(\log m - 1)+1}{m-1}$
		$\beta = 2$	$\frac{\log m}{2}$

$<k>$ is the mean word degree, $G(R|S)$ is the uncertainty per word associated with the interpretation of every word and $H(R|S)$ is the average uncertainty associated with the interpretation of every word. $<\ldots>$ is the expectation operator over k. m is the number of stimuli. β is the exponent of the power spectrum of Zipf's law.

$<k>$ when $\beta > 2$ and $m \to \infty$ is shown as dashed line in Fig. 1. $G(R|S)$ when $\beta > 1$ and $m \to \infty$ is shown as dashed line in Fig. 2. $H(R|S)$ when $\beta > 2$ and $m \to \infty$ is shown as a dashed line in Fig. 3. $H(S)$ when $\beta > 2$ and $m \to \infty$ is shown as a dashed line in Fig. 4.

When $\beta > 1$ and $m \to \infty$, we have $G(R|S) = 1/(\beta - 1)$. Zipf's law can be alternatively defined as $P(i) \sim i^{-\alpha}$, where $P(i)$ is the frequency of the i-th most frequent word in a piece of text and $\alpha = 1/(\beta - 1)$ (Chitashvili & Baayen, 1993). Notice that $G(R|S) = 1/(\beta - 1)$ tells us that $\alpha = G(R|S)$. The value of β where $\beta = \alpha$ (and thus $\beta = G(R|S)$) can be calculated solving $\beta = 1/(\beta - 1)$, which has two solutions

$$\beta_1 = \frac{1+\sqrt{5}}{2} \approx 1.61. \tag{16}$$

and

$$\beta_2 = \frac{1-\sqrt{5}}{2} \approx -0.61. \tag{17}$$

The first solution (Eq. 16) is the only valid one here since $\beta > 1$ is assumed in $\alpha = 1/(\beta - 1)$ (see also Ferrer i Cancho & Solé, 2001). There are two points of interest with regard to $\beta = \alpha \approx 1.61$. Firstly, ≈ 1.61 is close to the exponent found in certain children and schizophrenics. Only in those cases, β is also a direct measure for $G(R|S)$. Secondly, T. Hernández noticed that $(1+\sqrt{5})/2$ (Eq. 16) is the golden ratio, the value to which the fraction of two consecutive numbers of the Fibonacci series converges (Dunlap, 1997). The golden ratio has been the topic of many speculations about its role in nature and our sense of aesthetics (Ghyka, 1927). Future work should be devoted to investigating the origins of the appearance that striking coincidence.

Table 2
Summary of the relationship between the exponents of Zipf's law and various communication measures when $m \to \infty$.

Communication measures	$m \to \infty$			
	β		α	
	Approximation	*Condition*	*Approximation*	*Condition*
$\langle k \rangle$	$\frac{\beta-1}{\beta-2}$	$\beta > 2$	$\frac{1}{1-\alpha}$	$\alpha < 1$
$G(R\|S)$	$\frac{1}{\beta-1}$	$\beta > 1$	α	$\alpha > 0$
$H(R\|S)$	$\frac{1}{\beta-2}$	$\beta > 2$	$\frac{\alpha}{1-\alpha}$	$\alpha < 1$
$H(S)$	$\log\frac{n(1-\beta)}{2-\beta} - \frac{1}{\beta-2}$	$\beta > 2$	$\log\frac{n}{1-\alpha} - \frac{\alpha}{1-\alpha}$	$\alpha < 1$

$\langle k \rangle$ is the mean word degree, $G(R|S)$ is the uncertainty per word associated with the interpretation of every word and $H(R|S)$ is the average uncertainty associated with the interpretation of every word. β and α are, respectively, the exponents of the power spectrum and the frequency versus rank representation of Zipf's law. Recall $\alpha = 1/(\beta - 1)$ (Chitashvili & Baayen, 1993).

We can combine Zipf's law (Eq. 1) and the results in Table 2 in order to emphasize the relationship between communication and Zipf's law. When $\beta > 2$,

$$P(f) \sim f^{-\frac{1-2\langle k \rangle}{1-\langle k \rangle}} \tag{18}$$

and

$$P(f) \sim f^{-\frac{1}{H(R|S)}-2} \tag{19}$$

When $\beta > 1$,

$$P(f) \sim f^{-\frac{1}{G(R|S)}-1} \tag{20}$$

A rewritten version of Zipf's law in terms of quantitative communication measures is summarized in Table 3 for both the frequency spectrum and the frequency versus rank representation. We have seen that $G(R|S)$ is the exponent of Zipf's law in the frequency versus rank representation. β also tells us about $G(R|S)$ but $\beta = G(R|S)$ holds only when $\beta \approx 1.61$ (recall Eq. 16).

Table 3
Explicit relationship between Zipf's law and various communication aspects

Representation	Variable		
	$\langle k \rangle$	$G(R\|S)$	$H(R\|S)$
$P(f)$	$\sim f^{-\frac{1-2\langle k \rangle}{1-\langle k \rangle}}$	$\sim f^{-\frac{1}{G(R\|S)}-1}$	$\sim f^{-\frac{1}{H(R\|S)}-2}$
$P(i)$	$\sim i^{-1+\frac{1}{\langle k \rangle}}$	$\sim i^{-G(R\|S)}$	$\sim i^{-\frac{H(R\|S)}{1+H(R\|S)}}$

$P(i)$ is the frequency of the i-th most frequent word in a sample (e.g. a text). $P(f)$ is the proportion of words in a sample with frequency f. $\langle k \rangle$ is the mean word degree, $G(R|S)$ is the uncertainty per word associated to the interpretation of every word and $H(R|S)$ is the average uncertainty associated to the interpretation of every word. β and α are, respectively, the exponents of the power spectrum and the frequency versus rank representation of Zipf's law.

DISCUSSION

We have seen that β is an indicator of the degree of semantic precision of a communication system. Given the same values of m, the higher the value of β, the higher the precision. $\langle k \rangle$, $G(R|S)$ and $H(R|S)$ are decreasing functions of β (Figs. 1-3). In contrast, $H(S)$ has a global minimum between 1.5 and 2 for sufficiently large m (Fig. 4). $G(R|S)$ is a measure of vagueness that does not diverge for $1 < \beta < 2$. That is not surprising since $G(R|S)$ does not weight $H(R|s_i)$ by the probability of s_i. From the information theory point of view, $H(R|S)$ is the reference measure of semantic vagueness. We will leave $\langle k \rangle$ and $G(R|S)$ as alternative simpler measures which correlate with $H(R|S)$ for certain values of β and m. For instance, notice that $G(R|S)$ has no counterpart when $1 < \beta < 2$ and $m \to \infty$ (since $\langle k \rangle$ and $H(R|S)$ are not defined in that case).

Both $G(R|S)$ and $H(R|S)$ measure the semantic precision from an information theory approach. $G(R|S)$ and $H(R|S)$ can be defined as functions of a single parameter, β, for m large (recall Table 2). $\beta > 2$ is required for $H(R|S)$ while only $\beta > 1$ is required for $H(R|S)$. Thus, $H(R|S)$ cannot deal with the exponent of some children and schizophrenics having $\beta < 2$ if m is actually large. Briefly, $G(R|S)$ covers all the range of variations of β found in human language while $H(R|S)$ does not (provided m is large, of course).

There are strong constraints on the communication systems following Zipf's law (with our assumptions) that may exist when $m \to \infty$. First, $P(k)$ is a probability function only when $\beta > 1$, so systems with $\beta \leq 1$ are impossible at the thermodynamic limit. Second, it is easy to show that $H(R|S)$ and $<k>$ diverge when $m \to \infty$ and $1 < \beta < 2$. In other words, systems with finite vagueness are impossible if $1 < \beta < 2$. From the information theory point of view, if $H(R|S)$ is infinite, then communication is not possible due the infinite uncertainty associated with decoding a single word on average. Besides, we have seen that $H(R|S)$ takes finite values, regardless of how large m is, when $\beta > 2$. In a communication system with $\beta < 2$, m must be finite and not too large, otherwise vagueness is infinite, which contradicts the notion that our system is communicating. We can apply this to real problems. There are many cases where speakers are clearly communicating with $\beta < 2$: military combat texts with $\beta = 1.7$ and children with $\beta = 1.6$ (Piotrowski *et al.*, 1995). Some caution must be taken with schizophrenics with $\beta < 2$, where the assumption of communication may fail. There are reasons for thinking that the assumption is actually satisfied. If we assume that schizophrenics with $\beta < 2$ are communicating, then it follows that m should be small (it should be actually the smallest, since those schizophrenics take the smallest β among real speakers and the vagueness measures grow fast as β decreases). A dramatically reduced set of stimuli that can be perceived and thus can be conveyed using words (i.e. a dramatically low value of m) might explain the obsessive pattern found in that kind of schizophrenics. The same could be happening to children, whose perception of the world is under construction. We may synthesize the essence of the previous argument in a rule.

RULE 1. Suppose a communication system with exponent $\beta < 2$. Then m must be finite and not too large (otherwise the ambiguity would be too large). The chance of a large value of m decreases with β.

The predicted decrease in m in schizophrenics with $\beta < 2$ suggest that the apparent normality of category content and structure (Elvevåg *et al.*, 2005; Elvevåg *et al.*, 2002) may need to be revised.

The various results presented in this article allow us to face the following problem. What can be said about the communicative accuracy or the cost of a communication system when only the signal (e.g. word) frequency distribution is known? Candidates for this kind of analysis are atypical human speakers and the utterances of non-human species (McCowan *et al.*, 1999; McCowan *et al.*, 2002). If Zipf's law is found, the slope in log-log scale of the frequency spectrum is key to finding the answer. We will propose a series of lemmata that are helpful in determining which of two systems is more precise or economical (see Appendix B for outlines of proofs):

LEMMA 1. Suppose we have two communication systems A and B, with exponents β_A (or α_A) and β_B (or α_B) with $\beta_A, \beta_B > 1$, and the number of stimuli is m_A and m_B, respectively. If $\beta_A < \beta_B$ (or $\alpha_A > \alpha_B$) and $m_A \geq m_B$ then B is a strictly more precise communication system than A.

LEMMA 2. If we have two communication systems A and B, with exponents β_A (or α_A) and β_B (or α_B) with $\beta_A, \beta_B > 1$, their number of stimuli is m_A and m_B, and their lexicon size is n_A and n_B, respectively. We assume m_A, m_B, n_A and n_B are finite. If $\beta^* \leq \beta_A < \beta_B$ (or $\alpha^* > \alpha_A > \alpha_B$ with $\alpha^* = 1/(\beta^*-1)$) and $m_B \geq m_A$ and $n_A \geq n_B$ then it follows that A is a more economical communication system than B.

An apparently serious drawback to applying Lemma 2 is that we do not know, in general, if $\beta^* \leq \beta_A$. Interestingly, there are reasons for thinking that communication systems with $\beta^* \leq \beta_A$ are unlikely. The cost of word use and the vagueness increase at the same time as β decreases when $\beta < \beta^*$. It is hard to imagine how a communication system would tolerate decreasing the quality of communication and simultaneously expending more energy to communicate. Vagueness and cost of word use are in conflict for β above β^* (the former decreases with β and the latter grows with β), so it is reasonable to suppose that particular communication systems choose to favour one over the other. But it seems unlikely that a communication system would evolve against *both* factors below β^*. Since $\beta^* \leq \beta_A$ is apparently unlikely, we may propose a modified version of Lemma 2 that is likely to be true in natural communication systems:

> LEMMA 3. If we have two communication systems A and B, with exponents β_A (or α_A) and β_B (or α_B) with $\beta_A,\beta_B > 1$, their number of stimuli is m_A and m_B, and their lexicon size is n_A and n_B, respectively. We assume m_A, m_B, n_A and n_B are finite. If $\beta_A < \beta_B$ (or $\alpha_A > \alpha_B$) and $m_B \geq m_A$ and $n_A \geq n_B$ then it is likely that A is a more economical communication system than B.

We have seen in Eq. 15 that the cost of word use is a function of vagueness when the latter is measured using $H(R|S)$. When $\beta < \beta^*$, cost (of signal use) and vagueness decrease with β. In contrast, cost grows with β while vagueness decreases with β when $\beta > \beta^*$. Vagueness and cost are in conflict when $\beta > \beta^*$, which, as argued above, is likely to hold in natural communication systems.

Lemma 3 can be safely used if β_A is sufficiently large. Fig. 5 suggests that it is rather unlikely that a value of β very near to 2 minimizes the cost of word use. When m is about 10^{80}, a rough estimate of the number of atoms in the universe (Gribbin, 1986), we obtain $\beta^* \approx 1.923$ (recall here and later that our calculations are based on approximations using integrals). When m is about the number of neurons in the brain, about 10^{11} (Damasio, 1999), we get $\beta^* \approx 1.789$. If the number of neurons in the brain is taken, then β cannot minimize $H(S)$ if $\beta > 1.789$. In practice, if $\beta > 1.923$, that means that $\beta > \beta^*$. In a less compelling fashion, if $\beta > 1.789$, that means that $\beta > \beta^*$ is likely to be true. So, we do not need to worry about β^* if β is sufficiently large. In sum, imagine a communication system with exponent β. If $\beta > 1.923$ then $\beta > \beta^*$ is very likely and if $\beta > 1.789$ then $\beta > \beta^*$ is likely.

Now, let us try to apply the lemmata above to real problems. First, we may ask whether the correlation between β and semantic precision is consistent with the variation of β found in the real cases summarized in the introduction section, with regard to an ideal normal language with $\beta_A = 2$. Maybe nouns are the only unquestionable case of communication that is *a priori* more precise than normal language. There seems to be a certain consensus in philosophy and linguistics about the semantic rigidity of many nouns (Kripke, 1980; Mcbeth, 1995; Devitt & Sterelny, 1999). Let us define A as ideal normal language, and B as nouns. We have $\beta_A = 2$ and $\beta_B \in [2.15,2,32]$ (Balasubrahmanyan & Naranan, 1996). Thus, $\beta_A < \beta_B$ for the largest values of β_B. We will focus on that case. Since nouns are associated with a (probably strict) subset of all possible stimuli, that is, $m_A \geq m_B$, it follows from Lemma 1 that nouns are more precise that the entire set of words on average.

With the theory presented here and the support of the previous test, we can move to increasingly complicated cases. Imagine we take schizophrenics with $\beta < 2$ as A and ideal normal language as B. Lemma 1 can not be applied because we may have $m_A < m_B$ as discussed above. A similar problem is poised by children. Imagine we take ideal normal lan-

guage as A and schizophrenics with $\beta < 2$ as B. We do not know if $m_A \geq m_B$ holds so we cannot safely use Lemma 1. We must be conservative because the decrease in m that is predicted for schizophrenics with $\beta < 2$ could also happen in schizophrenics with $\beta > 2$.

Second, we may try to shed light on the cost of word use in real cases against ideal normal speakers. Let us define A as an ideal normal language and B as nouns. Since nouns are a strict subset of all words, we have $n_A > n_B$. We have seen above that $\beta_A < \beta_B$ and $m_A \geq m_B$. The latter means we cannot use Lemma 2. As for schizophrenic patients, we assume that lexicon size, n, is the same as in normal speakers, as it may be inferred that the lexicon is intact in schizophrenia (Goldberg *et al.*, 2000; Elvevåg *et al.*; 2001; Allen *et al.* 1993). Let us take A as schizophrenics with $\beta < 2$ and B as ideal normal language. We have $\beta_A < \beta_B$ and $m_B \geq m_A$, as we have deduced above. From Lemma 3, we discover that schizophrenics with $\beta < 2$ are more economical speakers. Let us take A as ideal normal language and B as schizophrenics with $\beta > 2$. We do not know if $m_B \geq m_A$. Again, recall that schizophrenics with $\beta > 2$ may have an anomalously low m_B, as schizophrenics with $\beta < 2$. If we take A as children with $\beta = 1.6$ and B as ideal normal speech, Lemma 3 cannot be applied because $n_A \geq n_B$ is not warranted. n_A, vocabulary size, could be one or more orders of magnitude smaller than that of normal adults (Johnson *et al.*, 1999). Since we know that older children eventually converge to $\beta = 2$ (Zipf, 1942), children with $\beta = 1.6$ must be sufficiently young. Knowing that vocabulary grows with age (Johnson *et al.*, 1999), children with $\beta = 1.6$ should have a significantly smaller vocabulary than adults. Although we do not know the exact value of their set of stimuli, and the size of that set depends on the level of their brain development, Lemma 3 cannot be safely used. Nonetheless, the expected significantly small vocabularies may reduce the value of $H(S)$ below that of normal adults.

Let us summarize all the inferences we have made till now:

- Nouns are more precise than mean words in ideal normal adults.
- Schizophrenics with $\beta < 2$ are likely to have a reduced value of m and a more economical communication system with regard to normal adult speakers.

We have assumed that $\beta = 2$ is the ideal exponent of normal adults. Are there reasons for thinking that ideal exponent should be very near 2? Imagine a communication system trying to transmit information about the largest set of stimuli possible. The latter would mean $m \rightarrow \infty$. In that case, what is the most economical communication system following Zipf's law? $m \rightarrow \infty$ imposes that $\beta > 2$ so that communication has finite vagueness. Since we have seen that the cost of communication grows with β when $\beta > 2$, communication should not go far above $\beta = 2$. Therefore, the ideal communication system minimizing the cost but avoiding infinite vagueness should have $\beta = 2+\varepsilon$, where ε is a small positive quantity (e.g. $\varepsilon = 0.01$ so $\beta = 2.01$). The effect of minimizing β when $\beta > 2$ admits a complementary view. We have considered the negative dimension of $H(R|S)$: the higher the value of $H(R|S)$, the higher the vagueness of words. We could make a positive complementary argument from the point of view of semantic versatility: the higher the value of $H(R|S)$, the higher the semantic versatility of words. It is important to avoid $\beta \leq 2$ to elude infinite vagueness, but it is important to remain near $\beta = 2$, since $H(R|S)$ decreases with β. The arguments above may shed light on the expected exponent of Zipf's law in ideal conditions.

In summary, this article has approached the relationship between the exponent of Zipf's law and various properties of communication systems. Given two communication systems, it is possible to infer which of them is the more economical or vague. It is clear that we need additional information, such as m or n, as well the exponent of Zipf's law, in order to know more about the features of a communication system. Interestingly, we have seen that the amount of extra information that is needed is reduced, and available in some cases. These

findings indicates that the little information provided by real communication systems can be squeezed to increase our knowledge about them.

ACKNOWLEDGMENTS

We thank Vito D.P. Servedio and Toni Hernández for helpful comments. We are grateful to Brita Elvevåg for advice about schizophrenia. This work was supported by the ECAgents project, funded by the Future and Emerging Technologies program (IST-FET) of the European Commission under the EU RD contract IST-1940. The information provided is the sole responsibility of the authors and does not reflect the Commission's opinion. The Commission is not responsible for any use that may be made of the data appearing in this publication.

APPENDIX A

Here, we shall obtain analytical approximations for $<k>$, $G(R|S)$ (Eq. 13), $H(R|S)$ (Eq. 12) and $H(S)$ (Eq. 15), assuming Eq. 5. We will approximate summations using integrals (Cormen *et al.*, 1990). When a summation can be expressed as

$$\sum_{k=k_{\min}}^{k_{\max}} f(k), \tag{21}$$

where $f(k)$ is a monotonically increasing function, we can approximate it by integrals (Cormen, 1990) holding

$$\int_{k_{\min}-1}^{k_{\max}} f(k)dk \leq \sum_{k=k_{\min}}^{k_{\max}} f(k) \leq \int_{k_{\min}}^{k_{\max}+1} f(k)dk . \tag{22}$$

When $f(k)$ is a monotonically decreasing function, then

$$\int_{k_{\min}}^{k_{\max}+1} f(k)dk \leq \sum_{k=k_{\min}}^{k_{\max}} f(k) \leq \int_{k_{\min}+1}^{k_{\max}} f(k)dk . \tag{23}$$

Here, we will use the approximation

$$\sum_{k=k_{\min}}^{k_{\max}} f(k) \approx \int_{k_{\min}}^{k_{\max}} f(k)dk , \tag{24}$$

which is used often in physics (e.g. Cohen & Havlin, 2002; Newman, 2005).

Before providing approximations for $<k>$, $G(R|S)$, $H(R|S)$ and $H(S)$, we need to introduce a function

$$F_x(\gamma, m) = \sum_{k=1}^{m} k^{-\gamma} \log^x k \approx \int_1^m k^{-\gamma} \log^x k dk , \tag{25}$$

which will be used recurrently later on. Interestingly, $F_0(\gamma,m) = H_m^{\gamma}$, where H_m^{γ} is the harmonic number of order γ.

When $\gamma \neq 1$, we have

$$F_0(\gamma, m) \approx \int_1^m k^{-\gamma} dk = \frac{m^{1-\gamma} - 1}{1-\gamma} \tag{26}$$

and

$$F_1(\gamma, m) \approx \int_1^m k^{-\gamma} \log k dk = \frac{1}{1-\gamma}\left[m^{1-\gamma}\left(\log m - \frac{1}{1-\gamma} \right) + \frac{1}{1-\gamma} \right]. \tag{27}$$

When $\gamma = 1$, we have

$$F_0(\gamma, m) \approx \log m \tag{28}$$

and

$$F_1(\gamma, m) \approx \frac{\log^2 m}{2}. \tag{29}$$

Table 4
Summary of definitions of different functions and their relationships

Function	Information theory	Definition
$F_x(\gamma, m)$	-	$\sum_{k=1}^{m} k^{-\gamma} \log^x k$
c	-	$1/F_0(\beta,m)$
$< k >$	-	$F_0(\beta\text{-}1,m)/F_0(\beta,m)$
$< \log k >$	$G(R\|S)$	$F_1(\beta,m)/F_0(\beta,m)$
$\frac{< k \log k >}{< k >}$	$H(R\|S)$	$F_1(\beta\text{-}1,m)/F_0(\beta\text{-}1,m)$

Table 4 summarizes the relationship between the auxiliary function $F_x(\gamma,m)$ and the functions of vagueness.

Thus, we may write,

$$< k > = c\sum_{k=1}^{m} k^{1-\beta} = cF_0(\beta - 1, m), \tag{30}$$

where c is the normalization constant of Eq. 8, defined as

$$c = \frac{1}{\sum_{k=1}^{m} k^{-\beta}} = \frac{1}{F_0(\beta, m)}. \tag{31}$$

Eqs. 30 and 31 together give

$$< k > = \frac{F_0(\beta - 1, m)}{F_0(\beta, m)}. \quad (32)$$

If $\beta \neq 1$, substituting Eq. 26 on Eq. 31 gives (Cohen & Havlin, 2002)

$$c \approx \frac{1-\beta}{m^{1-\beta} - 1}. \quad (33)$$

If $\beta = 1$, substituting Eq. 28 on Eq. 31 gives

$$c \approx \frac{1}{\log m}. \quad (34)$$

When $\beta \neq 1$ and $\beta \neq 2$, substituting $F_0(\beta - 1,m)$ with Eq. 26 with $\gamma = \beta - 1$ and $F_0(\beta,m)$ by Eq. 26 with $\gamma = \beta$ into Eq. 32 we obtain

$$< k > \approx \frac{(1-\beta)(m^{2-\beta} - 1)}{(2-\beta)(m^{1-\beta} - 1)}. \quad (35)$$

When $\beta = 1$, substituting $F_0(\beta - 1,m)$ with Eq. 26 (with $\gamma = \beta - 1 = 0$) and $F_0(\beta,m)$ by Eq. 28 (since $\gamma = \beta = 1$) into Eq. 32 we obtain

$$< k > \approx \frac{m-1}{\log m}. \quad (36)$$

When $\beta = 2$, substituting $F_0(\beta - 1,m)$ with Eq. 28 (since $\gamma = \beta - 1 = 1$) and $F_0(\beta,m)$ by Eq. 26 (with $\gamma = \beta = 2$) into Eq. 32 we obtain

$$< k > \approx \frac{\log m}{1 - \frac{1}{m}}. \quad (37)$$

When $\beta > 2$ and $m \to \infty$, Eq. 35 becomes

$$< k > \approx \frac{\beta - 1}{\beta - 2}. \quad (38)$$

The previous equation is shown as a dashed line in Fig. 1.

$G(R|S)$ can be written as

$$G(R \mid S) = < \log k > = c \sum_1^m k^{-\beta} \log k = \frac{F_1(\beta, m)}{F_0(\beta, m)}. \quad (39)$$

When $\beta \neq 1$, substituting $F_1(\beta,m)$ by Eq. 27 and $F_0(\beta,m)$ by Eq. 26 (both with $\gamma = \beta$) into Eq. 39 we get

$$G(R \mid S) \approx \frac{1}{m^{1-\beta}-1}\left[m^{1-\beta}\left(\log m - \frac{1}{1-\beta}\right) + \frac{1}{1-\beta}\right]. \tag{40}$$

When $\beta = 1$, substituting $F_1(\beta,m)$ by Eq. 29 and $F_0(\beta,m)$ by Eq. 28 (since $\gamma = \beta = 1$ in both cases) into Eq. 39 we get

$$G(R \mid S) \approx \frac{\log m}{2}. \tag{41}$$

When $\beta > 1$ and $m \to \infty$, Eq. 40 becomes

$$G(R \mid S) \approx \frac{1}{\beta - 1}. \tag{42}$$

(as in Ferrer i Cancho, 2005a). The previous equation is shown as a dashed line in Fig. 2.
As for $H(R|S)$, the numerator in Eq. 12 can be expressed as

$$< k \log k > = c\sum_{1}^{m} k^{1-\beta} \log k = cF_1(\beta - 1, m). \tag{43}$$

Substituting Eqs. 30 and 43 into Eq. 12 we obtain

$$H(R \mid S) = \frac{F_1(\beta - 1, m)}{F_0(\beta - 1, m)}. \tag{44}$$

If $\beta \neq 1$ and $\beta \neq 2$, substituting $F_1(\beta - 1,m)$ with Eq. 27 and $F_0(\beta - 1,m)$ with Eq. 26 (with $\gamma = \beta - 1$ in both cases) into Eq. 44 we get

$$H(R \mid S) \approx \frac{1}{m^{2-\beta}-1}\left[m^{2-\beta}\left(\log m - \frac{1}{2-\beta}\right) + \frac{1}{2-\beta}\right]. \tag{45}$$

If $\beta = 1$, substituting $F_1(\beta-1,m)$ with Eq. 27 and $F_0(\beta-1,m)$ with Eq. 26 (both with $\gamma = \beta - 1 = 0$) into Eq. 44 we get

$$H(R \mid S) \approx \frac{m(\log m - 1) + 1}{m - 1}. \tag{46}$$

If $\beta = 2$, substituting $F_1(\beta - 1,m)$ with Eq. 29 and $F_0(\beta - 1,m)$ with Eq. 28 (since $\gamma = \beta - 1 = 1$ in both cases) into Eq. 44 we get

$$H(R \mid S) \approx \frac{\log m}{2}. \tag{47}$$

Eq. 45 with $\beta > 2$ and $m \to \infty$ becomes

$$H(R \mid S) \approx \frac{1}{\beta - 2}. \tag{48}$$

The previous equation is shown as a dashed line in Fig. 3.

When $\beta > 2$ and $m \to \infty$, substituting Eqs. 38 and 48 into Eq. 15 we obtain

$$H(S) = \log\left(\frac{n(\beta - 1)}{\beta - 2}\right) - \frac{1}{\beta - 2}. \tag{49}$$

The previous equation is shown as a dashed line in Fig. 4. Since Eq. 49 is an approximation, and $<k>$ and $H(R|S)$ diverge for $\beta = 2$, it is convenient to keep $\beta >> 2$.

APPENDIX B

Here we give an outline of proof for Lemma 1 and 2.

LEMMA 1. Suppose we have two communication systems A and B, with exponents β_A (or α_A) and β_B (or α_B) with $\beta_A, \beta_B > 1$, and the number of stimuli is m_A and m_B, respectively. If $\beta_A < \beta_B$ (or $\alpha_A > \alpha_B$) and $m_A \geq m_B$ then B is a strictly more precise communication system than A.

Proof: The proof is based on $H(R|S)$, the reference measure for word vagueness. Assuming $\beta_A, \beta_B > 1$ we warrant that $P(k)$ is a probability distribution even when $m \to \infty$. In general, there are only four situations:

1) m_A and m_B are finite. It is easy to see from the approximate equations in Table 1 Appendix A (recall also Fig. 3) that $H(R|S)$ is a monotonically decreasing function of β (when $\beta > 0$) when m_A and m_B are finite. Given a particular β, the larger the value of m, the larger the value of the measure.
2) m_A is finite and m_B is not. That contradicts $m_A \geq m_B$.
3) m_A is infinite and m_B is not. That contradicts the notion that A is a communication system if $\beta_A \leq 2$. $\beta_A > 2$ must be satisfied and thus we can proceed as in 1).
4) m_A and m_B are infinite. That contradicts the notion that A and B are communication systems if $\beta_A \leq 2$ and/or $\beta_B \leq 2$. $\beta_A, \beta_B > 2$ must be satisfied and thus we can proceed as in 1).

LEMMA 2. If we have two communication systems A and B, with exponents β_A (or α_A) and β_B (or α_B) with $\beta_A, \beta_B > 0$, their number of stimuli is m_A and m_B, and their lexicon size is n_A and n_B, respectively. We assume m_A, m_B, n_A and n_B are finite. If $\beta^* \leq \beta_A < \beta_B$ (or $\alpha^* > \alpha_A > \alpha_B$ with $\alpha^* = 1/(\beta^* - 1)$) and $m_B \geq m_A$ and $n_A \geq n_B$ then it follows that A is a more economical communication system than B.

Proof: $\alpha^* = 1/(\beta^* - 1)$ comes from the equivalence between, α, the exponent of the frequency versus rank representation and β, the exponent of the frequency spectrum (Chitashvili & Baayen, 1993). It is easy to show from the approximate equations in Table 1 (recall Figs. 4-5) that $H(S)$, the measure of cost, is a monotonically increasing function of β when $\beta > \beta^*$ and

that given a particular β, the larger the value of m, the lower the cost, and that, the larger the value of n, the larger the cost.

REFERENCES

Allen, H. A., Liddle, P.F. & Frith, C. D. (1993). Negative features, retrieval processes and verbal fluency in schizophrenia. *British Journal of Psychiatry 163, 769-775*.

Akmajian, A., Harnish, R. M., Demers, R. A. & Farmer, A. K. (1995). *Linguistics. An introduction to language and communication*. Cambridge, MA: MIT Press.

Ash, R. B. (1990). *Information Theory*. New York: Dover Publications, Inc.

Balasubrahmanyan, V. K. & Naranan, N. (1996). Quantitative linguistics and complex systems studies. *Journal of Quantitative Linguistics 3, 177-228.*

Bollobás, B. (1998). Modern graph theory. New York: Springer.

Brillouin, L. (1960). *Science and theory of information* (Russian translation). Moscow: Gosudarstvennoe Izdatel´stvo Fiz.-Mat. Literatury.

Carroll, D. W. (1994). Psychology of language. Chapter 9, Conversational interaction. pp. 242-248. Pacific Grove, California: Brooks/Cole Publishing Company.

Chitashvili, R. J. & Baayen, R. H. (1993). Word frequency distributions. In: L. Hřebíček, G. Altmann (eds.), *Quantitative text analysis: 54-135.* Trier: Wissenschaftlicher Verlag Trier.

Cohen, R. & Havlin, S. (2002). Scale-free networks are ultrasmall. *Physical Review Letters 90, 057801.*

Cormen, T. H., Leisserson, C. E. & Rivest, R. L. (1990). *Introduction to algorithms*. Cambridge: MIT Press.

Damasio, A. (1999). *The Scientific American Book of the Brain.* New York: Scientific American. See http://hypertextbook.com/facts/2002/AniciaNdabahaliye2.shtml.

Devitt, M. & Sterelny, K. (1999). *Language and reality: an introduction to the philosophy of language.* Cambridge, MA: MIT Press.

Dunlap, R. A. (1997). *The golden ratio and Fibonacci series*. Singapore: World Scientific Publishing.

Elvevåg, B., Weinstock, M., Akil, M., Kleinman, J.E. & Goldberg, T. E. (2001) A comparison of verbal fluency tasks in schizophrenic patients and normal controls. *Schizophrenia Research 51, 119-126.*

Elvevåg, B., Fisher, J. E., Gurd, J.M. & Goldberg, T. E. (2002). Semantic clustering in verbal fluency: schizophrenic patients versus control participants. *Psychological Medicine 32, 909-917.*

Elvevåg, B., Storms G., Heit, E. & Goldberg, T. (2005). Category content and structure in schizophrenia: an evaluation using the instantiation principle. *Neuropsychology 19, 371-380.*

Ferrer i Cancho, R. (2005a). Decoding least effort and scaling in signal frequency distributions. *Physica A* 345, 275-284.

Ferrer i Cancho, R. (2005b). The variation of Zipf's law in human language. *European Physical Journal B 44, 249-257.*

Ferrer i Cancho, R. (2005c). The consequences of Zipf's law for syntax and symbolic reference. *Proceedings of the Royal Society of London B 272, 561-565.*

Ferrer i Cancho, R. (2005d). Core and peripheral lexicon from word length optimization. Submitted to the *Journal of Quantitative Linguistics.*

Ferrer i Cancho, R. (2005e). Zipf's law from a communicative phase transition. Submitted to *European Physical Journal B*.

Ferrer i Cancho, R. and Reina, F. (2002). Quantifying the semantic contribution of particles. *Journal of Quantitative Linguistics, 9, 35-47.*

Ferrer i Cancho, R. & Solé, R. V. (2001). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *J. Quantitative Linguistics, 8, 165-173*. First appeared as *Santa Fe Institute Working Paper 00-12-068*.

Ferrer i Cancho & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Science USA 100, 788-791.*

Ghyka, C. M. (1927). *Esthétique des proportions dans la nature et dans les arts*. Paris: Gallimard.

Gribbin, J. (1986). *In search of the big bang: Quantum physics and cosmology*. New York: Bantam Books. See http://www.sunspot.noao.edu/sunspot/pr/answerbook/universe.html.

Goldberg, T. E., Dodge, M., Aloia, M., Egan, M. F. & Weinberger, D. R. (2000). Effects of neuroleptic medications on speech disorganization in schizophrenia: biasing associative networks towards meaning. *Psychological Medicine 30, 1123-1130.*

Johnson, C., Davis. H. & Macken, M. (1999). Symbols and structure in language-acquisition. In: Lock, A. & Peters, C.R. (eds.), *Handbook of Human Symbolic Evolution: 686-746*. Oxford: Blackwell.

Kripke, S. A. (1980). Referring to artifacts. *Phylosophical Review 89, 109-114.*

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

Köhler, R. (1987). System theoretical linguistics. *Theoretical Linguistics 14, 241-257.*

Macbeth, D. (1995). Names, natural kind terms and rigid designation. *Philosophical Studies 79, 259-281.*

McCowan, B., Hanser, S. F. & Doyle, L. R. (1999). Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. *Animal Behavior 57, 409-419.*

McCowan, B., Doyle, L. R. & Hanser, S. F. (2002). Using information theory to asses the diversity, complexity and development of communicative repertoires. *Journal of Comparative Psychology 116, 166-172.*

Montemurro, M. (2001). Beyond the Zipf-Mandelbrot law in quantitative linguistics. *Physica A 300, 567-578.*

Montemurro, M. & Zanette, D. (2002). Frequency-rank distribution in large text samples: phenomenology and models. *Glottometrics 4, 87-98.*

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

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, in press. *cont-mat/0412004.*

Piotrowska, X., Pashkovska, W., & R. Piotrowski, R. (to appear). Pathological text and its statistical parameters.

Piotrowski, R.G., Pashkovskii, V.E., Piotrowski, V.R. (1995). Psychiatric linguistics and automatic text processing. In: *Automatic Documentation and Mathematical Linguistics, 28(5), 28-35.* [First published in Naučno-Techničeskaja Informacija, Serija 2, Vol. 28, No. 11. pp. 21-25, 1994].

Pulvermüller, F. (2003). *The neuroscience of language. On brain circuits of words and serial order.* Cambridge: Cambridge University Press.

Tuldava, J. (1996). The frequency spectrum of text and vocabulary. *Journal of Quantitative Linguistics 3, 38-50.*

Whitehorn, J.C. & Zipf, G.K. (1943). Schizophrenic language. *Archive of neurology and psychiatry 49, 831-851.*

Zipf, G. K. (1932). Selected studies of the principle of relative frequency in language. Cambridge, MA: Harvard University Press.

Zipf, G. K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Boston, MA: Houghton-Mifflin.

Zipf, G. K. (1942). Children's speech. *Science 96, 344-345*.

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