Glottometrics 6
2003

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. 6 (2003), Lüdenscheid: RAM-Verlag, 2003. 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 6 (2003) **ISSN 2625-8226**

Contents

Glottometrics 6, 2003, 1-8

Some Aspects of the Power Law

Luděk Hřebíček, Prague[1]

Abstract. This is a fragmentary attempt at seeking the modeling ideas of structures and their semantic validity on the basis of a certain linguistic experience; at seeking their connections with the approaches of several other sciences, for which brain functions are or may appear objects of analyses. Surpassing the boundary of linguistics is in favor of linguistics. The preliminary character of this reasoning is evident.

Keywords: power laws, compositeness, ranking

The search for a basic modeling idea of different forms of the power law seems to be a topical problem in linguistics and in the sciences relevant for natural languages. One can frequently find statements concerning the lack of a universal model explaining the respective power law and the affiliated structures like Zipf's law and Shannon's information theory (see, e.g., the explanation by Balasubrahmanyan and Naranan 2002). Zipf's law of least effort is usually referred to as the basic principle of Shannon's information theory (see, e.g., Schroeder 1991: 33 ff.). Mandelbrot (1953) is quoted as that author who first put these two points together. According to the cited mathematicians, not only the two but many other points can be understood as variants of the power law. Among the structures observed in the nature, however, many more relationships appear to be based on the same law. The frequent occurrence and many-sidedness of the law seems to be the main difficulty in seeking a unified model for all its aspects.

Language is a phenomenon anchored in biological objects, and therefore there is no wonder that linguists operate with structures described with the help of certain modifications of the power law. Among the items of the real world including language, Zipf's and Shannon's discoveries with their broad theoretical consequences belong to this category of relationships. Another fundamental linguistic principle is Menzerath-Altmann's linguistic law that has a pure appearance of a power law. Small wonder that linguists try to formulate some more general conditions for the power law using the concepts pertaining to the formulation of Menzerath-Altmann's law.

1. Menzerath-Altmann's law $y = Ax^{-b}$ defines the relation between the size of a language construct x and of its constituents y; A and b are constants. For details see Menzerath (1928, 1954), Altmann (1980), Köhler (1986), Altmann and Schwibbe et al. (1989). Some applications of the law to text semantics are presented in Hřebíček (1989, 1995, 1996, 1997). The general principle of this law can be (in English) called the *principle of compositeness* and *Prinzip der Konstituenz* (in German). To the best of our knowledge, the term was first used and discussed by K.-H. Best (2001a,b).

Let us formulate the following general statement: Any item of the real world always

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

contains one or more constituents.

This is, as a matter of fact, almost identical with the definition of a concrete system, cf. Bunge (1979: 6): "An object is a concrete system if it is composed of at least two different connected things".

Such a statement requires at least an informal analysis asking questions about its cogency. One can conjecture different degrees of incorporation proper to constructs and their constituents. Considering the constituents as components of a component system (cf. Csányi 1989, Kampis 1991) we must admit both different degrees of incorporation and the possibility of evolutionary addition or elimination of components. The zero degree then means that the supposed phenomena have nothing in common, i.e. A is not a component of B. The highest degree signifies the full adherence of a phenomenon to a given construct. Between these limits there is an arbitrary number of degrees whose measurement is in no case simple. It involves perhaps fuzzy mathematics or the admittance of attractors. This idea perhaps amplifies the discussed principle to all kinds of (hierarchical) relationships considered between different phenomena. In that case the discovery of a relationship of non-zero degree stimulates to search for the encompassing (higher) system. This imprecise picture indicates that the relations between constructs and constituents are sometimes more or less free and sometimes more or less close and it needs not be exprimable by a simple function. Any form of the power law seems to be a more precise expression of this principle. It concerns not only the intensity of the discussed relationship, but it can be presented as a function describing relationships between hierarchically different items. Consequently, reality is full of constructs and their constituents. Power law appears to be an attempt at a more exact approach to the generality contained in the idea declared by the above presented statement. The question to be solved in the future are the limits created by the principle of emergence, as well as the ability of a system to generate new structures.

The simple idea of constituents' affiliation to constructs calls the idea of an increasing pressure inside the space of a construct when the number of its constituents arises. This indicates the possibility of an attracting or, alternatively, pressing force participating in these processes and relations. This force (or "force") occurs in miscellaneous shapes, one of which doubtlessly is Zipf's 'effort' and 'least effort'.

Another possible appearance of this force is Newton's gravitation. To this form of affiliation the notions of a construct (for example, a planetary system) and constituents (its planets) can be applied. Manfred Schroeder (1991: 33) writes:

> 'Homogeneous power laws, like Newton's universal law of gravitational attraction $F \sim r^{-2}$, abound in nature – dead and alive alike.'

This confirms the generality of the principles formulated by Newton and Zipf, and perhaps also some grade of generality proper to the discussed axiom of compositeness.

2. The authors mentioned above, Balasubrahmanyan and Naranan as well as Schroeder, convincingly indicate that the power law represents the basic structure of Shannon's information theory. In linguistics this theory was many times proved in applications to texts. The question can be asked whether this aspect is in accordance with the above presented general modelling idea of compositness.

In an abstract imagination, inside a construct there can be very few constituents, for example, only one. (The zero number of constituents excludes the general starting idea of compositeness; many other laws are conjectured by sciences besides the power law.) The presence of constituents gives occasion for information transition among them or with some items outside the respective construct. A characteristic degree of organization proper to a given construct and the relevant system can be expressed as entropy; then certain configuration of a

construct with its constituents and with the other constructs is described in probabilities. If in the same text its constituents are reorganized and their number inside a construct increases, the average amount of information pertinent to its mean constituent decreases.

This point of view offers an explication not only for the informational aspect of Shannon's theory, but also for the indirect proportionality that is a characteristic peculiar to Zipf's rank-size relationship. When we ask why to the highest probability (frequency) the lowest rank number is ascribed, we can say that it is because of the principle of compositeness. The increase of the rank number *r* of Zipf's law (and, similarly, of the number of constituents) can be thus conceived by the general assertion discussed. Therefore we are inclined to see the deterministic succession of the explaining ideas in play in the following way:

The principle of compositeness → Decreasing probabilities → Zipf's increasing ranks (= increasing number of constituents)

3. In connection with the problem of word length in a text, Wimmer et al. (1994) and Wimmer and Altmann (1996) proceed from the supposition that the probability of a word-length class x in a text is proportional to the word-length class $(x - 1)$:

$$P_x \propto P_{x-1}. \tag{1}$$

With a constant coefficient of proportionality the following equation is obtained:

$$P_x = aP_{x-1}. \tag{2}$$

Instead of a linear proportionality, an arbitrary function can be assumed behaving as a variable coefficient:

$$P_x = g(x)P_{x-1}. \tag{3}$$

An identical conception was applied by Hřebíček (1989), where Menzerath-Altmann's law was substituted for $g(x)$. Consequently, the following conjectures are supplied:

$$y_x \propto y_1 \text{ and } y_x \propto \frac{1}{x},$$

or

$$y_x \propto \frac{y_1}{x}.$$

where x is the size of a construct and y the size of its constituent(s). If this proportionality is rewritten in a logharitmic form and supplemented by coefficient ln c the following equation is obtained:

$$\ln y_1 - \ln y_x = \ln c \ln x. \tag{4}$$

If then for constant c the expression of power law Ax^b is substituted, we have

$$(\ln y_1 - \ln y_x)\frac{1}{\ln x} = \ln A + b \ln x. \tag{5}$$

With $a = \ln A$ we obtain the formula:

$$y_x = y_1 x^{-(a+b \ln x)}, \qquad x = 1, 2, \ldots \tag{6}$$

It is evident that Menzerath-Altmann's law is the (linguistic) variant of the power law with negative *b*. If y_1 is the normalizing constant, function (6) is identical with the so-called Zipf-Alekseev distribution, see Alekseev (1978), Dolinskij (1988), Hammerl (1989), Altmann (1992), Wimmer and Altmann (1999: 665), otherwise it is a generalized power function.

The sense of the negative and positive value of *b* can be explained in the following way: The first approximation to Menzerath-Altmann's law can be also written in a logarithmic form, in which it is evident that *b* functions as a coefficient of proportionality:

$$\log y \propto -b \log x.$$

The mean size of *y* corresponds to a given *x*, and it can be written as y_x. If *x* increases by one, its relation to the value of y_x can be expressed as

$$(7) \qquad \frac{y_{x+1}}{y_x} = \frac{(x+1)^{-b}}{x^{-b}} \quad , \quad x = 1,\ 2,\ \ldots$$

and

$$(8) \qquad y_{x+1} = y_x\left(\frac{x+1}{x}\right)^{-b} = y_x\left(\frac{x}{x+1}\right)^{b} \ .$$

Two parallel progressions can be derived from these relations, with $y_1 = A$ each. So we have:

$$(9) \qquad y_2 = \left(\frac{2}{1}\right)^{-b} A \ \Rightarrow \ y_2 = \left(\frac{1}{2}\right)^{b} A, \quad y_3 = \left(\frac{3}{2}\right)^{-b}\left(\frac{2}{1}\right)^{-b} A \ \Rightarrow \ y_3 = \left(\frac{2}{3}\right)^{b}\left(\frac{1}{2}\right)^{b} A, \text{ etc.}$$

The general terms of the progressions are

$$(10) \qquad y_x = A\left(\frac{x!}{(x-1)!}\right)^{-b} \quad \text{and} \quad y_x = A\left(\frac{(x-1)!}{x!}\right)^{b}.$$

When the fractions are reduced, the power law with negative and with positive *b* is obtained. The difference evidently originates from the view of the relation, i.e., from constituent to construct, or inversely from construct to constituent; this means its ambivalence in general sense, its symmetry. (Further we will, however, indicate that in linguistic sense, in a given semantic frame, the direction of the relation cannot be arbitrarily changed.)

Consequently, when Menzerath-Altmann's law is specified by the above presented two additional conjectures (both of which are in a complete agreement with the presumptions of the law), the obtained mathematical structure is able to describe the relationships of constructs and their constituents. This is a more exact or minute kind of description, because it contains two parameters: *a* and *b*. When formula (6) is applied to the language constructs on different levels (and in different texts of different languages), it also appears to be a statistically relevant formula for the observed data concerning the construct–constituent relationship.

As was proved by the authors quoted above, the Zipf-Alekseev distribution appears a better fitting not only for the complex linguistic data but also for the word association data published by psychologists. A word also is a language category, of course; but in psychological tests words and word expressions are used as immediate symbols belonging to meanings. They are sufficiently free from contextual connections; they can be understood as certain semantic units behind word expressions. In texts, their concrete semantic values are specified in word forms and their contextual connections observable on all language levels. In

the psychological word-association tests words are almost completely removed from the majority of language contexts. This augments the domain of applicability of the power law, this time directly to one of the brain functions.

4. Another psychological test also concerns one of the brain functions, see Ebbinghaus (1919: 721-722). This author is noted for being the first psychologist to investigate learning and memory as the higher mental processes in experimental ways, see Seamon (1980: 7-8). Even today H. Ebbinghaus is frequently referred to, for example on the respective Internet addresses, e.g. Shulman (1997). Ebbinghaus' results concerning the memory processes are round up into the so-called *Ebbinghaus' curve*, also called *learning and forgetting curve*.

The process of forgetting was investigated by Ebbinghaus as human ability to retain certain units ordered in sequences. As these units "nonsense syllables" were used. In the first step of the test the percentage of retaining any sequence was 100%. In the subsequent steps the proportion of the remembered syllable sequences lowers. The approximate percents obtained by Ebbinghaus in the repeated tests are:

1	2	3	4	5	6	7	8
100	58	44	36	34	28	25	21

The second row are the percents, the first row are the ranks of the subsequent memory trials. The ranks correspond to time intervals of non-equal length. The intervals successively enlarge from one minute to one month. At this moment we are not able to obtain some more exact experimental data, therefore we use the above quoted average figures to test them; this approach is nothing but a preliminary inspection of Ebbinghaus' data. Let us use an arbitrary value as 100% of the memory items, for example 48, and seek the best fit with the help of Altmann-FITTER. An optimization method proceeds in considering equation (6) as a curve (not distribution). The most satisfactory correspondence of the "observed" and computed values can be obtained from two theoretical curves see Table 1.

Table 1
Fitting (1) and the simple power curve to the values corresponding to Ebbinghaus' curve

Rank	**"Observed"**	**Curve (6**	**Simple power curve**
1	48	48.00	48.00
2	28	29.37	28.95
3	21	21.72	21.54
4	17	17.42	17.46
5	16	14.63	14.84
6	13	12.66	12.99
7	12	11.18	11.61
8	10	10.02	10.53

Results:
Formula (1) yields $y = 48x^{-0.6863 - 0.0322 \ln x}$ with $R = 0.9978$.
The simple power curve yields $y = 48x^{-0.7293}$ with $R = 0.9972$.

As can be seen, here the contribution of the logarithmic part is not very relevant. It can be concluded that one of the mental processes represented by Ebbinghaus' curve seem to be distributed according to the power law. The same is valid for the structures observed in texts

that are anchored in the mental processes operating with meanings. This has been attested in the works quoted above.

5. Let us also remind Herbert Simon's (1955) mathematical model that describes the generation of a text as a process that leads asymptotically to Zipf's law. Its predominance consists in the better understanding of the origin of Zipf's law, as is stressed by Montemurro and Zanette (2002). These authors bring important modifications of Simon's model. The basic assertions of the model deserve a deeper discussion in connection with power law.

Simon's model starts with the assumption that at each time step t a new word is added to the text. New words are introduced into the text at a constant rate α, so that the vocabulary extent is the function of the independent variable t: $V_t = \alpha t$. The equation $V_t = at^v$ can be supposed as a more realistic form of the same functional relation, both of which are the forms of the power law.

Time, also in linguistics, cannot be completely excluded from the set of assumptions characterizing the process of text generation. The question whether the introduction of the time function turns easier the understanding of languge structures and their semantics may, however, appear controversial. In certain sense Simon's model is based on the identity of time steps and lexical units: the entire vocabulary of any text under those circumstances approximately equals the total time of the text generation. This is quite peculiar approach to the description of lexical units. (The linguistic comprehension is sometimes complicated by the absence of difference between lexical units and word forms in some applications of the model; vocabulary then consists from word forms differentiated by their total morphological form.)

The identity of a time step with a new-word introduction, naturally, can be inverted and we may say that any text is generated in steps that are reckoned in lexical units. Linguists thus need not infer the time function. Further presumption of a sharp difference between new lexical units and the units already used in the same text has a distinct linguistic aspect: it distinguishes *hapax legomena* from the repeated lexical units. According to the most developed text theory described in the work by Ziegler and Altmann (2002: 36), the denotations occurring in text at least in two different lexematic forms make up the semantic *nucleus* of the text; the other units are treated as *text periphery*.

The introduction of one new lexical unit into the text does not mean simply one new unit and nothing more. Such a unit is always set in (at least) two contexts:

(a) narrower context (NC) of a lexical unit, which is a text segment like sentence or some similar syntactic unit;

(b) larger context (LC) of a lexical unit, which are all text segments like sentences, in which a given lexical unit occurs.

It is evident that the number of the units inside each segment (e.g., sentence) indicate the number of LC's into which the same segment enters like the constituent of different LC('s). (The peculiarity of the text *hapax legomena* is that their NC's and LC's basing on a given lexical unit fuse with each other.)

An important pragmatic result: One can easily prove on any continuous text of an arbitrary language having an optimal size (which means that the text is not extremely short being thus able to give statistically measurable data and also is not extremely long being thus semantically homogeneous) that between NC and LC there is a relationship abiding by Menzerath-Altmann's law. This appears to be important for the semantic structure of the text.

The detachment of lexical units occurring only once from the entire text vocabulary in Simon's model can be comprehended as putting the accent on memory. Only language user's

memory can decide whether an introduced unit belongs to the category of the already used units or not. Memory is one of the brain functions, it is closely connected with language, but it has a deeper relationship with semantics (as it is proved, for example, in psychology) which need not always be connected with language. The "nonsense syllables" in the basis of Ebbinghaus' curve do not depend on some complicated hierarchy of before learned language expressions, one can take them as pure forms processed by the memory. The same curve implies that the properties observed as compositeness are generated somewhere in the deep structures of human brain. Their consequences are observable through the linguistic descriptions in which power law plays an important role. Thus we can conjecture that there is a property more substantial when one seeks more fundamental relations generating power law and its modifications.

In an approximate model of the biological structures forming the neural net of the human brain one can conceive of two sets: that of neurons and that of links between the neurons. Both the sets taken together remind physical fields such as electromagnetic or gravitational. Neurology describes actions in the neural nets as electric or chemical impulses; both are referred to be relevant for different types of the memory. Such a picture suggests connective graphs and their subgraphs of the theory of graph. Let us quote the work by J. G. Seamon (1980), where the great number of potential connections between the great number of neurons in human brain are mentioned as permutations. Their mutual relation points to the concepts of constructs and their constituents.

6. *Summary*. This was not an attempt at formulation of a unified theory, of course. The intention of the above presented notes is to indicate certain similarities between somewhat remote systems of different sciences and their descriptions. Our position is that their agreement can scarcely be treated as a pure coincidence. The seeking of their common background is doubtlessly justified. The question arises whether the principle of compositeness cannot be accepted as such a background, as something more general than, for example, the physical concept of the field of energy. Even energy is a phenomenon with difficulty handled in linguistics. The solution of certain scientific branches not to take language into account (e.g., "language is nothing but a code with random relationships betweeen language units and their meanings") can scarcely be henceforth accepted. In the frame of a given text the (semantic) relation between any construct and its constituent(s) inevitably is asymmetric. This relation cannot be reversed without a significant change of the meaning. Let us remind the famous Bertrand Russel's logical paradox and its solution, presented by its discoverer, which seems to have something in common with this irreversibility.

References

Alekseev, P. (1978). O nelinejnych formulirovkach zakona Zipfa. *Voprosy kibernetiki 41*, *53-65.*

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

Altmann, G. (1992). Two models for word association data. *Glottometrika 13*, *105-120.*

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

Balasubrahmanyan, V. K. and Naranan, S. (2002). Algorithmic information, complexity and Zipf's law. *Glottometrics 4* (*To honor G. K. Zipf), 1-26.*

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

Best, K.-H. (2001b). *Häufigkeitsverteilungen in Texten.* (Göttinger Linguistische Abhandlungen 4), Göttingen: Peust & Gutschmidt Verlag.

Bunge, M. (1979). *A world of systems*. Dordrecht: Reidel.

Csányi, V. (1989). *Evolutionary systems: a general theory.* Durham: Duke University.

Dolinskij, V. A. (1988). Raspredelenie reakcij v eksperimentach po verbal'nym associacijam. *Acta et Commentationes Universitatis Tartuensis 827, 89-101*.

Ebbinghaus, H. (1919). *Grundzüge der Psychologie*. Erster Band. Bearbeitet von Karl Bühler. Leipzig: von Veit.

Hammerl, R. (1989). Untersuchungen zur Verteilung der Wortarten im Text. *Glottometrika 11, 142-156.*

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

Hřebíček, L. (1995). Text levels. Language constructs, constituents and Menzerath-Altmann law. Wissenschaftliches Verlag Trier.

Hřebíček, L. (1996). Word associations and text. *Glottometrika 15*, *96-101.*

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

Kampis, G. (1991). Self-modifying systems in biology and cognitive science. Oxford: Pergamon.

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

Mandelbrot, B. B. (1953). An informational theory of the statistical structure of language. In: W. Jackson (ed.), *Communication theory: 486*. London: Butterworths [Quoted according to Balasubrahmanyan and Naranan 2002.]

Menzerath, P. (1928). Über einige phonetische Probleme. *Act du premier congrès international de linguistes* : 104-105. Leiden: Sijthhoff.

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

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

Schroeder, M. (1991). Fractals, chaos, power laws. Minutes from an infinite paradise. New York: Freeman.

Seamon, J. G. (1980). *Memory and cognition. An introduction.* New York-Oxford: Oxford UP.

Shulman, H. G. (1997). *Psychology 312: Memory and cognition.* The Ohio State University. [The Internet address: www.psy.ohio-state.edu/psy312.]

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

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

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

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

Ziegler, A. and Altmann, G. (2002). *Denotative Textanalyse. Ein textlinguistisches Arbeitsbuch.* Wien: Edition Praesens.

Software

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

Glottometrics 6, 2003, 9-34

Spracherwerb, Sprachwandel und Wortschatzwachstum in Texten. Zur Reichweite des Piotrowski-Gesetzes

Karl-Heinz Best, Göttingen[1]

Abstract. This paper presents a number of examples proving the fact that language change abides by the so-called Piotrowski Law, which takes, in simple cases, the shape of a sigmoid curve („S-curve"). This law of language change seems to go back to a hypothesis of Osgood & Sebeok (1954); Piotrovskaja & Piotrovskij (1974), Altmann (1983), Altmann, v. Buttlar, Rott & Strauss (1983) and others proposed mathematical models of this hypothesis and applied them to language changes successfully. The aim of this paper is to demonstrate how many different kinds of processes abide by the Piotrowski Law.

Keywords: Piotrowski law, language learning, language change, vocabulary growth

1. Die S-Kurve als Modell für den Verlauf von Sprachwandelprozessen

Einer der vielen Aspekte, mit denen sich die Sprachtheorie bei der Suche nach Sprachgesetzen (Best 2001b,c; Best [Hrsg.] 2001) zu befassen hat, ist die Frage danach, wie Sprachwandel verläuft. Entwickeln sich die Veränderungen chaotisch oder gesetzmäßig? Und wenn sie gesetzmäßig erfolgen: Welchem oder welchen Gesetz(en) entsprechen sie? In der Linguistik findet sich seit geraumer Zeit die Ansicht, dass Sprachwandel einen S-förmigen („sigmoiden") Verlauf nimmt; diese Ansicht vertreten bereits Osgood & Sebeok (1954/1965: 155):

> „*Language change in a community will be gradual and cumulative, representing a continous changing proportion of individuals who do or do not hear and produce a particular feature or set of features. The process of change in the community would most probably be represented by an S-curve.*"

Seitdem hat dieser Gedanke große Verbreitung und Akzeptanz gewonnen, wie sich aus der Tatsache ergibt, dass man ihn inzwischen in Crystals *Enzyklopädie* (Crystal 1993: 332) ebenso findet wie z.B. bei Bailey (1973: 77) und Labov (1994: 65); Aitchison (1991: 76ff.) widmet diesem Konzept sogar ein ganzes Kapitel ihres Buches und führt aus: „Recent research suggests that a typical change fits into a slow-quick-quick-slow pattern" (Aitchison 1991: 83). Es gibt auch hinreichend empirische Befunde, die dem Augenschein nach die Annahme eines S-förmigen Verlaufs mancher Sprachwandelprozesse stützen.

All diesen Überlegungen und Befunden fehlt aber eine mathematisch formulierte Hypothese, die es erst erlaubt, empirische Erhebungen daraufhin zu testen, ob sie die Theorie stützen oder widerlegen. Einen frühen Versuch zu einer solchen mathematischen Modellierung haben Piotrovskaja & Piotrovskij (1974) unternommen. Dieser Vorschlag wurde aber als unzulänglich kritisiert (Altmann u.a.1983: 106) und durch ein anderes, linguistisch besser begründetes Modell für den vollständigen, unvollständigen und reversiblen Sprachwandel ersetzt (Altmann 1983: 60-62; Altmann u.a.1983: 106f.; vgl. auch Leopold 1998: 99ff). Die

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

theoretische Herleitung für den speziellen Fall der Fremdwortübernahme wurde von Beöthy & Altmann (1982) sowie Best & Altmann (1986) entwickelt und anhand etlicher Fälle erfolgreich getestet. Die Grundidee der Modellierung dieser Fälle ist immer, dass sich neue Formen in Konkurrenz zu alten entwickeln bzw. dass sich die Fremdwortübernahme im Wechselspiel zwischen bereits entlehnten und noch nicht entlehnten, aber entlehnbaren Wörtern vollzieht. Diese Theorie hat sich mittlerweile vielfach bewährt (Best & Kohlhase [Hrsg.] 1983; Best 2001a; Best 2001c: 102-113; Best 2002a,b,c, 2003, 2003a,b,c,d; Körner 2002).

Auf der gleichen theoretischen Grundlage untersuchte Tuldava (1998: 138) das Wachstum des estnischen Wortschatzes und schlug dafür ein Modell vor, das genau dem des unvollständigen Sprachwandels entspricht (vgl. Altmann 1983: 61, Formel 7); auch das Wachstum des englischen Wortschatzes ließ sich auf diesem Wege modellieren (Best 2001c: 108f.).

Es geht nun hier nicht darum, die Entwicklung der Theorie des Verlaufs von Sprachwandelprozessen nachzuzeichnen; das kann in der angegebenen Literatur nachgelesen werden. Vielmehr soll gezeigt werden, welche Reichweite bisher dieser Theorie zugesprochen werden kann. Zunächst werden aber die Formen, die das Sprachwandelgesetz annehmen kann, kurz vorgestellt.

2. Das Piotrowski-Gesetz als Modell sprachlichen Wandels

In Übereinstimmung mit den Sprachwandelprozessen, für die hinreichende statistische Erhebungen greifbar waren, entwickelten Altmann u.a. (1983) sowie Altmann (1983) im Anschluss an Piotrovskaja & Piotrovskij (1974) einen Vorschlag, dem sie den Namen „Piotrowski-Gesetz“ gaben:

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

Dieses Gesetz modelliert den S-förmigen Verlauf von Sprachwandel in zwei Formen:

1. Mit $c = 1$ erhält man den sogenannten „vollständigen Sprachwandel“, bei dem alle alten Formen nach einer gewissen Zeit durch neue ersetzt werden; ein Beispiel dafür sind die Ersetzungen des -{t} für die 2. Person Sg. durch -{st} bei einigen Modalverben im Deutschen (Best 1983; Best 2001c: 102ff.).

2. Mit $c \neq 1$ kann man den „unvollständigen Sprachwandel“ modellieren; es handelt sich dabei u.a. um Entlehungs- und Wachstumsprozesse im Lexikon und in der Wortbildung (Beöthy & Altmann 1982; Best & Altmann 1986; Tuldava 1998), aber auch um Veränderungen in der Syntax (Best 2002b).

Neben diesen beiden Formen des Piotrowski-Gesetzes, die mit der Annahme des S-förmigen Verlaufs übereinstimmen, wurde für „reversible Sprachwandel“ eine weitere Form entwickelt:

$$p_t = \frac{1}{1+ae^{-bt+ct^2}}.$$

Damit lassen sich Sprachwandelprozesse modellieren, die zunächst wie der S-förmige Wandel einsetzen, dann aber irgendwann in ihrem Verlauf eine Trendwende erfahren und wieder ganz oder teilweise aus der Sprache verschwinden. Ein Beispiel hierfür ist die e-Epithese bei starken Verben (Imsiepen 1983) und bei Hilfsverben im Deutschen (Best 2001c: 111ff.). Diese Form des Sprachwandels findet oft keine Erwähnung − er fehlt z.B. bei Aitchison −, obwohl er bereits von Piotrovskaja & Piotrovskij (1974: 378) vorgestellt wurde.

Der Vollständigkeit halber sei erwähnt, dass im Anschluss an die Entwicklung dieser drei Grundformen des Piotrowski-Gesetzes noch eine verallgemeinerte Form entwickelt wurde, die alle drei Prozesse beinhaltet (Best, Beöthy & Altmann 1990). Die Anwendung dieses verallgemeinerten Piotrowski-Gesetzes erfordert jedoch die Abschätzung von 4 Parametern, was oft mangels hinreichender Daten nicht möglich ist. Aus diesem Grund wird darauf verzichtet, dieses Modell hier weiter zu verfolgen.

3. Zur „Reichweite“ des Piotrowski-Gesetzes

In diesem Abschnitt soll es nun darum gehen, zu zeigen, dass das Piotrowski-Gesetz in seinen drei vorgestellten Formen sich bei einer Vielfalt von Sprachveränderungsprozessen bewährt. Dabei handelt es sich teils um Prozesse, die das Sprachsystem umgestalten, teils aber auch um solche, die lediglich seine Verwendung betreffen oder auch Wachstumsprozesse darstellen. Alle diese Veränderungen verhalten sich offenbar prinzipiell gleich, indem sie sich als S-förmiger oder reversibler Sprachwandel darstellen. Aus diesem Grund werden sie in den folgenden Ausführungen auch nicht gesondert behandelt.

Die folgenden Tests werden mit dem Programm NLREG (2001) durchgeführt; a, b und c sind die Parameter, die geschätzt werden müssen; D ist der Determinationskoeffizient, der eine akzeptable Übereinstimmung zwischen Theorie und Beobachtung anzeigt, wenn $D \geq 0.80$, und eine sehr gute mit $D \geq 0.90$.

3.1. Verwendung des Wortakzents zur Unterscheidung von Verb und Substantiv

Im Englischen kann seit dem 16. Jhd. eine Entwicklung beobachtet werden, bei der zweisilbige Wortstämme zunehmend eine Wortartdifferenzierung durch einen unterschiedlichen Wortakzent erfahren (Sherman 1975; Chen & Wang 1975). Dies betrifft Wörter, die ursprünglich als Substantive wie auch als Verben immer auf der zweiten Silbe betont wurden („isotonic“); z.B. *affix* (Chen & Wang 1975: 261); seit 1755 ist dieses Wort als *áffix* als Substantiv nachgewiesen und damit „diatonic“. Auf diese Weise änderten sich noch etliche weitere Wörter; der Prozess stellt sich nach Chen & Wang (1975: 262) wie folgt dar, wenn man das Piotrowski-Gesetz in der Form des unvollständigen Sprachwandels anwendet:

Tabelle 1
Zunahme der „diatonen“ Wörter im Englischen

t	Zeitpunkt	diatone Wörter beobachtet	diatone Wörter berechnet
0	1570	3	8.73
12	1582	8	9.87
90	1660	24	21.45
130	1700	35	31.28
230	1800	70	71.96
364	1934	150	149.73
$a = 25.0246$		$b = 0.0107$ $c = 227.1361$	$D = 0.996$

t beginnt mit dem Jahr 1570 als Zeitpunkt 0 und zählt von da an die Jahre durch. Laut Chen & Wang (1975: 261) kämen für diese Entwicklung insgesamt 1315 Wörter infrage; tatsächlich wurden davon bisher nur 150 erfasst, gerade einmal 11%. Der Wendepunkt der Entwicklung

könnte überschritten sein.

Dieser Sprachwandel folgt dem Piotrowski-Gesetz in der Form

$$p_t = \frac{227.1361}{1+25.0246e^{-0.0107t}}.$$

Die Übereinstimmung zwischen dem Piotrowski-Gesetz und den Beobachtungen zu diesem Sprachwandel ist mit $D = 0.996$ hervorragend, wie auch die Graphik (Abb. 1) zeigt:

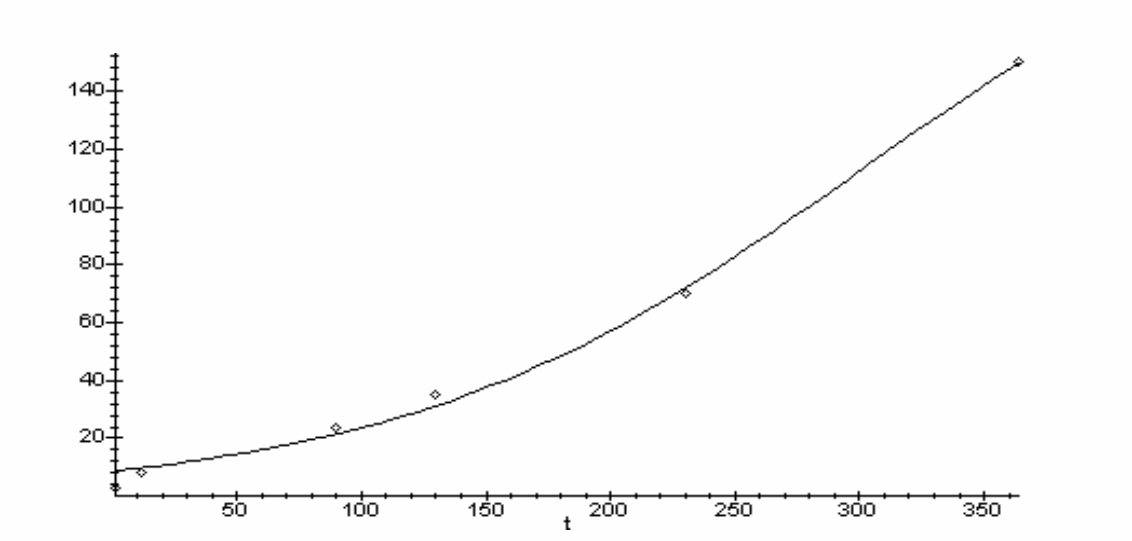

Abb. 1. Zunahme der „diatonen“ Wörter im Englischen

3.2. Flexion

Es gibt etliche morphologische Prozesse, für die gezeigt werden kann, dass sie einer der Formen des Piotrowski-Gesetzes folgen. Einer dieser Sprachwandel ist der Abbau der starken Verben im Deutschen. Dieser stellt sich nach Faust (1980: 400-404) wie folgt dar:

Tabelle 2
Erstes Auftreten schwacher Formen ehemals starker Verben

Jhd.	*t*	*x* (beobachtet)	*x* (kumulativ)	*x* (berechnet)
12.	1	1	1	0.0693
13.	2	2	3	0.3583
14.	3	1	4	1.8002
15.	4	2	6	7.9425
16.	5	18	24	23.0743
17.	6	12	36	36.4039
18.	7	5	41	40.9505
19.	8	1	42	41.9576
	$a = 3161.60$	$c = 42.2046$	$b = 1.6493$	$D = 0.99$

x: schwach gewordene, ehemals starke Verben.

Dem Verlauf entspricht die Formel:

$$p_t = \frac{42.2046}{1+3161.60\,e^{-1.6493t}}.$$

Das Resultat ist in Abb. 2 graphisch dargestellt.

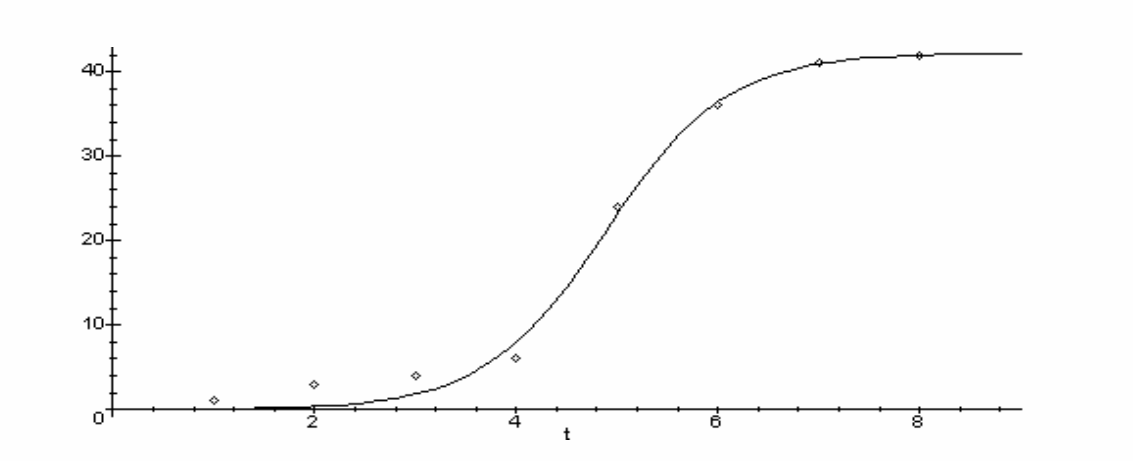

Abb. 2. Erstes Auftreten schwacher Formen ehemals starker Verben

Tabelle 3
Letztes Auftreten schwacher Formen ehemals starker Verben

Jhd.	*t*	*x* (beobachtet)	*x* (kumulativ)	*x* (berechnet)
15.	1	3	3	5.5645
16.	2	16	19	17.0253
17.	3	13	32	32.3567
18.	4	7	39	40.8113
19.	5	5	44	43.4214
20.	6	1	45	44.0698
	a = 30.2339	*c* = 44.2674	*b* = 1.4695	*D* = 0.99

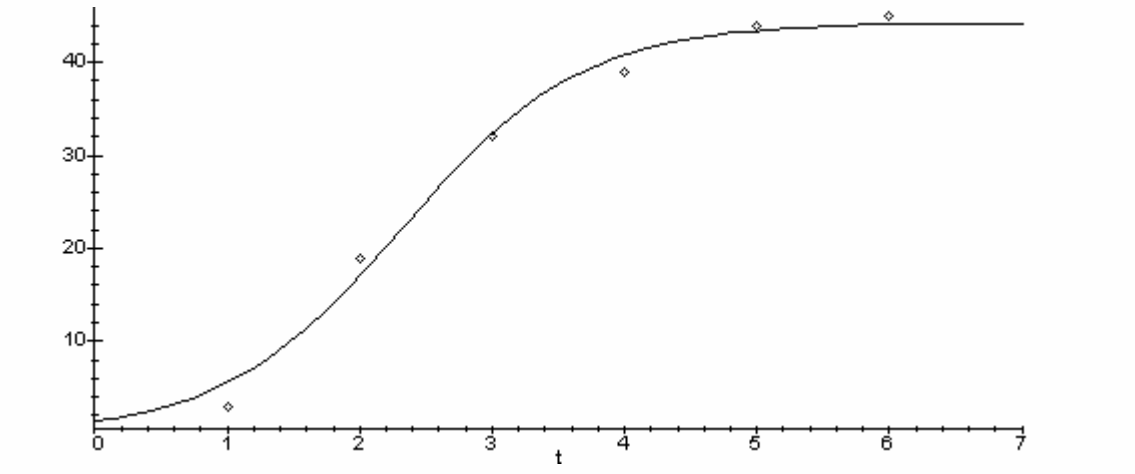

Abb. 3. Letztes Auftreten schwacher Formen ehemals starker Verben

Dieser Sprachwandel folgt damit der Formel:

$$p_t = \frac{44.2674}{1 + 30.2339e^{-1.4695t}}$$

Der graphische Verlauf ist in Abb. 3 zu sehen.

3.2.1. Exkurs: Zur Erklärung von Sprachwandel

An dieser Stelle sei ein Exkurs erlaubt, der auf das Problem der Erklärung von Sprachwandel eingeht, so wie verschiedene Ansätze versuchen, Lösungsmöglichkeiten zu entwickeln. Ohne auf wissenschaftshistorische Aspekte im Einzelnen eingehen zu wollen, lassen sich verschiedene Positionen nennen, darunter die „Natürlichkeitstheorie“ und die „Ökonomietheorie“ (s. dazu Nübling 2000: 249ff.). Sprachwandel erklärt sich gemäß den Vertretern der „Natürlichkeitstheorie“ dadurch, dass universelle, typologische oder systemspezifische Natürlichkeits- bzw. Markiertheitsprinzipien zur Wirkung kommen, darin in begründeten Fällen einbezogen die sog. „Markiertheitsumkehrung“; die Vertreter der „Ökonomietheorie“ wiederum finden die Erklärung für Sprachwandelprozesse darin, dass ökonomische Prinzipien bei hochfrequenten sprachlichen Erscheinungen andere Wirkungen zeitigen als bei niederfrequenten.

Ein unstrittiger Fall dürfte für die Spielarten beider Ansätze die Tatsache sein, dass im Falle eines analogischen Ausgleichs im Paradigma zwischen Singular- und Pluralformen der Singular sich gegen den Plural durchsetzt: der Singular ist semantisch und fast immer auch symbolisch die weniger markierte Kategorie und er ist auch fast immer die häufigere Kategorie. In diesem Zusammenhang sei die Umgestaltung der Paradigmen der Hilfsverben *sein* und *werden* im Präteritum aufgegriffen. Zunächst der analogische Ausgleich von *was* (Präteritum zu *sein*) zum heute verwendeten *war*. Die folgende Tabelle ergänzt die Daten, die bereits in Best (1983) vorgestellt wurden.

Tabelle 4
Die Ersetzung von *was(e)* durch *war(e)*

t	Zeitraum	*was(e)*	*war(e)*	Σ	beobachtet	berechnet
1	1430-1439	440	1	441	0.2268	0.3625
2	1440-1449	416	11	427	2.5761	0.5885
3	1450-1459	445	5	450	1.1111	0.9539
4	1460-1469	1898	8	1906	0.4197	1.5428
5	1470-1479	2596	16	2612	0.6126	2.4861
6	1480-1489	1977	22	1999	1.1006	3.9828
7	1490-1499	1673	23	1696	1.3561	6.3221
8	1500-1509	1754	31	1785	1.7369	9.8938
9	1510-1519	1930	90	2020	4.4554	15.1569
10	1520-1529	1172	408	1580	25.8228	22.5201
11	1530-1539	1212	687	1899	36.1769	32.1066
12	1540-1549	713	821	1534	53.5202	43.4837
13	1550-1559	1198	1467	2665	55.0468	55.5914
14	1560-1569	424	1157	1581	73.1815	67.0695
15	1570-1579	439	1124	1563	71.9130	76.8180
16	1580-1589	381	1053	1434	73.4310	84.3539
17	1590-1599	128	1241	1369	90.6501	89.7664
18	1600-1609	150	1039	1189	87.3843	93.4519

19	1610-1619	60	1324	1384	95.6647	95.8711
20	1620-1629	29	1187	1216	97.6151	97.4213
21	1630-1639	11	1147	1158	99.0500	98.3991
22	1640-1649	7	1487	1494	99.5315	99.0099
23	1650-1659	0	1338	1338	100.0000	99.3892
24	1660-1669	2	1613	1615	99.8762	99.6237
25	1670-1679	1	1619	1620	99.9381	99.7684
26	1680-1689	1	1586	1587	99.9369	99.8575
$a = 446.9696$ $b = 0.4867$ $D = 0.99$						

Dieser Sprachwandel, dessen graphische Darstellung man in Abb. 4 findet, folgt mit hervorragendem Testergebnis von $D = 0.99$ der Formel

$$p_t = \frac{1}{1+446.9696\,e^{-0.4867t}}.$$

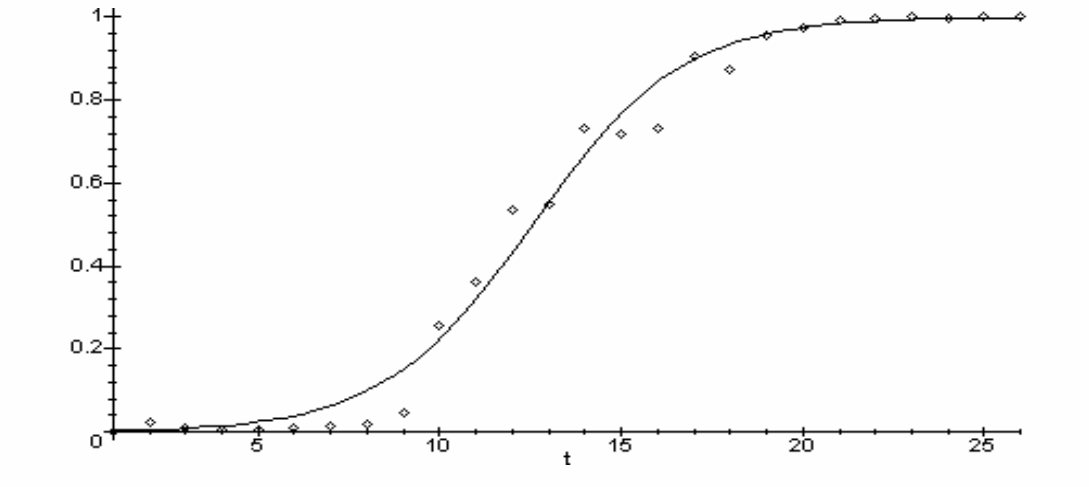

Abb. 4. Die Ersetzung von *was(e)* durch *war(e)*

Wie beim Verb *sein* wird auch bei der Ersetzung von *ward* durch *wurde* (Best & Kohlhase 1983) die Singularform nach dem Vorbild des Plurals umgestaltet. Diese Beobachtungen veranlassen zu den nun folgenden Überlegungen.

Der paradigmatische Ausgleich im Präteritum von *sein* und *werden* entwickelte sich im Gegensatz zu den eben skizzierten Annahmen der Natürlichkeits- und der Ökonomietheorie, obwohl in frühneuhochdeutscher Zeit, als diese Sprachwandel sich durchzusetzen begannen, auch die erwartete Tendenz zu beobachten ist: Es gab durchaus Formen wie *wasen* (statt *waren*) und *warden* (statt *wurden*), die den genannten Theorien entsprechen. So findet man in der *Schedelschen Weltchronik* (Nürnberg 1493) in ausgewählten Textpassagen 41 Belege für *warden* neben nur 2 Belegen für *wurden*; in *Ein kurtzweilig Lesen von Dil Ulenspiegel* (Straßburg 1515) finden sich in einer längeren Textpassage ebenfalls *warden* (5 Belege) neben *wurden* (17 Belege). In *Johann Knebel, Die Chronik des Klosters Kaisheim* von 1531 stößt man für den Plural von *sein* im Präteritum sowohl auf die Form *wasen* als auch auf *wasend*. Obwohl also auch die erwartbaren Ausgleichsprozesse zu beobachten sind, dass nämlich der seltenere bzw. markierte Plural nach dem Vorbild des Singulars umgestaltet wird, wird langfristig bei diesen beiden Verben der Singular an den Plural angeglichen.

Man kann für diese beiden Fälle wohl auch nicht Markiertheits- oder Häufigkeitsumkehrung geltend machen: Anders als bei der Ersetzung von *las* durch *lar* nach dem Plural *lares*

(Mayerthaler 1981: 49) kann nicht behauptet werden, dass bei diesen Verben fast ausschließlich der Plural verwendet werde, wie die folgende Übersicht zeigt:

Tabelle 5
Zur Häufigkeit der 3. Ps. Singular und Plural Indikativ im Präteritum von *werden* in einigen fnhd. Texten bzw. Textabschnitten

Jahr	Text	Singular	Plural
1430	Chronik v. Augsburg	153	64
1444	Chronik v. Nürnberg	87	35
1445	Chronik v. Köln	153	42
1466	Mentel-Bibel	128	38
1471	Steinhöwel: Appollonius v. Tyrus	102	20
1472	v. Eyb, Ehebüchlein	103	20
1493	Marquard v. Stein, Ritter v. Turn	128	24
1499	Chronik v. Köln	130	37
1509	Fortunatus	188	43
1515	Dil Ulenspiegel	183	22
1519	Wigalois	169	28
1520	Joh. Adelphus, Barbarossa	101	27

Auch wenn diese Tabelle sich nur auf wenige Texte stützt und keine Daten zu *was*, *war* und *waren* enthält, darf man davon ausgehen, dass auch in fnhd. Zeit etwa die gleichen Häufigkeitsrelationen gelten, die man in Frequenzwörterbüchern für die Gegenwartssprache findet. Auch die „Ökonomietheorie“, die sich bei ihren Annahmen ausdrücklich auf die Frequenz beruft, dürfte mit diesen beiden Fällen analogischen Ausgleichs ihre Probleme haben. Mańczaks Gesetz 2a) (1980: 40) und Fenk-Oczlons „frequentistische Gesetzeshypothese“ (1991: 362) fordern ebenfalls eine andere Richtung des analogischen Ausgleichs als die beobachtete, lassen aber Ausnahmen zu; erklärt sind diese damit nicht. In diesem Zusammenhang sei auf Köhlers Konzept einer linguistischen Synergetik hingewiesen, der für die Gestaltung des Sprachsystems eine ganze Reihe von Einflussgrößen verantwortlich macht, auch solche, die in keiner der genannten Theorien eine Rolle spielen (Köhler 1986): die „Systembedürfnisse“ (gemeint: die Bedürfnisse, die die Sprecher und Hörer gegenüber ihrer Sprache zur Geltung bringen) und die Wechselbeziehungen, die zwischen den Entitäten der Sprache herrschen. Es muss vorläufig offen bleiben, ob die Erweiterung dieses Modells einmal zu der Einflussgröße führt, die für die beiden fraglichen Prozesse verantwortlich ist. Vielleicht muss man in einigen Fällen auch in Betracht ziehen, dass es manchmal nur wichtig ist, dass ein Ausgleich stattfindet, nicht aber, in welcher Richtung er sich vollzieht.

Weitere Beispiele aus der Morphologie, die einer der Formen des Piotrowski-Gesetzes folgen, sind folgende: Beim Verb *sein* ist noch ein zweiter Prozess zu beobachten: die e-Epithese. Ebenso wie bei den starken Verben (Imsiepen 1983) und dem Verb *werden* (Best & Kohlhase 1983) treten über mehrere Jahrhunderte hinweg im Präteritum Formen auf, die ein zusätzliches -<e> aufweisen: *wase* statt *was* bzw. *ware* statt *war* (Best 1983; 2001c: 112f.; 2003b: 119-121). Dieser Fall und auch die Umgestaltung der 2.Ps.Sg. der Modalverben (Best 1983; Best 2001c: 103-105) ebenso wie die Umstrukturierung des Gen.Pl. bei einigen Maßeinheiten im Russischen (Piotrovskaja & Piotrovskij 1974: 367; Altmann u.a. 1983: 114) entsprechen dem Piotrowski-Gesetz.

In einem Fall konnte nachgewiesen werden, dass auch einzelne Autoren ihre Sprachformen entsprechend dem Piotrowski-Gesetz ändern. So hat Kohlhase (1983) gezeigt, dass der Nürnberger Chronist Heinrich Deichsler im Lauf der Arbeit an seiner Chronik die Form *ward*

immer mehr durch die Form *wurd* ersetzte. (Vgl. dazu und zu einem weiteren Fall idiolektalen Wandels: Best 2003d.)

3.3. Wortbildung

Aus dem Bereich der Wortbildung sind ebenfalls einige Übernahme- und Zuwachsprozesse so gut dokumentiert, dass untersucht werden kann, ob sie dem Piotrowski-Gesetz entsprechen. Dabei handelt es sich um den Ausbau der Wörter auf *-ität* (Best 2001c: 107), auf *-ical* (Best 2002c) sowie auf *-ion* (Körner 2002). Die Übereinstimmung der Theorie mit diesen Fällen ist jedesmal überzeugend. Bei den *-bar*-Adjektiven (Flury 1964:93) liegt ein etwas anders gelagerter Fall vor: Es handelt sich nicht um einen Prozess, bei dem Entlehnung aus einer anderen Sprache eine Rolle spielt; die langfristige, allmähliche Zunahme dieser Adjektive wird überlagert von einem kurzfristig im 19. Jhd. auftretenden sehr starken Zuwachs, der aber offenbar zum 20. Jhd. hin wieder zurückgeht. Man kann in solchen Fällen testen, ob diese kurzfristige Entwicklung dem Piotrowski-Gesetz entspricht, wenn genügend Daten für kurze Zeiträume vorliegen; das ist aber leider nicht der Fall. Wenn man den extremen Zuwachs im 19. Jhd. bei der Berechnung einmal außer Acht lässt, stellt sich der langfristige Trend wie folgt dar:

Tabelle 6
Zuwachs der *-bar*-Adjektive im Deutschen: alle Wörter (n. R. Flury 1964: 93)

t	Jahrhundert	beobachtet	berechnet
1	15.	105	83.9998
2	16.	122	161.9427
3	17.	305	276.8348
4	18.	400	405.5129
5	19.	889	-
6	20.	575	580.0398
	a = 15.0649	b = 0.8046	c = 650 D = 0.98

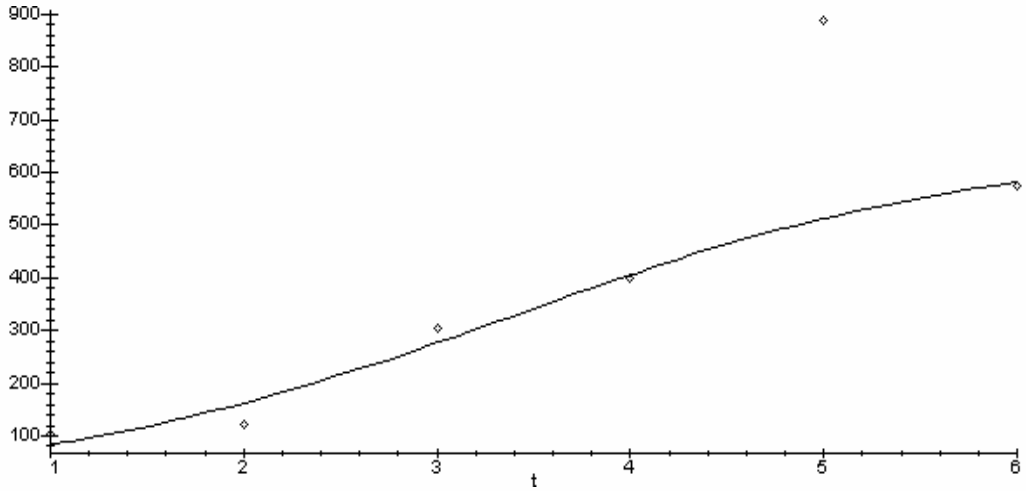

Abb. 5. Zuwachs der *-bar*-Adjektive im Deutschen

Die Graphik in Abb. 5 zeigt, wie sehr der Wert für das 19. Jhd. aus dem allgemeinen Trend ausbricht. Dieser Sprachwandel folgt der Form

$$p_t = \frac{650}{1+15.0649e^{-0.8046t}}.$$

Man kann das Piotrowski-Gesetz auch an die Datei einschließlich des 19. Jhds. anpassen, erhält dann mit $D = 0.79$ aber ein wesentlich schlechteres, nicht ganz zureichendes Ergebnis.

3.4. Syntax

Auch einige syntaktische Phänomene ändern sich im Lauf der Zeit und folgen dabei dem Piotrowski-Gesetz. Einer dieser Fälle ist die Zunahme des bestimmten Artikels in Sprachdenkmälern aus romanischer und altfranzösischer Zeit, der sich nach Piotrowski, Bektaev, & Piotrowskaja (1985: 44) folgendermaßen entwickelt:

Tabelle 7
Zunahme der Textfrequenz des bestimmten Artikels
(Anteil möglicher Vorkommen)

t	Zeitraum	beobachtet	berechnet
1	8./9. Jhd.	0.096	0.0559
2	10. Jhd.	0.106	0.1164
3	11. Jhd.	0.169	0.2057
4	12. Jhd.	0.319	0.2973
5	13. Jhd.	0.392	0.3622
6	14. Jhd.	0.370	0.3970
$a = 16.3933$	$b = 0.9118$	$c = 0.4244$	$D = 0.94$

Abweichend von Piotrowski, Bektaev, & Piotrowskaja wurde für jedes Jhd. nur ein Wert für die relative Textfrequenz des bestimmten Artikels angegeben. Dieser Sprachwandel folgt dem Piotrowski-Gesetz in der Form (vgl. auch Abb. 6)

$$p_t = \frac{0.4244}{1+16.3933\,e^{-0.9118t}}.$$

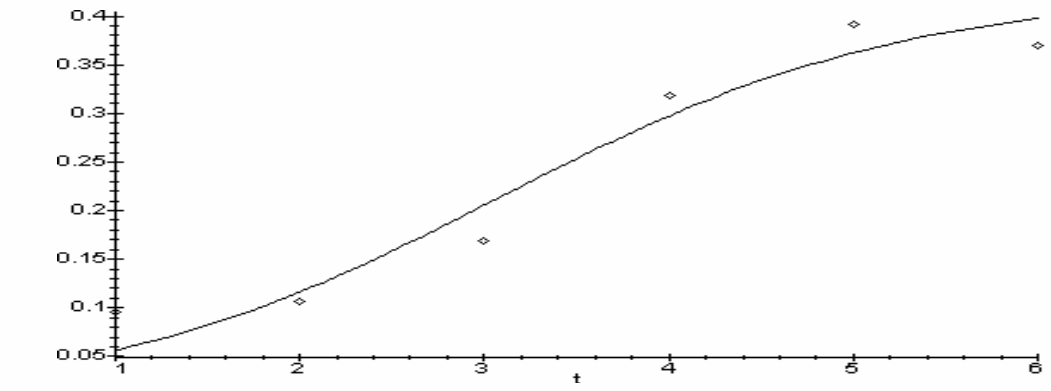

Abb. 6. Zunahme der Textfrequenz des bestimmten Artikels

Als reversibler Sprachwandel erweist sich die Verwendung des lat. Pronomens *hic* im Vergleich zu anderen Demonstrativpronomina (Piotrowski, Bektaev, & Piotrowskaja 1985: 87):

Tabelle 8
Textfrequenz des Pronomens *hic* (Anteil an Demonstrativpronomen)

t	Zeit	beobachtet	berechnet
1	Jahr 0	23.00	23.76
2	nach 100 Jahren	28.25	27.71
3	nach 200 Jahren	32.00	31.49
4	nach 300 Jahren	34.25	34.32
5	nach 400 Jahren	35.00	35.42
6	nach 500 Jahren	34.25	34.43
7	nach 600 Jahren	32.00	31.66
a = -0.9503	b = -0.0089	c = -0.0009	D = 0.99

Der Sprachwandel verläuft nach der Formel (vgl. Abb. 7)

$$p_t = \frac{1}{1 - 0.9503\, e^{0.0089t - 0.0009t^2}}.$$

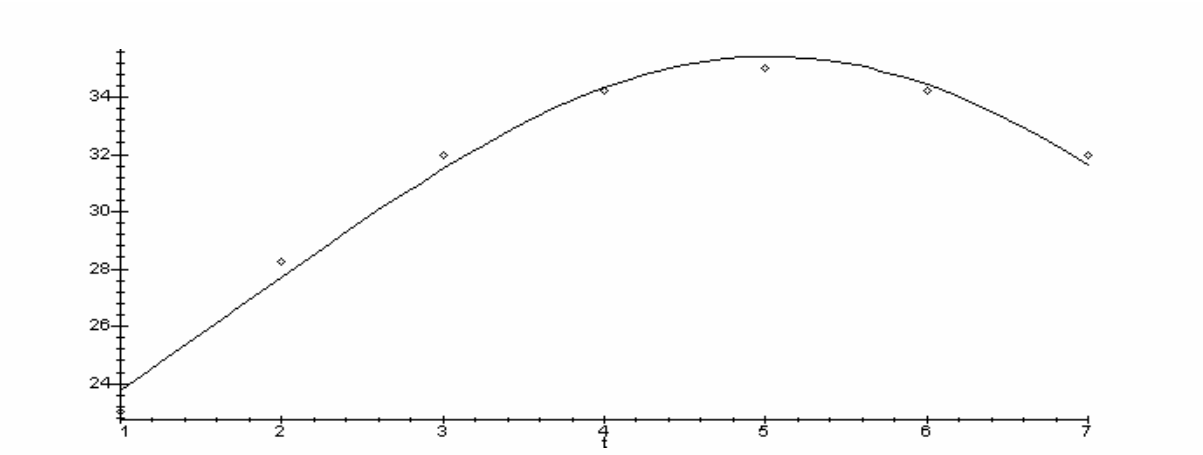

Abb. 7. Textfrequenz des Pronomens *hic*

Zur Entwicklung erweiterter Adjektiv- und Partizipialattribute im Deutschen wurden folgende Verhältnisse festgestellt (n. Weber 1971: 125):

Tabelle 9
Textfrequenz des erweiterten Adjektiv- und Partizipialattributs
(Belege je 10000 Druckzeichen)

t	Jahrhundert	beobachtet	berechnet
1	16.	1.84	1.8286
2	17.	9.4	9.2637
3	18.	9.85	11.2772
4	19.	13.0	11.3546
5	20.	9.75	9.9505
a = 428.0497	b = 5.0471	c = 0.7038	D = 0.93

Dieser Sprachwandel hat die Form (vgl. auch Abb. 8)

$$p_t = \frac{12}{1+428.0497\, e^{-5.0471t+0.7038t^2}}.$$

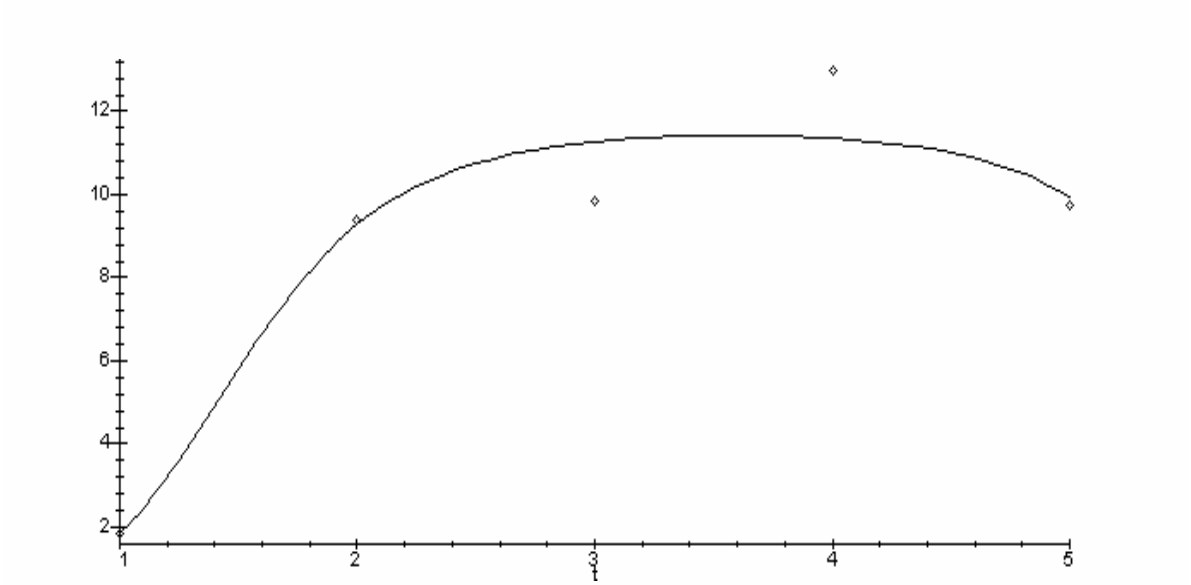

Abb. 8. Textfrequenz des erweiterten Adjektiv- und Partizipialattributs

Als Ergebnis dieses Abschnitts kann festgestellt werden, dass alle syntaktischen Wandel, für die hinreichende Daten zur Verfügung stehen, einen Verlauf gemäß dem Piotrowski-Gesetz in einer seiner Formen aufweisen.

Weitere Untersuchungen betreffen die langfristigen Kürzungstendenzen bei deutschen Sätzen und Teilsätzen. Alle untersuchten Fälle folgen ebenfalls dem Piotrowski-Gesetz in einer seiner Formen (Best 2002b).

3.5. Wortschatzwachstum einer Sprache

Altmann u.a. (1983: 111) stützen das Piotrowski-Gesetz neben anderen Argumenten mit dem Hinweis, dass es sich auch in außerlinguistischen Bereichen bewähre; einer davon sei die „theory of growth“. Auch in der Linguistik gibt es Bereiche, in denen Wachstumsprozesse beobachtet werden können. So nahm Tuldava (1998: 136ff.) an, dass die Zunahme des Wortschatzes der estnischen Literatursprache ebenfalls als Wachstumsprozess aufgefasst und entsprechend dem Piotrowski-Gesetz modelliert werden könne.

> „Die estnische Literatursprache trat im 16. Jahrhundert in Erscheinung und durchlief die Stadien des ‚Entstehens, Formierens und Stabilisierens‘. Diese Stadien spiegeln sich in der Zusammensetzung und im Wachstum der repräsentativen (für ihre Zeit vollständigsten und normativen) Wörterbücher wieder“ (Tuldava 1998: 137).

Anders als bei Tuldava sind hier bei den Berechnungen auch die Wörterbücher von 1660 und 1930 berücksichtigt worden; die Übereinstimmung zwischen dem theoretischen Modell und den Beobachtungen ist mit $D = 0.94$ dennoch sehr gut. Dieser Wachstumsprozess folgt der Formel (vgl. auch Tabelle 10 und Abb. 9):

$$p_t = \frac{155000}{1+127.2652\, e^{-0.1897\, t}}.$$

Verfährt man wie Tuldava und lässt die Wörterbücher der Jahre 1660 und 1930 unberücksichtigt, erhält man mit $D = 0.99$ ein noch deutlich besseres Ergebnis.

Tabelle 10
Das Anwachsen der estnischen Lexik

t	Jahr: tatsächlich	Jahr: festgesetzt	Anzahl der Wörter: beobachtet	Anzahl der Wörter: berechnet
1	1660		10000	1458.50
13	1780		14000	13129.32
17	1818	1820	21000	25580.35
22	1869	1870	50000	52373.51
24	1893	1890	60000	66216.20
28	1930		120000	95221.26
31	1960		105000	114362.36
33	1976	1980	115000	124683.67
$a = 127.2652$ $b = 0.1897$ $c = 155000$ $D = 0.94$				

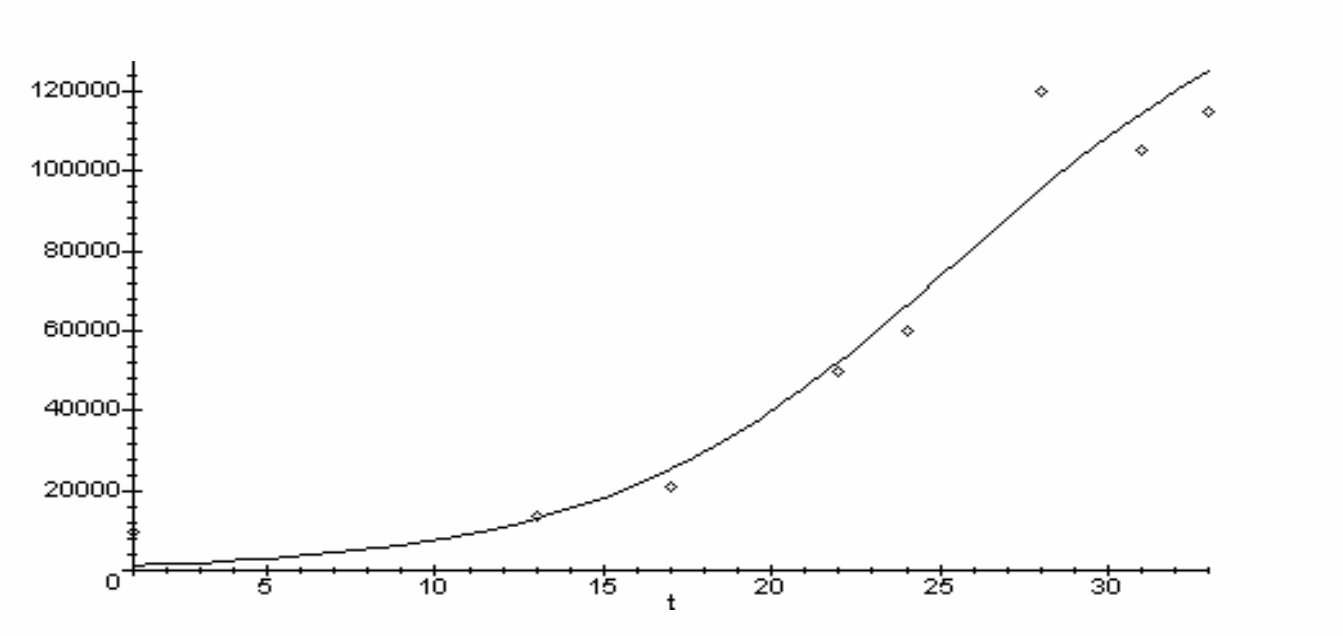

Abb. 9. Das Anwachsen der estnischen Lexik

In Best (2001c: 108-110) konnte am Beispiel von zwei Datensätzen zum Wachstum des englischen Wortschatzes gezeigt werden, dass auch diese sich dem Piotrowski-Gesetz entsprechend entwickeln.

3.6. Verlust von Teilen des Wortschatzes

Ebenso, wie der Aufbau eines Wortschatzes dem Piotrowski-Gesetz folgt, sollte auch der Abbau verlaufen. D.h., die Zunahme obsoleter Wörter einer Sprache wird sich auf die gleiche Weise entwickeln, so die Hypothese. Dies kann an englischen Daten geprüft werden. Man

findet eine entsprechende Zusammenstellung bei Dike (1935: 364): „I classified 16,018 obsoletisms: 1126 OE; 3005 ME; 8612 Early Modern; 3275 Later Modern (1660ff.).“ Unter *t* werden die Jahrhunderte, beginnend mit dem Jahr 1000 aufgelistet, das für die Endphase des Altenglischen angesetzt wurde. Unter Verwendung der Periodisierung des Englischen nach Viereck, Viereck & Ramisch (2002: 70) ergab sich folgende Tabelle:

Tabelle 11
Zum Wortschatzverlust des Englischen (n. Dike 1935: 364)

t	Zeitpunkt	Verlust	noch vorhanden beobachtet	noch vorhanden: berechnet
1	700		16018	16029.98
6	1200	1126	14892	15981.85
9	1500	3005	11887	11837.73
10.5	1650	8612	3275	3312.24
13.3	1930	3275	0	48.67
	a = 2.1804	*c* = 16030	*b* = -1.5889	*D* = 0.99

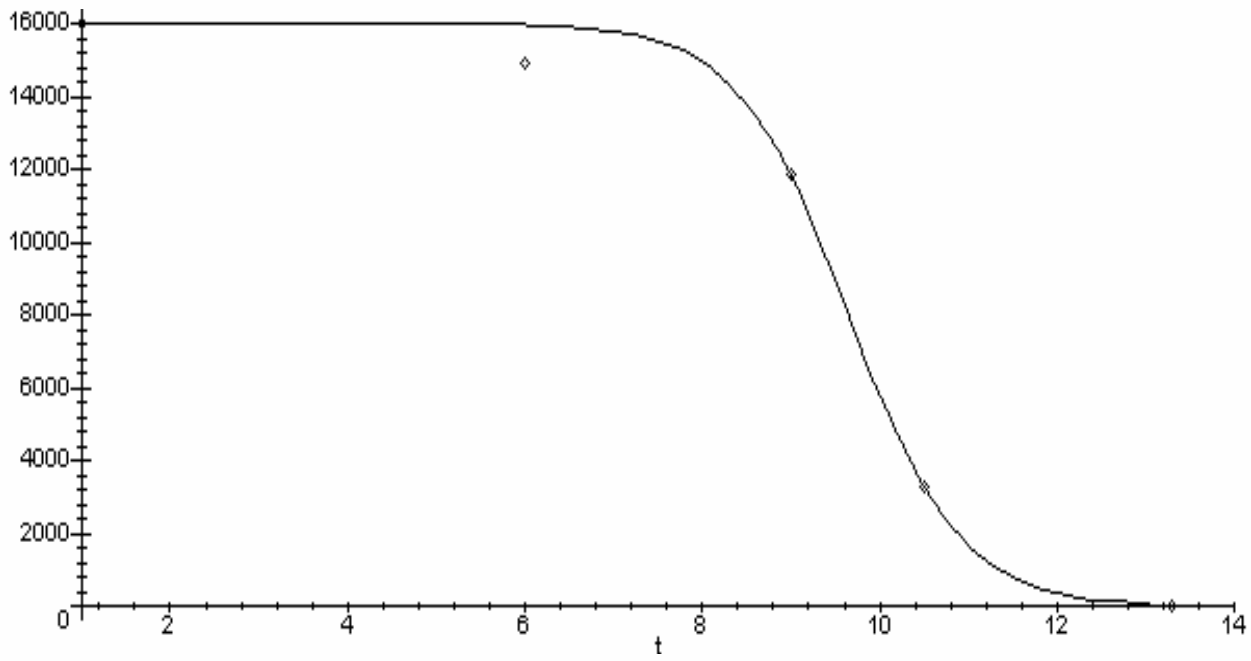

Abb. 10. Zum Wortschatzverlust des Englischen

Der Wortschatzverlust im Englischen verläuft nach der Formel (vgl. auch Abb. 10):

$$p_t = \frac{16030}{1+2.1804\,e^{1.5889\,t}}.$$

3.7. Entlehnungen

Dass Entlehnungen von Wörtern und Affixen gemäß dem Piotrowski-Gesetz verlaufen, konnte schon anhand einer ganzen Reihe solcher Prozesse nachgewiesen werden (Altmann u.a.

1983; Best 2001, 2001a, Best 2001c: 106f.; Best 2002a,b,c, 2003, 2003a,c; Müller-Hasemann 1983). Die spezifischen theoretischen Annahmen hierzu wurden in Beöthy & Altmann (1982) sowie in Best & Altmann (1986) entwickelt. Als weiteres Beispiel seien die Anglizismen in 3000 deutschen Werbetexten angeführt (Schütte 1996: 174):

Tabelle 12
Anteil der Anglizismen am Wortschatz in deutschen Werbetexten

t	Jahr	beobachtet	berechnet
1	1951	1.1	0.95
2	1961	2.1	1.55
3	1971	2.1	2.52
4	1981	3.8	4.07
5	1991	6.7	6.52
$a = 171.9970$ $b = 0.4969$ $c = 100$ $D = 0.97$			

Es wurde $c = 100$ gesetzt, da dies der äußerst unwahrscheinliche, aber denkbare Grenzwert ist. Die Übereinstimmung zwischen dem theoretischen Modell und den Beobachtungen ist mit $D = 0.97$ sehr gut. Dieser Entlehnungsprozess folgt der Formel (vgl. auch Abb. 11):

$$p_t = \frac{100}{1+171.9970\ e^{-0.4969\,t}}.$$

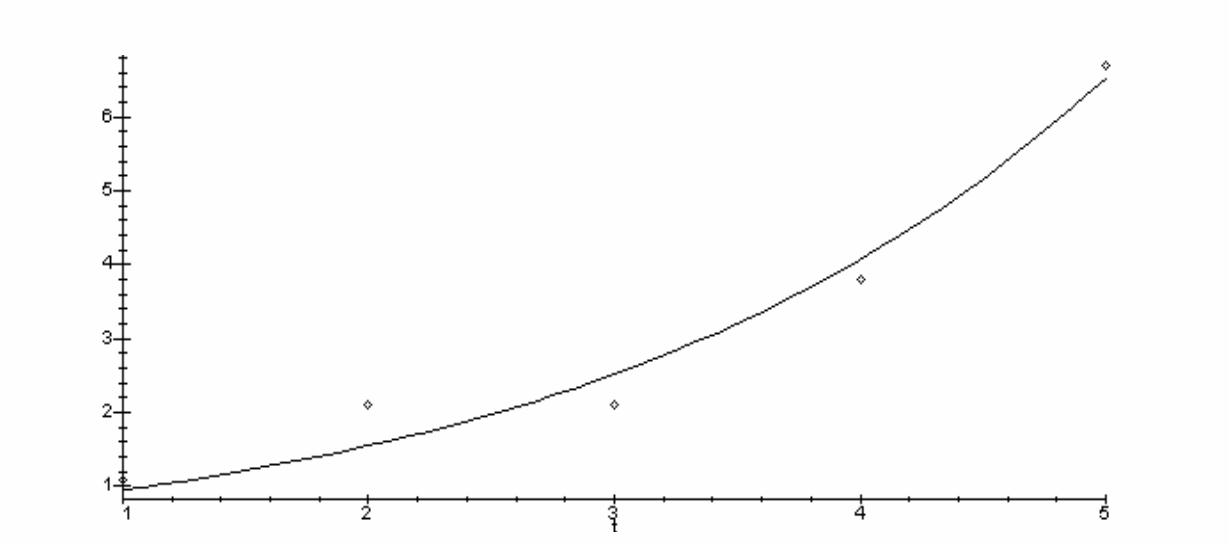

Abb. 11. Anteil der Anglizismen am Wortschatz in deutschen Werbetexten

Das Ergebnis stimmt für 1981 sehr gut mit dem Befund von Müller-Hasemann (1983: 158) überein, der am Beispiel des *Quelle-Katalogs* auf einen Anteil von rund 4% Anglizismen kam, während im *Spiegel* (ohne Anzeigentexte) lediglich 1.2% Anglizismen für 1979 festgestellt wurden.

3.8. Die Durchsetzung des etymologischen Prinzips in der Schreibung von Wörtern

In nichtbiblischen Texten des 16. - 18. Jahrhunderts kann bei der Schreibung von Wörtern des Typs „Geste“ vs. „Gäste“ die Durchsetzung des etymologischen Prinzips beobachtet werden.

Der Sprachwandel folgt der Formel (vgl. auch Abb.12):

$$p_t = \frac{100}{1+0.0154\,e^{0.06t}}.$$

Tabelle 13
Durchsetzung des etymologischen Prinzips in der Schreibung (n. Besch 1984: 126)

t	Alter	Abweichung in Prozent von der nhd. Norm: beobachtet	Abweichung in Prozent von der nhd. Norm: berechnet
1	1544	82	98.39
15	1558	85	96.34
28	1571	99	92.35
33	1576	85	89.95
47	1590	91	79.43
87	1630	24	25.95
95	1638	8	17.82
99	1642	5	14.57
105	1648	16	10.63
107	1650	10	9.55
117	1660	4	5.47
120	1663	11	4.61
122	1665	2	4.11
124	1667	10	3.67
127	1670	10	3.08
128	1671	19	2.91
135	1678	10	1.93
136	1679*	2.5	1.82
156	1699	6	0.55
178	1721	0.8	0.15
224	1767	0	0.01
a = 0.0154 *b* = -0.06 *c* = 100 *D* = 0.95			

*Hier handelt es sich um den Mittelwert für 2 Texte aus dem gleichen Jahr.

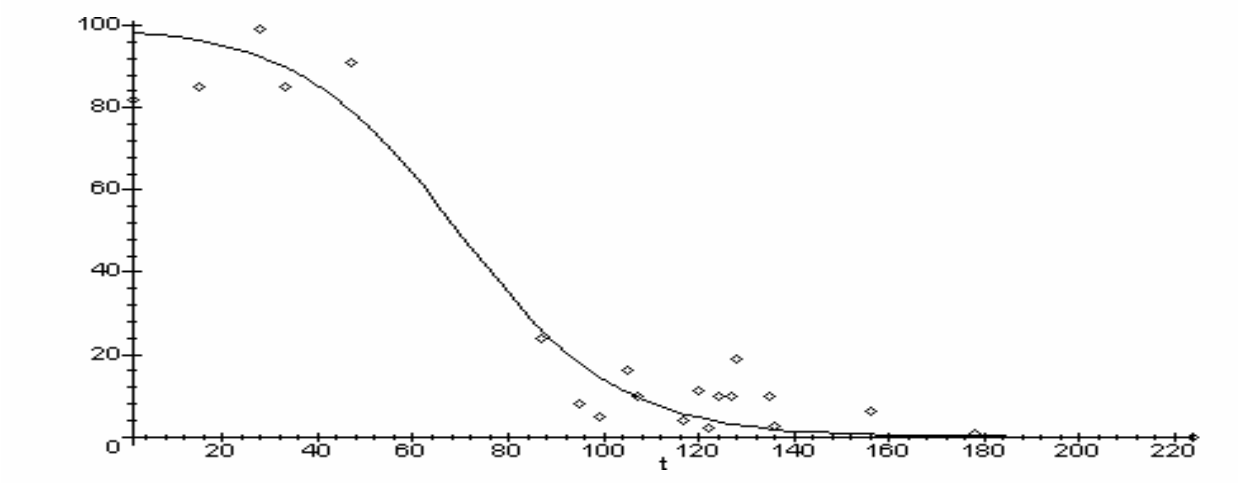

Abb. 12. Durchsetzung des etymologischen Prinzips in der Schreibung

3.9. Die Kennzeichnung der Vokallänge

Ein Prozess, der hinsichtlich seines Verlaufs besonders interessant erscheint, ist die Entwicklung der Kennzeichnung der Vokallänge in der Schreibung ostmitteldeutscher Luther-Bibeldrucke (Besch 1984: 131). Es handelt sich dabei um einen Sprachwandel, bei dem die Schreibung sich zunächst von der späteren Norm entfernt, um sich ihr danach fast bis zu ihrer Durchsetzung zu nähern.

Tabelle 14
Kennzeichnung der Vokallänge in der Schreibung (n. Besch 1984: 131)

t	Jahr	Abweichung in Prozent von der nhd. Norm: beobachtet	Abweichung in Prozent von der nhd. Norm: berechnet
1	1522	78.08	75.24
3	1545	56.90	61.75
5	1569	56.66	54.16
11	1626	76.03	75.95
18	1694	92.17	99.74
22	1736	95.72	99.99*
28	1797	96.71	100.00
$a = 0.2128$ $b = -0.4761$ $c = -0.04$ $D = 0.93$			

*abgerundet. *t* ist nach Jahrzehnten berechnet

Auch dieser Fall verläuft in Übereinstimmung mit dem Piotrowski-Gesetz in seiner Form für den reversiblen Sprachwandel nach der Formel (vgl. auch Abb. 13):

$$p_t = \frac{100}{1+0.2128e^{0.4761t-0.04t^2}}.$$

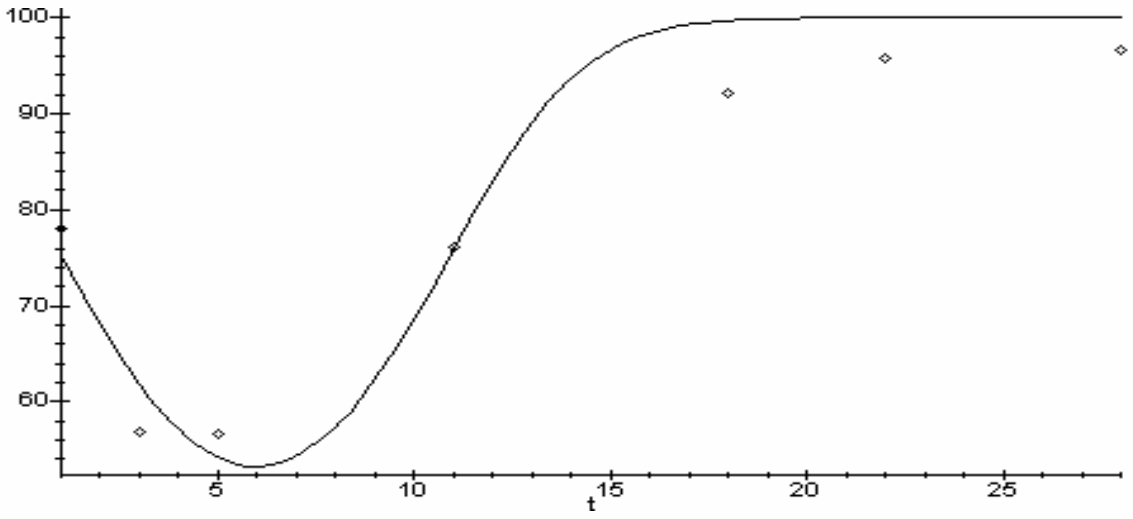

Abb. 13. Kennzeichnung der Vokallänge in der Schreibung

3.10. Wortschatzzuwachs im Spracherwerb

Ein Wachstumsprozess, der meist nicht im Zusammenhang mit Sprachwandel gesehen wird, ist das Erlernen des muttersprachlichen Wortschatzes durch Kinder. Der Gedanke, dass es sich hierbei um einen ganz normalen Wachstumsprozess handelt, der im Prinzip wie andere derartige Entwicklungen verläuft, liegt jedoch nicht allzu fern. Als Modell dazu, wie man sich den Wortschatzerwerb von Kindern vorstellen kann, führen Wagner, Altmann & Köhler (1987: 138) aus: „Am Anfang wächst der Wortschatz langsam, dann beschleunigt sich der Prozess in dem Maße wie soziale Kontakte und Kreativität anwachsen, und zum Schluß, im sprachlich ‚erwachsenen' Alter, verlangsamt es sich wieder, da man das gesamte Vokabular der Sprache langsam erschöpft." Es ist das schon hinlänglich bekannte Muster, das sich anhand der folgenden Daten wieder einmal nachweisen lässt:

Tabelle 15
Wortschatzerwerb englisch sprechender Kinder (Fries & Traver 1940: 49)

t	Alter	Zahl der Wörter beobachtet	Zahl der Wörter berechnet
1	2.0	215	625.67
8	2.7	642	746.36
42	5.5	1528	1704.34
54	6.5	2500	2240.13
66	7.5	2600	2904.58
78	8.5	3960	3704.03
90	9.5	5000	4631.66
102	10.5	6000	5663.81
114	11.5	6100	6760.01
126	12.5	7700	7868.12
134	13.0	8800	8586.38
$a = 21.9499$ $b = 0.0265$ $c = 14000$ $D = 0.99$			

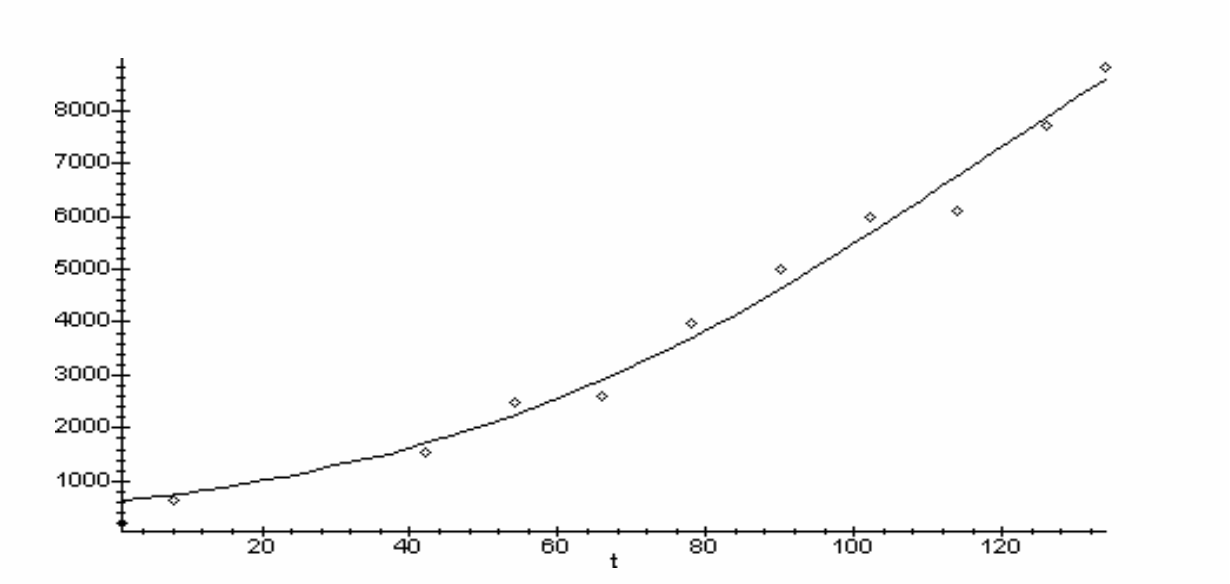

Abb. 14. Wortschatzerwerb englisch sprechender Kinder

Auch dieser Prozess folgt damit dem Piotrowski-Gesetz für den unvollständigen Sprachwandel, wie am Beispiel der Angaben zu englisch sprechenden Kindern (Fries & Traver 1940: 49; Tab. n. Tracy u.a.) demonstriert werden kann; es handelt sich in diesem Fall um geschätzte Werte (*t*: Monate, gezählt von 2;0 an.), vgl. Tabelle 15.

Dieser Wachstumsprozess entspricht der Formel (vgl. auch Abb. 14):

$$p_t = \frac{14000}{1 + 21.9499\ e^{-0.0265\ t}}.$$

Auch dieser Fall unterstützt die Hypothese, dass der Wortschatzzuwachs entsprechend dem Piotrowski-Gesetz verläuft. In diesen Daten steckt aber ein zusätzlicher Effekt, der darin besteht, dass die Beobachtungsdaten um die errechnete Kurve oszillieren, ein Effekt, den schon Köhler (1986: 137ff.) entdeckte, und der auch in anderen Zusammenhängen nachgewiesen werden kann (Best 2001a: 109).

Mit Morley (²1967) sei auf eine weitere Untersuchung zum Spracherwerb hingewiesen. Sie untersuchte an 114 englischsprachigen Kindern, wann bei ihnen die ersten Wörter auftraten, wann sie die ersten 2-3-Wort-Sätze bildeten und wann ihre Sprechweise für Fremde verstehbar wurde. Auch diese Prozesse verlaufen entsprechend dem Piotrowski-Gesetz (Best 2003a: 122-126).

Es gibt noch eine Reihe weiterer Auswertungen zum Wortschatzerwerb von Kindern, denen eine eigene Untersuchung gewidmet wird (Best 2002a). Alle diese und noch andere Fälle gehorchen dem Piotrowski-Gesetz.

3.11. Wortschatzwachstum in Texten

Zerlegt man einen Text in gleich lange Textblöcke und untersucht, wie der Wortschatz vom ersten bis zum letzten Textblock anwächst, so scheint dieser Prozess einem Modell zu folgen, das ebenfalls dem Piotrowski-Gesetz entspricht, wie Wimmer & Altmann (1999: 7) vorschlugen. Diese Hypothese wurde am Beispiel von Corneilles *Nicomède* überprüft; die entsprechenden Daten wurden Muller (1972: 225) entnommen:

Tabelle 16
Zuwachs neuer Wörter in Corneilles *Nicomède* (kumulierte Werte)

Textblock (je 103 Verse)	neue Wörter (beobachtet)	neue Wörter (berechnet)	Textblock (je 103 Verse)	neue Wörter (beobachtet)	neue Wörter (berechnet)	Textblock (je 103 Verse)	neue Wörter (beobachtet)	neue Wörter (berechnet)
1	315	431.31	7	994	977.75	13	1334	1358.75
2	504	513.17	8	1056	1062.81	14	1369	1393.05
3	620	602.03	9	1113	1139.94	15	1413	1420.94
4	750	695.74	10	1175	1208.13	16	1464	1443.42
5	853	791.56	11	1229	1267.08	17	1498	1461.40
6	927	886.52	12	1291	1317.07	18	1528	1475.69
$a = 3.2693$			$b = 0.2514$			$c = 1528$	$D = 0.99$	

Dieser Prozess verläuft damit gemäß der Formel (vgl. auch Abb. 15):

$$p_t = \frac{1528}{1+3.2693\, e^{-0.2514\, t}}.$$

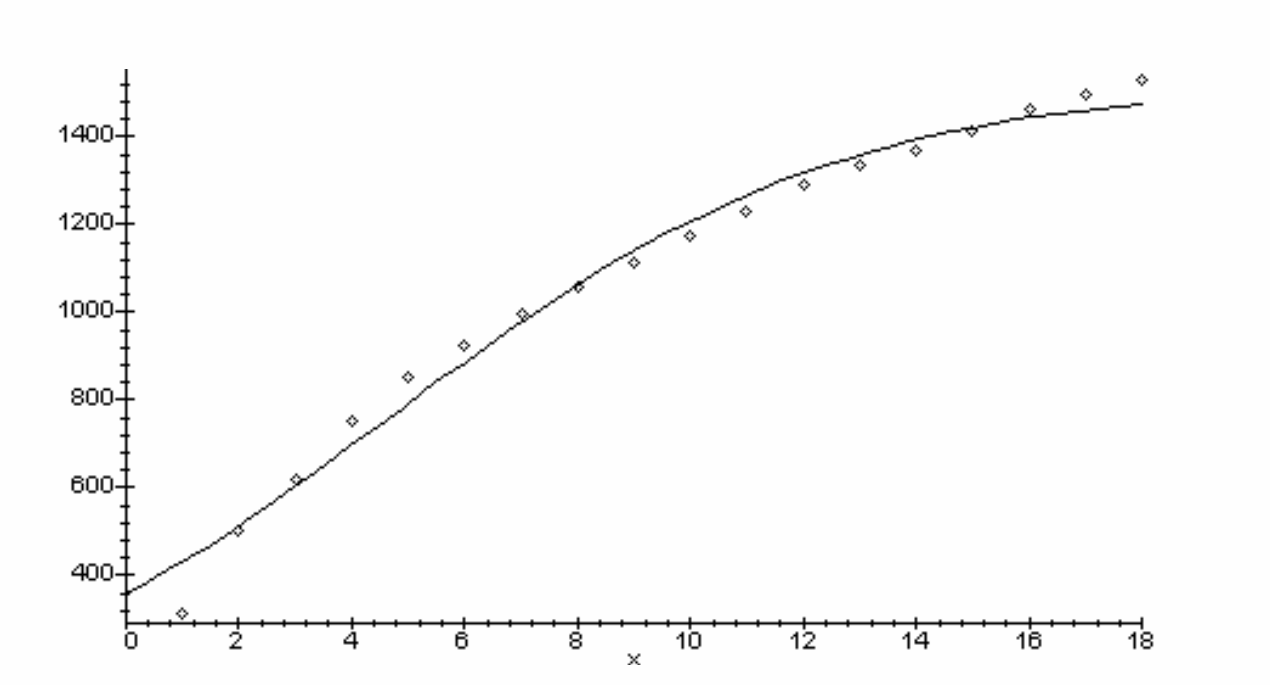

Abb. 15. Zuwachs neuer Wörter in Corneilles *Nicomède*

Tabelle 17
Zuwachs neuer Wörter in den historisch geordneten Stücken Corneilles (kumulierte Werte)

Nr.	Stück*	neue Wörter (beobachtet)	neue Wörter (berechnet)	Nr.	Stück	neue Wörter (beobachtet)	neue Wörter (berechnet)
1	Mélite	1766	2160.38	17	Théodore	4237	4224.75
2	Clitandre	2270	2332.30	18	Héraclius	4264	4283.84
3	La Veuve	2607	2504.41	19	Andromède	4286	4336.22
4	La Galérie...	2802	2674.86	20	Don Sanche	4311	4382.49
5	La Suivante	2936	2841.91	21	Nicomède	4356	4423.26
6	Place Royale	3018	3003.91	22	Pertharite	4370	4459.07
7	Médée	3227	3159.44	23	Oedipe	4398	4490.47
8	L'illusion C.	3449	3307.33	24	La Toison ...	4438	4517.95
9	Le Cid	3532	3446.65	25	Sertorius	4472	4541.95
10	Horace	3600	3576.76	26	Sophonisbe	4491	4562.88
11	Cinna	3687	3697.30	27	Othon	4520	4581.11
12	Polyeucte	3770	3808.13	28	Agésilas	4542	4596.97
13	Pompée	3850	3909.32	29	Attila	4567	4610.76
14	Le Menteur	4016	4001.15	30	Tite et B.	4579	4622.74
15	Suite du M.	4157	4083.99	31	Pulchérie	4596	4633.13
16	Rodogune	4202	4158.33	32	Suréna	4606	4642.14
		$a = 1.3612$	$b = 0.1467$		$c = 4700$	$D = 0.98$	

* Die Titel konnten nicht alle vollständig genannt werden.

Was sich im Fall von *Nicomède* am Beispiel eines einzelnen Textes beobachten lässt, gilt auch, wenn man das Wortschatzwachstum in den 32 Stücken Corneilles betrachtet, deren Werte Muller (1967: 82) historisch geordnet anführt. Kumuliert man die Anzahl der Wörter, die von Stück zu Stück neu auftreten, so erhält man Tabelle 17.

Das Wortschatzwachstum bei Corneille entspricht mit sehr gutem $D = 0.98$ der Formel

$$p_t = \frac{4700}{1+1.3612\, e^{-0.1467\, t}}.$$

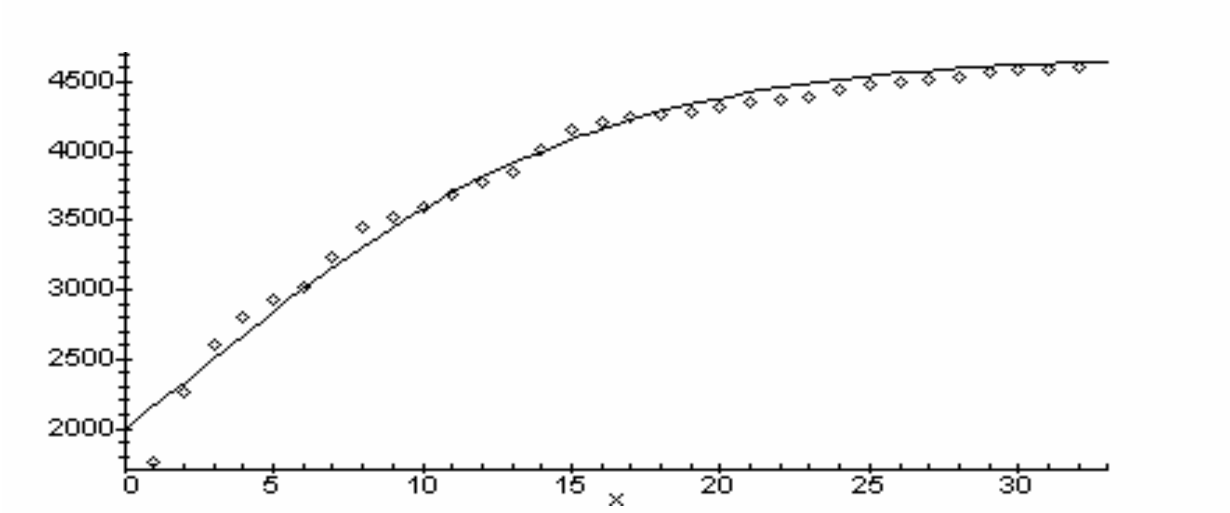

Abb. 16. Zuwachs neuer Wörter in den Stücken Corneilles (kumulierte Werte)

Ein entsprechendes Ergebnis lässt sich auch für die deutsche Zeitungssprache erzielen, wie ein Test anhand von Billmeiers Erhebung (1968: 167) zu *Neues Deutschland* (18 Seiten verschiedener Sparten, Februar 1964) zeigt:

Tabelle 18
Zuwachs neuer Wörter in *Neues Deutschland* (kumulierte Werte)

Textblock (in Wörtern)	neue Wörter (beobachtet)	neue Wörter (berechnet)	Textblock (in Wörtern)	neue Wörter (beobachtet)	neue Wörter (berechnet)	Textblock (in Wörtern)	neue Wörter (beobachtet)	neue Wörter (berechnet)
3000	1295	1840.91	24000	5900	6036.05	45000	10460	10659.34
6000	2296	2248.10	27000	6799	6805.28	48000	10981	11080.31
9000	3099	2723.38	30000	7472	7566.04	51000	11532	11436.66
12000	3633	3268.60	33000	8022	8297.93	54000	11945	11734.43
15000	4280	3881.74	36000	8683	8983.57	57000	12331	11980.53
18000	4982	4556.04	39000	9230	9610.07	60000	12832	12182.12
21000	5329	5279.66	42000	9850	10169.64			
$a = 7.6829$			$b = 0.0001$		$c = 13000$	$D = 0.99$		

Dieser Prozess verläuft damit gemäß der Formel (vgl. auch Abb. 17):

$$p_t = \frac{13000}{1 + 7.6829\, e^{-0.0001\, t}}.$$

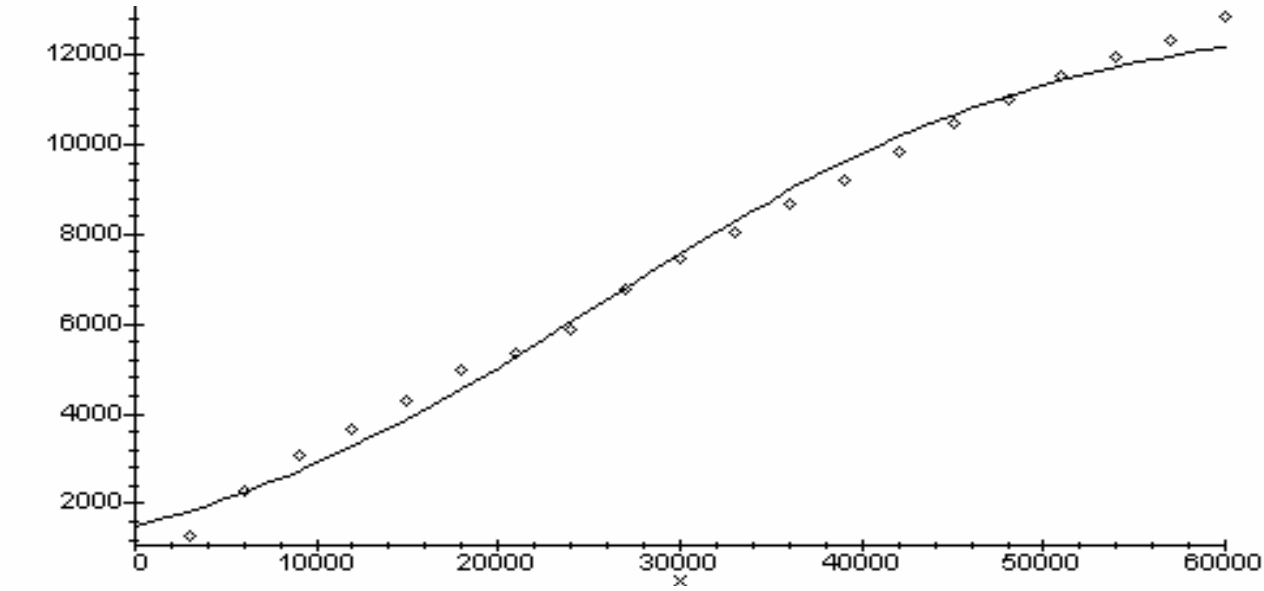

Abb. 17. Zuwachs neuer Wörter in *Neues Deutschland*

4. Zusammenfassung und Perspektiven

Es war die Aufgabe dieser Ausführungen, zu zeigen, dass Sprachwandelprozesse sehr unterschiedlicher Art und auch Wachstumsprozesse im Spracherwerb sowie in einzelnen Texten und Textgruppen nicht chaotisch verlaufen, sondern theoretisch begründbaren und empirisch überprüfbaren Gesetzen folgen. Dabei ist zu beachten, dass außer S-förmigen Sprachwandelprozessen, die in der einschlägigen Literatur anerkannt werden, auch reversible Entwicklungen vorkommen, für die ein mathematisches Modell vorgeschlagen und inzwischen mehrfach erfolgreich getestet wurde. Ein recht breites Spektrum von Phänomenen fügt sich diesen Gesetzeshypothesen offensichtlich. Es sei darauf verwiesen, dass es noch deutlich mehr derartige Entwicklungen gibt, für die dasselbe gilt; einige davon werden in absehbarer Zukunft vorgestellt. Für den jetzigen Zeitpunkt kann festgestellt werden, dass bisher für jeden Prozess, für den hinreichend dicht gestreute Daten vorliegen, die Übereinstimmung mit dem Piotrowski-Gesetz nachgewiesen werden kann. Wenn, wie im Fall der *-bar-*Adjektive, ein einzelner Wert dem generellen Trend nicht zu folgen scheint, so darf vorerst angenommen werden, dass bei einer angemessenen Verdichtung der Messpunkte gezeigt werden kann, dass kurzfristige Entwicklungen als zeitlich begrenzte Sonderentwicklungen ebenfalls dem Piotrowski-Gesetz folgen.

Damit soll nicht behauptet werden, dass nicht auch einmal Entwicklungen beobachtet werden können, die nicht den typischen S-förmigen oder reversiblen Verlauf zeigen. Das Piotrowski-Gesetz ist eine Gesetzeshypothese für den einigermaßen natürlich verlaufenden Sprachwandel, bei dem nicht durch bewusste Entscheidung vieler Mitglieder der Sprachgemeinschaft oder gar durch massive Eingriffe abrupte Veränderungen herbeigeführt werden. Eine Widerlegung des vorgeschlagenen Sprachgesetzes stellen solche Erscheinungen nicht dar. Gesetze können durch Einwirkung von außen oder durch konkurrierende Gesetze in ihren Wirkungen eingeschränkt werden.

Eine andere Frage stellt sich, wenn man einige der graphisch dargebotenen Sprachwandel betrachtet: Die Berechnungen zu Tab. 6 ergaben, dass die Textfrequenz des bestimmten Artikels sehr gut mit dem Piotrowski-Gesetz modelliert werden kann; in der Graphik dazu zeigt sich aber ein zusätzlicher Effekt in diesen Daten, der darin besteht, dass die Beobachtungswerte sich in einer bestimmten Phase eindeutig unterhalb, in einer anderen Phase ebenso eindeutig oberhalb des errechneten Verlaufs bewegen. Köhler (1986: 137ff.) hat diesen Trend in der deutschen Lexik entdeckt und als „Oszillation“ bezeichnet. Ein besonders eindrucksvoller Fall dieser Art wurde beim Wachstum des englischen Wortschatzes beobachtet (Best 2001: 109). Bisher ist diese Oszillation nicht so gravierend, dass sie die Möglichkeit, einen Sprachwandel als dem Piotrowski-Gesetz folgend zu erkennen, infrage stellt. Möglicherweise treten jedoch einmal Fälle auf, die dazu zwingen, diesem Effekt noch mehr Aufmerksamkeit zu widmen.

Abschließend darf festgestellt werden, dass das Piotrowski-Gesetz − eine logistische Funktion − sich bei sprachlichen Wandel- und Wachstumsprozessen sehr gut bewährt hat. Diese Auffassung wird dadurch zusätzlich bestätigt, dass gleichartige Prozesse von Wachstum und Ausbreitung innerhalb ebenso wie außerhalb der Linguistik zu beobachten sind, z.B. bei der Modellierung der Entwicklung der Wissenschaft selbst (Tuldava 1998: 140). Sherrod verwendet in der von ihm entwickelten Software *NLREG* (1991-2001) eine ganz ähnliche Funktion, um die Ausbreitung von Aids in den USA zu modellieren (vgl. auch Leopold 1998: 100). Der Erratbarkeit gekürzter Wörter scheint wiederum eine ähnliche Funktion zugrunde zu liegen (Rumpel, Goldenberg & Boucsein 1984: 23). Alle Untersuchungen unterstützen Tuldavas Auffassung:

> „Das Gesetz des logistischen Wachstums in seiner allgemeinen Form (Beschleunigung − Wendepunkt − Verlangsamung) hat mit großer Wahrscheinlichkeit eine allgemeine sozial-linguistische Relevanz“ und gehört „zu den grundlegenden Gesetzen der Entwicklung von selbstorganisierenden Systemen“ (Tuldava 1998: 140).

Nachtrag Februar 2003: Anlässlich des Artikels “Ansteckung durch Worte. Das Piotrowski-Gesetz: Die Kurve sprachlicher Neuerungen” von Wolfgang Krischke (FAZ 22.1.03, S. N3, “Geisteswissenschaften”) weist Rainer Schimming, Greifswald, in einer E-Mail vom 6.2.03 darauf hin, dass das Gesetz des logistischen Wachstums bereits im 19. Jhd. von Pierre François Verhulst vorgeschlagen und von R. Pearl im 20. Jhd. wiederbelebt wurde; es sei daher auch als Verhulst-Pearl-Modell bekannt und auf “unübersehbar viele Fälle” mit Erfolg angewendet worden. Die dargestellten linguistischen Prozesse finden also in reichem Maße Parallelen in anderen Wissenschaften.

Literatur

Aitchison, J. (1991). *Language change: progress or decay?* Second edition. Cambridge: Cambridge University Press.

Altmann, G. (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, K.-H., & Kohlhase, J. (Hrsg.), *Exakte Sprachwandelforschung: 54-90.* Göttingen: edition herodot.

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

Bailey, Ch.-J. N. (1973). *Variation and Linguistic Theory*. Arlington, Virginia: Center for Applied Linguistics.

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

Besch, W. (1984). Sprachliche Änderungen in Lutherbibel-Drucken des 16.-18. Jahrhunderts. In: Joachim Schildt (Hrsg.), *Luthers Sprachschaffen. Gesellschaftliche Grundlagen - Geschichtliche Wirkungen: 108-133*. Berlin: Akademie der Wissenschaften der DDR, Zentralinstitut für Sprachwissenschaft. Linguistische Studien, Reihe A, 119/1.

Best, K.-H. (1983). Zum morphologischen Wandel einiger deutscher Verben. In: Best, K.-H., & Kohlhase, J. (Hrsg.), *Exakte Sprachwandelforschung: 107-118*. Göttingen: edition herodot.

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

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

Best, K.-H. (2001b). Probability distributions of language entities. *Journal of Quantitative Linguistics 8, 1-11.*

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

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

Best, K.-H. (2002a). Zur Entwicklung von Wortschatz und Redefähigkeit bei Kindern. Mskr.

Best, K.-H. (2002b). Satzlängen im Deutschen: Verteilungen, Mittelwerte, Sprachwandel. *Göttinger Beiträge zur Sprachwissenschaft 7, 7-31.*

Best, K.-H. (2002c). Der Zuwachs der Wörter auf *-ical* im Deutschen. *Glottometrics 2, 11-16.*

Best, K.-H. (2003). Anglizismen − quantitativ. *Göttinger Beiträge zur Sprachwissenschaft 8* (erscheint).

Best, K.-H. (2003a). Slawische Entlehnungen im Deutschen. In: Berger, T., Kempgen, S., & Schweier, U. (Hrsg.), *Festschrift für Werner Lehfeldt zum 60. Geburtstag*. München: Verlag Otto Sagner. (Im Druck).

Best, K.-H. (2003b). *Quantitative Linguistik. Eine Annäherung*. 2., überarb. Aufl. Göttingen: Peust & Gutschmidt (im Druck).

Best, K.-H. (2003c). Wie verläuft Sprachwandel? *Naukovyj Visnyk Černivec'koho Universitetu. Serija „Hermans'ka Filolohija". Vypusk 155, 86-94.*

Best, K.-H. (2003d). Zum Wandel von Idiolekten. *Naukovyj Visnyk Černivec'koho Universitetu. Serija „Hermans'ka Filolohija". Vypusk 156* (im Druck).

Best, K.-H., & Altmann, G. (1986). Untersuchungen zur Gesetzmäßigkeit von Entlehnungsprozessen im Deutschen. *Folia Linguistica Historica 7, 31-41*.

Best, K.-H., Beöthy, E., & Altmann, G. (1990). Ein methodischer Beitrag zum Piotrowski-Gesetz. In: Hammerl, R. (Hrsg.), *Glottometrika 12, 115-124*. Bochum: Brockmeyer.

Best, K.-H., & Kohlhase, J. (1983). Der Wandel von *ward* zu *wurde*. In: Best, K.-H., & Kohlhase, J. (Hrsg.), *Exakte Sprachwandelforschung: 91-102*. Göttingen: edition herodot.

Best, K.-H., & Kohlhase, J. (Hrsg.) (1983). *Exakte Sprachwandelforschung*. Göttingen: edition herodot.

Billmeier, G. (1968). Über die Signifikanz von Auswahltexten. Untersuchung auf der Grundlage von Zeitungstexten. In: Moser, Hugo u.a. (Hrsg.), *Forschungsberichte des Instituts für deutsche Sprache 2, 126-171.*

Chen, M. Y., & Wang, W. S.-Y. (1975). Sound Change: Actuation and Implementation. *Language 51*: 255-281.

Crystal, D. (1993). *Die Cambridge Enzyklopädie der Sprache*. Frankfurt-New York: Campus.

Dike, E. B. (1935). Obsolete English Words: Some Recent Views. *The Journal of English and Germanic Philology 34, No. 3, July 1935, 351-365.*

Faust, M. (1980). Morphologische Regularisierung in Sprachwandel und Spracherwerb. *Folia Linguistica 14, 387-411*.

Fenk-Oczlon, G. (1991). Frequenz und Kognition − Frequenz und Markiertheit. *Folia Linguistica 25, 361-394*.

Flury, R. (1964). *Struktur- und Bedeutungsgeschichte des Adjektivsuffixes -bar.* Winterthur: Verlag P.G. Keller. (diss. phil., Zürich)

Fries, Ch. C., & Traver, A. A. (1940). *English Word List. A Study of their Adaptability for Instruction*. Washington, D.C.: American Council on Education.

Imsiepen, U. (1983). Die e-Epithese bei starken Verben im Deutschen. In: Best, K.-H. & Kohlhase, J. (Hrsg.), *Exakte Sprachwandelforschung: 119-141.* Göttingen: edition herodot.

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

Kohlhase, J. (1983). Die Entwicklung von *ward* zu *wurde* beim Nürnberger Chronisten Heinrich Deichsler. In: Best, K.-H., & Kohlhase, J. (Hrsg.), *Exakte Sprachwandelforschung 103-106.* Göttingen: edition herodot.

Körner, H. (2002). Der Zuwachs der Wörter auf *-ion* im Deutschen. *Glottometrics 2*, 82-86.

Labov, W. (1994). *Principles of Linguistic Change.* Oxford, UK, & Cambridge, USA: Blackwell.

Leopold, E. (1998). *Stochastische Modellierung lexikalischer Evolutionsprozesse.* Hamburg: Kovač.

Mańczak, W. (1980). Frequenz und Sprachwandel. In: Lüdtke, H. (Hrsg.), *Kommunikationstheoretische Grundlagen des Sprachwandels: 37-79.* Berlin-New York: de Gruyter.

Mayerthaler, W. (1981). *Morphologische Natürlichkeit.* Wiesbaden: Athenaion.

Morley, Muriel E. (²1967). *The Development and Disorders of Speech in Childhood.* Edinburgh-London: E. & S. Livingstone Ltd.

Muller, Ch. (1967). *Étude de statistique lexicale. Le vocabulaire du théatre de Pierre Corneille.* Paris: Librairie Larousse.

Muller, Ch. (1972). *Einführung in die Sprachstatistik.* München: Hueber.

Müller-Hasemann, W. (1983). Das Eindringen englischer Wörter ins Deutsche ab 1945. In: Best, K.-H., & Kohlhase, J. (Hrsg.), *Exakte Sprachwandelforschung: 143-160.* Göttingen: edition herodot.

Nübling, D. (2000). *Prinzipien der Irregularisierung.* Tübingen: Niemeyer.

Osgood, C.E., & Sebeok, T.A. (eds.) (1954/ 1965). *Psycholinguistics.* Bloomington/ London: Indiana University Press.

Piotrovskaja, A.A., & Piotrovskij, R.G. (1974). Matematičeskie modeli diachronii i tekstoobrazovanija. In: *Statistika reči i avtomatičeskij analiz teksta: 361-400.* Leningrad: Nauka.

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

Rumpel, D., Goldenberg, D., & Boucsein, W. (1984). Die Erkennbarkeit von abgekürzten Wörtern − experimentelle Untersuchungen und mathematische Modelle. In: Rothe, U. (Hrsg.), *Glottometrika 7, 15-44.* Bochum: Brockmeyer.

Schütte, D. (1996). *Das schöne Fremde. Anglo-amerikanische Einflüsse auf die Sprache der deutschen Zeitschriftenwerbung.* Opladen: Westdeutscher Verlag.

Sherman, D. (1975). Noun-Verb Stress Alternation: An Example of the Lexical Diffusion of Sound Change in English. *Linguistics 159, 43-71.*

Tuldava, J. (1998). *Probleme und Methoden der quantitativ-systemischen Lexikologie.* Trier: Wissenschaftlicher Verlag Trier (russ. 1987).

Viereck, W., Viereck, K., & Ramisch, H. (2002). *dtv-Altas Englische Sprache.* München: dtv.

Wagner, K. R., Altmann, G., & Köhler, R. (1987). Zum Gesamtwortschatz der Kinder. In: Wagner, K. R. (Hrsg.), *Wortschatz-Erwerb*: *128-142.* Bern, Frankfurt, New York, Paris: Peter Lang.

Weber, H. (1971). *Das erweiterte Adjektiv- und Partizipialattribut im Deutschen.* München: Hueber.

Wimmer, G., & Altmann, G. (1999). Review Article: On Vocabulary Richness. *Journal of Quantitative Linguistcs 6, 1-9.*

Software

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

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

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

Adresse des „Göttinger Projekts“ im Internet (mit ausführlicher, ständig aktualisierter Bibliographie): http://www.gwdg.de/~kbest/projekt.htm.

Glottometrics 5, 2003, 35-39

Word-Length Distribution in Modern Welsh Prose Texts

Andrew Wilson[1]
Lancaster University

Abstract. This paper examines the distribution of word lengths in 12 prose texts written in modern Welsh (a P-Celtic language). The texts belong to the genres of new articles and Bible translation. For all texts, the observed frequencies can best be fitted by the 1-displaced Singh-Poisson distribution. This differs from published results on a Q-Celtic language (Scottish Gaelic) and suggests a P-celtic/Q-Celtic difference in word-length distribution. Further work is required to investigate other genres of Welsh as well as the other P- and Q-celtic languages.

Keywords: word length, Welsh, Celtic, Singh-Poisson distribution

Introduction

The Celtic language family is made up of two distinct sub-groups: P-Celtic (or Brythonic) and Q-Celtic (or Goidelic). The P-Celtic group consists of Welsh, Cornish and Breton, whilst the Q-Celtic group consists of Irish, Scottish Gaelic and Manx[2].

Although very little Celtic data has yet been examined within the Göttingen project on word-length distributions[3], one set of Q-Celtic data has already been processed – a set of 31 Scottish Gaelic e-mails, for which the best-fit distribution was the 1-displaced hyperpoisson distribution (Drechsler 2001). This study will add data from a P-Celtic language – Welsh – in order to obtain a preliminary impression of whether Celtic is likely to show the same distribution for all its member languages, or whether there are likely to be differences, perhaps along the Q-Celtic versus P-Celtic dimension.

Data

The data for this study consisted of twelve Welsh prose texts. All were written within the past twenty years and most of them within the past one to two years. They were selected from two main genres: Bible texts and news reports. Within the category of Bible texts, two psalms and two short epistles from the *Beibl Cymraeg Newydd* (1985) were processed; these translations date from the late 1970s. Within the category of news reports, four texts from the Welsh weekly newspaper *Y Cymro* were processed (two general news items and two sports items) as well as

[1] Address correspondence to: Andrew Wilson, Dept. of Linguistics and Modern English Language, Lancaster University, Lancaster LA1 4YT, GB. E-mail: eiaaw@exchange.lancs.ac.uk
[2] For more details on the Celtic languages, see, e.g., Ball (1992), Macaulay (1992).
[3] On the Göttingen project, see, e.g., Best (1999; 2001).

four texts from the Welsh-language version of the University of Wales Bangor's in-house newsletter, *Newyddlen*.

The texts used were as follows:

Beibl Cymraeg Newydd:

Text B1	2 John
Text B2	3 John
Text B3	Psalm 20
Text B4	Psalm 21

Y Cymro:

Text C1	Tachwedd 24, 1999:	Stamp Cristnogaeth ar y Mileniwm
Text C2	Tachwedd 24, 1999:	Diffyd deddf addysg i Gymru yn araith fawr y Frenhines
Text C3	Tachwedd 24, 1999:	Inter yn brawf i Llanelli (sport)
Text C4	Rhafgyr 1, 1999:	Cadw safonau y clybiau (sport)

Newyddlen:

Text N1	Hydref 1999:	Cyfnod Newydd o Hanes ar Safle'r George
Text N2	Mehefin 2000:	Gradd i Paxman y Dyfodol
Text N3	Mehefin 2000:	Glinfyrddau i Nyrsys
Text N4	Tachwedd 2000:	Argyfwng – Pa Argyfwng?

Method

For each text analysed, the number of words falling into each word-length class was counted.

Using the standard rules of pronunciation for spoken Welsh (Rowland 1857, Williams 1980), the word lengths were determined in accordance with the usual principles of the Göttingen project, i.e., in terms of the number of spoken syllables per orthographic word.

In line with the general guidelines of the project, abbreviations were treated as instances of the full word for which the abbreviation stands – so, for example, *Dr* is counted as a two-syllable word (*doctor*). Acronyms, in contrast, are treated as single words and the number of syllables counted as pronounced; thus, for instance, *PCB* is counted as one word with three syllables.

Like abbreviations, numerals are treated as instances of the fully spelled out forms. Thus, *111* is treated as three words of one, two and one syllables respectively (*cant undeg un*). It should be noted here that a special feature of the Welsh language is that it has two counting systems: a decimal system and a vigesimal system (King 1993: 111-114). This means that a number such as *21* can be spoken as either *dauddeg un* (decimal) or *un ar hugain* (vigesimal). For the purposes of the present work, the decimal system was used throughout.

Clitics (e.g. the *i* in *o'i*) are pronounced as an integral part of the preceding word and were treated as such here.

The word-length frequency statistics for each text were then run through the Altmann-Fitter

software[4] at Göttingen to determine which probability distribution was the most appropriate model.[5]

Statistics

The Altmann Fitter compares the empirical frequencies obtained in the data analysis with the theoretical frequencies generated by the various probability distributions (Wimmer, Altmann 1996, 1999). The degree of difference between the two sets of frequencies is measured by the chi-squared test and also by the discrepancy coefficient *C*; the latter is given by X^2/N and is used especially where d.f. = 0. A probability distribution is considered an appropriate model for the data if the difference between the empirical and theoretical frequencies is not significant, i.e., if $P(X^2) > 0.05$ and/or $C < 0.02$. The best distribution is that which shows the highest *P* and/or lowest *C*.

Results

The best results were achieved by fitting the 1-displaced Singh-Poisson distribution, which is given by:

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

Although two other distributions – the positive Singh-Poisson distribution and the 1-displaced hyperpoisson distribution – could also be fitted, these showed poorer goodness of fit.

The individual results for the 1-displaced Singh-Poisson distribution are shown in the tables below, where:

x = number of syllables in the word
f_x = frequency of words with x syllables in the text
NP_x = expected frequency calculated by the relevant probability formula
X^2 = chi-square value
d.f. = number of degrees of freedom
P = probability of chi-square value
C = the coefficient X^2/N
α, a = parameters in the above equation.

In the case of text N4, it was necessary to merge two length classes in order to obtain a satisfactory fit.

[4] RST Rechner- und Softwaretechnik GmbH, Essen.
[5] I am grateful to Karl-Heinz Best for running the data through the Altmann Fitter.

Table 1
Fitting the 1-displaced Singh-Poisson distribution to word lengths in Welsh texts

	Text B1		Text B2		Text B3		Text B4	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	202	201.60	201	201.06	77	78.03	114	112.64
2	69	71.15	84	84.34	48	46.72	45	48.68
3	29	25.12	27	26.20	15	13.98	29	22.66
4	5	7.13	6	6.40	2	3.26	5	7.03
5							0	1.99
a	0.7061		0.6214		0.5987		0.93083	
α	0.6694		0.7946		0.9999		0.6874	
X^2	1.298		0.049		0.6123		4.6516	
DF	1		1		1		2	
P	0.25		0.82		0.43		0.10	

	Text C1		Text C2		Text C3		Text C4	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	106	105.42	124	123.21	183	182.93	179	178.69
2	53	49.53	81	85.20	100	100.09	99	101.11
3	20	25.00	43	35.38	30	29.97	35	30.83
4	8	8.41	7	9.79	7	7.01	5	7.37
5	4	2.64	1	2.42				
a	1.0095		0.8304		0.5989		0.6099	
α	0.7050		0.9195		0.9506		0.9594	
X^2	1.97		3.49		0.00		1.36	
DF	2		2		1		1	
P	0.37		0.17		0.99		0.2434	

	Text N1		Text N2		Text N3		Text N4	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	125	125.08	84	84.01	112	110.97	127	128.32
2	63	63.91	43	43.53	60	64.81	65	67.25
3	23	21.97	29	28.14	41	33.23	51	38.43
4	4	5.04	12	12.13	10	11.36	11	20.00
5	2	1.00	5	5.20	1	3.63		
a	0.6876		1.2928		1.0255		1.1430	
α	0.8519		0.7090		0.7868		0.7264	
X^2	0.06		0.04		25		0.57	
DF	1		2		2		-	
P	0.80		0.98		0.12		-	
C							0.0022	

Conclusion

These results suggest that the 1-displaced Singh-Poisson distribution is the best-fit probability distribution for word lengths in modern Welsh prose texts.

As this distribution could be fitted to all the text-types treated in the study, it seems that genre and domain of discourse are unlikely to affect the distribution of word lengths in Welsh prose. However, further studies are required to confirm this hypothesis. Word-length distributions in Welsh verse also deserve attention, since, in some languages (such as Latin – Wilson 2001) the prose/verse distinction can be significant in determining the distribution of word lengths.

Comparing this set of Welsh data with Drechsler's (2001) Scottish Gaelic data, it seems that the split between P- and Q-Celtic may also have led to different patterns of word-length distribution in the two branches. However, in order to obtain a fuller and more representative picture of word-length distributions in Celtic, further studies also need to be carried out on Breton, Irish, Cornish and Manx, as well as on other text-types of Welsh and Scottish Gaelic.

References

Ball, M.J. (ed.) (1992). *The Celtic languages*. London: Routledge.

Beibl Cymraeg Newydd (1985). *Y Beibl Cymraeg Newydd. Y Testament Newydd. Y Salmau.* Y Gymdeithas Feiblaidd Frytanaidd a Thramor, Llundain.

Best, K-H. (1999). Quantitative Linguistik: Entwicklung, Stand und Perspektive. *Göttinger Beiträge zur Sprachwissenschaft* 2, 7-23.

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

Drechsler, J. (2001). Häufigkeitsverteilungen von Wortlängen in gälischen Texten. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 115-123*. Göttingen: Peust & Gutschmidt.

King, G. (1993). *Modern Welsh: A comprehensive grammar*. London: Routledge.

Macaulay, D. (1992). *The Celtic languages.* Cambridge: Cambridge University Press.

Rowland, T. (1857). *A grammar of the Welsh language, based on the most approved systems*. London: Hughes & Butler.

Williams, S.J. (1980). *A Welsh grammar*. Cardiff: University of Wales Press.

Wilson, A. (2001). Word Length Distributions in Classical Latin Verse. *Prague Bulletin of Mathematical Linguistics 75, 69-84.*

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

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

Glottometrics 6, 2003, 40-51

Satztypen und Satzlängen im Funktional- und Autorenstil

T.V. Dshurjuk, V.V. Levickij[1]

Abstract. Different kinds of texts of three German authors taking into account sentence types and sentence lengths are compared. Using statistical tests different tendencies could be shown.

Keywords: Sentence type, sentence length, text analysis

1. Ziele

Die Länge und die Struktur des Satzes wurden vielfach Objekt quantitativer Analysen (für eine Literaturübersicht s. Admoni 1966, Lesskis 1962, 1963, Altmann 1988; von den Arbeiten der letzteren Jahre sind Grzybek 2001, Levickij et al. 2001, Niehaus 2001, Uhlířová 2001, Best 2001 etc. zu nennen) .

Bei der Erforschung der Satzlänge wurden am häufigsten drei Probleme behandelt: (1) Satzlänge als Stilcharakteristikum; (2) Satzlänge als eines der Merkmale des Autorenstils und als Kriterium zur Entscheidung über die strittige Autorschaft eines Textes; (3) das Modellieren der Satzlängenverteilung.

In diesem Beitrag behandeln wir nur Probleme der Punkte (1) und (2).

2. Das Material und die Methoden

Ausgewählt wurden 22 abgeschlossene literarische Texte der Gegenwart (Romane und Erzählungen) von H. Böll, M. Walser und Ch. Wolf (siehe Literaturverzeichnis), definiert als Zeitraum nach 1945, und Artikel aus dem Nachrichtenmagazin „Der Spiegel“ (1998-2001), in denen die Sparten „Deutschland“ (Politik), Wirtschaft, Wissenschaft und Technik, Kultur betrachtet wurden. Aus den Texten wurden alle Sätze (Einheiten, die voneinander durch die Satzzeichen Punkt, Frage- oder Ausrufezeichen getrennt sind) berücksichtigt.

Die Sätze der erhaltenen Stichproben wurden in vier strukturelle Typen eingeteilt :

a) *Einfache Sätze* sind Satzgebilde, die nur aus einem Hauptsatz bestehen, in denen also kein Nebensatz vorhanden ist, z. B. „Er ist dreizehn Jahre alt“; „Kein Problem“.

b) *Satzreihen* sind zwei oder mehr vollständige Hauptsätze mit allen grammatisch erforderlichen Satzgliedern, die aneinander gereiht sind, z.B. „Er trägt eine Krawatte um den Hals, er ist schon erwachsen, er möchte schreien, aber er tut es nicht.“

c) *Satzgefüge* sind alle diejenigen Satzgebilde, die außer dem Hauptsatz mindestens einen Nebensatz enthalten, z.B. „Es ist kein Wunder, daß er diesen Sommer bei ihr bleibt“ oder „Für die Ermittler geht es darum nur noch um die Frage, wer aus der Hierarchie der Bundesbahn und des Radherstellers auf die Anklagebank muß“.

d) *Syntaktisch-komplizierte Sätze* sind Sätze, die Parenthesen, Ellipsen, satzwertige Parti-

[1] Address correspondence to: V.Levickij, Radiščev-Str. 6/5, Ukr-58000 Černivci, Ukraine

zipien, satzwertige Infinitive, Herausstellungen, Semikolon, Klammern u.s.w. enthalten. Zum Beispiel „Sollten die Ermittlungen, möglicherweise schon im kommenden Jahr, zu Anklagen führen, dann müssten die über 100 Verletzten und die Angehörigen der 101 tödlich Verunglückten nicht länger von „einer Art Naturkatastrophe“ (Löwen) reden; dann gäbe es nicht nur Opfer, sondern auch Schuldige − und die Bahn müsste mit einer Welle von Zivilklagen rechnen“.

Weiter werden die einfachen Sätze im Artikel abgekürzt durch das Symbol ES, Satzgefüge durch SG, Satzreihen werden durch SR und syntaktisch-komplizierte durch SK bezeichnet.

Die gesamte Stichprobe beträgt 36825 Sätze: 15283 Sätze aus Prosawerken und 21542 aus der Publizistik.

3. Die Satztypenverteilung in den Werken von M. Walser

Das Verfahren illustrieren wir anhand der Werke von M. Walser. Für die quantitative Analyse der Prosawerke von M. Walser wurden die folgenden Texte ausgewählt: der Roman „Die Ehen in Philippsburg“ und die Erzählungen „Der Umzug“, „Die Klagen über meine Methoden häufen sich“, „Eigentlich müßte Post da sein“, „Ein Flugzeug über dem Haus“, „Erlebnis“, „Gefahren-voller Aufenthalt“, „Ich als Malteser“, „Ich suchte eine Frau“, „Mißempfindungen“, „Krankheitsempfindungen“, „Symbiose“. Die Verteilung der Häufigkeit verschiedener Satztypen in den Prosawerken von M.Walser ist in der Tabelle 1 angegeben.

Tabelle 1
Die Häufigkeitsverteilung verschiedener Satztypen
im Roman und in den Erzählungen von M.Walser

Textarten	Satztypen				
	ES	*SR*	*SG*	*SK*	Gesamt
Roman	1421	357	1239	115	3132
Erzählungen	815	141	459	63	1478
Gesamt	2236	498	1698	178	4610

Anhand dieser Tabelle wird deutlich, dass M.Walser am häufigsten einfache Sätze (etwa 48%) und Satzgefüge (etwa 37%) gebraucht. Die dritte Stelle nach der Gebrauchshäufigkeit besetzen die Satzreihen; der Anteil der syntaktisch-komplizierten Sätze beträgt nur 4 % (siehe Abb. 1)

Die vorläufige Analyse der Werte in Tabelle 1 zeigt, dass die Häufigkeiten verschiedener Satztypen in dieser Tabelle ungleichmäßig verteilt sind. In diesem Zusammenhang entsteht die Frage, ob die Satztypenverteilung mit dem Genre des Werkes (Roman versus Erzählung) verbunden ist, anders gesagt, ob sich die Häufigkeiten, die in den Zeilen der Tabelle 1 angegeben sind, signifikant voneinander unterscheiden? Die Antwort auf diese Frage kann man z.B. mit Hilfe des Chi-Quadrat-Tests erhalten. Für die Datenverarbeitung der Tabelle 1 wurde die Formel

$$(1) \qquad X^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - E_{ij})^2}{E_{ij}},$$

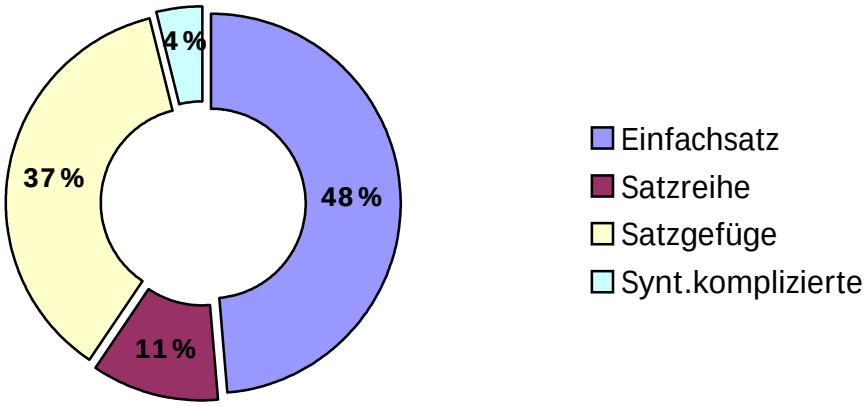

Abbildung 1. Die Gebrauchshäufigkeit verschiedener Satztypen in den Werken von M.Walser

benutzt, wobei n_{ij} die beobachteten und E_{ij} die erwarteten Werte sind. Hier ist $k = 2$ (Zahl der Textarten), $m = 4$ (Zahl der Satztypen). Die Summe $X^2 = 43.51$ mit 3 Freiheitsgraden zeigt, dass in Tabelle 1 der Unterschied zwischen den darin angegebenen Häufigkeiten signifikant ist.

Trotzdem bleibt noch eine ganze Reihe von wichtigen Fragen ungeklärt. Erstens: die festgestellten Unterschiede zwischen der Häufigkeitsverteilung der Satztypen im Roman und in den Erzählungen von M. Walser können nur den Autorenstil dieses Schriftstellers charakterisieren. Wir wissen nicht, ob die festgestellten Unterschiede auch in den Werken anderer Autoren gelten; anders gesagt, es bleibt ungeklärt, ob diese Unterschiede auch allgemeiner gelten. Wir wissen auch nicht, ob irgendeine Abhängigkeit zwischen dem strukturellen Satztyp und dem Genre des Werkes besteht: es ist unklar, ob im Roman der Vorzug dem einen Satztyp und in den Erzählungen einem anderen gegeben wird. Falls solche Abhängigkeiten bestehen, entsteht eine andere Frage: besteht die Möglichkeit, die Assoziation zwischen

(a) dem Satztyp und dem Genre des Werkes ode
(b) dem Satztyp und dem Autorenstil oder
(c) dem Satztyp und verschiedenen funktionalen Stilen (in unserer Untersuchung zwischen der Prosa und Publizistik)

zu testen?

Von verschiedenen Möglichkeiten, die sich hier ergeben, wurde bereits die Partitionierung großer Tabellen in Vierfeldertafeln verwendet, um mit einem Chiquadrat die Assoziationen zu testen und mit dem Φ-Koeffizienten zu charakterisieren (s. Levickij, Romanova 1997; Levickij et al. 2001). An dieser Stelle benutzen wir den Test für einzelne Zellen, der es ermöglicht, ohne Partitionierung die einzelnen Zellen (d.h. ihre Assoziationen) zu testen und zu charakterisieren (vgl. z.B. Altmann, Lehfeldt 1980: 301). Da die erwarteten Werte immer größer als 30 sind, können wir für die einzelnen Zellen das Quantil der Normalverteilung berechnen, nämlich

$$(2) \qquad z = \frac{n_{ij} - E_{ij}}{\sqrt{\frac{n_{i}.n._{j}(n - n_{i}.)(n - n._{j})}{n^2(n-1)}}}$$

wobei $n_{i.}$ = die Randsummen auf der rechten Seite der Tabelle
$n._j$ = die Randsummen unterhalb der Tabelle
n = Summe aller Häufigkeiten in der Tabelle.

Beim Testen treffen wir folgende Entscheidungen auf der 0.05-Ebene:

$z \geq 1.96$ bedeutet eine Assoziation (A)
$z \leq -1.96$ bedeutet eine Dissoziation (D)
$-1.96 < z < 1.96$ bedeutet eine neutrale Beziehung (N)

So ergibt sich beispielsweise für Tabelle 1, für die Beziehung zwischen einfachen Sätzen und dem Roman

$n_{ij} = n_{\text{Roman,Einfache Sätze}} = 1421$
$E_{ij} = 3132(2236)/4610 = 1519.12$
$n = 4610.$

Setzen wir diese Zahlen in Formel (2) ein, so bekommen wir

$$z = \frac{1421 - 1519.12}{\sqrt{\frac{3132(2236)(4610-3132)(4610-2236)}{4610^2(4610-1)}}} = -6.20$$

Da diese Zahl kleiner als -1.96 ist, bedeutet das, dass in Romanen von Walser im allgemeinen einfache Sätze vermieden werden. Auf diese Weise können wir die ganze Tabelle 1 durchtesten und erhalten Resultate, wie in Tabelle 2 angegeben. Der Test ist asymptotisch, liefert aber bei so großen Zahlen zuverlässige Resultate.

Tabelle 2
Beziehungen zwischen Satztyp und Prosagenre bei Walser

	ES	*SR*	*SG*	*SK*
Roman	-6.20 D	1.90 N	5.59 A	-0.97 N
Erzählung	6.20 A	-1.90 N	-5.59 D	0.97 N

Praktisch sind nur ES und SG signifikant mit dem Genre verbunden, jeweils in entgegengesetztem Sinne: der Roman bevorzugt SG und vermeidet ES, die Erzählungen verhalten sich umgekehrt.

4. Die Satztypenverteilung in den Werken von drei Autoren

Das Korpus der Sätze in den Werken von den drei Autoren (M.Walser, Ch.Wolf, H.Böll) besteht aus 15283 Einheiten. Das Prozentverhältnis von vier Satztypen in diesem Korpus veranschaulicht Abb.2.

Die Häufigkeitsverteilung der Satztypen in den Werken von drei Autoren ist in Tabelle 3 angeführt.

Der Homogenitätstest (1) ergibt $X^2 = 280.5$, was die kritische Größe signifikant überschreitet (bei 6 FG ist $\chi^2_{0,05} = 12,6$). Die Verteilungen in Tabelle 3 unterscheiden sich daher voneinander signifikant.

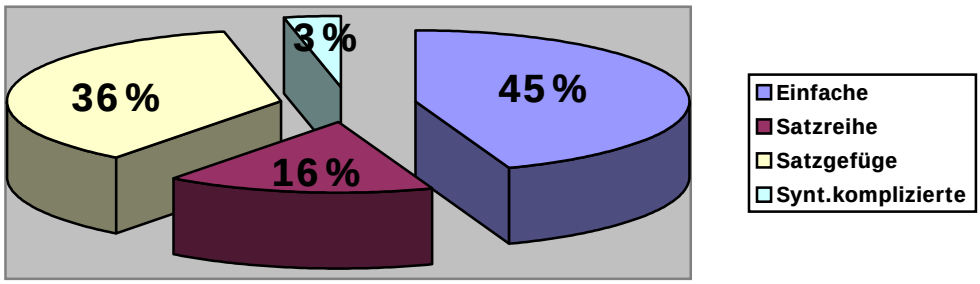

Abbildung. 2. Die Verteilung von 4 Satztypen in der Prosa von den drei Autoren

Tabelle 3
Die Häufigkeitsverteilung der Satztypen in den Werken von drei Autoren

Autoren	Satztypen				Gesamt
	ES	*SR*	*SG*	*SK*	
Walser	2236	498	1698	178	4610
Wolf	3632	1161	2498	131	7422
Böll	919	766	1370	196	3251
Gesamt	6787	2425	5566	505	15283

Beim Test für Einzelzellen muß man beachten, dass hier der Kontrast nicht im Genre besteht, sondern zwischen einzelnen Autoren. Die Resultate sind daher nicht allgemein gültig, sondern nur für die Präferenzen der drei Autoren untereinander.

Berechnet man (2) für jede Zelle der Tabelle 3, so erhält man Resultate, die in Tabelle 4 angegeben sind.

Tabelle 4
Beziehungen zwischen Autor und Satztyp

Autoren	Satztypen			
	ES	*SR*	*SG*	*SK*
Walser	6.96 **A**	-11.26 **D**	0.70 **N**	2.53 **A**
Wolf	10.94 **A**	-0.74 **N**	-6.90 **D**	-10.34 **D**
Böll	-20.87 **D**	13.53 **A**	7.64 **A**	9.79 **A**

Das bedeutet, daß in den Werken von M. Walser die Gebrauchshäufigkeit der einfachen und syntaktisch komplizierten Sätze die theoretisch ermittelten Werte signifikant überschreitet; Ch. Wolf gibt den Vorzug den einfachen, und H.Böll, im Gegensatz zu Ch.Wolf, den zusammengesetzten Sätzen. Walser bevorzugt entweder einfache oder sehr komplizierte Sätze. Die Dissoziationen bestehen für (Walser, SR) und (Wolf, SG), (Wolf, SK) und (Böll, ES). Es ist nochmals zu betonen, dass hier eine Tendenz nur im Vergleich zu den anderen Autoren, nicht absolut, gilt.

Für die Untersuchung der eventuellen Kontingenz zwischen dem Genre des Prosawerkes und dem Satztyp wurden die Daten der Tabelle 3 anders gruppiert (s. Tabelle 5) und wieder

statistisch bearbeitet. Dies ist eine Verallgemeinerung des Genre-Problems.

Die Ergebnisse der Analyse haben gezeigt, dass sich die Häufigkeiten, die in Tabelle 5 angegeben sind, signifikant voneinander unterscheiden ($X^2 = 31.6$; FG = 3), d.h. die beiden Textarten sind nicht homogen.

Tabelle 5
Gebrauchshäufigkeit verschiedener Satztypen in Romanen und Erzählungen

Textarten	Satztypen				Gesamt
	ES	*SR*	*SG*	*SK*	
Romane	4841	1734	3839	290	10704
Erzählungen	1946	691	1727	215	4579
Gesamt	6787	2425	5566	505	15283

Den Homogenitätstest kann man statt mit (1) alternativ und etwas schneller mit der Informationsstatistik durchführen, indem man

$$2I = 2\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij} \ln n_{ij} + 2n \ln n - 2\sum_{i=1}^{k} n_{i.} \ln n_{i.} - 2\sum_{ji=1}^{m} n_{.j} \ln n_{.j} \quad (3$$

berechnet, was auch mit einem Taschenrechner schnell geht. Für Tabelle 5 erhalten wir

$$2\sum_{i=1}^{2}\sum_{j=1}^{4} n_{ij} \ln n_{ij} = 2[4841 \ln 4841 + 1734 \ln 1734 + ... + 215 \ln 215] = 241238.2483$$

$$2n \ln n = 2(15283) \ln 15283 = 294488.0163$$

$$2\sum_{i=1}^{2} n_{i.} \ln n_{i.} = 2[10704 \ln 10704 + 4579 \ln 4579] = 275826.3470$$

$$2\sum_{j=1}^{4} n_{.j} \ln n_{.j} = 2[6787 \ln 6787 + ... + 505 \ln 505] = 259853.0790 .$$

Daraus ergibt sich

$$2I = 275826.3470 + 294488.0163 - 275826.3470 - 259853.0790 = 46.84$$

Das Resultat ist etwas höher als mit dem klassischen Test, aber die Aussage ist die gleiche.

Die Einzelheiten kann man mit (2) ermitteln. Dabei werden die einfachen Sätze (s. Tabelle 6) im Roman öfter, als man das theoretisch erwarten kann, verwendet, und in den Erzählungen werden Satzgefüge und syntaktisch komplizierte Sätze entsprechend öfter eingesetzt.

Die erhaltenen Ergebnisse sind etwas unerwartet. Man könnte z.B. erwarten, dass in Romanen komplizierte Sätze und in Erzählungen einfache Sätze öfter vorkommen. Aber wir haben etwas Umgekehrtes festgestellt: Je komplexer der Satz, desto weniger wird er in Romanen benutzt, und umgekehrt in Erzählungen. Dieses Ergebnis kann man dadurch erklären, dass im Roman die direkte Rede mit den relativ kürzeren Erwiderungen der Personen und verhältnismäßig kurzen Sätzen der Autorenrede, die die direkte Rede einführen, öfter gebraucht wird. In den Erzählungen wird, in der Regel, das ganze Erzählen vom Autor geführt; dabei werden bei der Beschreibung der Natur oder der inneren Gemütsbewegungen der han-

delnden Personen nicht selten komplizierte und superkomplizierte Sätze gebraucht.

Tabelle 6
Beziehungen der Satztypen zu den Textarten

Textarten	Satztypen			
	ES	*SR*	*SG*	*SK*
Romane	3.11 **A**	1.72 **N**	-2.18 **D**	-6.29 **D**
Erzählungen	-3.11 **D**	-1.72 **N**	2.18 **A**	6.29 **A**

Und schließlich wurden für die Untersuchung der Verteilung der Satztypen in den verschiedenen funktionalen Stilen die entsprechenden Häufigkeiten in der Prosa und Publizistik einander (s. Tabelle 7) gegenübergestellt.

Tabelle 7
Gebrauchshäufigkeit von vier Satztypen in der Prosa und Publizistik

Textarten	Satztypen				Gesamt
	ES	*SR*	*SG*	*SK*	
Prosa	6787	2425	5566	505	15283
Publizistik	11679	2799	6919	145	21542
Gesamt	18466	5224	12485	650	36825

Die Analyse der Dateien der Tabelle 7 mit Hilfe von (2) oder (3) hat gezeigt, dass eine offenbar ungleichmäßige Verteilung der Satztypen in den Prosawerken und der Publizistik (s. Tabelle 8) besteht.

Tabelle 8
Satztypen in Prosa und Publizistik

Textarten	Satztypen			
	ES	*SR*	*SG*	*SK*
Prosawerke	-18.54 **D**	7.79 **A**	8.59 **A**	18.89 **A**
Publizistik	18.54 **A**	-7.79 **D**	-8.59 **D**	-18.89 **D**

In der Publizistik werden, öfter als man erwartet, die einfachen Sätze, und in der Prosa kompliziertere gebraucht. Man kann zulassen, dass in verschiedenen Arten der publizistischen Texte (der politische Text, der wirtschaftliche Text usw.) die Häufigkeit der von uns analysierten Satztypen auf eine andere Weise verteilt ist, d.h., nicht so wie in der Publizistik im Vergleich zu den Prosawerken. Für die Überprüfung dieser Hypothese wurden die Dateien der Tabelle 7 so umgruppiert, dass man die Häufigkeitsverteilung in zwei Arten der Prosatexte und in vier Arten der publizistischen Texte (siehe Tabelle 9) erhält.

Tabelle 9
Gebrauchshäufigkeit von vier Satztypen in 6 Arten der Texte

Textarten	Satztypen				Gesamt
	ES	*SR*	*SG*	*SK*	
Romane	4841	1734	3839	290	10704
Erzählungen	1946	691	1727	215	4579
Die Lexik der Politik	4268	1117	2677	62	8124
Wirtschaftslexik	2631	575	1373	12	4591
Wissenschaftslexik	2110	460	1100	10	3680
Die Lexik der Kultur	2670	647	1769	61	5147
Gesamt	18466	5224	12485	650	36825

Die Ergebnisse der Analyse (s. Tabelle 10) zeigen, dass sich die früher nachgewiesenen Dateien über die Kontingenz der Merkmale in der Prosa und Publizistik (s. Tabelle 8) im Grunde nicht ändern.

Tabelle 10
Tests für unterschiedliche Textarten

Textarten	*ES*	*SR*	*SG*	*SK*
Romane	-12.09	7.09	5.09	8.81
Erzählungen	-11.06	1.87	5.82	16.09
Die Lexik der Politik	4.88	-1.28	-2.05	-7.77
Wirtschaftslexik	10.37	-3.45	-6.12	-8.27
Wissenschaftslexik	9.20	-3.09	-5.42	-7.25
Die Lexik der Kultur	2.63	-3.58	0.76	-3.41

Die einfachen Sätze kommen öfter in der Publizistik vor, die komplizierten Sätze werden öfter in der Prosa gebraucht.

5. Die Satzlänge

5.1. Die Satzlänge im Autorenstil

Satzlänge kann entsprechend der Anzahl verschiedener Einheiten wie Phoneme, Morpheme, Wörter u.s.w. gemessen werden. In unserer Untersuchung ist die Messeinheit das Wort. Die durchschnittliche Satzlänge ist gegeben als die Anzahl der Wörter im analysierten Text, in der Gruppe der Texte oder im Textfragment, dividiert durch die Gesamtzahl der Sätze im Textmaterial. Es wird angenommen, dass der Forscher über einen Satz von Kriterien verfügt, mit dessen Hilfe er den strukturellen Satztyp und das, was man als Wort bezeichnet, eindeutig bestimmen kann. Die Durchschnittslänge von vier Satztypen in den Texten der drei Autoren ist in Tabelle 11 dargestellt.

Da die Quote der verschiedenen Satztypen in der künstlerischen Prosa (s. Abb. 2) nicht gleich ist, wurde in Tabelle 11 die *gewichtete* durchschnittliche Satzlänge im Text jedes Autors auf Grund der Gesamtheit aller Texte (Romane und Erzählungen) berechnet (s. letzte Zeile der Tabelle 11). Daher ist die durchschnittliche Satzlänge in den Werken von M. Walser

gleich 14.62, und nicht 26.75, was sich durch Mittelung der Spalten in Tabelle 11 ergäbe.

Tabelle 11 zeigt, dass die durchschnittliche Satzlänge bei M. Walser und Ch. Wolf fast übereinstimmt, während sie bei H. Böll annähernd 1.5 mal größer ist als bei den anderen Autoren. Hier kann man diese Feststellung auch ohne Test akzeptieren.

Da Länge und Satztyp stark korrelieren, ist im Grunde nur die letzte Zeile der Tabelle 11 entscheidend. Im weiteren werden daher Typ und Länge nicht als zwei Faktoren berücksichtigt.

Tabelle 11
Durchschnittslänge der Sätze in den Werken von drei Autoren

Arten der Sätze	Walser	Wolf	Böll	Durchschnittslänge einzelner Satztypen
ES	7.09	7.37	8.09	**7.51**
SR	15.93	14.68	15.82	**15.48**
SG	19.68	21.03	29.23	**23.31**
SK	64.31	51.52	67.02	**60.88**
Gewichtete durchschnittliche Satzlänge bei Autoren	**14.62**	**14.25**	**23.31**	**17.39**

5.2. Die Satzlänge und die Art des Textes

Um weitere Rechnungen zu vereinfachen, haben wir die Satzlängen auf drei Klassen reduziert:

- Sätze mit einer Länge von 1 bis 12 Wörtern (kurze Sätze)
- Sätze mit einer Länge von 13 bis 24 Wörtern (mittlere Sätze)
- Sätze mit über 24 Wörtern Länge (lange Sätze)

Demnach ergaben sich die einzelnen Häufigkeiten in unterschiedlichen Textarten wie in Tabelle 12 dargestellt.

Tabelle 12
Gebrauchshäufigkeit von drei Unterklassen der Sätze in sechs Textarten

Satzklassen	Textarten						Summe
	Roman	Erzählung	Lexik der Politik	Wirtschaftslexik	Wissenschaftslexik	Lexik der Kultur	
Kurze	5770	2334	3113	1861	1418	2156	16652
Mittlere	2948	1242	3640	2104	1786	2041	13761
Lange	1986	1003	1371	626	476	950	6412
Summe	10704	4579	8124	4591	3680	5147	36825

Tabelle 17
Tests für Assoziation der Satzlängenklassen mit der Textart

Satz-klassen	Textarten					
	Roman	Erzählung	Lexik der Politik	Wirtschafts-lexik	Wissen-schaftslexik	Lexik der Kultur
Kurze	21.44	8.36	-14.16	-6.81	-8.59	-5.18
Mittlere	-24.95	-15.31	15.69	12.66	14.76	3.65
Lange	3.70	8.57	-1.44	-7.21	-7.55	2.13

Bei so großen Zahlen empfiehlt es sich, das Signifikanzniveau lieber zu erhöhen, falls man die Zahlen als Charakteristika benutzt. Nicht zu vergessen ist die Tatsache, dass diese Resultate nur Vergleichsmaße, keine absoluten Charakteristika sind. Als Charakteristika eignen sich dann besser die durchschnittlichen Längen mit der Angabe der Dispersion.

Zusammenfassung

Die Gebrauchshäufigkeit der unterschiedlichen strukturellen Typen des Satzes und die Satzlänge können von vielen Faktoren abhängen. In unserer Studie wurde die Abhängigkeit der Satzlängen- und Satztypenverteilungen von der Art des Textes, der syntaktischen Struktur des Satzes und dem Faktor, den wir als „Autorenstil“ bezeichnen, erforscht.

Es wurde festgestellt, dass die einfachen Sätze öfter, als man das erwarten könnte, in der Publizistik vorkommen; komplizierte Sätze sind häufiger in den Romanen und Erzählungen anzutreffen. M. Walser und Ch. Wolf geben den Vorzug einfachen Sätzen, H. Böll – den Satzreihen und Satzgefügen.

Die Durchschnittslänge des Satzes in der Prosa ist 17.39, in der Publizistik 16.06 Wörter. Je nach dem strukturellen Typ des Satzes schwankt seine Durchschnittslänge in den nachgeprüften Texten von 7.51 (die einfachen Sätze in der Prosa) bis 60.9 (syntaktisch komplizierte Sätze in den Prosawerken). Ch. Wolf gibt in ihren Werken kurzen und mittleren Sätzen den Vorzug, während M. Walser und H. Böll den langen Sätzen den Vorzug geben. Die Satzkomplexität und die Satzlänge sind automatisch stark korreliert. Die Verteilung der mittleren und langen Sätze entspricht aber nicht immer den intuitiven Vorstellungen: als lang erweisen sich meistens nicht syntaktisch komplizierte Sätze, sondern Satzgefüge; die Sätze der mittleren Länge haben meistens die Struktur einer Satzreihe oder eines Satzgefüges. Die Länge und der Gebrauch verschiedener struktureller Satztypen erfordern weiteres Studium. Es wäre besonders zweckmäßig, solche Faktoren wie den Autorenstil, das Genre, die chronologischen Rahmen des Schaffens des Schriftstellers und möglichst auch den „Gender“- Faktor zu erforschen.

Literatur

Admoni,V.G. (1966). *Razvitie predloženia v period formirovania nemeckogo nacionalnogo jazyka*. Leningrad: Nauka.

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

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

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

Grzybek, P. (2001). Zur Satz- und Teilsatzlänge zweigliedriger formelhafter Sprichwörter In: Uhlířová, L. et al. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs: 64-75.* Trier: Wissenschaftlicher Verlag.

Lesskis, G.A. (1962). O razmerach predloženij v russkoj naučnoj i chudožestvennoj proze 60-ch godov XIX v. *Voprosy jazykoznanija No 2, 78-95.*

Lesskis, G.A. (1963). O zavisimosti meždu razmerom predloženia i charakterom teksta. *Voprosy jazykoznanija, No. 3, 92-112.*

Levickij, V.V., Pavlycko, O.O., Semenyuk, T.G. (2001). Sentence Length and Sentence Structure as Statistical Charakteristics of Style in Prose In: Uhlířová, L. et al. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs: 177-186.* Trier: Wissenschaftlicher Verlag.

Levickij, V.V., Romanova, T.A. (1997). Use of tenses of verbs and adverbs in the English language: a statistical study. *J. of Quantitative Linguistics 4, 135-138.*

Niehaus, B. (2001). Die Satzlängenverteilung in literarischen Prosatexten der Gegenwart In: Uhlířová, L. et al. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs: 196-213.* Trier: Wissenschaftlicher Verlag.

Uhlířová, L. (2001). On Word Length, Clause Length and Sentence Length in Bulgarian. In: Uhlířová, L. et al. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs: 266-282.* Trier: Wissenschaftlicher Verlag.

Quellentexte

a) Prosawerke:

1. Böll, H. An der Brücke. //Und sagte kein einziges Wort. Erzählungen. Moskau, Verlag für fremdsprachige Literatur, 1958.- S.310

2. Böll, H. Das Brot der frühen Jahre.//Und sagte kein einziges Wort. Erzählungen. Moskau, Verlag für fremdsprachige Literatur, 1958.- S.310

3. Böll, H. Der Mann mit den Messern.// Und sagte kein einziges Wort. Erzählungen. Moskau, Verlag für fremdsprachige Literatur, 1958.- S.310

4. Böll, H. Die Botschaft. //Und sagte kein einziges Wort. Erzählungen. Moskau, Verlag für fremdsprachige Literatur, 1958.- S.310

5. Böll, H. Lohengrins Tod. //Und sagte kein einziges Wort. Erzählungen. Moskau, Verlag für fremdsprachige Literatur, 1958.- S.310

6. Böll, H. Und sagte kein einziges Wort. Erzählungen. Moskau, Verlag für fremdsprachige Literatur, 1958.- S.310

7. Böll, H. Wanderer, kommst du nach Spa...//Und sagte kein einziges Wort. Erzählungen. Moskau, Verlag für fremdsprachige Literatur, 1958.- S.310

8. Walser, M. Der Umzug. // Werke in zwölf Bänden. Bd.8.- Frankfurt am Main: Suhrkamp Verlag, 1997.- S. 39-45.

9. Walser, M. Die Klagen über meine Methoden häufen sich. // Werke in zwölf Bänden. Bd.8.- Frankfurt am Main: Suhrkamp Verlag, 1997.- S. 110-121.

10. Walser, M. Ehen in Philippsburg. Suhrkamp Verlag,1985.- 343 S.

11.Walser, M. Eigentlich müsste Post da sein. // Werke in zwölf Bänden. Bd.8.- Frankfurt am Main: Suhrkamp Verlag, 1997.- S. 238-241

12. Walser, M. Ein Flugzeug über dem Haus. // Werke in zwölf Bänden. Bd.8.- Frankfurt am Main: Suhrkamp Verlag, 1997.- S. 9-20.

13. Walser, M. Erlebnis. // Werke in zwölf Bänden. Bd.8.- Frankfurt am Main: Suhrkamp Verlag, 1997.- S. 351-361.

14. Walser, M. Gefahrenvoller Aufenthalt. // Werke in zwölf Bänden. Bd.8.- Frankfurt am Main: Suhrkamp Verlag, 1997.- S. 20-30.

15. Walser, M. Ich als Malteser. // Werke in zwölf Bänden. Bd.8. Frankfurt am Main: Suhrkamp Verlag, 1997.- S. 345-351.

16. Walser, M. Ich suchte ein Frau. // Werke in zwölf Bänden. Bd.8. Frankfurt am Main: Suhrkamp Verlag, 1997.- S. 31-39.

17. Walser, M. Mißempfindungen. Krankheitsempfindungen. // Werke in zwölf Bänden. Bd.8.- Frankfurt am Main: Suhrkamp Verlag, 1997.- S. 15-25

18. Walser, M. Symbiose. // Werke in zwölf Bänden. Bd.8. Frankfurt am Main: Suhrkamp Verlag, 1997.- S. 33-42

19. Wolf, Ch. Der geteilte Himmel. Halle/Saale, Mitteldeutscher Verlag,1965.- 317 S.

20. Wolf, Ch. Neue Lebensansichten eines Katers. // Unter den Linden. Berlin und Weimar: Aufbau-Verlag, 1975.- 133 S.

21. Wolf, Ch. Selbstversuch. // Unter den Linden. Berlin und Weimar: Aufbau-Verlag, 1975.- 133 S.

22. Wolf, Ch. Unter den Linden. //Unter den Linden. Berlin und Weimar: Aufbau-Verlag,1975.- 133 S.

b) Pressetexte

a. Sparte Deutschland:

23. Der Spiegel, № 29, 52. Jahrgang, 1998, S.17-85.
24. Der Spiegel, № 6, 53. Jahrgang, 1999, S.17-85.
25. Der Spiegel, № 17, 54. Jahrgang, 2000, S.17-71.
26. Der Spiegel, № 20, 54. Jahrgang, 2000, S.17-90.
27. Der Spiegel, № 21, 54. Jahrgang, 2000, S.17-83.
28. Der Spiegel, № 23, 54. Jahrgang, 2000, S.17-75.
29. Der Spiegel, № 42, 55. Jahrgang, 2001, S.17-71.
30. Der Spiegel, № 44, 55. Jahrgang, 2001, S.17-73.

b. Sparte Wirtschaft:

31. Der Spiegel, № 29, 52. Jahrgang, 1998, S.85-105.
32. Der Spiegel, № 6, 53. Jahrgang, 1999, S.85-99.
33. Der Spiegel, № 17, 54 . Jahrgang, 2000, S. 71-93.
34. Der Spiegel, № 20, 54. Jahrgang, 2000, S.91-115.
35. Der Spiegel, № 21, 54. Jahrgang, 2000, S.83-121.
36. Der Spiegel, № 23, 54. Jahrgang, 2000, S.75-95.
37. Der Spiegel, № 42, 55. Jahrgang, 2001, S.123-158.
38. Der Spiegel, № 44, 55. Jahrgang, 2001, S.73-105.

c. Sparte Wissenschaft und Technik:

39. Der Spiegel, № 29, 52. Jahrgang, 1998, S.139-155.
40. Der Spiegel, № 6, 53. Jahrgang, 1999, S.169-193.
41. Der Spiegel, № 17, 54. Jahrgang, 2000, S.225-247.
42. Der Spiegel, № 20, 54. Jahrgang, 2000, S.213-239.
43. Der Spiegel, № 21, 54. Jahrgang, 2000, S.225-247.
44. Der Spiegel, № 23, 54. Jahrgang, 2000, S.217-249.
45. Der Spiegel, № 42, 55. Jahrgang, 2001, S.275-294.
46. Der Spiegel, № 44, 55. Jahrgang, 2001, S.203-222.

d. Sparte Kultur:

47. Der Spiegel, № 29, 52. Jahrgang, 1998, S.155-190.
48. Der Spiegel, № 6, 53. Jahrgang, 1999, S.193-230.

Glotometrics 6, 2003, 52-60

Word lengths in the Baltic languages – are they of the same type as the word lengths in the Slavic languages?

Otto A. Rottmann, Bochum[1]

Abstract: In our present analysis we found that word length in the two living Baltic languages and the majority of the Slavic languages (see 4.0) is controlled by the Extended Positive Binomial distribution (EPB) (with the rest of the Slavic languages being governed by members of the distribution family to which the EPB distribution belongs) which leads to the assumption that Baltic and Slavic languages do not only stem from a common evolutionary branch and are therefore members of a diachronically oriented language family, but it is also possible to find individual phenomena at synchronous level subsuming the languages concerned under one type, in our case the type of <word length>.

Key words: Baltic languages, Slavic languages, word length, Extended Positive Binomial distribution, unified theory

0. Introduction

The idea for this paper resulted from a publication by Poljakov (1995) in which he confirmed the findings by the great Endzelin (1911) according to which the Baltic and the Slavic languages actually originate from the ramification of a common linguistic branch: the Baltic-Slavic branch of the big Indo-European language tree.

The fact that the allocation to such classes is based on history does not prevent us from analysing whether the Baltic languages also belong to common types at synchronous level, in our case the type of word length (for the difference between "class" and "type" or "classification" and "typology" see Rottmann (2003).

This paper has the following objectives:

1 Finding the mechanism governing word length in Latvian texts.
2 Finding the mechanism governing word length in Lithuanian texts.
3 Allocation of the word lengths in the Latvian and Lithuanian languages to one type or types.
4 Comparison of the type(s) with that of the Slavic languages, in other words, answer the question whether or not word lengths in the Baltic and the Slavic languages are governed by one common type.

This paper does not discuss theoretical problems (such a discussion can be found in the above mentioned publication), but is meant to present empirical results with conclusions as to typology.

[1] Address correspondence to: Otto Rottmann, Behrensstr. 19, D-58099 Hagen.
E-mail: Otto.Rottmann@t-online.de

1.0. General

The analyses are based on written prose texts randomly chosen with half of them being fictional, the other half non-fictional. We considered <word> to be a unit at orthographic level, i.e. its end is marked by a subsequent blank or punctuation. Word length is defined by the number of syllables in a word. Generally, the number of syllables in a word is identical with the number of vowels. A special situation occurs with entities like the Russian prepositions *с*, *в* or *к*: though those entities are without a vowel (due to phonetic changes in the history of the Russian language), we consider them one-syllable words because they are entries in the lexicon and meet the definition given above, because they (being prepositions) are followed by a blank. The apodictic opinion − according to which <word> *must* be defined as a phonetic entity, an opinion borrowed from Mel'čuk and occasionally found in literature is expressly rejected due to the choice of the orthographic level, which is deemed appropriate for written texts.

In our study counting was based on the following additional criteria:

a) Initials of first names and patronymics were counted as one syllable.
b) Abbreviations occurring in the texts were dissolved and counted as if the text included the non-abbreviated forms.
c) Abbreviation words were counted in compliance with the potential inflection (e.g. *ГУМ* consists of one syllable, *ГУМе* comprises two and *колхозе* three syllables).
d) Words were taken as we found them in the texts, not evaluated according to any differing correct spelling (a criterion which is e.g. especially important for Old Bulgarian where the written language reflects phonetic changes in the spoken language).
e) Numbers, decimal numbers, years written in figures were counted as if written in full words.
f) Headings and captions were counted, as in the case of word length it is irrelevant if words are part of a full sentence, an ellipsis or a word combination.
g) Quotations were only taken into consideration if they were worded in the same language as the text not being part of a quotation (e.g. a quotation in Latvian would not be counted if occurring in a Lithuanian text).
h) Proper names were included in the counting, if they were part of the language of the text.

Word length data were processed by means of the software *Fitter*. The evaluation of the test results is mainly based on the *P*-level of the chi-square criterion: if $P \geq 0.01$, the result can at least be called "satisfactory". Surely, this value does not have the character of a statistical law, but just a conventional decision. In many sciences, e.g. in the social sciences or in metallurgy, the threshold taken as the basis for decision is $P \geq 0.05$. In linguistics, however, the criterion could even be lowered, since the number of data processed is considerably high: it is a known fact that the chi-square grows with the size of the sample. We stay with the above decision. In those cases in which the value of *P* is not acceptable – e.g. the sample size is too great – or if the number of degrees of freedom is zero ($DF = 0$) and it can-not be computed, another discrepancy criterion is used. The software *Fitter* computes $C = X^2/N$. *C* is considered satisfactory, if its value is $C \leq 0.01$. Values in the range $0.01 \leq C \leq 0.02$ are weak, but tolerable. *C* does not depend on degrees of freedom, it merely relativizes the observed discrepancy.

2.0. Word length in the Latvian language (objective 1)

The analysis is based on 24 Latvian prose texts [with LAT 1 to LAT 12 being the fictional texts (mainly excerpts/sections from so-called "long forms") and LAT 13 to LAT 24 the non-fictional ones (articles from journals and papers)].

Those texts are the following:

LAT 1 Virza, Edvarts. Straumēni. Vecā Zemgales māja gada gaitās. Rīga 1989, p. 64-68.
LAT 2 Klīdzējs, Jānis. Cilvēka bērns. Rīga 1991. p. 15-17.
LAT 3 Klīdzējs, Jānis. Cilvēka bērns. Rīga 1991. p. 132-135.
LAT 4 Pumpurs, Andrejs. Lāčplēsis. Rīga 1988. p. 146-152.
LAT 5 Viks. Dinīts nāk! Rīga 1990. p. 38-44.
LAT 6 Ezera, Regīna. Pūķa ola. Rīga 1995. p. 7-12.
LAT 7 Poruks, Jānis. Mājās und ceļā. Rīga 1989. p. 111-115.
LAT 8 Poruks, Jānis. Mājās un ceļā. Rīga 1989. p. 124-127.
LAT 9 Blaumanis, Rūdolfs. Brīnuma zālīte. Rīga 1976. p. 69-73.
LAT 10 Blaumanis, Rūdolfs. Brīnuma zālīte. Rīga 1976. p. 407-411.
LAT 11 Sakse, Anna. Zāles stiebrs, Rīga 1987. p. 169-173.
LAT 12 Sakse, Anna. Zāles stiebrs, Rīga 1987. p. 197-200.
LAT 13 Rancāns, Antons. „Vasaras izskaņa 'Andrupenes lauku sētā'". In: Vietējā Nedēļas Avīze Nr. 34/ 119 (2001), p. 24.
LAT 14 Valtere, Edīte. „Neatkarīgie prasa uzmanību". In: Mājas un Dārzs, februāris 2001, p. 92.
LAT 15 Zelenkovs, Andris. „Uzvara, kas neglāba zviedrus". In: Lauku Avīze Nr. 78-1298 (2001), p. 9.
LAT 16 Mārtuža, Eva. „Slazdu vilinājums". In: Lauku Avīze Nr. 80-1300 (2001), p. 30.
LAT 17 Samauska, Leva. „Dienišķā und Svētā". In: Leva Nr. 20-188 (2001), p. 30.
LAT 18 Metuzāls, Sandris. „Fotofilmu ēras noriets". In: Klubs, aprīlis 2001, p. 46.
LAT 19 Pelūde, Anija. „Šūpojies līganā bērzā". In: Leva Nr. 15/16-183/184 (2001), p. 33.
LAT 20 Mukāne, Ināra. „Kur meklēt Laimes Lāci?". In: Lauku Avīze Nr. 79-1299 (2001), p. 17.
LAT 21 Niedre, Ojārs. „CK un ČK. LKP loma 1949. gada deportācijā". In: Lauku Avīze Nr. 35-1255 (2001), p. 11.
LAT 22 Svīre, Māra. „Iztikas minimums? Un vēl pilns?". In: Lauku Avīze Nr. 12/ 1232, p. 10.
LAT 23 Zemberga, Kaija. „Kuldīga – pasaules mantojuma sarakstā?". In: Lauku Avīze Nr. 84-1304 (2001), p. 7.
LAT 24 Medvedeva, Marina. "Cilvēks 'aiz borta'". In: Rēzeknes Vēstis Nr. 98/ 9068 (2001), p. 3.

The analysis performed by means of the software *Fitter* on the basis of the number of syllables resulted in the 1-displaced EPB distribution as the suitable model for the mechanism controlling Latvian word length. The relevant formula is as follows:

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

The parameter α is estimated $\hat{\alpha} = 1 - f_1 / N$ and it remains constant throughout the iteration procedure; q = 1-p.

In the tables of data LAT 1 to LAT 24 for the Latvian language and LIT 1 to LIT 20 for the Lithuanian language listed below *x* indicates the variable (classes), f_x the number of events observed for the relevant class in the text analyzed; NP_x specifies the theoretical, i.e. computed value for the relevant class. *n, p* and α are the parameters of the distribution. X^2 is the chi-square, the index of the X^2 specifies the number of the degrees of freedom; *C* is the alternatively exploited evaluation criterion (see above).

Analysis of the individual data:

	LAT 1		LAT 2		LAT 3	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	355	355.00	556	556.00	597	597.00
2	504	498.77	519	519.03	500	510.33
3	271	282.36	216	214.08	252	227.32
4	88	79.92	41	44.15	31	45.00
5	9	11.31	6	4.74	3	3.34
6	1	0.64				
	$n = 5$ $p = 0.2206$		$n = 5$ $p = 0.1710$		$n = 4$ $p = 0.2290$	
	$\alpha = 0.7109$		$\alpha = 0.1710$		$\alpha = 0.5683$	
	$X_1^2 = 1.65$ $P = 0.20$		$X_1^2 = 0.58$ $P = 0.45$		$C = 0.005$	

	LAT 4		LAT 5		LAT 6	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	205	205.00	296	296.00	320	320.00
2	370	392.34	319	306.30	300	290.86
3	369	262.15	229	233.67	202	221.48
4	0	77.85	83	101.87	108	96.37
5	2	8.67	42	27.76	27	26.21
6			2	5.40	3	5.09
	$n = 4$ $p = 0.3082$		$n = 8$ $p = 0.1790$		$n = 8$ $p = 0.1787$	
	$\alpha = 0.7833$		$\alpha = 0.6952$		$\alpha = 0.6667$	
	$C = 0.0028$		$X_1^2 = 2.93$ P = 0.09		$X_2^2 = 4.29$ $P = 0.12$	

	LAT 7		LAT 8		LAT 9	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	405	405.00	399	399.00	331	331.00
2	421	412.62	363	347.59	387	376.93
3	210	226.88	161	187.76	137	168.89
4	79	69.31	76	56.35	61	40.36
5	12	12.70	3	11.31	5	5.43
6	1	1.49			2	0.40
	$n = 7$ p = 0.1549		$n = 7$ $p = 0.1526$		$n = 6$ $p = 0.1520$	
	$\alpha = 0.6410$		$\alpha = 0.6018$		$\alpha = 0.6414$	
	$X_2^2 = 2.98$ $P = 0.23$		$C = 0.0016$		$C = 0.0012$	

	LAT 10		LAT 11		LAT 12	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	449	449.00	292	292.00	304	304.00
2	349	346.28	364	367.42	335	342.46
3	160	165.34	255	249.65	183	166.13
4	42	39.46	76	75.39	31	40.29
5	5	4.94	6	8.54	4	4.89
6					1	0.24
	$n = 5$, $p = 0.1927$, $\alpha = 0.5532$ $X_1^2 = 0.36$ $P = 0.55$		$n = 4$ $p = 0.3118$ $\alpha = 0.7059$ $X_1^2 = 0.34$ $P = 0.91$		$n = 5$ $p = 0.1952$ $\alpha = 0.6457$ $C = 0.0005$	

	LAT 13		LAT 14		LAT 15	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	172	172.00	195	195.00	160	160.00
2	259	250.12	294	274.80	309	301.88
3	191	192.73	146	173.88	265	278.06
4	71	86.63	65	62.87	150	146.35
5	31	25.03	21	14.21	52	48.14
6	8	5.50	2	2.25	9	10.14
7					0	1.33
8					1	0.10
	$n = 9$ $p = 0.1615$ $\alpha = 0.7650$ $X_1^2 = 2.31$ $P = 0.13$		$n = 8$ $p = 0.1531$ $\alpha = 0.7303$ $X_1^2 = 3.52$ $P = 0.06$		$n = 8$ $p = 0.2083$ $\alpha = 0.8309$ $X_3^2 = 1.44$ $P = 0.70$	

	LAT 16		LAT 17		LAT 18	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	403	403.00	247	247.00	208	208.00
2	313	300.60	328	310.96	227	224.75
3	148	172.70	200	219.98	230	212.40
4	77	61.42	86	92.22	100	125.46
5	14	15.12	33	25.37	52	51.87
6	0	2.73	7	5.47	25	15.93
7	1	0.42			0	3.76
8					1	0.82
	$n = 15$ $p = 0.0758$ $\alpha = 0.5785$ $X_1^2 = 2.42$ $P = 0.12$		$n = 10$ $p = 0.1358$ $\alpha = 0.7259$ $X_2^2 = 5.89$ $P = 0.05$		$n = 17$ $p = 0.1057$ $\alpha = 0.7533$ $X_1^2 = 1.67$ $P = 0.20$	

	LAT 19		LAT 20		LAT 21	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	285	285.00	298	298.00	128	128.00
2	312	318.02	360	374.25	242	256.44
3	257	238.61	273	248.69	300	274.68
4	95	108.50	87	91.81	153	163.46
5	39	33.30	16	20.34	62	58.36
6	2	7.27	1	2.70	9	12.50
7	1	1.16	0	0.20	0	1.49
8	0	0.14	1	0.01	1	0.08
9	0	0.01				
10	1	0.00				
	$n = 12$ p = 0.1200 $\alpha = 0.7127$ $X_2^2 = 3.59$ $P = 0.17$		$n = 7$ $p = 0.1813$ $\alpha = 0.7124$ $X_2^2 = 4.38$ $P = 0.11$		$n = 7$ $p = 0.2631$ $\alpha = 0.8570$ $X_3^2 = 5.23$ $P = 0.16$	

	LAT 22		LAT 23		LAT 24	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	341	341.00	228	228.00	373	373.00
2	377	385.55	310	315.76	359	353.22
3	298	280.85	278	280.78	226	233.40
4	106	113.65	149	142.67	89	89.97
5	26	27.60	46	45.31	24	22.29
6	3	4.02	11	9.21	5	4.12
7	2	0.34	1	1.26		
	$n = 7$ $p = 0.1954$ $\alpha = 0.7042$ $X_2^2 = 1.94$ $P = 0.38$		$N = 8$ $p = 0.2026$ $\alpha = 0.7771$ $X_3^2 = 0.82$ $P = 0.84$		$n = 9$ $p = 0.1418$ $\alpha = 0.6533$ $X_2^2 = 0.66$ $P = 0.72$	

Thus the first objective is met: word length in the Latvian language is obviously controlled by the EPB distribution, though in some cases an escaping tendency is perceptible whose future direction is not yet predictable. Preliminarily, it merely has a local character, i.e. it merely concerns some frequency classes.

3.0. Word length in the Lithuanian language (objective 2)

The following texts – being of the same kind as those used for the analysis in Latvian - were taken as the basis for the analysis of word length in Lithuanian:

LIT 1 Misevičius, Vytautas. Karaliūnas gargaliūnas, Vilnius 1987. p. 22-25.
LIT 2 Tamulaitis, Vytautas. Greitutės nuotykiai, Vilnius 1992. p. 5-8.
LIT 3 Miliūnas, Viktoras. Skrisk, žuvėdra, Vilnius 1980. p. 44-48.
LIT 4 Misevičius, Vytautas. Juodojo džentelmeno galas, Vilnius 1981. p. 6-9.
LIT 5 Nekrašius, Jonas. „Fluxus ir laiško menas". In: Literatūra ir menas no. 17 (2845), 2001. p. 9.
LIT 6 Jauniškis, Bronius. Kur lapinas? Vilnius 1991. p. 3-4.
LIT 7 Jauniškis, Bronius. Kur lapinas? Vilnius 1991. p. 6-8.
LIT 8 Kašauskas, Raimondas. Vakaris vėjas. Vilnius 1989. p. 76-79.
LIT 9 Ignatavičius, Eugenijus. Chrizantemų autobuse. Vilnius 1988. p. 5-8.

LIT 10 Ignatavičius, Eugenijus. Chrizantemų autobuse. Vilnius 1988. p. 17-20.
LIT 11 Kašauskas, Stasys. Meilė ir kiti žaidimai. Vilnius 1987. p. 6-10.
LIT 12 Dautartas, Vladas. Senojo gluosnio pasaka. Vilnius 1987. p. 24-27.
LIT 13 Dautartas, Vladas. Senojo gluosnio pasaka. Vilnius 1987. p. 28-32.
LIT 14 Astas, Vydas. Tik svajotojai. Vilnius 1988. p. 49-52.
LIT 15 Šernas, Pranas. „Raganaitė". In: Literatūrinis kultūrinis almanachas "BALTIJA" 1994. p. 45– 48.
LIT 16 Žemgulytė, Paulina. „Senojoje jotvingių žemėje". In: Literatūra ir menas no. 21 (2849) - 2001. p. 1.
LIT 17 Tyruonis, Uldis. „Jūra yra jūra" In: Literatūra ir menas no. 20 (2848) - 2001. p. 6.
LIT 18 Šernas, Pranas. „Raganaitė". In: Literatūrinis kultūrinis almanachas „BALTIJA" (1994). p. 50– 52.
LIT 19 Tumelis, Juozas. „Gegužės 3-iosios konstitucija 1791–2001". In: Literatūra ir menas no. 19 (2847) – 2001. p. 1.
LIT 20 Žemgulytė, Paulina. „Susikalbėjimo vardan". In: Literatūra ir menas Nr. 18 (2846) – 2001. p. 1.

Again the 1-displaced EPB distribtion turned out to be the suitable mechanism controlling word length in Lithuanian (objective 2). The results in detail:

	LIT 1		LIT 2		LIT 3	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	166	166.00	253	253.00	277	277.00
2	214	226.88	278	276.24	279	266.33
3	221	199.80	213	220.00	164	190.03
4	93	93.84	103	93.44	92	72.31
5	17	24.79	20	22.32	11	17.33
6	3	3.49	0	2.85		
7	1	0.21	1	0.15		
	$n = 6$	$p = 0.2605$	$n = 6$	$p = 0.2416$	$n = 6$	$p = 0.2220$
	$\alpha = 0.7678$		$\alpha = 0.7085$		$\alpha = 0.6634$	$C = 0.014$
	$X_2^2 = 5.46$	$P = 0.07$	$X_2^2 = 2.78$	$P = 0.25$		

	LIT 4		LIT 5		LIT 6	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	181	181.00	164	164.00	202	202.00
2	282	276.96	286	262.78	252	258.40
3	196	211.56	182	216.79	219	204.23
4	104	86.19	115	107.31	80	89.67
5	12	19.75	35	35.41	25	23.62
6	3	2.54	12	8.18	3	3.73
7			2	1.52	1	0.34
	$n = 6$	p = 0.2340	$n = 11$	$p = 0.1416$	$n = 7$	$p = 0.2085$
	$\alpha = 0.77$		$\alpha = 0.79$		$\alpha = 0.74$	
	$X_2^2 = 8.04$	$P = 0.02$	$X_3^2 = 10.12$	$P = 0.02$	$X_2^2 = 2.35$	$P = 0.31$

	LIT 7		LIT 8		LIT 9	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	360	360.00	219	219.00	229	229.00
2	414	423.19	290	288.58	273	272.05
3	367	356.44	203	209.65	214	219.51
4	162	150.11	85	87.03	105	94.46
5	20	31.61	29	22.58	17	22.87
6	1	2.66	5	4.16	3	3.11
	$n = 5$	$p = 0.2963$	$n = 8$	$p = 0.1719$	$n = 6$	$p = 0.2440$
	$\alpha = 0.73$		$\alpha = 0.74$	$X_2^2 = 2.26$	$\alpha = 0.73$	$X_2^2 = 2.83$
	$X_2^2 = 6.76$	$P = 0.03$	$P = 0.32$		$P = 0.24$	

	LIT 10		LIT 11		LIT 12	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	289	289.00	212	212.00	303	303.00
2	356	357.13	257	254.24	375	359.33
3	238	246.91	238	245.57	220	248.69
4	103	94.84	134	126.51	98	86.06
5	24	21.86	34	36.66	16	14.89
6	3	3.26	6	6.03	1	1.03
	n = 7	p = 0.1873	n = 6	p = 0.2787	n = 5	p = 0.2571
	α = 0.72	X_2^2 =1.26	α = 0.76	X_2^2 = 0.90	α = 0.70	X_2^2 = 5.73
	P = 0.53		P = 0.64		P = 0.06	

	LIT 13		LIT 14		LIT 15	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	301	301.00	269	269.00	416	416.00
2	423	410.01	292	280.83	473	451.70
3	241	268.13	192	214.78	323	355.41
4	112	93.52	104	91.26	171	169.49
5	15	18.35	24	23.27	66	54.56
6	1	2.00	2	3.87	12	12.49
7					0	2.08
8					1	0.28
	$n = 6$	$p = 0.2074$	$n = 7$	$p = 0.2032$	$n = 12$	$p = 0.1252$
	$\alpha = 0.73$		$\alpha = 0.70$		$\alpha = 0.72$	
	$X_2^2 = 7.92$	P = 0.02	$X_2^2 = 5.57$	$P = 0.06$	$X_3^2 = 7.18$	P =0.07

	LIT 16		LIT 17		LIT 18	
X	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	130	130.00	340	340.00	264	264.00
2	199	192.29	351	343.20	351	362.30
3	201	208.97	241	255.61	275	270.30
4	112	113.54	114	108.78	119	112.03
5	32	30.85	32	28.94	24	27.86
6	5	3.35	4	5.48	7	4.16
7					1	0.36
	$n = 5$	$p = 0.3521$	$n = 8$	$p = 0.1755$	$n = 7$	$p = 0.1992$
	$\alpha = 0.81$		$\alpha = 0.69$		$\alpha = 0.75$	
	$X_2^2 = 1.41$	$P = 0.49$	$X_2^2 = 1.99$	$P = 0.37$	$X_2^2 = 4.10$	$P = 0.13$

	LIT 19		LIT 20	
X	f_x	NP_x	f_x	NP_x
1	173	173.00	125	125.00
2	236	244.69	186	185.80
3	268	237.36	219	212.15
4	116	136.45	136	148.03
5	44	51.47	56	70.43
6	21	13.32	46	24.13
7	1	2.71	4	6.12
8			1	1.35
	n = 10	$p = 0.1773$	$n = 13$	$p = 0.1599$
	$\alpha = 0.80$	$C = 0.016$	$\alpha = 0.84$	
			$X_4^2 = 1.44$	$P = 0.23$

4.0. Conclusions

The answer to the question if word lengths in the Latvian and Lithuanian languages belong to one type is definitely positive: in both languages word length is controlled by the EPB distribution. This means: word lengths in the two Baltic languages belong to one type of distribution (objective 3).

Objective 4 is met as well: in our analysis of word length in the Slavic languages (Rottmann 2003) it was found that all Slavic languages are controlled by one distribution family (cf. Wimmer/Altmann 2003), in which the EPB distribution is by far the one occurring most frequently in Slavic languages. So, the question is also answered positively, Baltic and Slavic languages are controlled by one type of distribution.

References

a) Literature

Endzelin, J. (1911). *Slavjano-baltijskie ėtjudy*. Char'kov.

Holst, J.H. (1911). *Lettische Grammatik*. Buske: Hamburg.

Poljakov, O. (1995). *Das Problem der balto-slavischen Sprachgemeinschaft*. Peter Lang: Frankfurt (Main).

Rottmann, O.A. (2003). *Zur Struktur und Form von Wort und Satz in den slavischen Sprachen – ein Beitrag zur quantitativ-typologischen Analyse* (forthcoming).

Wimmer, G., Altmann, G. (2003). Unified derivation of some linguistic laws (forthcoming).

b) Software for the computation of distributions

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

Glottometrics 6, 2003, 61-64

Age and polysemy of words

Udo Strauß, Gescher[1]
Gabriel Altmann, Lüdenscheid

Abstract. The older a group of words all of which came into existence in the same time interval the greater its mean polysemy, i.e. polysemy increases with time. The dependence will be demonstrated on the unique but slightly distorted data set created by D. Wolff (1972) using the English dictionary.

Keywords: polysemy, English

The problem of the semantic expansion of words with increasing age can be captured by a simple hypothesis but is joined by a very tiresome extraction of data. A good historical dictionary of language is necessary in order to obtain reliable data. However, since no dictionary can exactly indicate the time of birth of all words, the problem must be treated statistically, taking say centuries as time intervals.

A second problem is the fact that not all words behave in the same way. Some of them are more inclined to expand semantically than other ones and some of them have no expansion propensity at all. This propensity need not be associated with the frequency or polytexty or length of the word – though in many cases it is – it can be, so to say, inherent. For example "I" has a great frequency, an enormous polytexty and is extremely short and though it is very old, its polysemy did not change considerably. Thus the link between age and polysemy cannot be the same for every word, but it must hold on the average, i.e. for ensembles of words which came into existence at the same time, e.g. in the same century.

The only investigation known to us is Dieter Wolff´s (1972) analysis of the polysemy of English words showing the results in a peculiar cumulative graphical form, verbal description and cutting the data after polysemy $y > 10$. Thus, even if the following theoretical model may be adequate, the testing by means of Wolff´s data must be considered very preliminary and must be repeated using different, more complete data.

Wolff´s hypothesis "the older a word the more meanings it seems to accumulate" will be modificated here in "the older an ensemble of words that arose in the same time interval, the greater their average polysemy". This ensemble can be composed of any kind of words and its age is ascertained for the available texts. The derivation of the hypothesis can be considered a special case of Wimmer-Altmann´s (2002) "unified theory" and the differential equation can be set up on the following assumptions:

The relative rate of polysemy increase is proportional to the relative rate of time interval change. The intervals are, unfortunately, neither disjoint – nobody can tell whether a word appearing for the first time in texts of the 14th century did not exist much earlier – nor are the words arising in them of the same frequency, length and polytexty. Thus we take into account a ceteris paribus condition which is proportional to the time interval rate. In the *simplest case* – using Occam´s razor – both proportionalities are considered equal and we obtain

[1] Address correspondence to: Udo Strauss, AIS, Schuckertstr. 25-27, D-48712 Gescher, Germany.
E-mail: strauss@medsorga.de

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

yielding the curve

$$(2) \qquad y = ax^b e^{bx}$$

where $a = e^C$, C being the integration constant. Evidently, for $b > 0$ the limit of (2) is infinity but it is impossible to suppose an empirical finite limit. In semantics and lexicology this assumption is not unknown, neither explicitly (cf. Krylov, Jakubovskaja 1977; Piotrowski, Bektaev, Piotrowskaja 1985: 71; Kornai 2002) nor implicitly, even if it is in conflict with the Zipfian idea of self-regulation (cf. Zipf 1949; Köhler 1986). However, no empirical limit is determinable. Wolff notes that there are words with up to 33 meanings and in irregular intervals up to 95. Thus the problem cannot be solved theoretically without setting up a model for each word separately splitting up the ceteris paribus condition into variable factors, which leads to partial differential equations.

Since the data at our disposal do not allow such a procedure we merely test (2) on Wolff´s data shown in Table 1. Here the youngest words (from the 20th century) obtain time index 1, those of the 19th century index 2 etc. The words of the 14th century, those of LME and ME are pooled and obtain time index 7.5; OE obtains index 11. This partitioning and pooling corresponds roughly with the dating of English (cf. Finkenstaedt, Wolff 1973).

Table 1
Wolff´s (1972) data on the age and polysemy of English words

Century or epoch	Time index	Polysemy y									
	x	1	2	3	4	5	6	7	8	9	10
20	1	501	11	-	-	-	-	-	-	-	-
19	2	14365	1822	365	57	13	5	1	-	-	1
18	3	5705	1629	503	130	38	17	1	1	1	1
17	4	9503	3834	1541	493	199	77	23	16	4	2
16	5	7118	3885	1960	918	432	253	136	66	34	20
15	6	1441	946	496	245	151	90	52	33	15	6
14+LME+ME	7.5	4341	3347	2285	1513	968	595	427	269	166	136
OE	11	1163	870	600	464	274	202	151	103	88	69

It can easily be seen that from Old English up to the 16th century the tail of the distribution does not end at $y = 10$. Nevertheless, even these data are sufficient to test (2). If one computes the mean polysemy in each line of Table 1, one obtains the results in Table 2 (second column). By iterative fitting we obtained a curve in which $a \approx 1$, thus at last

$$\hat{y} = x^{0.0928} e^{0.0928x}$$

The values of this theoretical curve are given in the third column of Table 2. The result is graphically shown in Fig. 1. As can be seen, the parameter a is practically 1 and b is about 0.1. If we had complete data, the fit would be still better or b would be somewhat greater because the means of older words would increase.

Table 2
Fitting (2) to the age-polysemy relation in English

Time index x	Observed polysemy mean y	Computed polysemy mean (2) $\hat{y}$
1	1.02	1.09
2	1.16	1.28
3	1.41	1.46
4	1.63	1.64
5	2.03	1.85
6	2.28	2.06
7.5	2.87	2.42
11	3.14	3.47
	$b = 0.0928$, $R = 0.91$	

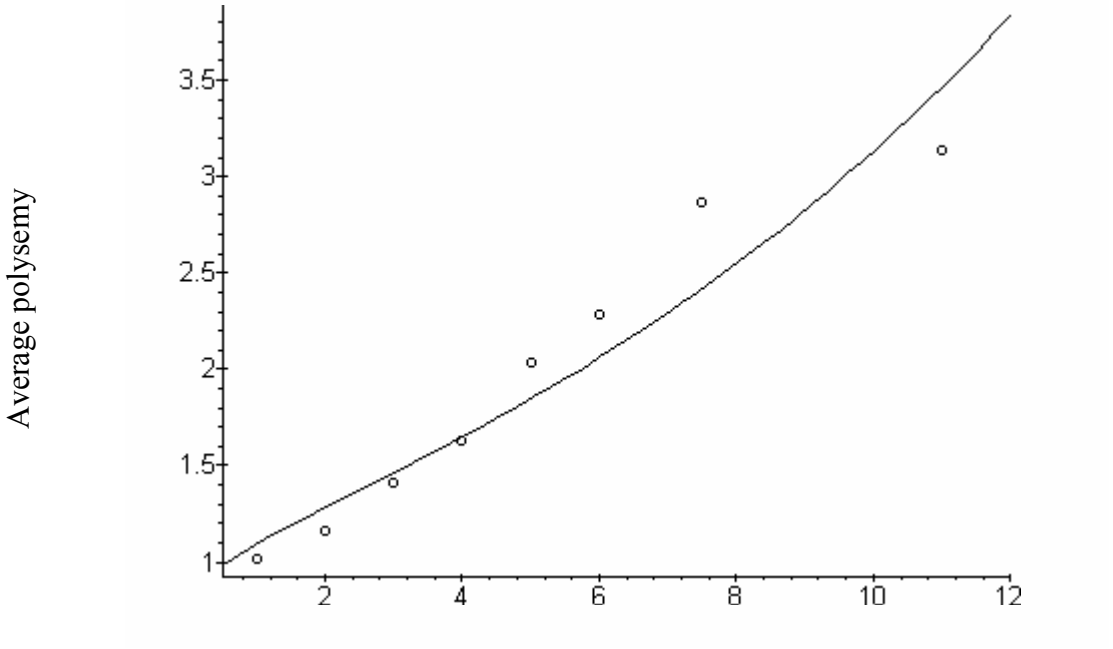

Fig. 1. Empirical and computed values of the age-polysemy relation

The status of this hypothesis is, of course, preliminary but it is realistic and must be tested further.

The usual approach to developmental phenomena using Piotrowski´s law (cf. Altmann, Buttlar, Rott, Strauß 1983) yields slightly better results ($R^2 = 0.97$), too, but only if one uses the so called incomplete change alternative (cf. Altmann 1983)

$$(3) \qquad y = \frac{c}{1 + ae^{-bx}}$$

with a fixed parameter a (since the curve must not begin below $y_1 = 1$). In this case one obtains c (≈ 3) as the asymptote which is not realistic even for the given data. If Wolff´s data

were complete we had even higher values for older words. In addition, the interaction model giving rise to (3) is not applicable in this case.

References

Altmann, G. (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, K.-H., Kohlhase, J. (eds.), *Exakte Sprachwandelforschung: 54-90.* Göttingen: Herodot.

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

Finkenstaedt, T., Wolff, D. (1973). *Ordered profusion. Studies in dictionaries and the English lexicon with contributions by H. Joachim Neuhaus and Winfried Herget.* Heidelberg: Winter.

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

Kornai, A. (2002). How many words are there? *Glottometrics 4, 61-86.*

Krylov, J.K., Jakubovskaja, M.D. (1977). Statističeskij analiz polisemii kak jazykovoj universalii i problema semantičeskogo toždestva slova. *Naučno-techničeskaja informacija Ser. 2, No. 3, 1-6.*

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

Wimmer, G., Altmann, G. (2002). Unified derivation of some linguistic laws. *Paper at the Graz Conference on Word Length, August 2002.*

Wolff, D. (1972). Bedeutungshäufigkeit und ihr statistisches Verhalten. *Beiträge zur Linguistik und Informationsverarbeitung 22, 33-44.*

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

Glottometrics 6, 2003, 65-69

Multidimensional Scaling to Visualize Text Separation

Eric S. Wheeler,Toronto[1]

Abstract. One can count the occurrences of each type of character in a text, and arrive at a text profile. Does such a profile give enough information to separate texts, such as wanted emails from unwanted ones? Multidimensional scaling (MDS) provides a means of visualizing such profile data, so that one can make an informed assessment for a given set of circumstances. By extension, MDS can be applied beyond the problem of text separation.

Keywords: Multidimensional scaling, MDS, text separation, data visualization methods.

Some email is welcome and other email is not. Can we use the quantitative properties of an incoming email either to distinguish between the two cases absolutely, or at least to rate the likelihood of the email belonging to one category or another?

This problem will serve as an example of the more general problem of separating texts according to some criteria. Given a method for separating texts, we advocate using multi-dimensional scaling (MDS) to visualize how well the method is working on a given set of texts, because the MDS technique is visually informative and relatively easy to employ.

1. A simple approach to quantifying a text is to count the occurrences of each character type in the text. For example, we can categorize the characters by their ASCII or Unicode representation, giving (in the ASCII case) 256 different types, corresponding to the values from 0 to 255. The result is a profile for a given text that has 256 dimensions. The profile can be normalized by dividing each count by the total number of characters in the text (see Table 1). Hence, texts of any size can be compared one to another.

As one might expect, our initial counts on sample texts showed a large number of ASCII codes with no occurrences, and a few ASCII codes that accounted for much of the text (cf. Wheeler 2002). Blanks, line feeds, and carriage returns are frequent, reflecting the overall shape of the text. Upper and lower case letters are distinct, possibly allowing us to capture the more frequent use of upper case ("shouting" in email) in one kind of text or another.

At the same time, this measure is independent of the order of presentation (an "A" at the beginning counts as much as an "A" in the middle or at the end), and therefore will give the same measure for texts that are mere rearrangements of other texts. It does not require sophisticated parsing of the language, and only relies on information that in some sense is below the conscious control of the author. Whether or not that proves to be useful, this method stands in contrast to (say) the methods of anti-spam programmes that parse for known expressions, vocabulary and constructs believed to indicate "junk" email. It is our opinion that the authors of junk email learn to adapt to these anti-spam programmes, but cannot adapt easily to the kinds of text properties that our proposed method measures.

[1] Address correspondence to: Eric S. Wheeler, 33 Peter Street, Markham, Ontario, Canada L3P 2A5.
E-mail: wheeler@wheeler-and-young.on.ca or ewheeler@yorku.ca

Table 1
A portion of the profile of a sample text

ASCII code	Character type	Occurrences	Occurrences normalized
60	<	13571	.7 %
61	=	14820	.7 %
62	>	27364	1.4 %
63	?	4191	.2 %
64	@	9686	.5 %
65	A	10730	.5 %
66	B	4468	.2 %
67	C	7959	.4 %
68	D	8543	.4 %
69	E	11002	.5 %
70	F	6048	.3 %

ASCII code 60 represents the character "<". It occurred 13571 times in the text, which is 7 times per thousand characters or 0.7%.

2. With a method for quantifying texts (the one proposed here, or any other), one can separate texts by establishing a boundary that divides a known test set correctly. Future emails can then be measured, placed on one side of the boundary or the other, and hence dealt with accordingly.

For instance, with our method, we have measured email logs for emails that were saved over a period from 1999 to 2002, and emails that were disposed of in 2002 (with a conscious effort to only dispose of emails that were unwanted. Welcome emails that were no longer needed were filed separately). The result is 6 profiles for welcome email, and 1 for unwanted email.

In theory, future emails that have a similar profile to the 6 can be kept, and those that are similar to the unwanted email can be disposed of immediately.

However, what is the boundary between these profiles? Impressionistically, it appears that the junk email has many more blanks than wanted email. At the same time, there are many characters that do not appear at all or only infrequently (in either of the kinds of text). So, one might look for a boundary defined in terms of the few characters that seem to differ most. For example, the rule might be: "If an incoming text has more than 10% blanks, dispose of it. Otherwise, keep it."

Or, we could see the 256 numbers of the profile as a point in 256-space, and imagine a geometric figure, such as a sphere, around a given point. Anything within the sphere belongs to the point. But are either of these constructs the real distinguishing boundary?

Also, one might ask whether the boundary is an absolute dividing line, or do the two kinds of text overlap. (Consider the parallel case: men are typically taller than women, but nonetheless, there are some women who are taller than most men, and some men who are shorter than most women. What height boundary distinguishes men from women?). If the texts overlap, by how much do they overlap?

Finally, we must allow for the possibility that the two kinds of text are not distinguished by the property we are measuring. But how do we discover this easily?

3. Multidimensional scaling (MDS) is a statistical technique to "systematize data by representing the similarities of object spatially as in a map" (Shiffman et al. 1981: xv). We have used the technique with some success to represent large, complex sets of dialect data in readily viewed maps (Embleton and Wheeler 1997a, 1997b, 2000).

In essence, MDS takes an *n*-dimensional space and reduces it by one dimension in such a way as to minimize the distortion created. Compare the shadow of your hand (a 2-dimensional image) to the hand itself (a 3-dimensional image). One can see that a certain orientation of the hand will produce a shadow which best represents the hand without hiding the relative positions of the fingers and thumb. Other positions would be less clear.

Starting with a high-dimensional space, MDS repeats this process until there is a 2-dimensional map (or sometimes a 3-dimensional image) of the original data, and in some sense it is the "best" representation. For our purpose here, that is not critical, but when required, it is possible to get a numeric measure of the amount of stress introduced by the process.

MDS is not the only statistical technique that can reduce data in high dimensions to small dimensional spaces. Shiffman et al. (1981: 13-14) compare MDS to Factor Analysis. While MDS works with the distances between points, Factor Analysis works with the angles between vectors. In most of its procedures, Factor Analysis assumes a linear relationship between variables. "The MDS approach does not contain this assumption, and the result is that it normally provides more readily interpretable solutions of lower dimensionality" (13).

For us, the value of MDS is that it readily produces a 2-dimensional picture of the points from a high-dimensional space. Using this picture, we can easily look for a boundary to separate two subsets of points, and get an intuitive grasp on whether or not the points separate readily.

4. We applied the MDS procedure to the sample emails described above. First, we created a distance matrix between points, based on the square of the Euclidean distance between the normalized coordinates of the points in 256-space. For points $X = (x_i)$ and $Y = (y_i)$:

$$\text{distance}(X, Y) = \sum_i (x_i - y_i)^2 \quad \text{for } i = 0 \text{ to } 255$$

There is no need to take the square root to get a useful metric.

The MDS procedure in the SAS online statistics package (Shiffman et al. (1981) describe a number of other such packages) gives the following coordinates. Points **a** to **f** are the welcomed emails; the Trash point is the unwanted email (cf. Table 2).

Table 2

Point	**Dimension 1**	**Dimension 2**
a	-0.44612	0.094678
b	1.92095	1.314669
c	-0.23406	-0.71237
d	-0.08026	-0.46194
e	0.607701	-0.07722
f	-2.16815	1.003454
Trash	0.399927	-1.16126

The data, plotted by the graphics component of a spreadsheet, is in Figure 1. I have added a putative boundary line to separate the welcome and the unwanted emails.

With a picture of the data, it is relatively easy to see that there may be a separation between the two sets of text. However, it is also equally easy to see that the putative boundary may not hold up under further examination. For example:

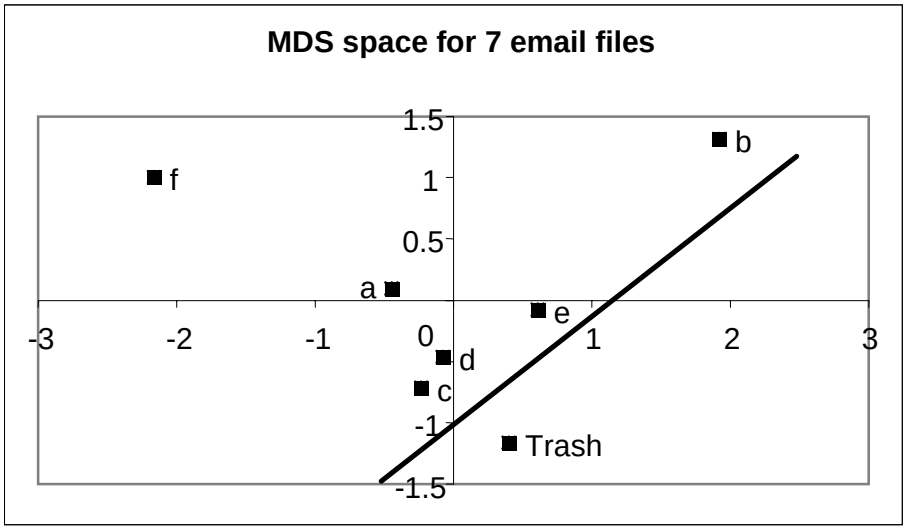

Figure 1. Six email logs of welcome email, and 1 log of unwanted email

- The Trash point is much closer to points c, d, and e than are the points b and f.
- The boundary could have been drawn in a variety of positions, such as horizontally through the point (0,-1).
- The "good" points extend over a large area, while the Trash point has just one position, and that position is close to the other texts.

Even if we believe there is a divide, it is not clear how broad an area the Trash point covers, and how much overlap, if any, exists between the two text types.

It is the picture produced by applying MDS to the profile data that allows us to quickly see the state of our analysis. Without the picture, it is entirely unclear whether or not the texts are separate. With the picture, one can see the possibilities, and also the next steps to confirming (or not) the possibilities.

5. The email logs are large files. In order to get a sense of the spread covered by this data, we divided the files into 72, equal-sized subsets (not necessarily on the boundaries of a particular email) and calculated the resulting points again.

The new numeric coordinates do not match those from the earlier test because the MDS process produces the "best" projection that it can, and that varies when the number and position of the points changes. However, the points themselves are consistent with the earlier picture.

In this new view, the "welcome" emails group in two places along one direction (we omitted from the chart a few of the extreme outliers that were even further out.). The "trash" emails all grouped closely together, between the two welcome groups. Figure 2 shows the result.

Clearly, there is something distinctive about the unwanted emails. They group together, and not with the welcome emails, but the boundary between them is not simple. Two lines (as drawn in) might be the distinguishing criteria, but we might do well to test some more before we relied too heavily on that hypothesis. Nonetheless, the two types of text do separate on the basis of the measures we have taken.

That is a conclusion that can be seen directly from the MDS picture of the data, even though it is not in the form that we anticipated, and not in a form that would be easy to arrive at from a simple examination of the numbers.

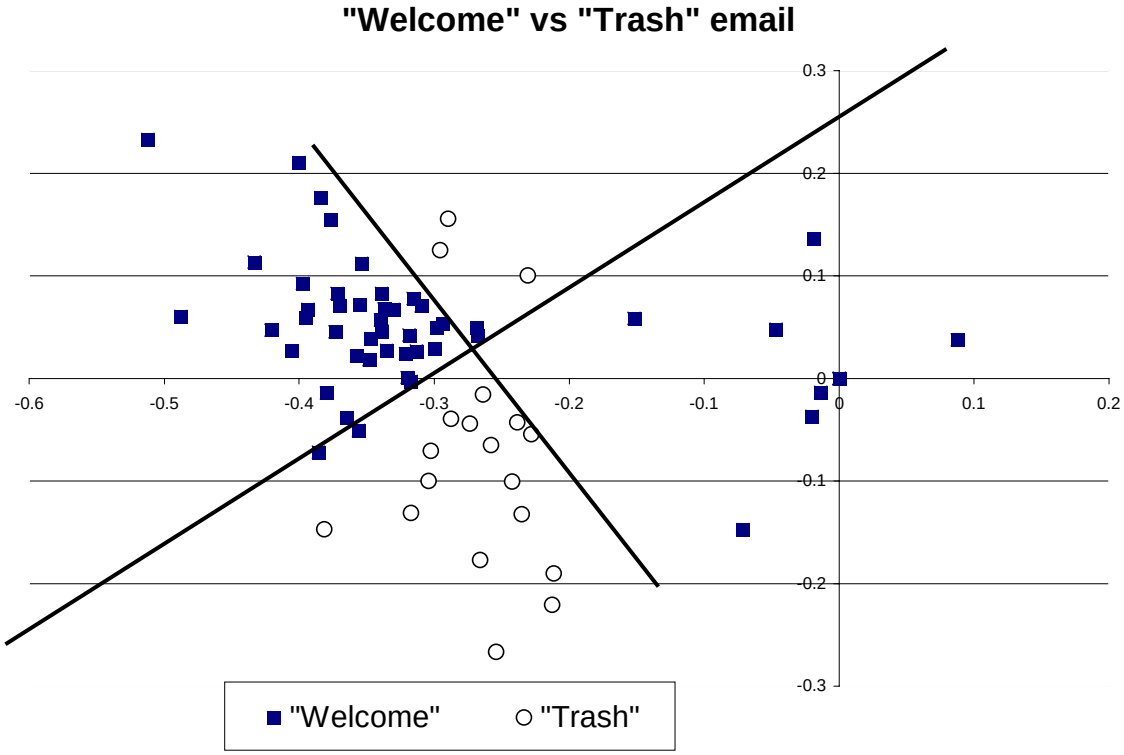

Figure 2. Files 1 to 53 of "Welcome" email, and files 54 to 72 of "Trash" email, with two boundary lines added

Yet, the MDS procedure does not force us to make assumptions about the data. The data is still the data, and the investigator is free to interpret the data appropriately.

6. Our modest conclusion is, first, that the method of profiling a text by counting characters is interesting because it is simple to do, and yet it produces a usable quantification of a text.

More significantly, the MDS method readily reduces large data sets with multiple dimensions to a 2-dimensional visualization of the data − and hence allows a researcher to "see" hypotheses, to reason about them, and to communicate that reasoning to a (perhaps less technical) audience, easily and quickly.

References

Embleton, Sh.M., Wheeler, E.S. (1997a). Multidimensional Scaling and the *SED* Data. In: W.Viereck, H. Ramisch (eds.), *The Computer Developed Linguistic Atlas of England 2, 5-11.* Tuebingen: Max Niemeyer Verlag.

Embleton, Sh.M., Wheeler, E.S. (1997b). Finnish Dialect Atlas for Quantitative Studies. *J. of Quantitative Linguistics 4, 99-102.*

Embleton, Sh.M., Wheeler, E.S. (2000)**.** Computerized Dialect Atlas of Finnish: Dealing with Ambiguity. *J. of Quantitative Linguistics 7, 227-231.*

Schiffman, S.S., Reynolds, M.L., Young, F.W. (1981). *Introduction to Multidimensional Scaling. Theory, Methods, and Applications.* New York: Academic Press. 411pp.

Wheeler, E.S. (2002). Zipf's Law and Why It Works Everywhere. *Glottometrics 4, 45-48.*

Glottometrics 6, 2003, 70-82

Python for linguistics?

Ralf Jüngling, Portland[1]
Gabriel Altmann, Lüdenscheid

Abstract. The present article tries to point out the advantages offered by the programming language Python in solving simple computational and quantitative linguistic problems. Several examples illustrate the features of Python and show its simplicity.

Keywords: Python, quantitative linguistics, computational linguistics

Introduction

As a matter of fact, today no linguistic discipline can do without the mechanical aid represented by the computer. The amount of data placed at our disposal by the Internet, by text collectors or by machine-readable dictionaries is so overwhelming that manual processing is rather illusory. In this situation, the Hamletian question is not "to compute or not to compute" but merely "with which progamming language?" Programming languages come and go and once one has learned to use one of them one adheres to it – there are more important things to be learnt. However, newer languages remove the weak points of older ones and enable us to do more in shorter time and with less effort, i.e. with less programming. All of them have their advantages and disadvantages; they can be very general or specialized for particular tasks.

In the following we shall describe the rather recent language *Python* which seems to be very suitable for processing *any* problem arising in computational and quantitative linguistics. Since just the enumeration of possible tasks would take a whole article we restrict ourselves to some simple ones illustrating the elegance and power of Python.

The programming language Python was created in the early nineties by the Dutch computer scientist Guido van Rossum. Although affiliated with several different institutions during the years, van Rossum has led the development of the language until today. The main language features that make programming in Python so productive are not unique to Python but are shared by a family of programming languages for which the name s*cripting languages* was coined (Ousterhout 1998). We explain these features shortly in the next section. Other prominent scripting languages are for instance *Perl*, *Tcl* and *Ruby*.

Although grown up in Perl's shadow, Python possesses a strong community as well. Its users appreciate their language of choice for several outstanding features, i.e. its ease to grasp syntax, its compact but well readable expressions, its arsenal of modern language constructs which facilitate to follow advanced software design principles and a variety of high quality special purpose libraries and extensions. For these reasons we believe that among today's scripting languages Python is best suited for quite a number of scientific disciplines where programming has become a daily task to "get the job done" in reasonable time. In what follows we demonstrate how to exploit Python to get the job done in computational or quantitative linguistics. The problem domains we will highlight by example are

[1] Address correspondence to: Ralf Jüngling, juenglin@cse.ogi.edu

- acquisition of text from sources like databases and internet (e.g. the Gutenberg projects maintain archives of copyright free texts; the English Gutenberg project can be found at http://promo.net/gb, the German Gutenberg project at http://www.gutenberg2000.de)
- converting text into basic entities (sentences, words, syllables)
- determining statistic characteristics of text entities

Some features of Python

The name "scripting language" mistakenly suggests, that these languages are merely sufficient for simple tasks. While this is true for their ancestors, Unix shells which provide limited capabilities for automating system administration and file management tasks, Python and its relatives can be regarded as general purpose programming languages.

As the most important aspect, scripting languages are *interpreter languages* rather than *compiler languages* (like e.g. C, C++, also called "System programming languages" by Ousterhout (1998)). In order to execute a program, an interpreter language needs an *interpreter* installed on the target computer which executes each expression of the program at runtime, line by line. The main drawback attributed to interpreter languages is a lower speed of execution comparative to compiler languages. On the other hand, writing programs takes much less time. The first reasons why interpreter languages are more productive is a shortened "program-try-correct cycle" because there is no compilation step as in compiler languages. The second is that the interpreter can be used in an *interactive mode*: if one does not remember syntax details like e.g. the number or the order of arguments to a function, one can look them up quickly by "trying it out in the interpreter". Assume, for instance, that we forgot the name of the function which computes the arcus tangens, but remember that it is contained in the *module*[2] "math"*:*[3]

[2] Any piece of Python source code contained in a single file is a *module* in Python parlance. The contents of a module can be accessed by *importing* them via Python's IMPORT statement.

[3] The token ">>>" is Python's "prompt"; whenever it is shown, it indicates that we print an extract of a "dialogue" with the interpreter in interactive mode. In order to make the first steps with Python, we recommend to novice users to type the statements next to the prompt directly into the interpreter to experience how such a "dialogue" works. (When doing so, one does not have to type in the comments, of course.)

```
>>> # Note: any text preceded by a '#' will not be interpreted --  we use this to comment the code!
>>> import math              # we import the “math” module
>>> dir(math)                # we ask to see the names of everything that is contained in the module
['__doc__', '__name__', 'acos', 'asin', 'atan', 'atan2', 'ceil', 'cos', 'cosh', 'e', 'exp', 'fabs', 'floor',
'fmod', 'frexp', 'hypot', 'ldexp', 'log', 'log10', 'modf', 'pi', 'pow', 'sin', 'sinh', 'sqrt', 'tan', 'tanh']
>>>                # there are two likely candidates: ‘atan’ and ‘atan2’; we will check them
>>> help(math.atan)
Help on built-in function atan:

atan(...)
   atan(x)
      Return the arc tangent (measured in radians) of x.

>>> help(math.atan2)
Help on built-in function atan2:

atan2(...)
   atan2(y, x)

   Return the arc tangent (measured in radians) of y/x.
   Unlike atan(y/x), the signs of both x and y are considered.

>>> # sometimes using the interpreter in interactive mode is even sufficient to find the answer:

>>> print "A day has %d seconds." % (24*60*60)   # ‘%d’ is a number-placeholder within a string
A day has 86400 seconds.
```

Scripting languages are said to be *dynamically typed* or *typeless* in contemporary computer science parlance. It simply means that the *type* of a datum (e.g. “int” for the integral number 42) is not tied to a variable but to the datum itself. Variables are merely names temporarily assigned (*bound*) to data. Thus it is not necessary to spend code on variable declarations, one simply uses a variable − without declaration − by assigning a datum to the desired variable name:

```
>>> a = 42                      # now ‘a’ denotes the integral number 42
>>> type(a)                     # what is the type of the datum referred to by ‘a’?
<type 'int'>

>>> a = "I am a sentence."      # now ‘a’ denotes a character string
>>> type(a)
<type 'str'>
```

The typeless property is related to the fact that scripting languages typically completely overtake *memory management*. As an example, if a datum in the interpreter’s memory gets “nameless”, the interpreter knows that it is no longer of interest to a program and that its associated memory may tacitly be reused for other data:

```
>>> a = "I am a sentence."
>>> b = a                       # a datum may have more than one name
>>> c = "I am another one."
>>> a = c = 1                   # the names ‘a’ and ‘c’ got re-assigned to the number ‘1’
>>>                             # now the string “I am another one.” has silently been
>>>                             # forgotten by the interpreter while “I am a sentence.”
>>>                             # is still accessible through variable ‘b’
>>> print a, c, b
1 1 I am a sentence.
```

Scripting languages have powerful *abstract data types* built-in. Most often used in Python programs are *lists* and *dictionaries*:

```
>>> l = [1, 2, "carrera", 3.1415]          # lists may contain items of any type
>>> type(l)
 <type 'list'>>>> print l[0]               # list items are accessed by 'indexing',
1                                          # the first item has index 0
>>> b = l[2]; print b                      # several expressions may be separated by ';'...
a string
>>>                                        # note that there are two variables currently referring to
>>>                                        # "a string": 'l[2]' and 'b'
>>> l[2] = 3; print l, b
[1, 2, 3, 3.1415000000000002] carrera

>>> d = {}; type(d)                        # 'd' is a dictionary; it may also contain data of any type
<type 'dict'>                              # but unlike lists, they can be indexed by arbitrary indices
>>> d["fred"] = "Simpson street 217"       # here we use a dictionary to associate strings to strings
>>> d["geraldo"] = "Garlekin avenue 16"
>>> print d                                # unlike in lists, items in a dictionary are not ordered
{'geraldo': 'Garlekin avenue 16', 'fred': 'Simpson street 217'}
>>> # note that the interpreter used other quotation marks to indicate a string; in fact both are valid
```

Scripting languages offer rich *libraries* of ready-to-use code for special purposes:

```
>>> # we want to retrieve a certain HTML page from the internet; provided we are online…
>>> import urllib                          # … all we need is contained in the module 'urllib':
>>> url = urllib.urlopen("http://www.python.org")  # we connect to the web server …
>>> htmltext = url.read()                  # … and fetch the HTML page
```

Scripting languages are easy to enrich by *extensions.* To understand what an extension is and in what it is different to a library, remember the distinction between compiler and interpreter languages: Because their level of machine abstraction is in general lower, programs written in compiler languages usually are much faster. The Python interpreter itself is written in the compiler language C. A consequence of this is that certain *built-in operations*, like e.g. seeking for a character-sequence within a string are actually as fast as in a compiled program. Python falls behind in execution time as soon as the problem at hand cannot be solved by a single (or a few) built-in operations. When execution speed does matter (e.g. in numerical calculations or in processing huge amounts of text), rather than to switch entirely to a compiler language, one has the option to enrich the Python language by new problem-specific built-in operations. These new operations are brought in by "extensions" which are again modules, but written in a compiler language. The point here is, that it is not necessary to change the interpreter in order to enrich the language.

As an example, there is an extension which enriches Python by basic matrix and linear algebra operations similar to the core capabilities of matrix languages like MATLAB, IDL or PV-Wave. Other examples are extensions providing access to "Graphical User Interface (GUI) Toolkits". Typically extensions actually come along with some Python code, thus as a mixture of ordinary Python modules and compiled extension modules.

Python at work

As already mentioned, Python code is organized in *modules*. All one needs to create a code module is a simple text editor.[4] A module may contain a program which can be brought to

[4] A word processor is not a suitable editor, because it interweaves text formatting codes with the actual text.

execution by typing

```
python myprogram.py
```

at the system's command shell. This command invokes the interpreter which in turn tries to interpret as Python source code (that means: to execute) the data found in file "myprogram.py".

But a module may also contain a collection of function definitions and constants. The rationale for collecting functions in such "non-executable modules" is code re-use: It frequently happens that one writes a program to solve a problem which is similar to a problem already solved (i.e. another program was written to solve it). Typically the program as well will look similar to the one already written. It is good practice in such cases to identify and extract the similar portions of the program to move into another module which serves as a library. To illustrate this idea, we dive into Python programming by presenting some short piece of code from our own "misc" module, which we extracted from our example code presented in subsequent sections[5]:

```
 1  """Collection of miscellaneous auxiliary functions
 2  """
 3
 4  from operator import add
 5
 6  def sign(x):
 7      """Sign of number
 8
 9      sign(x) -> -1/0/1
10      """
11      if x<0:
12          return -1
13      elif x>0:
14          return 1
15      else:
16          return 0
17
18  def sum(l):
19      """Sum of a list of numbers
20
21      sum(l) -> l[0]+l[1]+...+l[-1]
22      """
23      return reduce(add, l)
```

There is already much to learn about Python within these few lines:

In line 4 we use the statement IMPORT to import the function "add" from another module "operator", because we need it somewhere in the module (in line 23).

Lines 6 and 18 each contain a DEF statement, which is used to define a function. All code below the head of a function definition which is indented by one or more indentation levels, makes up the "function body". The function body is a piece of code which is executed on invocation of the function. A function may expect an arbitrary number of arguments (including none), and may give back a result via RETURN. To become concrete again, we define two functions in lines 6 to 23, "sign" and "sum". The purpose of the function "sign" is to determine and give back the sign of a number; it expects a single argument which is denoted by "x" in the function body. The purpose of "sum" is to add up all numbers within a list; its argument is denoted by "l".

[5] The line numbers are not part of the code but an aid when explaining it.

While we believe the code of "sign" can easily be read and understood, the code of "sum" surely cannot without an understanding of the function "reduce"[6]. We want to draw attention to the following points:

- In Python indentation is used to determine the grouping of code into code blocks.
- Unlike in typed languages, the author of a function has no control over the kind (type) of argument a user passes to it. As an option, he can add code that checks the argument's type at runtime, though.
- It shows good practice to briefly explain the purpose and usage of a function in the *doc-string* of the function. A doc-string is an optional means to add little documentation to the code; if present, it must precede the actual code within the function body (above, the doc-string of "sign" spans lines 7 to 10 and that of "sum" lines 19 to 22). As shown above, a module can also have a doc-string (lines 1-2).

Now the functions "sign", "sum" and "add" can be imported from module "misc":

```
>>> from misc import sum
>>> help(sum)
Help on function sum in module misc:

sum(l)
   Sum of a list of numbers

   sum(l) -> l[0]+l[1]+...+l[-1]

>>> sum([1, 2, 3])                    # this should yield 6
6
>>> "ABC" + "DEF"                     # adding strings means concatenating them
'ABCDEF'
>>> sum(["What", "a", "surprise"])    # thus, this should yield "Whatasurprise"
'Whatasurprise'
>>>
>>> sum(["What", 5])                  # is adding of strings and integers defined?
Traceback (most recent call last):
  File "c:\python\for\linguistics\misc.py", line 23, in sum
    return reduce(add, l)
TypeError: cannot concatenate 'str' and 'int' objects
```

Here comes a surprise: our aim was to write a function which adds up all numbers contained in a list, but we actually wrote a function which allows adding up whatever can be added in Python (i.e. for whatever the "+" operator is defined). This is an example of the strength of the typeless property and the seasoned Python programmer would have written another doc-string (e.g. "Sum of items in sequence") rather than checking the type of the items contained in the argument in order to refuse adding non-numeric items[7]. But this style of programming needs some getting used to, since there are surprises of the other kind also:

```
>>> from misc import sign
>>> sign(12)                          # this should yield 1
1
>>> sign("another surprise")          # this should abort with an error message
1
```

We now turn to concrete code examples from the linguistic domain.

---

[6] Exercise: Ask the Python interpreter about „reduce".

[7] Note that we even received a meaningful error message when we attempted to add a string and an integer.

Frequency dictionary

Perhaps the most popular problem in computer linguistics is the setting up of a frequency dictionary. At the first glance the problem seems to be very simple but linguists know that nothing is so imperfect and coarse like a mechanically set up frequency dictionary! Unfortunately, theoretical consequences are drawn and testing of models is performed just based on problematic data of this kind. However, here we shall not care for the linguistic background. We read a text, segment it into "words" and order the words according to a sorting key, e.g. alphabetically, by decreasing frequency, by word length etc. At the beginning of the program we let three possibilities to read in the text (a) from a saved file, (b) from the system's "standard input" and (c) from the Internet − either alternative can be used.

```
"""Determine word frequencies in texts.
"""

from misc import sign, printDict          # we need the functions "sign" and "printDict" below

text = file("mytext.txt").read()          # a: read in text from file "mytext.txt"

from sys import stdin                     # b: read in text from standard input
text = stdin.read()

from urllib import urlopen                # c: read in text from the internet
url = ftp.ibiblio.org/pub/docs/books/gutenberg/etext00/7ljw110.txt
text = urlopen(url).read()

for character in '.,;!?\'\t\n"/([{}]}§$%&':   # we eliminate all spurious characters,
    text = text.replace(token, ' ')
allwords = filter(None, text.split(' '))  # break up the text into words,
allwords = map(str.lower, allwords)       # and turn all words to lowercase

frequency = {}                            # we build up the frequency dictionary by use of a
for word in allwords:                     # python dictionary "{}" – don't get confused, this is a
    if word not in frequency:             # name coincidence!
        frequency[word] = 1
    else:
        frequency[word] += 1

words = frequency.keys()                  # at first we sort the list of word
words.sort()                              # alphabetically with the built-in sort method
print "Dictionary sorted alphabetically:"
printDict(frequency, words)               # we print the dictionary in the established order

def compareByFrequency(word1, word2, freq=frequency):
    return sign(freq[word2] - freq[word1])

words.sort(compareByFrequency)            # then we sort by decreasing frequency
print "Dictionary sorted by frequency:"
printDict(frequency, words)

def compareByFrequencyAndLength(word1, word2, freq=frequency):
    return sign(freq[word2]+len(word2) - (freq[word1]+len(word1)))

words.sort(compareByFrequencyAndLength)  # finally we sort by frequency and word length
print "Dictionary sorted by frequency and length:"
printDict(frequency, words)
```

Some more comments on the code are in order. We use a simple rule to identify the words of

a text in line 17: any character sequence surrounded by space tokens (‘ ‘) and which itself does not contain a space token, is a word. To make this simple algorithm work, we have to prune all non-letter characters (i.e. punctuation, quotation marks and special characters like new-line, '\n', or a tabular space, '\t') beforehand (line 15f). What we get in line 17 (and denote by “allwords”) is a list of strings; the function FILTER eliminates all empty strings, which may stem from adjacent space tokens in the input text. In the loop line 21ff we go through all the words, check for each whether we have seen it before (line 22) and count accordingly. After this loop, we are actually done – the rest of the code is concerned with producing output.

Our function “printDict” (invoked in lines 30, 37 and 44) prints the contents of the first argument (a dictionary) as a table, sorted in the order of the items in the second argument (a list). For the first output table, we sort the words in the dictionary (“frequency”) alphabetically (by the built-in method SORT of the list “words”, see line 28).

We need to sort in different, “non-standard” ways. However, this is easy to achieve by telling the SORT method how to compare two items of the list: The function “compareByFrequency”, defined in lines 32f yields 1 if the first argument “word1” has a higher frequency (according to our frequency dictionary) than the second argument “word2”, -1 if “word2” has a higher value than “word1” and 0 if both arguments have equal frequency. Given this comparison function, SORT knows how to sort the list by frequency (line 35).

With the following piece of text (the sonnet „To make my days impatient with unrest“ by R.S. Hillyer (1895-1961) that can be found in Internet (http://www.sonnets.org/hillyer.htm)) as input,

```
To make my days impatient with unrest,
To filch the quiet of the darks repose,
Seeking forever what my soul well knows
Is ever far beyond my farthest quest;

So this is love; swift joys and lingering woes,
A wistful kiss beneath the ashen west,
Farewell and greeting, mouth to mouth once pressed,
And then the empty darkness onward flows.

The heights that I have won do not endure,
They shrink beneath the stars I yearn to win,
The triumphs of my passion only lure

My vagrant feet to tread the verge of sin;
Though well I know that when I fall thereover,
Love will fly hence; the loved and the lover.
```

the output of the program looks like this (large parts omitted):

```
Dictionary sorted alphabetically:
a               :1
and             :4
ashen           :1
beneath         :2
beyond          :1
darkness        :1
.
.
.
woes            :1
won             :1
yearn           :1
```

```
Dictionary sorted by frequency:
the             :10
my              :5
to              :5
and             :4
i               :4
of              :3
.
.
```

The Type-Token relation

The type-token analysis examines the increase of new words (units) in the text, a problem connected with the flow of information in text. E.g. in didactical texts the flow is slower than in poetic texts. The program presented below takes a text and determines two quantities: X = the position of the word in text, Y = the number of different words up to position X. There is a great number of theoretical curves that can be fitted to this course. The program merely prepares the data and takes word forms into account. For the above example text, the type-token analysis yields the following result:

```
to                1  1
make              2  2
my                3  3
days              4  4
impatient         5  5
.                 .  .
.                 .  .
.                 .  .
hence           107 79
the             108 79
loved           109 80
and             110 80
the             111 80
lover           112 81
```

In the second column the counter (X) is increased with every word or token. In the third column the counter (Y) is increased by 1 only if the given word did not yet occur.

```
 1  """Type-Token relation
 2  """
 3
 4  text = file("mytext.txt").read()
 5
 6  for token in '.,;!?\n\t\'"/([{}]}§$%&':
 7      text = text.replace(token, '')
 8  allwords = text.split(' ')
 9  allwords = map(str.lower, allwords)
10
11   types, x, y = [], 0, 0
12   for token in allwords:
13       x = x + 1
14       if word not in types:
15           types.append(word)
16           y = y + 1
17       print "%-15s%3d%3d" % (word, x, y)
```

Lines 6 to 9 of the program appeared identically in our previous program and is actually a

perfect piece-of-code candidate to become factored out into our library module. Unlike in the previous program, where we used a dictionary to count up the word frequencies, here a list ("types") suffices to memorize each encountered word (lines 14 and 15).

The PRINT statement in line 17 demonstrates fancy output generation with the % operator. The % operator takes two arguments – a left-hand side and a right-hand side – and yields a string. Its left-hand side argument is a format string which consists in general of normal text and "placeholders" for data, the right-hand side argument is a tuple of variables (the data) to be embedded in the format string; the number of placeholders on the left must meet the number of variables on the right. The format string in line 17 for instance contains three placeholders: "%-15s" for a string and two times "%3d" for two integers. The numbers in the placeholders are optional and cause the string substituted for the placeholder to have a minimum size of 15 respectively 3 characters, which yields the tabular-like appearance of the output.

Calculating statistics

In Python also the computation of statistical characteristics for a series of numbers is extremely simple. The next program shows an example.

```
 1  """Basic statistics
 2  """
 3
 4  from math import sqrt                    # we make use of the functions "sqrt"
 5  from misc import sum                     # and "sum" below
 6
 7  data = [2,5,8,6,3,10,1,0,5]              # for simplicity we write the data directly down here
 8  data = map(float, data)                  # we make sure that the numbers are real
 9  n = len(data)                            # n is the number of items in the list
10
11  data.sort()                              # we sort the numbers
12  min = data[0]                            # to easily determine the minimum,
13  max = data[-1]                           # the maximum
14  median = data[n/2]                       # and the median of the data
15  range = max – min
16  mean = sum(data)/n
17  variance = sum([(d-mean)**2 for d in data])/(n-1)
18  stddev = sqrt(variance)
19
20  print "data      : %s" % data           # we use the % operator again to control the output
21  print                                    # formatting
22  print "min       : %g" % min
23  print "max       : %g" % max
24  print "range     : %g" % range
25  print "median    : %g" % median
26  print "mean      : %g" % mean
27  print "var       : %g" % variance
28  print "stddev    : %g" % stddev
```

Converting HTML to raw text

Some publicly available electronic text archives provide the texts only as HTML. This is meant as a convenience for the readers, because one needs nothing more than a web browser to read the texts. The linguist, however, needs the text in a raw form, i.e. without any HTML

tags contained. We show the simplest way to convert HTML to raw text with Python. In this example we encounter a bit of what is called *object oriented programming* (OOP). While we have no room to give even a sketchy introduction into OOP, the example can be understood with a few explaining remarks.

```
1  """html2txt converts html to raw text
2  """
3  from htmllib import HTMLParser
4
5  class MyHTMLParser(HTMLParser):
6
7     def do_hr(self, attrs):  pass
8     def handle_image(*args): pass
9
10    def anchor_end(self):
11       if self.anchor: self.anchor = None
12
13
14 def html2txt(htmltext):
15    """Convert HTML to raw text
16
17    html2txt(htmltxt) -> rawtext
18    """
19    from formatter import AbstractFormatter, DumbWriter
20    from StringIO import StringIO
21
22    sbuffer = StringIO()
23    f = AbstractFormatter(DumbWriter(sbuffer))
24    p = MyHTMLParser(f)
25    p.feed(htmltext)
26    rawtext = sbuffer.getvalue()
27    p.close()
28
29    return rawtext
```

The new concept associated with the CLASS statement in line 5 is that of a *class.* A class is another way of assembling code, which conceptually belongs together, into a single entity. Organizing code in a class allows somewhat more flexibility than organizing it in a module. However, these concepts do not rule out each other: in the example above, a class is contained in a module, and in the Python standard library there are many modules containing even several classes (which belong together conceptually).

There are merely two new peculiarities of classes that need to be understood:

1. Class code can be partially reused by *inheriting* it.
2. The class code has to be "started up", before it can be used[8].

In the above example we import a class "HTMLParser" from the standard library module "htmllib" (line 3). If the "HTMLParser"-code would do exactly what we want – i.e. going through a HTML text and collecting all headings and body text – we could simply use it. However, it does only almost exactly what we want. Now we could copy the module file "htmllib.py" to, say "myhtmllib.py" and alter the code to make it behave as we need it. Because the code is organized in a class, there is a more economical way: In line 5 above we declare to inherit the code (or better: the behaviour) of "HTMLParser" in our class "MyHTMLParser". In the class definition (spanning from line 5 to line 11) of

[8] In OOP parlance one creates an *instance* or *object* of the class; it amounts to initialising a data structure.

"MyHTMLParser" we define three functions. These functions (and many more) were already defined in class "HTMLParser" we inherit from, but because their behaviour does not fit our needs, we substitute them in "MyHTMLParser" by ones that do. The two functions "do_hr" and "handle_image" simply do nothing (this is the purpose of the "pass" keyword) – thus we virtually "switch off" these functions inherited from "HTMLParser". The function "anchor_end" on the other hand is a slight derivation of the version in "HTMLParser".

In lines 22 to 24, we "start up" four classes: "StringIO", "AbstractFormatter", "DumbWriter" and "MyHTMLParser". The result of starting up a class is an *object* of that class. Since the "HTMLParser" needs a "formatter object" as an argument (and likewise, the "AbstractFormatter" needs a "writer object" and the "DumbWriter" needs a "file object"), we import the classes in line 19 and 20 and create the desired objects.

From the lines above, the purpose of the several classes we import and how they work cannot be inferred. What shall be demonstrated here is, that functionality for dealing with HTML text is contained in the standard library. All we do is picking out the necessary parts and compose them appropriately in the function "html2txt" which serves the only purpose to convert HTML to raw text[9]: it expects a string of HTML text as input and yields a string − the raw text − as output.

Conclusion

As a conclusion one might draw from the last example, learning a modern programming language is twofold: on the one hand it means learning the language (i.e. the language concepts and its syntax), on the other it means becoming familiar with its standard library. In case of Python we recommend to take a textbook to learn the language and to use the online documentation to look up what is in Python's standard library. Be aware and keep in mind, that the library source code is also part of the documentation – the online documentation just describes the purpose, if one needs to know how something works, one looks into the code.[10]

What we brought in the examples above are just the simplest of the day-to-day tasks in linguistics. E.g. statistical analysis usually involves much more sophisticated methods, and analysis results shall be visualized. There are interesting extensions and libraries available via the internet and even efforts to build a library for linguistics are underway. We resisted the temptation to employ those extensions here, for we wanted to demonstrate what can be done with the Python distribution alone. However, readers seriously interested in using Python for their own work should consult the internet sources we give in the appendix.

Finally we want to remind that Python and most Python software spread via internet stems from voluntary efforts and are made available for others' benefits. While authors of such software are in general pleased to get feedback from users, they usually are not willing (nor do they have the time) to give support to users. Therefore, in case of problems, it is good practice to try to get help from the *community* which usually grows around any open-source software and which typically maintains a mailing list or a newsgroup in the internet.

---

[9] A limitation with this "parser approach" is intolerance against HTML formatting errors. A tolerant approach was to prune anything from the text that looks like an HTML tag, e.g. by use of "regular expression" (see module "re").

[10] For instance, we looked into the "HTMLParser" code to determine which functions we need to substitute in "MyHTMLParser".

Internet Sources

Python can be obtained from the Python homepage, http://www.python.org/.

Python-Teaching material (books/courses):

- http://www.python.org/doc/Newbies.html
- http://www.python.org/cgi-bin/moinmoin/IntroductoryBooks
- http://greenteapress.com/thinkpython.html

Some extensions/libraries for Scientific Computing ...

- http://www.scipy.org
- http://sourceforge.net/projects/numpy
- http://matpy.sourceforge.net/
- http://www.biopython.org

Interfaces to GUI-Toolkits …

- http://www.wxpython.org
- http://www.daa.com.au/~james/software/pygtk/
- http://www.riverbankcomputing.co.uk/pyqt/index.php

Software for linguists…

- http://nltk.sourceforge.net/
- http://www.cogsci.princeton.edu/~wn/links.shtml#Python

References

Articles:
Ousterhout, J.K. (1998). Scripting: Higher Level Programming for the 21st Century, *IEEE Computer*, March 1998.

Introductory texts on the Python language (for people with some programming experience):
Harms, D., McDonald, K. (2000). *The Quick Python.* Manning Publications.
Lutz, M., Ascher, D. (1999). *Learning Python.* O'Reilly.

Other, non-introductory texts on the Python language:
Martelli, A., Ascher, D. (2002). *Python Cookbook.* O'Reilly.

Glottometrics 6, 2003, 83-93 *To honor G. K. Zipf*

On a Zipf's Law Extension to Impact Factors

Ioan-Iovitz Popescu[1]

Abstract. The Lavalette's law is further promoted with empirical arguments from its original area of impact factors of scientific journals. Alike its famous precursory Zipf's and Mandelbrot's rank-frequency laws, the Lavalette's law offers the promise of various applications also beyond its original meaning. Thus, an alternate reduced rank-frequency distribution is introduced by assigning equal ranks to the words with the same frequency. Also the fractal behavior of self-similarity of actual rank-frequency curves belonging to different scales is revealed.

1. Introduction to Zipfian laws

As it is well known, the Zipf's law is an empirical law set up for linguistics in the early 1930s by the Harvard linguistic professor George Kingsley Zipf (1902-1950). This heralded the power law $q(n) \propto 1/n$, now commonly called Zipf's law, which states that the frequency q of occurrences of some event (such as of a word in a text sample) is inversely proportional to its rank n. As often happens, there are forerunners, as displayed in a time table of bibliometrics by Ronald Rousseau (2001). Actually, G. K. Zipf (1935, 1949) originally described a broad statistical regularity of natural languages and proposed two complementary empirical laws of word frequencies, as highlighted by Landini (2000), namely:

1. "The *rank-frequency* law. This is the most famous one; unfortunately many people call it "Zipf's law" as if it was the only one. […] The procedure to estimate this relation is very simple: the words in a text are sorted by decreasing frequency and a rank number is assigned to each word. For words with the same frequency, the sub-sorting and ranking is arbitrary. The plot of log (frequency) versus log (rank) approximates a straight line of slope -1."

2. "The *number-frequency* law. […] The plot of log (frequency) versus log (number of words with the same frequency) approximates a straight line of slope -0.5. While the rank-frequency law tends to occur for the high frequency words (although not necessarily for the first few ranking positions), the number-frequency law is observed for the low frequency words."

Let us first discuss the Zipf's rank-frequency law as currently expressed by the more general power-law function

$$q(n) = c\, n^{-b}$$

with the scaling constant $c = q(1)$ and the exponent b close to unity ($b = 1$ in the original Zipf's

[1] Address correspondence to: P.O. Box MG-18**,** RO-077125 Bucharest-Magurele, Romania. Email: iovitz@pcnet.ro. Cf. http://www.geocities.com/iipopescu/

expression). In other words, the rank-frequency data should lay on a straight line with slope -*b* when plotted in a double-logarithmic log(*n*), log(*q*) graph. Generally, *q*(*n*) can be any quantity used in ordering a set of occurrences, such as the frequency of natural or randomly generated words, size of cities or other settlements, income size, frequency of access to web sites, size of oil and other mineral deposits, earthquake magnitudes, galactic intensities, up to genetic ranking for cancer classification. Indeed, there is an impressing list of natural and social phenomena revealing a Zipf's power-law behavior (Li 2003). However, the explanation, modeling and meaning of this mysterious law represents a permanent intellectual and interdisciplinary challenge from Zipf's times up to the present days (Laherrère 1996; Laherrère, Sornette 1998; Landini 1997/ 2000; Li 1998, 2002; Manrubia, Zanette 1998; Marsili, Zhang 1998; Powers 1998; Redner 1998; Troll, beim Graben 1998; Tsallis, de Albuquerque 2000, Altmann et al. 2002; Debowski 2002).

Alternately, the law can be expressed as well by the probability. Thus, defining the *text length* (*L*) by the *total number of running words* of the considered text, the ratio $p(n) = q(n)/L$ represents the probability to find the word with rank *n*. For instance, in the English language, the probability of encountering the n^{th} most common word is given roughly by $p(n) = 0.1/n$ for *n* up to about 1000, or better by (Weisstein 2003)

$$p(n) = 1/[n \ln(1.78N)]$$

where *N* is the *vocabulary size*, i.e. the *total number of different words* of the given text. However, the simple hyperbolic Zipf's law $q = c/n$ cannot hold generally true and breaks down for less frequent words or when the vocabulary increases indefinitely since the harmonic series diverges. Indeed, we have the constraint that the probabilities $p(n) = q(n)/L$ must sum to 1, inasmuch as the frequencies $q(n)$ sum to *L*. From here results the above divergence assertion, since summing over this probability distribution gives a non-convergent series. Therefore, faster converging probability distributions have to be used to model Zipf-like distributions in this limit, such as the *Riemann zeta function*, ζ, defined by the series

$$\zeta(b) = \sum_{n=1}^{\infty} n^{-b}$$

converging for $b > 1$ (but diverging for $b \leq 1$).

One of the earliest extensions of the Zipf's law, intended to account for the observed typical downward deviation of the higher-ranked words, has been performed by Benoit Mandelbrot (1954). This well-known mathematician of *fractals* (a term coined by him in 1975) modified the original Zipf's law $q(n) = c/n$ in the form

$$q(n) = [(N + \rho)/(n + \rho)]^{(1+\varepsilon)}$$

containing three adjustable empirical corrections to estimate, namely, a slight correction (already added above) to the power 1, which became the exponent $(1 + \varepsilon)$, a number ρ added to the rank *n*, and the size *N* of the vocabulary of the considered text. All these three parameters *N*, ρ and ε depend on the text length and, for very large texts, $0 < \varepsilon << 1$ and $0 < \rho < 10$ (Debowski 2002).

The interest in the Zipf's law formulation has also been rejuvenated by Laherrère (1996), Laherrère, Sornette (1998), Redner (1998), Tsallis, de Albuquerque (2000) and others. Thus, the

main results of the studies addressed to citations of publications (Redner), or to citations of authors (Laherrère, Sornette) were that the stretched exponential

$$q(n) \propto \exp[-(n/n_0)^{\beta}]$$

fits reasonably well the data for relatively small n-values. However, the needed asymptotic behavior to fit actual data is the inverse power law $q(n) = c\, n^{-b}$ with $b = 3$, a shape which can not be provided by the exponential. Better results have recently been obtained (Tsallis and de Albuquerque 2000) with a function of the power-law type, namely

$$q(n) \propto [1 + (b-1)^{-1}\lambda n]^{-b}$$

with the exponent $b = 2.89$ close to the previous one.

2. The Lavalette's law

In the following we will be concerned with Lavalette's extension of the Zipf's law and its excellent fitting with actual data of journal impact factors. This is a new ranking power-law established by the French biophysicist Daniel Lavalette (1996), barely more complex than the Zipf's law, $q(n) = c\, n^{-b}$. Actually, the role of n as independent variable is taken by the ratio $n/(N - n + 1)$ between the descending and the ascending ranking numbers. Finally, Lavalette's law states that the impact factor q (in the role of *frequency*) of a set of N scientific journals, ordered by the descending ranking number n, obeys the relationship

$$q(n) = c\, [Nn/(N - n + 1)]^{-b}$$

with two fitting parameters, namely the exponent b and the scaling constant $c = q(1)$. Fig. 1 shows the normalized Lavalette function $q(n)/q(1) = [Nn/(N - n + 1)]^{-b}$ as represented by three different plots, namely linear (top), semi-logarithmic (middle), and double-logarithmic (bottom). Perhaps the linear plot could be confused with a Zipf's curve, but the semi-logarithmic graph follows a characteristic *sigmoidal* S-shape which by no means can be provided by the Zipf's law. The downwards deviation from the Zipf's straight line at higher-ranked words in a double-logarithmic diagram appears striking, too. Obviously, when viewed on a $\log[Nn/(N - n + 1)]$, $\log(q)$ plot, the relationship is linear with slope $-b$, and precisely this property allowed Lavalette to guess and test his law. Fig. 2 schematically summarizes the essential features of the competing distributions presented above: Zipf's, Mandelbrot's, Laherrere's, Tsallis/de Albuquerque's, and Lavalette's distribution. Note that also log-normal functions naturally bend in a convex form in a double-logarithmic plot.

Actually, empirical Zipf curves follow only roughly a straight line with slope $-b$ on a double-logarithmic graph, excepting the words of the low end (with highest ranks) when the actual data drop off quite steeply. Also the frequency of the most frequent words (with lowest ranks) do not necessarily follow as fast as expected by the original Zipf's law, that is proportional to 1, 1/2, 1/3, 1/4, and so on. A typical double-logarithmic rank-frequency plot and its Lavalette fitting for 917 distinct words (i.e. vocabulary) out of 7404 running words (i.e. text length), occurring in the text of the USA Constitution, are given in Fig. 3. For this purpose it will be instructive to discriminate

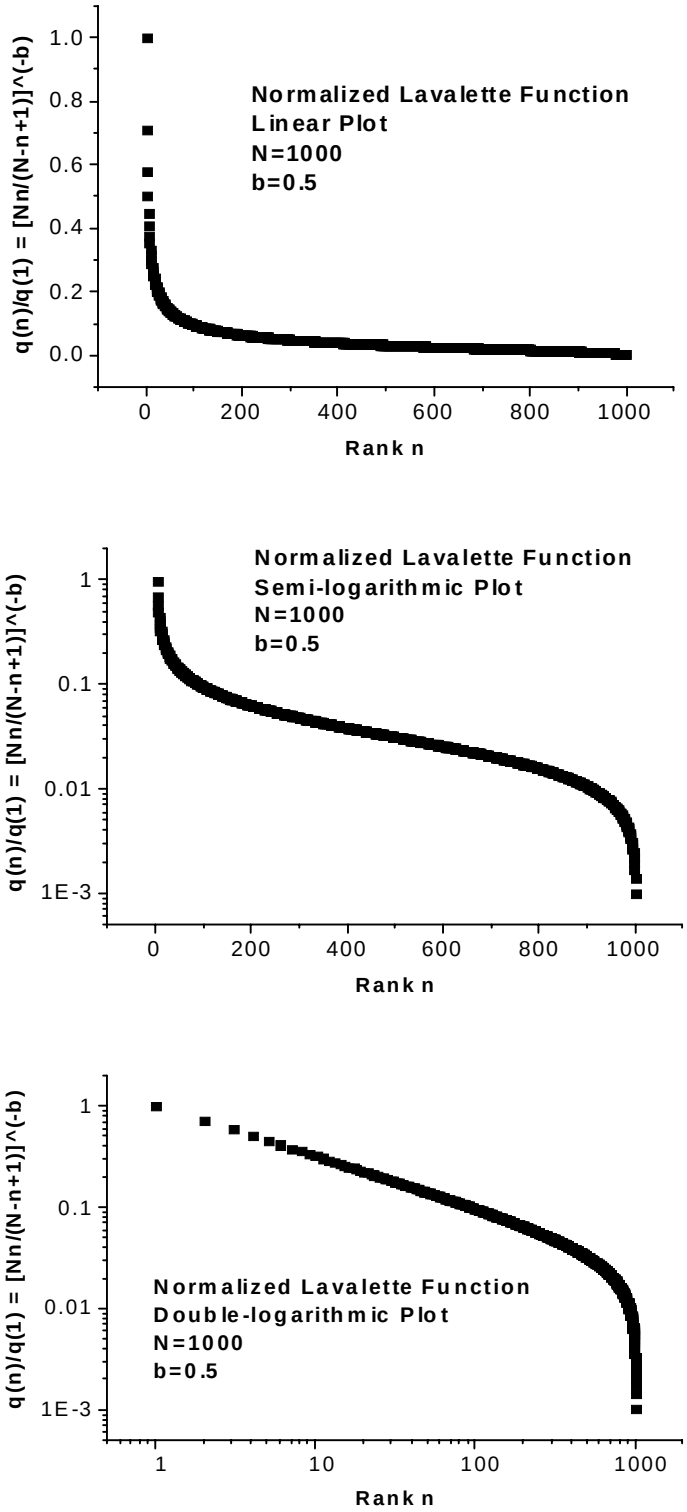

Fig.1 A typical normalized Lavalette function $q(n)/q(1) = [Nn/(N–n+1)]^{–b}$ for N=1000 and b = 0.5 in linear (top), semi-logarithmic (middle) and double-logarithmic (bottom) plot.

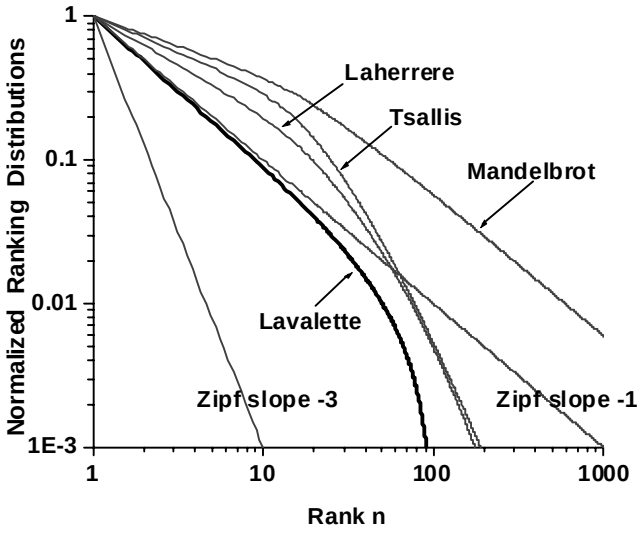

Fig.2. Illustrating essential shapes of competing ranking distributions: Zipf's, Mandelbrot's, Laherrère/Sornette's, Tsalis/de Albuquerque's, and Lavalette's (thicker) curve. Note that also log-normal functions naturally bend in a convex form in a double-logarithmic plot.

between two possible rules concerning the ranks, namely allotting the ranks either *distinctly* or *equally* to the words with the same frequency. Consequently, we have to consider two types of rank-frequency distributions as illustrated in Fig.3, that is:

1. The *ordinary rank-frequency distribution* (upper curve in Fig.3) by *assigning distinct ranks to the words with the same frequency* (ranking within frequencies being otherwise arbitrary, e.g. alphabetical). In a double-logarithmic scale this leads to a slight convex bending and broadening towards the low end of the ranking distribution, a shape that is characteristic for any text and contributes much to the illusion of a general linear decrease. Though the deviation from the Zipf's law for the higher-ranked words is still a matter of controversy (Li 1998), the convex bending is, however, almost always manifest, as we highlighted also in this case with the help of a Lavalette fitting. As usual, the meaning of *N* in the ordinary distribution is the *total number of different words* (the vocabulary).

2. The *reduced rank-frequency distribution* (lower curve in Fig.3) by *assigning equal ranks to the words with the same frequency*. Obviously, the result of this rank rearrangement is a pronounced downwards bending of the ranking distribution, yet very well fitted by a Lavalette function. Also *N* means in this new remodeled ranking the *total number of different frequencies* occurring in the vocabulary spectrum when sorted by counts. From now on the link to the complementary *number-frequency* Zipf's law is straightforward and the result is shown in Fig.4 by a plot of log (frequency) versus log (number of words with the same frequency), approximating a straight line with negative slope.

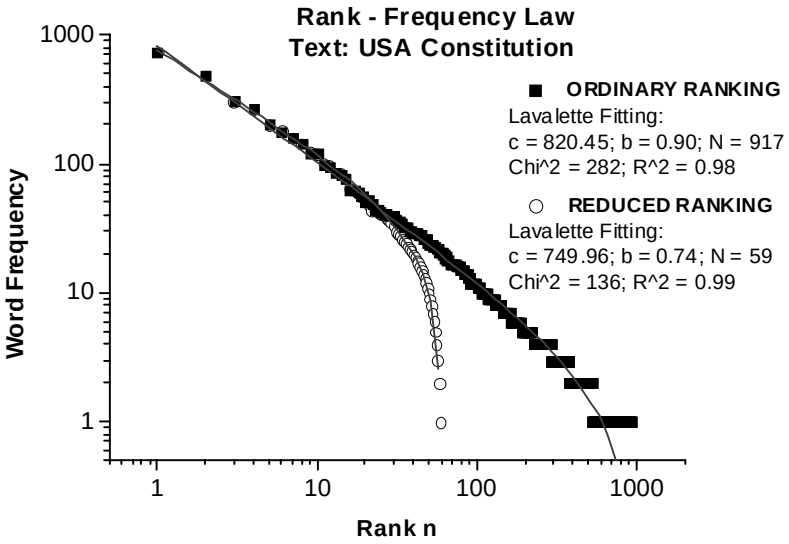

Fig.3 Illustrating ordinary and reduced rank-frequency distributions and their Lavalette fitting for the text of the USA Constitution, vocabulary size = 917 words, text length = 7404 words. Notice the earlier higher-ranked distribution bending of the reduced ranking as compared with the ordinary ranking. *N* means total number of different words for the ordinary ranking and total number of different frequencies for the reduced ranking.

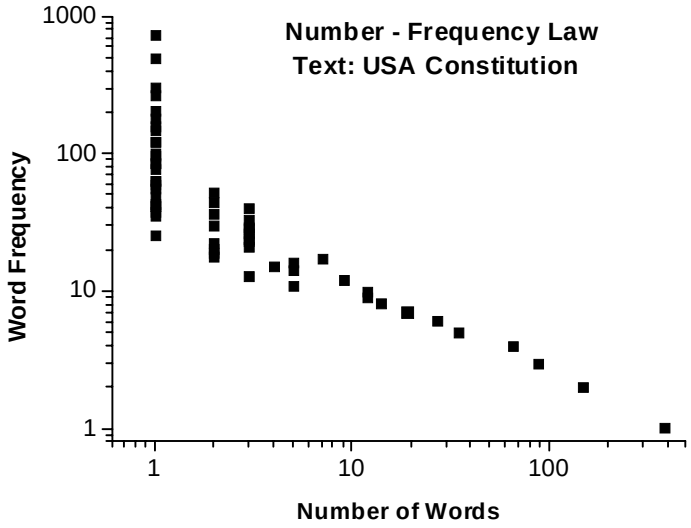

Fig.4 Illustrating the word number-frequency distribution for the text of the USA Constitution, vocabulary size = 917 words, text length = 7404 words.

3. Further arguments and conclusions

Returning to the *rank-frequency law*, we will apply the same analysis to the *rank-impact law* as originally proposed by Lavalette (1996). As already shown in Fig. 3, the Lavalette fitting can be achieved with the help of a single two-parameter (*b* and *c*) function along the entire range of frequency count. If necessary, also *N* can be used as a third tuning parameter in order to complete the missing data and to fix the needed set size. Empirical arguments for Lavalette's distribution were previously illustrated in an *addendum* on the Lavalette ranking law to the web-article (Popescu 2002) for journals ranked by average impact factors and sorted by *scientific fields*, by *title initial letters*, or by *uniform random* sub-sorting. The main conclusion is that the Lavalette's distribution appears the best suited to fit the impact factor data among all the competing functions of Fig. 2. In the present article an updated impact factor database will be used, as gathered in Popescu (2003), for a further empirical support of the Lavalette's extension of the Zipf's law. For this purpose, Fig. 5 and Fig. 6 illustrate the ordinary and the reduced rank-impact distributions and their Lavalette fitting for the average impact factors of a whole set of 8011 scientific journals, respectively of an arbitrary subset of 1018 scientific journals with given title initial (here the letter A by mere chance preferred). In this case *N* means the total number of journals for ordinary ranking and the total number of different frequencies for reduced ranking.

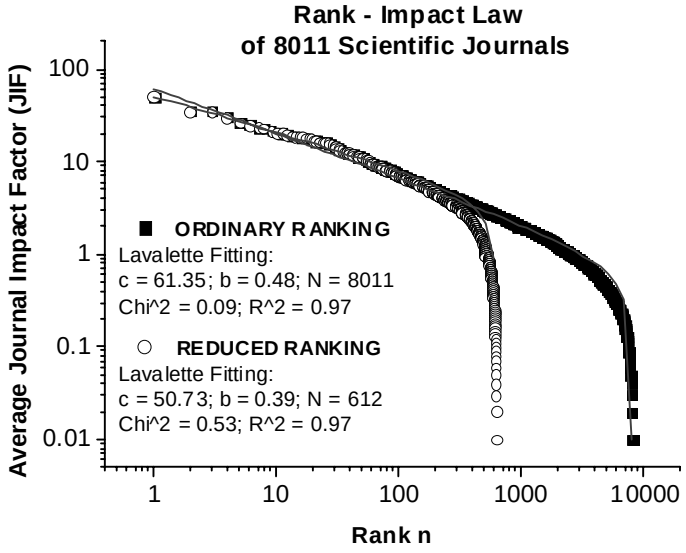

Fig.5 Illustrating the ordinary and the reduced rank-impact distributions and their Lavalette fitting for the average impact factors (JIF) of a set of 8011 scientific journals. Notice the self-similarity of the rank-impact curves of Fig. 5 and Fig. 6. *N* means total number of journals for the ordinary ranking and total number of different frequencies for the reduced ranking. For a direct link to the used impact factor database look at Popescu (2003), where JIF = average journal impact factor over all years of ISI quotation (1974-2001) and ISI = Institute for Scientific Information (http://www.isinet.com).

One may conclude that perhaps the major feature of actual rank-frequency or rank-impact curves of various subsets is that these look the same on any scale, including the curve describing the whole set. The striking fractal behavior of functional self-similarity of Lavalette's curves is non-trivial, as it is the case with Zipf's straight lines, and again the name of Mandelbrot and of

his fundamental books on fractals should be recalled (Mandelbrot 1977, 1983; 1997). Self-similarity is clearly manifest in actual data whenever one compares the Lavalette distributions of subsets between them or with the whole set distribution, as proven by the pair of Fig.5 (whole set of 8011 journals) and Fig.6 (subset of 1018 journals having title initial letter A). Self-similarity is further illustrated in Fig.7 for the ordinary ranking curves of the whole set of 8011 SCIENCE journals and of three successive subsets of 609 PHYSICS journals, out of which 85 OPTICS journals, and out of which 44 journals containing the phonemes OPT in the title. Obviously, if the ordinary curves are self-similar, the reduced ones, not shown in this figure, are likewise. Notice that the initial coalescence of OPTICS and OPT curves is caused by the coincidence of the first few ranking positions. Also the self-similarity is rather approximate than perfect and the statistics gets poorer and poorer according as we magnify by successive sub-sorting. Generally, from a massive empirical evidence one may conclude that self-similar Lavalette's rank-frequency or rank-impact distributions govern the ranking of any kind of sub-sorting. Moreover, alike its famous precursory Zipf's and Mandelbrot's laws, Lavalette's law offers the promise of various applications also beyond its original meaning of merely citation frequency.

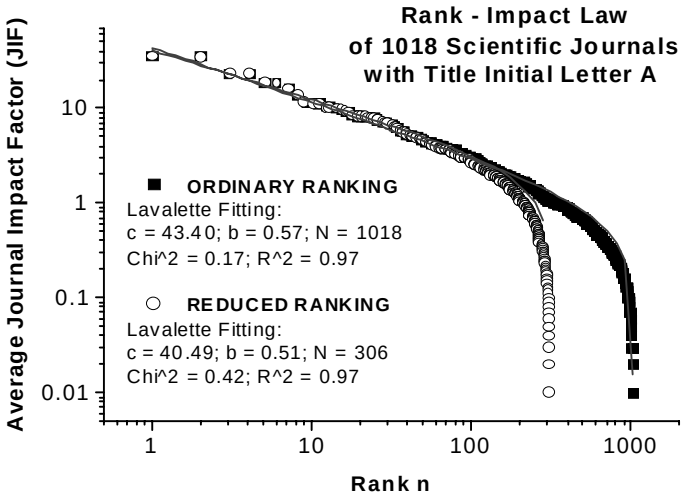

Fig. 6. Illustrating the ordinary and the reduced rank-impact distributions and their Lavalette fitting for the average impact factors (JIF) of a subset of 1018 scientific journals with the same title initial (here letter A), out of a whole set of 8011. Notice the self-similarity of the rank-impact curves of Fig. 5 and Fig. 6. *N* means total number of journals for the ordinary ranking and total number of different frequencies for the reduced ranking. For a direct link to the used impact factor database look at Popescu (2003), where JIF = average journal impact factor over all years of ISI quotation (1974-2001) and ISI = Institute for Scientific Information (http://www.isinet.com).

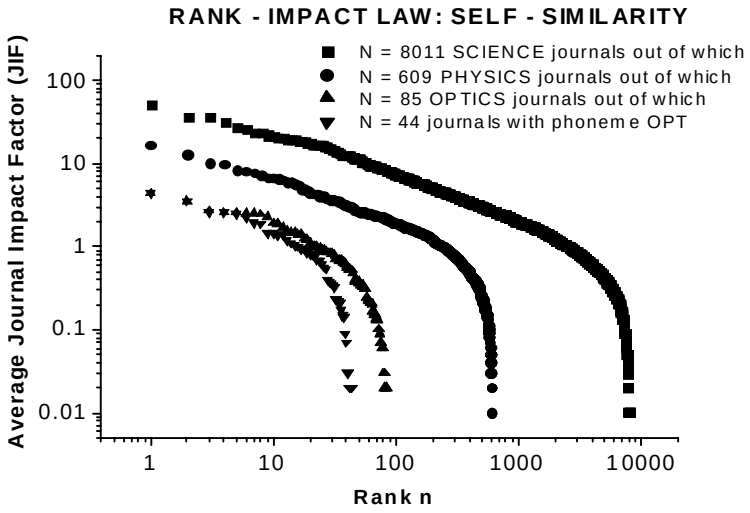

Fig. 7. Illustrating the self-similarity of ordinary rank-impact curves for the average impact factors (JIF) of the whole set of 8011 SCIENCE journals and of its three successive subsets of 609 PHYSICS journals, out of which 85 OPTICS journals, and out of which 44 journals containing the phonemes OPT in the title. The reduced rank-impact curves, not shown in this figure, are self-similar likewise. Obviously, the initial coalescence of OPTICS and OPT curves is caused by the coincidence of the first few ranking positions. For a direct link to the used impact factor database look at: Popescu (2003), where JIF = average journal impact factor over all years of ISI quotation (1974-2001) and ISI = Institute for Scientific Information (http://www.isinet.com).

Acknowledgments. The author is highly grateful to Professors Gabriel Altmann, Daniel Lavalette, and Mircea Oncescu for their interest in this work. Hearty thanks are also due to Dr. Magdalena Nistor and Drd. Sorin Vizireanu for their valuable help in computers and homepage. Since my first stage in Germany as a Humboldt Dozentenstipendium fellow (October 1967 – March 1969), I have always been pleased to acknowledge the *Alexander von Humboldt-Foundation* for generous donations and computer facilities.

References

Altmann, G. et al. (ed.) (2002). *Glottometrics 3, 4*, volumes dedicated *"To Honor G. K. Zipf"* at his 100th birthday anniversary. Lüdenscheid: RAM-Verlag. http://www.ram-verlag.de/

Debowski, L. (2002). Zipf's Law against the text size: A half-rational model. *Glottometrics 4, 49-60.*

Laherrère, J. (1996).”Parabolic fractal” distributions in Nature. *Comptes Rendus de l'Académie des Sciences, Série II a, 322, n.7, 535-541.*

Laherrère, J., Sornette, D. (1998). Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *European Journal of Physics B, 2, 525-539*

Landini, G. (1997, 2000). Zipf's laws in the Voynich Manuscript. http://web.bham.ac.uk/G.Landini/evmt/zipf.htm

Lavalette, D. (1996) *Facteur d'impact: impartialité ou impuissance?* Internal Report, INSERM U350, Institut Curie - Recherche, Bât. 112, Centre Universitaire, 91405 Orsay, France (November 1996), see URL http://www.curie.u-psud.fr/U350/

Li, W. (1998). Comments on "Zipf's Law and the Structure and Evolution of Languages" by Tsonis A.A., Schultz C., Tsonis P.A., (1997). *Complexity 2(5), 12-13* (letter to the editor), *Complexity 3(5), 9-10*, see URL http://linkage.rockefeller.edu/wli/pub/comp98_zipf.html

Li, W. (2002). Zipf's Law in Importance of Genes for Cancer Classification Using Microarray Data. http://linkage.rockefeller.edu/wli/pub/

Li, W. (2003), W. Li's references on Zipf's Law. http://linkage.rockefeller.edu/wli/zipf/

Mandelbrot, B.B. (1954). Structure formelle des textes et communication: deux études. (Formal structure of texts and communication: two studies). *Word 10, 1-27.*

Mandelbrot, B.B. (1977, 1983). *The Fractal Geometry of Nature.* San Francisco: Freeman. Section 38, *Scaling and Power Laws without Geometry*. For a comprehensive bibliography visit Math Archives at URL http://archives.math.utk.edu/topics/fractals.html and the Spanky Fractal Database at URL http://spanky.triumf.ca/www/welcome1.html

Mandelbrot, B.B. (1997). *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk.* Berlin: Springer.

Manrubia, S.C., Zanette, D.H. (1998). Intermittency model for urban development. *Physical Review E 58, 295.*

Marsili, M., Zhang, Y.-C. (1998). Interacting Individuals Leading to Zipf's Law. *Physical Review and Letters 80, 2741.*

Popescu, I.-Iovitz, Ganciu, M., Penache, M. C., Penache, D. (1997). On the Lavalette Ranking Law. *Romanian Reports in Physics 49, 3-27.*

Popescu, I.-Iovitz (2002). Science Journal Ranking by Average Impact Factors. http://www.geocities.com/iipopescu/Jo_rankingb.htm; Addendum on the *Lavalette Ranking Law*, http://www.geocities.com/iipopescu/Jo_rankingb.htm#references

Popescu, I.-Iovitz (2003). Direct links to databases used for the graphs of the present article: http://www.geocities.com/iipopescu/USA_Constitution_Word_Frequency.xls (for Fig.3 and Fig.4) and http://www.geocities.com/iipopescu/Jo_rankingb.htm (for Fig.5, Fig.6, and Fig.7).

Powers, D.M.W. (1998). Applications and Explanations of Zipf's Law. In: *New Methods in Language Processing and Computational Natural Language Learning*, ACL, pp 152-160, http://www.uia.ac.be/conll98/pdf/151160po.pdf

Redner, S. (1998). *How* popular is your paper? An empirical study of the citation distribution, *European Journal of Physics B 4, 131-134.*

Rousseau, R. (2001). *Bibliometrics Timetable (Ronald Rousseau)* . For major links of interest in bibliometric research see the website http://apollo.iwt.uni-bielefeld.de/mw/bibliometrics/

Troll, G., beim Graben P. (1998). Zipf's law is not a consequence of the central limit theorem. *Physical Review E 57, 1347.*

Tsallis, C., de Albuquerque, M.P. (2000). Are citations of scientific papers a case of nonextensivity? *European Physical J. B, 13, 777-780*, http://tsallis.cat.cbpf.br/ biblio.htm

Weisstein, E.W. (2003). Eric Weisstein's World of Mathematics, Zipf's Law. http://mathworld.wolfram.com/ZipfsLaw.html

Zipf, G. K. (1935). *The Psycho-biology of Language: An Introduction to Dynamic Philology.* Houghton Mifflin Co, Boston, the first clear formulation of Zipf's law. **George Miller** (1965), a renowned linguist, summarized these studies in *"Introduction" in Psycho-Biology of Languages by G. Zipf*, MIT Press.

Zipf, G. K. (1949, 1965). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley (1949), 2nd edition, New York, Hafner (1965); a comprehensive bibliography on *Zipf's Law* has been gathered by **Wentian Li** from Rockefeller University, http://linkage.rockefeller.edu/wli/zipf/

Glottometrics 6, 2003, 94-102

Das Grazer Projekt zu Wortlängen(häufigkeiten)

Emmerich Kelih, Peter Grzybek, Ernst Stadlober[1]

1. Projektbeschreibung: Allgemeiner Hintergrund

Mit 1.4.2002 fördert der Österreichische *Fonds zur Förderung der wissenschaftlichen Forschung* (FWF, Wien; Projektnummer: P-15485) finanziell ein auf drei Jahre ausgerichtetes Projekt zur Erforschung von Wortlängen(häufigkeiten) in Texten slawischer Sprachen. An dem interdisziplinären und interuniversitären Projekt, das verantwortlich von Peter Grzybek (Institut für Slawistik, Universität Graz) in Zusammenarbeit mit Ernst Stadlober (Institut für Statistik, TU Graz) geleitet wird, arbeiten SpezialistInnen aus Textwissenschaft, Informatik und Statistik mit.

Der allgemeine theoretische Hintergrund des Projekts ist ziemlich weitreichend und anspruchsvoll: Man geht von nicht mehr und nicht weniger aus, als dass die komplementäre Anwendung quantitativer und qualitativer Methoden im Bereich der Geistes- bzw. Kulturwissenschaften eine Möglichkeit darstellt, den immer noch lebendigen Mythos von den "zwei Kulturen" zu überwinden, wie er in den 50er Jahren von Snow ins Leben gerufen und von seinen Anhängern immer wieder künstlich wiederbelebt wurde. Sobald man jedoch 'Natur' und 'Kultur' als spezifische (kulturelle) Konstrukte versteht, ändert sich die Sichtweise und es zeichnen sich zwei Perspektiven ab, wie man mit der vermeintlichen Gegenüberstellung dieser beiden Konzepte umgehen könnte: Einerseits lassen sich die historisch wechselnden Definitionen von 'Kultur' und 'Natur' (und diskursstrategische Gründe für diese Definitionen) selbst zu einem wissenschaftlichen Thema machen, andererseits muss es dann darum gehen, die Konvergenzen zwischen 'Natur' und 'Kultur' selbst zu fokussieren. Und genau diese Annahme kennzeichnet den theoretischen Hintergrund des Grazer Projekts: Während Sprache und sprachlicher Text als spezifische kulturelle Produkte und als spezifische Zeichensysteme innerhalb kultureller Gefüge angesehen werden, lassen sich die zu ihrer Analyse angewendeten statistischen Verfahren als geeignete Meta-Sprache für Kulturstudien im allgemeinen, für linguistische Studien als einer von deren Spezialbereichen im besonderen ansehen.

In Anbetracht dieser eher verwegenen, wenn nicht vermessenen allgemeinen Annahmen nimmt sich das konkrete Forschungsziel des Grazer Projekts vergleichsweise bescheiden aus, insofern "nur" Wortlängen, Wortlängenhäufigkeiten und Faktoren, die hierauf Einfluss haben, betrachtet werden. Ungeachtet dessen ist die eingeschlagene Perspektive innovativ, wenn man sich die relativ junge Geschichte von Wortlängenforschungen vor Augen hält.

Das Wort ist, ebenso wie der Satz, zentraler Bestandteil eines jeden Prozesses der Textkonstruktion. Ungeachtet dieser zentralen Rolle ist die Wortlänge als eigenständige theoretische Kategorie von der Linguistik und Textwissenschaft lange Zeit vernachlässigt worden. Erst in jüngster Zeit, insbesondere im Zuge des Aufkommens der synergetischen Linguistik, ist die Frage des Häufigkeitsvorkommens von Wörtern einer bestimmten Länge ("Wortlängenhäufigkeiten") in Texten (einer gegebenen Sprache, eines individuellen Autors, eines be-

[1] Address correspondence to: Peter Grzybek, Universität Graz, Institut für Slawistik, Merangasse 70, A-8010 Graz. E-mail: grzybek@uni-graz.at

stimmten Genres, usw.) theoretisch in systematische Kontexte integriert worden, und erst kürzlich ist eine spezifische Theorie der Häufigkeitsverteilung(en) von Wortlängen erarbeitet worden.

Sicherlich hat es, insbesondere in den 50er und 60er Jahren, eine ganze Reihe von Ansätzen gegeben, vor allem im Kontext von Strukturalismus und Informationstheorie, bei denen quantitative Aspekte des Wortes in erster Linie als relevant für stilistische Studien autoren- oder diskursspezifischer Charakteristika angesehen wurden. Diese Studien konzentrierten sich in erster Linie auf die mittlere Wortlänge – ein Vorschlag, der bekannterweise bereits 1851 von dem englischen Mathematiker und Logiker Augustus de Morgan (1806-1871) unterbreitet wurde. Natürlich wurde auch schon in diesen früheren Untersuchungen dem Umstand Rechnung getragen, dass Mittelwerte aufgrund von sehr unterschiedlichen Voraussetzungen zustande kommen, d.h. auf sehr unterschiedlichen Häufigkeitsverteilungen basieren können. Folglich konzentrierte sich auch die sogenannte quantitative Stilistik nicht nur auf Mittelwerte, sondern auch auf Varianzen als spezifische Textcharakteristika. Aus heutiger Sicht sind diesem Vorgehen aber zumindest zwei wesentliche Einwände entgegenzusetzen:

1. Mittelwert und Varianz sind nur zwei spezifische Kenngrößen der Verteilung; ohne Zweifel erlauben diese Maße einen korrekten, dennoch aber nur eingeschränkten Blick auf das gesamte Datenmaterial. Um zu aussagekräftigeren Ergebnissen zu gelangen, ist es deshalb notwendig, zusätzliche charakteristische Kenngrößen (wie z.B. Standardfehler des Mittelwerts, der Standardabweichung, des Medians, ebenso weitere Variations- und Dispersionsmaße wie Variationskoeffizient und dessen Standardfehler, Dispersionsindex und dessen Streuung, Schiefe, Kurtosis und deren Standardfehler, und viele andere mehr) zu verarbeiten. Auch verschiedene auf der Entropie und ihrer Varianz basierende Maße, wie z.B. (relative) Redundanz und Wiederholungsrate, usw. müssen im Detail analysiert werden. Antić/Djuzelic/Grzybek/Stadlober (2003) haben eine Liste entsprechender Kenngrößen erarbeitet, deren Relevanz und Zusammenhänge sowohl theoretisch als auch empirisch auszuloten sein werden.
2. In der Regel hat sich die Analyse von Wortlängen auf die mittlere Wortlänge und/oder Varianz nicht individueller Texte, sondern (mehr oder weniger klar definierter) Textkorpora, sei es eines Autors, einer bestimmten literarischen Periode, einer Gattung, eines spezifischen Funktionalstils o.ä. konzentriert. Ausgehend von der (falschen) Annahme, dass durch die Akkumulation von möglichst vielen Texten (einer Sprache, eines Genres, eines Autors, usw.) so etwas wie eine "Norm" etabliert werden könnte, hat man die Tatsache verdrängt, dass jeder Text das spezifische (individuelle) Ergebnis eines Prozesses der Textgenerierung ist, der durch bestimmte linguistische und/oder psycholinguistische Regularitäten gesteuert wird. Deshalb geht man in der gegenwärtigen Quantitativen Linguistik davon aus, dass ein Textkorpus nichts anderes als eine heterogene Textmischung (ein "Quasi-Text") ist. Daher kann es keine Textakkumulation geben, die so homogen ist, als dass sie durch ein einheitliches Verteilungsmodell beschrieben werden könnte. Für die konkrete Wortlängenforschung bedeutet die Existenz solcher "lokaler" Einflussfaktoren (wie Autorschaft, Gattung, Funktionalstil, usw.) die getrennte Analyse vollständiger Texte, und weder die Analyse von Textkorpora noch von Textauszügen.
3. Keine spezifische Kenngröße (oder eine Kombination von Kenngrößen) erlaubt eine Antwort auf die Frage, wie diese Kenngrößen durch die Häufigkeitsverteilung selbst motiviert sind, d.h. auf die Frage, wie sich die jeweiligen Häufigkeiten der *i*-silbigen Wörter innerhalb einer Verteilung ausnehmen. Bislang verfügbare empirische Ergebnisse zeigen in der Tat, dass die Häufigkeiten, mit der ein-, zwei-, drei- usw. mehrsilbige Wörter in Texten vorkommen, nicht chaotisch, sondern nach bestimmten Regularitäten organisiert sind; die Kenntnis dieser Gesetzmäßigkeiten erlaubt tiefe Einsicht in die Textstruktur und in die Textverarbeitung. Im Gegensatz zu früheren Annahmen, denen zufolge ein einziges, einheitliches Gesetz für die Wortlängenhäufigkeiten verantwortlich sein könnte (Čebanov, Fucks, u.a.), geht man heutzutage von einem flexiblen System eines übergeordneten Basismodells mit verschiedenen Modifikationen aus, die mit spezifischen (sprach-, autoren-, gattungs-, usw. bedingten) Faktoren in Zusammenhang stehen.

Die Frage jedoch, welche Faktoren auf die Wortlänge und deren Häufigkeit in Texten

Einfluss haben, und wie diese Faktoren möglicherweise interagieren, ist bislang noch nie systematisch untersucht worden. Theoretisch bestehen zwei Möglichkeiten, wie solche Einflussfaktoren ins Spiel kommen können:

a. Gemäß der ersten Annahme führt der Einfluss solcher Faktoren wie Autorschaft, Gattung, Diachronie usw. zu verschiedenen Typen von Häufigkeitsmodellen. Falls diese durchaus plausible Hypothese sich bestätigen sollte, dann wäre es als nächstes wichtig zu wissen, inwiefern sich die für die Texte einer gegebenen Sprache relevanten Modelle gegebenenfalls auf ein gemeinsames, übergeordnetes Modell zurückführen und als dessen Modifikationen interpretieren lassen.
b. Gemäß der zweiten Annahme führen die genannten Faktoren zu unterschiedlichen Modellen; in diesem Fall würden sich Einflussfaktoren wie die genannten nicht auf das spezifische Häufigkeitsmodell, wohl aber auf die spezifischen Parameter eines gegebenen Modells auswirken.

Beide Varianten lassen sich in der konkreten Textrealität beobachten; dennoch sind noch keine systematischen Untersuchungen durchgeführt worden, die dieser Frage konkret nachgehen. Deshalb konzentriert sich das Projekt in erster Linie auf die folgenden drei Fragebereiche:

I. Die systematische Untersuchung von Wortlänge und Wortlängenhäufigkeiten in Texten aus drei verschiedenen slawischen Sprachen (Kroatisch, Russisch, Slowenisch) zielt auf die Unterscheidung von sprach(en)-spezifischen und sprach(en)-übergreifenden Faktoren;
II. Die systematische Untersuchung bestimmter Autorenstile, texttypologischer Besonderheiten usw. zielt auf die Isolierung von Faktoren, die möglicherweise die Wortlänge und deren Häufigkeitsverteilung in Texten beeinflussen; die Relevanz dieser Faktoren und mögliche Zusammenhänge zwischen ihnen wird in einem weiteren Schritt zu untersuchen sein.
III. Vorausgesetzt, es lassen sich auf die beiden zuletzt genannten Fragen Antworten finden, lässt sich in einem nächsten Schritt die Richtung der Fragestellung umdrehen und danach fragen, ob sich bestimmte individuelle Texte einem bestimmten Autor, einem bestimmten Genre etc. mit einer bestimmten Wahrscheinlichkeit zuordnen lassen. Bescheidener formuliert lautet die Frage: Welchen Beitrag können Wortlängenuntersuchungen zur Beantwortung dieser Frage(n) beitragen?

Das insgesamt auf drei Jahre ausgerichtete Programm lässt sich in drei aufeinander folgende Phasen untergliedern:

1. Als erstes gilt es, eine umfangreiche Textdatenbank mit jeweils ca. 1000 Texten in jeder der drei zur Diskussion stehenden slawischen Sprachen aufzubauen, einhergehend mit entsprechenden Meta-Daten. Bei der Zusammenstellung dieses Korpus werden insbesondere texttypologische Faktoren zu berücksichtigen sein, damit eine geeignete Basis für die anschließend durchzuführenden statistischen Analysen geschaffen ist. Es wird zu überlegen sein, inwiefern diese Datenbank auch für den externen Gebrauch zugänglich gemacht werden kann, wofür freilich eine geeignete Korpusschnittstelle zu erstellen ist. Weiters ist in dieser Phase geeignete Textanalyse-Software zu erstellen (die nach Möglichkeit zukünftige, auch über slawische Sprachen hinausgehende Möglichkeiten der Weiterentwicklung vorsieht).
2. Im nächsten Schritt kommt es darauf an, die Texte für die anstehenden Analysen aufzubereiten; diese ausschließlich textbezogene Präparation beinhaltet neben einer einheitlichen Behandlung von Abkürzungen, Überschriften, Zahlen, Fremdwörtern, u.a. auch Fragen der Satzdefinition, der Unterscheidbarkeit von narrativen, deskriptiven, dialogischen und anderen Sequenzen. Im Anschluss an diese Textaufbereitungen lassen sich mit den in der ersten Phase erstellten Analyseprogrammen erste statistische

Auswertungen durchführen, die für jeden Text die entsprechenden Rohdaten liefern, die in für die weiterführenden statistischen Analysen geeignete Datenfiles abzuspeichern sind.

3. Die letzte Phase beinhaltet die quantitativen und qualitativen Auswertungen. In diesem Abschnitt wird es darauf ankommen, ein passendes Verteilungsmodell (oder, in Abhängigkeit von den Ergebnissen, mehrere geeignete Modelle) zu finden. Weitere Analysen werden sich dann auf die Faktoren richten, die entweder die konkreten Parameter des Verteilungsmodells oder aber den Verteilungstyp insgesamt beeinflussen. Dieser Arbeitsschritt wird zur Anwendung von Diskriminanz- und Clusteranalysen führen, was tiefe Einsicht in Fragen der Texttypologie, Textklassifikation und Textdiskrimination verschaffen sollte.

Wenn sich in der Tat relevante Regularitäten beobachten lassen sollten, so werden diese auch allgemein für Prozesse der Informationsverarbeitung als relevant anzusehen sein. In diesem Falle stellen die angewendeten statistischen Verfahren und Methoden eine optimale Basis für weiterführende interdisziplinäre Studien dar, denen es darauf ankommt, die vermeintliche Kluft zwischen den "zwei Kulturen" von Natur- und Kulturwissenschaften zu überbrücken.

2. Das Grazer Projekt: Phase I (2002-2003)

Die folgende Darstellung hat zum Ziel, das im ersten Projektjahr (1.4.2002 – 31.3.2003) Erreichte vor dem Hintergrund der oben vorgestellten Projektziele zu referieren.

2.1. Erstellung der Datenbank und Entwicklung von Analyseprogramme

Wie oben dargestellt, ist es in der ersten Projektphase ein zentrales Ziel gewesen, die *systematische* und *automatisierte* Untersuchung der Wortlängen(häufigkeiten) in den erwähnten slawischen Sprachen vorzubereiten.

In diesem Zeitraum wurde neben der Recherche nach elektronisch verfügbaren kroatischen, slowenischen und russischen Texten (bzw. dem zusätzlichen Einscannen weiterer Texte) die einheitliche unicode-fähige Codierung der Texte (CP 1250 bzw. CP 1251) bewältigt. Nach derzeitigem Stand wurde die folgende Anzahl von Texten (Analyseeinheit: "Inhaltlich abgeschlossene Einheit") in einer zentralen Datenbank gespeichert und zur weiteren Bearbeitung vorbereitet:

Sprache	**Anzahl von Texten**
Kroatisch	1050
Russisch	1303
Slowenisch	1290
Gesamt	**3643**

Für die systematische Verwaltung der einzelnen Texte wurde ein Linux-Server eingerichtet, wobei darauf geachtet wurde, strukturell zwischen Daten (einzelne Texte, abgespeichert als Files) und Metadaten (Informationen über die Autoren, Titel, Textquelle, Texttyp, abgelegt in einer POSTGRESQL-Datenbank) zu trennen. Das verwendete Metadaten-Schema wurde unter anderem daraufhin optimiert, Textgruppen nach bestimmten Kriterien aus dem Korpus zu

extrahieren. Um die Administration der Textdatenbank (Speicherung der einzelnen Texte, Eingabe und Verwaltung der Metadaten) zu erleichtern, wurde ein projektintern zugänglicher Web-Zugang implementiert, der den Zugriff auf die Daten von beliebigen Orten aus garantiert. Dieses Verfahren sieht bereits die Möglichkeit späterer externer Zugänge vor.

Da keine Analyse-Software zur Bestimmung der Wortlänge in Texten und die daraus notwendigerweise abzuleitenden statistischen Kenngrößen verfügbar ist, wurde innerhalb des Projektes eine eigene Software auf der Basis der Programmiersprache PERL entwickelt. Geplant ist, dass bei endgültiger Fertigstellung folgende Analyseschritte automatisiert durchgeführt werden können:

a.) Statistische Auswertung von Graphemsystemen (Vorkommenshäufigkeit);
b.) Auswertungen der Silbenstruktur;
c.) Automatische Bestimmung der Wortlänge aufgrund unterschiedlicher Wortdefinitionen und Maßeinheiten (Graphem, Silbe).

In einem weiteren Schritt werden diese – auf kroatische, russische und slowenische Texte ausgerichteten – Analyseprogramme direkt auf dem Server ausgeführt und garantieren somit eine bedienerfreundliche und effiziente statistische Textanalyse. Vorbereitet ist sowohl die Möglichkeit der lokalen (externen) Applikation als auch der Erweiterung auf andere Sprachen.

2.2. Theoretische Grundlagenarbeiten

2.2.1. Wortlängenforschungen im Kontext der Geschichte der Quantitativen Linguistik

Mit dem Ziel, die spezifischen Untersuchungen zur Wortlänge bzw. zu Wortlängenhäufigkeiten in den allgemeinen Kontext der Quantitativen Linguistik und Quantitativen Textanalyse einzuordnen, wurden eine Reihe synoptischer Studien zu diesem Bereich durchgeführt. So geht Grzybek (2003c) detailliert auf die Frage der Geschichte der Wortlängenforschungen ein: Neben der chronologischen Aufarbeitung der einzelnen Analysen seit Ende des 19. Jahrhunderts finden sich Darstellungen bis hin zu den neuesten Ansätzen im Lichte einer synergetischen Linguistik, in der die Evolution der in Betracht zu ziehenden statistischen Modelle für Wortlängen näher untersucht werden.

Ausgehend von dieser konkreten Fragestellung wurden zwei weitere Überblicksartikel zum Status und zur Entwicklung der quantitativen Linguistik in Russland (bzw. der Sowjetunion) verfasst, mit denen eine Einbettung in einen breiteren Forschungskontext vorgenommen werden kann. So beschäftigen sich Grzybek/Kelih (2003a) mit den Anfängen quantitativer Sprachanalysen in Russland seit Mitte des 19. Jahrhunderts bis hin zu den quantitativ orientierten Studien im Umfeld des russischen Formalismus. Daran anknüpfend beschäftigen sich Grzybek/Kelih (2003b) mit dem neuerlichen Start der quantitativen Linguistik in Russland nach 1956 im Kontext von strukturalistischen, kybernetischen, und semiotischen Untersuchungen bzw. im Lichte der maschinellen Übersetzung, um sodann den Bogen zu den aktuellen Forschungen der russischen quantitativen Linguistik zu spannen.

2.2.2. Tagung zur Quantitativen Textanalyse (Graz, 21.-23. Juni 2002)

Unmittelbar nach Projektbeginn wurde vom 21. bis 23. Juni 2002 eine international besetzte Konferenz unter dem Titel "Wortlängen in Texten. Internationales Symposium zur quanti-

tativen Textanalyse" veranstaltet. Ziel dieser interdisziplinär ausgerichteten Veranstaltung war es, ein Forum für die Präsentation von aktuellen Forschungsergebnissen auf dem Gebiet der quantitativen Linguistik zu schaffen. Dieses sollte die Möglichkeit bieten, mit führenden VertreterInnen der quantitativen Linguistik bereits zu Projektanfang die Untersuchungen zur Wortlängen(häufigkeiten) in slawischen Sprachen auf breiter Ebene zu diskutieren und die Perspektiven bzw. die inhaltliche Richtung einschlägiger Forschungen zu koordinieren.

Da Verlauf und Ergebnisse dieser Veranstaltung andernorts im Rahmen zweier Konferenzberichte vorgestellt wurden (Kelih/Grzybek 2002, Antić/Kelih/Grzybek 2002), kann hier auf eine eingehende Darstellung verzichtet werden. Anzumerken bleibt, dass ein Großteil der bei diesem Symposium gehaltenen Beiträge unter dem Titel "Word Length Studies and Related Topics" erscheint (Grzybek ed. 2003).

2.2.3. Wortlängenforschungen – Fragen der Wort-Definitionen

In unmittelbarer Relevanz für die Wortlängenforschung wurden im ersten Projektjahr vor allem Fragen der Definition und Quantifizierung der Einheit ‚Wort' in einer Reihe von Arbeiten reflektiert und diskutiert. Bekanntlich ist für quantitative Untersuchungen in jedem Fall eine exakt nachvollziehbare Definition der zu untersuchenden Einheiten notwendig, wobei auf unterschiedliche theoretische Konzeptionen der Linguistik zurückgegriffen werden kann.

Ausgehend von einer *graphematischen* Konzeption des Wortes, derzufolge ein Wort eine in (schriftlichen) Texten durch Leerstellen abgegrenzte Einheit darstellt, wurde diese Konzeption durch eine *graphematisch-phonologische* Definition erweitert: Wird die Wortlänge in slawischen Sprachen in der Anzahl der Silben pro Wort bestimmt, ergibt sich bei der Bestimmung der Wortlänge aufgrund des graphematischen Kriterium eine als problematisch anzusehende Gruppe von "0-silbigen" Wörtern (Gruppe von Präpositionen, die aus phonologischer Sicht als Enklitika behandelt werden können). Diskutiert wurde diese Problematik im Rahmen einer empirischen Studie unter dem Titel: "On the question of so-called 0-syllable words in determining word length" (vgl. Antić/Kelih 2003). Hierbei ging es vor allem um die Auswirkungen dieser beiden unterschiedlichen Wortdefinitionen auf übliche statistische Kenngrößen.

Auf der Basis dieser Ergebnisse wurde in einer weiteren Studie auf ähnliche Art und Weise eine andere Konzeption des Wortes diskutiert, und zwar im Hinblick auf die Implikationen einer *phonologischen* Wortdefinition (Taktgruppen) in Texten. In einer entsprechend angelegten vergleichend-statistischen Analyse der Wortlänge aufgrund von drei unterschiedlichen Konzeptionen (graphematisch, graphematisch-phonologisch, phonologisch) konnte anhand der Wortlängen in russischen Texten die systematische Verschiebung von statistischen Kenngrößen (s. Kelih/Grzybek 2003a) gezeigt werden.

Zusammengefasst zeigt es sich, dass die Bestimmung der Wortlänge unter Berücksichtigung von unterschiedlichen linguistischen Konzeptionen stringent vollzogen werden kann. Insgesamt stellt es sich heraus, dass die Wahl von unterschiedlichen Wortdefinitionen zwar – wie nicht anders zu erwarten – zu unterschiedlichen quantitativen Größen führt, dass diese jedoch eine ***systematische Verschiebung*** bewirken (was in letzter Konsequenz eine wechselseitige Transformation der Ergebnisse unabhängig von der jeweils gewählten Basiseinheit erlaubt).

2.2.4. Frage der konstituierenden Elemente des Wortes

Im Zusammenhang mit der Definition und Bestimmung der Wortlänge ist die Frage der konstituierenden Elemente des Wortes von unmittelbarer Relevanz. Denn bei der Berechnung der Wortlänge ist prinzipiell die Wahl unterschiedlicher konstituierender Einheiten möglich, die jeweils – hierarchisch gesehen – auf verschiedenen Ebenen anzusetzen sind: so ist ein Wort z.B. messbar in Graphemen, Phonemen, in Silben oder in Morphemen.

In diesem Zusammenhang wurde die grundsätzliche Frage der statistischen Modellierung von Graphemsystemen untersucht (vgl. Grzybek/Kelih 2003a,b); als ein wesentliches Ergebnis konnte festgestellt werden, dass es sich – zumindest bei den im Projekt zu berücksichtigenden Sprachen – beim graphematischen System um ein äußerst stabiles System handelt, und dass Einflussfaktoren (wie Textlänge, Textgattung, Stichprobengröße usw.) offensichtlich keine wesentliche modifizierende Rolle spielen.

Neben der prinzipiellen Frage der Wortdefinition (vgl. 2) wird weiteres zu überprüfen sein – wie eingangs erwähnt – welche Auswirkungen unterschiedliche Maßeinheiten (Graphem, Silbe, Morphem) auf die Bestimmung der Wortlänge zeigen. Zu überprüfen wird sein, inwiefern – ebenso wie für unterschiedliche Wortdefinitionen (s.o.) – ***systematische Verschiebungen*** nachgewiesen werden können (vgl. Kelih 2003b, Kelih/Grzybek 2003b).

2.2.5. Fragen von Einflussfaktoren

Bereits im Vorfeld systematischer Analysen der Wortlängen(häufigkeiten) in Texten slawischer Sprachen ist es von Bedeutung, unterschiedlichste Einflussfaktoren auf die Wortlänge in Erwägung zu ziehen.

Nicht nur im Hinblick auf die beabsichtigten Datenanalysen, sondern vor allem auch in Anbetracht der Datenbankstruktur (werden ganze Romane oder einzelne Romankapitel als ‚Text' definiert?) ist die Frage der Datenhomogenität von zentraler Bedeutung. Strauss/Grzybek/Altmann (2003) haben bei der Untersuchung des Zusammenhangs von Wortfrequenz und Wortlänge Argumente dafür bereitgestellt, Texte ausschließlich auf der Ebene homogener Texteinheiten zu berücksichtigen; in einer sich daraus ergebenden Folgestudie haben Grzybek/Altmann (2003) darüber hinaus in einer methodologisch ausgerichteten Untersuchung auf die Frage der Datenaufbereitung für quantitative Analysen aufmerksam gemacht.

Neben Faktoren wie Autorschaft, Zeitraum der Entstehung der jeweiligen Texte, u.ä. wird besondere Rücksicht auf die Frage des Texttyps zu legen sein. In ausführlicher Weise wird die Frage einer quantitativen Texttypologie – auch in Hinblick auf die strukturelle Gliederung der zu analysierenden Texte in der dazu angelegten projektinternen Datenbank – aus theoretischer Sicht in Grzybek (2003b) diskutiert. Entsprechende empirische Untersuchungen zu diesem Fragenkomplex – auf der Basis von slowenischen Texten – werden in Kelih (2003a) durchgeführt.

Abgesehen von den theoretischen Begründungen wurden eine Reihe vorbereitender Untersuchungen im Hinblick auf die Frage der Texttypologie als einer möglichen Einflussvariable durchgeführt.

Wie in einer Untersuchung von Grzybek/Stadlober (2003) zur Frage der Wortlänge in tschechischen Texten von Karel Čapek gezeigt werden konnte, sprechen die erhaltenen Ergebnisse dafür, dass offensichtlich weniger die Autorschaft oder der Zeitpunkt der Entstehung der analysierten literarischen Werke eine Rolle spielt, als vielmehr der – ebenfalls notwendigerweise einer Definition unterliegende – Texttyp.

In diesem Zusammenhang war es unter anderem auch von Bedeutung, die Anwendbarkeit der eingeschlagenen Untersuchungsdesigns auch in spezifischen Textsorten zu prüfen. In Be-

zug auf diese Frage liegen zwei Detailstudien von Grzybek (2002, 2003a) zu poetischen (metrisch gebundenen) Verstexten von A.S. Puškin sowie eine weitere Untersuchung an Sprichwort-Material (Grzybek 2003d) vor, die als paradigmatische Basis für weitere Studien dienen können.

2.2.6. Vorbereitung der Analyse statistischer Textgrößen

Abgesehen von den im ersten Projektjahr gelösten textwissenschaftlichen und programmiertechnischen Fragestellungen wurde eine Reihe von statistisch-theoretischen Arbeiten durchgeführt.

So konnten Stadlober/Djuzelic (2003) im Rahmen der oben angeführten Konferenz der Frage nachgehen, inwiefern – unter Einsatz von multivariaten Verfahren der Statistik – bestimmte aus der Wortlänge ableitbare statistische Kenngrößen als potentielle Variablen für eine quantitative Texttypologie eingesetzt werden können. Ausgehend von diesen methodologischen Überlegungen wird eine Ausarbeitung von relevanten statistischen Kenngrößen zur Beschreibung von Häufigkeitsverteilungen, die unmittelbar aus der Wortlänge her- und ableitbar sind (vgl. Antić/Djuzelic/Grzybek/Stadlober 2003). Selektiv aufgearbeitet sind einige dieser Kenngrößen in Djuzelic (2002), wo unter anderem statistische Kenngrößen der Wortlänge als Diskriminationsfaktoren für automatische Textklassifizierungen verwendet werden.

Abschließend sei angemerkt, dass die erwähnten statistischen Kenngrößen in die oben angeführte Analyse-Software eingebaut werden sollen, um so eine effiziente und systematische Analyse der Wortlänge und den damit in Zusammenhang stehenden Einflussfaktoren in den nächsten zwei Projektjahren ermöglichen werden.

2.2.7. Internationale Kooperationen

Im Rahmen der bisherigen Projektarbeit verstärkt sich derzeit die Zusammenarbeit auf internationaler Ebene. Abgesehen von einzelnen Kontakten in verschiedene europäische Länder kristallisiert sich insbesondere eine verstärkte Zusammenarbeit mit einer Reihe von slowakischen Institutionen heraus. Konkret handelt es sich um das Institut für Mathematik der Slowakischen Akademie der Wissenschaften Bratislava (Gejza Wimmer), das Institut für Mathematik der Matej Bela Universität Banská Bystrica (Jana Kusendová), sowie das Institut für Slowakische Sprache der Universität Trnava (Emília Nemcová). Um die sich abzeichnenden Kooperationen auf institutionalisierter Ebene realisieren zu können, wird derzeit ein Austausch über verschiedene europäische Förderprogramme eingeleitet.

Literatur

(Die folgenden Angaben beziehen sich ausschließlich auf die im Grazer Projekt durchgeführten Arbeiten und sollen keinen Literaturbericht zum Thema darstellen)

Antić, G.; Djuzelic, M.; Grzybek, P.; Stadlober, E. (2003). *Statistische Kenngrößen zur Beschreibung von Wortlängenhäufigkeitsverteilungen*. [Ms.]

Antić, G.; Kelih, E.; Grzybek, P. (2002). Word Length in Texts. An International Symposium on Quantitative Text Analysis. In: *Journal of Quantitative Linguistics* [Im Druck].

Antić, G.; Kelih, E. (2003). On so-called 0-syllable words in determining word length. In: Grzybek, P. (ed.), *Word Length Studies and Related Topics*. [In print]

Djuzelic, M. (2002): Einflussfaktoren auf die Wortlänge und ihre Häufigkeitsverteilung am Beispiel von Texten slowenischer Sprache. Technische Universität Graz, Diplomarbeit. [http://www.cis.tugraz.at/stat/dthesis/djuz02.zip]

Grzybek, P. (2002). Versuchen wir einmal, die Kräfte aus dem Gleichgewicht zu bringen… Quantitative Aspekte von Puškins *Evgenij Onegin* und *Domik v Kolomne*. In: J. Bernard; P. Grzybek; A. Pokrivčák; G. Withalm (eds.), *Form – Struktur – Komposition. Pragmatik und Rezeption.* Wien, 305-335. [= Special Issue of: *Semiotische Berichte, 26,1-4.*]

Grzybek, P. (2003a). Quantitative Aspekte slawischer Texte (am Beispiel von Puškins *Evgenij Onegin.* (Beitrag zum XIII. Internationalen Slawistenkongress, Ljubljana 2003.) In: *Wiener Slawistisches Jahrbuch, 49.* [Im Druck]

Grzybek, P. (2003b). Empirische Texttypologie in quantitativer Sicht. [In Vorb.].

Grzybek, P. (2003c): History and Status of Word Length Frequency Studies. In: Grzybek, P. (ed.) (2003). *Word Length Studies and Related Topics.*[In print]

Grzybek, P. (2003d): Zur Wortlänge und ihrer Häufigkeitsverteilung in Sprichwörtern (Am Beispiel slowenischer Sprichwörter, mit einer Re-Analyse estnischer Sprichwörter). In: Palm-Meister, Christine (Hg.), *Europhras 2000*. Tübingen. [Im Druck]

Grzybek, P.; Altmann, G. (2003): Oscillation in the frequency-length relationship. In: *Glottometrics 5, 97-107*.

Grzybek, P. (ed.) (2003). *Word Length Studies and Related Topics. Proceedings of the Graz Conference on Quantitative Text Analysis, June 21-23, 2002*. [In print]

Grzybek, P.; Kelih (2003a): Zu den Anfängen der quantitativen Linguistik in Russland. *International Handbook of Quantitative Linguistics*. [In print]

Grzybek, P.; Kelih, E. (2003b). Quantitative Linguistik in Russland seit 1956. *International Handbook of Quantitative Linguistics*. [In print]

Grzybek, P.; Kelih, E. (2003a). Häufigkeit russischer Grapheme. Teil I: Zur Geschichte der Untersuchung russischer Graphemhäufigkeiten. In: Deutschmann, P.; Höller, H. (eds.), *Datenverarbeitung in Sprach-, Literatur- und Kulturwissenschaft.* Graz. [Im Druck].

Grzybek, P., Kelih, E. (2003b). Häufigkeit russischer Grapheme. Teil II: Modelle von Häufigkeitsverteilungen. In: Deutschmann, P.; Höller, H. (eds.), *Datenverarbeitung in Sprach-, Literatur- und Kulturwissenschaft.* Graz. [Im Druck].

Grzybek, P., Stadlober, E. (2002): The Graz Project on Word Length (Frequencies). In: *Journal of Quantitative Linguistics, 9;* 187-192.

Grzybek, P.; Stadlober, E. (2003). Zur Prosa Karel Čapeks – Einige quantitative Bemerkungen. In: Kempgen, S. (ed.), *Sprach- und Textstudien*. [Im Druck]

Kelih, E. (2003a): Empirische Texttypologie aus quantitativer Sicht. [In Vorb.].

Kelih, E. (2003b): Wortlänge in Silben und Morphemen (am Beispiel des Russischen). [In Arbeit]

Kelih E.; Grzybek, P. (2002). Wortlängen in Texten. Internationales Symposium zur quantitativen Textanalyse. In: *etc. Empirische Text- und Kulturforschung / Empirical Text and Culture Research*, *2*; 89-91.

Kelih, E.; Grzybek, P. (2003a): Wortdefinitionen und Wortlängenforschung. [Im Druck]

Kelih, E.; Grzybek, P. (2003b): Wortlänge in Silben und Graphemen (am Beispiel des Russischen). [In Arbeit]

Stadlober, E., Djuzelic, M. (2003): Multivariate Statistical Methods of Quantitative Text Analysis. In: Grzybek, P. (ed.) (2003). *Word Length Studies and Related Topics.* [In print]

Strauss, U.; Grzybek, P.; Altmann, G. (2003): The more the better? Word Length and Word Frequency. [Submitted]

Glottometrics 6, 2003, 103-106

History of Quantitative Linguistics

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

I. Viktor Jakovlevič Bunjakovskij

Viktor Jakovlevič Bunjakovskij was an important Russian mathematician (3.12.1804 - 30.11.1889), who played an important role in improving the level of mathematical education, as did other mathematicians of his time, such as M.S. Ostrogradskij or P.L. Čebyšev.- Bunjakovskij received his primary education in Moscow in the home of Count A.P. Tormasov, a friend of his father's who had died in 1809. From 1820, Bunjakovskij stayed abroad together with the latter's son; first he was in Coburg, then in Lausanne, finally in Paris, where he studied at the Sorbonne und the Collège de France, and heard lectures of Laplace, Fourier, and Poisson, among others. In 1824 Bunjakovskij received the bachelor's degree, in 1825, he was appointed doctor of mathematics by the Paris Faculté des sciences. Immediately after his return to Russia (1826), Bunjakovskij began to teach mathematics in Petersburg. From 1846-1859 he held lectures on analytical mechanics, differential and integral calculus, as well as on probability theory at Persburg University. In 1828, Bunjakovskij was elected to an adjunct for pure mathematics of the Academy of Sciences. In 1830, he was elect to an extraordinary, and in 1836 to an ordinary member of the Academy. From 1864 until shortly before his death Bunjakovskij was Vice President of the Academy of Sciences. In 1883, Bunjakovskij, who mostly worked in the realm of number theory and probability theory, published his catalogue of scientific works (*Liste des travaux mathématiques de Victor Bouniakowsky*; SPb 1883), which points out 108 scholarly titles.

In 1847, Bunjakovskij published an article in the third volume (part II) of the journal *Sovremennik*, which was entitled: „On the possibility to apply determining measures of confidence to the results of some observing sciences, particularly statistics". Irrespective of the importance this article has, in a historical perspective, with regard to establishing statistics as a methodological discipline in its own right, Bunjakovskij's article represents – not only for Russia – one of the earliest quotations where the possibility and reasonability of applying statistics to, among others, philological questions is being discussed.

О ВОЗМОЖНОСТИ

ВВЕДЕНІЯ ОПРЕДѢЛИТЕЛЬНЫХЪ МѢРЪ ДОВѢРІЯ КЪ РЕЗУЛЬТАТАМЪ НѢКОТОРЫХЪ НАУКЪ НАБЛЮДАТЕЛЬНЫХЪ, И ПРЕИМУЩЕСТВЕННО СТАТИСТИКИ.

Thus, at the very ending of this article, Bunjakovskij writes:

"[…] It would be time now, to finish the present article; it may be allowed, however, in accordance with the analogy of the matter at stake, to add a couple of words with regard to another application of probability analysis, to which obviously no-one has ever before drawn the attention. The new application includes grammatical and etymological studies of a language, as well as comparative philology. However strange such studies may, at first sight, seem for a mathematical analysis, yet I am fully convinced that a broad field for strictly mathematical ruminations opens up before us. My claim is not so much based on more or less uncertain conjectures and suppositions but on a critical evaluation of the object, on a number of attempts I have already made, and some analytical formulae, which I have introduced in order to define the *numerical* probability of particular word formations. In this way one can, for example, approximately determine the measure of confidence of a given etymology, and, depending on the proximity of this number to 1 and its confidence, one has to judge about the proposed authenticity. This is not the place to discuss details of this matter to with I merely wanted to draw the attention. But in order to directly explain in which way corresponding studies may find their way into the realm of applied mathematics, I regard it to reasonable to briefly point out some of the numerical assertions and materials which need to be elaborated. If we talk about a language, we predominantly assume that we have its detailed *arithmetical description* or, if we will, its statistics, i.e. numerical assertions about the complete inventory of the words of that language, about the distribution of these words according to the parts of speech, about the number of letters, about the initial letters, the endings, etc. etc. Also general rules have to be named here, exceptions of various kinds, words undoubtedly stemming from different languages, and the like. This is the numerical material the strict analysis of which quite naturally demands mathematical considerations. If one has the relevant statistical data as to two or more languages, one can compare them in various aspects, and the results obtained take the status of an authority which philologists, at the present state of science, do not use to have available.

Of course the elaboration of what I have called the *statistics of language* is a very painstaking endeavor, and most probably philologists will tend to call this kind of effort almost wasted, because the assumed gain of the exactness of the conclusions about language does not pay for the loss of time. We do not take on the responsibility of solving the question to which degree such an assertion may be justified.

Maybe I will publish my theoretical studies on the matter which I have only mentioned here, at some other opportunity. However, as far as the practical applications of the general formulae are concerned, it will be necessary, given the general lack of detailed arithmetical data about languages, to confine oneself to a few examples. By the way, in order to devote the proper degree of completeness and soundness to such a work with regard to the philological conditions, it goes without saying that a mathematician has to establish direct contact with specialists in this field, more or less foreign to himself."

Peter Grzybek

II. Bohumil Trnka: The first bibliography

Czech scholar Bohumil Trnka (3.6.1895 in Kletečná – 14.2.1984 in Prague), professor of English and older English literature at the Faculty of Arts, Charles University, Prague, was one of the founders of the Prague Linguistic Circle (1926), its first secretary, and one of the most important representatives of the Prague Linguistic School.

The representative English-translated selection of Trnka's fifty contributions to English, Czech and general linguistics, written in the years 1928-1978, was published by the Czech anglicist Fried (Trnka 1982), along with a brief survey of Trnka's life and work (written by Vilém Fried), and with an afterword by Roman Jakobson. This comprehensive volume covers all main fields of his linguistic research: general linguistics, synchronic phonology, historical linguistics (diachronic phonology and morphology), synchronic morphology, syntax and style, and last but not least, statistical linguistics. Trnka's studies are a significant testimony to the development of linguistic thought in the twentieth century.

Below we would like to remember that one of Trnka's favourite scientific topics was phonology: As early as in 1935 he published the first systematic phonological description of Modern English (which was – 31 years later – also published in Japan, 1966); he concentrated on the qualitative as well as quantitative analysis (functional load) of phonemes, both in the language system and in speech. In his research, he particularly emphasized one point: while the presence vs. the absence of an element in a language system is a matter of quality, the functional load of an element is a quantitative complementary feature. Also Trnka's scientific contribution to structural and functional morphology, as well as to other levels (plans, subsystems) of language, to the theory of the linguistic sign, to the historical evolution of language, etc. is well-known.

Trnka was the author of the most comprehensive programmatic integration of quantitative studies in the Prague structuralist theory of language. Here is a quotation from one of his later works (1950b: 3): "Before any speech units can be counted, we must have them, and it is clear that the correctness of statistical results depends entirely – provided that no omissions or other errors in counting are made – on that of structural linguistics." Trnka believed in the existence of general quantitative laws which govern the structure of all languages, and he considered attempts to formulate them as a major task of future quantitative linguistics. On the 6th International Linguistic Congress in Paris (1948), Trnka was one of the nine elected members of the committee for linguistic statistics, which was established at the congress to promote quantitative research (together with M. Cohen, W. Doroszewski, V. Georgiev, B. Migliorini, F. Mossé, A.S.S. Ross, H. Spang-Hanssen and G.K. Zipf). As the secretary of this committee, he began to organize the work in accordance with the recommendations of the Congress, by starting to compile „a provisional bibliography" of „works devoted specially to the statistical method in linguistic matters" for early publication with financial aid provided by UNESCO. This bibliography – with an introduction by Marcel Cohen – is the first bibliography of quantitative linguistics ever (Trnka 1950a): it includes 235 items divided into ten sections:

I. General works on linguistic statistics.
II. Frequency of phonemes. General laws of phonemic frequency.
III. Frequency of words and general laws of their distribution.
IV. Frequency dictionaries and frequency word counts for the purpose of learning modern languages.
V. Morphological, syntactic, metrical and semantic studies based on counts.
VI. Concordances and word frequency counts in vocabularies referring to individual authors.

VII. Statistical studies preparatory to the construction of auxiliary languages or to the rationalization of spelling. Basic English.
VIII. Statistical study preparatory to the construction of shorthand and typewriter systems. Telephone conversations.
IX. The growth of the vocabulary of children's speech . Schizophrenic language.
X. Statistical studies referring to problems of historical grammar and classification of languages.

It is worth mentioning that among those who collaborated with Trnka on the bibliography was another member of the committee for quantitative linguistics, G.K. Zipf. Trnka also reviewed Zipf's studies: he was the first to acquaint Czech linguists with Zipf's work (1950b). Trnka pointed out the independence of and the differences between the starting points and aims of the Prague School's and Zipf's quantitative researches. He was certainly not wrong when he highly praised Zipf's work in his English review on the one hand, expressing his conviction that "it will not fail to influence the linguistic thought of today" (1950: 5), nor was he wrong when, on the other hand, he stressed that Zipf's laws were only partially applicable, and when he demanded that they be revised. He accurately described Zipf's contribution showing the advantages of a statistical method compared with qualitative analysis in the sense that statistical analysis "is being able to afford to neglect the narrow limits of one language and to concentrate on linguistic problems of a general character".

Comparing the picture of the modest title page of Trnka's bibliography with big bibliographies of quantitative linguistics compiled later on (either accessible in print or on the internet now), we can see and appreciate the enormous development of the field of quantitative studies during last decades.

Trnka, Bohumil (1935), *A Phonological Analysis of Present-Day Standard English*. Praha: Universita Karlova.

Trnka, Bohumil (1950a), *A Tentative Bibliography*. Utrecht etc.: Publication of the Committee of Linguistic Statistics.

Trnka, Bohumil (1950b), review of: G. K. Zipf, The psychobiology of language; Human behavior and the principle of least effort. In: *Časopis pro moderní filologii 33, 3-5*.

Trnka, Bohumil (1982), *Selected papers in structural linguistics*. (Ed. V. Fried.). Berlin: Mouton.

Ludmila Uhlířová

Books received

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

Baayen, R.H. (2001). *Word frequency distributions*. Dordrecht: Kluwer.

Ballod, Matthias (2001). *Verständliche Wissenschaft. Ein informationsdidaktischer Beitrag zur Verständlichkeitsforschung* (= Forum für Fachsprachen-Forschung, Bd. 57). Tübingen: Narr.

Hřebíček, Luděk (2002). *Vyprávění o lingvistických experimentech s textem*. Prague: Academia.

Ondrejovič, S., Považaj, M. (eds.). (2001). *Lexicographica ´99. Zborník na počest´ Kláry Buzássyovej.* Bratislava: Veda.

Roelcke, Th. (2002). *Kommunikative Effizienz. Eine Modellskizze*. Heidelberg: Universitätsverlag C. Winter.

Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S., Wimmerová, S. (2003). *Úvod do analýzy textov.* Bratislava: Veda.

Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse.* Wien: Edition Praesens.

Available issues

Glottometrics 1, 2001

Glottometrics 2, 2002

Glottometrics 3, 2002 (To honor G.K: Zipf)

Glottometrics 4 (To honor G.K. Zipf)

Glottometrics 5 (To honor G.K. Zipf)