Glottometrics 13
2006

RAM-Verlag

ISSN 2625-8226

Glottometrics

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

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

Herausgeber – Editors

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

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

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

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 13 (2006), Lüdenscheid: RAM-Verlag, 2006. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet (Open Access) unter der Adresse
https://www.ram-verlag.eu/journals-e-journals/glottometrics/ verfügbar.
Bibliographische Deskription nach 13 (2006) **ISSN 2625-8226**

Glottometrics 13

Contents

Glottometrics 13, 2006, 1-22

Zum Problem der Entstehung des syllabotonischen Versmaßsystems im europäischen Vers [1]

Evgeny Kazartsev, St. Petersburg–Berlin[2]

Abstract. The present study is dedicated to the problem of the rise of the syllabotonic verse in early modern European poetry. It is part of the research on regularities in the genesis of a common versification system (the syllabotonics) within the poetry of different European nations. Here we investigate the rhythmic features of the German iambic tetrameter in the time of its emergence and early development. We compare some specifics of the German iambus with those of Dutch and later Russian verse. The investigation is performed using quantitative metrical methods. The result yields a picture of the development of the German verse in the years 1620 to 1720.

Einführung

Im 17. sowie in der ersten Hälfte des 18. Jahrhunderts vollzog sich in der Versdichtung mehrerer europäischer Völker ein Wandel, und zwar zum syllabotonischen Verssystem, das die früheren rein syllabischen oder rein tonischen Verse verdrängte bzw. ersetzte. Der Syllabotonik liegt der regelmäßige Wechsel betonter und unbetonter metrischer Positionen innerhalb des Gedichts zugrunde. Es handelt sich also um ein Versmaßsystem mit regulierten Versfüssen wie z.B. Jamben und Trochäen.

Zu Beginn des 17. Jahrhunderts beherrschte die Syllabotonik den literarischen Vers des niederländischen Barock. Ein wenig später, im zweiten Viertel des 17. Jahrhunderts, festigte sich die Syllabotonik, angeregt durch die niederländischen Quellen und dank der Tätigkeit von Martin Opitz[3] und seinen Anhängern, auch im deutschen Vers. Im weiteren Verlauf öffnete sich unter dem Einfluss der deutschsprachigen Dichtung der literarische Vers in Nord-, Mittel-, und Osteuropa der Syllabotonisierung (Gasparov 1996). Die Entstehung des Jambus in der russischen Literatur des 18. Jahrhunderts ist ebenfalls unter dem deutschen Einfluss zu sehen. In der russischen Versforschung stellt man traditionell die Frage nach Einflüssen der deutschen Quellen auf die Dichtung von Michail V. Lomonosov, der als erster russischer Dichter „reine" jambische Verse geschrieben hat.[4]

---

[1] Der Aufsatz ist im Rahmen eines Stipendiums der Alexander-von-Humboldt-Stiftung entstanden: diese Forschung wurde vom Oktober bis Dezember 2002 von dem Deutschen Akademischen Austauschdienst (DAAD) und in den Jahren 2003–2004 von dem Russischen Bildungsministerium (Projekt № PD 02-3.17-115) zum Teil unterstützt (einige Ergebnisse davon wurden in Kazartsev 2004b publiziert); eine Erweiterung sowie Beendigung dieser Arbeit im Jahre 2006 wurde dank der finanziellen Unterstützung der A. von Humboldt-Stiftung möglich.

[2] Address corespondence to: kazar@list.ru

[3] Martin Opitz, *‚der Stammvater'* der deutschen Syllabotonik, begann seine Reform des Verses, indem er im Jahre 1624 sein „Buch von der deutschen Poeterey" veröffentlichte, das unter dem Einfluss des Buches „Die niederduytsche Poemata" (1616) des niederländischen Dichters Daniel Heinsius entstand (Heinsius 1615).

[4] Die Frage nach dem nationalen oder nicht-nationalen („importierten") Charakter der Rhythmik der Jamben Lomonosovs wurde in den Forschungen von K.F. Taranovskij und V.M. Žirmunskij gestellt (Taranovskij 1953, 1975; Žirmunskij 1968). In der Dissertation des Autors dieser Studie zum Thema „Rhythmus und Sprache bei der Entstehung der russischen Syllabotonik (auf der Grunglage des russischen und deutschen Verses)" wurde

Ein grundlegendes wissenschaftliches Problem dieser Studie besteht in der Suche nach Regularitäten bei der Entstehung eines gemeinsamen Verssystems innerhalb einer Periode in der Dichtung verschiedener europäischer Völker. Angesichts dessen wendet man sich der Untersuchung des frühen deutschen Jambus zu, da die deutsche Syllabotonik, deren Entstehung mit dem niederländischen Vers verbunden ist, den Impuls zur Entwicklung analoger Systeme in anderen, darunter einigen slawischen Sprachen gegeben hat.

Von besonderem Interesse ist die Entstehungsphase des Jambus bei Martin Opitz sowie die frühe Entwicklung des jambischen Rhythmus innerhalb der Barockperiode. Die genaue quantitative Erforschung der Syllabotonik in der deutschen Poesie von Beginn an eröffnet uns die Möglichkeit einer exakten Erfassung ihrer Entstehung im Deutschen sowie der Ausbreitung dieser Verstechnik im mittel- und osteuropäischen Literaturraum, die dort in wesentlichem Maße unter dem deutschen Einfluß entstanden ist.

Die vorliegende Arbeit ist Teil einer Forschung, die der Untersuchung der Entstehung des syllabotonischen Verses in der europäischen Dichtung des 17. und 18. Jahrhunderts gewidmet ist. In dieser Studie werden Ergebnisse der rhythmischen Analyse, die an Beispielen des deutschen Jambus der Barockzeit durchgeführt wurde, vorgestellt. Sie gründet sich auf die gängigen deskriptiv-statistischen Methoden in der Metrik.[5] Als Textbasis der Untersuchung dienten 4-hebige Jamben (Ja4)[6] von Martin Opitz (Beginn der analysierten Periode), geistliche Oden von Simon Dach und Andreas Gryphius (mittlere Periode) sowie auch die bereits rhythmisch weiterentwickelten Jamben von Johann Christian Günther (Ende der Periode).[7] Bei den Dichtern wird eine Unterscheidung zwischen *‚Praktikern‘* und *‚Theoretikern‘* vorgenommen: als Theoretiker kann Martin Opitz betrachtet werden, die übrigen Autoren zählen zu den Praktikern.

Es ist zu hoffen, dass durch die hier vorgestellten Ergebnisse eine Perspektive für quantitative Untersuchungen der Syllabotonik im europäischen Vers eröffnet werden kann.

Quantitative Erforschung der Versrhythmik

In der slavischen bzw. russischen Verslehre hat die formale deskriptiv-statistische (quantitative) Analyse hat eine lange Tradition. Diese so genannte „russische Methode“ (ein Begriff des amerikanischen Slavisten J. Bailey) in der Versforschung gründet sich auf die Anwendung verschiedener statistischer Methoden bei der metrischen Textanalyse (Bailey 1979; 1987). Die folgenden Daten wurden durch die formale Analyse der rhythmischen Einheiten des Verses gewonnen - in Betracht gezogen sind zwei Merkmale des rhythmischen Texts: die Akzentprofile und die rhythmischen Formen. Die Akzentprofile zeigen, wie oft eine be-

---

das Problem der Entstehung der russischen Jamben unter dem Einfluss der deutschen Sprache und des deutschen Versrhythmus untersucht (Kazartsev 2001a, 2001c).

[5]Die wissenschaftliche Konsultation dieser Forschung wurde von M.A. Krasnoperova durchgeführt. Der Autor dankt H.-W. Jäger, C. van Bree, K. Bostoen für die wissenschaflichen Ratschläge und die Unterstützung bei der Auswahl des Materials sowie Ch. Küper für wichtige Hinweise und Anregungen. Herzlich gedankt sei F. Simmler und M. Hüning für die Förderung im Rahmen des AvH-Stipendiums.

[6] Wir nehmen an, dass der 4-hebige Jambus (Ja4) unter den anderen syllabotonischen Metren zu dieser Zeit in der deutschen Poesie keine führende Position einnahm. Beispiele für den Ja4 finden sich im Laufe des 17. Jahrhunderts relativ selten. Sein Gebrauch ist am Übergang des 17. zum 18. Jahrhundert wesentlich auffälliger, vor allem im Werk J.-Ch. Günthers.

[7] Untersucht wurden folgende Verstexte (siehe „Quellen: Gedichte“): alle Verse von M. Opitz aus dem „Buch von der deutschen Poeterey“ (1624), die im 4-hebigen Jambus verfasst sind [Opitz (1624)], sowie seine früheren und späteren Gedichte (seine erste Ja4 aus dem Jahre 1619, Jamben vom 1631 sowie seine letzte Ja4 aus dem 1637), zwei Oden von S. Dach (1650a,b), die rhythmisch relativ gleich sind, eine geistliche Ode von A. Gryphius (1663) und die Ode an Prinz Eugen von J.-Ch. Günther (1730). Der gesamte Umfang des deutschen Materials beträgt 1504 Verszeilen (die Beschreibung der niederländischen Quellen siehe in: Kazartsev 2006).

stimmte Hebung im ganzen Gedicht betont ist. Die rhythmischen Formen stellen die Betonungskonfigurationen der Verszeilen dar, und zwar lassen sich elf rhythmische Konfigurationen des Ja4 im deutschen Vers unterscheiden (s. Tabelle 1).

Tabelle 1
Rhythmische Konfigurationen des Ja4 im deutschen Vers

Form Nr.	**Konfigurationstyp**	**Textbeispiele**
I	v–´v–´v–´v–´ (v)	Er st**e**ht, beschl**eu**st und f**i**cht schon w**ie**der
II	v–v–´v–´v–´ (v)	Um von Eug**e**n Best**a**nd zu l**e**rnen
III	v–´v–v–´v–´ (v)	Aus M**ö**rsern und Carth**au**nen kr**a**chen
IV	v–´v–´v–v–´ (v)	Er r**au**scht wie P**a**nzer und Gew**e**hr
V	v–v–v–´v–´ (v)	Als bis ein Alex**a**nder w**ei**nt
VI	v–v–´v–v–´ (v)	Und von Geb**u**rth an eine Kr**a**ft
VII	v–´v–v–v–´ (v)	Das s**o**ll ihr, um euch zu erg**ö**tzen
VIII	v–v–v–v–´ (v)	Und bis zur einen Bäker**ei**
IX	v–´v–´v–´v– (v)	Und schr**ey**n: O **a**rmer **Bo**sphorus!
X	v–´v–v–´v– (v)	Der gn**ä**digste El**i**sabeth
XI	v–v–´v–´v– (v)	Und nach vollbr**a**chter M**i**ssethat

Anmerkung: «v» – Senkung, «–´» – betonte Hebung, «–» – unbetonte Hebung, «(v)» – fakultative 9-te Silbe.

Die Untersuchung der genannten Merkmale zeigt, dass die Hebungen im deutschen Jambus durch die ganze Barockperiode sehr häufig realisiert sind: bei den analysierten Versbeispielen sinkt die Häufigkeit der vollbetonten rhythmischen Konfiguration (Form I) nicht unter 70% (s. Tabelle 2).

Tabelle 2
Rhythmik des deutschen 4-hebigen Jambus
(AH = Absolute Häufigkeit; RH = Relative Häufigkeit)

Formen	**Opitz**		**Dach**		**Gryphius**		**Günther**	
	AH	RH	AH	RH	AH	RH	AH	RH
I	101	0,732	150	0,721	280	0,700	371	0,742
II	9	0,065	13	0,063	33	0,083	9	0,018
III	14	0,102	9	0,043	32	0,080	50	0,100
IV	7	0,051	30	0,144	43	0,108	56	0,112
V	0	-	0	-	1	0,003	2	0,004
VI	1	0,007	1	0,005	2	0,005	0	-
VII	0	-	1	0,005	0	-	1	0,002
IX	3	0,022	3	0,014	6	0,015	9	0,018
X	3	0,022	1	0,005	2	0,005	0	-
XI	0	-	0	-	1	0,003	2	0,004
Insgesamt	138	1,000	208	1,000	400	1,000	500	1,000
Hebungen								
1.	128	0,928	194	0,933	363	0,908	487	0,974
2.	121	0,877	197	0,947	365	0,913	447	0,894
3.	130	0,942	176	0,846	355	0,888	443	0,886
4.	132	0,957	204	0,981	391	0,978	489	0,978

Dabei ist bei Martin Opitzens frühesten jambischen Versen im Unterschied zum späteren deutschen Vers ein charakteristisches Übergewicht der Form III im Vergleich zur Form IV zu beobachten (s. Tabelle 2).[8] Verbunden damit ist eine (beim einseitigen Test signifikante) Erhöhung der prozentualen Betontheit der dritten Hebung gegenüber der zweiten. Diese rhythmische Tendenz entsteht in seinen ersten 4-hebigen Jamben und bleibt praktisch bis zum Ende seines Lebens bestehen. Im späteren deutschen Vers ist diese Tendenz nicht mehr zu beobachten. Wie entsteht diese Besonderheit in der Rhythmik von M. Opitz? Der Grund hierfür könnte im Einfluss des niederländischen Verses, der als Muster für M. Opitz diente, liegen. Im niederländischen 4-Heber hatte sich nämlich um 1600 eine ähnliche rhythmische Tendenz entwickelt (siehe Kazartsev 2006). Im opitzschen Vers tritt sie sogar deutlicher auf als in den niederländischen Mustern (s. Abbildung 1 und 2, sowie Tabelle 3).[9]

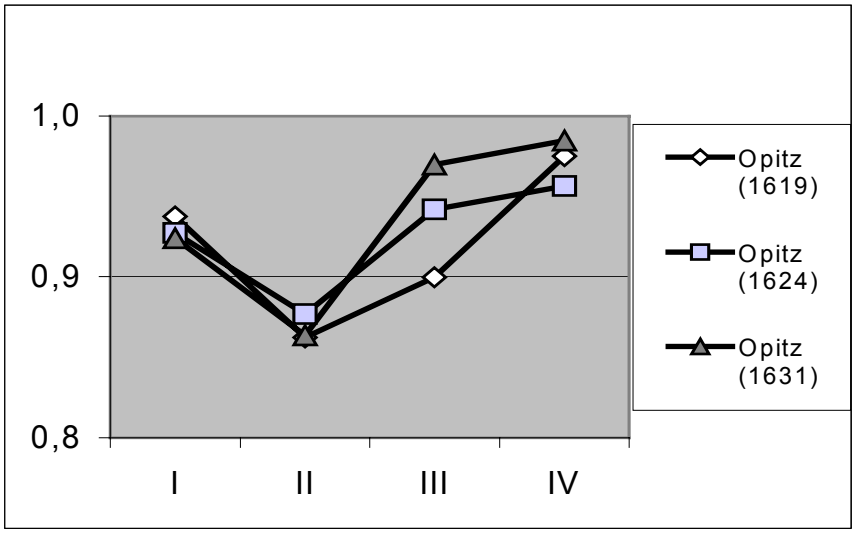


Abbildung 1. Akzentprofile des jambischen 4-Hebers bei M. Opitz

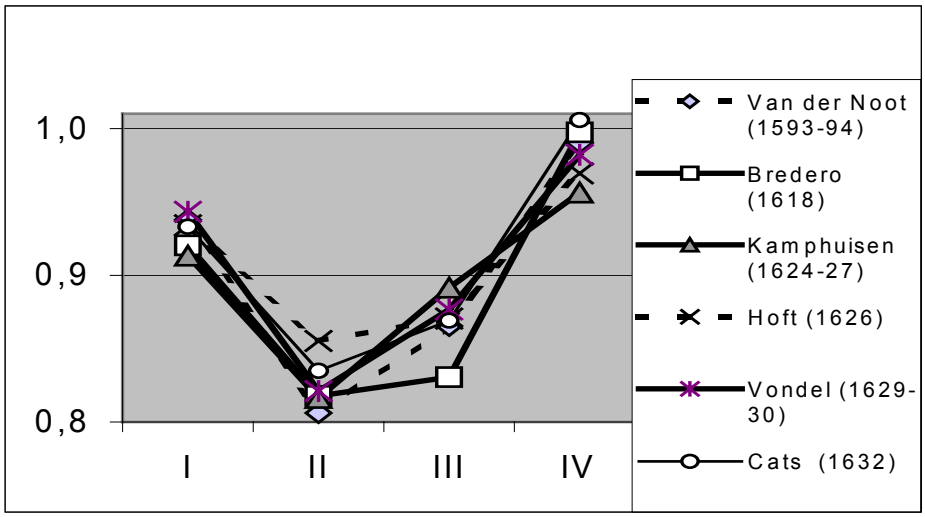


Abbildung 2. Akzentprofile des niederländischen 4-Hebers

[8] Aufgrund der Binomialverteilung wird die Hypothese der Zufälligkeit der Verteilung der Formen III und IV in diesem Fall mit relativ niedriger Fehlerquote 9,5% abgelehnt.

[9] Da die Rhythmik des 4-Hebers von M. Opitz innerhalb verschiedener Perioden seines Lebens relativ gleich war (s. Tabelle 3), werden nur die Angaben seines Ja4 aus dem „Buch von der deutschen Poeterey“ (1624) für die Vergleichsanalyse mit dem späteren deutschen Vers genommen (s. Tabelle 2).

Tabelle 3
Akzentprofile des 4-hebigen Jambus bei M. Opitz und im niederländischen Vers

Autoren \ Hebungen		1.		2.		3.		4.		Zeilen Gesamt
		Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	
M. Opitz	1619	75	0,938	69	0,863	72	0,900	78	0,975	80
	1624	128	0,928	121	0,877	130	0,942	132	0,957	138
	1631	61	0,924	57	0,864	64	0,970	65	0,985	66
	1637	106	0,946	99	0,884	106	0,946	108	0,964	112
Niederländische Autoren	V. der Noot (1593-94)	141	0,928	121	0,796	130	0,855	149	0,980	152
	Bredero (1618)	71	0,910	63	0,808	64	0,821	77	0,987	78
	Kamphuisen (1624-27)	364	0,924	333	0,845	339	0,860	378	0,959	394
	Hooft (1626)	84	0,903	75	0,807	82	0,882	88	0,946	93
	Vondel (1629-30)	168	0,933	146	0,811	156	0,867	175	0,972	180
	Cats (1632)	216	0,923	193	0,825	201	0,859	233	0,996	234

Die „Distanzen“ zwischen den Angaben der Akzentprofile in den Werken von M. Opitz und von den niederländischen Autoren (s. Tabelle 3) mit Hilfe der Euklidischen Norm[10] sowie die Analyse dieser Abstände in Kartesischen Koordinaten (s. Tabelle 4) weisen die Verbindung der Rhythmik der früheren deutschen und der niederländischen Jamben auf: dem niederländischen Vers ist die Rhythmik der ersten 4-Hebers von M. Opitz aus dem Jahre 1619 besonders nah. Man kann sagen, dass dieser Vers eine Verbindungsrolle zwischen den niederländischen und den späteren deutschen Jamben spielt (s. Abbildung 3).

Tabelle 4
Vergleich der Akzentprofile (erste drei Hebungen) von M. Opitz und von den niederländischen Autoren mit Hilfe der Euklidischen Norm

	Op19	Op24	Op31	Op37	Noot	Bred	Kamp	Hooft	Vondel	Cats
Opitz 19	-	4,55	7,10	5,15	8,07	10,03	6,82	4,53	6,14	5,75
Opitz 24		-	3,09	3,03	11,84	14,08	9,58	8,76	10,01	9,80
Opitz 31			-	3,81	13,29	15,99	10,70	11,08	11,60	11,73
Opitz 37				-	12,76	15,12	10,94	9,65	10,85	10,77
Noot					-	4,06	3,74	4,95	1,97	2,93
Bredero						-	6,16	5,64	5,17	4,40
Kamphuysen							-	4,88	3,39	3,53
Hooft								-	3,59	2,05
Vondel									-	1,87
Cats										-

[10] Die Euklidische Norm wurde nur für erste drei Positionen der Akzentprofile nach folgender Formel berechnet:

$$D(h,z) = \sqrt{\sum_{i=1}^{3}(h_i - z_i)^2}$$

Dabei bezeichnen wir die Größe der Hebungen in Positionen 1 bis 3 in einem Profil als h_i und im anderen Akzentprofil z_i und benutzen dafür die Daten aus Tabelle 3, wobei wir die relative Häufigkeit mit 100 multiplizieren, damit wir prozentuale Angaben sowie die Zahlen > 1 bekommen.

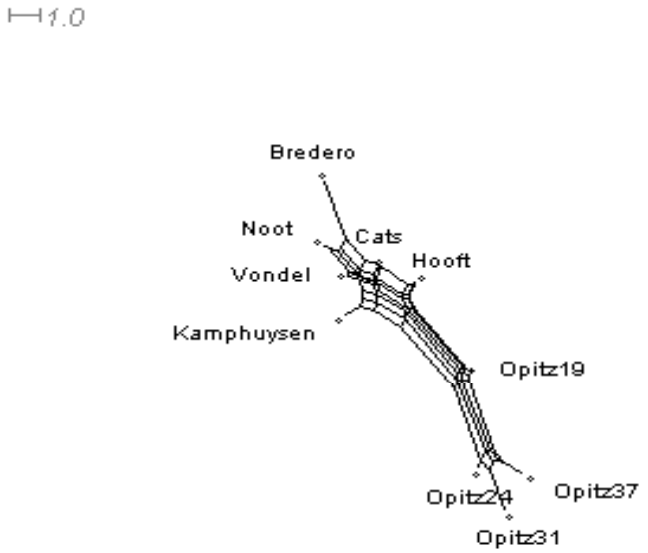


Abbildung 3. Graphische Darstellung der Abstände zwischen dem 4-Hebiger von M. Opitz und dem niederländischen Vers[11]

Im späteren deutschen Vers bei Simon Dach und, nicht ganz so ausgeprägt bei Andreas Gryphius (Mitte des 17. Jahrhunderts), wurde eine gegensätzliche Tendenz festgestellt: hier ist Form IV häufiger als Form III, d.h. die zweite Hebung im Akzentprofil des Verses wird öfter betont als die dritte. Daher stellt, nach den gegebenen Merkmalen, der Vers Dachs einen Gegensatz zum Vers Opitzens dar (s. Abbildung 4).[12]

Die Verteilung der Betonungen im Jambus Dachs ergibt ein alternierendes Akzentprofil: die zweite Hebung wurde viel häufiger realisiert als die Nachbarhebungen, die erste und die dritte. Einige spätere deutsche Verse von Gryphius erscheinen bereits weniger alternierend. In der Poesie J.-Ch. Günthers (Ende der Periode) wurde eine „Glättung" des aufgetretenen Kontrastes in der Rhythmik der Vorgänger festgestellt, was man als „Wechselwirkung" der beobachteten Tendenzen interpretieren kann: die Unterschiede der Häufigkeit der Formen III und IV werden geringer, die Betontheit der inneren Hebungen wird ausgeglichen (Tabelle 2; Abbildung 4). Der Vers von Günther ist auch den frühesten Jamben von M. Opitz ähnlich: das Vollbetonungsniveau des Verses (die Häufigkeit der Form I) sowie die Verteilung der Konfiguration mit der unakzentuierten zweiten Hebung (Form III) liegt in den Werken von beiden Autoren nahe beieinander ($X^2 = 0{,}008$ bei $\chi^2_1 = 3{,}84$). Es ist wahrscheinlich, dass Günther einige Charakterzüge der rhythmischen Gestalt der frühesten deutschen Jamben übernahm.

[11] Diese Darstellung wurde mit dem taxonomischen Programm „Splitstree4" nach den Angaben der Tabelle 4 gemacht.

[12] Der Vergleich der rhythmischen Formen im Vers von Opitz und Dach mithilfe des χ^2-Kriteriums ergab die Inhomogenität der Stichproben: $X^2 = 11{,}22$ beim kritischen Wert 7,82 (Signifikanzniveau 5%; 3 FG).

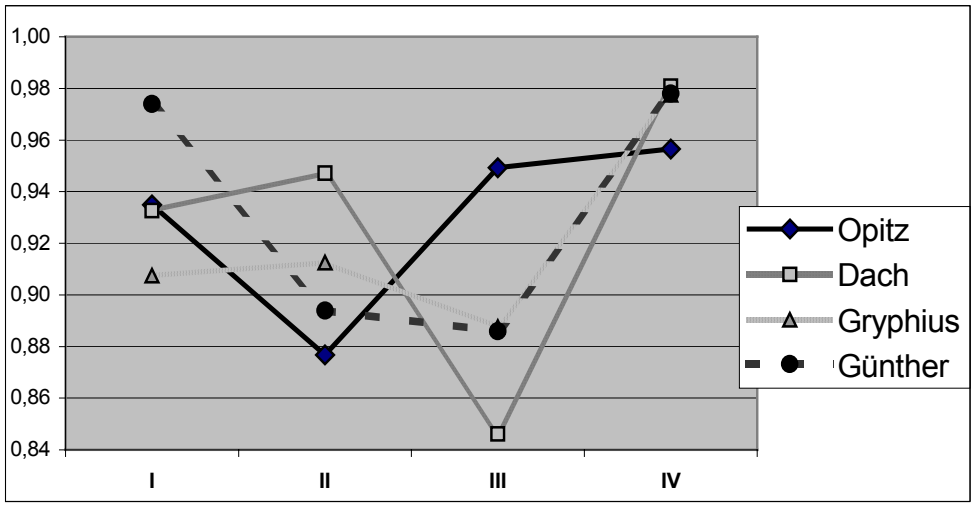


Abbildung 4. Akzentprofile des deutschen 4-Hebers

Vers- und Sprachrhythmik:
Rhythmische Gestalt des Jambus vor dem Hintergrund der Versmaßmodelle

Die untersuchten Verstendenzen spiegeln offensichtlich einige verborgene Versifikationsprozesse wider, welche die Entstehung und die rhythmische Entwicklung des Jambus begleiten und die möglicherweise mit der Wahl der rhythmischen Wörter (RW)[13] zusammenhängen. Für die Untersuchung dieser Frage ist eine besondere Methode der Versanalyse zu verwenden, die in der Modellierung der rhythmisch organisierten Texte besteht.

Im Rahmen der *russischen Methode* wird die Modellierung der Versrhythmik verwendet. Das erste wahrscheinlichkeits-statistische Versmaßmodell wurde von Boris W. Tomaševskij in den 1920er Jahren entwickelt (Tomaševskij 1929). Es zeigte, wie sich die Verteilung der rhythmischen Strukturen im Vers gestaltet, wenn sich die rhythmischen Wörter *kontextfrei* – also unabhängig von den vorangehenden Wörtern sowie von der metrischen Position – verbinden. In den 60er Jahren überarbeitete Andrej N. Kolmogorov das Modell wesentlich, indem er zur Berechnung einen rhythmischen Wortschatz verwendete, der der Prosa entnommen ist und nicht wie bei Tomaševskij den Verstexten (Kolmogorov 1968; Kolmogorov, Prochorov 1985). Dabei wurde die Prosa als sprachlicher Hintergrund für die Erforschung des Verses betrachtet. Das Modell von Tomaševskij–Kolmogorov wurde „Sprachmodell", oder auch „Sprachmodell der Unabhängigkeit" (SpM)" genannt. Es wurde von Marina A. Krasnoperova in ihrer modernen Theorie der rekonstruktiven Modellierung der Versifikation (RM-Teorie) weiter entwickelt und theoretisch bearbeitet (Krasnoperova 2000: 88–94). Unter dem Begriff "Sprachmodell" oder "Sprachmodelle" ist in der Metrik die theoretische Darstellung der Versrhythmik zu verstehen. Das Sprachmodell der Unabhängigkeit voraussagt die Verteilung der rhythmischen Formen im Vers, wenn der Dichter den Vers „wie die Prosa" schreibt, d.h. unabhängig von den rhythmischen Tendenzen seines Verstextes, und

[13] Unter dem Begriff „rhythmisches Wort" (RW) wird die Silbenanzahl innerhalb der Grenzen eines phonetischen bzw. phonologischen Wortes gemeint, wobei eine Silbe den Hauptakzent des Wortes trägt. Man unterscheidet z.B. folgende Typen der rhythmischen Wörter: der einsilbige (1.1: *dort*), die zweisilbigen mit dem Akzent auf der ersten (2.1: *schr**ei**ben*), sowie auf der zweiten Silbe (2.2: *dan**a**ch*), dreisilbigen dementsprechend (3.1: *St**u**die,* 3.2: *begr**ei**fen;* 3.3: *Theor**ie***) u.s.w., sowie die Komposita (z.B.: 3.1,3: ***Bu**chsensch**u**ss*; 4.2,4: *zu W**i**ttemb**e**rg* u a.) .

sich nur um die Einhaltung eines bestimmten Versmaßes bemüht. Die Sprachmodelle beruhen auf dem Prosa-Wortschatz, in dem jedem Typ eines rhythmischen Wortes eine bestimmte Wahrscheinlichkeit aufgrund seiner Frequenz (Häufigkeit) in der Prosa zugeordnet ist. Im SpM wird die Vorkommenswahrscheinlichkeit einer rhythmischen Konfiguration (Ja4-Form) nach dem Theorem der Multiplikation der Wahrscheinlichkeiten bestimmt, das der einen oder anderen Variante einer Ja4-Form entspricht, z.B. der folgenden rhythmischen Variante (V-n) der vollbetonten Form (s. oben, Tabelle 1, F. Nr. I):

v–´v–´v–´v–´(v): Er st**e**ht,| beschl**eu**st| und f**i**cht schon| w**ie**der|

Diese V-n-Verszeile besteht aus zwei zweisilbigen rhythmischen Wörtern (s. die Wortgrenze „|“) mit dem Akzent auf der zweiten Silbe (2.2), aus einem dreisilbigen rhythmischen Wort, wobei die zweite Silbe einen Wortakzent trägt (3.2), sowie aus einem zweisilbigen Wort, in welchem die erste Silbe akzentuiert ist (2.1). Also wird die Wahrscheinlichkeit dieser rhythmischen Variante von der Form I im SpM:

$$P_{2.2}*P_{2.2}*P_{3.2}*P_{2.1} = P_{In.}$$

Nach Angaben der Tabelle 4 (*Volksbuch*)

$$P_{In} = 0{,}1755*0{,}1755*0{,}1704*0{,}2534 = 0{,}0013 .$$

Eine Verszeile (V-m) der Form II (mit der unbetonten ersten Hebung):

v–v–´v–´v–´ (v): Um von Eug**e**n |Best**a**nd| zu l**e**rnen|

hätte im Sp-Modell die folgende Darstellung:

$$P_{4.4}*P_{2.2}*P_{3.2} = P_{IIm.}$$

Nach Angaben der Tabelle 4 (*Volksbuch*) ergibt sich

$$P_{IIm} = 0{,}0146*0{,}1755*0{,}1704 = 0{,}0004.$$

Jede rhythmische Form hat eine bestimmte Anzahl der Varianten, daher ist die Wahrscheinlichkeit jeder Form im Modell die Summe der Wahrscheinlichkeiten ihrer Varianten.

Im Rahmen der RM-Theorie wurden Ende der 80er und Anfang der 90er Jahre eine Reihe weiterer wahrscheinlichkeits–statistischer Modelle erarbeitet, darunter auch *das Abhängigkeitsmodell* (*SpMA*) (Krasnoperova 1981; 1992; 2000: 99–105). Im Gegensatz zum Sprachmodell ist die Wortwahl im Vers nach diesem Modell kontextsensitiv. Der Hauptunterschied zwischen beiden Modellen besteht also darin, dass die Auswahl der rhythmischen Wörter beim Verfassen einer Verszeile nach SpMA sowohl von der metrischen Position, Hebung (bzw. Senkung), als auch vom vorangehenden rhythmischen Kontext abhängig ist, wohingegen im SpM diese Auswahl unabhängig davon getroffen wird. Das heisst, dass die Wahrscheinlichkeit der Auswahl eines rhythmischen Wortes im SpMA eine bedingte Wahrscheinlichkeit ist: sie wurde nicht nur durch den Prosa-Wortschatz, wie im SpM, sondern auch durch die oben genannten kontextuellen Bedingungen bestimmt.

Dieser Studie liegt partiell der Apparat des Simulationsystems der RM-Theorie zugrunde. Der Versrhythmus wird also vor dem Hintergrund zweier Arten von wahrscheinlichkeitsstatistischen Versmaßmodellen (SpM und SpMA) untersucht. Die Verwendung von Sprach-

modellen soll es ermöglichen, die rhythmischen Besonderheiten der Dichtung aufgrund der theoretischen, modellierten Bedingungen zu erforschen.

Tabelle 5
Verteilung der rhythmischen Wörter in der deutschen Prosa

Typ des RW	Wortschätze											
	Volksbuch		Opitz		Grimmelsh.		Günther		Gottsched		Goethe	
	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.
1.1	234	0,2003	214	0,1486	251	0,1970	170	0,1420	404	0,1345	259	0,1881
2.1	296	0,2534	392	0,2722	344	0,2700	318	0,2657	836	0,2784	380	0,2760
2.2	205	0,1755	116	0,0806	135	0,1060	107	0,0894	303	0,1009	171	0,1242
3.1	32	0,0274	67	0,0465	60	0,0471	77	0,0643	161	0,0536	50	0,0363
3.2	199	0,1704	318	0,2208	229	0,1797	225	0,1880	595	0,1981	264	0,1917
3.3	51	0,0437	20	0,0139	35	0,0275	28	0,0234	43	0,0143	27	0,0196
4.1	7	0,0060	9	0,0063	12	0,0094	26	0,0217	40	0,0133	8	0,0058
4.2	30	0,0257	70	0,0486	41	0,0322	61	0,0510	141	0,0469	41	0,0298
4.3	37	0,0317	88	0,0611	58	0,0455	65	0,0543	123	0,0410	59	0,0428
4.4	17	0,0146	11	0,0076	12	0,0094	4	0,0033	17	0,0057	5	0,0036
5.1 5.2	3	0,0026	7	0,0049	10	0,0078	14	0,0117	33	0,0110	10	0,0073
5.3	20	0,0171	25	0,0174	20	0,0157	20	0,0167	51	0,0170	22	0,0160
5.4 5.5	11	0,0094	41	0,0285	17	0,0133	13	0,0109	27	0,0090	23	0,0167
6.2	0	0,0000	0	0,0000	1	0,0008	1	0,0008	0	0,0000	0	0,0000
6.3	0	0,0000	2	0,0014	1	0,0008	2	0,0017	17	0,0057	3	0,0022
6.4	4	0,0034	12	0,0083	2	0,0016	6	0,0050	18	0,0060	4	0,0029
6.5 6.6	1	0,0009	5	0,0035	2	0,0016	2	0,0017	8	0,0027	2	0,0015
7.3 7.4	0	0,0000	1	0,0007	1	0,0008	0	0,0000	3	0,0010	2	0,0015
7.5 7.6	3	0,0026	1	0,0007	0	0,0000	4	0,0033	4	0,0013	1	0,0007
8.4 8.6 8.7	2	0,0017	0	0,0000	1	0,0008	0	0,0000	0	0,0000	1	0,0007
3.1,3	3	0,0026	8	0,0055	5	0,0039	6	0,0050	16	0,0053	10	0,0073
4.1,3	6	0,0051	16	0,0111	16	0,0125	18	0,0150	91	0,0303	10	0,0073
4.2,4	4	0,0034	4	0,0028	3	0,0024	4	0,0033	12	0,0040	8	0,0058
5.1,3 5.1,4	0	0,0000	4	0,0028	3	0,0024	11	0,0092	11	0,0037	4	0,0029
5.2,4	2	0,0017	4	0,0028	6	0,0047	5	0,0042	19	0,0063	6	0,0044
5.2,5	0	0,0000	0	0,0000	0	0,0000	0	0,0000	0	0,0000	0	0,0000
5.3,5	1	0,0009	0	0,0000	2	0,0016	0	0,0000	2	0,0007	1	0,0007
6.2,4 6.3,5	0	0,0000	5	0,0035	6	0,0047	8	0,0067	22	0,0073	4	0,0029
Ande-re	0	0,0000	0	0,0000	1	0,0008	2	0,0017	6	0,0020	2	0,0015
Σ	1168	1,0001	1440	1,0001	1274	1,0000	1197	1,0000	3003	1,0000	1377	1,0002

Hinweis. Der Unterschied vom 1,0000 in 0,0001(2) in der Summe von relativen Zahlen entsteht durch das Aufrunden der Berechnungen im Excel-Programm.

Für die Berechnung der Modelle wurden rhythmische ‚Wortschätze' anhand deutscher Prosatexte zusammengestellt. Dabei wurden der stilistische Charakter und der zeitliche Rahmen des Materials berücksichtigt. Es kamen folgende Werke zur Verwendung: Volksbuch von „Historian und Geschicht Doctor Johannis Fausti" (als Beispiel für volkstümliche Literatur um 1580), gelehrte Prosa von Martin Opitz (mittlerer Stil, 1624), der Abenteuer- und Schelmenroman „Simplicissimus" von Hans Jacob Christoph von Grimmelshausen (niederer

Stil, Mitte des 17. Jahrhunderts), Briefe Johann Christian Günthers (epistolare Form, erstes Viertel des 18. Jahrhunderts) und die Lobrede auf Martin Opitz von Johann Christoph Gottsched (hoher Stil, 1739). Zur Analyse der entwickelten Literatursprache wurde ein Auszug aus Johann Wolfgang Goethes Roman „Wilhelm Meisters Wanderjahre“ (1821) herangezogen.[14] Der gesamte Umfang des zusammengestellten lexikalischen Korpus beträgt 9459 rhythmische Wörter. Die Verteilung der rhythmischen Wörter in den für die Analyse geeigneten Werken der deutschen Prosa ist in Tabelle 5 dargestellt.

Die vergleichende Analyse der rhythmischen Wortschätze mithilfe des χ^2-Kriteriums,[15] das den Grad der Abweichung (Divergenz) der Wortschätze anzeigt, stellt eine statistische Inhomogenität der Rhythmik der Prosatexte fest, was offenbar von der stilistischen Inhomogenität der analysierten Werke abhängig ist. Dabei zeigt der rhythmische Wortschatz des *Volksbuchs* von „Doktor Faustus“ die größte Abweichung von Erwartungswerten der verglichenen Prosa (Tabelle 6: $X^2 = 147{,}87$, FG = 70). Wenn man den Text des Faustbuches von der Analyse ausschlösse und die Rhythmik der verbleibenden Werke untereinander vergliche, bliebe dennoch eine bedeutende Divergenz (s. Tabelle 7: $X^2 = 184{,}74$**;** FG = 56).

Tabelle 6
Vergleichanalyse rhythmischer Wortschätze[16]

Autor/Wortschatz	**Anteil von X^2**
Faustbuch	147,87
Opitz	48,82
Grimmelshausen	22,39
Günther	39,35
Gottsched	71,29
Goethe	26,34
Insgesamt	**356,06**
Der kritische Wert von χ^2*	**90,53**
Freiheitsgrade	70

*Hier sowie in Tabelle 7 wird der kritische Wert der Größe χ^2 in Entsprechung zum Signifikanzniveau bei 5% angenommen.

Tabelle 7
Vergleich der rhythmischen Wortschätze in Gruppen

Autor/Wortschatz	**X^2**	**Der kritische Wert von χ^2**	**Freiheits-grade**
Opitz–Grimmelshausen–Günther–Gottsched–Goethe	184,74	74,47	56
Faustbuch–Grimmeshausen–Goethe	77,67	41,34	28
Grimmelshausen–Goethe	12,54	23,69	14

[14] Quellenbeschreibung ist bei „Quellen: Prosa“ zu sehen.

[15] Da die langen rhythmischen Wörter (5-silbige oder längere) in den analysierten Prosatexten nicht immer häufig sind, wurden einige bestimmte RW-Typen für die Vergleichanalyse der Wortschätze mit Hilfe des χ^2 gruppiert: alle 5-silbigen (ohne 5.3); alle 6-,7- und 8-silbigen; die Komposita 3.1,3 und 4.1,3; andere Komposita von 4.2,4 bis 6.3,5.

[16] Hier wurde die Vergleichanalyse der rhythmischen Wortschätze aufgrund der Angaben der Tabelle 5 mit Hilfe des Chiquadrat-Kriteriums durchgeführt: die Angaben der Tabelle 6 zeigen die entsprechenden Anteile der Abweichungen einzelner Wortschätze von den Erwartungswerten der Verteilung der rhythmischen Wörter im gesamten Prosamassiv.

Nach den Angaben der Tabelle 7 kann die Prosa Grimmelshausens und Goethes als homogen ($X^2 = 12{,}54$ beim kritischen Wert 23,69) aufgefasst werden, was die Homogenität der rhythmischen Stilistik des niederen Stils des 17. Jahrhunderts und der entwickelten Literatursprache signalisiert. Doch der Vergleich der Wortschätze von den beiden rhythmisch homogenen Prosawerken mit dem Wortschatz des Faustbuches zeigt wieder die Inhomogenität (s. Tabelle 6: der Wert von X^2 ist höher als der kritische Wert). Es ist also festzustellen: dieser zusätzlich durchgeführte Vergleich unter Berücksichtigung der oben vorgelegten Resultate ergibt, dass man die Rhythmik des Volksbuches als ‚eigenständig' bezeichnen muss, da sie sich von den übrigen rhythmischen Wortschätzen unterscheidet. An dieser Stelle übergehen wir die Frage nach der Partitionierung des Chiquadrats, da sich schon bei der Zusammensetzung der Klassen einige Verzerrungen ergeben.

Tabelle 8
Akzentprofile des Verses vor dem Hintergrund
der Sprachmodelle (SpM und SpMA)*

Hebung / Vers/Modell	1.	2.	3.	4.
Opitz	0,928	0,877	0,942	0,957
Dach	0,933	0,947	0,846	0,981
Gryphius	0,908	0,913	0,888	0,978
Günther	0,974	0,894	0,886	0,978
SpM Faustbuch	0,857	0,779	0,730	0,973
SpM Opitz	0,832	0,674	0,677	0,973
SpM Grimmelshausen	0,855	0,738	0,721	0,973
SpM Günther	0,882	0,698	0,653	0,973
SpM Gottsched	0,886	0,735	0,699	0,973
SpM Goethe	0,858	0,764	0,749	0,973
SpMA Faustbuch	0,921	0,913	0,880	0,973
SpMA Opitz	0,900	0,847	0,864	0,973
SpMA Grimmelshausen	0,914	0,870	0,875	0,973
SpMA Günther	0,933	0,815	0,854	0,973
SpMA Gottsched	0,937	0,839	0,881	0,973
SpMA Goethe	0,925	0,889	0,894	0,973

*Das SpM sowie das SpMA wurde mit Rücksicht auf die mittlere Betontheit der letzten (vierten) Hebung im Vers (97,32%) sowie unter Berücksichtigung der Anzahl männlicher und weiblicher Kadenzen (MK & WK) in den gegebenen Verstexten berechnet: der Anteil der MK beträgt 55,54%, der Anteil der WK dementsprechend 44,46%.

Die wahrscheinlichkeits-statistischen Modelle (das SpM und das SpMA) wurden auf der Basis der rhythmischen Wortschätze der Prosa konstruiert und mit dem realen Vers verglichen (s. Tabelle 8). In allen Fällen lagen die Werte des SpMA näher an der Versrhythmik als die des SpM: das demonstriert die Berechnung der Euklidischen Norm, die die Distanz zwischen den Vers- und Modellangaben in den Akzentprofilen zeigt (s. Tabelle 8, 9 sowie Abbildung 5).[17] Das vorliegende Resultat unterstützt eine der Thesen der RM-Theorie, die

[17] In der Praxis der russischen Verslehre ist es möglich, das SpM nach der höheren Akzentuierung des Jambus umzurechnen, weil dieses Modell oft die niedere Betontheit des Verses im Vergleich mit dem konkreten Text voraussieht, obwohl es sonst die Versrhythmik relativ gut beschreibt. Doch in dieser Untersuchung ist es auch nicht der Fall. Selbst wenn man hier die höhere Betonung der Hebungen im jambischen Vers gegenüber dem SpM berücksichtigt, indem man die Modelle nach den Versangaben so umrechnet, dass man jeden Träger der

besagt, dass sich das Modell der Abhängigkeit besser in Relation zu einer weniger entwickelten Verstechnik setzten lässt als das SpM (Krasnoperova 1989: 68–70; 2000: 109).

Tabelle 9
Vergleich der Arzentprofile des Verses und der Modelle (SpM&SpMA) mit Hilfe der Euklidischen Norm[18]

	Opitz	Dach	Gryphius	Günther
SpM(Fa)	24,34	21,77	21,23	22,63
SpM(Op)	34,72	33,66	32,69	33,51
SpM(Gri)	27,05	25,56	24,66	25,61
SpM(Got)	34,2	31,95	31,90	31,82
SpM(Gün)	28,48	26,26	26,01	26,12
SpM(Goet)	23,46	22,09	20,96	22,22
SpMA(Fa)	7,21	4,94	1,55	5,66
SpMA(Op)	8,79	10,68	7,00	9,04
SpMA(Gri)	6,87	21,92	4,48	21,43
SpMA(Got)	10,77	13,23	10,62	9,46
SpMA(Gün)	7,24	11,36	7,95	6,65
SpMA(Goet)	4,95	7,56	3,00	4,99

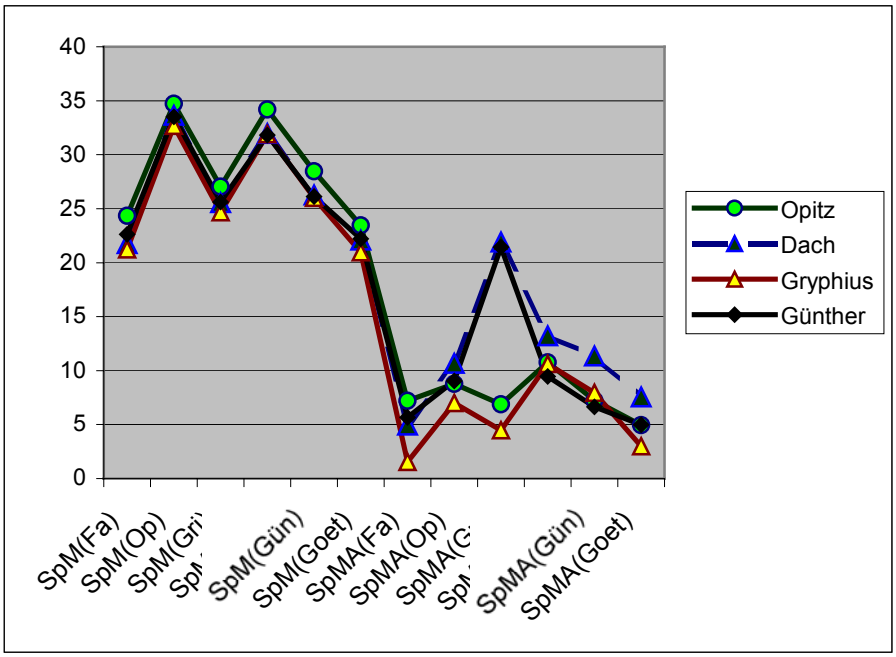


Abbildung 5. Graphische Darstellung der Abstände zwischen dem 4-Heber der deutschen Autoren und den Sprachmodellen (SpM&SpMA) nach Angaben der Tabelle 9.

Hebungen der Modelle nach dem Koeffizienten der Erhöhung der Betontheit (mittlerer Betontheit) des Verses multipliziert – wird eine solche Annäherung des SpM an den Vers immer noch schlechter, verglichen mit den unverändert gebliebenen Angaben des SpMA (vgl. Kazartsev 2002: 422).

[18] Die Euklidische Norm (Distanz) wurde nur für die ersten drei Hebungen nach Angaben der Tabelle 9 berechnet. Die Berechnung ist oben beschrieben (s. Anmerkung 10).

Die gewonnene Erkenntnis bezieht sich nicht nur auf den frühesten Jambus (M. Opitz), sondern auch auf die wesentlich spätere Dichtung (J.-Ch. Günther). Daher kann man annehmen, dass die Beispiele des deutschen Jambus, die untersucht werden sollen, unter den Bedingungen eines noch nicht entwickelten bzw. in der Entwicklung begriffenen Systems der Verstechnik entstanden sind. Diese Tatsache deckt sich mit den literatur-historischen Fakten, wonach der jambische 4-Heber im Laufe des 17. Jahrhunderts keine weitere Verbreitung in der deutschen Poesie gefunden hat.

Es lässt sich außerdem eine Entsprechung in der Genese des deutschen und des russischen syllabotonischen Verses beobachten. Die Resultate der Untersuchungen von M.A. Krasnoperova unter der Mitarbeit des Autors dieser Studie erlauben die Behauptung, dass der Rhythmus des frühen russischen 4-hebigen Jambus Michail V. Lomonosovs (1739–1743) eine größere Nähe zum SpMA als zum SpM aufweist (Krasnoperova 2000: 111, 118; Krasnoperova, Kazartsev, Muchin 2000a, 2000b; Kazartsev, 2001c, 2004b). Demzufolge lassen sich der frühe deutsche und der frühe russische Jambus mit ein und demselben Modell-Typ (SpMA) beschreiben.[19] Das heißt, dass der Versifikationmechanismus, der diesem Modell entspricht, für den deutschen sowie für den russischen 4-Heber in seinen Anfängen gleich wäre.

Man kann also annehmen, dass die grundlegenden Dichtungsprozesse im deutschen und russischen Vers eine Analogie aufweisen. Es sei aber eine weitere Frage gestellt: gibt es Gemeinsamkeiten beim deutschen und russischen Vers hinsichtlich der Auswahl der rhythmischen Quelle? Allerdings ist es im russischen Vers nicht gelungen, für die ganze Entstehungsphase eine einzige rhythmische Quelle zu finden (Kazartsev 2004a: 51). Welche rhythmischen Wortschätze haben aber die deutschen Dichter verwendet? Um diese Frage zu beantworten, sollte man eine Vergleichsanalyse der Versrhythmik mit den aufgestellten Modellen (SpMA) durchführen, die nach den verschiedenen rhythmischen Wortschätzen konstruiert wurden. Für den Vergleich von SpMA mit dem Texstkorpus der Verse wurde das χ^2-Kriterium verwendet (s. Tabellen 10 und 11).

Tabelle 10
Verteilung der rhythmischen Formen im Vers und im SpMA

Formen	**Vers**				**Modell (Prosa)**					
	Opitz	**Dach**	**Gryphius**	**Günther**	**Faustbuch**	**Opitz**	**Grimmels.**	**Günther**	**Gottsched**	**Goethe**
I	0,732	0,721	0,700	0,742	0,708	0,620	0,657	0,604	0,660	0,706
III	0,102	0,043	0,080	0,100	0,079	0,142	0,121	0,174	0,151	0,103
IV	0,051	0,144	0,108	0,112	0,105	0,108	0,107	0,123	0,096	0,089
andere	0,116	0,091	0,112	0,046	0,107	0,131	0,116	0,099	0,093	0,103

[19] Wie für den deutschen, so eignet sich auch für den russischen Vers nicht nur ein Typ der wahrscheinlichkeitsstatistischen Modelle (SpMA), sondern es tritt ein Untertyp auf: das „symmetrische" SpMA, das auf der Kombination zweier analoger Reihenfolgen zur Bildung der Verszeile basiert (Krasnoperova 2000; Krasnoperova, Kazartsev, Muchin 2000a,b). Dementsprechend wird bei der einen zuerst der Beginn der Zeile, danach das Ende, und schließlich die Mitte ausgefüllt. Bei der anderen ist es genau umgekehrt: zuerst wird das Ende, dann der Beginn, schließlich die Mitte der Zeile ausgefüllt. Die Wahrscheinlichkeit für die Ausfüllung der Zeile „vom Beginn", oder „vom Ende", liegt in beiden Fällen bei 0,5. Die Mitte der Zeile im Ja4 besteht aus zwei Positionen. Die Auswahl der Reihenfolge ihrer Ausfüllung realisiert sich in Analogie zur Reihenfolge der Ausfüllung der Anfangspositionen.

Tabelle 11
χ^2-Homogenitätstest für Vers- und SpMA-Angaben*

Modell (Prosa)	**Vers**				
	Opitz	**Gryphius**	**Dach**	**Günther**	**Günther****
Goethe	2,58	3,94	14,75	20,00	2,21
Faustbuch	5,02	0,16	6,85	21,41	1,70
Grimmelshausen	5,60	6,66	15,39	28,34	4,57
Opitz	8,78	15,90	22,69	45,81	13,45
Günther	14,12	27,84	26,00	46,05	30,01
Gottsched	7,04	16,49	22,24	27,15	13,85
Kritisch. χ^2 bei 5%	7,82				5,99
Freiheitsgrade	3	3	3	3	2

* Verglichen wurden die Werte der Verteilung der rhythmischen Formen (s. Tabelle 10).

** Hier sind die Ergebnisse des Vergleichs im Bezug auf die Verteilung nur der I, III und IV rhythmischen Form angegeben.

Durch Vergleichanalyse vom modellierten und realen Versrhythmus mit Hilfe von chi-Quadrat ließ sich feststellen, welches von den nach den verschiedenen Wortschätzen konstruierten Abhängigkeitsmodellen die Verteilung der rhythmischen Formen im Vers adäquater voraussagt. Es ist wichtig anzumerken, dass die Abweichung aller SpMA vom Vers vernachlässigbar erscheint, wenn man für die Ablehnung der Hypothese der Homogenität ein Signifikanzniveau mit der Fehlerquote von 5% zulässt. In diesem Fall erhalten alle dem Vers besser entsprechenden Modelle allerdings Konkurrenz von Seiten anderer Modelle. So eignet sich z.B. für Opitz neben dem Modell nach dem rhythmischen Wortschatz der Prosa von Goethe auch das SpMA nach den Angaben des Faustbuches und der Prosa von Gottsched.[20] Daher ist die Nähe eines konkreten Modells zu diesem oder jenem jambischen Vers, unter Berücksichtigung eines so niedrigen Signifikanzniveaus, nicht eindeutig zu interpretieren. Erhöht man jedoch das Signifikanzniveau, werden die wahrscheinlichsten Quellen der Versrhythmik offenbar. Es wäre sinnvoll in Entsprechung mit den vorherigen Forschungen in der Metrik[21] auch für diese Untersuchung ein relativ hohes Signifikanzniveau von 30% annehmen. Diese Festlegung stellte sich als zweckmäßig heraus, da in diesem Fall jedem analysierten Vers ein „entsprechendes" Modell aus der Gruppe der mehr oder weniger „homogenen" Modelle zugeordnet werden kann (s. weiter).

So zeigt sich bei der Analyse der Tabelle 12, dass für drei der vier Autoren (M. Opitz, A. Gryphius und J.-Ch. Günther) bei einem $P > 0.30$ die Auswahl des passenden Modells möglich wird. Nur für die Zuordnung der Verse Dachs zu einem konkreten Modell scheint dieser Schwellenwert zu hoch zu sein. Senkt man aber aus diesem Grund – hypothetisch – das Signifikanzniveau auf 20%, so stellt sich heraus, dass die Dichtung von Gryphius nicht nur zur Rhythmik des Faustbuches eine gewisse Nähe aufweist, sondern auch zur Rhythmik der Prosa Grimmelshausens und Goethes. Dies kann jedoch nicht stimmen, da bekannt ist, dass

[20] Andererseits ergeben die Modelle, die auf den rhythmischen Wortschätzen Goethes und Grimmelshausens basieren, keine bedeutende Abweichung vom Jambus Gryphius', so dass sie hier mit dem am besten für den Vers geeigneten SpMA, das auf dem Wortschatz des Faustbuches aufbaut (s. weiter), konkurrieren können. Bei diesem Signifikanzniveau stellen die SpMA's Goethes und Grimmelshausens bei der Versanalyse Günthers auch eine durchaus ernstzunehmende Konkurrenz für das Modell des Faustbuches dar.

[21] In der russischen Metrik wurde eine Homogenitätsschwelle mit einer Fehlerquote von 30% als Signifikanzniveau in den Untersuchungen von M.A. Krasnoperova und E.V. Kazartsev angenommen (siehe z.B. Kazartsev 2001b).

die Rhythmik des Faustbuchs von den Wortschätzen von Grimmelshausen und Goethe (die untereinander rhythmisch homogen sind) deutlich abweicht (s. oben).

Die Analyse der Angaben in Tabelle 12 zeigt, dass die rhythmische Sprachquelle bei allen deutschen Dichtern (außer sei M. Opitz) vermutlich das Faustbuch war, da das Modell, das nach dem Wortschatz des Faustbuches konstruiert wurde, die geringsten Abweichungen zum Vers der drei übrigen Dichter aufweist (das zeigt auch die euklidische Distanz, s. oben: Tabelle 9, Abbildung 5).

Tabelle 12
Reihenfolge von den SpMA, die dem Vers am nächsten sind

Versuchsreihen		**X²**	Kritisch. χ^2 bei 30%**	**P***	**FG**
Vers	**Modell (Prosa)**				
Gryphius	Faustbuch	0,16	3,67	0,98	3
	Goethe	3,95		0,27	3
	Grimmelshausen	6,66		0,08	3
Günther	Faustbuch	1,70	2,41	0,43	2
	Goethe	2,21		0,33	2
	Grimmelshausen	4,57		0,10	2
Opitz	Goethe	2,58	3,67	0,46	3
	Faustbuch	5,02		0,17	3
	Grimmelshausen	5,70		0,13	3
Dach	Faustbuch	6,85		0,08	3

* Fehlerquote bei der Ablehnung der Hypotese der Homogenität.

Wie aus Tabelle 12 hervorgeht, erhält man den niedrigsten Wert für χ^2 (0,16), wenn man die rhythmischen Konfigurationen, die dieses Modell voraussagt, mit dem Vers von Andreas Gryphius vergleicht. Bei dem angenommenen Signifikanzniveau ist eine solche Abweichung unbedeutend. Offensichtlich entspricht der von Gryphius beim Dichten verwendete rhythmische Wortschatz der Rhythmik des Faustbuches. Eine Ablehnung der Hypothese von der Homogenität von Vers und Modell würde in diesem Fall eine unwahrscheinlich große Fehlerquote (98%) zur Folge haben (s. Tabelle 12). Aus der Abbildung 6 wird ersichtlich, wie nahe die Akzentprofile des 4-Hebers von A. Gryphius und diesem SpMA beieinander liegen.

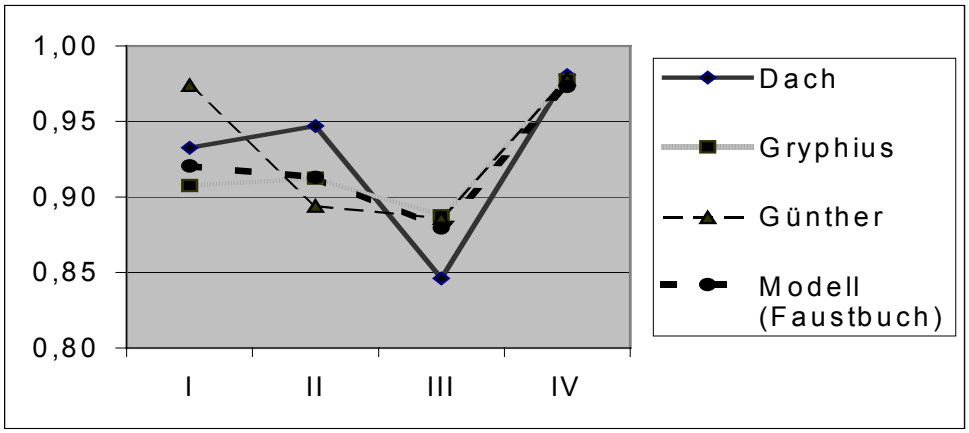


Abbildung 6. Akzentprofile vom Ja4 und vom SpMA

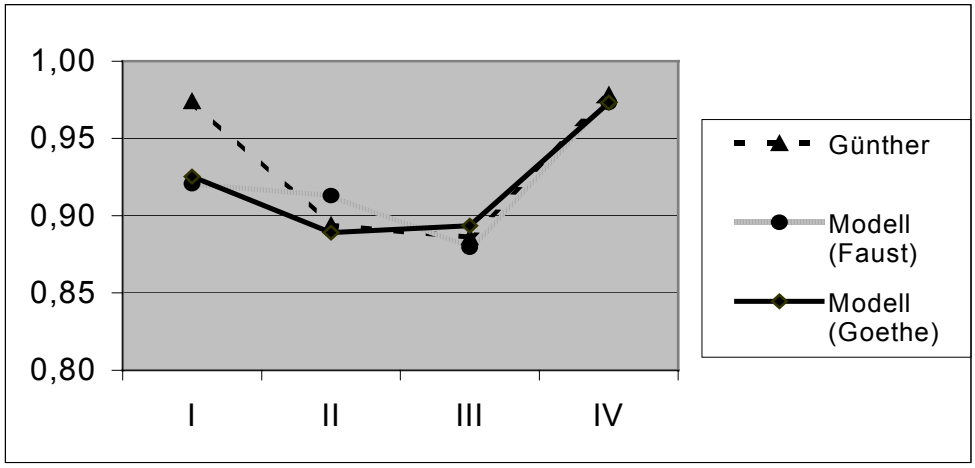


Abbildung 7. Akzentprofile vom Ja4 J. Ch. Günthers und vom SpMA

Auf den ersten Blick entspricht dem Vers von Johann Christian Günther kein bestimmtes Modell: der Wert von X^2 ist überall relativ hoch. Dabei können die Modelle, die auf der Prosa von Goethe und dem Faustbuch aufgebaut sind, seine Versrhythmik wesentlich besser beschreiben. Das Akzentprofil des Verses hat mit beiden Modellen Ähnlichkeit (s. Abbildung 7). Die größte Abweichung Güntherscher Verses von den Modellen kommt in den rhythmischen Konfigurationen II und VI vor, in denen die erste Hebung unbetont bleibt. Da die erste Hebung bei der Rezeption des Verses sehr wichtig ist, spielt die Betonung an dieser Stelle eine metrisch-konstruktive Rolle, und die ganze Verszeile kann durch ihr Nichtvorhandensein als eine „schlecht-metrische" markiert werden.[22] Die Anzahl dieser beiden Formen in den SpMA ist wesentlich höher als im Vers Günthers. Es ist ganz unverkennbar, dass Günther im Vergleich zu seinen Vorgängern diese Konfigurationen, besonders die zweite Form, vermeiden will (Tabelle 2).[23] Lassen wir diese Formen außer Acht und vergleichen nur die Verteilung der drei rhythmischen Hauptkonfigurationen – das sind die Formen I, III und IV – im Vers Günthers mit dem Modell (Tabelle 12), so ist festzustellen, dass das Modell, das nach dem rhythmischen Wortschatz vom Faustbuch berechnet ist, bei der Vergleichsanalyse anhand dreier Parameter auch dem Güntherschen Jambus am nächsten kommt: der Wert für X^2 ist sehr niedrig (1,70). Beim angenommenen Signifikanzniveau ist diese Abweichung unbedeutend (s. Tabelle 12). Die gewonnene Erkentniss unterstützt auch die Vergleichanalyse der Akzentprofile des Verses und des Modells mit Hilfe der Euklidischen Distanz (s. oben: Tabelle 9, Abbildung 5).

Allerdings kann dieses Modell doch eine Konkurrenz von Seiten der eigenen Verstendenz erhalten. So wurde oben bereits darauf verwiesen, dass sich gewisse Ähnlichkeiten zwischen den Jamben von Opitz und Günther konstatieren lassen, die sich in der Verteilung der vollbetonten rhythmischen Konfiguration (Form I) und in der Konfiguration mit dem Pyrri-

[22] Die Vorstellung, dass die Formen II und VI im Vers vermieden werden, lässt sich aus den Arbeiten von M. A. Krasnoperova ableiten, die der Verwendung von Typen rhythmischer Wörter mit einem langen unbetonten Anfang (LA-Wörter, wie z. B. „*Magaz***i**n") oder einer langen unbetonten Endung (LE-Wörter, wie z. B. „Kr**eu***zungen*") gewidmet sind. LA- und LE-Wörter sind auf bestimmten Positionen des Verses unterschiedlich häufig zu finden : LA-Wörter trifft man am Verszeilenanfang viel seltener an als die LE-Wörter (Krasnoperova, 2000: 148).

[23] Darin kann man eine Analogie mit dem russischen Jambus erkennen, da die Formen mit der unbetonten Anfangshebung auch im Vers von Lomonosov 1745-1764 herausgefiltert wurden (Krasnoperova 2000: 140, 143; 178-179). Der Terminus „herausfiltern" (auch „filtrieren" und „Filter") wurde in diesem Sinne erst in den Arbeiten von M.A. Krasnoperova im Rahmen von der RM-Theorie verwendet (siehe: Krasnoperova 2000: 69).

chius auf dem zweiten Versfuß (Form III) äußern. Was die andere wichtige rhythmische Konfiguration angeht, die einen Pyrrichius in der dritten Hebung zeigt (Form IV), kann die Vergrößerung der Anzahl dieser Form im Vers von Günther im Vergleich mit dem Vers von Opitz aufgrund der Vermeidung der anderen alternationsbildenden Konfigurationen (Formen II und VI) passieren (s. Tabelle 2). Nehmen wir also hypothetisch an, dass Günther von der Rhythmik des Opitzschen Verses ausgegangen sei, habe aber die Formen II und VI vermieden. Dann nehmen wir jenen Teil der Form II, welcher beim Dichten von Günther (verglichen mit dem Jambus von M. Opitz) herausgefiltert wurde, bekommen dann die Summe dieser Form und der Häufigkeit der Form VI bei M. Opitz und ergänzen damit die Frequenz der vierten Form in seinem Vers. Im Ergebnis bewegt sich die Frequenz der letzteren im Jambus von M. Opitz (0,105) nahe dem analogen Merkmal des Verses von J.-Ch. Günther (0,100).[24]

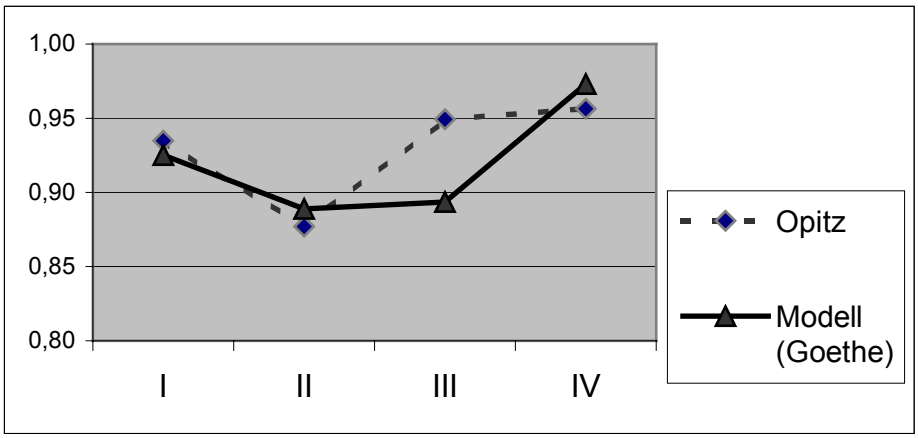


Abbildung 8. Akzentprofile vom Ja4 M. Opitzens und vom SpMA

Das Modell, welches auf der Grundlage des *besonderen* rhythmischen Wortschatzes des Faustbuches konstruiert wurde, eignet sich für die Dichtung S. Dachs ($X^2 = 6,85$, FG = 3) nicht so sehr wie für den Vers der übrigen *Praktiker* (Gryphius oder Günther), aber immer noch besser als die anderen Modelle (s. Tabelle 12). Die frühesten Beispiele für den Ja4 beim *Theoretiker* M. Opitz beschreibt nicht das SpMA, das nach den Angaben des Faustbuches berechnet wurde, sondern das andere Modell, welches auf dem rhythmischen Wortschatz Goethes aufgebaut ist: $X^2 = 2,58$, FG = 3 (nichtsignifikante Abweichung; s. Tabelle 12); der Vergleich der Akzentprofile zeigt auch die Ähnlichkeit von Vers und Modell in den zwei ersten Hebungen (s. Abbildung 8). Dieses Resultat unterstützt auch die Berechnung der Euklidischen Norm (s. Tabelle 9, Abbildung 5).

Ungeachtet dessen, dass die Rhythmik des Faustbuches und des Verses von Dach und Opitz divergiert, sollte man die Haltung von beiden Autoren zu dem Modell, das nach dem Wortschatz des Faustbuches gebaut ist, nicht als dieselbe betrachten. Wenn Opitzens Vers von den Daten des genannten Modells abweicht, so übertreibt der 4-Heber Dachs eine rhythmische Tendenz, die dieses Modell voraussagt. Eine charakterisierende Besonderheit dieses

[24] Praktisch verfahren wir folgendermaßen: Aufgrund der Tabelle 2 nehmen wir Unterschied zwischen der relative Häufigkeit der Form II im Vers von Opitz und von Günther plus relative Häufigkeit der anderen alternationsbildenen F VI (Opitz) und ergänzen damit das Merkmal der F IV beim Opitz (0,065 – 0,018 + 0,007 + 0,051 = 0,105). Das ist eine Rekonstruktion, die uns zeigen kann, wie häufig die F IV im 4-Heber von M. Opitz sein würde, wenn er sich nur auf diese alternationsbildende Form konzentrieren und andere solche Konfigurationen (eben so wie Günther) vermeiden würde.

Modells, die die Reduzierung der rhythmischen Konfiguration mit der unbetonten zweiten Hebung (Form III) und die statistische Erhöhung der Konfiguration mit der unbetonten dritten Hebung (Form IV) enthält, ist im Vers einfach überrepräsentiert: Dach reduziert die dritte Form und erhöht die vierte noch stärker als theoretisch vorgesehen (siehe Tabelle 8, was zur Abweichung des Verses vom Modell führt. Man kann also feststellen, dass die oben beobachtete deutliche Alternation der Rhythmik vom Jambus Dachs als die Übertreibung einer rhythmischen Tendenz betrachtet werden kann, die das ausgewählte metrische Modell (d.h. das SpMA nach der Rhythmik des Faustbuches) vorherbestimmt.

Zusammenfassung

Die durchgeführte quantitative Untersuchung gibt eine neue Vorstellung von der Rhythmik des deutschen Verses von 1620–1720. Zunächst wurden einige Tendenzen seiner rhythmischen Entwicklung innerhalb der Barockperiode festgestellt sowie eine Deutung dafür versucht. Es lässt sich eine Veränderung des rhythmischen Charakters des Jambus am Ende des Untersuchungszeitraums (Vers von J.-Ch. Günther) beobachten, die offenbar von einem Wechsel des rhythmischen Wortschatzes hervorgerufen wird. Bei dieser Analyse zeigt sich auch, dass der deutsche und der russische Jambus in ihrer Entstehungsphase bestimmte typologische Merkmale gemeinsam haben. Das gilt zum Beispiel für die Verstechnik, die das Abhängigkeitsmodell abbilden soll. Es gibt aber auch wesentliche Unterschiede, wie oben gezeigt wurde: im deutschen 4-Heber - im Gegensatz zu dem russischen - ist es möglich, eine gemeinsame rhythmische Sprachquelle (Faustbuch) für den gesamten Untersuchungszeitraum zu finden.

Zum Schluss fassen wir also alle Ergebnisse dieser Untersuchung zusammen:

1. Die Versrhythmik aller Dichter, die zu den Praktikern gezählt werden, lässt sich am besten mit dem Modell beschreiben, das auf der Rhythmik der „*Volkssprache“* basiert, die durch das Faustbuch repräsentiert ist. Dabei kann das verwendete Modell die Verteilung der rhythmischen Strukturen bei S. Dach erheblich schlechter voraussagen als bei A. Gryphius oder J.-Ch. Günther. Es gibt jedoch Gründe zur Vermutung, dass die Rhythmik von S. Dach sich doch durch das erwähnte Modell beschreiben lässt: die Abweichung von dem Modell lässt sich nämlich als eine übertriebene Repräsentation der theoretisch berechneten Daten interpretieren. Am besten sagt dieses Modell die Verteilung der rhythmischen Strukturen bei den Versen von A. Gryphius voraus.

2. Auch die Rhythmik des Jambus bei Günther entspricht am besten dem SpMA, das auf der Sprachrhythmik des Faustbuches basiert, sie ist aber auch ähnlich dem Modell, das auf der Grundlage der Prosa Goethes berechnet ist (20er Jahre des 19. Jahrhunderts). Demnach stellt das Wörterbuch der Literatursprache eine „Konkurrenz“ für das Faustbuch dar. Man kann Folgendes annehmen: Ungeachtet dessen, dass der Großteil des poetischen Wortschatzes des „letzten deutschen Barockdichters“ nach wie vor der *‚Volkssprache‘* verbunden bleibt, spiegelt sein Werk offensichtliche *‚progressive‘* Tendenzen in der Entwicklung der rhythmischen Stilistik der Sprache wider.

3. Bei der Weiterentwicklung der Syllabotonik wurden einige Besonderheiten des Verses von J.-Ch. Günther im Zusammenhang mit der Vermeidung bestimmter rhythmischer Konfigurationen ohne Akzent auf der ersten Hebung (die Formen II und VI) festgestellt, welche von seinen Vorgängern verstärkt fortgeführt wurden. Es zeigte sich auch, dass die

Rhythmik des *‚späteren‘* Jambus von J.-Ch. Günther gut aus den Daten der frühesten Muster des deutschen Jambus (M. Opitz) abgeleitet werden kann.

4. Der jambische Vers von M. Opitz – der als frühestes Beispiel für den deutschen Ja4 gilt – hat eine charakteristische rhythmische Tendenz (Übergewicht der Form III, im Vergleich zur Form IV und damit verbundene Erhöhung der prozentualen Betontheit der dritten Hebung gegenüber der zweiten). Diese Tendenz entsteht wahrscheinlich unter dem Einfluss der niederländischen Muster.

5. Die 4-hebigen Jamben in Martin Opitzens „Buch von der deutschen Poeterey“ stehen in der stilistischen Charakteristik des Rhythmus ebenfalls dem Modell nahe, das auf der Basis des rhythmischen Wortschatzes aus Goethes Prosa konstruiert wurde.

6. Die Nähe des Modells, das auf dem Wortschatz der späteren Literatursprache (Prosa von Goethe) aufgebaut ist, zu den Versen von Opitz und zum Teil auch zu denen von Günther, erlaubt die Annahme, dass die Entwicklung der rhythmischen Stilistik die Entwicklung der natürlichen Sprache vorwegnimmt. Diese Vermutung wird von den Untersuchungen, die M.A. Krasnoperova am frühen russischen Jambus vorgenommen hat, insofern unterstützt, als der Ja4 M.V. Lomonosovs (1745–1764) eine größere Nähe zum Modell aufweist, das auf dem rhythmischen Wörterbuch der späteren Literatursprache (Prosa Puškins) basiert (Krasnoperova 2000: 133). Ähnliche Unterschiede bei den stilistischen Tendenzen zeigen sich offensichtlich auch beim rhythmischen Charakter des deutschen Verses.

Literatur

Bailey, J. (1979). The Russian linguistic-statistical method for studying poetic rhythm: A review article. *Slavic and East European Journal 23(2), 251–261.*

Bailey, J. (1987). An exploration in comparative metrics. *Style 21(3), 359–376.*

Gasparov, M.L. (1996). *A History of European Versification*. Oxford, NY: Clarendon Press; Oxford University Press.

Heinsius, D. (1615). *Nederduytsche Poëmata*. Herausgegeben und eingeleitet von Barbara Becher-Cantarino. Faksimiledruck nach der Erstausgabe. Amsterdam 1616. Nachdrucke deutscher Literatur des 17. Jahrhunderts. Bd. 31: 3–66. Bern/Frankfurt M: Peter Lang (1983).

Kazartsev, E.V. (2001a). Novyje materialy k issledovaniju genezisa russkoj sillabo-toniki In: *Slavjanskij stich: lingvističeskaja i prikladnaja poetika: 63–71*. Moskva.

Kazartsev, E.V. (2001b). Ritmika pervoj duchovnoj ody M.V. Lomonosova v kontekste problemy genezisa russkoj sillabo-toniki In: *Formal'nye metody v lingvističeskoj poetike. Sbornik naučnych trudov, posv. 60-letiju professora M.A. Krasnoperovoj: 37–48.* St.-Petersburg: Izdatel’stvo S.-Peterburgskogo gosudarstvennogo universiteta.

Kazartsev, E.V. (2001c). *Ritm i jazyk v genezise russkoj sillabo-toniki (na materiale russkogo i nemeckogo sticha).* Avtoreferat dissertacii na soiskanie učenoj stepeni kandidata filologičeskich nauk. St. Petersburg.

Kazartsev, E.V. (2002). Die Anwendung linguistischer Statistik bei der Analyse des deutschen Verses. In: Christoph Küper (ed.), *Meter, Rhythm and Performance - Metrum, Rhythmus, Performanz. Proceedings of the International Conference on Meter, Rhythm and Performance: 413–424*. Frankfurt: Peter Lang.

Kazartsev, E.V. (2004a). Ritmika od M. V. Lomonosova i teorija trech stilej. In: *Lotmanovskij sbornik 3, 41–58.* Tartusskij universitet, Rossijskij gosudarstvennyj gumanitarnyj universitet, Moskva: Ob'edinennoje gumanitarnoje izdatel'stvo.

Kazartsev, E.V. (2004b). K voprosu o tipologii stanovlenija sillabo-toniki v nemeckom i russkom stichosloženii. In: *Fonetičeskie čtenija v čest' 100-letija so dnja roždenija L. R. Zindera (Essays in Phonetics in honour of the 100th anniversary of L. R. Zinder): 142–149*. St. Petersburg.

Kazartsev, E.V. (2006). K voprosu o stanovlenii sillabo-toničeskogo stichosloženija v Niderlandach: nekotorye osobennosti ritmiki 4-stopnogo jamba na fone jazykovych modelej razmera. In: *Formal Methods in Linguistic Poetics. A Collection of Scholary Works Dedicated to Professor M. A. Krasnoperova on the Occasion of her 65th Birthday.* St. Petersburg.

Kolmogorov, A.N. (1968). Primer izučenija metra i ego ritmičeskich variantov. In: *Teorija sticha: 145–167.* Leningrad: Nauka.

Kolmogorov, A.N.; Prochorov, A.V. (1985). Model' ritmičeskogo strojenija russkoj reči, prisposoblennaja k izučeniju ritmiki klassičeskogo russkogo sticha (vvedenije). In: *Russkoje stichosloženije. Tradicii i problemy razvitija: 113–133*. Moskva: Nauka.

Krasnoperova, M.A. (1981). *Model' vosprijatija i poroždenija ritmičeskoj struktury stichotvornogo teksta.* Avtoreferat kandidat. dissertacii. Leningrad.

Krasnoperova, M.A. (1989). *Modeli lingvističeskoj poetiki. Ritmika: Učebnoje posobie*. Leningrad: Izdatel'stvo S.-Peterburgskogo gosudarstvennogo universiteta.

Krasnoperova, M.A. (1992). *Rekonstruktivnoe modelirovanie stichosloženija na materiale ritmiki russkogo sticha.* Dissertacija doktora filolog. nauk. St. Petersburg.

Krasnoperova, M.A. (2000). *Osnovy rekonstruktivnogo modelirovanija stichosloženija (na materiale ritmiki russkogo sticha)*. St. Petersburg: Izdatel'stvo S.-Peterburgskogo gosudartsvennogo universiteta.

Krasnoperova, M.A.; Kazartsev, E.V.; Muchin, A.S. (2000a). Technika stichosloženija v modeljach razmera (k voprosu o stanovlenii russkoj sillabo-toniki). In: *Materialy XXIX mežvuzovskoj naučno-metodičeskoj konferencii prepodavatelej i aspirantov. Sekcija strukturnoj i prikladnoj lingvistiki: 23–27*. St.Petersburg.

Krasnoperova, M.A.; Kazartsev, E.V.; Muchin, A.S. (2000b). K istokam ritmiki russkogo sillabo-toničeskogo sticha (problema mežjazykovoj kommunikacii). In: Z. E. Žuravleva, V.A. Kopcik, G.J. Rizničenko. *Materialy meždunarodnoj naučnoj konferencii „Jazyki nauki – jazyki iskusstva"* (Moskva-Suzdal, 1999): *353–358*. Moskva: Progress-Tadicija.

Taranovskij, K.F. (1953). *Ruski dvodelny rytmovi.* Beograd: Naučna kniga.

Taranovskij, K.F. (1975). Rannie russkie jamby i ich nemeckie obrazcy. In: *Russkaja literatura 18 veka i ee meždunarodnye svjazi. Sbornik 10, 18 vek*: Nauka.

Tomaševskij, B.V. (1929*). O stiche. Statji*. Leningrad: Priboj.

Žirmunskij, V.M. (1968). O nacional'nych formach jambičeskogo sticha // *Teorija sticha.* Leningrad: Nauka.

Quellen:

Gedichte

Dach, S. (1650a), Christliches Denckmal dem Weiland Ehrn Vesten Achtbahrn und Wolgelarten HERRN DAVID TAUTEN etc. Welcher nach langwiriger vielerley Kranckheit newlich 3 Augustin. 1650 / sanfft und seelig sein Leben geendet / und darauff den 7. Christlich und ehrlich in der Tuhmkirchen im Kneiphof zur Erden bestattet worden. Königsberg in Preussen, 1650. – 7 S.

Dach, S. (1650b), Einfältiges Trost-Schrifftchen bey frühzeitigem und kläglichem Aber dennoch seligem Abschied aus dieser Welt des Liebreichen und Holdseligen Kindes SOPHIA Herrn Iohann Lösels Beydes der Philosophiae Artzney D. und Professoris publ. allhie etc. Und Frawen CATHARINE gebohrnenen Lepnerin Hertzlieben Töchterleins Welches / da es vierzeyn wochen alt war / sanfft und selig dieser Welt gutte Nacht gegeben / und indas ewige Frewden-Leben versetzet worden / den Hochbetrübten Eltern zu Trost verfertiget. Königsberg, 1650. – 8 S.

Gryphius, A. (1663), Andreae Gryphii Gedancken ueber den Kirchhof und Ruhestaedte der Verstorbenen (1657) // Andreae Griphii Trauer-Spiele auch Oden und Sonnette. Leipzig: gedruckt bey Johann Erich Hahn, S. 482–496.

Günther, J.-Ch. (1730), Auf den zwischen Ihro Kayserl. Majestaet und der Profte an 1718. geschlossenen Frieden // Sammlung von Christian Günthers, aus Schlesien, bis anhero edirten deutschen und lateinischen Gedichten, auf das neue übersehen, wie auch in einer bessern Wahl und Ordnung an das Licht gestellet. Breslau – Leipzig. – S. 123–137.

Opitz, M. (1619), Nachtklang // Martini Opicii Teutsche Poemata und Aristarchus. Strasburg, 1624. – S. 52–54.

Opitz, M. (1624), Ode „Derselbe welcher diese Nacht…“, Στροφη α, Αντιστροφη α, Στροφη β, Αντιστροφη β, Στροφη γ, Αντιστροφη γ // Buch von der Deutschen Poeterey. In welchem alle ihre eigenschafte zugehör gründlich erzehlet / und mit Exempeln ausgeführet wird. Breslau. – S. 44–45; 47–50.

Opitz, M. (1631), Auf Herzen David Müllers geliebten Söhnleins Davids Begräbnis, Breslau. Sonderdruk.

Opitz, M. (1637), der 118 Psalm // Die Psalmen Davids nach den Franzoesischen Weisen gesetzt durch Martin Opitzen. Danzig, 1637 – S. 334–339.

Prosa

Goethe, J.-W. (1894), Wilhelm Meisters Wanderjahre oder die Entsagenden // Goethes Werke. Herausgegeben im Auftrage der Großherzogin Sophie von Sachsen. Bd. 4 Weimar: Heman Böhlau. – S. 3–12.

Gottsched, J.Ch. (1749), Lob- und Gedächtnissrede auf den Vater der deutschen Dichtkunst, Martin Opitzen von Boberfeld, Nachdem selbiger vor hundert Jahren in Danzig Todes verblichen, zur Erneurung seines Andenkens im 1739sten Jahre den 20 August auf der philosophischen Catheder zu Leipzig gehalten // Herrn Johann Christoph Gottscheds… gesamelte Reden… Leipzig. – S. 156–165.

Grimmelshausen, H.J.Ch. von (1670), Der abendtheurliche Wiederum gantz neu umgegossene und mit seinem ewigwehrenden wunderbarlichen Calender / auch anderen zu seinem verbesserte SIMPLICISSIMUS Teutsch: Das ist die Beschreibung des Lebens eines seltzamen Vaganten genant Melchior Sternfels von Fuchshaim… Mompelgart: Fillion. – S. 3–13.

Opitz, M. (1624), Buch von der Deutschen Poeterey. In welchem alle ihre eigenschafte zuegehoer gruendlich erzehlet / und mit exempeln ausgefuehret wird. Breslau. – 56 S.

Günther, J.-Ch. (1934), Briefe: an Frau Dauling (?); an Herrn Christian Klugen Jun; an den jüngsten Herrn von Beuchelt; an Melchior Michael; an Gottlieb Rasper; Schreiben an Herrn M(ichael) in Landeshutt; an Herrn Hanns Gottfried von Beuchelt // Johann Christian Günthers Sämtliche Werke. Historisch-kritische Gesamtausgabe. Herausgegeben von Wilchelm Krämer. Bd. III: Freundschaftsgedichte und –Briefe in zeitlicher Folge. Leipzig: Verlag K. W. Hiersemann, 1934. – S. 143–150; 155–156.

Faustbuch (1963), Historian und Geschicht Doctor Johannis Fausti. Handschrift aus der Herzog-August-Bibliothek Wolfenbüttel (1580) // Das Faustbuch: nach der

Wolfenbüttler Handschrift. Herausgegeben von H. G. Haile. Berlin: Philologische Studien und Quellen; Schmidt. – S. 31–36.

Glottometrics 13, 2006, 23-46

Some aspects of word frequencies

Ioan-Iovitz Popescu, Bucharest
Gabriel Altmann, Lüdenscheid

Abstract. In the present article some new aspects of word frequency distributions are presented, namely the h-point, the k-point, the m-point, the n-point, the Lorenz curve and Gini's coefficient, all of which characterise some properties of texts. Their relationship to the measurement of vocabulary richness is scrutinized. It is an attempt at transferring some views from other domains of science to linguistics.

Keywords: word frequency, vocabulary richness, h-point, k-point,m-point, n-point, Lorenz curve, Gini's coefficient

1. Introduction

Calculation and presentation of word frequencies has many aspects, of which only a few have so far been taken into account: plain frequency, rank-frequency and type-token relation. Different models have been derived, and some characteristics have been set, in relation to vocabulary richness. However, the problem is much more complex. In this paper we will present a sketch of the field. Some kinds of counting are adequate for solving some problems but irrelevant for solving other ones, although in many cases no difference is apparent.

The first major distinction is between counting (a) word forms and (b) lemmas. Counting word forms is the usual procedure both for linguists and non-linguists because the written form affords the simplest and the most secure access to language. Lemmatisation is a complex job for professional linguists writing long programs, which work correctly up to about 90% of the time and thus may require some months of work correcting the program's errors made by the program. But even if the count is undertaken with pencil and paper, the identity of a lemma will remain an eternal problem and at least 10% of the result will be criticized by other linguists as incorrect. Hence, we may draw the interesting conclusion that lemmatisation programs are not so bad whatever they do. With lemma-counting, the vagueness of word identity becomes very conspicuous; while with word-form-counting the problem is only disguised, not solved. Consider for example German separable verbal prefixes which are described (by some linguists) as adverbs if they are not joined with the verb, but identified with prepositions in mechanical word-form-counting. Thus a verb can have double the number of forms in text that it actually has. Or consider the problem of suppletion existing in all European languages: we consider the cases of the German first person pronoun in German *ich, meiner, mir, mich* as forms of the same lemma but hesitate to include the plural form *wir* or the possessive form *mein*. In some languages there are two forms of *we* namely "we without you" and "we with you" (e.g. Indonesian *kami* and *kita*); do they belong to "I"? Hungarian has the lexeme *én* ("I") and forms like *nekem, hozzám, tőlem, engem,...* and a different plural. But in the third person *ő* ("he, she, it") there is a regular plural *ők*. In other languages there are other problems solvable only by using a set of criteria which are not inherent in data but applied as analytical means, i.e. our decisions.

Even if both forms of counting have their problems, for want of anything better they are used to describe some linguistic distribution problems (e.g. Zipf's law), today a very advanced discipline, some vocabulary richness problems, and type-token progressions. However, there are still other aspects that must be mentioned. (I) There are several units whose counting might be reasonable, namely (c) morphemes, accessible only to specialized linguists, and causing enormous difficulties in some languages; and (d) denotative units mostly called hrebs (c.f. e.g. Ziegler, Altmann 2002) which consist of all elements (morphemes, lexemes, phrases) denoting or referring to the same real or textual entity. They are very complex and are used for quite different purposes. Nevertheless, they are able to display a different aspect of vocabulary or style richness, namely the richness in expressing the same with different means. (II) The problem of plain word frequencies alone does in no case cover the full spectrum of frequency problems. It is merely the first step, the surface of the problem. Words of certain frequency can appear at special places in the sentence or in the text signalising their status (cf. e.g. Niemikorpi 1997; Uhlířová 1997); words of special frequency can be repeated at random or in quasi-regular distances (cf. Hřebíček 2000) that can be modelled; words of special frequency can have special forms or meanings – the association of length and meaning complexity with frequency are two of the oldest problems of quantitative linguistics (cf. Zipf 1935; Köhler 1986); and above all, frequency of words plays a basic role in the development of language and establishment of structures (cf. e.g. Bybee, Hopper 2001).

Here we shall restrict ourselves to the presentation of plain frequencies and searching for some possible characterizations. The presentation of frequencies can also be made in several ways. Up to now ten different ways have be proposed, though not all directly for word frequencies. Nevertheless, we shall discuss them all in turn.

(1) The *rank-frequency distribution of word forms* is the usual and the most popular presentation. It was originated by Zipf who supposed that *rank × frequency = constant*. Many empirical cases for which the harmonic series (truncated at the right side) holds can be found, and the theory has been strongly developed in the last decades. Here we ascribe rank 1 to the most frequent word, rank 2 to the second most frequent etc. It is believed that there is a law-like mechanism behind this order, even if in many cases it displays different forms. Some researchers have shown that even random texts (ape-typing) display the same behaviour (cf. Miller 1957; Li 1992). There is no contradiction in these conceptions because a background stochastic process working both in language and in non-language can result in the same distribution. Some researchers called the attention to the fact that synsemantics are situated at the beginning of the distribution, do not add much to the contents of the text, and have nothing to do with the problem of vocabulary richness. This has, again, two aspects: (I) There is no clear linguistic boundary between synsemantics and autosemantics (even using strict criteria); (II) in the rank-frequency presentation they are not strictly separated; there are always some autosemantics occurring at low ranks and some synsemantics occurring at high ranks. Hence, either one separates the word stock of the text and sets up two different curves for the two kinds of words, or one finds a point which approximately shows the separation line. As a matter of fact, such a point has been found by Hirsch (2005) and introduced in linguistics by Popescu (2006). It is very simple to find. In the rank-frequency presentation it is the point at which $r = f_r$, the point nearest to the origin [0,0], as illustrated in Figure 1.

The *h*-point can be set up approximately or one can find it by interpolation or by taking means of the neighbouring values. It is called *h-point* or *Hirsch-point* or, in linguistics, *Hirsch-Popescu-point*. See further below in 3.1.

Again there are several questions:

(α) has this point something to do with vocabulary richness?

(β) has it something to do with word class differentiation? and

(γ) does it depend on *N*, the text size?

The first question can be answered positively: the smaller the proportion of words occurring frequently, the higher the word proportion in the distribution tail in which infrequent words and hapax legomena occur. Hence, the smaller *h* is, the greater the richness.

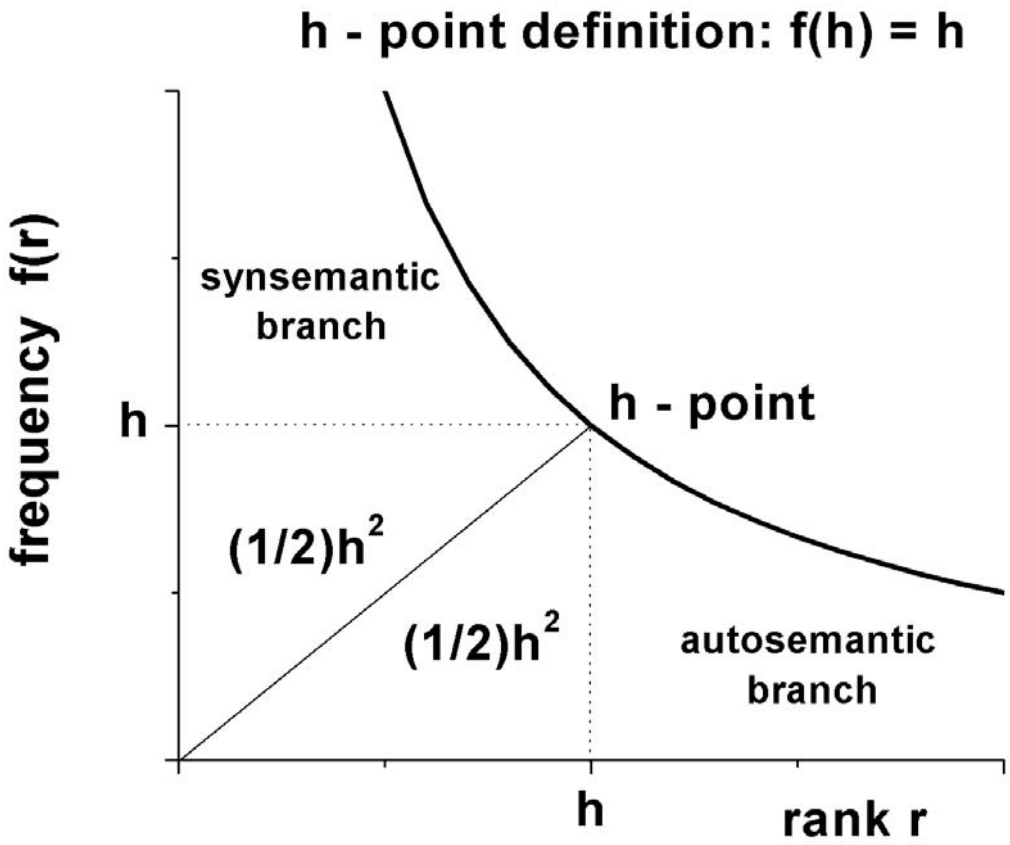


Figure 1. The *h*-point definition: the "bisector" point of the rank-frequency distribution at which rank = frequency

The second question has been analysed by several authors in different ways. Some assume the existence of two layers (auxiliary words and autosemantics); others even assume the existence of three layers, an idea which is not unrealistic; as a matter of fact, one may assume the existence of an arbitrary number of classes and in the infinity of texts some corroborating cases can always be found. We easily see that this problem can be solved by setting up confidence (or other) intervals for the *h*-point and employingh them as individual text characteristics.

The third problem is not yet solved. Theoretically, *h* must increase with increasing *N* because with increasing text, ever more words pass the $r = f(r)$-point while the addition of hapax legomena slows down. But this depends not only on the text but also on the analysed language. In highly analytic languages the situation will differ from that in highly synthetic ones. According to a private communication from B.D. Jayaram (2006), in two analysed Indian languages the share of the cumulative rank distribution up to the *h*-point does not change with increasing *N*. If this result could be ascertained in different languages, the *h*-coverage, or better the 1 - *F*(*h*) coverage, where *F*(*h*) denotes *the relative cumulative distribution up to h*, could be considered a stable characteristic of the vocabulary richness.

Since we obtain a different result (cf. 3.1), evidently this depends on the language, on the sample of texts used, on boundary conditions that are not known, and so on.

(2) The *rank-frequency distribution of lemmas* means a drastic change of circumstances. While in English the lemma of the article *the* is identical with its only word form, in German 7 forms of the article (of gender, case, number) belong to one lemma. Consequently, it is a very frequent word in all its forms, thus the *h*-point must lay at a greater rank, the corresponding $F(h)$ being greater and 1- $F(h)$ (which can also be considered as an index of vocabulary richness) respectively smaller. For the sake of illustration let us consider a German newspaper article (communication by R. Köhler 2006) in which the word-form rank-frequency distribution yielded $N = 1076$, $h = 13$ and $F(h) = 0.3383$, hence 1- $F(h) = 0.6617$. However, in the same text, there were found $N = 892$ lemmas with $h = 16$ and $F(h) = 0.7152$, i.e. there is a difference that must be levelled out. Thus no direct comparison is possible.

The next presentations are counted doubly, one holding for word forms, the other one for lemmas. From the linguistic point of view this difference is relevant.

(3) and (4). *The cumulative rank-frequency distribution* is usually presented in form of cumulative relative frequencies. They yield a concave curve (sequence). Since cumulative relative frequencies sum up to 1, arguing analogically we can find on this sequence again a point which is nearest to the point [0, 1]. However, here the rank itself does not play any role, nevertheless, $1 - F(h)$ could be used as an index of vocabulary richness. But again, counting lemmas we obtain a result which is not reasonably interpretable. Now, if we relativize also the ranks (calling them *rr*), i.e. divide each rank by the highest rank (V), we obtain the relation $<rr, F(r)>$, an analogon of the Lorenz curve used in economy and sociology. We shall present it below. The point we seek is given as the minimum of $[rr^2 + (1 - F(r))^2]^{1/2}$. The point of nearest distance will be called *m*-point. Its computation is presented in Figure 2.

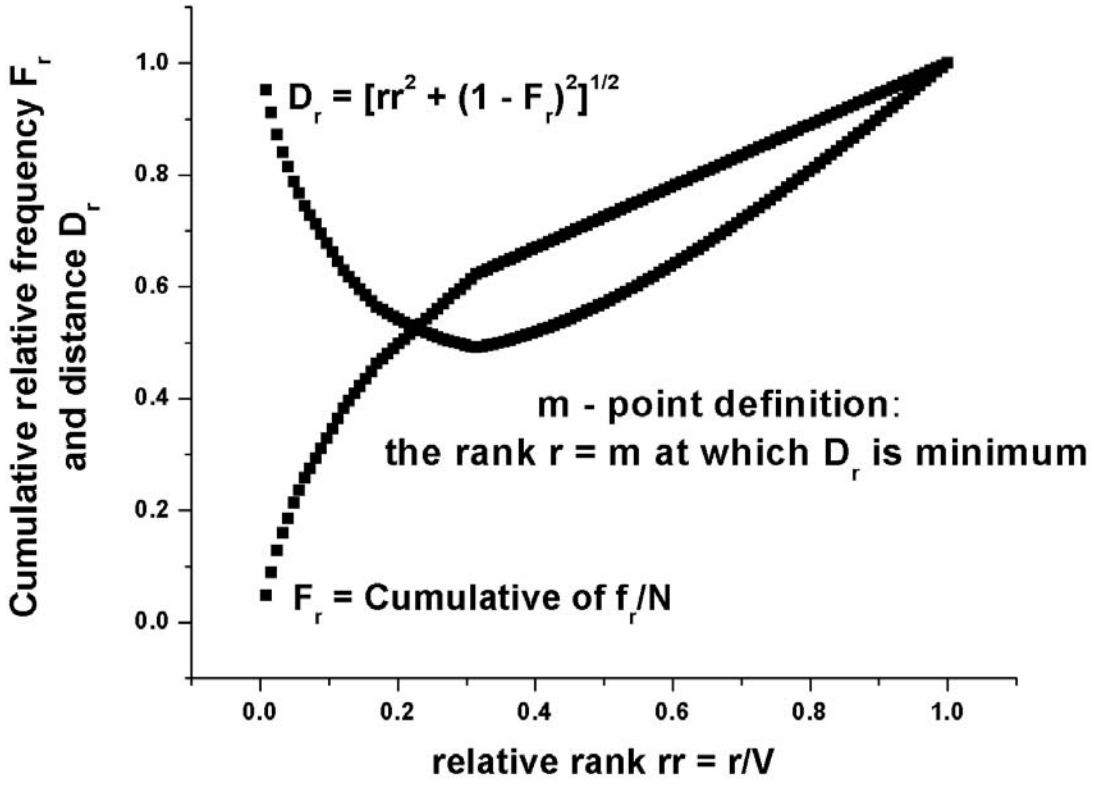


Figure 2. Computation of the *m*-point

Again, we can perform the same procedure for word forms and lemmas, obtaining two different results, and we can construct intervals for the *m*-point (minimum of the D_r curve) in Fig. 2. Further, we can approximate F_r using a continuous curve; the same can be done for D_r.

(5) and (6). *Reversed ranking*. Under some circumstances we begin to ascribe ranks in reverse order, i.e. the most infrequent word obtains rank 1, the next one rank 2, etc., and the most frequent word gets the highest rank. Then both ranks and frequencies are relativized: the ranks are divided by the highest rank yielding, say R_i ; and the relative frequencies are cumulated, yielding, say S_i (see 3.3). Thus we obtain the relation $<R_i, S_i>$ yielding a monotone convex sequence called Lorenz curve. If we add the point [0,0] to the data, then the sequence begins at zero and ends at 1. Again, we may seek the analogy to the *h*-point as that which is nearest to [1,0] using the same Euclidian distance as above; but there has been no investigation of this up to now. We do not know how text length affects the sequence even if both ranks and frequencies are relativized. Instead, we can proceed in the same way as it is done in other sciences. We start from the fact that if every word occurred in the text exactly once, i.e. if the text had maximal possible vocabulary richness, the Lorenz "curve" would be a straight line from [0,0] to [1,1], since relativized ranks would be equal to the cumulative relative frequencies, i.e. $R_i = S_i$. In this way we get at least a concept of maximal vocabulary richness. However, real texts differ from this maximum and the difference consists in the area between this straight line and the real Lorenz curve of the text. The area is usually called *Gini's coefficient* and can be computed as the sum of small trapezoids between the two lines. Since, in this way, everything is relativized, texts should be comparable independently of *N*. But even in this case, *G* can depend on *N* as will be shown below. However, in any case, the greater Gini's coefficient, the greater the disparity of word participation in text building, i.e. the smaller the vocabulary richness is. Since rank-frequency distributions of words follow a kind of Zipf law, this conclusion holds generally true; but there could be some not quite natural texts contradicting this interpretation. A wide-ranging investigation in many languages may bring a solution. Again, the difference between word-form counting and lemma counting could provide some surprises.

In long-tail distributions encompassing word counts, the point next to [1, 0] would be the point containing the hapax legomenon with the highest rank (remember that the ranking is reversed) or some of the words with frequency 2.

(7) and (8). *Ranking of differences.* This representation has been introduced by Balasubrahmanyan and Naranan (1996), as far as we know. Here the frequencies are decreasingly ordered and relativized, i.e. one obtains first all p_r. Then one defines a new variable $d_r = p_r - p_{r+1}$ and analyses its form. In this way one can obtain partial sums distributions used sporadically in linguistics and musicology (Wimmer, Altmann 2001). Again, word forms and lemmas can yield different results.

(9) and (10). *Frequency spectrum.* Every rank-frequency distribution can be easily transformed in a frequency spectrum – though the formal transformation is not always simple. But in empirical data it is easy to state that there are exactly f_1 words occurring once, f_2 words occurring twice, …, f_x words occurring *x*-times, simply by adding "from below". Some linguists call this representation as "frequency of frequency" or "lexical frequency". The random variable is "number of occurrences" and the frequency is the "number of words having the given number of occurrences". Again, there is a difference between word forms and lemma distributions. Since these monotone decreasing sequences have a hyperbolic form, the same operations can be performed as with ranked distributions. We can find the *h-point* analogue, named in the following *k-point*, and the analogous nearest point to [0, 1] for the cumulative distribution. But here the interpretation of the *k*-point is just the opposite: the greater $F(k)$ (denoting here *the relative cumulative distribution up to k*), the greater the vocabulary richness. Some authors constructed different indices based only on hapax leg-

omena ($x = 1$), other ones used the first 50 frequency classes ($x = 1,\ldots,x = 50$), but the k-point is a unique, objective point telling something about the dynamics of the sample. It is, of course, different for word forms and for lemmas but further investigation will show the relation between them based on the synthetism of language. For the time being, the study of texts is separated from the study of language but step by step we will bring them together and will be able to draw conclusions across the two fields.

2. Tasks and problems

In "plain" word frequency research different tasks have been formulated: (a) *Text characterization*, fulfilled using different statistics whose sampling properties are not always known, thus possibly rendering any conclusions irrelevant. If known statistics are used, we can draw relatively safe conclusions. Characterization is used, for example, for solving disputed authorship problems, for psychiatric purposes, for the study of language ontogenesis, for decisions about the affiliation of a text to a genre (and generally for establishing genres), for discourse analysis, for typologically characterizing a given language, and for drawing conclusions about other properties of language, e.g. vocabulary richness. The last task is so problematic that it has developed as an independent discipline encompassing a number of approaches, extensive literature and countless case studies (see point *d* below).

(b) A special part of the above task is the *comparison of texts* either by word for word comparison, global comparison using indices of some properties or comparing the frequency distributions (using e.g. chi-square statistics, information statistics, nonparametric tests etc.)

(c) *Information flow* is the study of increase of information in the course of the text. Usually we try to capture it by a curve or by an index of type-token-ratio. There are a great number of curves established by different arguments – and their number can still be indefinitely increased (they are monotonously increasing and concave) – but the background problems are so serious that there are always years of stagnation in the research. First of all, type-token relation (TTR) can be measured in three ways, of which only one can be set in connection with information flow (cf. Wimmer 2005); the others produce fractals. Second, it measures only a part of the new information because not only new words (types) furnish new information but also a new connection of old ones. Third, the type-token ratio reproduces a part of the hearer's information; that of the speaker is quite different (cf. Andersen, Altmann 2006). Fourth, there is an enormous difference between counting the increase of form-types and lemma-types. Counting form-types, strongly synthetic languages (such as Hungarian) seem to have a more rapid information flow than strongly isolating languages (such as English), which is nonsense. Thus lemma-type counting would be the only correct way; but most investigators count word-form-types because it is simpler. Fifth, the type-token curve or index is frequently used to measure the vocabulary richness of a text. On one hand, if counting forms, only intra-language comparisons are possible; on the other, all indices depend on N and their confidence intervals are enormous, hence judgements about vocabulary richness using TTR are problematic. Sixth, what *is* vocabulary richness? Does it have something in common with the velocity of increase of new types, or only their number in the text? The first drops absolutely with increasing text length, the second relatively. Does it have something to do with the extent of author's vocabulary? Surely not, because writing adults know all words of a language (except specialist terminology); it depends rather on the theme, the author's selectivity, the purpose of the text (e.g. didactic text, poetic text, newspaper text…). But how these entities can be measured? Thus discussion about the concept itself can be continued *ad infinitum* (cf. Wimmer, Altmann 1999). Intuitively we know that there should

be something like vocabulary richness but its operationalization is a matter of definitions and criteria. Below we shall show some possible ways differing from the known ones.

(d) *Modeling word frequencies* is the hobby of mathematicians shared seldom by linguists unless they want to show the adequate application of a model in a language. The problem goes beyond the boundaries of linguistics and Zipf's law is a well known concept in a great number of scientific disciplines. There are different approaches leading to different models and whole families of distributions. Baayen (2001, 2005) mentions urn models, LNRE (Large number of rare events) models and power models to which one can add the proportionality model from which curves and distributions, both discrete and continuous can be derived (cf. Wimmer, Altmann 2005). All models contain some truth but the problem of the meaning of parameters and the boundary conditions are no problems for mathematicians. Sometimes, they may be put in the drawer of ceteris-paribus conditions but this is no definitive solution.

A further problem of this research is the *homogeneity* of the analysed text, seldom taken into account even by linguists. Some linguistic problems can be solved using corpora, e.g. the study of grammatical usage, but for other problems, e.g. the study of vocabulary richness, a corpus is irrelevant. It is well known that in a novel consisting of several chapters, each chapter can turn out to be different from the previous ones in general or special aspects. The boundary conditions change not only in dependence on the theme but also in dependence on the pause in writing. The disposition of the writer is changed after a pause (e.g. coffee break or night), it brings new rhythms, e.g. word length or sentence length rhythm; the theme change causes jumps in the type-token curve, i.e. the new chapter brings a new stock of types which can destroy the regularity of the distributions found in the previous chapters. The heavy problem in textology is the fact that there are no populations having fixed values of properties (cf. Orlov, Boroda, Nadarejšvili 1982). There is nothing like "Goethe's sentence length" or "Shakespeare's word-frequency distribution". If such populations existed, then every sample from them should reflect the given property within an admitted confidence interval. That means that samples from that population must not differ significantly from each another and a pooled sample must still reflect the property of the population. It has been shown that in textology this is not true. Though some properties may remain constant in the course of a novel – and in that case it must be demonstrated – other ones may change, building a time series, a chaotic sequence, quasi-regular runs, a complex oscillating sequence or something else. In a drama even the speech of individual persons may display properties which are not in conformity with those of the drama as a whole (which is a very problematic population). Hence it is safer to study some textual properties in sufficiently large self-contained "natural" parts and if they are homogeneous (within the admitted confidence interval) to pool them and make statements about the text as a whole. Unfortunately, studies of the sequential character of texts are not popular and we still do not know which of the infinite number of text properties may be considered to be the stable ones (cf. e.g. Hřebíček 1997, 2000).

3. Characterizations

In the following sections we shall take into account word forms, allowing us an access to texts in different languages without the necessity of insecure lemmatisation; we shall strive for text characterization using some new methods and we shall try to draw conclusions concerning other text properties, e.g. vocabulary richness.

In the following we shall investigate four characteristic points of word frequency distribution:

1. The *h*-point, defined as that point on the rank-frequency distribution which is the nearest to the [0, 0], i.e. to the origin (see Figure 1)

2. The *k*-point which is defined in the same way but for the frequency spectrum. It differs from the *h*-point in its interpretation in connection with vocabulary richness.

3. The *m*-point concerns the cumulative distribution $F(m)$ of frequencies and one can, again distinguish two points, one for the rank-frequencies, the other for frequency spectra. A further differentiating property is taking ranks in absolute or in relative values. The *m*-point, in any form, is the nearest to the [0,1] point.

4. The *n*-point is computed only for rankings but the ranking is performed in reverse order, i.e. the smallest frequency has rank 1, the second smallest rank 2 etc. The ranks themselves are considered in relative values, i.e. $rr = r/V$. The *n*-point is the nearest to [1,0]. The cumulative values of relative frequencies yield the Lorenz curve, and the area between the bisector and the Lorenz curve is called Gini's coeffcient.

3.1. The *h*-point

The history of the *h*-index is short. It has been recently introduced by Hirsch (2005) in scientometrics as a tool for evaluating individual scientific output. Subsequently, one of us proposed its extension to linguistic analysis (Popescu 2006). The main reason requiring the introduction of the *h*-index in text evaluation matters may be summarized as follows. Let us consider a (more or less) Zipfian, hyperbolic rank-frequency distribution, as schematically shown in Fig.1. The area covered by the distribution curve represents the total word count or the text length (size), *N*, while the maximal rank gives the total number of unique, distinct, different words, that is the text vocabulary, *V*. It is obvious that there always exists a very special (rank, frequency) point, a "crossing point", at which the frequency is nearest to its rank, a value denoted by *h*. This is the definition of the *h*-point having the extraordinary property that it splits any vocabulary (*V*) into two word classes, namely in (1) a number of the order of magnitude of *h* of highly frequent synsemantic or auxiliary words (such as prepositions, conjunctions, pronouns, articles, adverbs, and others that are likewise meaningful only in the company of or reference to other words) and (2) usually a much larger number ($V - h$) of lowly frequent, seldom, autosemantic words, with $V >> h$, actually building the very vocabulary of the text. In other words, the *h*-point appears as a natural cutting point of the word-frequency distribution into two very distinct branches, namely the "rapid" branch, of relatively few (about *h*) unique words that are highly frequent, and the "slow" branch, of many (about $V - h$) unique words that are of low frequency. The role played by the *h*-point in separating autosemantics from synsemantics of a text reminds us of the tole of "Maxwell's demon" in physics in separating high velocity from low velocity gas molecules. In this connection, of particular interest in text analysis are relative cumulative frequencies up to the *h*-point, denoted by $F(h)$, and defined as the ratio of the area under the distribution curve from rank = 1 up to the rank = *h*, and the total area under this curve (the text length *N*). Let us illustrate this point by a simple example using the rank-frequency distribution of word forms in J.W.v.Goethe's poem "Erlkönig" as shown in Table 1.

As can easily be seen, the *h*-point is at rank $r = 6$ because $f(6) = 6$, too. Since $N = 225$, the $F(h) = (11 + 9 + 9 + 7 + 6 + 6)/225 = 0.2133$. If we want to express very approximately the vocabulary richness of the text, we must use $1 - F(h)$ yielding $1 - 0.2133 = 0.7867$ and add (or subtract) some index of synthetism constructed specially for this purpose. (For some synthetism indices see Altmann, Lehfeldt 1973).

Table 1
Rank-frequency distribution of word forms in Goethe's "Erlkönig"

Rank	Frequency	Rank	Frequency	Rank	Frequency	Rank	Frequency
1	11	11	4	21	3	31	2
2	9	12	4	22	2	32	2
3	9	13	4	23	2	33	2
4	7	14	4	24	2	34	2
5	6	15	4	25	2	35	2
6	*6*	16	3	26	2	36	2
7	5	17	3	27	2	37	2
8	5	18	3	28	2	38	2
9	4	19	3	29	2	39	2
10	4	20	3	30	2	40-124*	1

* The ranks 40 to 124 have frequency 1

As already Hirsch argued, there exists a simple relationship between h and the total area under the rank–frequency curve, namely

(1) $$N = a\, h^2$$

a being a constant of the word distribution and N the total word count of the considered text (the text length). From this it follows the difference equation

(2) $$\Delta N = 2ah\, \Delta h$$

which gives how the location of the h-point changes with increasing text length. Hence, by division, we eliminate the constant a and obtain a quite general relationship, namely

(3) $$\Delta N/N = 2\, \Delta h/h$$

In particular, in order to obtain an increase of $\Delta h = 1$, we need an increase of N given by $(\Delta N)_{\Delta h=1} = 2\, N/h$.

In Table 2 we show some results from self-contained complete Nobel lecture English texts of which, perhaps, only some were written in genuine English (see reference on *The Nobel Lecture*). However, a good sign that the selected text sample is quite homogeneous can be considered the stability of the ratio $a = N/h^2$ at an average value of 7.5 and a standard deviation of only 0.74. This quality can be appreciated by taking into consideration the large interval of the a-value, ranging from 4.5 to 9.5, for a great variety of literary texts (Popescu, 2006).

As discussed above, a reliable measure of vocabulary richness appears to be the quantity 1- $F(h)$, where $F(h)$ is *the relative cumulative distribution up to h*, and this value is given in Table 2. However, a fundamental correction should be made to the F-values as defined previously. This is related to the well known fact that there is always some ambiguity in separating synsemantics from autosemantics. The effective overlapping area of this genuine "linguistic Gordian knot" is of the order h^2, as it can be evaluated by a naked eye inspection of the rank-frequency distribution. It is clear that we overestimated so far the $F(h)$ synsemantics area and, correspondingly, underestimated the 1 - $F(h)$ autosemantics area. As a basic correction of our former $F(h)$ data, it appears quite naturally to cut the h^2 "knot" exactly into two parts, as suggested in Fig.1, one part to be subtracted from synsemantics and one part to be

added to autosemantics. Consequently, we finally have to correct the $F(h)$ of the rank-frequency distribution as follows

$$\underline{F(h)} = [Areah - h^2/2] / N = F(h) - h^2/2N \quad (4)$$

and, similarly, the $F(k)$ of the frequency spectrum distribution

$$\underline{F(k)} = [Areak - k^2/2] / V = F(k) - k^2/2V \quad (5)$$

Table 2
A sample of 33 Nobel lectures sorted by $1 - \underline{F(h)}$

Year	Field	Nobel Awardee	*N*	*V*	*h*	*k*	1 - *F*(*h*)	1 - $\underline{F(h)}$	*F*(*k*)	$\underline{F(k)}$	*a* = N/h^2
1996	Lit	Wislawa Szymborska	1982	826	16	6	0.7321	**0.7967**	0.9395	0.9177	7.74
2002	Peace	Jimmy Carter	2330	939	16	6	0.6996	**0.7545**	0.9414	0.9222	9.10
1986	Peace	Elie Wiesel	2693	945	19	6	0.6755	**0.7425**	0.9280	0.9090	7.46
1935	Chem	Irène Joliot-Curie	1103	390	12	6	0.6745	**0.7398**	0.9333	0.8871	7.66
1993	Lit	Toni Morrison	2971	1017	22	7	0.6382	**0.7197**	0.9351	0.9110	6.14
1976	Lit	Saul Bellow	4760	1495	26	7	0.6317	**0.7027**	0.9472	0.9308	7.04
1975	Med	Renato Dulbecco	3674	1005	22	8	0.6353	**0.7012**	0.9284	0.8966	7.59
1930	Lit	Sinclair Lewis	5004	1597	25	7	0.6325	**0.6950**	0.9474	0.9321	8.01
1959	Lit	Salvatore Quasimodo	3695	1255	21	7	0.6327	**0.6924**	0.9474	0.9279	8.38
1989	Econ	Trygve Haavelmo	3184	830	21	9	0.6209	**0.6902**	0.9373	0.8885	7.22
1986	Econ	James M. Buchanan Jr.	4622	1232	23	7	0.6326	**0.6898**	0.9221	0.9022	8.74
1989	Peace	Dalai Lama	3597	1030	23	7	0.6122	**0.6857**	0.9388	0.9150	6.80
1950	Lit	Bertrand Russell	5701	1574	29	8	0.6102	**0.6840**	0.9428	0.9225	6.78
1905	Med	Robert Koch	4281	1066	24	9	0.6157	**0.6830**	0.9306	0.8926	7.43
1953	Peace	George C. Marshall	3247	1001	19	8	0.6255	**0.6811**	0.952	0.9200	8.99
1970	Lit	Alexandr Solzhenitsyn	6512	1890	32	9	0.6023	**0.6809**	0.9524	0.9310	6.36
1975	Econ	Leonid V. Kantorovich	3923	1042	22	8	0.6179	**0.6796**	0.9367	0.9060	8.11
1983	Peace	Lech Walesa	2586	769	19	6	0.6079	**0.6777**	0.9129	0.8895	7.16
1902	Phys	Pieter Zeeman	3480	908	21	8	0.6118	**0.6752**	0.9273	0.8921	7.89
1973	Lit	Heinrich Böll	6088	1672	28	9	0.6107	**0.6751**	0.9474	0.9232	7.77
1991	Peace	Mikhail Gorbachev	5690	1546	26	11	0.6062	**0.6656**	0.9592	0.9201	8.42
1920	Phys	Max Planck	5200	1342	24	10	0.6002	**0.6556**	0.9508	0.9135	9.03
1984	Lit	Jaroslav Seifert	5241	1325	26	9	0.5903	**0.6548**	0.9404	0.9098	7.75
1963	Peace	Linus Pauling	6246	1333	28	10	0.5908	**0.6536**	0.9347	0.8972	7.97
1925	Med	John Macleod	4862	1176	24	9	0.5907	**0.6499**	0.9379	0.9035	8.44
1925	Med	Frederick G. Banting	8193	1669	32	11	0.5871	**0.6496**	0.9401	0.9039	8.00
2004	Lit	Elfriede Jelinek	5746	1038	33	8	0.5522	**0.6470**	0.8863	0.8555	5.28
1979	Peace	Mother Teresa	3820	636	26	9	0.5571	**0.6456**	0.8789	0.8152	5.65
1911	Chem	Marie Curie	4317	1016	25	9	0.5691	**0.6415**	0.9409	0.9010	6.91
1902	Phys	Hendrik A. Lorentz	7301	1423	31	9	0.565	**0.6308**	0.9178	0.8893	7.60
1938	Lit	Pearl Buck	9088	1825	39	10	0.5453	**0.6290**	0.9326	0.9052	5.98
1908	Chem	Ernest Rutherford	5083	985	26	12	0.5448	**0.6113**	0.9442	0.8711	7.52
1965	Phys	Richard P. Feynman	11265	1659	41	11	0.5337	**0.6083**	0.9066	0.8701	6.70

where N = text size, V = text vocabulary, and the wording *Areah* and *Areak* stands for the area under the distribution curve up to the h-point and k-point respectively. The correctness of this procedure can be appreciated by its consequences, that is by the excellent ranking rearrangement of the tabulated data, Table 2, as shown in the graphs of Fig. 3. The important conclusion is that the $F(k)$ index does not depend on N, while the $1 - F(h)$ index varies monotonously in the expected direction (downwards) in dependence on N (see Fig. 4) as can easily be computed from the data in Table 2.

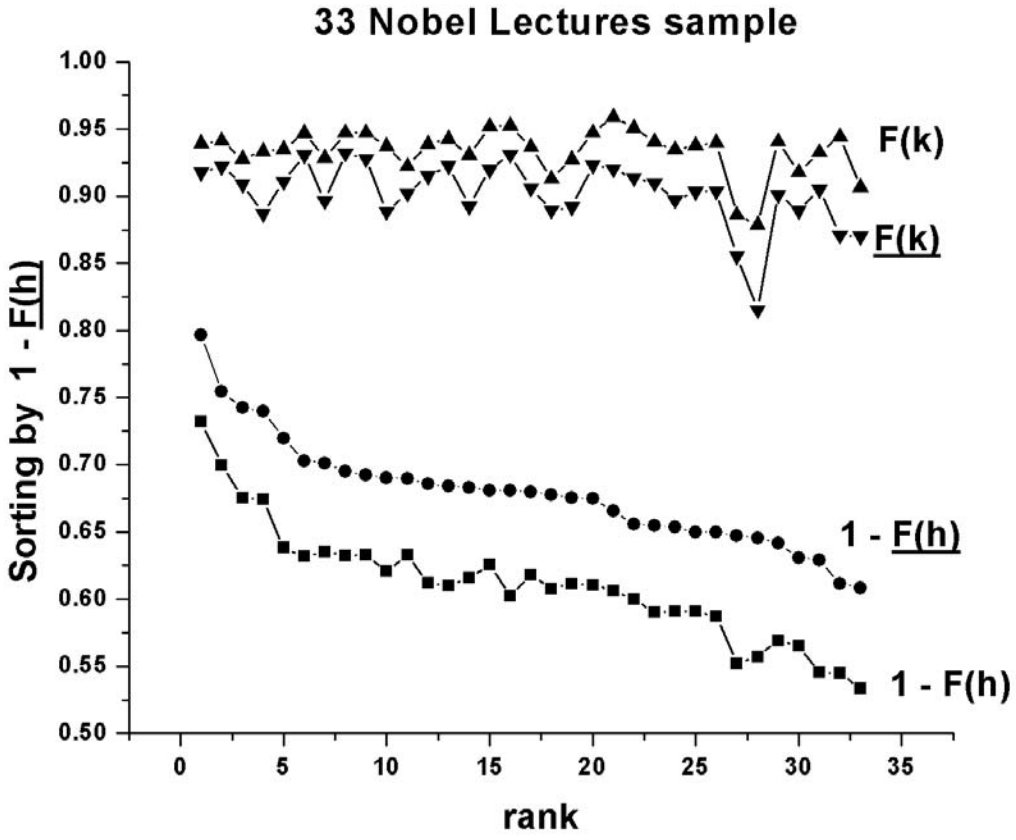


Figure 3. Some characteristics and ranking of a 33 Nobel lectures text sample

Generally, we shall point out that the boundary between auxiliary words and content words in actual texts is a fuzzy one rather than a sharp one, as it would appear from the above definition of the h-point at $f(h) = h$, with the corresponding relative cumulative frequency $F(h)$ up to h. Thus, for instance, for Goethe's "Erlkönig" rank-frequency distribution we found $h = 6$, $f(6) = 6$, and $F(6) = 48/225 = 0.2133$; while summing up all extant, real auxiliary words of this distribution, we get a value close to $F(17) = 92/225 = 0.4089$. Consequently, we have an ideal h-point, $h = 6$, defined merely on ideal symmetry grounds, and an effective H-point, $H = 17$, expressing the real boundary location and self-diffusion of synsemantics and autosemantics. Actually, any word frequency analysis in these terms should establish the relationship between these two cardinal points, the ideal, symmetric one, h-point, and the effective, real one, H-point.

The problem can be solved in different ways, e.g. setting up asymmetrical intervals, using order statistics or considering the two domains as fuzzy sets whose elements have different degrees of belonging in one of the domains. Here we shall consider only the problem of vocabulary richness. While $F(k)$ and $\underline{F(k)}$ do not depend (in our sample) of N, they can be

directly used as richness indices. In Table 3 we present an ordering of 33 Nobel lectures by *F(k)*. Though many natural scientists are in the first half and writers in the second, no conclusions can be drawn from this observation. In the same way, there is no relation of vocabulary richness to the year of origin. *F(k)* seems to be a quite independent index.

Table 3
33 Nobel lectures sorted by *F(k)*

Year	Field	Nobel Awardee	F(k)	N	Year	Field	Nobel Awardee	F(k)	N
1979	Peace	Mother Teresa	0,8152	3820	1975	Econ	Leonid V. Kantorovich	0,9060	3923
2004	Lit	Elfriede Jelinek	0,8555	5746	1986	Peace	Elie Wiesel	0,9090	2693
1965	Phys	Richard P. Feynman	0,8701	11265	1984	Lit	Jaroslav Seifert	0,9098	5241
1908	Chem	Ernest Rutherford	0,8711	5083	1993	Lit	Toni Morrison	0,9110	2971
1935	Chem	Irène Joliot-Curie	0,8871	1103	1920	Phys	Max Planck	0,9135	5200
1989	Econ	Trygve Haavelmo	0,8885	3184	1989	Peace	Dalai Lama	0,9150	3597
1902	Phys	Hendrik A. Lorentz	0,8893	7301	1996	Lit	Wislawa Szymborska	0,9177	1982
1983	Peace	Lech Walesa	0,8895	2586	1953	Peace	George C. Marshall	0,9200	3247
1902	Phys	Pieter Zeeman	0,8921	3480	1991	Peace	Mikhail Gorbachev	0,9201	5690
1905	Med	Robert Koch	0,8926	4281	2002	Peace	Jimmy Carter	0,9222	2330
1975	Med	Renato Dulbecco	0,8966	3674	1950	Lit	Bertrand Russell	0,9225	5701
1963	Peace	Linus Pauling	0,8972	6246	1973	Lit	Heinrich Böell	0,9232	6088
1911	Chem	Marie Curie	0,9010	4317	1959	Lit	Salvatore Quasimodo	0,9279	3695
1986	Econ	James M. Buchanan Jr.	0,9022	4622	1976	Lit	Saul Bellow	0,9308	4760
1925	Med	John Macleod	0,9035	4862	1970	Lit	Alexandr Solzhenitsyn	0,9310	6512
1925	Peace	Frederick G. Banting	0,9039	8193	1930	Lit	Sinclair Lewis	0,9321	5004
1938	Lit	Pearl Buck	0,9052	9088					

On the other hand, the quantity *F(h)* depends on N as shown in Fig. 4. The trend can be captured by the curve $1\text{-}F(h) = 1.5467N^{-0.0985}$ which, though yielding only a low determination coefficient $R = 0.60$, shows highly significant F and t values. Though this trend will surely be modified by adding several other texts in different languages, knowing the dependence we can use it for estimating the vocabulary richness of a text in a very simple way. Ignoring the absolute value of 1 – *F(h)* we consider only its difference to the computed curve which is a kind of norm (a kind of mean) for Nobel lectures. The results are shown in Table 4.

Table 4
Vocabulary richness measured by 1- *F(h)*

Nobel Awardee	*N*	1-*F(h)*	Theor	Diff	Nobel Awardee	*N*	1-*F(h)*	Theor	Diff
Irène Joliot-Curie	1103	0.7398	0.7757	-0.0359	Saul Bellow	4760	0.7027	0.6716	0.0311
Wislawa Szymborska	1982	0.7967	0.7322	0.0645	John Macleod	4862	0.6499	0.6702	-0.0203
Jimmy Carter	2330	0.7545	0.7206	0.0324	Sinclair Lewis	5004	0.6950	0.6683	0.0267
Lech Walesa	2586	0.6777	0.7132	-0.0355	Ernest Rutherford	5083	0.6113	0.6673	-0.0560
Elie Wiesel	2693	0.7425	0.7103	0.0321	Max Planck	5200	0.6556	0.6658	-0.0102
Toni Morrison	2971	0.7197	0.7035	0.0162	Jaroslav Seifert	5241	0.6548	0.6653	-0.0105
Trygve Haavelmo	3184	0.6902	0.6988	-0.0086	Mikhail Gorbachev	5690	0.6656	0.6599	0.0057
George C. Marshall	3247	0.6811	0.6974	-0.0163	Bertrand Russell	5701	0.6840	0.6598	0.0242
Pieter Zeeman	3480	0.6752	0.6927	-0.0175	Elfriede Jelinek	5746	0.6470	0.6593	-0.0123
Dalai Lama	3597	0.6857	0.6904	-0.0047	Heinrich Böll	6088	0.6751	0.8555	0.0196
Renato Dulbecco	3674	0.7012	0.6890	0.0122	Linus Pauling	6246	0.6536	0.6539	-0.0029
Salvatore Quasimodo	3695	0.6924	0.6886	0.0038	Alexandr Solzhenitsyn	6512	0.6809	0.6512	0.0299
Mother Teresa	3820	0.6456	0.6863	-0.0407	Hendrik A. Lorentz	7301	0.6308	0.6439	-0.0131
Leonid V. Kantorovich	3923	0.6796	0.6845	-0.0049	Frederick G. Banting	8193	0.6496	0.6366	0.0130
Robert Koch	4281	0.6830	0.6787	0.0043	Pearl Buck	9088	0.6290	0.6312	-0.0012
Marie Curie	4317	0.6415	0.6781	-0.0366	Richard P. Feynman	11265	0.6083	0.6170	-0.0087
J.M. Buchanan Jr.	4622	0.6898	0.6736	0.0162					

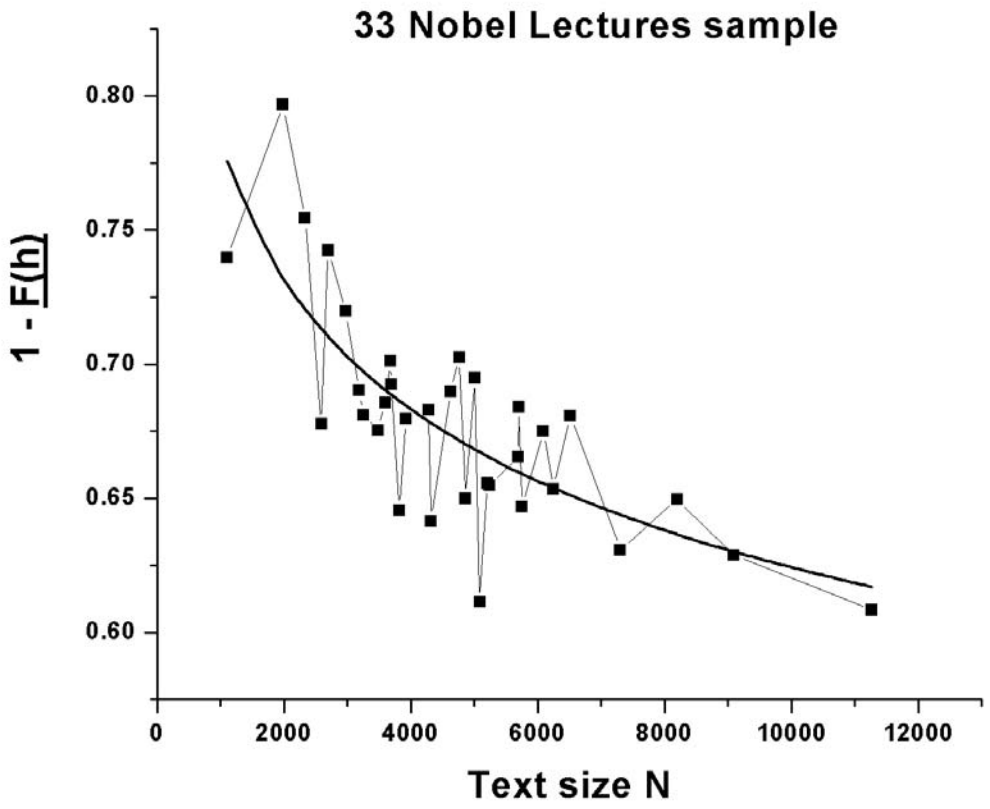


Figure 4. Dependence of $\underline{F(h)}$ on N.

The difference values can be normalized or relativized. However, the computed "theoretical" values hold only for the given data. Every new text would slightly change both the curve and the differences. Thus whatever the meaning of the above difference, vocabulary richness must be ascertained by means of several independent procedures.

3.2. The m-point

The *m*-point can be found if we directly scrutinize the cumulative rank-frequency distribution taking into account both relative frequencies and *relative* ranks. Since the sum of relative frequencies is 1 and ranks go from $1/V$ to $V/V = 1$, the curve approaches the point [1,1]. The *F*-curve has in our case a monotonously increasing concave form. Let us transform the values from Table 1 in *F* and present them as Table 5. In our case the relative ranks are $rr = r/124$ and F_r is the cumulative relative frequency.

Table 5
Cumulative distribution of ranked word forms of Goethe's "Erlkönig"

rr	F_r	rr	F_r	rr	F_r	rr	F_r
0.0081	0.0489	0.0887	0.3111	0.1693	0.4622	0.2500	0.5511
0.0161	0.0889	0.0968	0.3289	0.1774	0.4711	0.2581	0.5600
0.0242	0.1289	0.1048	0.3467	0.1855	0.4800	0.2661	0.5689
0.0322	0.1600	0.1129	0.3644	0.1935	0.4889	0.2742	0.5778
0.0403	0.1867	0.1210	0.3822	0.2016	0.4978	0.2823	0.5867
0.0484	0.2133	0.1290	0.3956	0.2097	0.5067	0.2903	0.5956
0.0565	0.2356	0.1371	0.4089	0.2177	0.5156	0.2983	0.6044
0.0645	0.2578	0.1452	0.4222	0.2258	0.5244	0.3065	0.6133
0.0726	0.2756	0.1532	0.4356	0.2339	0.5333	0.3145	0.6222
0.0807	0.2933	0.1613	0.4489	0.2419	0.5422	*	**

*from 40 up to 124 by step 1/124 = 0.0081
** by step 0.00444444

To find the given point we compute the Euclidian distance from [0, 1] as

$$D_r = \sqrt{rr^2 + (1 - F_r)^2} \qquad (6)$$

where rr is the relative rank, yielding $D_1 = [0.0081^2 + (1 - 0.0489)^2]^{1/2} = 0.9511$, $D_2 = 0.9113$, $D_3 = 0.8714$, $D_4 = 0.8406$, …, $D_{38} = 0.4934$, $D_{39} = 0.4916$, $D_{40} = 0.4934$; …, $D_{124} = 1.0000$. The smallest distance is at rank $m = 39$ where frequency 2 occurred for the last time. The $F_{39} = 0.6222$. The computation is shown in Fig. 5

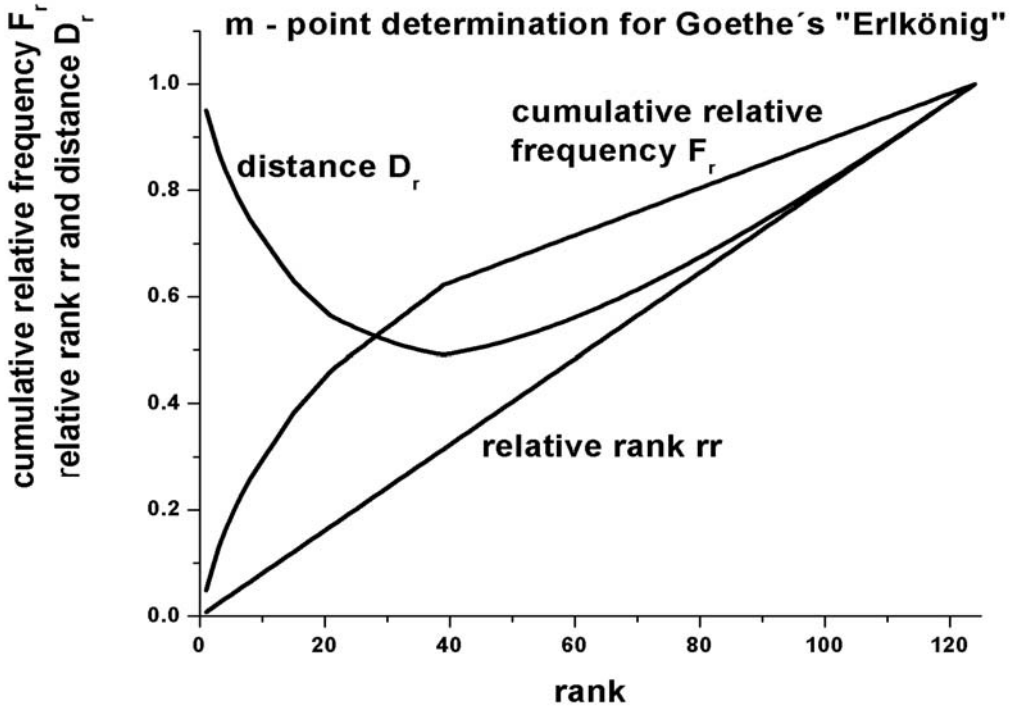


Figure 5. The m-point determination using the minimal distance to [0,1]

In Table 6, there are some results from the same texts as in Table 2, showing the *m*-point and the corresponding cumulative frequency *F*(*m*).

Table 6
The m-point and the cumulative frequencies

Year	Field	Nobel Awardee	*N*	*V*	*m*	1 - *F(m)*
1996	Lit	Wislawa Szymborska	1982	826	215	0,3133
2002	Peace	Jimmy Carter	2330	939	239	0,3163
1986	Peace	Elie Wiesel	2693	945	208	0,313
1935	Chem	Irène Joliot-Curie	1103	390	93	0,3191
1993	Lit	Toni Morrison	2971	1017	207	0,2975
1976	Lit	Saul Bellow	4760	1495	293	0,3008
1975	Med	Renato Dulbecco	3674	1005	225	0,2727
1930	Lit	Sinclair Lewis	5004	1597	307	0,3006
1959	Lit	Salvatore Quasimodo	3695	1255	258	0,3031
1989	Econ	Trygve Haavelmo	3184	830	179	0,2754
1986	Econ	James M. Buchanan Jr.	4622	1232	271	0,2746
1989	Peace	Dalai Lama	3597	1030	240	0,2722
1950	Lit	Bertrand Russell	5701	1574	325	0,2619
1905	Med	Robert Koch	4281	1066	213	0,2684
1953	Peace	George C. Marshall	3247	1001	207	0,2978
1970	Lit	Alexandr Solzhenitsyn	6512	1890	341	0,2816
1975	Econ	Leonid V. Kantorovich	3923	1042	229	0,2763
1983	Peace	Lech Walesa	2586	769	165	0,2769
1902	Phys	Pieter Zeeman	3480	908	191	0,269
1973	Lit	Heinrich Böell	6088	1672	337	0,2654
1991	Peace	Mikhail Gorbachev	5690	1546	334	0,265
1920	Phys	Max Planck	5200	1342	278	0,2671
1984	Lit	Jaroslav Seifert	5241	1325	264	0,2624
1963	Peace	Linus Pauling	6246	1333	278	0,2448
1925	Med	John Macleod	4862	1176	234	0,2742
1925	Med	Frederick G. Banting	8193	1669	334	0,2457
2004	Lit	Elfriede Jelinek	5746	1038	198	0,2183
1979	Peace	Mother Teresa	3820	636	129	0,2427
1911	Chem	Marie Curie	4317	1016	193	0,2699
1902	Phys	Hendrik A. Lorentz	7301	1423	273	0,2461
1938	Lit	Pearl Buck	9088	1825	337	0,2294
1908	Chem	Ernest Rutherford	5083	985	194	0,2506
1965	Phys	Richard P. Feynman	11265	1659	298	0,2274

It can easily be shown that *m* and 1-*F*(*m*) depend on *N* even if the determination coefficient is not too great. Hence it can be used in the same way as $\underline{F(h)}$ for estimating the vocabulary richness taking the difference between the computed and the observed points. However, here a theoretical approach would be more appropriate.

3.3. Gini's coefficient

The overall characterisation of word frequency distribution can be further performed by means of the Gini-coefficient used mostly in economics. As a matter of fact it shows us the deviation of the maximum vocabulary richness which would be attained if all words occurred exactly once. There are other indices doing this work, e.g. Herfindahl's repeat rate or Shannon's entropy or simply the skewness of the frequency distribution, etc.; but Gini's coefficient has not been used up to now.

If we present the distribution in cumulative form such that all frequencies are 1 (maximum vocabulary richness), then the cumulative frequencies would equal to the relative ranks, i.e. $rr_i = F_i$. In Cartesian coordinates this would yield a straight line of 45^o from [0,0] to [1,1]. Since in other sciences the ranking is performed in reversed form, we shall show it here, too. In programming this can be done mechanically, we shall present it explicitly. Starting from Table 1 we leave the ranks as they are but begin to rank frequencies "from below", i.e. the smallest frequency obtains rank 1, the second smallest rank 2 etc. At the same time we take both the relative values of ranks and the cumulative relative values of frequencies. The beginning of such a table is given in Table 7.

Table 7
Reverse ranking of frequencies in Goethe's "Erlkönig"

Rank r	Frequency f_r	Relative rank rr	Relative frequency p_r	Relative cumulative frequency F_r
1	1	1/124 = 0.00806	1/225 = 0.00444	0.00444
2	1	0.01613	0.00444	0.00888
3	1	0.02419	0.00444	0.01333
4	1	0.03226	0.00444	0.01778
5	1	0.04032	0.00444	0.02222
.....				
124	11	1.0000	11/225 = 0.04889	1.00000

As can be seen, the empirical values (of F_r) are positioned below this line (represented by rr). Hence Gini's coefficient is defined as the proportion of the area between the empirical F_r-values (last column) and the straight line between [0,0] and [1,1] to the whole area under the straight line, as shown in Fig. 3. The sequence of small straight lines $<rr_i, F_r>$ is usually called Lorenz curve. The whole area under the straight line is 0.5 and in our terms it means maximal vocabulary poverty (there is only one word steadily repeated in the text).

The area between the Lorenz curve and the straight line consists of small trapezoids. The area of a trapezoid is given as

$$A = \frac{a+b}{2}h \tag{7}$$

where a and b are the unequal sides and h is the height. The height $h = rr_{i+1} - rr_i$ while the two sides are given as

$$a = rr_i - F_i$$
$$b = rr_{i+1} - F_{i+1}$$

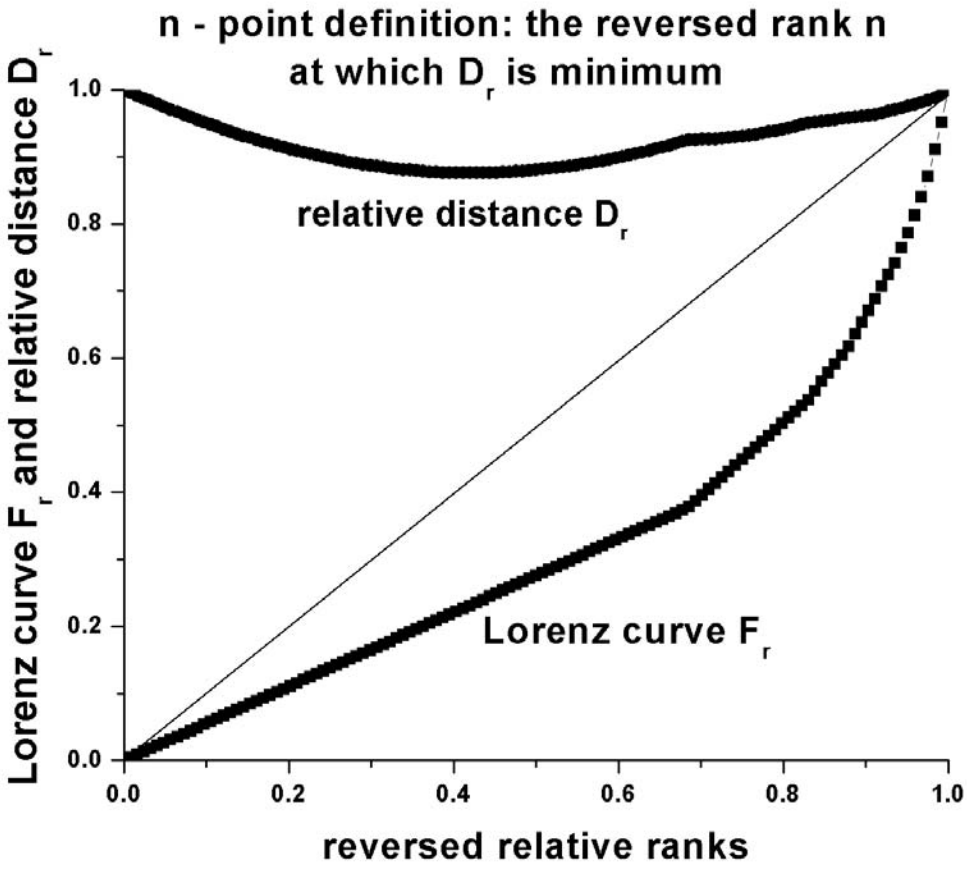


Figure 6. Lorenz curve. Reversed relative ranks ($rr = x$) against cumulative relative frequencies (y)

hence

$$(8) \quad A = \frac{(rr_i - F_i) + (r_{i+1} - F_{i+1})}{2}(rr_{i+1} - rr_i)$$

Consider the first trapezoid in the above scheme. We have

$$A_1 = \frac{(rr_1 - F_1) + (rr_2 - F_2)}{2}(rr_2 - rr_1)$$

$$= [(0.00806 - 0.00444) + (0.0161 - 0.00888)](0.0161 - 0.00806)/2 = 0.0000435.$$

and Gini's coefficients will be computed as follows. First we compute the sum of the trapezoids

$$(9) \quad G_1 = \sum_{i=1}^{V-1} \frac{(rr_i - F_i) + (r_{i+1} - F_{i+1})}{2}(rr_{i+1} - rr_i)$$

yielding $G_1 = 0.1828$. Then we compute the proportion of this area to the whole area under the straight line, i.e.

$G = 0.1828/0.5 = 0.3656.$

Fortunately, there are other methods for computing G without the necessity of reversing the frequencies and computing relative frequencies and cumulative frequencies. Consider V as the highest rank and N as text length, i.e. sum of all frequencies as given in Table 1 ($V = 124$, $N = 225$). Then we obtain G directly as

$$G = \frac{1}{V}\left(V + 1 - \frac{2}{N}\sum_{r=1}^{V} rf_r \right) \tag{10}$$

yielding exactly the same value as the first procedure. Still other variants are known. For $V >> 1$ it is approximately

$$G = 1 - \frac{2}{VN}\sum_{r=1}^{V} rf_r \,. \tag{11}$$

Gini's coefficient shows the position of the text between maximal and minimal vocabulary richness. Of course, the situation will differ with lemmatized texts and with the frequency spectrum. Some values for selected texts are given in Table 8.

Table 8
Gini's coefficient for 33 Nobel texts (rank-frequency) ordered according to N

N	1-G	N	1-G	N	1-G
1103	0.4479	3695	0.3959	5241	0.3337
1982	0.4786	3820	0.3067	5690	0.3504
2330	0.4605	3923	0.3610	5701	0.3453
2586	0.3741	4281	0.3415	5746	0.2823
2693	0.4216	4317	0.3281	6088	0.3455
2971	0.3975	4622	0.3615	6246	0.3163
3184	0.3629	4760	0.3826	6512	0.3514
3247	0.3846	4862	0.3372	7301	0.2982
3480	0.3517	5004	0.3796	8193	0.3047
3597	0.3740	5083	0.3068	9088	0.2826
3674	0.3579	5200	0.3422	11265	0.2640

Formula (9) is very practical for further computation and evaluation. Since the last expression in the formula is the arithmetic mean of the rank frequency distribution, one can derive the variance of G in a straightforward way as

$$Var(G) = \frac{4\sigma^2}{V^2 N} \tag{11}$$

and use it, for example, for setting up confidence intervals, comparisons with other texts, classifications or for studying text evolution, etc.

Needless to say, the Lorenz curve can be approximated by a continuous curve whose parameters can be used for comparisons, too.

In Table 8 we can see that 1-G depends in a high degree on N. A simple dependence yields 1-G = 2.68069$N^{-0.24196}$ with a determination coefficient R = 0.70 (and highly significant F and t tests) which can be improved by an additive constant; but since the computation is preliminary we show only the evaluation method. Again, we consider not the absolute value of 1-G but its difference to the "expected" value given by the above formula. Thus, e.g. for N = 3820 yielding 1-G = 0.3067 we obtain 1-G_t = 2.68069(3820)$^{-0.24196}$ = 0.3643. The difference (1-G) - (1-G_t) = G_t - G = 0.3067 - 0.3643 = -0.0576, i.e. the vocabulary richness of this text is smaller than expected. In the same way we can estimate the other texts and obtain the results in Table 9. Another possibility would be setting up confidence intervals for the above curve but its incessant change when adding new texts would force us to unending evaluations. Hence we show only the method.

Table 9
Evaluation of Gini's coefficient for 33 Nobel lectures
(ordered according text length N)

Year	Field	Awardee	N	1-G	1-G_t	Diff(G)
1935	Chem	Irène Joliot-Curie	1103	0.4479	0.4921	-0.0442
1966	Lit	W. Szymborska	1982	0.4786	0.4270	0.0516
2002	Peace	Jimmy Carter	2330	0.4605	0.4106	0.0499
1983	Peace	Lech Walesa	2586	0.3741	0.4004	-0.0263
1986	Peace	Elie Wiesel	2693	0.4216	0.3965	0.0251
1993	Lit	Toni Morrison	2971	0.3975	0.3872	0.0103
1989	Econ	Trygve Haavelmo	3184	0.3629	0.3808	-0.0179
1953	Peace	G.W. Marshall	3247	0.3846	0.3790	0.0056
1902	Ohys	Pieter Zeeman	3480	0.3517	0.3727	-0.0210
1989	Peace	Dalai Lama	3597	0.3740	0.3697	0.0043
1975	Med	Renato Dulbecco	3674	0.3579	0.3678	-0.0099
1959	Lit	S. Quasimodo	3695	0.3959	0.3673	0.0286
1979	Peace	Mother Teresa	3820	0.3067	0.3643	-0.0576
1975	Econ	L.V. Kantorovich	3923	0.3610	0.3620	-0.0010
1905	Med	Robert Koch	4281	0.3415	0.3544	-0.0129
1911	Chem	Marie Curie	4317	0.3281	0.3537	-0.0256
1986	Econ	J.M.Buchanan Jr.	4622	0.3615	0.3479	0.0136
1976	Lit	Saul Bellow	4760	0.3826	0.3455	0.0371
1925	Med	John Macleod	4862	0.3372	0.3437	-0.0065
1930	Lit	Sinclair Lewis	5004	0.3796	0.3413	0.0383
1908	Chem	E. Rutherford	5083	0.3068	0.3400	-0.0332
1920	Phys	Max Planck	5200	0.3422	0.3381	0.0041
1984	Lit	Jaroslav Seifert	5241	0.3337	0.3375	-0.0038
1991	Peace	M. Gorbachev	5690	0.3504	0.3309	0.0195
1950	Lit	Bertrand Russell	5701	0.3453	0.3307	0.0146
2004	Lit	Elfriede Jelinek	5746	0.2823	0.3301	0.0478
1973	Lit	Heinrich Böll	6088	0.3455	0.3255	0.0200
1963	Peace	Linus Pauling	6246	0.3163	0.3235	-0.0072
1970	Lit	A. Solzhenitsyn	6512	0.3514	0.3202	0.0312
1902	Phys	H.A. Lorentz	7301	0.2982	0.3115	-0.0133
1925	Med	F. G. Banting	8193	0.3047	0.3029	-0.0018
1938	Lit	Pearl Buck	9088	0.2826	0.2954	-0.0128
1965	Phys	R.P. Feynman	11265	0.2640	0.2805	-0.0165

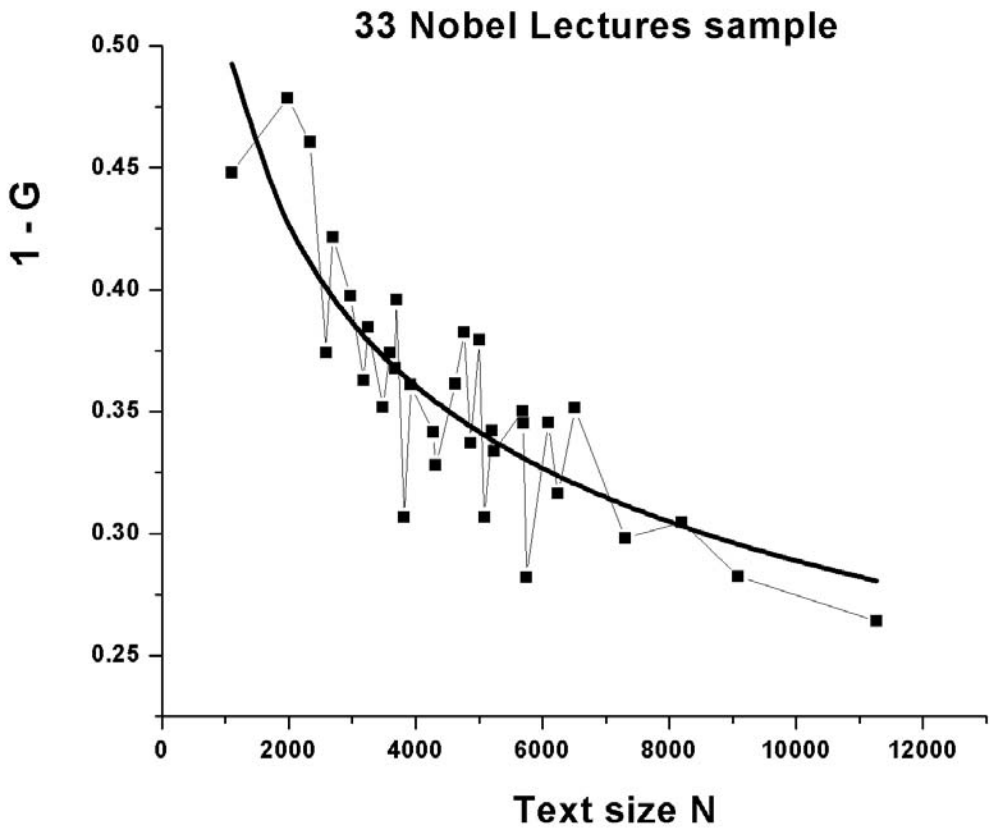


Figure 7. 1 – G in terms of text size N

As can be seen, the differences are not identical with those measured with F(h) but display the same trend. This is probably caused by the different expected trend. If we compare the two Figures (4 and 7), we can see the same course of observed values, hence one of the two coefficients is sufficient to characterize the vocabulary.

A complete table of all results and a graph are shown in the Appendix.

4. Conclusions

Our results, which may be considered first steps in a slightly different characterization of word frequencies, can be summarised as follows.

The possibility exists of distinguishing auxiliary words from content words using the *h*-point. With the aid of intervals, we could for every text ascertain the domain of pure auxiliary words, the mixed domain, and the domain of content words. On the other hand, we could use this for typological purposes, showing these domains in strongly analytical or synthetic languages. Though there will be always a difference between individual texts, it must be possible to ascertain general tendencies. Since h and $F(h)$ change with increasing N, this increase can be different in different languages.

As to vocabulary richness, the *k*-point of the frequency spectrum seems to be a relatively stable index not changing with increasing N; hence, $F(k)$ can be used for this purpose. It shows approximately the proportion of content words in the frequency spectrum. It would be interesting to compare texts of parents with those of their children, texts of people with

psychiatric disorders with those of "healthy" individuals, the texts of primitive literary forms with those of modern novels, poetic texts with scientific texts, texts of the same sort from different diachronic periods, or texts of a single writer at different points in his development, and so on.

The dependence of different text indices on N is a pitch for mathematicians as well as linguists. While researchers trying to characterize vocabulary richness strive for indices which are independent of N, other researchers want to see which characteristics are dependent on N. If we know how something changes in dependence on something else, we can eliminate the influence in an adequate mathematical way. But this is rather a task for mathematicians. Linguists usually take as many logarithms as necessary until they obtain curves whose differences look very insignificant. It would be better to study the sampling behaviour of indices, characterizing a well defined property. But how is vocabulary richness defined at all? No definition which could be directly operationalized can be found. While probability distributions of words can easily be compared, it is not so easy to compare the existing indices of vocabulary richness. Though the quantity $F(k)$ above seem to be very stable, and though we have a well-relativized variance of Gini's coefficient so that texts would be comparable in spite of their different sizes, we postpone this task, hoping to have inspired other researchers.

For typological purposes, all these indices are very useful. One might proceed as follows. Take a text, translate it into several languages, and compare the above indices at least by ordering the languages according to their magnitude. Then take other properties, and study their relation to the textual properties. Try to draw conclusions about the behaviour of $F(h)$, $F(k)$, G, etc. based on other properties of language.

References

Altmann, G., Lehfeldt, W. (1973). *Allgemeine Sprachtypologie*. München: Fink.

Andersen, S., Altmann, G. (2006). Information content of words in text. In: Grzybek, P. (ed.), *Contributions to the science of language: Word length studies and related issues: 93-117*. Boston: Kluwer.

Baayen, R.H. (2001). *Word frequency distributions*. Dordrecht/Boston/London: Kluwer.

Baayen, R.H. (2005). Word frequency distributions. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 397-409*. Berlin/New York: de Gruyter.

Balasubrahmanyan, V.K., Naranan, S. (1996). Quantitative linguistics and complex *s*ystem studies*. Journal of Quantitative Linguistics 3(3): 177-228.*

Bybee, J., Hopper, P. (eds.) (2001). *Frequency and the emergence of linguistic structure*. Amsterdam/Philadelphia: Benjamins.

Gini, C. (1912). "Variabilità e mutabilità" Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi (1955)., see at http://en.wikipedia.org/wiki/Gini_coefficient

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. http://arxiv.org/PS_cache/physics/pdf/0508/0508025.pdf, see more at http://en.wikipedia.org/wiki/Hirsch_number

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

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

Jayaram, B.D. (2006), private communication.

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

Köhler, R. (2006), private communication.

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

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

Niemikorpi, A. (1997). Equilibrium of words in the Finnish frequency dictionary. *J. of Quantitative Linguistics 4(1-3), 190-196.*

Orlov, J.K., Boroda, M.G., Nadarejšvili, I.Š. (1982). *Text, Sprache Kunst. Quantitative Analysen.* Bochum: Brockmeyer.

Popescu, I.-Iovitz (2006). Text ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text: 553-562.* Berlin/New York: de Gruyter.

The Nobel Lectures, http://nobelprize.org/nobel/

Uhlířová, L. (1997). Length vs. order: word length and clause length from the perspective of word order. *J. of Quantitative Linguistics 4(1-3), 266-275*

Wimmer, G. (2005). The type-toke-relation. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 361-368*. Berlin/New York: de Gruyter.

Wimmer, G., Altmann, G. (1999). On vocabulary richness. *Journal of Quantitative Linguistics 6(1), 1-9.*

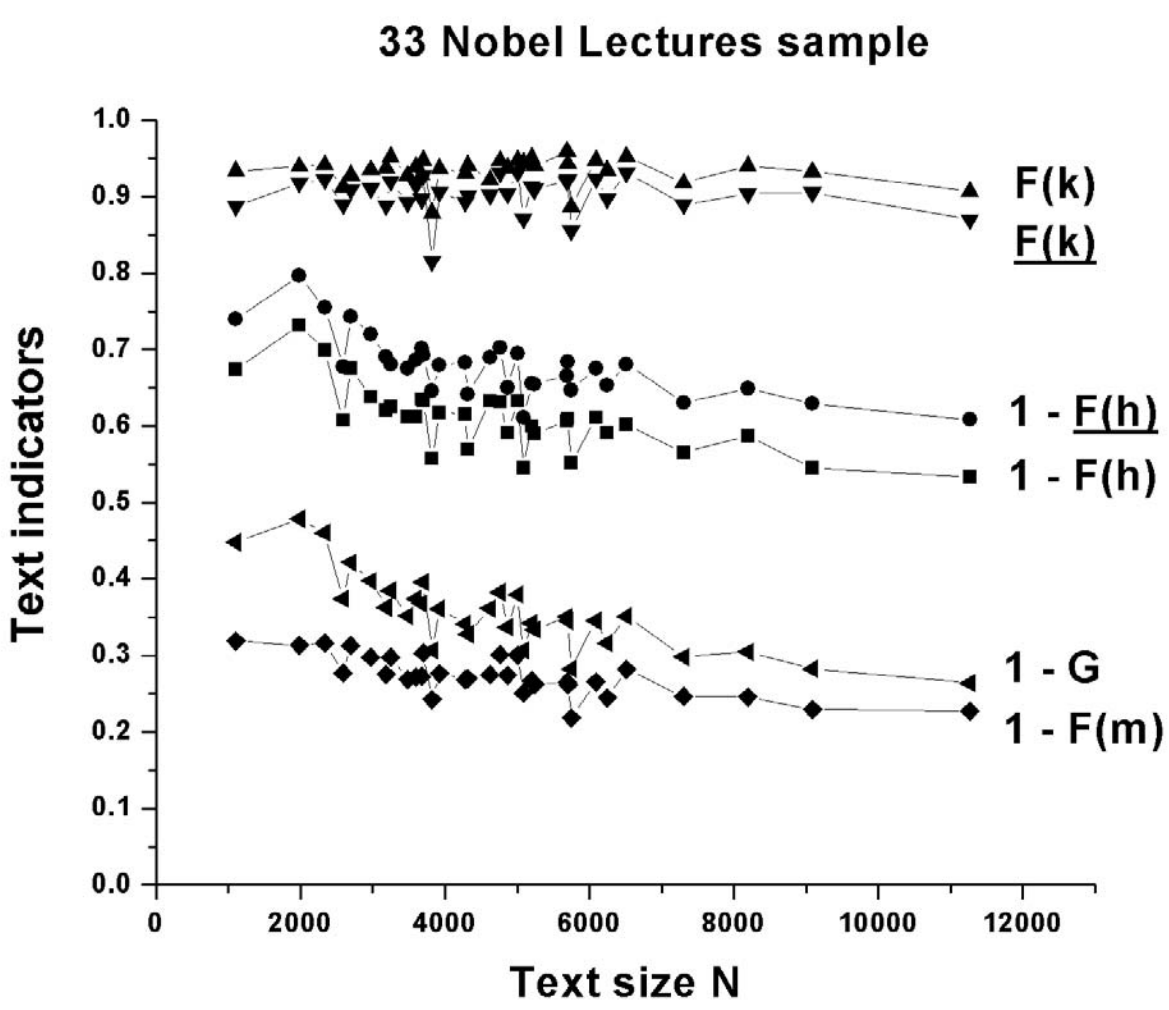
Wimmer, G., Altmann, G. (2001). Models of rank-frequency distributions in language and music. In: Uhlířová, L. et al. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček: 283-294*. Trier: WVT.

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 791-807*. Berlin/New York: de Gruyter.

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

Zipf, G.K. (1935). *The psycho-biology of language. An introduction to dynamic philology*. Boston: Houghton Mifflin.

Indicators of 33 Nobel lectures sorted by 1 - *F(h)*													
Year	Field	Nobel Awardee	*N*	*V*	*h*	*k*	*m*	1 - *F(h)*	1 - *F(h)*	*F(k)*	*F(k)*	*1 - F(m)*	*1 - G*
1996	Lit	Wislawa Szymborska	1982	826	16	6	215	0,7321	0,7967	0,9395	0,9177	0,3133	0,4786
2002	Peace	Jimmy Carter	2330	939	16	6	239	0,6996	0,7545	0,9414	0,9222	0,3163	0,4605
1986	Peace	Elie Wiesel	2693	945	19	6	208	0,6755	0,7425	0,928	0,9090	0,313	0,4216
1935	Chem	Irène Joliot-Curie	1103	390	12	6	93	0,6745	0,7398	0,9333	0,8871	0,3191	0,4479
1993	Lit	Toni Morrison	2971	1017	22	7	207	0,6382	0,7197	0,9351	0,9110	0,2975	0,3975
1976	Lit	Saul Bellow	4760	1495	26	7	293	0,6317	0,7027	0,9472	0,9308	0,3008	0,3826
1975	Med	Renato Dulbecco	3674	1005	22	8	225	0,6353	0,7012	0,9284	0,8966	0,2727	0,3679
1930	Lit	Sinclair Lewis	5004	1597	25	7	307	0,6325	0,6950	0,9474	0,9321	0,3006	0,3796
1959	Lit	Salvatore Quasimodo	3695	1255	21	7	258	0,6327	0,6924	0,9474	0,9279	0,3031	0,3959
1989	Econ	Trygve Haavelmo	3184	830	21	9	179	0,6209	0,6902	0,9373	0,8885	0,2754	0,3629
1986	Econ	James M. Buchanan Jr.	4622	1232	23	7	271	0,6326	0,6898	0,9221	0,9022	0,2746	0,3615
1989	Peace	Dalai Lama	3597	1030	23	7	240	0,6122	0,6857	0,9388	0,9150	0,2722	0,374
1950	Lit	Bertrand Russell	5701	1574	29	8	325	0,6102	0,6840	0,9428	0,9225	0,2619	0,3453
1905	Med	Robert Koch	4281	1066	24	9	213	0,6157	0,6830	0,9306	0,8926	0,2684	0,3415
1953	Peace	George C. Marshall	3247	1001	19	8	207	0,6255	0,6811	0,952	0,9200	0,2978	0,3846
1970	Lit	Alexandr Solzhenitsyn	6512	1890	32	9	341	0,6023	0,6809	0,9524	0,9310	0,2816	0,3514
1975	Econ	Leonid V. Kantorovich	3923	1042	22	8	229	0,6179	0,6796	0,9367	0,9060	0,2763	0,3610
1983	Peace	Lech Walesa	2586	769	19	6	165	0,6079	0,6777	0,9129	0,8895	0,2769	0,3741
1902	Phys	Pieter Zeeman	3480	908	21	8	191	0,6118	0,6752	0,9273	0,8921	0,269	0,3517
1973	Lit	Heinrich Böell	6088	1672	28	9	337	0,6107	0,6751	0,9474	0,9232	0,2654	0,3455
1991	Peace	Mikhail Gorbachev	5690	1546	26	11	334	0,6062	0,6656	0,9592	0,9201	0,265	0,3504
1920	Phys	Max Planck	5200	1342	24	10	278	0,6002	0,6556	0,9508	0,9135	0,2671	0,3422
1984	Lit	Jaroslav Seifert	5241	1325	26	9	264	0,5903	0,6548	0,9404	0,9098	0,2624	0,3337
1963	Peace	Linus Pauling	6246	1333	28	10	278	0,5908	0,6536	0,9347	0,8972	0,2448	0,3163
1925	Med	John Macleod	4862	1176	24	9	234	0,5907	0,6499	0,9379	0,9035	0,2742	0,3372
1925	Med	Frederick G. Banting	8193	1669	32	11	334	0,5871	0,6496	0,9401	0,9039	0,2457	0,3047
2004	Lit	Elfriede Jelinek	5746	1038	33	8	198	0,5522	0,6470	0,8863	0,8555	0,2183	0,2823
1979	Peace	Mother Teresa	3820	636	26	9	129	0,5571	0,6456	0,8789	0,8152	0,2427	0,3067
1911	Chem	Marie Curie	4317	1016	25	9	193	0,5691	0,6415	0,9409	0,9010	0,2699	0,3281
1902	Phys	Hendrik A. Lorentz	7301	1423	31	9	273	0,565	0,6308	0,9178	0,8893	0,2461	0,2982
1938	Lit	Pearl Buck	9088	1825	39	10	337	0,5453	0,6290	0,9326	0,9052	0,2294	0,2826
1908	Chem	Ernest Rutherford	5083	985	26	12	194	0,5448	0,6113	0,9442	0,8711	0,2506	0,3068
1965	Phys	Richard P. Feynman	11265	1659	41	11	298	0,5337	0,6083	0,9066	0,8701	0,2274	0,2640


33 Nobel Lectures sample
Text indicators
Text size N
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
2000
4000
6000
8000
10000
12000
F(k)
F(k)
1 - F(h)
1 - F(h)
1 - G
1 - F(m)

Glottometrics 13, 2006, 47-58

Wie viele Morphe enthalten Wörter in deutschen Pressetexten?

Karl-Heinz Best, Göttingen

Abstract. The paper presents word length data in German press texts with word length being based on the number of morphs. It can be shown that they follow the positive Cohen-negative binomial distribution (cf. Wimmer & Altmann 1999). In addition an excursus deals with the relation of the number of morphs and syllables in words.

Keywords: Morph distribution, German

0. Ziel der Untersuchung

Im Göttinger *Projekt Quantitative Linguistik* (http://wwwuser.gwdg.de/~kbest/) wurden bereits viele Untersuchungen zur Wortlänge in ca. 50 Sprachen durchgeführt (Best 2001; 2005b). In fast allen Fällen wurde dabei die Wortlänge nach der Zahl der Silben pro Wort bestimmt. In dieser Arbeit wird dagegen untersucht, wie viele Morphe in deutschen Wörtern vorkommen und wie sich die so gemessenen Wortlängen in Texten verteilen. Textgrundlage sind Pressemeldungen im *Eichsfelder Tageblatt (ET),* einer Lokalausgabe des *Göttinger Tageblatts.* Nur einmal, in Best (2001b), wurde eine entsprechende Untersuchung am Beispiel der Fabeln Pestalozzis durchgeführt. Es ist also sinnvoll, die Datenbasis zu erweitern.

1. Definition von „Wort“ und „Morph“

Zur Durchführung einer solchen Untersuchung ist es nötig, die Einheiten „Wort“ und „Morph“ zu definieren. „Wort“ wird wie auch in den anderen Untersuchungen des Göttinger Projekts als orthographisches Wort verstanden, d.h. als Buchstabenkette, die durch Leerstellen und/ oder Interpunktionszeichen begrenzt ist.

„Morph“ kann recht unterschiedlich bestimmt werden. Das „Morph“ ist eine Einheit der parole, die in dieser Arbeit nach folgendem Prinzip analysiert wird (vgl. dazu auch Best 2005a): Ein Wort lässt sich immer dann in zwei oder mehr Morphe oder Morphketten zerlegen, wenn wenigstens eines der gefundenen Segmente phonetisch und semantisch in gleicher oder mindestens ähnlicher Form und mit gleicher oder mindestens ähnlicher Bedeutung auch außerhalb dieses Wortes vorkommt.

Beispiele: *Besen-stiel*, *Ein-heit*, *geh-st*, *hör-bar*, *kind-lich*, *Kind-er*. Diese Wörter bestehen aus Morphen, die in genau der gleichen Form und mit genau der gleichen Bedeutung auch außerhalb dieser Wörter allein oder in Kombination mit anderen Morphen vorkommen. Das genannte Prinzip wird in diesen Fällen also in idealer Weise erfüllt.

Ein Regelsystem für die Morph-Segmentierung ist in Best (2001a: 2-5) entwickelt. Es tendiert in Zweifelsfällen wie *Trif-t*, *Kun-st* eher zur Segmentierung als dagegen, geht dabei aber nicht ganz so weit wie Kandler & Winter (1992ff.), die auch *Pet-er* und *eng-l-* (in „englisch“) vorsehen. Nach diesem Regelsystem wurde auch für die folgenden Analysen verfahren. Zur Veranschaulichung diene folgendes Beispiel aus dem *Eichsfelder Tageblatt* (= *ET*):

Text 2: Strom abgeschaltet – Frau stirbt in Klinik
Wladiwostok (ap). (*ET* 6.3.97, S. 8, Sparte: "Welt im Spiegel")

Kurz nach d-er Ge?bur-t ihr-es Kind-es is-t ein-e drei-ßig-jähr-ig-e Frau in Wladiwostok ge-storb-en, weil ihr-er Klin-ik wegen un-be-zahl-t-er Rechn-ung-en d-er Strom ab-ge-schalt-et wurd-e. D-ie Ge-rät-e fiel-en aus, als i-m An-schluß an ein-en Kaiser-schnitt ein-e weit-er-e Oper-at-ion not-wend-ig war. Es dauer-t-e fast ein-e Stund-e, bis d-er Energ-ie-ver-sorg-er auf d-ie dring-end-en Hilf-e-ruf-e d-es Klin-ik-person-al-s re-ag-ier-t-e.

Das Fragezeichen deutet eine für problematisch erachtete Segmentierung an, die aber bei der Datenerhebung akzeptiert wird. Bei allen Texten wird immer nur der laufende Text ohne Überschrift, Agentur-, Orts- und sonstige zusätzliche Angaben ausgewertet. Um die Text-dateien zu erstellen, wird anschließend ausgezählt, aus wie vielen Morphen die Wörter bestehen.

2. Anpassung eines geeigneten Modells

Nachdem die Dateien der zu untersuchenden Texte gemäß dem genannten Segmentierungsverfahren gewonnen sind, stellt sich die Frage, welche der theoretisch begründeten Verteilungen (Wimmer u.a., 1994; Wimmer & Altmann, 1996) getestet werden sollen. Es hat sich erwiesen, dass die positive Cohen-negative Binomialverteilung für die Pressetexte die vorläufig besten Ergebnisse ermöglicht. Die entsprechende Formel lautet:

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

Die Anpassung der positiven Cohen-negativen Binomialverteilung an die Wortlängen der 21 Pressetexte – bestimmt durch die Zahl der Morphe pro Wort – wurde mit dem *Altmann-Fitter* (1997) durchgeführt und erbrachte Ergebnisse wie in den folgenden Tabelle angegeben. Für alle Texte gelten folgende Vereinbarungen:

x: Zahl der Morphe pro Wort;
n_x: beobachtete Zahl der Wörter mit *x* Morphen Länge im betreffenden Text;
NP_x: aufgrund der positiven Cohen-negativen Binomialverteilung berechnete Anzahl der Wörter mit Länge *x*;
X^2: Chiquadrat;
P: Überschreitungswahrscheinlichkeit des berechneten X^2;
FG: Freiheitsgrade;
k, *p*, α: Parameter der Verteilung.

Die senkrechten Striche in den Tabellen zeigen an, dass in diesen Fällen die entsprechenden Wortlängenklassen zusammengefasst wurden. Die Testergebnisse der Anpassungen gelten als zufriedenstellend, wenn $P \geq 0.05$; diese Bedingung ist in allen Fällen mit Ausnahme von Text 17 erfüllt. Bei $FG = 0$ wird der Diskrepanzkoeffizient *D* als Testkriterium verwendet, der mit $C \leq 0.01$ ein gutes Ergebnis signalisiert.

Die Ergebnisse:

	Text 1		Text 2		Text 3	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	10	9.80	16	16.25	9	7.26
2	22	20.74	21	18.95	15	13.68
3	5	7.48	9	9.33	8	7.20
4	4	2.61	2	4.48	5	4.11
5	1	1.37	5	3.99	0	2.46
6					0	1.52
7					1	0.95|
8					1	1.82|
k =	1.4525		1.3332		0.2321	
p =	0.6864		0.5566		0.2924	
α =	0.8183		0.5566		0.7686	
X^2 =	1.730		1.882		4.514	
FG =	1		1		3	
P =	0.19		0.17		0.21	

Text 1: Sieben Deutsche im Jemen entführt (*ET*, 6.3.97, S. 8)
Text 2: Strom abgeschaltet - Frau stirbt in Klinik (*ET*, 6.3.97, S. 8).
Text 3: Hessen will Einbürgerung (*ET*, 6.3.97, S. 3)

	Text 4		Text 5		Text 6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	6	5.83	20	19.59	22	23.39
2	16	14.96	35	33.11	23	22.79
3	4	6.16	15	19.69	15	12.51
4	4	2.46	13	8.92	9	6.20
5	0	0.96|	2	3.28	0	2.88
6	0	0.37|	1	1.41	0	1.27|
7	1	0.26|			0	0.54|
8					0	0.22|
9					1	0.20|
k =	1.4081		62.9711		2.8678	
p =	0.6375		0.9725		0.6618	
α =	0.8298		0.4801		0.3286	
X^2 =	2.003		3.706		5.384	
FG =	1		2		2	
P =	0.16		0.16		0.07	

Text 4: Entschädigung in Ungarn (*ET*, 6.3.97, S. 3)
Text 5: Deserteure müssen weiter bangen (*ET*, 6.3.97, S. 3)
Text 6: Kämpfe erschüttern Albaniens Süden (*ET*, 6.3.97, S. 1)

	Text 7		Text 8		Text 9	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	30	27.60	17	16.71	10	9.70
2	24	25.16	29	26.11	22	20.00
3	13	15.84	6	10.10	5	9.11
4	11	8.64	4	3.81	6	4.05
5	4	4.28	2	1.41|	2	1.77
6	1	1.98	1	0.86|	0	0.77|
7	2	1.50			1	0.60|
k =	4.4233		1.3274		1.3104	
p =	0.7059		0.6510		0.5873	
α =	0.1251		0.7401		0.7686	
X^2 =	2.104		2.259		3.119	
FG =	3		1		2	
P =	0.55		0.13		0.21	

Text 7: Seehofer bleibt bei höherer Zuzahlung (*ET*, 7.3.97, S. 1)
Text 8: Doch kein Nutzer für Schürmann-Bau? (*ET*, 7.3.97, S. 1)
Text 9: Finanzprobleme der Städte wachsen (*ET*, 7.3.97, S. 1)

	Text 10		Text 11		Text 12	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	11	11.14	16	18.06	34	35.11
2	28	27.14	25	21.67	42	42.03
3	12	11.83	6	9.05	19	19.42
4	3	4.52	4	3.70	12	9.12
5	3	2.37	3	1.49|	4	4.32
6			1	1.03|	1	2.06
7					1	0.98|
8					1	0.96|
k =	3.8706		1.2498		0.8163	
p =	0.7772		0.6146		0.5078	
α =	0.7772		0.6387		0.6266	
X^2 =	0.730		2.698		1.529	
FG =	1		1		3	
P =	0.39		0.10		0.68	

Text 10: Workshop berät über “Expo am Meer” (*ET*, 7.3.97, S. 5)
Text 11: Bombenanschlag in Peking (*ET*, 8.3.97, S. 1)
Text 12: Absage der SPD/ Konsens mit den Kumpeln (*ET*, 8.3.97, S. 1)

	Text 13		Text 14		Text 15	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	41	43.29	20	20.36	30	36.11
2	36	36.04	33	30.25	31	31.32
3	26	21.73	18	18.73	18	14.62
4	12	10.71	8	9.60	9	5.83
5	3	4.56	3	4.30	0	2.09|
6	0	1.74|	2	1.74|	1	0.69|
7	0	0.61|	1	0.65|	2	0.34|
8	1	0.32|	0	0.23|		
9			1	0.14|		
k =	9.1963		7.6874		5.1476	
p =	0.8384		0.8082		0.8040	
α =	0.0105		0.4393		0.3056	
X^2 =	2.675		1.523		3.532	
FG =	2		2		1	
P =	0.26		0.47		0.06	

Text 13: Rebellen lehnen Amnestie ab (*ET*, 8.3.97, S. 1)
Text 14: Merkel stoppt die Stillegung von Biblis (*ET*, 8.3.97, S. 2)
Text 15: Kaschmir/ Indien und Pakistan wollen Annäherung (*ET*, 8.3.97, S. 2)

	Text 16		Text 17		Text 18	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	27	26.23	17	15.43	7	6.80
2	46	40.76	39	33.24	19	20.74
3	10	16.93	1	11.85	10	8.66
4	6	7.49	9	4.75	6	3.95
5	6	3.44	3	3.73	2	3.85
6	2	1.62				
7	1	1.53				
k =	0.3721		0.0022		0.1832	
p =	0.4747		0.4660		0.4262	
α =	0.7680		0.8758		0.8887	
X^2 =	5.992		15.032		2.295	
FG =	3		1		1	
P =	0.11		0.0001		0.13	

Text 16: Bundesministerien/ Personalzuwachs. Immer mehr Chefs an der Spitze (*ET*, 10.3.97, S. 1)

Text 17: Rechtschreibreform: Kritik auch in Wien (*ET*, 10.3.97, S. 1). Bei diesem Text ist die Anpassung der positiven Cohen-negativen Binomialverteilung nicht möglich. Die 1-verschobene Hyperpoisson-Verteilung, die nur 2 Parameter enthält, lässt sich bei 0 Freiheitsgraden mit $C = 0.0000$ erfolgreich anpassen, indem man die Wörter, die 3-5 Morphe enthalten, zu einer Klasse zusammenfasst.

Text 18: Luftangriffe im Libanon (*ET*, 10.3.97, S. 3)

	Text 19		Text 20		Text 21	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	10	11.13	28	28.00	35	34.13
2	13	10.76	36	31.92	25	26.41
3	4	5.87	6	11.65	14	12.79
4	3	3.23	5	4.16	5	6.47
5	4	1.78	1	1.46|	5	3.35
6	0	0.98|	0	0.51|	2	1.77
7	0	0.54|	1	0.17|	1	2.08
8	0	0.30|	1	0.13|		
9	1	0.41|				
k =	0.9237		1.2799		0.5481	
p =	0.4397		0.6662		0.4298	
α =	0.4423		0.6662		0.4298	
X^2 =	1.788		3.679		1.926	
FG =	2		1		3	
P =	0.41		0.06		0.59	

Text 19: Tibeter fordern Freiheit (*ET*, 10.3.97, S. 3)
Text 20: Auto in zwei Teile zerrissen (*ET*, 10.3.97, S. 5)
Text 21: Start des Euro ein Jahr verschieben (*ET*, 10.3.97, S. 3)

3. Ergebnis

In dieser Untersuchung wurde Wortlänge einmal anders bestimmt als sonst fast immer, nämlich durch die Zahl der Morphe pro Wort. In 20 von 21 Fällen kann die positive Cohen-negative Binomialverteilung erfolgreich angepasst werden. Nur bei Text 17 versagt das Modell. Das Problem ist in diesem Fall ein mathematisches: Da die positive Cohen-negative Binomialverteilung 3 Parameter aufweist, kann man der Unregelmäßigkeit bei den Wörtern mit 3 Morphen nicht durch eine Zusammenfassung der Wortlängen 3-5 begegnen. Stattdessen kann man mit dieser Methode aber die 1-verschobene Hyperpoisson-Verteilung, die nur 2 Parameter aufweist, erfolgreich auch an diesen Text anpassen.

Die 1-verschobene Hyperpoisson-Verteilung lässt sich übrigens an alle Pressetexte anpassen; sie wurde hier deshalb nicht vorgezogen, weil bei Verwendung dieser Verteilung in etlichen Fällen Zusammenfassungen nötig sind, die die positive Cohen-negative Binomialverteilung nicht erfordert. Die Erfahrungen mit Wortlängen, gemessen nach der Zahl ihrer Morphe, sind bisher so gering, dass man noch kein Urteil darüber fällen kann, welches Modell im Deutschen zu bevorzugen ist.

Als Ergebnis kann nun festgestellt werden, dass in allen Fällen eines der von Wimmer u.a. (1994) und Wimmer & Altmann (1996) theoretisch begründeten Modelle angepasst werden kann. Probleme, die sich bei der Anpassung der positiven Cohen-negativen Binomialverteilung bei Text 17 und bei der Verwendung der 1-verschobenen Hyperpoisson-Verteilung bei mehreren Texten zeigen, könnten auch darauf zurückzuführen sein, dass die behandelten Texte doch oft sehr kurz sind.

4. Das Ordsche Kriterium

Um herauszufinden, ob die bearbeiteten Texte hinsichtlich der Wortlängenverteilungen untereinander ähnlich sind, kann man eine empirische Abwandlung des Ordschen Kriteriums (Ord, 1972: 98f, 133ff) verwenden, das sich auf die "Momente" der benutzten Verteilungen stützt (Altmann, 1988: 48ff). Dabei handelt es sich um

$$m_1 = \frac{1}{N}\sum_x x f_x \quad \text{(Mittelwert), und}$$

$$m_r = \frac{1}{N}\sum_x (x - m_1)^r f_x, \quad r \geq 2 \quad \text{(zentrale Momente)}$$

wobei m_2 die Varianz und m_3 die Schiefe oder Asymmetrie der Verteilung darstellen. Hieraus lassen sich nun zwei Größen

$I = m_2 / m_1$ und $S = m_3 / m_2$

berechnen. Die Werte hierzu finden sich für die Pressetexte in Tabelle 1.

Tabelle 1
Das Ordsche Kriterium für Wortlängen in Pressetexten

Text	m_1	m_2	m_3	I	S
1	2.1429	0.9320	0.8834	0.4349	0.9479
2	2.2264	1.4204	1.7406	0.6380	1.2254
3	2.5128	2.2498	6.1474	0.8953	2.7324
4	2.3548	1.5193	3.3381	0.6452	2.1971
5	2.3605	1.2770	1.0954	0.5410	0.8578
6	2.2571	1.6767	4.5181	0.7429	2.6946
7	2.3547	2.0905	3.4642	0.8878	1.6571
8	2.1186	1.2232	1.9189	0.5774	1.5688
9	2.3913	1.6295	2.8533	0.6814	1.7510
10	2.2807	1.0089	0.9879	0.4424	0.9792
11	2.2000	1.4327	2.1687	0.6512	1.5137
12	2.3158	1.7775	3.4951	0.7676	1.9663
13	2.2017	1.4551	2.2879	0.6609	1.5723
14	2.5000	2.1105	5.3023	0.8442	2.5123
15	2.2198	1.6000	3.0443	0.7208	1.9027
16	2.2653	1.6847	3.1915	0.7437	1.8944
17	2.1594	1.1485	1.4322	0.5319	1.2470
18	2.4773	1.1131	0.6839	0.4493	0.6144
19	2.5429	2.8767	8.5551	1.1313	2.9739
20	2.0256	1.5891	4.8265	0.7845	3.0373
21	2.1954	1.9273	3.5409	0.8779	1.8372

Die Größen *I* und *S* kann man nun in ein Koordinatensystem <*I,S*> eintragen und zur Veranschaulichung der in diesem Fall recht geringen Homogenität der untersuchten Textgruppe verwenden, wie in der Graphik (s. Abb. 1) zu sehen ist.

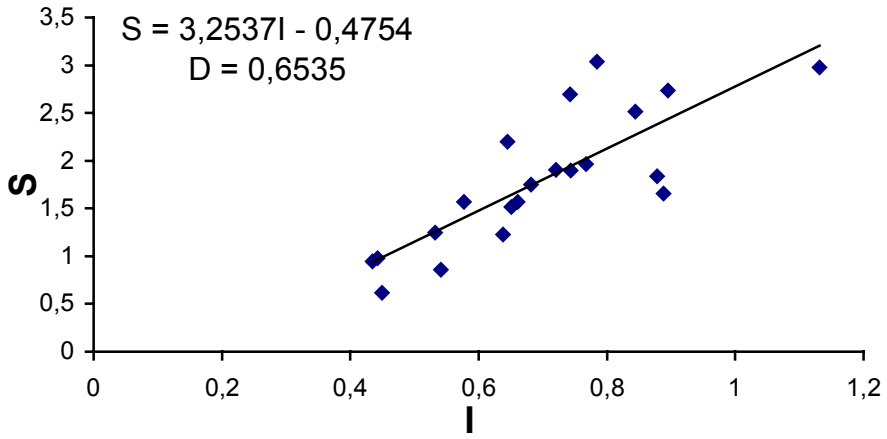


Abb. 1. Das Ordsche Kriterium für Wortlängen in Pressetexten

Für *I* und *S* lässt sich bei den Pressetexten eine lineare Regression $S = 3.2537I - 0.4754$ mit $D = 0.65$ bestimmen (s. Abb. 1). Die Texte weisen also eine hohe Streuung aus, obwohl sie zu einer einzigen Textsorte gehören. Dieses Ergebnis stimmt mit dem überein, das mit Fabeln von Pestalozzi erzielt wurde (Best 2001b).

5. Exkurs: Zur Relation Morphe – Silben im Deutschen

Gelegentlich taucht die Frage auf, ob man von der Zahl der Morphe auf die Zahl der Silben je Wort in einem Text schließen kann. Darauf kann im Folgenden eine vorläufige Antwort gegeben werden, vorläufig deshalb, weil insgesamt nur recht wenige Daten vorliegen.

Das Verhältnis zwischen der Zahl der Morphe und der Zahl der Silben stellt sich bei den 21 Pressemeldungen dieser Arbeit wie in Abb. 2 und Tabelle 2 dargestellt dar.

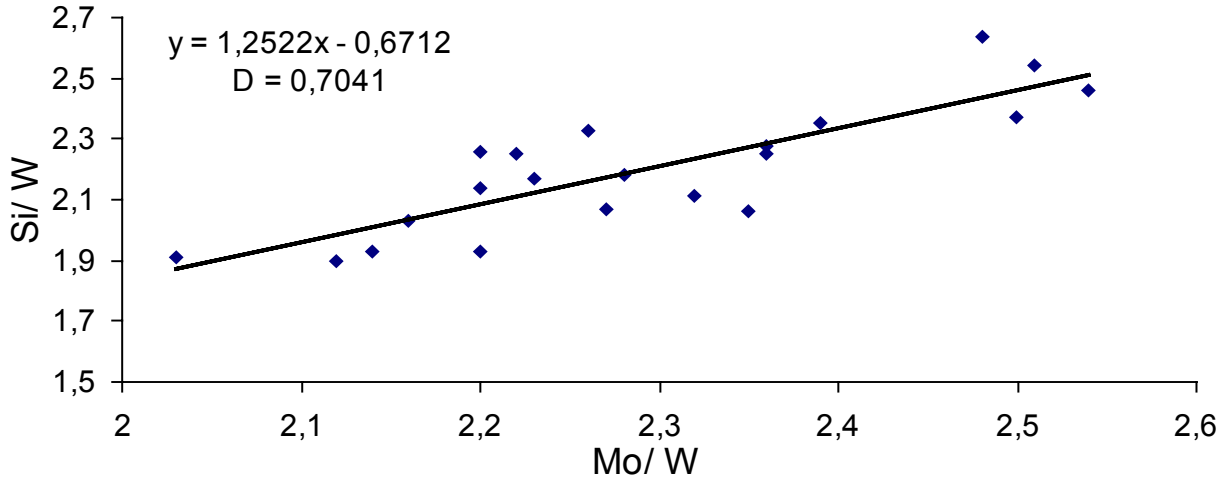


Abb.2. Morph-Silben-Relation in Pressetexten

Tabelle 2
Das Verhältnis Morphe – Silben in den Pressetexten

Text	Wörter	Morphe	Silben	Morphe/ Wort	Silben/ Wort
1	42	90	81	2.14	1.93
2	53	118	115	2.23	2.17
3	39	98	99	2.51	2.54
4	31	73	64	2.35	2.06
5	86	203	196	2.36	2.28
6	70	159	163	2.26	2.33
7	85	201	191	2.36	2.25
8	59	124	112	2.12	1.90
9	46	110	108	2.39	2.35
10	57	130	124	2.28	2.18
11	55	121	106	2.20	1.93
12	114	263	241	2.32	2.11
13	119	261	269	2.20	2.26
14	86	213	204	2.50	2.37
15	91	202	205	2.22	2.25
16	98	223	203	2.27	2.07
17	69	149	140	2.16	2.03
18	44	108	116	2.48	2.64
19	35	89	86	2.54	2.46
20	78	157	149	2.03	1.91
21	87	189	186	2.20	2.14
Σ	1444	3281	3158		

Die Graphik mit der Trendlinie und dem Determinationskoeffizienten $D = 0.70$ zeigt, dass es einen linearen Zusammenhang zwischen der Zahl der Morphe und der Zahl der Silben in diesen Texten gibt, der allerdings relativ großen Schwankungen unterliegt. Insgesamt gibt es knapp 4% mehr Morphe als Silben bei diesen Pressetexten.

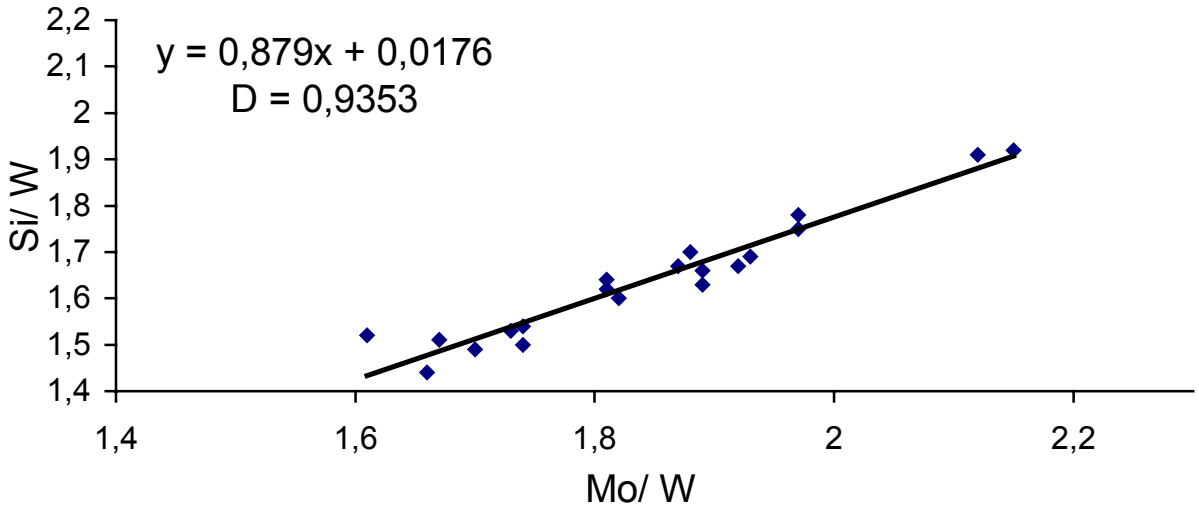


Abb.3. Morph-Silben-Relation in Fabeln von Pestalozzi

Tabelle 3
Das Verhältnis Morphe – Silben in Fabeln von Pestalozzi

Text	Wörter	Morphe	Silben	Morphe/ Wort	Silben/ Wort
1	156	261	236	1.67	1.51
2	142	269	231	1.89	1.63
3	319	576	517	1.81	1.62
4	137	248	225	1.81	1.64
5	230	494	441	2.15	1.92
6	107	195	171	1.82	1.60
7	137	265	232	1.93	1.69
8	105	169	160	1.61	1.52
9	111	193	171	1.74	1.54
10	96	180	163	1.88	1.70
11	143	282	254	1.97	1.78
12	118	232	207	1.97	1.75
13	110	190	168	1.73	1.53
14	255	476	425	1.87	1.67
15	136	257	226	1.89	1.66
16	133	232	200	1.74	1.50
17	177	294	255	1.66	1.44
18	175	336	292	1.92	1.67
19	139	294	266	2.12	1.91
20	234	397	348	1.70	1.49
Σ	3160	5840	5188		

In den Fabeln sind über 12% mehr Morphe als Silben enthalten. Hier ist die Abweichung der einzelnen Texte vom generellen Trend wesentlich geringer als bei den Pressetexten. Die folgende Graphik zeigt die beiden unterschiedlichen Trends (vgl. Abb. 4):

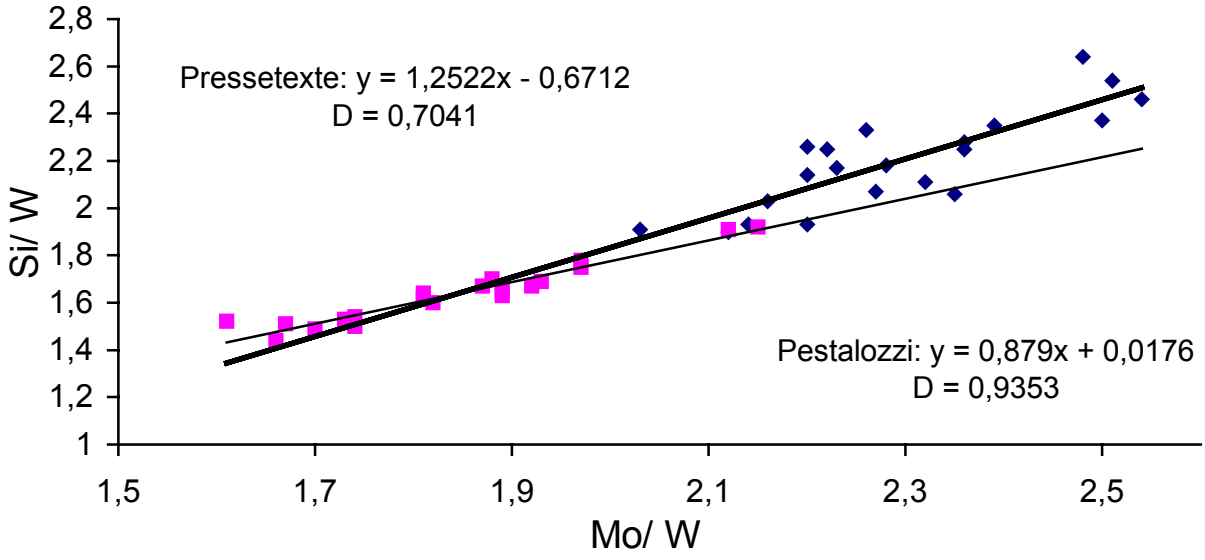


Abb. 4. Morph-Silben-Relation in zwei Textsorten (für beide Textsorten getrennt)

Die steiler verlaufende, fettere Linie steht für die Pressetexte. Die kleinen Quadrate stehen für die Fabeln, die Rauten für die Pressetexte. Man sieht, dass beide Textgruppen eine etwas verschiedene Steigung aufweisen. Es gibt kaum Überschneidungen bei der Morph-Silben-Relation. Wenn man den Unterschied der Textsorten vernachlässigt, erhält man einen klaren Trend für das Deutsche; die Bereiche der Fabeln und der Pressetexte sind auch in der gemeinsamen Graphik gut zu erkennen (vgl. Abb. 5):

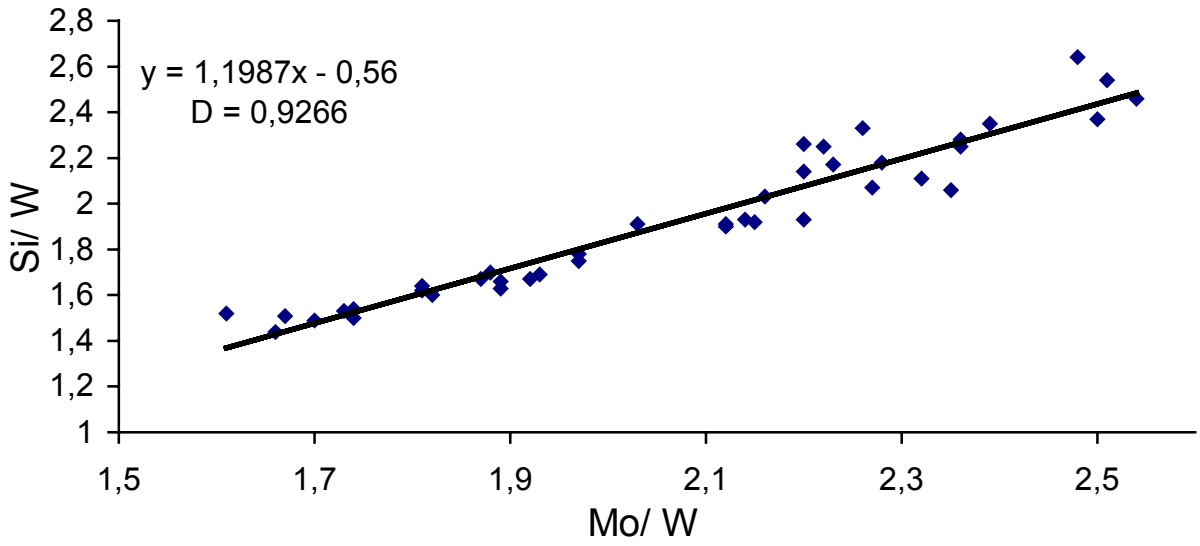


Abb. 5. Morph-Silben-Relation in zwei Textsorten (Gesamttrend)

Zusammenfassung

Die Untersuchung unterstützt die Hypothese, dass Wortlängen in Texten gesetzmäßig verteilt sind. Waren es in den meisten Fällen Texte, in denen die Wortlängen danach bemessen wurde, wie viele Silben die Wörter enthalten, so ist hier erst zum zweiten Mal die Zahl der Morphe pro Wort als maßgebend verwendet worden. Die Verwendung der positiven Cohen – negativen Binomialverteilung hat lediglich tentativen Charakter; auch die Anpassung der 1-verschobenen Hyperpoisson-Verteilung ist möglich. Eine Entscheidung darüber, welches Modell bei Wortlängen im Deutschen, nach der Zahl der Morphe bestimmt, am angemessensten ist, kann bei bisher nur 2 Datensätzen noch nicht gefällt werden.

Die Anwendung des Ordschen Kriteriums zeigt, dass die Pressetexte nicht als eine in sich homogene Textgruppe gewertet werden können. Dieses Ergebnis stimmt mit dem überein, das bei Fabeln von Pestalozzi erzielt wurde (Best 2001b: 266f.).

Der Versuch, eine Antwort auf die Frage zu geben, ob man bei deutschen Texten von der Zahl der Morphe auf die der Silben schließen kann, lässt sich aufgrund der nur wenigen Daten vorläufig wie folgt beantworten: Es sieht so aus, als ob es einen linearen Zusammenhang gäbe. Allerdings ist der Trend bei den Pressetexten deutlich weniger klar als bei den Fabeln. Nimmt man alle 41 Texte zusammen, so stellt sich ein enger linearer Zusammenhang heraus. Dieser Befund sollte aber noch untermauert werden.

Polikarpov (2006: 216) beklagt in einem ähnlichen Zusammenhang, dass in vielen Untersuchungen die semiotische Ebene der Organisation der Sprache zu wenig Beachtung finde. Die vorliegende Arbeit hat sich dieser Sichtweise angeschlossen. Es bleibt aber nach wie vor ein Desideratum, der Empfehlung Polikarpovs weiter nachzugehen.

Literatur

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

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

Best, Karl-Heinz (2001a). Zur Länge von Morphen in deutschen Texten. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 1-14*. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (2001b). Wie viele Morphe enthalten deutsche Wörter? Am Beispiel einiger Fabeln Pestalozzis. In: Ondrejovič, Slavomír & Považaj, Matej (eds.), *Lexicographica '99. Sborník na Počest Kláry Buzássyovej: 258-270.* Bratislava: Veda.

Best, Karl-Heinz (2005a). Morphlänge. In: Köhler, Reinhard, Altmann, Gabriel & Piotrowski, Rajmund G. (eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch: 255-260.* Berlin: de Gruyter.

Best, Karl-Heinz (2005b). Wortlänge. In: Köhler, Reinhard, Altmann, Gabriel & Piotrowski, Rajmund G. (eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch: 260-273.* Berlin: de Gruyter.

Kandler, Günther, & Winter, Stefan (1992ff.). *Wortanalytisches Wörterbuch. Deutscher Wortschatz nach Sinn-Elementen.* Bd. 1ff. München: Fink.

Ord, J. K. (1972). *Families of frequency distributions*. London: Griffin.

Polikarpov, Anatolij A. (2006). Towards the foundations of Menzerath's Law. On the functional dependence of affix length on their positional number within words. In: Grzybek, Peter (ed.), *Contributions to the Science of Text and Language: Word length studies and related issues:. 215-240*. Dordrecht: Springer.

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

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

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

Software

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

Glottometrics 13, 2006, 59-65

Wortlängen im Gotischen

Svitlana Kiyko[1]

Abstract. The aim of this paper is to show that word lengths in Gothic texts follow the 1-displaced hyper-Poisson distribution; in the lexicon they abide by the Conway-Maxwell-Poisson distribution.

Keywords: Gothic, word length

Vorbemerkung

Nachdem im Rahmen des Göttinger *Projekt Quantitative Linguistik* (Best 2001) Untersuchungen zu Wortlängenverteilungen in rund 50 Sprachen durchgeführt wurden, sollen hiermit auch Daten zum Gotischen, einer altgermanischen Sprache, bearbeitet werden. Das Gotische ist vor allem aus einem einzigen Text, der sog. Ulfila-Bibel, und zusätzlich der Skeireins, Erläuterungen zum Johannes-Evangelium, bekannt. Darüber hinaus gibt es einige wenige Texte und Inschriften (Braune/Ebbinghaus 1961: 2ff.). Wenn einmal neue Textbruchstücke auftauchen, kann es Probleme bereiten, sie zu lesen (Garbe 1972).

Wortlänge wird in dieser Arbeit wie auch in den meisten anderen des Göttinger Projekts nach der Zahl der Silben pro Wort bestimmt. Als „Wort" gilt das orthographische Wort, also die ununterbrochene Zeichenkette, die durch Leerstellen oder Interpunktionszeichen begrenzt ist; die Zahl der Silben pro Wort wird danach bestimmt, wie viele Vokale oder Diphthonge das Wort aufweist.

Theoretischer Hintergrund

Wie bei allen Arbeiten im Göttinger Projekt bilden auch hier die Arbeiten von Wimmer u.a. (1994) sowie Wimmer & Altmann (1996) den theoretischen Hintergrund. In diesen Untersuchungen wurde in Auseinandersetzung mit früheren Vorschlägen ein Gesetz der Wortlänenverteilung entwickelt, das als eine der wichtigsten Varianten auch die Hyperpoisson-Verteilung enthält, die sich gerade auch bei älteren Sprachen immer wieder als Modell der Wahl bewährt hat (Best 1998: 158). Dieses Modell soll daher auch an den gotischen Texten erprobt werden.

Es hat sich mehrfach gezeigt, dass bei Wortlängen eines Wörterbuchs aufgrund der andersartigen Datenstruktur mit anderen Verteilungsmodellen zu rechnen ist (Grotjahn & Altmann 1993). Wimmer u.a. (1994: 102) konnten an ein ungarisches Wörterbuch die Conway-Maxwell-Poisson-Verteilung erfolgreich anpassen. Auch bei einem deutschen alphabetischen Wörterbuch konnte diese Verteilung erfolgreich angewendet werden (Best 2003: 46). Dieses Modell wird daher auch für das gotische Wörterbuch getestet.

[1] Für Berechnungen und Unterstützung bei der Manuskripterstellung danke ich herzlich Herrn Dr. K.-H. Best.
Address correspondence to : jurkij@sacura.net.

Kriterien der Textauswahl

Die Textauswahl richtete sich nach folgenden Kriterien:

1. Der zu untersuchende Text sollte keine Lücken haben und keine problematischen Formen wie Abkürzungen oder Zahlwörter beinhalten. So wurden z.B. keine Kapitel aus dem ersten Brief des Paulus an die Korinther analysiert, weil diese alle lückenhaft sind.
2. Die Anzahl der Strophen sollte nicht allzu unterschiedlich sein, so dass Texte mehr oder weniger gleicher Länge untersucht wurden.
3. Es sollten alle 4 Evangelienverfasser und die Briefe des Paulus berücksichtigt werden.

Genaue Angaben zu den Texten finden sich im Anhang.

Das Modell der Wortlängenverteilungen im Gotischen

An die Dateien der Texte wurde die Hyperpoisson-Verteilung in 1-verschobener Form angepasst, da die Texte keine nullsilbigen Wörter enthalten. Die Formel der Verteilung lautet wie folgt:

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

Dabei sind a und b Parameter; ${}_1F_1(1;\ b;\ a)$ ist die konfluente hypergeometrische Funktion, d.h. ${}_1F_1(1;b;a)=1+\frac{a}{b}+\frac{a(a+1)}{b(b+1)}+\cdots$ und $b^{(x\text{-}1)} = b(b+1)(b+2)...(b+x\text{-}2)$.

Die Ergebnisse der Anpassung der 1-verschobenen Hyperpoisson-Verteilung an die Bibeltexte finden sich in den folgenden Tabellen. Die Anpassungen wurden mit einer geeigneten Software, dem Altmann-Fitter (1997) durchgeführt.

In den Tabellen sind folgende Angaben enthalten: x ist die Silbenzahl der Wörter, n_x die Zahl der Wörter mit der entsprechenden Silbenzahl in dem jeweiligen Text, NP_x die aufgrund der 1-verschobenen Hyperpoisson-Verteilung zu erwartende Anzahl der Wörter dieser Länge; a und b sind die Parameter der Hyperpoisson-Verteilung; X^2 ist das Chiquadrat, P die Überschreitungswahrscheinlichkeit für das berechnete Chiquadrat; FG gibt die Zahl der Freiheitsgrade an. In einigen Fällen mussten die höheren Wortlängenklassen so zusammengefasst werden, dass mangels Freiheitsgraden kein P bestimmt werden kann; in diesen Fällen gilt der Diskrepanzkoeffizient $C = X^2/N$ als Prüfgröße für die Güte der Anpassung. Eine Anpassung mit $P \geq 0.05$ bzw. $C \leq 0.01$ gilt als zufriedenstellend. Diese Bedingungen sind in allen Fällen erfüllt.

Die Ergebnisse der Anpassung der 1-verschobenen Hyperpoisson-Verteilung an die gotischen Texte sind in Tabelle 1 angegeben.

Tabelle 1
Anpassung der Hyperpoisson-Verteilung an gotische Texte

	R 13		M 25		M 8		R 10	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	88	87.71	54	53.63	224	221.83	107	106.88
2	93	92.69	50	46.94	184	186.22	123	116.29
3	42	43.72	21	27.14	91	93.09	48	59.64
4	16	13.27	15	11.72	37	33.13	26	20.01
5	2	3.61	4	4.03	9	9.15	5	6.18
6			1	1.54	1	2.58		
a =	0.8522		1.7043		1.2360		0.9702	
b =	0.8063		1.9472		1.4723		0.8916	
X^2 =	1.331		2.686		1.481		4.673	
FG =	2		3		3		2	
P =	0.51		0.44		0.69		0.10	

	2Kor 6		2Kor 9		Marcus 2		2Kor 2	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	116	110.40	88	89.46	219	218.10	103	100.41
2	70	82.45	96	88.96	166	179.52	99	96.51
3	50	44.42	48	52.30	105	84.59	40	47.51
4	24	18.72|	21	21.82	23	27.92	22	15.72|
5	5	9.01|	5	7.06	5	7.09	1	4.85|
6			4	2.40	1	1.78		
a =	1.9343		1.4383		1.1022		1.0091	
b =	2.5900		1.4464		1.3391		1.0499	
X^2 =	2.923		2.675		7.761		1.605	
FG =	1		3		3		1	
P =	0.09		0.44		0.05		0.21	

Die senkrechten Striche hier und in weiteren Tabellen zeigen an, dass die betreffenden Längenklassen zusammengefasst wurden.

	Gal 4		Titus 1		2Timoth 1		Lucas 15	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	115	115.17	96	100.00	76	73.79	219	213.84
2	112	107.81	58	60.41	69	78.44	192	205.62
3	43	49.64	52	35.57	57	48.33	103	89.30
4	18	15.15	28	20.43	24	20.96	24	25.05
5	4	4.23	7	24.59	5	9.48	2	6.19
a =	0.9062		23.2073		1.4653		0.7921	
b =	0.9680		38.4121		1.3784		0.8237	
X^2 =	1.595		0.769		5.310		5.991	
FG =	2		0		2		2	
P =	0.45		-		0.07		0.05	
C =	-		0.0032		-		-	

	Kollos 4		Ephes 3		Phil 3		Tim 1	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	106	105.56	97	97.31	119	118.97	117	114.81
2	98	97.59	121	115.31	138	137.97	93	102.15
3	47	48.29	45	54.32	60	60.73	64	55.54
4	22	21.56	21	15.97	18	16.50	25	21.74
5			2	3.41|	3	3.83	4	8.76
6			1	0.68|				
a =	1.0649		0.7820		0.7096		1.3981	
b =	1.1518		0.6599		0.6119		1.5714	
X^2 =	0.048		3.749		0.318		5.217	
FG =	1		2		2		2	
P =	0.83		0.15		0.85		0.07	

	1 Thes 3		Galat 6		2 Kor 13		Joh 1	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	79	77.20	93	90.94	82	84.36	183	182.96
2	82	80.13	113	110.50	112	101.35	204	203.96
3	42	47.53	41	48.01	32	42.89	48	63.80|
4	28	19.74|	18	15.55	12	11.02	22	11.61|
5	2	8.40|			4	2.38	7	1.67|
a =	1.3843		0.6765		0.6535		0.4349	
b =	1.3337		0.5567		0.5440		0.3901	
X^2 =	0.856		1.522		5.176		0.000	
FG =	1		1		2		0	
P =	0.36		0.22		0.08		-	
C =	-		-		-		0.0000	

Modellierung der Wortlängenverteilung im gotischen Wörterbuch

Die Wörterbuchdaten wurden anhand von Stamm (1872) erarbeitet. Als Modell wurde die 1-verschobene Conway-Maxwell-Poisson-Verteilung

$$(2) \qquad P_x = \frac{a^{x-1}}{[(x-1)!]^b T}, \quad x = 1, 2, 3, \ldots$$

mit $T = \sum_{j=0}^{\infty} \frac{a^j}{(j!)^b}$ angepasst. Das Ergebnis ist in Tabelle 2 zu sehen.

Tabelle 2
Anpassung der Conway-Maxwell-Poisson-Verteilung an Gotische Daten

x	n_x	NP_x
1	519	447.21
2	1572	1636.44
3	1398	1362.39
4	461	477.07
5	58	90.36
6	12	10.63
7	4	0.85
8	1	0.05
a =	3.6592	
b =	2.1359	
C =	0.0074	

Bei größeren Dateien wie denen von Wörterbüchern kommt als Testkriterium nur der Diskrepanzkoeffizient C infrage, der hier eine gute Anpassung des Modells anzeigt.

Die Daten des Wörterbuchs wurden auch für jeden Buchstaben des Alphabets gesondert aufgeführt. Für einige davon kann ein Modell gefunden werden, für andere nicht. Ursache ist in mehreren Fällen eine zu geringe Datenmenge. Wahrscheinlich ist der Umfang des ausgewerteten Wörterbuchs für eine solche Untersuchung der Wortlängenverteilung einzelner Buchstaben doch allzu gering.

Ergebnis

An alle 20 Texte kann die 1-verschobene Hyperpoisson-Verteilung angepasst werden. Sie erweist sich damit wiederum als ein geeignetes Modell für die Wortlängenverteilungen in einer weiteren germanischen Sprache.

Auch das gotische Wörterbuch weist eine Wortlängenverteilung auf, die mit einer für solche Zwecke bereits bewährten Verteilung modelliert werden kann.

Insgesamt darf damit festgestellt werden, dass Wortlängenverteilungen im Gotischen sich nicht anders verhalten als in vielen weiteren Sprachen. Die Idee, dass sowohl das Sprachsystem als auch seine Verwendung Gesetzen unterliegt, hat sich damit ein weiteres Mal bewährt.

Literatur

Best, Karl-Heinz (1998). Results and Perspectives of the Göttingen Project on Quantitative Linguistics. *Journal of Quantitative Linguistics 5, 155-162.*

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

Best, Karl-Heinz (2003). *Quantitative Linguistik. Eine Annäherung*. 2., überarb. u. erw. Aufl. Göttingen: Peust & Gutschmidt.

[**Braune/ Ebbinghaus**]: Braune, Wilhelm, *Gotische Grammatik mit Lesestücken und Wörterverzeichnis.* Fortgeführt von Karl Helm. 16. Auflage neu bearbeitet von Ernst A. Ebbinghaus. 2. Druck. Tübingen: Niemeyer 1961.

Garbe, Burckhard (1972). Die Verso-Seite des Speyerer Codex-Argenteus-Blatts. *Zeitschrift für deutsches Altertum und deutsche Literatur 101, 225-226.*

Grotjahn, Rüdiger, Altmann, Gabriel (1993). Modelling the Distribution of Word Length: Some Methodological Problems. In: Köhler, Reinhard, & Rieger, Burghard B. (eds.), *Contributions to quantitative linguistics:. 141-153.* Dordrecht u.a.: Kluwer.

Friedrich Ludwig Stamms *Ulfilas oder die uns erhaltenen Denkmäler der gotischen Sprache. Texte, Wörterbuch und Grammatik.* Ungekürzte Neuauflage der Ausgabe aus dem Jahre 1872. Stuttgart: Magnus o.J. (ca. 1980).

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

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

Software

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

Informationen zum *Göttinger Projekt:* Homepage: http://wwwuser.gwdg.de/~kbest/

Anhang

Es wurden folgende Texte untersucht:

R 13	Du Rumonim (Der Brief des Paulus an die Römer), Kapitel 13, Verse 1-14.
M 25	Aivaggeljo þairh Maþþaiu (Das Evangelium nach Matthäus), Kapitel 25, Verse 38-46.
M 8	Aivaggeljo þairh Maþþaiu (Das Evangelium nach Matthäus), Kapitel 8, Verse 1-34.
R 10	Du Rumonim (Der Brief des Paulus an die Römer), Kapitel 10, Verse 1-21.
2Kor 6	Du Kaurinþium anþara (Der zweite Brief des Paulus an die Korinther), Kapitel 6, Verse 1-18.
2Kor 9	Du Kaurinþium anþara (Der zweite Brief des Paulus an die Korinther), Kapitel 9, Verse 1-15.
Marcus 2	Aivaggeljo þairh Marcus (Das Evangelium nach Markus), Kapitel 2, Verse 1-28.
2Kor 2	Du Kaurinþium anþara (Der zweite Brief des Paulus an die Korinther), Kapitel 2, Verse 1-17.
Gal 4	Du Galatim (Der Brief des Paulus an die Galater), Kapitel 4, Verse 1-31.
Titus 1	Du Teitau (Der Brief des Paulus an Titus), Kapitel 1, Verse 1-16.
2Timoth 1	Du Teimauþaiau anþara (Der zweite Brief des Paulus an Timotheus), Kapitel 3, Verse 1-17.
Lucas 15	Aivaggeljo þairh Lucas (Das Evangelium nach Lukas), Kapitel 15, Verse 1-32.
Kollos 4	Du Kaulaussaium (Der Brief des Paulus an die Kolosser), Kapitel 4, Verse 1-19.
Ephes 3	Aipistaule Pavlaus du Aifaisium (Der Brief des Paulus an die Epheser), Kapitel 3, Verse 1-21.

Phil 3	Du Filippisium (Der Brief des Paulus an die Kolosser), Kapitel 3, Verse 1-21.
Tim 1	Du Teimauþaiau frumei (Der erste Brief des Paulus an Timotheus), Kapitel 1, Verse 1-20.
1Thes 3	Aipistaule Pavlaus du Þaissalauneikaium „a“ (Der erste Brief des Paulus an die Thessalonicher), Kapitel 3, Verse 1-13.
Galat 6	Du Galatim (Der Brief des Paulus an die Galater), Kapitel 6, Verse 1-18.
2Kor 13	Du Kaurinþium anþara (Der zweite Brief des Paulus an die Korinther), Kapitel 13, Verse 1-13.
Joh 1	Aivaggeljo þairh Johannen (Das Evangelium nach Johannes), Kapitel 15, Verse 1-27.

Glottometrics 13, 2006, 66-72

Deutsche Entlehnungen im Englischen

Karl-Heinz Best, Göttingen[1]

Abstract. The paper deals with two topics: The process of borrowing of German words in English and the spectrum of fields. Our aim is to show that the borrowings abide by the logistic law and the spectrum of fields follows a regular rank-frequency distribution.

Keywords: Borrowings, English, German

Zur Diskussion um die Entlehnungen

Die Diskussion zum Thema „Fremdwörter" dreht sich seit geraumer Zeit vorwiegend um den Einfluss des amerikanischen und britischen Englisch auf das Deutsche. Als weniger anstößig werden inzwischen die Einflüsse des Lateinischen und des Französischen betrachtet, obwohl sie immer noch in vielen Lexika einen wesentlich höheren Anteil haben als die Anglizismen (Best 2001, Körner 2004). Gelegentlich wird in diesem Zusammenhang allerdings auch die Frage gestellt, ob denn das Deutsche seinerseits andere Sprachen beeinflusst habe. Einige Daten dazu sind bekannt: So der Einfluss des Deutschen auf das Französische (Best 2007, Sarcher 2001), das Polnische (Hentschel 1995) und das Ungarische (Beöthy & Altmann 1982). Eine Fülle von Beispielen hat Stiberc (1999) zusammengetragen; die Aktualität des Themas belegt die populäre Sammlung, die der *Deutsche Sprachrat* unter dem Titel *Ausgewanderte Wörter* veranlasst und teilweise veröffentlicht hat (Limbach [Hrsg.] 2007).

Mit dem Einfluss fremder Sprachen im englischen Wortschatz hat sich bereits Herdan (1966: 335) befasst; er erhebt auf der Basis des *OED* (*Oxford English Dictionary* 1933) die Daten differenziert für Französisch und Latein, für die anderen Sprachen jedoch nur als „Mischung", so dass der Einfluss des Deutschen nicht gesondert erfasst wird. Einige Daten dazu sind dank der Bemühungen von Finkenstaedt & Wolff (1973: 136ff.) bekannt; erst Pfeffer (1987) sowie Pfeffer & Cannon (1994) erlauben es, die Entwicklung der Entlehnungen nachzuvollziehen. Sie bilden die Grundlage der vorliegenden Untersuchung, die insofern nichts Neues zu bieten hat. Die Frage, um die es nun aber gehen soll, ist zweifacher Natur: 1. kann man für die deutschen Entlehnungen ins Englische nachweisen, dass sie dem logistischen Gesetz folgen, wie das im Falle des Französischen und des Ungarischen bereits geschehen ist? Und 2.: lässt sich zeigen, dass das Themenspektrum der deutschen Entlehnungen ebenfalls gesetzmäßig geregelt ist, wie das für das Französische (Best 2007) und für die etymologischen Spektra des Deutschen, Englischen (Best 2005a) und Türkischen (Best 2005) belegt werden konnte? Um beide Fragen zu bearbeiten, wurden die von Pfeffer und Cannon vorgelegten Daten den entsprechenden statistischen Tests unterworfen.

Die Entwicklung der deutschen Entlehnungen im Englischen

Die theoretische Grundlage für diese Untersuchung hat Altmann (1983) gelegt, als er in Auseinandersetzung mit früheren Vorschlägen das logistische Gesetz (unter dem Namen

[1] Address correspondence to: kbest@gwdg.de

„Piotrowski-Gesetz“) entwickelte. Die Übertragung dieses Konzepts auf das Problem der Entlehnungen wurde in Beöthy & Altmann (1982) sowie Best & Altmann (1986) begründet. Daraus ergibt sich die Hypothese, dass auch die deutschen Entlehnungen im Englischen dem logistischen Gesetz folgen sollten. Um dies zu prüfen, folgt eine Übersicht zur Zunahme deutscher Wörter im Englischen (n. Pfeffer & Cannon 1994: 3; Cannon 1998), wobei das logistische Gesetz in der Form für den unvollständigen Sprachwandel

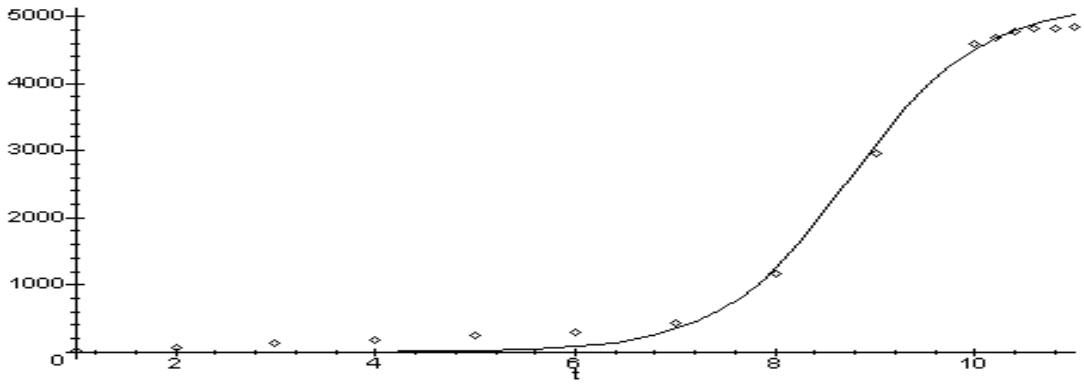
$$(1) \qquad p_t = \frac{c}{1 + ae^{-bt}}$$

(Altmann 1983: 60f.) mit Hilfe der Software *NLREG* (Graphik: *MAPLE*) angepasst wurde.

Legende zur Tabelle 1: *a*, *b* und *c* sind die Parameter des Modells; *c* stellt den Grenzwert dar, gegen den der Sprachwandel, also der Prozess der Übernahme deutscher Wörter ins Englische, strebt. *D* ist der Determinationskoeffizient, der im optimalen Fall $D = 1.00$ erreichen kann und mit $D \geq 0.90$ eine sehr gute Übereinstimmung zwischen Beobachtung und Modell anzeigt; dies ist hier gegeben.

Tabelle 1
Entlehnungen aus dem Deutschen ins Englische

t	Jahre	beobachtet	kumuliert	berechnet
1	-1500	12	12	0.0363
2	1501-1550	50	62	0.1676
3	1551-1600	65	127	0.7729
4	1601-1650	62	189	3.5630
5	1651-1700	64	253	16.3922
6	1701-1750	50	303	74.7388
7	1751-1800	137	440	327.4123
8	1801-1850	722	1162	1226.2397
9	1851-1900	1793	2955	3029.3034
10	1901-1950	1629	4584	4447.0311
10.2	1950er	97	4681	4604.1383
10.4	1960er	88	4769	4727.2039
10.6	1970er	44	4813	4822.1189
10.8	1980er	17	4830	4894.5052
11	1990er	7	4837	4949.2287
$a = 648430.60$ $b = 1.5287$ $c = 5108.95$ $D = 0.9967$				



Graphik 1. Entlehnungen aus dem Deutschen ins Englische

Anmerkung: Da die Periodisierung in den beiden zugrundeliegenden Werken geringfügig verschieden ist, sind die Daten für das Jahr 1950 doppelt aufgeführt; es lässt sich nicht ermitteln, um wie viele Wörter es sich dabei handelt.

Datenbasis der Erhebung sind das 20bändige *Oxford English Dictionary (OED)* (2. Auflage 1989) einschließlich Ergänzungsbände und viele weitere Wörterbücher, darunter vor allem Fachwörterbücher. Aufgenommen wurden „Lehnwörter", in einem weiten Sinne verstanden, darunter auch Lehnübersetzungen (z.B. „airship" von „Luftschiff") und Übernahmen von einzelnen Bedeutungen zu einem bereits vorhandenen englischen Wort (so die Bedeutung „Kurfürst" zum englischen Wort „elector"). Namen wurden nur dann berücksichtigt, wenn sie wie „Mindel" mehr als nur rein biographischer oder geographischer Natur sind. Es wurden nur direkte Übernahmen aus dem Deutschen aufgenommen, nicht auch solche, die wie „Vampir" über andere Sprachen (in diesem Fall über das Französische) ins Englische fanden. Viele der Übernahmen sind Fachwörter, die letztlich auf griechische oder lateinische Ursprünge zurückgehen, aber aus dem Deutschen ins Englische gelangten (Pfeffer 1987: 3ff.; Pfeffer & Cannon 1994: XXIIff.). Cannon (1998: 20) beziffert die deutschen Entlehnungen im Englischen auf 5457 Wörter; davon sind 696 undatiert (Pfeffer & Cannon 1994: 6).

Nach Cannon (1998: 19) ist Deutsch unter mindestens 84 Sprachen, von denen das Englische in neuer Zeit Fremdwörter übernommen hat, an 8. Stelle zu nennen; nach Best (2003: 94f.; Daten n. Wolff 1969) sogar nur an 13. Stelle (s. auch Scheler 1977: 72). Im etymologischen Spektrum des Englischen mit 85 Sprachen bzw. Sprachgruppen, das Finkenstaedt & Wolff (1973: 119f.) vorstellen, rangieren Deutsch an 12. und Niederdeutsch an 13. Stelle, wenn man einmal Germanisch, Alt- und Mittelenglisch als „Geber"-Sprachen unberücksichtigt lässt. Dennoch meint Stanforth (1996: 171): „Wir können Pfeffers eingangs zitierter Behauptung, daß der deutsche Einfluß auf die englische Sprache vielfach unterschätzt worden ist, besonders in den englischen Sprachgeschichten, nur beipflichten. Zwar hat Pyles in gewisser Weise recht, wenn man die absolute Zahl der Germanismen schätzt und sie als Anteil der englischen Lexik ausdrückt. Aber der Beitrag des Deutschen zum englischen Fachwortschatz ist, wie Pfeffer gezeigt hat, beachtlich, wenn auch die Zahl der ‚Alltagsgermanismen' nach wie vor gering bleibt."

„Mit einem Lehnwortanteil von über 70 % (*SOED/ CED*), rund 70% (*ALD*) und über 50% (*GSL*) steht das heutige Englisch innerhalb der germanischen − und wohl auch innerhalb des Gros der idg. Sprachen − einzigartig da. Es ist der Prototyp einer gemischten Sprache." (Scheler 1977: 74; Abkürzungen: verschiedene Wörterbücher.)

Ein Modell für das Themenspektrum der deutschen Entlehnungen

Interessant ist auch die Frage, aus welchen Kommunikationsbereichen das Englische seine deutschen Entlehnungen genommen hat. Darauf gibt das Themenspektrum (Pfeffer & Cannon 1994: 5) Antwort. Die Frage ist also, ob die Verteilung der deutschen Wörter auf verschiedene Themen gesetzmäßig geregelt ist. Für solche Spektra gibt es verschiedene Möglichkeiten der Modellierung; wie schon im Fall des Französischen (Best 2007) wird hier wiederum mit der Software *NLREG* (Graphik: *MAPLE*) Altmanns (1993: 62) Modell

$$(2) \qquad y_x = \frac{\binom{b+x}{x-1}}{\binom{a+x}{x-1}} y_1, \qquad x = 1, 2, 3, \ldots$$

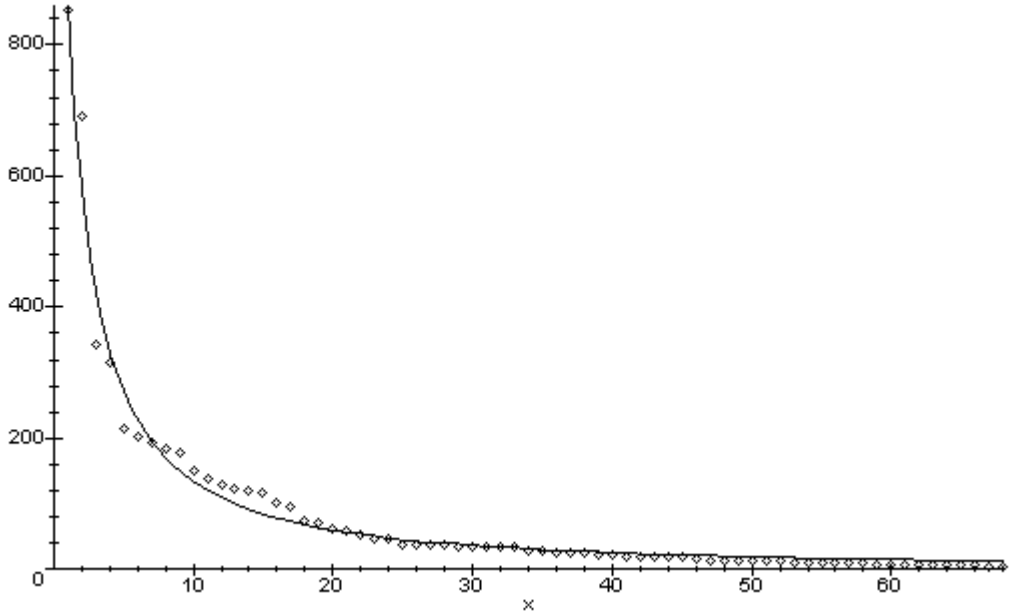
für beliebige Rangordnungen an die Daten angepasst (s. Tabelle 2).

Tabelle 2
Themenspektrum der deutschen Entlehnungen im Englischen

x	Herkunft	n_x beob.	n_x ber.	x	Herkunft	n_x beob.	n_x ber.
1	Mineralogy	854	854.00	35	Optics	27	30.28
2	Chemistry	690	577.68	36	Geography	26	29.24
3	Biology	343	427.79	37	Entomology	24	28.26
4	Geology	316	335.13	38	Meteorology	24	27.33
5	Botany	215	272.85	39	Industry	22	26.46
6	Politics	201	228.44	40	Mining	21	25.63
7	Music	193	195.36	41	Ichthyology	20	24.85
8	Medicine	183	169.89	42	Printing	20	24.11
9	Biochemistry	178	149.75	43	Metallurgy	19	23.41
10	Philosophy	150	133.46	44	Ethnology	18	22.75
11	Psychology	139	120.06	45	Ornithology	18	22.11
12	Military	130	108.87	46	Economics	16	21.51
13	Zoology	122	99.39	47	Theatre	13	20.94
14	Food	119	91.28	48	Trades	13	20.39
15	Physics	118	84.26	49	Administration	12	19.86
16	Linguistics	101	78.14	50	Dance	12	19.36
17	Beverages	96	72.77	51	Immunology	12	18.89
18	Literature	73	68.01	52	Pottery	12	18.43
19	Theology	70	63.78	53	Apparel	11	17.99
20	Mathematics	63	59.99	54	Travel	10	17.57
21	Sports	58	56.58	55	Textiles	10	17.17
22	Pathology	52	53.50	56	Archeology	9	16.78
23	Sociology	48	50.70	57	Astronomy	9	16.41
24	Physiology	46	48.15	58	Transportation	9	16.05
25	Art	39	45.82	59	Crystallography	8	15.70
26	Anthropology	37	43.68	60	Architecture	7	15.37
27	Currency	37	41.71	61	Commerce	7	15.05
28	Pharmacology	37	39.89	62	Law	7	14.75
29	Technology	36	38.21	63	Aeronautics	6	14.45
30	Games	35	36.64	64	Forestry	6	14.16
31	Anatomy	35	35.19	65	Furniture	6	13.88
32	Education	34	33.84	66	History	6	13.62
33	Mythology	33	32.57	67	Agriculture	5	13.36
34	Ecology	29	31.39	68	Paleontology	5	13.11
$a = 2.0493$ $b = 0.7391$ $D = 0.98$							

Anmerkung: n_x beob.: beobachtete Werte; n_x ber.**:** aufgrund der Anpassung des Modells berechnete Werte**;** a und b sind wieder die Parameter des Modells; der Determinationskoeffizient $D = 0.98$ zeigt an, dass die Beobachtungen sehr gut erfasst werden.

Die Termini für die Kategorien wurden übernommen; in Einzelfällen wurden die Beobachtungswerte aufgrund der eigenen Auszählungen der Listen von Pfeffer & Cannon (1994: 8ff.) leicht korrigiert. Die Sammelkategorie „Miscellany“ (n_x = 221) wurde nicht berücksichtigt.



Graphik 2. Themenspektrum der deutschen Entlehnungen im Englischen

Ergebnis und Perspektive

Als erstes kann festgestellt werden, dass das Deutsche auch im englischen Wortschatz erhebliche Spuren hinterlassen hat. Allerdings betrifft dies weniger den Alltagswortschatz, dafür umso mehr die speziellen Wortschätze der Fach- und Wissenschaftssprachen. Sehr viele der entlehnten Wörter gehören zur Gelehrtensprache; das Deutsche dient dem Englischen also mehr als Vermittlersprache, weniger als Herkunftssprache. Dass das Deutsche sich aber durchaus auch im Alltagswortschatz bemerkbar macht, zeigen Beispiele wie „Gemütlichkeit", „Poltergeist" und „Wunderkind", die von Limbach ([Hrsg.] 2007) neben vielen anderen aufgeführt werden.

Zweitens lässt sich aufgrund der durchgeführten Tests belegen, dass wie in anderen Fällen auch die Übernahme deutschen Lehnguts in Englische in seiner historischen Entwicklung dem logistischen Gesetz folgt. Für das Französische und Ungarische wurde der gleiche Nachweis bereits geführt; im Falle des Polnischen sind die entsprechenden Daten nicht veröffentlicht; die graphische Darstellung (Hentschel 1995: 76) lässt aber kaum einen Zweifel zu, dass auch dieser Entlehnungsprozess dem gleichen Gesetz unterliegt.

Drittens war die Frage nach dem Themenspektrum der Entlehnungen gestellt. Auch in diesem Fall gibt es keinen Grund, daran zu zweifeln, dass sich hier gesetzmäßige Verhältnisse eingestellt haben. Da bisher nur wenige Entlehnungsspektren auf ihre Gesetzmäßigkeit hin untersucht sind, kann man allerdings noch darüber nachdenken, ob für solche Fälle immer das gleiche Modell anzuwenden ist und ob nicht noch ein theoretisch besser begründetes zu entwickeln sein wird.

Literatur

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

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

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

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

Best, Karl-Heinz (2003). *Quantitative Linguistik: eine Annäherung*. 2., überarb. u. erw. Aufl. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (2005). Diversifikation der Fremd- und Lehnwörter im Türkischen. *Archív Orientální 73, 291-298.*

Best, Karl-Heinz (2005a). Ein Modell für das etymologische Spektrum des Wortschatzes. *Naukovyj Visnyk Černivec'koho Universytetu: Hermans'ka filolohija. Vypusk 266, 11-21.*

Best, Karl-Heinz (2007). Kann man Wissenstransfer messen? In: Wichter, Sigurd, & Stenschke, Oliver (Hrsg.), *Wissenstransfer – Wissenstransfer und Diskurs.* (In Arbeit)

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

Cannon, Garland (1998). Post-1949 German loans in written English. *Word 49, 19-54.*

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

Gebhardt, Karl (1975). Gallizismen im Englischen, Anglizismen im Französischen: ein statistischer Vergleich. *Zeitschrift für Romanische Philologie 91, 292-309.*

Hentschel, Gerd (1995). Zur ‚Seuche' des deutschen Lehnwortes im Polnischen und zu den ‚Selbstheilungskräften' dagegen. In: *Munus amicitiae. Studia Linguistica in honorem Witoldi Mańczak k septuagenari: 69-78*. Ed. A. Bochnakowa et S. Wiďak. Universitas Iagellonica, Ser. Varia CCCLVI, Cracowiae.

Herdan, Gustav (1966). Haeckels Biogenetisches Grundgesetz in der Sprachwissenschaft. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 19, 321-338.*

Kann, Hans-Joachim (2000). Neue Germanismen in *Time* 1999. *Der Sprachdienst 44, 173-176.* (Für 1998: *Der Sprachdienst 43/ 1999: 104-108*; etc.)

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

Limbach, Jutta [Hrsg.] (2007). *„Ausgewanderte Wörter"*. Ismaning: Hueber.

Pfeffer, J. Alan (1987). *Deutsches Sprachgut im Wortschatz der Amerikaner und Engländer. Vergleichendes Lexikon mit analytischer Einführung und historischem Überblick.* Tübingen: Niemeyer.

Pfeffer, J. Alan, & Cannon, Garland (1994). *German Loanwords in English. An Historical Dictionary.* Cambridge: Cambridge University Press.

Sarcher, Walburga (2001). *Das deutsche Lehngut im Französischen als Zeugnis für den Wissenstransfer im 20. Jahrhundert.* Hamburg: Kovać (= diss. phil., Augsburg 2000).

Scheler, Manfred (1977). *Der englische Wortschatz.* Berlin: Erich Schmidt Verlag.

Stanforth, Anthony W. (1996). *Deutsche Einflüsse auf den englischen Wortschatz in Geschichte und Gegenwart*. Tübingen: Niemeyer.

Stiberc, Andrea (1999). *Sauerkraut, Weltschmerz, Kindergarten und Co.* Freiburg/ Basel/ Wien: Herder.

Wolff, Dieter (1969). *Statistische Untersuchungen zum Wortschatz englischer Zeitungen.* Saarbrücken, diss. phil.

Software

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

Für weitere Informationen: http://wwwuser.gwdg.de/~kbest.

Glottometrics 13, 2006, 73-96

History of Quantitative Linguistics

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

XIX. August Schleicher (1821-1868)

Geb. 19.2.1821 in Meiningen, Studium der Theologie und semitischen Philologie in Leipzig und Tübingen, ab 1843 der Altphilologie in Bonn. Promotion und Habilitation in Bonn 1846, Privatdozent in Bonn. 1850 Prof. für klassische Philologie und Literatur in Prag; 1857-1868 Prof. der deutschen und vergleichenden Sprachwissenschaft und des Sanskrit in Jena, wo er am 6.12.1868 verstarb.

Es ist hier nicht der Ort, die Bedeutung Schleichers für die Linguistik zu entwickeln; dazu mögen u.a. Hinweise auf Arens (²1969: 248ff., passim), Gardt (1999: 279-282) und für seine Nähe zu den Naturwissenschaften Koerner (1981) genügen. Dass er auch für die Quantitative Linguistik nicht ohne Bedeutung ist, kann man den einschlägigen Nachschlagewerken oft nicht entnehmen.

Auf diesen Aspekt weisen nun aber Grzybek & Kelih (2003: 136; 2004: 95; 2005: 24) hin: Schleicher (1852: 20f.) stellte die Häufigkeiten der Buchstaben getrennt für Vokale und Konsonanten auf der Basis eigener Auszählungen von 2000 Einheiten zusammen. Er verstand seine Untersuchung altkirchenslavischer Laute/Grapheme als Ergänzung zu Förstemann (1852; vgl. dazu Best 2006). Es ging Schleicher darum, aufgrund von Berechnungen zu Vokal- und Konsonantensystemen präzise, zahlenmäßig ausgedrückte Aussagen dazu machen zu können, wie groß der Abstand des Slawischen zum Gotischen, Griechischen und Lateinischen ist. Ein Verfahren zur Berechnung der Differenz zwischen Sprachpaaren stellt er (1852: 24f.) vor. Dabei war er sich bewusst, dass man keine generellen Aussagen nur aufgrund der Befunde in der Phonetik der Sprachen treffen darf, meinte aber, die Ergebnisse passten gut zu den anderen Beobachtungen. Wie ein Plädoyer für Sprachstatistik liest sich: „es ist übrigens nicht zu leugnen, dass durch solche zälung sich vieles scharf herausstellt, was sonst leicht übersehen werden kann“ (1852: 17). Ähnlich heißt es später: „Die speciellen Lautlehren haben zu geben die Lautstatistik der Sprachen, d.h. die Aufzählung der die Sprachen bildenden Laute und ihrer Verbindungen, so wie auch der in den Sprachen als Wurzelformen der Beziehungselemente (wo sie vorhanden sind) vorkommenden Silben“ (Schleicher ³1874: 127). Budilovič übernahm 1883 - nicht ganz korrekt - Schleichers Daten (Grzybek & Kelih 2003: 136f.).

Kaeding (1897: 39) führt Schleicher (1888: 209f.) in seiner „Zusammenstellung der bisher bekannt gewordenen ‚Häufigkeitsuntersuchungen‘...“ an; es findet sich an dieser Stelle aber nichts, was seinen Zählungen in Schleicher (1852) nahe käme. Stattdessen kann man allenfalls Aussagen wie „Viel Einbuße hat h erlitten“ (Schleicher ³1874: 210) lesen.

Nachdem Schleicher sich in den genannten Zitaten derart positiv zu statistischen Erhebungen geäußert hat, ist es verwunderlich, wie wenig davon in anderen Werken erkennbar wird. Er befasst sich immer wieder mit der Klassifikation der Sprachen, meint auch, dass ihre Entwicklung ebenso wie ihr Verfall bestimmten Gesetzen unterliegen (1848: 25), und betrachtet Sprachen entschieden „als reale Naturwesen“ (1865: Vorwort; ähnlich 1863: 6); eine quantitative Perspektive wird dabei kaum erkennbar. Eine Vorstufe dazu kann man evt. in

seiner Idee von den allmählichen Übergängen erkennen, wie sie u.a. in folgendem Zitat zum Ausdruck kommt (Schleicher 1863: 19):

„Uebrigens ist...der Unterschied bezüglich des Beobachtungsmaterials zwischen der Sprachwelt und der Pflanzen- und Thierwelt nur ein quantitativer, nicht aber ein specifischer, denn bekanntlich ist ja die Abänderungsfähigkeit in gewissem Grade auch für Thiere und Pflanzen eine anerkannte Tatsache.

Aus dem bisher über die Differenzierung einer Grundform in mehrere zuerst wenig dann allmählich stärker von einander abweichende Formen Dargelegten folgt, dass wir auf sprachlichem Gebiete zwischen den Bezeichnungen für die verschiedenen Stufen der Unterschiede, d.h. zwischen Sprache, Dialekt, Mundart, Untermundart keine festen und sicheren Begriffsunterschiede aufzustellen vermögen. Die Verschiedenheiten, welche durch diese Worte bezeichnet werden, haben sich allmählich gebildet und gehen ineinander über...“

Zumindest im Bereich der Lautstatistik, verbunden mit Bemühungen um eine quantitativ ausgedrückte Bestimmung der Differenzen zwischen Sprachen, ist Schleicher in die Annalen der Quantitativen Linguistik aufzunehmen.

Literatur

Arens, Hans (²1969). *Sprachwissenschaft. Der Gang ihrer Entwicklung von der Antike bis zur Gegenwart. Zweite, durchgesehene und stark erweiterte Auflage.* Freiburg/ München: Karl Alber.

Best, Karl-Heinz (2006). Ernst Wilhelm Förstemann (1822-1906). *Glottometrics 12, 77-86.*

Dietze, Joachim (1966). *August Schleicher als Slawist. Sein Leben und sein Werk in der Sicht der Indogermanistik.* Berlin: Akademie.

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

Gardt, Andreas (1999). *Geschichte der Sprachwissenschaft in Deutschland. Vom Mittelalter bis ins 20. Jahrhundert.* Berlin/ New York: de Gruyter.

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

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

Grzybek, Peter, & Kelih, Emmerich (2005). Zur Vorgeschichte quantitativer Ansätze in der russischen Sprach- und Literaturwissenschaft. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch* (S. 23-64). Berlin/ New York: de Gruyter.

Kaeding, Friedrich Wilhelm [Hrsg.] (1897). *Häufigkeitswörterbuch der deutschen Sprache. Festgestellt durch einen Arbeitsausschuß der deutschen Stenographie-Systeme. Erster Teil: Wort- und Silbenzählungen. Zweiter Teil: Buchstabenzählungen.* Steglitz bei Berlin: Selbstverlag des Herausgebers. Teilabdruck: *Grundlagenstudien aus Kybernetik und Geisteswissenschaften. Bd. 4/ 1963.*

Killy, Walther, & Vierhaus, Rudolf [Hrsg.] (1998). *Deutsche Biographische Enzyklopädie (DBE). Band 8.* München: K.G. Saur.

Koerner, E.F. Konrad (1981). Schleichers Einfluß auf Haeckel: Schlaglichter auf die wechselseitige Abhängigkeit zwischen linguistischen und biologischen Theorien im 19. Jahrhundert. In: Scharf, Joachim-Hermann, & Kämmerer, Wilhelm (Hrsg.), *Leo-*

poldina-Symposion Naturwissenschaftliche Linguistik (S. 731-745). Halle: Deutsche Akademie der Naturforscher LEOPOLDINA.

Schleicher, August (1848). *Zur vergleichenden Sprachengeschichte.* Bonn: König.

Schleicher, August (1852). Die Formenlehre der kirchenslawischen Sprache, erklärend und vergleichend dargestellt. Bonn: König/ Wien: Gerold & Sohn/ Prag: Credner & Kleinbub.

Schleicher, August (1863). *Die Darwinsche Theorie und die Sprachwissenschaft.* Weimar: Böhlau.

Schleicher, August (1865). *Über die Bedeutung der Sprache für die Naturgeschichte des Menschen.* Weimar: Böhlau.

Schleicher, August (³1874). *Die deutsche Sprache*. Stuttgart: Cotta. (Die Auflagen [4]1879 und [5]1888 haben laut Vorwort, VIII „keinerlei Aenderungen erfahren.“)

Schmidt, Johannes (1890). August Schleicher. In: *Allgemeine deutsche Biographie. Einunddreißigster Band* (S. 402-416). Hrsg. Historische Commission bei der Königlichen Akademie der Wissenschaften. Leipzig: Duncker & Humblot.

Syllaba, Theodor (1995). *August Schleicher und Böhmen.* Prag: Karls-Universität.

Karl-Heinz Best

XX. Hans Arens (1911-2003)

Hans Arens, vollständig: Hans Karl Wilhelm Arens, geb. 31.1.1911 in Köln, 1913 nach Berlin verzogen, Abitur 1929 in Berlin, Studium der Germanistik, Anglistik und Romanistik in Berlin, 1936-1938 Hilfsassistent am Germanischen Seminar der Friedrich-Wilhelm-Universität Berlin, 14.2.1939 Promotion zum Dr. phil., 1938 - 1945 im Dienst des Forschungsamts des Reichsluftfahrtministeriums mit Dekodierungsaufgaben befasst, Staatsexamen 1947 in Marburg, Pädagogische Prüfung 1950 in Marburg, Assessor in Triberg, seit 13.4.1953 Gymnasiallehrer an der Alten Klosterschule in Bad Hersfeld, 1974 Studiendirektor, 31.1.1976 aus dem Dienst ausgeschieden. 1987 Bundesverdienstkreuz am Bande. 1975-1984 1. Vorsitzender des Arbeitskreises für Musik in Bad Hersfeld. Arens verstarb am 12.1.2003 in Bad Hersfeld. Viele literaturwissenschaftliche und sprachwissenschaftliche Publikationen, darunter umfangreiche Kommentare zu Goethe, Faust I und Faust II. Vermutlich sein bekanntestes Werk ist Arens (1955, ²1969), eine Geschichte der Sprachwissenschaft in Dokumenten, mit eigenen einleitenden Kommentaren. (Angaben teilweise n. Kürschner (1994: 18f.). Für weitere Informationen danke ich Beate M. Schwarz, Stadtarchiv von Bad Hersfeld, Dorothea Stahl, geb. Arens, Bad Hersfeld, und bes. Dr. Michael Fleck, Bad Hersfeld, der mir eine Reihe Kopien zugänglich gemacht hat, darunter einen selbstverfassten Lebenslauf und den Entwurf für seinen Vortrag an der Universität Trier 1982.)

Arens ist für die Quantitative Linguistik ein bedeutsamer Autor, weil er sich um eine statistische Fundierung literaturwissenschaftlicher Erkenntnisse bemühte: „Das statistische Verfahren [kann] in allen Bereichen der Sprache zu Einsichten verhelfen [...], die auf eine andere Weise nicht zu gewinnen wären“ (Arens ²1969: 630). Programmatisch heißt es ferner in Arens (1964: 7):

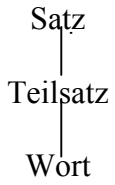
> „4. Die Wörter und Sätze sind als Stoff und Elemente der Dichtung auch meßbare Größen.
>
> 5. Meßbar oder zählbar sind z.B. Längen (der Wörter und Sätze), das Vokabular, die Anteile der Hauptwortarten in ihm, Bedeutungsfülle oder -armut (Begriffs- und Formwörter), die Kategorien der einzelnen Wortarten (z.B. abstrakt – konkret, transitiv – intransitiv), ihre Ursprungsbereiche, das Sprachniveau (Grad der Geläufigkeit oder Seltenheit der Wörter in ihrem vorliegenden Gebrauch), die

> Sinndichte des Textes (Verhältnis der Begriffs- und Formwörter), Satzarten (Anzahl der Haupt- und Nebensätze), Bauformen der Sätze und Perioden, die rhythmische Gliederung.
> 6. Die Gesamtheit der mit den Mitteln der Stilometrie gefundenen Charakteristika bestimmt *objektiv* seine Besonderheit im Vergleich mit anderen Texten, vermag jedoch nicht, seine dichterische Qualität zu erfassen."

Dies bildet den Hintergrund für eine vielseitige Untersuchung eines 347 Wörter langen Satzes aus Thomas Mann, *Joseph und seine Brüder. Der erste Roman: Die Geschichten Jaakobs*, bei der u.a. Satzbau und -länge, Wortlänge, Wortarten, Wortschatz, das Verhältnis heller zu dunklen Vokalen etc. untersucht und in eine Gesamtschau eingebracht werden, wobei in vielen Fällen Vergleiche mit anderen Sprachkunstwerken angestellt werden, um das Besondere dieses Textausschnitts zu erfassen. Arens führt damit einen Ansatz literaturwissenschaftlicher Arbeit vor, der die Texte so objektiv wie möglich erfasst und die gewonnenen Daten interpretiert, ohne deshalb auf weiterreichende Fragen zu verzichten.

Auch in Arens (1965: 6) wird „Objektivität als Ziel auch in der Beschreibung künstlerischer Formen" gefordert. Der Aspekt, um den es Arens in diesem Buch geht, ist der Zusammenhang zwischen Satz- und Wortlänge; er hat nämlich an 117 Texten von 52 Autoren deutscher literarischer Prosa des 17. bis 20. Jahrhunderts festgestellt, dass in Texten mit größerer durchschnittlicher Satzlänge auch die Wortlängen größer sind. Seine Erklärungsversuche qualifiziert er selbst als unvollständig (Arens 1965: 85); er nutzt die Befunde, um unterschiedliche Arten von Prosa stilistisch zu charakterisieren (Arens 1965: 91f.; vgl. dazu auch Arens ²1969: 629f.).

Es ist nicht notwendig, auf dieses Buch hier näher einzugehen; das hat in aller Ausführlichkeit bereits Altmann (1983; s. auch Altmann & Schwibbe 1989: 46-48; Referenzen: Aichele 2005: 17f.; Cramer 2005: 666, 676; Hug o.J.) getan. Arens' Befunde wurden von ihm als Konsequenz aus dem sog. Menzerath-Altmann-Gesetz dargestellt: Betrachtet man in der Sprache das Verhältnis zwischen Konstrukt und Konstituente, so gilt: *„Je größer das Ganze, um so kleiner die Teile"* (Menzerath 1954: 101). „Linguistischer" formuliert: *„Je größer ein sprachliches Konstrukt, desto kleiner seine Konstituenten"* (Altmann & Schwibbe 1989: 5). Das Menzerath-Altmann-Gesetz ist auf vielfache Weise erfolgreich geprüft worden (Asleh & Best 2005; Best 2003: 98-102, 128; Best 2006). Geht man von einer Hierarchie sprachlicher Größen in der Form



aus, dann bedeutet das: Je größer (länger) ein Satz, desto kleiner die Teilsätze; je kleiner die Teilsätze, desto größer die Wörter. Also: Je länger die Sätze, desto länger die Wörter. Genau das hat Arens festgestellt. Altmann hat damit den Beobachtungen von Arens ein neues Fundament gegeben; er hat das Gesetz anhand der Daten von Arens erfolgreich mit Hilfe des F-Tests getestet und „Arenssches Gesetz" genannt. (Anmerkung: Testet man die Daten von Arens mit dem Chiquadrat-Test, sind die Ergebnisse nicht so gut, ohne dass der Trend aber damit infrage gestellt würde. Hug (o.J.) stellt ebenso wie Altmann & Meyer (2005: 44) infrage, ob die Hierarchie der Einheiten in der angegebenen Weise gelten kann.)

Der Vollständigkeit halber muss erwähnt werden, dass bereits Fucks (1955) diesem Zusammenhang zwischen Wort- und Satzlänge ansatzweise auf der Spur war, das hier als Arenssches Gesetz firmiert. In dieser Arbeit wurden für 27 Texte von 23 Prosadichtern und

27 Schriftstellern (Politikern, Wissenschaftlern,...) die durchschnittlichen Satz- und Wortlängen erhoben; es zeigte sich, dass die beiden Stile sich deutlich unterscheiden. Dies gilt auch für den Stilindex, den Fucks als Produkt aus mittlerer Wortlänge und mittlerer Satzlänge bildet. Fucks (1955; 241) kommentiert seine Befunde:

> „Generell könnte man versucht sein, das Gesamtresultat...dadurch zu erklären, daß die Schriftsteller in ihren F a c h s p r a c h e n schreiben, deren W ö r t e r meist überdurchschnittlich l a n g sind, und daß Hand in Hand damit eine T e n d e n z z u l ä n g e r e n S ä t z e n g e h e. Diese Erklärung ist, wie Beispiele zeigen, teils zutreffend, teils unzutreffend. Auf eine genauere Erörterung sei hier jedoch verzichtet."

Fucks beschränkt seine Idee auf den Gegensatz von zwei Stilen, während Arens einen generellen Trend jenseits aller Stilunterschiede sieht, womit er recht hat. Allerdings wird der Zusammenhang bei den reichhaltigeren Daten von Arens auch deutlicher als bei denen von Fucks; testet man Fucks' Daten, so fallen die Ergebnisse wesentlich schlechter aus, gleich, ob man Dichter und Schriftsteller je für sich oder kombiniert bearbeitet.

Man findet in den angegebenen Werken von Arens noch weitere statistische Daten, z.B. Wortlängenverteilungen bei verschiedenen Dichtern (Arens 1964: 16; 1965: 70f.), Satzlängenverteilungen (Arens 1965: 69f.), Attributstrukturen (Arens 1964: 56), Angaben zur Textlänge (Arens 1964: 18f.; 1981: 11), und manches andere mehr. In den Notizen für seinen Trierer Vortrag am 25.11.1982 finden sich außer Wort- und Satzlänge: Gliederung des Wortmaterials in 5 Schichten, das Verhältnis von types und tokens in Texten, das Verhältnis von Begriffs- und Formwörtern, Relation von Verben zu Substantiven und Adjektiven und umgekehrt, verschiedene Subklassen dieser Wortarten, Verhältnis von Haupt- und Nebensätzen und das Problem der Autorenidentifikation. Einige Informationen zur Sprachstatistik finden sich in Arens (²1969: 628-630).

Auch weniger bedeutsame Fragen behandelt er mit statistischen Mitteln; so den Wortschatz der 17. Auflage von *Duden. Rechtschreibung der deutschen Sprache und der Fremdwörter* von 1973, der laut Buchumschlag „über 160000 Stichwörter und Beispiele" enthalten soll. Arens dagegen: „Ich habe mir die Mühe der Berechnung gemacht und 110000 ermittelt" (Arens 1979: 70); auf der nächsten Seite vergleicht er den Umfang dieses Wörterbuchs mit einem anderen.

Arens' Haltung, sein Streben nach möglichst objektiver Erkenntnis bei der Analyse sprachlicher Kunstwerke – aber auch ganz alltäglicher Dinge – kann man als eine der wenigen Manifestationen dessen sehen, was Fucks (1968: 77, 88) als „Quantitative Literaturwissenschaft" bezeichnete. Perspektiven, wie man seine Ideen fortführen kann, lassen sich u.a. Altmann (1988) und Altmann & Altmann (2005) entnehmen.

Das Literaturverzeichnis nennt auch Arbeiten, die für die Quantitative Linguistik nicht direkt bedeutsam sind, um Hans Arens in seinen vielfältigen wissenschaftlichen Interessen erkennbar werden zu lassen. Eine Bibliographie seiner Werke scheint es nicht zu geben; die folgende Liste mag einen Ersatz dafür bieten, obwohl sie nicht ganz vollständig ist.

Literatur

Aichele, Dieter (2005). Quantitative Linguistik in Deutschland und Österrreich. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch: 16-23.* Berlin/ N.Y.: de Gruyter.

Altmann, Gabriel (1983). H. Arens' „Verborgene Ordnung" und das Menzerathsche Gesetz. In: Faust, Manfred, Harweg, Roland, Lehfeldt, Werner, & Wienold, Götz (Hrsg.);

Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik. Festschrift für Peter Hartmann: 31-39. Tübingen: Narr.

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

Altmann, Gabriel, & Schwibbe, Michael H. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim/ Zürich/ New York: Olms.

Altmann, Vivien, & Altmann, Gabriel (2005). *Erlkönig und Mathematik.* http://ubt.opus.hbz-nrw.de/volltexte/2005/325.

Altmann, Gabriel, & Meyer, Peter (2005). Physicist's look at language. In: Altmann, Gabriel, Levickij, Viktor, & Perebyinis, Valentina (Hrsg.), *Problemy kvantytatyvnoï lingvistyky / Problems of Quantitative Linguistics: zbirnyk naukovych prac': 42-59.* Černivci: Ruta.

Arens, Hans (1939). *Ulrich von Lichtenstein ‚Frauendienst': Untersuchungen über den höfischen Sprachstil.* Leipzig: Akademie. (Diss.; Repr.: New York u.a.: Johnson 1970.)

Arens, Hans (1955, ²1969). *Sprachwissenschaft. Der Gang ihrer Entwicklung von der Antike bis zur Gegenwart. Zweite, durchgesehene und stark erweiterte Auflage.* Freiburg/ München: Karl Alber. (Taschenbuchausgabe: Frankfurt: Fischer Athenäum Taschenbücher 1974.)

***Arens, Hans** (1961). Gedanken zur Rechtschreibung und ihrer Reform. In: *Duden. Gedenkschrift zu seinem 50. Todestag: 91-99*. Hrsg. von der Alten Klosterschule Bad Hersfeld. Bad Hersfeld: Hoehlsche Buchdruckerei. (Diese Gedenkschrift ist aktualisiert in der unter Arens 1979 genannten Gedenkschrift erschienen.)

Arens, Hans (1964). *Analyse eines Satzes von Thomas Mann.* Düsseldorf: Schwann.

Arens, Hans (1965). *Verborgene Ordnung.* Düsseldorf: Schwann.

Arens, Hans (1977). Zur neueren Geschichtsschreibung der Linguistik. *Historiographia Linguistica 4, 319-382.* (Forschungsbericht)

Arens, Hans (1977). Rez. zu den ersten beiden Bänden der *Historiographia Linguistica. Zeitschrift für germanistische Linguistik 5, 353-364.*

Arens, Hans (o.J.; wohl 1979). Duden heute. In: *Duden. Gedenkschrift zu seinem 150. Geburtstag am 3. Januar 1979: 64-79.* Hrsg. v. Stadt Bad Hersfeld. Bad Hersfeld: Hoehl- Druck. (Diese Gedenkschrift aktualisiert (S. 82) die unter Arens 1961 genannte Festschrift.)

Arens, Hans (1980). Geschichte der Linguistik. In: *Lexikon der germanistischen Linguistik. Band 1. 2., vollständig neu bearbeitete und erweiterte Auflage: 97-107*, hrsg. von Hans Peter Althaus, Helmut Henne & Herbert Ernst Wiegand. Tübingen: Niemeyer. ([1]1973)

Arens, Hans (1980). „Verbum cordis": Zur Sprachphilosophie des Mittelalters. *Historiographia Linguistica VII, 13-27.*

Arens, Hans (1981). Gastvorträge (9. und 10.2.1981) an der Universität Antwerpen: *Verborgene Ordnung* und *Sprache und Schrift*.

Arens, Hans (1982). *Kommentar zur Goethes Faust I.* Heidelberg: Winter.

Arens, Hans (1982). Gastvortrag (25.11.1982) an der Universität Trier: *Quantitative Linguistik und Stilanalyse.*

Arens, Hans (1984). *Aristotle's Theory of Language and ist Tradition: Texts from 500 to 1750. Selection, Translation and Commentary.* Amsterdam: Benjamins.

Arens, Hans (1987). Gedanken zur Historiographie der Linguistik. In: Peter Schmitter (Hrsg.), *Zur Theorie und Methode der Geschichtsschreibung der Linguistik. Analysen und Reflexionen: 3-19.* Tübingen: Narr. (= Geschichte der Sprachtheorie I)

Arens, Hans (1989). *Kommentar zur Goethes Faust II.* Heidelberg: Winter.

***Arens, Hans** (1990). ‚De Magistro'. Analyse eines Dialogs von Augustinus. In: *De ortu grammaticae: Studies in medieval grammar and linguistic theory in memory of Jan Pinborg: 17-33.* Ed. by Geoffrey L. Bursill-Hall et alii. Amsterdam: Benjamins.

Arens, Hans (1994). *E. Marlitt: eine kritische Würdigung.* Trier: Wissenschaftlicher Verlag Trier.

Arens, Hans (2000). Sprache und Denken bei Aristoteles. In: *History of Language Sciences/ Geschichte der Sprachwissenschaften. 1. Band, 1. Teilband: 367-375.* Hrsg. v. Sylvain Auroux, E.F.K. Korner, Hans-Josef Niederehe & Kees Versteegh. Berlin/ New York: de Gruyter.

Asleh, Laila, & Best, Karl-Heinz (2004/05). Zur Überprüfung des Menzerath-Altmann-Gesetzes am Beispiel deutscher (und italienischer) Wörter. *Göttinger Beiträge zur Sprachwissenschaft 10/11, 9-19.*

Best, Karl-Heinz (2003). *Quantitative Linguistik. Eine Annäherung*. 2., überarb. u. erw. Aufl. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (2006). Sind Wort- und Satzlänge brauchbare Kriterien der Lesbarkeit von Texten? In: Wichter, Sigurd, & Busch, Albert (Hrsg.), *Wissenstransfer - Erfolgskontrolle und Rückmeldungen aus der Praxis: 21-31.* Frankfurt/ M. u.a.: Lang.

Cramer, Irene M. (2005). Das Menzerathsche Gesetz. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch:. 659-688.* Berlin/ N.Y.: de Gruyter.

Duden. *Rechtschreibung der deutschen Sprache und der Fremdwörter. 17., neu bearbeitete und erweiterte Auflage.* Bibliographisches Institut/ Mannheim/ Wien/ Zürich: Dudenverlag 1973.

Frühe deutsche Lyrik. Ausgewählt und erläutert von Hans Arens. Mit einer Einleitung von Arthur Hübner. Berlin: Weidmann 1935.

Fucks, Wilhelm (1955). Unterschied des Prosastils von Dichtern und anderen Schriftstellern. Ein Beispiel mathematischer Stilanalyse. *Sprachforum 1, 234-244.*

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

Handbuch der Linguistik: allgemeine und angewandte Sprachwissenschaft. Aus Beiträgen von Hans Arens [u.a.] unter Mitarbeit von Hildegard Janssen zusammengestellt von Harro Stammerjohann. Darmstadt: Wissenschaftliche Buchgesellschaft. (Die Beiträge sind nicht namentlich gekennzeichent.)

Hug, Marc (oJ., 2003 oder später). La loi de Menzerath appliquée à un ensemble de textes. cavi. univ-paris3.fr/lexicometrica/article/numero5/lexicometrica-hug.pdf.

Kürschner, Wilfried (Hrsg.) (1994). *Linguisten-Handbuch. Biographische und bibliographische Daten deutschsprachiger Sprachwissenschaftlerinnen und Sprachwissenschaftler der Gegenwart. Band 1: A-L.* Tübingen: Narr.

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

* Angaben übernommen aus Kürschner 1994.

Karl-Heinz Best

XXI. Adolf Lucas Bacmeister (1827-1873)

Geb. 9.7.1827 in Esslingen, gest. 25.2.1873 Stuttgart. Ab 1841 Besuch des evangelisch-theologischen Seminars in Blaubeuren, ab 1845 Studium der Theologie, später der Philologie in Tübingen. B. war 1848 in die Revolution verwickelt und wurde verhaftet und zu mehrmonatiger Haft verurteilt. 1853 philologisches Examen, danach Lehrer. 1857-64 Gymnasiallehrer in Reutlingen, 1864 Schriftleiter der *Augsburger Allgemeinen Zeitung*, 1865 Promotion zum Dr.phil. in Tübingen, 1871 Schriftleiter der Zeitschrift *Ausland*. Ab 1872 freier Feuilletonist. (Nach DBE, NDB und Schmid 1886.) Arbeitsbereiche: eigene literarische Werke; Bearbeitungen und Übersetzungen älterer deutscher und lateinischer Literatur; Untersuchungen

zu Orts- und Personennamen sowie zu weiteren linguistischen Themen. In Bacmeister (1870) erschien eine Auswahl seiner Zeitungsbeiträge. Eine weitere Auswahl journalistischer und literarischer Arbeiten findet sich in Bacmeister (1886), von den Herausgebern als „populärwissenschaftlich“ (S. III) charakterisiert. Obwohl Bacmeister mehrere sprachstatistische Untersuchungen durchgeführt hat, ist er in der Quantitativen Linguistik bisher so gut wie unbekannt. Einen Hinweis gibt es aber:

Auf Bacmeisters Häufigkeitsbestimmung von Buchstaben (1870b) macht Menzerath (1954: 1, Fußn. 1) aufmerksam und nennt sie „eine der ältesten und interessantesten Zählungen.“ In diesem Beitrag geht Bacmeister auf das Problem der Kosten für die Übermittlung von Nachrichten ein (1870b: 74): „Welche Art der Beförderung ist theurer, die welche sich für je 100 Buchstaben 20 Pf. St. zahlen läßt? oder die, welche nach continentaler Rechnung für 20 Worte als einfache Depesche den gleichen (denn so müssen wir natürlich vorläufig annehmen) Betrag fordert?“ Um das Problem zu lösen, zählt er zunächst die Wörter und Buchstaben eines bestimmten Textes (einer Thronrede) aus, der in Deutsch als Original, in Englisch, Französisch, Italienisch, Schwedisch und Spanisch als Übersetzung vorlag, und kann so am Beispiel eines Textes das Verhältnis von Wörtern und Buchstaben in diesen Sprachen numerisch vergleichen, außerdem auch noch dadurch, dass er einen Index Buchstaben/ Wort bildet. Damit liegt lange vor Greenberg (1960) bereits ein Wortlängenindex vor. Bacmeisters (1870b: 74f.) Ergebnisse:

Tabelle 1
Wörter und Buchstaben eines Textes (Thronrede) im Vergleich

Sprache	Wörter	Buchstaben	Buchstaben/ Wort
Deutsch (Original)	601	etwa 3300	5.5
Englisch	744	3572	4.8
Französisch	719	3456	4.8
Italienisch	563	etwa 3600	6.4*
Schwedisch	509	etwa 3060	6
Spanisch	713	3440	4.8

(*Hier gibt Bacmeister falsch berechnet 5.5 an.)

Es folgt im gleichen Aufsatz die von Menzerath erwähnte Zusammenstellung, welcher deutsche Buchstabe unter insgesamt 1000 wie häufig vertreten ist, wobei Bacmeister (1870b: 77) auch die Umlaute berücksichtigt. <ch> wertet er als eigenen Buchstaben, <ß> fehlt allerdings. Den Abschluss des Artikels bilden Überlegungen dazu, wie man die Kosten der Nachrichtenübertragung ggfs. verringern kann. Seine Auswertung, für die Bacmeister einen kleinen Zählungsfehler einräumt, findet sich in Tabelle 2, wobei zusätzlich wie schon in Best (2006) die 1-verschobene negative hypergeometrische Verteilung

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

angepasst wurde. Erläuterungen zu Tabelle 2:

n_x: beobachtete Anzahl der Buchstaben;
NP_x: Anzahl der Buchstaben nach der negativen hypergeometrischen Verteilung;
K, M, n: Parameter der Verteilung;
X^2: Werte des Chiquadrates;

FG: Freiheitsgrade;
P: Überschreitungswahrscheinlichkeit für das entsprechende Chiquadrat.

Die Anpassung der negativen hypergeometrischen Verteilung an die Textdateien wird als erfolgreich betrachtet, wenn $P \geq 0.05$ bzw. $C \leq 0.01$.

Tabelle 2
Anpassung der negativen hypergeometrischen Verteilung
an die 1000-Buchstaben-Zählung von Bacmeister

Rang	<...>	n_x	NP_x	Rang	<...>	n_x	NP_x
1	e	192	187.38	16	w	20	19.39
2	n	119	112.98	17	b	17	17.32
3	i	73	87.96	18	z	16	15.38
4	s	65	73.53	19	k	14	13.57
5	t	60	63.52	20	f	13	11.89
6	a	60	55.88	21	p	9	10.31
7	d	54	49.71	22	ü	8	8.83
8	r	54	44.53	23	v	8	7.46
9	u	40	40.08	24	ä	5	6.18
10	g	35	36.16	25	c	4	4.99
11	l	34	32.67	26	j	4	3.89
12	ch	31	29.53	27	ö	4	2.88
13	o	24	26.67	28	q	2	1.98
14	m	24	24.04	29	x	0	1.18
15	h	23	21.63	30	y	0	0.50
$K = 2.9476$, $M = 0.6303$, $n = 29$, $X^2 = 10.3533$, $FG = 25$, $P = 0.996$							

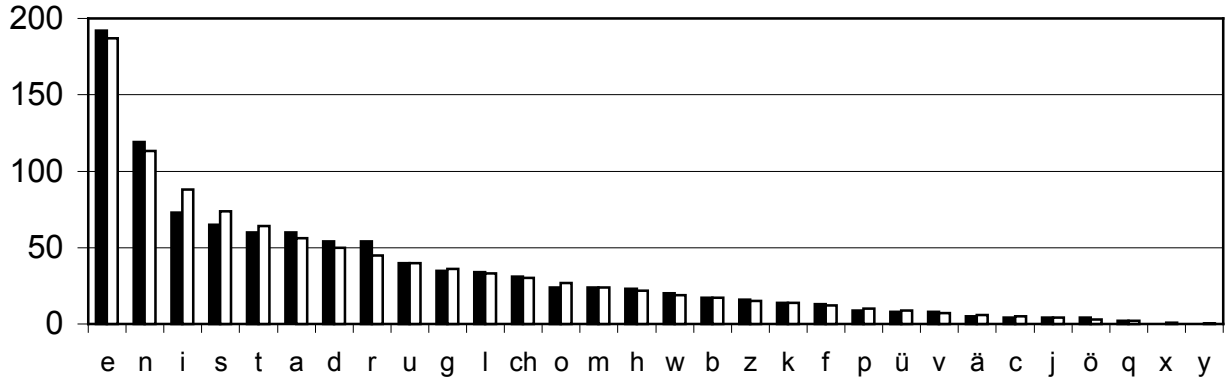


Abbildung zu Tabelle 2.

Man sieht, Bacmeisters Daten sind so gut, dass man das Modell, das sich auch sonst bei Buchstaben bewährt (Best 2006), mit sehr gutem Testergebnis anpassen kann. Testkriterium ist $P = 0.996$, da die Datei mit insgesamt 1000 Buchstaben nicht allzu groß ist.

Zwei weitere Themen werden statistisch behandelt:

In seinem Beitrag (1870a: 5) gibt er zunächst eine statistische Übersicht der 25 häufigsten englischen Nachnamen und bedauert, dass es eine solche umfassende Übersicht deutscher Familiennamen nicht gibt: „Hätten wir eine ähnliche Statistik aus Deutschland, sie würde ein eigenthümliches Gegenstück der obigen [gemeint: die englische, Verf.] bieten.“ Im weiteren Verlauf des Artikels nennt er dann ersatzweise eigene Zählungen zu den Nachnamen

von Reutlingen, das 1859 etwa 13000 Einwohner hatte, und von Eningen mit ca. 4000 Einwohnern.

Auch an die von Bacmeister (1870a: 7-8) erhobenen Namen der beiden Städte kann man die 1-verschobene negative hypergeometrische Verteilung anpassen:

Tabelle 3
Anpassung der negativen hypergeometrischen Verteilung
an die häufigsten Familiennamen von Reutlingen im 19. Jahrhundert

Rang	Name	n_x	NP_x	Rang	Name	n_x	NP_x
1	Votteler	78	96.54	46	Buck	17	18.03
2	Hohloch	67	65.66	47	Gruoner	17	17.80
3	Klein	52	55.07	48	Hirschburger	17	17.57
4	Lamparter	47	49.08	49	Weinmann	17	17.35
5	Hummel	45	45.05	50	Wucherer	17	17.13
6	Grüninger	43	42.05	51	Schaupp	16	16.91
7	Bihler/ Bühler	41	39.70	52	Schneider	16	16.69
8	Fucks	41	37.78	53	Trißler	16	16.47
9	Göbel	40	36.17	54	Vohrer	16	16.26
10	Benz	39	34.78	55	Elwert	15	16.05
11	Braun	35	33.56	56	Engel	15	15.84
12	Grözinger	34	32.48	57	Keim	15	15.63
13	Reicherter	33	31.52	58	Knapp	15	15.42
14	Kurtz	31	30.64	59	Wendler	15	15.21
15	Röhm	31	29.84	60	Kenngott, Könn-	14	15.00
16	Walz	31	29.11	61	Mössinger	14	14.79
17	Beck	28	28.43	62	Zindel	14	14.58
18	Faißt	28	27.80	63	Ammer	13	14.37
19	Göppinger	28	27.21	64	Aickelin	13	14.16
20	Maier	28	26.66	65	Geckeler	13	13.95
21	Fischer	27	26.14	66	Kiefner	13	13.74
22	Kalbfell	27	25.64	67	Dentzel	12	13.53
23	Astfalk	26	25.18	68	Rall	12	13.31
24	Schmid	25	24.73	69	Schleicher	12	13.09
25	Faßnacht	24	24.31	70	Uber	12	12.87
26	Hecht	24	23.90	71	Epp	11	12.65
27	Weiß	24	23.51	72	Helbling	11	12.42
28	Finckh	23	23.14	73	Mauerhan	11	12.19
29	Hammer	23	22.78	74	Metzger	11	11.95
30	Lachmann	23	22.44	75	Weckler	11	11.71
31	Rösch	23	22.10	76	Bauer, Baur	10	11.46
32	Spannagel	23	21.78	77	Döttinger	10	11.20
33	Tochtermann	23	21.46	78	Jäger	10	10.93
34	Vollmer	23	21.16	79	Launer	10	10.66
35	Eisenlohr	22	20.87	80	Lumpp	10	10.37
36	Ruoff	22	20.58	81	Maurer	10	10.06
37	Zwißler	22	20.30	82	Roth	10	9.74
38	Ankele	20	20.02	83	Rupp	10	9.39
39	Gminder	20	19.76	84	Werwag	10	9.02
40	Heß	20	19.50	85	Horrwarth	10	8.61

41	Müller	20	19.24	86	Bantlin	9	8.15
42	Schäfer	20	18.99	87	Bertsch	9	7.62
43	Hirrlinger	18	18.74	88	Beutel	9	6.98
44	Ochs	18	18.50	89	Brucklacher	9	6.15
45	Schauwecker	18	18.26	90	Sautter	9	4.86
$K = 1.9441$, $M = 0.6822$, $n = 89$, $X^2 = 15.0466$, $FG = 86$, $P \approx 1.00$							

Die ungewöhnliche Reihenfolge der Namen erklärt Bacmeister (1870a:7) damit, dass Reutlingen „bis in die neuere Zeit sich ziemlich abgeschlossen gegen fremden Zuzug gehalten" habe. Für das benachbarte Eningen gibt Bacmeister (1870a: 8) an:

Tabelle 4
Anpassung der negativen hypergeometrischen Verteilung
an die häufigsten Familiennamen von Eningen im 19. Jahrhundert

Rang	Name	n_x	NP_x	Rang	Name	n_x	NP_x
1	Rall	131	140.64	14	Eger	17	18.01
2	Koch	70	71.44	15	Groß	17	16.87
3	Hummel	53	53.31	16	Fausel	13	15.79
4	Leuze	51	44.00	17	Hoffmann	11	14.77
5	Maier	44	38.03	18	Keck	11	13.78
6	Jäger	37	33.75	19	Knieß	11	12.82
7	Sautter	36	30.46	20	Schaufler	11	11.87
8	Mühleisen	26	27.81	21	Beck	10	10.91
9	Kuhn	25	25.60	22	Dürr	10	9.92
10	Eitel	22	23.71	23	Passauer	10	8.89
11	Lotterer	22	22.05	24	Bader	8	7.74
12	Wick	21	20.58	25	Kromer	8	6.41
13	Hespeler	20	19.24	26	Mader	8	4.59
$K = 1.8847$, $M = 0.5155$, $n = 25$, $X^2 = 9.9121$, $FG = 22$, $P = 0.99$							

Statt der 1-verschobenen negativen hypergeometrischen Verteilung kann man an die Rangordnung der Nachnamen auch Altmanns Modell für beliebige Rangordnungen (Altmann 1993) mit ebenfalls sehr guten Ergebnissen anpassen.

In einem weiteren Artikel (Bacmeister 1870c: 97) geht er auf die Frage ein, wie viele Wurzeln man für eine Sprache annehmen muss. Zum Chinesischen führt er aus, diese Sprache habe 450 „begriffsbildende Lautgruppen", die aufgrund der Töne „1200 bis 1500 verschiedene Klanggruppen" bilden. Damit glaubt er in Übereinstimmung mit anderen Autoren die Dimension des Wortschatzes getroffen zu haben. Er verweist in diesem Zusammenhang u.a. auf Pott, der die durchschnittliche Zahl der Wurzeln je Sprache auf ca. 1000 geschätzt habe.

Bacmeister gehört aufgrund seiner Erhebungen zu den frühen deutschsprachigen Autoren, die sich der Statistik bedienen, um sprachliche Verhältnisse zu charakterisieren, im Falle der Nachrichtenübertragung aber darüber hinaus auch mit dem Ziel, ganz praktische, in diesem Fall ökonomische Probleme anzugehen. Grund genug, ihn in die Annalen der Quantitativen Linguistik aufzunehmen.

(Die folgende Literaturliste verzeichnet auch Publikationen, die für die Quantitative Linguistik nicht relevant sind; sie soll dazu beitragen, das Bild von Bacmeister zu ergänzen.)

Literatur

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

Bacmeister, Adolf (1858). *Das Nibelungenlied für die Jugend bearbeitet.* Stuttgart: Neff. (Weitere Auflagen 1874, 1886)

Bacmeister, Adolf (1860). *Gudrun. Altdeutsches Heldengedicht neudeutsch bearbeitet.* Reutlingen. (2. Aufl. Stuttgart: Neff o.J. [1874].)

Bacmeister, Adolf (1860). *Die Oden des Qu. Horatius Flaccus im Versmaß des Urtextes verdeutscht.* Stuttgart: Neff.

Bacmeister, Adolf (1861). *„freidanks bescheidenheit". Spruchsammlung aus dem 13. Jahrhundert. Neudeutsch bearbeitet.* Reutlingen: Palm.

Bacmeister, Adolf (1861/ 1867). *Margret More, Tagebuch 1522-1535.* Deutsch von Adolf Bacmeister. Stuttgart: Mäcken.

Bacmeister, Adolf (1867). *Alemannische Wanderungen. 1. Ortsnamen der keltisch-römischen Zeit. Slavische Siedlungen.* Stuttgart: Cotta.

Bacmeister, Adolf (1870). *Germanistische Kleinigkeiten.* Stuttgart: Kröner.

Bacmeister, Adolf (1870a). Alte Familiennamen. In: Bacmeister (1870), S. 1-52.

Bacmeister, Adolf (1870b). Stenotelegraphie. In: Bacmeister (1870), S. 73-80.

Bacmeister, Adolf (1870c). Der Ursprung der Sprache. In: Bacmeister (1870), S. 94-102.

Bacmeister, Adolf (1872). *Cornelius Tacitus. Das Leben des Julius Agricola.* Uebersetzt von Adolf Bacmeister. Stuttgart: Neff.

Bacmeister, Adolf (1874). *Keltische Briefe.* Hrsg. v. Keller, Otto. Straßburg: Trübner.

Bacmeister, Adolf (1886). *Abhandlungen und Gedichte. Mit einer Biographie Bacmeisters* hrsg. v. Hartmann, Julius, Klaiber, Julius, & Schmid, Rudolf. Stuttgart: Kohlhammer. (Das Buch enthält im Anschluss an die Biographie ein Verzeichnis der Bücher Bacmeisters.)

Bacmeister, Adolf, & Keller, Otto (1891). *Die Briefe des Quintus Horatius Flaccus. Im Versmaß der Urschrift verdeutscht.* Leipzig: Teubner.

Best, Karl-Heinz (2006). Zur Häufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten. *Glottometrics 11, 9-31.*

DBE: *Deutsche Biographische Enzyklopädie (DBE).* Bd. 1. Hrsg. v. Walter Killy. München u.a.: K.G. Sauer 1995.

Greenberg, Joseph H. (1960). A Quantitative Approach to the Morphological Typology of Languages. *International Journal of American Linguistics 26*: 178-194.

Haering, Hermann, & Hohenstatt, Otto (Hrsg.) (1942). *Schwäbische Lebensbilder. 3. Bd.* Stuttgart: Kohlhammer. (Enthält eine Kurzbiographie Bacmeisters.)

Lernoff, Theobald [= Pseudonym für **Bacmeister, Adolf**] (1860). *Deutsche Sonette*. Ulm: Nübling.

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

NDB: *Neue deutsche Biographie. 1. Bd.* Hrsg. von der historischen Kommission bei der bayerischen Akademie der Wissenschaften. Berlin: Duncker & Humblot 1953.

Schmid, Rudolf (1886). Adolf Bacmeisters Biographie. In: Bacmeister (1886), S. VII-XXXV.

Karl-Heinz Best

XXII. Siegfried Behn (1884-1970)

Behn wurde am 3.6.1884 in Hamburg geboren. 1897-1903 Gymnasium in Hamburg und Worms; 1903-1908 Studium der Philosophie, Psychologie und weiterer Fächer in München und Heidelberg; 1908 Promotion in Heidelberg; 1908-1913 weitere Studien in München, Heidelberg, Zürich und Bonn; 1913 venia legendi für Philosophie und Psychologie in Bonn; 1913-1922 Privatdozent für Philosophie in Bonn; 1922-1933 a.o. Professor für Philosophie und zeitweise überschneidend Dozent, später Professor an der Pädagogischen Akademie in Bonn; 1931 o. Professor für Philosophie unter besonderer Berücksichtigung der experimentellen Pädagogik an der Universität Bonn, 1937-1949 o. Professor für Philosophie und Psychologie an der Universität Bonn. Im Mai 1945 übernimmt Behn die Leitung der Nachrichtenkommission, die das Verhalten der Universitätsangehörigen in der Zeit des Nationalsozialismus zu überprüfen hatte. 18.10.1949 Emeritation. Gest. 27.11.1970 in Bonn-Bad Godesberg.

Behn ist für die Quantitative Linguistik von Bedeutung, weil er als Psychologe in seinen experimentellen Untersuchungen zum Rhythmus literarischer Werke schon früh dem Zusammenhang auf die Spur gekommen ist, der später als „Menzerath-Altmann-Gesetz" (Altmann & Schwibbe 1989; Asleh & Best 2004/05) bekannt geworden ist: „Es besteht also eine Tendenz, die Dauer verschiedener Zeilen einander anzugleichen." Und: „Zeilen mit mehr Silben werden gegen solche mit weniger Silben [in ihrer Dauer; Verf.] verkürzt, oder umgekehrt: Zeilen mit weniger Silben werden gegen solche mit mehr Silben verlängert" (Behn 1912: 97). Mit dieser Erkenntnis gehört Behn als einer der frühesten Autoren in die Vorgeschichte des Menzerath-Altmann-Gesetzes, dessen Entwicklung von Altmann & Schwibbe (1989: 60) und Cramer (2005: 659f.) skizziert wird. Allerdings wirft seine Version möglicherweise theoretische Probleme auf, da er Silben in einen Zusammenhang mit Gedichtzeilen bringt. Hier muss man die Frage stellen, wo in der Hierarchie sprachlicher Einheiten die „Zeile" anzusiedeln ist? Setzt man „Zeile" einmal mit „Satz" gleich, stimmt das Verhältnis: Je länger der Satz, desto kürzer die Teilsätze; je kürzer die Teilsätze, desto länger die Wörter. Bis hierher hat man es mit dem sog. Arensschen Gesetz zu tun (Altmann & Schwibbe 1989: 46ff.), das zumindest als Trend hinreichend empirisch abgesichert ist. Die nächste Verbindung ist dann: Je länger die Wörter, desto kürzer die Silben. Damit wäre Behns Feststellung bestätigt: Je länger die Zeile (≈ Satz), desto kürzer die Silbe. Der Trend müsste allerdings noch schwächer ausgeprägt sein, als es beim Arensschen Gesetz der Fall ist, da ja noch eine hierarchische Ebene hinzukommt. Nun kann man aber keineswegs „Zeile" immer mit „Satz" gleichstellen. Womit aber dann? Eine Zeile wird oft weniger als einen Satz, manchmal aber auch mehr als einen solchen enthalten. Man tut wohl gut daran, Behns Feststellung als Hypothese aufzufassen und einer Prüfung auf der Grundlage hinreichender Daten zu unterziehen.

Behn ist aber noch in einem weiteren Zusammenhang interessant, ohne damit selbst als Quantitativer Linguist in Erscheinung zu treten: Er unterscheidet bei der Analyse von Gedichten 5 Betonungsstufen: 1: leicht betont, 2: schwach betont, 3: voll betont, 4: stark betont, 5: schwer betont (Behn 1921: 278; ähnlich 1912: 40) und analysiert einige Beispiele (1912: 115ff.; 1921: 279). Interpretiert man diese Betonungsgrade einmal als Komplexitätsstufen und stellt sie für die Gedichte, die Behn entsprechend markiert hat, in einer Tabelle zusammen, kann man zeigen, dass sie der 1-verschobenen Hirata-Poisson-Verteilung (Wimmer & Altmann 1999: 256)

$$P_x = \sum_{i=0}^{\left[\frac{x}{2}\right]} \binom{x-i}{i} \frac{e^{-a}\, a^{x-i}}{(x-i)!} b^i (1-b)^{x-2i}, \qquad x = 1,2,...$$

folgen. Dieses Modell gehört zu den Verteilungen, die für das Vorkommen von Einheiten unterschiedlicher Komplexität vorgeschlagen wurden (Wimmer & Altmann 1996: 129).

Tabelle 1
Heine, Was will die einsame Träne

x	Betonungsgrad	n_x beobachtet	n_x berechnet
1	leicht betont	31	29.00
2	schwach betont	9	7.28
3	voll betont	32	33.08
4	stark betont	4	8.15
5	schwer betont	37	35.50
$a = 1.3601$ $b = 0.8155$ $X^2 = 2.7561$ $FG = 2$ $P = 0.25$			

(Behn 1912, 115f., zu: Heinrich Heine, Was will die einsame Träne, in: Buch der Lieder. Die Heimkehr 1823-1824, Nr. 27.)

Legende zu den Tabellen:
x = Betonungsgrade
n_x = beobachtete Zahl der Silben mit Betonungsgrad x
FG = Freiheitsgrade
a, b = Parameter der Hirata-Poisson-Verteilung
X^2 = Wert des Chiquadrats
P = Überschreitungswahrscheinlichkeit des betreffenden Chiquadrats
Die Verteilung erweist sich als geeignetes Modell für die beobachteten Daten, wenn $P \geq 0.05$, was in allen drei Fällen gegeben ist.

Tabelle 2
Heine, Ich wollt‘ meine Schmerzen ergössen sich

x	Betonungsgrad	n_x beobachtet	n_x berechnet
1	leicht betont	16	20.89
2	schwach betont	7	5.09
3	voll betont	32	25.81
4	stark betont	4	6.18
5	schwer betont	30	31.03
$a = 1.4492$ $b = 0.8320$ $X^2 = 4.1545$ $FG = 2$ $P = 0.13$			

(Behn 1912, 116f., zu: Heinrich Heine, Ich wollt‘ meine Schmerzen ergössen sich, in: Buch der Lieder. Die Heimkehr 1823-1824, Nr. 61.)

Tabelle 3
Keller, Nacht

x	Betonungsgrad	n_x beobachtet	n_x berechnet
1	leicht betont	57	57.49
2	schwach betont	54	53.92
3	voll betont	43	40.69
4	stark betont	20	22.35
5	schwer betont	18	17.54
$a = 1.2058$ $b = 0.2222$ $X^2 = 0.3951$ $FG = 2$ $P = 0.82$			

(Behn 1921, 279, zu: Gottfried Keller, Nacht)

Behns Unterscheidung von 5 Betonungsgraden ermöglicht es, ihre Verteilung in den von ihm selbst akzentuierten Gedichten zu überprüfen. Es erweist sich in allen drei Fällen als möglich, eines der bekannten und theoretisch begründeten Modelle an diese Texte anzupassen, die Hirata-Poisson-Verteilung, die sich bisher bei Wortlängenverteilungen im modernen Französischen (Dieckmann & Judt 1996; Feldt, Janssen & Kuleisa 1997) sowie im Schweizerdeutschen (Stark 2001) bewährt hat. Auffällig ist, dass bei den Heine-Gedichten nicht, wie sonst üblich, eine kontinuierliche Abnahme von den weniger komplexen zu den aufwendigeren Einheiten zu beobachten ist, sondern eine erhebliche Schwankung. Eine solche Datenstruktur wurde bisher nur selten, z.B. bei Wortlängen in chinesischen Texten vorgefunden (Zhu & Best 1992; Best & Zhu 1994: 28; 2001). Damit zeigt sich, dass die Theorie sich nach Wort- und Satzlängen und etlichen anderen Entitäten mit den Betonungsgraden anscheinend in noch einem, bisher nicht behandelten Bereich bewährt. Da es sich nur um Daten zu 3 Texten handelt, muss man natürlich vorsichtig sein, zumal die Entscheidung über die Betonungsgrade sicher nicht immer unproblematisch ist, auch wenn Behn (1912: 102f.) ein möglichst operationales Verfahren dazu angibt.

(Die Darstellung der wichtigsten biographischen Daten erfolgte aufgrund der Informationen der Universität Bonn und des *Biographisch-Bibliographischen Kirchenlexikons*, die problemlos im Internet zugänglich sind. In diesen Quellen sind auch hinreichend Informationen über die weiteren wissenschaftlichen Aktivitäten und Werke von Behn erhältlich; die Literaturliste beschränkt sich daher auf die wenigen einschlägigen Arbeiten.)

Literatur

Altmann, Gabriel, & Schwibbe, Michael H. (1989). *Das Menzerathsche Gesetz in informationverarbeitenden Systemen*. Hildesheim: Olms.

Asleh, Laila, & Best, Karl-Heinz (2004/05). Zur Überprüfung des Menzerath-Altmann-Gesetzes am Beispiel deutscher (und italienischer) Wörter. *Göttinger Beiträge zur Sprachwissenschaft 10/11, 9-19.*

Behn, Siegfried (1912). *Der deutsche Rhythmus und sein eigenes Gesetz. Eine experimentelle Untersuchung.* Straßburg: Trübner.

Behn, Siegfried (1921). *Rhythmus und Ausdruck in deutscher Kunstsprache.* Bonn: Verlag v. Friedrich Cohn.

Best, Karl-Heinz, & Zhu, Jinyang (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In: Klenk, Ursula (Hrsg.), *Computatio Linguae II: 19-30*. Stuttgart: Steiner. (= ZDL-Beiheft 83)

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

Cramer, Irene M. (2005). Das Menzerathsche Gesetz. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch: 659-688.* Berlin/ N.Y.: de Gruyter.

Dieckmann, Sandra, & Judt, Birga (1996). Untersuchung zur Wortlängenverteilung in französischen Pressetexten und Erzählungen. In: Schmidt, Peter (Hrsg.), *Glottometrika 15, 158-165.* Trier: Wissenschaftlicher Verlag Trier.

Feldt, Sabine, Janssen, Marianne, & Kuleisa, Silke (1997). Untersuchung zur Gesetzmäßigkeit von Wortlängenhäufigkeiten in französischen Briefen und Pressetexten. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16, 145-151*. Trier: Wiss. Verlag Trier.

Stark, Alexandra B. (2001). Die Verteilung von Wortlängen in schweizerdeutschen Texten. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 153-161*. Göttingen: Peust & Gutschmidt.

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

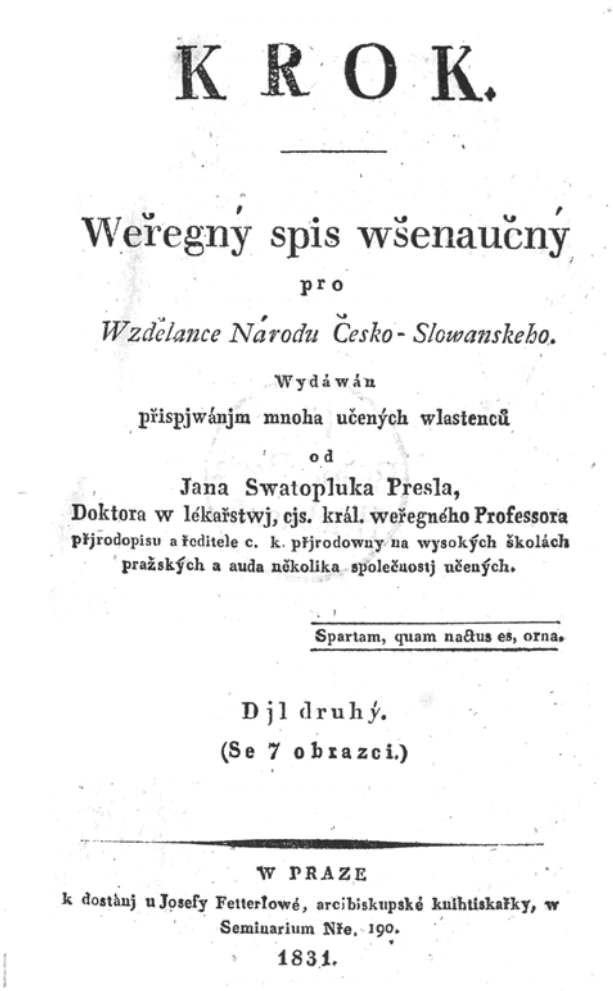
Wimmer, Gejza, & Altmann, Gabriel (1999). *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

Zhu, Jinyang, & Best, Karl-Heinz (1992). Zum Wort im modernen Chinesisch. *Oriens Extremus 35, 45-60.*

Karl-Heinz Best

XXIII. Jan Svatopluk Presl (1791-1849)

K R O K. ***Weřegný spis wšenaučný pro wzdělance národu Česko-Slowanského. Wydáwán přispjwánjm mnoha učených wlastenců od Jana Swatopluka Presla, Doktora w lékařstwj, c. k. weřegného professora přjrodopisu a ředitele c. k. přjrodowny na wysokých školách pražských.***



K R O K.

Weřegný spis wšenaučný

pro

Wzdělance Národu Česko-Slowanskeho.

Wydáwán

přispjwánjm mnoha učených wlastenců

od

Jana Swatopluka Presla,

Doktora w lékařstwj, cjs. král. weřegného Professora přjrodopisu a ředitele c. k. přjrodowny na wysokých školách pražských a auda několika společnostj učených.

Spartam, quam nactus es, orna.

Djl druhý.

(Se 7 obrazci.)

W PRAZE

k dostánj u Josefy Fetterlowé, arcibiskupské knihtiskařky, w Seminarium Nře. 190.

1831.

Letter and sound frequencies obtained an increasing interest in the 19th century , relevant research arising in the first half of the century and being extended later. To a certain extent, the overall interest of this increase was motivated linguistically; in this case, it has to be seen as a specific continuation of comparative approaches in the field of linguistics, ultimately referring back to the Neogrammarians' interest in regularities, or even "laws", underlying language and language development. Later, this predominantly linguistic interest was increasingly accompanied by technical needs, starting from shorthand and stenography, includeing cyphering and decyphering techniques, different type-writer systems, extending to information transfer, etc.

It goes without saying that in the earliest studies, sound and letter statistics were not always strictly distinguished, and if so, results were highly dependent on the sound definitions chosen. As far as the earlist Slavic statistics are concerned, August Schleicher's data on Old Church Slavonic sound frequencies tend to be regarded as one of the earliest letter or sound statistics (cf. Grzybek/Kelih 2003, 2004 2005): Schleicher had published these data in his *Formenlehre der kirchenslawischen Sprache* (1852: 20f.), understanding them as a complement to the earlier data on German, Greek, Latin and Gothic, which Ernst Förstemann (1846, 1852) had provided in the seminal studies.

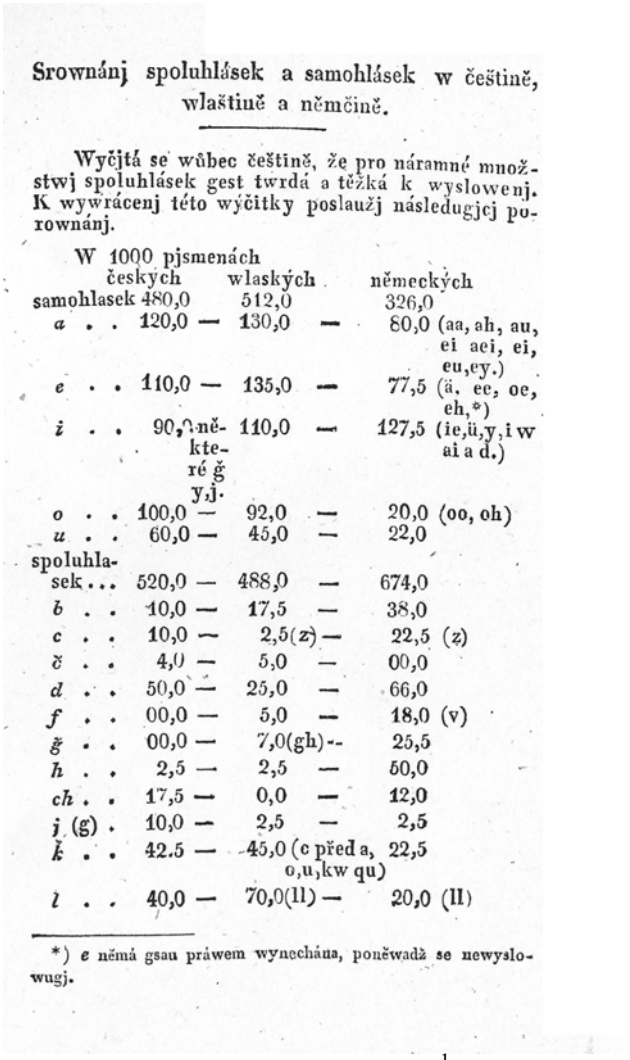
In this context, an almost unknown sound statistic must be presented here which was published almost two decades before the studies mentioned aboved . This statistic seems to be one of the very first letter and sound statistics in general, not only in the Slavic area. It was published in 1831, in the first Czech scientific journal *Krok*, edited by Jan Svatopluk Presl (1791-1849).

Jan Svatopluk Presl was born in Prague. His early interest in medicine culminated in his study at Karl-Ferdinand University in Prague from which he graduated as a doctor of medical sciences, in 1816. After being assistant for zoology and mineralogy, from 1816 to 1818, he became professor for these subjects at Oloumuc university, and was then appointed professor for zoology and mineralogy at Prague University, in 1820. He soon became one of the most important scientists of his country, who made important contributions to the (Czech) scientific terminology in botanics, chemics, zoology, mineralogy, geology, paleontology. In fact, with these works Presl inspired the development of Czech scientific terminology, thus making it possible for original Czech research to develop. To be sure, at a time when not Czech, but rather German was the standard language for official communication (e.g., in Czech elementary schools and the administrative bureacracy), Presl became very active in promoting Czech language and culture, hereby being in closest contact with the leading thinkers of his time, who were promoters and representatives of the Czech National Revival (České národní obrození). This cultural movement – the main purpose of which was to revive Czech language, culture and national identity – was headed by Josef Dobrovský (1753-1829). and Josef Jungmann (1743-1847). Whereas Dobrovský has often been called one of the founders of slavistics as an academic discipline, Jungmann is prediminantly famous for his *Historie literatury i jazyka českého* (1825) and his five-volume *Slovnik jazyka českého* (1835-39), a major lexicographical work, which had a great formative influence on the Czech language.

Together with Dobrovský and Jungmann, but also with Václav Hanka, Jan E. Purkyně, and others, Presl founded a society aiming at the promotion of Czech in academic life, which

initially pursued the idea of founding a Slavic Academy in Prague. Presl became actively engaged in founding the Czech National Museum; together with Jungmann, Hanka, Pavel J. Šafárik (1795-1876) and František Palacký (1798-1876) Presl also founded the Committee for the Scientific Development of the Czech Language and Literature at the museum, from which the literary foundation *Matica česká* emerged in 1831.

In extensive discussions with the group of scholars mentioned above, Presl attempted to prove Czech being an adequate language for scientific discourse. It was on this general background that, in 1821, first Czech scientific journal *Krok* was founded. Most of the above-mentioned scholars contributed to this journal, which appeared with irregular intervals until 1840. Despite his pronounced engagement in Czech cultural affairs, it was only towards the end of his life that Presl became increasingly politically involved: In the revolutionary year of 1848, he was elected into the Czech National Committee; he also became one of the 341 delegates to the All-Slav Congress in Prague, where 42% of the participants were of Czech or Slovak origin. The Congress was headed by Palacký – one of the regular contributors to *Krok.*

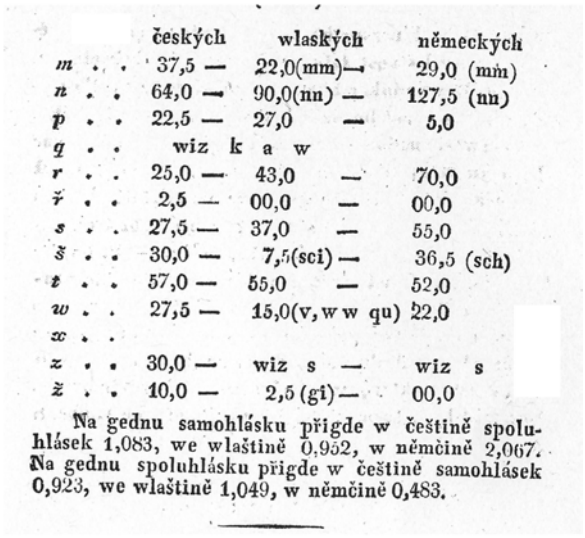


Srownánj spoluhlásek a samohlásek w češtině, wlaštině a němčině.

Wyčjtá se wůbec češtině, že pro náramné množstwj spoluhlásek gest twrdá a těžká k wyslowenj. K wywrácenj této wýčitky poslaužj následugjcj porownánj.

	W 1000 pjsmenách českých	wlaských	německých
samohlasek	480,0	512,0	326,0
a . .	120,0 —	130,0 —	80,0 (aa, ah, au, ei aei, ei, eu,ey.)
e . .	110,0 —	135,0 —	77,5 (ä, ee, oe, eh,*)
i . .	90,0 (některé ğ y,j.)	110,0 —	127,5 (ie,ü,y,i w ai a d.)
o . .	100,0 —	92,0 —	20,0 (oo, oh)
u . .	60,0 —	45,0 —	22,0
spoluhlasek ...	520,0 —	488,0 —	674,0
b . .	10,0 —	17,5 —	38,0
c . .	10,0 —	2,5(z) —	22,5 (z)
č . .	4,0 —	5,0 —	00,0
d . .	50,0 —	25,0 —	66,0
f . .	00,0 —	5,0 —	18,0 (v)
ğ . .	00,0 —	7,0(gh) —	25,5
h . .	2,5 —	2,5 —	50,0
ch . .	17,5 —	0,0 —	12,0
j (g) .	10,0 —	2,5 —	2,5
k . .	42.5 —	45,0 (c před a, o,u,kw qu)	22,5
l . .	40,0 —	70,0(ll) —	20,0 (ll)

*) e němá gsau právem wynechána, poněwadž se newyslowugj.



	českých	wlaských	německých
m . .	37,5 —	22,0(mm) —	29,0 (mm)
n . .	64,0 —	90,0(nn) —	127,5 (nn)
p . .	22,5 —	27,0 —	5,0
q . .	wiz k a w		
r . .	25,0 —	43,0 —	70,0
ř . .	2,5 —	00,0 —	00,0
s . .	27,5 —	37,0 —	55,0
š . .	30,0 —	7,5(sci) —	36,5 (sch)
t . .	57,0 —	55,0 —	52,0
w . .	27,5 —	15,0(v, w w qu)	22,0
x . .			
z . .	30,0 —	wiz s —	wiz s
ž . .	10,0 —	2,5 (gi) —	00,0

Na gednu samohlásku přigde w češtině spoluhlásek 1,083, we wlaštině 0,952, w němčině 2,067. Na gednu spoluhlásku přigde w češtině samohlásek 0,923, we wlaštině 1,049, w němčině 0,483.

The name of Krok is a homonym[1] in Czech, and the name seems to be carefully chosen: On the one hand, it means "step"; and in fact, the foundation of this journal was an important step in Czech cultural development. On the other hand, Krok refers far back to Czech (Bohemian) history, being the name of an Ancient Czech (Bohemian) legendary ruler who reigned in the 7th century. Accordings to the legends – cf. the *Chronica Boemorum* by Bohemian writer and

[1] My thanks to Ludmila Uhlířová who directed my attention to this ambivalence.

historian Cosmas of Prague (ca. 1045-1125) – , one of his three daughters, Libuše is the foundress not only of Prague, but also, by her marriage with Přemysl, the foundress of the Přemyslid dynasty who ruled in the Czech lands from 873 or earlier until the 14th century.

The sound statistics published in volume 3 of *Krok* includes data on Czech, Italian, and German. The author's introductory words very clearly put the whole publication into the cultural background described above: after all, the statistics are meant to provide counter-evidence as to the alleged difficulty of pronunciating Czech. Therefore, consonant-vowel combinations are computed, resulting in similar relations for Czech and Italian, as opposed to German.

It goes without saying, that these data need more thorough re-analysis than this can be done in this context, the more since the data are accompanied by some methodological flaws. For example, no information is given as to the text material analysed, or as to the sample size. Furthermore, no raw data are given, but only percentages (or rather promille, ‰), summing up to 1 only in case of Czech and German, thus obviously containing some errors as to Italian. Anyway, the data are supposed to represent sound, not letter frequencies. This can be seen, for example, from the treatment of German where not letter combinations such as <aa>, <ah>, <au> are treated as [a], but also <ei> or <eu>; similarly, not only <e>, but also the <ee>, <ä> or <oe> are counted as [e]. Also, double consonants like <ll>, <mm> or <nn> are counted as [l], [m], or [n], respectively. The letter <z> is counted as [c] if realized as affricate (usually being denoted as [ts], today), voiceless <v> as [f], etc. etc. – for further details see the reproduction above.

As far as we know, the publication in *Krok* is one of the earliest sound statistics we know of, at least with regard to Slavic languages. For the history of quantitative linguistics, it would be interesting and valuable to find earlier references.

References

Hoffmannová Eva (1973). *Jan Svatopluk Presl, Karel Borivoj Presl.* Praha.

Kimball, Stanley B. (1973). The Austro-Slav Revival: A Study of Nineteenth-Century Literary Foundations. *Transactions of the American Philosophical Society*, *NS, 63(4), 1-83.*

Kolari, Veli (1973). Notes on Jan Svatopluk Presl as Terminologist. *Scando-Slavica 19, 187-196.*

Kolari, Veli (1981). *Jan Svatopluk Presl und die tschechische botanische Nomenklatur: Eine lexikalisch-nomenklatorische Studie*. Helsinki.

Moritsch, Andreas (Hrsg.) (2000). *Der Prager Slavenkongress 1848*. Wien etc.

Moritsch, Andreas (2000). Revolution 1848 – Österreichs Slaven wohin? In: Moritsch (Hg.) (2000): *5-18.*

Pokorný, Jiří (2000). Der Prager Slavenkongress und die Tschechen. In: Moritsch (Hg.) (2000): *63-70.*

Weitenweber, Wilhelm Rudolf (1854). Denkschrift über die Gebrüder Joh. Swat. Und Carl Bor Presl. *Abhandlungen der Königlich-Böhmischen Gesellschaft der Wissenschaften, Folge 5, Bd. 8, 4-27*.

Peter Grzybek

XXIV. Tomo Maretić's first Croatian and/or Serbian Sound Statistics (1899)

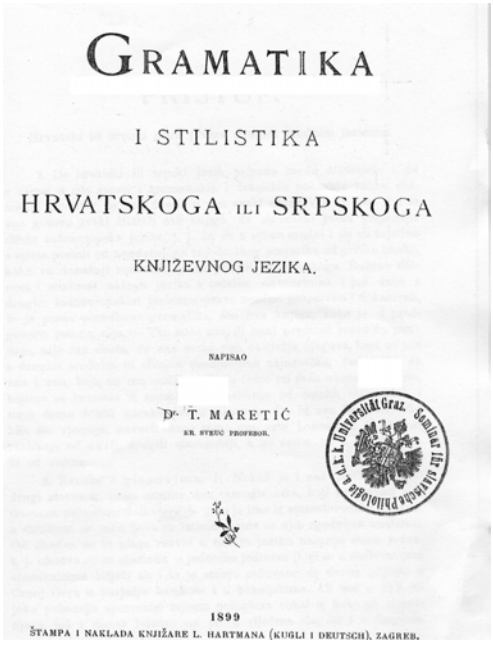
The first Croatian and/or Serbian sound statistic we know of was published by the renowned linguist Tomo Maretić in 1899.

Tomo Maretić is considered to be one of the greatest of Croatian linguists. He was born on October 13, 1854 in Virovitica, a small Croatian town in the region of Slavonia. Here, he attended primary school, and then attended high school at Varaždin, Slavonska Požega and Zagreb. In 1875, he enrolled to study Slavic languages, and simultaneously Latin and Greek, at the Philosophical Faculty of Zagreb University.

Having passed his exams in Classical Philology, intending to become a middle school teacher, he began his Ph.D. studies in Slavistics, and received his doctorate in 1883.

After spending some time at well-known European universities, namely in the Neogrammarians' center of Leipzig (where Leskien was the leading Slavist), and Prague, Tomo Maretić first was appointed extraordinary professor for "Slavic philology with particular emphasis on Croatian and Serbian history of language and literature", in 1886, and ordinary professor three years later. Also in 1886, he became a corresponding member of the Yugoslav Academy of Science and Art, to which he was elected a full member four years later. Later, he twice became head of the philological-historical class of the academy, first from 1906-1913, then a second time from 1919-1928, after being president of the Academy in the years 1915-1918. Tomo Maretić died on January 15, 1938.

Maretić's linguistic oeuvre comprises more than a hundred important contributions to Slavic languages and literatures in general, and to Serbian and Croatian philology in detail. It is impossible to mention all his works here; only some major achievements can be cited by way of example, in order to demonstrate the broad spectrum of his interests. Maretić's interest in linguistics can be traced back to the late 1870s when he published his first work on accentology. Basically, this study was an addendum to one of his many translations from Greek; in his commentary, Maretić attempted to prove an earlier claim (by writer, translator and theoretician Ivan Trnski) that Croatian prosody is based on a particular type of accent (*štokavski*), rather than on the quantity of syllables. Later, Maretić continued this line of research, for example in his study *O nekim pojavama kvantitete i akcenta u jeziku hrvatskom ili srpskom* (1883). In the early 1880s, he published some important works on folkore, *O narodnoj zagonetci hrvatskoj* (1881) and *Studije iz pučkog vjerovanja i pričanja Hrvata i Srba* (1882). A third major field of interest, in addition to accentology and folklore, was lexicology; this is well documented in works like *O narodnim imenima i prezimenima u Hrvata i Srba* (1886), *Ruske i češke riječi u književnom hrvatskom jeziku* (1892) or *Imena rijeka i potoka u hrvatskim i srpskim zemljama* (1893). Maretić's lexicological and lexicographic competence is best proven, of course, in his engagement in the *Rječnik hrvatskoga ili srpskoga jezika*: Maretić edited six of the 23 volumes of this outstanding dictionary, the compilation of which had begun in 1880 on the initiative of Đuro Daničić. As a professor at the Philosophical Faculty of Zagreb University, Maretić was, however, obliged to work not only on linguistics, but on literature, as well; thus, his early interest in all aspects of philology in a broad understanding of this term, continued in his later works, as well. In his literary studies, however, Maretić rather concentrated on folk literature, as for example in his well-known *Metrika narodnih naših pjesama* (1909). Maretić's official academic carreer ended in 1914; yet, he returned to university


GRAMATIKA

I STILISTIKA

HRVATSKOGA ILI SRPSKOGA

KNJIŽEVNOG JEZIKA.

NAPISAO

Dr. T. MARETIĆ

1899

ŠTAMPA I NAKLADA KNJIŽARE L. HARTMANA (KUGLI I DEUTSCH). ZAGREB.

from 1919-24, teaching Indo-European studies Irrespective of the high value of all these studies, which are very much appreciated still today, Maretić's magnum opus remains his *Gramatika i stilistika hrvatskoga ili srpskoga jezika* (1899, ²1932, ³1963). This book served as the basis for linguistic education of several generations of Croatian and Serbian linguists. It was in this book that, as far as we know, the first Croatian or Serbian statistics of this kind were published.

In his introductory ruminations on the overall position of Croatian or Serbian in the context of other Slavic languages, Maretić mainly concentrates on differences with regard to sounds and cases. The discussion of sounds is then intensified in the first chapter (p. 9ff). According to his presentation, Croatian or Serbian is characterized by 31 sounds, one of which has no alphabetic correspondence, neither in the Latin nor the Cyrillic variant. This particular sound [ç], according to Maretić, occurs quite rarely; we are concerned here with an assimilating consonant, which may precede the sound [c] – we would rather denote it as [ts] today –, just like [d] preceding [t] or [g] preceding [k]. Maretić conseqeuntly assumes the basic sound inventory to consist of the following 30 items, which are identical with the corresponding letters:

a, b, c, č, ć, d dž, đ, e f, g, h i, j, k, l, lj, m, n, nj, o, p, r, s, š, t, u, v, z, ž

а, б, в, г, д, ђ, е, ж, з, и, ј, к, л, љ, м, н, њ, о, п, р, с, т, ћ, s, ф, х, ц, ч, џ, ш

Starting from the more or less basic observation that the frequency of individual sounds differs for any given language, Maretić presents a table of the frequency of Croatian or Serbian sounds. Maretić selected ten passages per 1,000 sounds (i.e., random samples) from Vuk Karadžić's translation of the *New Testament* and then counted the frequency of all sounds in the order of their occurence. The authenticity of this dana material is not unproblematic. After all, Vuk's translation, published in Vienna in 1847, was based on an earlier «Serbian» translation by Atanasije Stoiković, a Serbian writer who worked as a professor at the Russian university of Kharkov. Stojković's translation, however, which had been published by the Russian Bible Society in Saint Petersburg in 1824, was not written in the vernacular, but represented a mixture of Church Slavonic and Serbian (Slavenoserbian). In contrast to Maretić's problematic choice of his text basis, he displayed an enormous carefulness and a suprisingly high degree of methodological reflection in treating this material: according to his information, the total sum of 10,000 sounds was based on approximately 2,000 words; assuming that it would take about half an hour to pronounce this amount of material, Maretić assumed it to be sufficient for his purposes, referring to the similar approach by William Dwight Whitney in his analysis of Old Indian sounds.[1] In order to prevent particular biases, Maretić selected only passages in which one and the same words were repeated as rarely as possible, referring to the fact that in the specific text of the *New Testament*, the occurrence and repetition of specific proper nouns may significantly change the frequency structure; for

[1] Maretić referred to the German translation of Whitney's *Sanskrit Grammar: Including Both the Classical Language, and the Older Dialects, of Veda and Brahmana,* published under the title of *Indische Grammatik: umfassend die klassische Sprache und die älteren Dialecte* (Leipzig, 1879).

the same reason, he tried to avoid passages with foreign words. Further, in order to prevent ordinary errors, he counted all passages more than once. As to the definition of the items counted, it seems important to take into consideration that Maretić counted all sounds *as they are pronounced, not as they are written*. Therefore he interpreted *s njim* as [š njim], *bratski* and *gradski* as [bracki] or [gracki], respectively; furthermore, words like *donio* or *mislio* are interpreted as [donijo] and [mislijo]. The sound [r] is counted separately, depending on whether it fulfils a vowel or a consonant function *prst* vs *ruka*.

As a result, the sound frequencies indicated in Table 1 are obtained.

	I	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**total**
a	99	127	105	103	114	101	106	104	120	100	1079
b	11	21	10	16	17	15	21	21	23	18	173
c	11	8	5	8	11	2	2	1	3	5	56
ć	24	9	18	12	12	11	16	5	9	4	120
d	2	9	11	9	18	6	8	8	7	13	91
dž (џ)	41	30	45	44	41	50	42	40	53	42	428
đ	0	0	0	0	0	0	1	0	0	0	1
e	104	108	120	113	116	115	89	116	107	111	1099
g	15	17	23	21	11	17	19	23	24	22	192
h	4	9	14	8	9	7	6	7	3	6	73
i	113	119	109	120	103	100	108	81	89	93	1035
j	62	56	42	48	47	52	37	58	51	50	503
k	30	41	28	24	38	34	28	28	33	34	318
l	15	25	16	21	11	18	15	31	23	19	194
lj (љ)	9	12	5	13	6	6	12	11	9	7	90
m	29	36	46	37	54	43	33	34	25	30	367
n	40	41	46	43	48	49	55	47	50	38	457
nj (њ)	3	5	3	10	2	11	10	5	5	9	63
o	104	79	87	91	99	100	100	101	84	115	960
p	30	12	18	26	10	26	30	22	18	34	226
r (vokal.)	5	2	7	5	2	7	5	1	3	8	45
r (kons.)	44	35	41	18	27	36	34	38	32	38	343
s	47	31	42	45	40	44	47	62	59	47	464
š	22	14	13	20	14	13	21	10	9	14	150
t	37	49	29	34	46	31	43	49	38	44	400
u	51	43	43	49	40	33	45	35	53	37	429
v	32	43	41	41	46	48	39	35	38	39	402
z	11	16	21	10	13	11	18	19	16	12	147
ž	2	2	6	5	3	9	8	7	12	10	64
	997	999	994	994	998	995	998	999	996	999	

Table 1 represents the frequency for each of the ten passages, as well as the total of the combined samples. The sounds [f] and [ç] are not listed, since they did not occur once in any of the samples.

Mentioning that the frequency of the sounds [i] would be slightly higher, and the frequency of [j] slightly lower, in case the *ekavian* or *ikavian* variant of the analyzed texts were taken, Maretić in conclusion concentrates on the vowel-consonant proportions. For the

totality of all then texts, this proportion is 46.47% vowels as compared to 53.53% consonants. Maretic interprets this ratio in terms of a language's weakness or hardness, based on the assumption that a higher percentage of consonants renders a language "harder", whereas a higher percentage of vowels makes it "weeker" – meaning that it is more or less easy to pronounce a word. In this respect, Maretić then compared the results obtained with similar data he obtained for German (38.86), Polish (41.43), Russian (41.69), Lithuanian (43.00), Czech and Latin (43.09), French (43.36) Greek (46.01), Italian (47.73), and Old Church Slavonic (48.37), as well as for the Old Indian (43.52) data provided by Whitney (see above). According to Maretić's interpretation, the weakness of Croatian or Serbian is not very different from Italian, and even exceeds Greek, which usually is considered to be one of the most euphonic languages. Having to concede in this context that there may be hard-to-pronounce consonant clusters in Croatian or Serbian, such as *k bratu*, *k zdravome*, *kumstvo*, and others, Maretić refers to the rareness of these cases, on the one hand, and to equivalent German examples such as *Angst*, *herbstlich*, on the other.

Despite Maretić's enormous productivity in the field of lingistics, his analysis of sound frequency reported here remained his only systematic study in this direction; comparable follow-up studies were conducted only decades later. It would lead too far, here, to deal with these later studies; rather, by way of a conclusion, a cautious first attempt to re-analyse Maretić's data shall be presented, showing some remaining problems and pointing out future tasks for systematic study.

Figure 1 represents the result of fitting the negative hypergeometric function to the whole corpus of texts. This distribution is chosen since it has repeatedly turned out to be the best model for Slavic letter, sound, and phoneme frequencies. As can easily be seen, the fit is far from being convincing, with $C = X^2 / N = 0.034$. To be sure, not any one of the otherwise discussed models (such as geometric, Zipf, Zipf-Mandelbrot, and others) yields more satisfying results. Figure 1 very clearly shows the crucial positions responsible for the largest deviations: positions 2, 3, and 4, being clearly underestimated, and positions 1, 5, and 6, clearly overestimated.

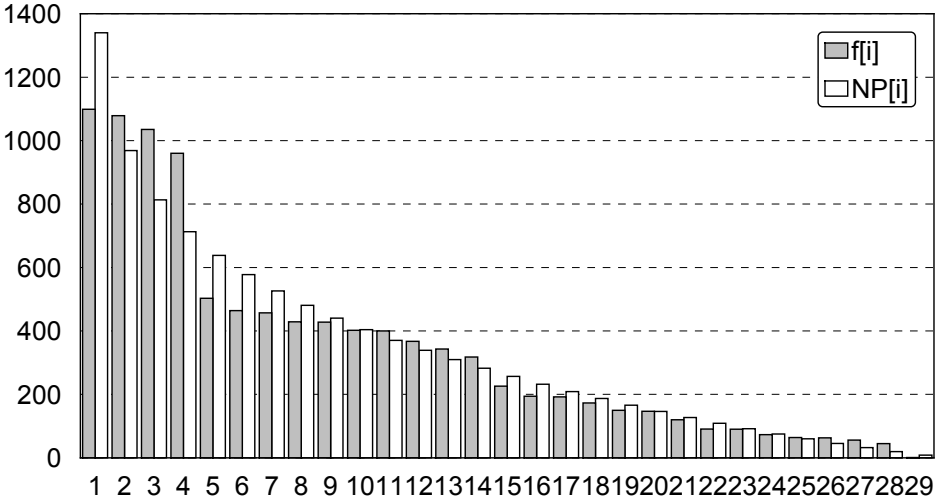


Fig. 1: Fitting the negative hypergeometric distribution to the combined corpus of 10 samples

Yet, analyzing all ten samples individually provides a surprising result: fitting the negative hypergeometric distribution under these circumstances, statistically satisfying results are obtained in all ten cases, though with extremely varying X^2 values (ranging from $0.07 \leq X^2 \leq 0.99$. These findings raise two important questions of rather general kind:

1. The question of data homogeneity comes into play, taking into consideration the good results for the individual texts, on the one hand, and the poor result for the combined corpus.
2. Taking into consideration the diverging X^2 values of the ten individual samples, the question of sample size might turn out to be important, perhaps being the reason for the varying goodness of fit.

With this perspective in mind, a closer look at sample #9 (cf Fig. 2), representing the relatively best statistical result, clearly shows that, despite the good value, it is still the same range of positions which displays the same kind of deviations characteristic of the combined corpus (cf. the deviations in either direction, with the significant overcrossing at position 5).

Thus, in addition to the obviously arising questions of data homogeneity and sample size, mentioned above, we seem to be concerned with some additional problems leading us back to the issues discussed above:

3. Is the chosen text basis authentic language material for Croation or Serbian sound frequencies?
4. Are the definitions of 'sound' valid, and, as a consequence, is the sound inventory as a whole adequately defined?

It seems that there are more questions than answers. In any case, it was Tomo Maretić who, in 1899, asked these important questions with regard to Croatian and / or Serbian, questions which deserve further attention.

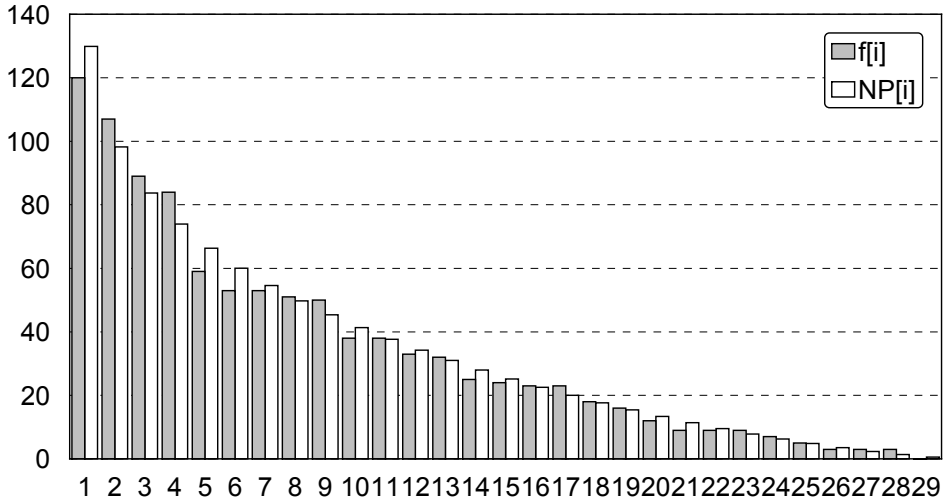


Fig. 2: Fitting the negative hypergeometric distribution to the sample #9

References

Babić, Stjepan (1993). Tomo Maretić 1854-1938. In: *Portreti hrvatskih jezikoslovaca*. Zagreb, 137-143.

Belić, Aleksandar (1939). Dr. Tomislav Maretić. *Naš jezik 6*(*3*)*, 65-69.*

Belić, Aleksandar (1938/39). Dr. Tomislav Maretić. 13.X.1854 – 15.1.1938. *Južnoslovenski filolog 17, 218-221*.

Ivšić, Stjepan (1940). *Prof. dr. Tomislav Maretić*. Zagreb.

Silić, Josip (1984). Tomo Maretić – jedan od najvećih hrvatskih lingvista. In: *Virovitički zbornik 1234-1984.* Virovitica, *395-403*.

Skok, Petar (1949). Tomo Maretić 1854-1938. *Ljetopis jugoslavenske akademije znanosti i umjetnosti 54, 310-335*.

Skok, Petar (1952). Tomo Maretić (1854-1938). In: *Ljetopis jugoslavenske akademije znanosti i umjetnosti za godine 1949-1950, 56, 319-326.*

Peter Grzybek

Book Reviews

Grzybek, Peter (ed.), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. Dordrecht: Springer, 2006. 352 pp. Reviewed by **Ján Mačutek.**

The book under review contains the proceedings of the conference held in Graz and Seggau (Austria), 21-23 June 2002. The contributions are dedicated, or at least related, to word length studies. The 17 articles, with the exception of two introductory contributions, appear in the alphabetical order. Hence, they are divided into thematic sections in the *Preface* by P. Grzybek:

- introduction, history and contemporary approaches (two articles by P. Grzybek),
- theoretically oriented studies (G. Wimmer and G. Altmann; V.V. Kromer),
- works on text corpora (R. Köhler; P. Jakopin; M. Tadić; D. Vitas, G. Pavlović – Lažetić and C. Krstev),
- analyses of some particular questions (A. Wilson; G. Antić, E. Kelih and P. Grzybek; E. Stadlober and M. Djulzelic; O.A. Rottmann),
- contributions discussing word length in a broader context (S. Andersen and G. Altmann; A. Fenk and G. Fenk - Oczlon),
- articles on the interdependence of word length and other linguistic units (W. Lehfeldt; A.A. Polikarpov; U. Strauss, P. Grzybek and G. Altmann).

In the following, the articles are described in the same order as they appear in the book.

Peter Grzybek (Austria): *Introductory Remarks: On the Science of Language in Light of Language of Science* (pp. 1-14).
Some very basic problems of the science of text and language are presented. Like any other science, also this one needs a theory (or theories). They must consist of systems of hypotheses, which are grounded on previous knowledge and are empirically confirmed. If, in addition, the hypotheses are general and systemic, they can be named laws. However, it should be noted that there are no deterministic laws and every law is of a stochastic character. Consequently, after answering objections against quantification in linguistics, there are no obstacles to formulate statistical hypotheses, which can and must be tested. Next, word length is set into a synergetic context. If one accepts a hierarchical structure of linguistic units (beginning with phonemes/graphemes, continuing up to sentences), mutual dependencies of these levels must be considered.

Peter Grzybek: *History and Methodology of Word Length Studies* (pp. 15-90).
The first attempts to study word length are dated as early as 1850s. The works from the first hundred years can be submitted to criticism, however, the pioneers in this field established a basis for the research. Roughly speaking, two lines of mathematical models, which appeared in 1940s, can be noticed: the geometric and the Poisson distributions. Especially the latter one was modified and substantially generalized. Finally, in 1990s the synergetic approach resulted in a unified theory of linguistic laws.

Simone Andersen and Gabriel Altmann (Germany): *Information Content of Words in Texts* (pp. 91-115).

Information content of words is examined from speaker's and hearer's points of view. The authors formulate and test several hypotheses, mainly on the relationship between word length and its information content. They also try to build first mathematical models. A (preliminary) conclusion is that the influence of word length on the speaker's information content is more important than the influence of word frequency.

Gordana Antić, Emmerich Kelih (Austria) and Peter Grzybek: *Zero-Syllable Words in Determining Word Length* (pp. 117-156).

The problem of zero-syllable words in Slavic languages (with an emphasis on Slovenian) is investigated. The orthographic word definition was chosen and word length was measured in syllables. Slovene texts of three types (literary prose, poetry and journalistic prose) were analyzed. There are statistically significant differences between the mean values of mean words length with and without zero-syllable words. One can assume that the decision to take them into account or to neglect them affects word length.

August Fenk, Gertraud Fenk-Oczlon (Austria): *Within-Sentence Distribution and Retention of Content Words and Function Words* (pp. 157-170).

The study, motivated by psychological experiments with recall of listed words, introduces several hypotheses. According to one of them (which was tested and confirmed), content words are more probable to be recalled than function words. Other hypotheses are closely related to word length topic, i.e., content word have the tendency to increase and function words to decrease in the course of a sentence, or, word length seems to increase in the course of a sentence. Empirical tests of these assumptions were not finished at the time of writing the contribution.

Primož Jakopin (Slovenia): *On Text Corpora, Word Lengths, and Word Frequencies in Slovenian* (pp. 171-185).

A survey on Slovene text corpora is presented. The author opens a questions what to do with tokens which do not satisfy definition(s) of word. He lists the most frequent words from different corpora and also graphically compares word length frequencies (measured in letters) for three sources of texts – fiction, newspaper and internet search engine.

Reinhard Köhler (Germany): *Text Corpus as an Abstract Data Structure* (pp. 187-197).

The article discusses some weaknesses of text corpora, and, following the principles of software engineering, suggests (in a very general form) improvements which can help prevent information loss, enable a broader use of a particular corpus, etc. An architecture of a corpus interface is highlighted.

Victor V. Kromer (Russia): *About Word Length Distribution* (pp. 199-210).

The Poisson-uniform distribution is proposed for a model of word length. The model is theoretically derived and there are attempts to interpret it, taking into account the synergetic organization of language. Mutual relations, limits and some special values of parameters are described.

Werner Lehfeldt (Germany): *The Fall of the Jers in the Light of Menzerath's Law* (pp. 211-213).

In this short resume, sound changes in Slavic languages are modelled in accordance with Menzerath's law. Fitting gives very good results for texts written after the fall of the jers.

Anatolij A. Polikarpov (Russia): *Towards the Foundations of Menzerath's Law* (pp. 215-240).

The contribution demonstrate the correlation between average affix length and its position within a word. Prefixes and suffixes behave quite differently, which is explained to be a consequence of their different functions in new words formation process. The derived (and observed) regularities

are set into the Menzerathian context. Theoretical predictions are confirmed on a sample from Russian.

Otto A. Rottmann (Germany): *Aspects of the Typology of Slavic Languages* (pp. 241-258).
The author emphasizes that a typology using arbitrary criteria should be avoided, and, on the contrary, he tries to establish a typology on criteria which are theoretically based. The distributions which fit word length in texts from twelve Slavic languages are interpreted as attractors in the area of word length. The number of attractors is suggested for a typology criterion. Languages with only one attractor (i.e., with one fitting distribution) are supposed to be stabilized; the more attractors a language has, in the greater extent it is "in motion".

Ernst Stadlober, Mario Djulezic (Austria): *Multivariate Statistical Methods in Quantitative Text Analyses* (pp. 259-275).
Several multivariate statistical methods (statistical distances, discriminant functions, canonical discrimination) are used to distinguish three types of Slovene texts (literary prose, poetry and journalistic prose). Word length distributions are described by six variables – four of them depend on moments, the other two on text length. Relevance of particular variables to discrimination is investigated.

Udo Strauss (Germany), Peter Grzybek, Gabriel Altmann: *Word Length and Word Frequency* (pp. 277-294).
Problems related to the Zipf's hypotheses on the relationship between word length and frequency are presented. The appropriateness of the model is tested on eleven texts of different types and languages. Tests of homogenous texts give satisfactory results. On the other hand, some "artificial phenomena" occur in text mixtures.

Marko Tadić (Croatia): *Developing the Croatian National Corpus and Beyond* (pp. 295-300).
The Croatian National Corpus, which has been developed since 1998, is described. Some technical details (sources of texts, encoding standards, programming languages, etc.) are given. Complementarity of corpus and quantitative linguistics is emphasized.

Duško Vitas, Gordana Pavlović-Lažetić, Cvetana Krstev (Serbia): *About Word Length Counting in Serbian* (pp. 301-317).
After discussing problems with word definition in Serbian corpora, the article presents analyses of word length (measured both in characters and syllables) in Serbian. Vowel-consonant structures of different text types are investigated. Frequencies of single words and sequences of two and three words are compared.

Andrew Wilson (United Kingdom): *Word-Length Distribution in Present-Day Lower Sorbian Newspaper Texts* (pp. 319-327).
The contribution brings the first attempt to analyze word length in Sorbian texts. The 1-displaced hyper-Poisson distribution (which is a special case of a theoretically derived model) fits well data from ten newspaper texts.

Gejza Wimmer (Slovakia), Gabriel Altmann: *Towards a Unified Theory of Some Linguistic Laws* (pp. 329-337).
The authors introduce a very general approach which unifies many linguistic laws, forces, hypotheses, distributions, etc. Two parallel models, a discrete and a continuous one, are derived. In both cases, also two-dimensional formulas are suggested. The approach enables us to see a unique mechanism behind linguistic phenomena.

The volume (although concentrating on word length studies) provides an interesting insight into the present state of linguistics, which (finally) begins to use mathematics as its tool in a large scale. It can serve also as a rich source of impulses for future research in different domains of

linguistics, mutatis mutandis. Word length which is an easily accessible variable also for non-linguists was one of the first aspects of language attracting the attention of mathematicians. Though today its study represents a well developed discipline, we are far from having a ripe theory in which all parameters are uniquely interpreted. This will, perhaps, be the task of the future theoretical research. The "normal" research will test the existing hypotheses extending the inventory of languages both geographically and historically.

****

Skalička, Vladimír, *Souborné dílo. III. díl [Collected works. Part III]***.** Praha: Nakladatelství Karolinum 2006. P. 925-1401. Reviewed by **L. Hřebíček** and **G. Altmann.**

The last volume of Skalička´s collected works contains his writings from 1964 to 1994 including also book reviews and small contributions to a Czech encyclopedy. The last article "Praguian typology of languages" was written together with Petr Sgall who systematized Skalička´s doctrine and published the article in 1994 (Skalička died 1991). The editors who performed a heroic work by collecting, translating and ordering 1304 pages of text, added Skalička ´s complete bibliography using different sources. Three registers (Names, Objects, Languages) yield an excellent guide through Skalička´s work. Though the three volumes are now available in Czech, the interested reader can find the majority of the articles in world languages from which they have been translated.

The first most conspicuous feature of all his works is their shortness. He did not like many words, he said what was necessary and passed to the next idea. This was both his speaking and writing custom.

The second important feature is Skalička´s place in the history of linguistics. He is known as typologist but his typology is rather of a synergetic kind, distinguishing him even from present-day typologists. Zipf and Skalička can be considered predecessors of language synergetics. Zipf as quantifier, discoverer of quite new dimensions of language, Skalička as typologist with enormous knowledge of languages seeing everywhere interrelations between language properties. As a matter of fact, he can be called the pope of linguistic interrelations.

Though in more than 300 publications of the three volumes one can find articles on language development, syntax, glottochronology, hyposyntax, sociolinguistics, history of linguistics, morphology, phonemics, style, translation, book reviews, encyclopaedic articles etc., his main interest concentrated to typology and we shall restrict ourselves to this dominant feature of his work.

Though one finds in his work allusions to very modern concepts like redundancy, compensation (in a self-regulating systems we call it functional equivalents, e.g. morphology and syntax are mutual compensations), necessity, probability, possibility, probabilistic relations, etc. he did not use any mathematics. Nevertheless, he gave excellent hints at the study of relations at all levels of language. Skalička defines the tasks of typology as follows: (1) Study of phenomena, (2) study of interrelations, (3) quantification of interrelations (p. 981) which are probabilistic (p. 986). As a matter of fact, he himself came at the second stage and developed it better than any typologist before him, stage three is the initial object of modern synergetic linguistics and is performed with mathematical means.

However, looking retrospectively at his work we see a lot of difficulties connected with the quantification of his interrelations. As a matter of fact, phenomena themselves, e.g. properties, must be quantified, and then interrelations can be ascertained almost automatically.

There are still several other levels above his requirements, namely the theoretical derivation of the interrelations, i.e. their embedding in a theory and at last their explanation by means of laws (subsumption under laws). Peter Sgall asks the essential question "why some feature combine with one another?" (p. 1205). Skalička tried to give some reason or causes for it but the realisation of his program is confronted with enormous problems. Though in typology and language theory we cannot avoid his results and a time must come when everything he said must be "quantified", considering some features that can be found throughout his work we are rather desperate. Let us mention only some of them: root length; word length; distinguishing word classes; word complexity; conversion; number of affixes; extent of derivation; number of synsemantics; affix length; building of compounds; number of preposition and postpositions; homonymy of affixes; synonymy of affixes; fixedness of word order; inflection; internal inflexion; number of clauses; number of endings in the word; morpheme discontinuity; existence of infinitives, participles, verbal nouns; number of vowels and consonants; extent of congruency; differentiation of root and auxiliary elements; differentiation of inflection and derivation; sentence markedness; vowel harmony; existence of suppletivism; article formation; possessivity; extent of declination, which is only a modest extract from the diversity of features. Some of them can easily be quantified, e.g. word length, but the majority of them is a source of enormous problems even from the qualitative point of view. We know, for example, what is word order but what should be quantified? Quantification is not a classification, e.g. in word order types like SVO, OVS etc. it is at least an ordering, scaling etc. How to quantify the "strength of word order" or the "extent of conversion" (theoretical or observed)?

Skalička considers a language type a non-existing theoretical construct, a set (system) of mutually cooperating phenomena (p. 985) realized in a real language to a certain extent. Each language has components of all types. Hence, type is not a class of languages and typology is not classification. Today, we would say that a real language is a kind of steady state in which a function of the vector of all (quantified) properties results in a constant (in different terminology: an attractor). This recalls the first steps of G.K. Zipf in the domain of word frequency distributions. If a feature changes its value, the functional equivalents must change their own values in order to restore the equilibrium. Zipf came to control cycles of maximally four features, Köhler´s system containing also the requirements and mathematical relations contains about 20 features, Skalička´s system without any quantification contains about 100 features. It would be lost of energy trying to enlarge his system, the most important task would be to incorporate his features step by step in Köhler´s system. Even the results of classical typologists like Finck, Steinthal, Misteli, Lewy, Sapir etc. and of all new ones should be embedded in Köhler´s system.

One sees that the fulfilling of this task is not only the highest aim of typology but at the same time a step towards language theory. Language theory does not consist of rules or sentence structures but of stochastic laws whose possible existence has been hinted at by Skalička. He assumed a development in this direction and spoke of the possibility of a deductive typology (p. 1024) which is nothing else but language theory.

Unfortunately, this way will not be simple. Taking two properties whose interrelation will be corroborated in several languages we necessarily meet a language in which it does not hold. Does it mean that the preliminary theory is false? No, it means that a part of the variance (deviations) must be captured by a third variable. But which is the third variable under say 100 possible ones? One cannot check all, either one has luck or one has ingenuous ideas. Skalička showed us a direction but not a way. Besides, no single linguist is competent in so many language as are necessary for corroborating a theory. He himself brought examples from 368 languages and language families, an enormous achievement for a linguist, but examples do not

corroborate anything. A systematic theoretical work accompanied by empirical work of specialists in individual languages is necessary in order to accomplish his program. No future typology can be built without his results unless it remains a torso. Thus a translation of at least his basic works in English would be a very meritorious deed.

Influenced by the typology of his predecessors, Skalička kept the old names of types like inflective, agglutinating, polysynthetic, isolating and introflective but did not consider them as classes but rather as a set of cooperating properties. A language does not belong to a type, in almost all languages almost all types are present. The three volumes are full of properties of individual types, here we present only P. Sgall´s summary (p. 1270f.) of the polysynthetic type (we found more than 30 properties in the work): (a) Some lexical words are used also as auxiliary words and there are no other grammatical means in the morphology; (b) no distinguishing of word classes; (c) word building is performed by composition; (d) there are no affixes and endings, hence they cannot be in opposition; (e) the phonemic form of grammatical morphemes is similar to that of lexical words, hence they can be homonymous; (f) the lack of grammatical means is associated with the grammaticalized word order; (g) compounds fulfil also the functions which in other types are fulfilled by dependent clauses. One easily sees that these properties are extreme cases. In individual languages they are realized "to a certain extent". And this "extent" is exactly what is searched for in synergetic linguistics. Thus starting from Skalička´s enormous number of hypotheses about language structure, one could step by step set up a very ample theory of language. We hope that the edition of his collected work in Czech will stimulate his fellow countrymen to further develop the most promising and the most vital theory that ever arose in typology. But here we see an analogous inscription to that at the entrance of Dante´s hell: "Give up every hope if you enter without mathematics!"