Glottometrics 9
2005

RAM-Verlag

ISSN 2625-8226

Glottometrics

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

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

Herausgeber – Editors

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

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

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

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 9 (2005), Lüdenscheid: RAM-Verlag, 2005. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet (Open Access) unter der Adresse
https://www.ram-verlag.eu/journals-e-journals/glottometrics/ verfügbar.
Bibliographische Deskription nach 9 (2005) **ISSN 2625-8226**

Contents

History of Quantitative Linguistics

Book Review

Glottometrics 9, 2005, 1-12

Sprachliche Einheiten in Textblöcken

Karl-Heinz Best, Göttingen

Abstract. If one segments a text in passages of equal size then it can be shown that entities of different kind abide by a special law known as *Frumkina's law*. In German linguistics this idea has been proposed already by Zwirner & Zwirner about 30 years before Frumkina (1962) as „Blockmethode" (Zwirner 1967: 2449). In this article some further evidence will be given for this law.

1. Entwicklung und Stand des Gesetzes der Textblöcke

Der Gedanke, Textblöcke zu bilden und die Häufigkeit, mit der sprachliche Einheiten in ihnen auftreten, zu untersuchen, ist bereits in den 30er Jahren des 20. Jahrhunderts vorgestellt worden. Die älteste mir bekannte Arbeit hierzu ist Zwirner & Zwirner (1935), in der die beiden Autoren davon berichten, dass sie einen Text von 20000 Lauten in 200 Textblöcke einteilten. Sie stellen in diesem Artikel lediglich dar, welche Verteilung von [œ] sie beobachteten. Diese Werte „folgen nun einem Gesetz", dessen Formel „von Bortkiewicz abgeleitet und mit dem Namen ‚Gesetz der kleinen Zahlen' belegt" wurde (Zwirner & Zwirner 1935: 44), was nur ein anderer Name für die Poisson-Verteilung ist. Zwirner & Zwirner (1935) geben auch an, welche Werte nach diesem Gesetz zu erwarten wären; ihr Ergebnis weicht nur geringfügig von meinen eigenen Berechnungen ab (Best ^{2}2003: 107). Ältere Arbeiten von Zwirner & Zwirner oder anderen Autoren zur „Blockmethode" sind mir nicht bekannt. In einer weiteren Untersuchung legen Zwirner & Zwirner (1938) Daten zu zwei weiteren Lauten vor, die dem gleichen Gesetz folgen.

Die weitere Entwicklung dieses Gesetzes lässt sich wie folgt skizzieren: B. Brainerd (1972) untersuchte das Vorkommen des englischen Artikels; er stellte fest, dass die Poisson-Verteilung hierbei nicht ausreicht und benutzte stattdessen eine Mischung von Poisson-Verteilungen. Angeregt durch die Untersuchung von Frumkina (1962) zur Verteilung russischer Wörter im Werk Puschkins und andere vergleichbare Arbeiten schlugen Altmann & Burdinski (1982: 152f.) vor, dass die Einheiten gemäß der negativen hypergeometrischen Verteilung vertreten sein sollten, die sie aus der theoretischen Verteilung von Einheiten in Textblöcken ableiteten, welche als Grenzfälle auch die Binomialverteilung, die negative Binomialverteilung und die Poisson-Verteilung umfasst (weiter dazu Leopold 1998: 91-97). Das Gesetz wurde dementsprechend *Frumkina-Gesetz* (Altmann 1988: 175) genannt. Altmann & Burdinski (1982) testeten Daten aus den Untersuchungen von Frumkina (1962) und Brainerd (1972) und konnten zeigen, dass diese sich gemäß den von ihnen neu vorgeschlagenen Gesetzmäßigkeiten verhalten. Zur weiteren Erprobung der Theorie wurde die Verteilung einiger Wörter in deutschen, bulgarischen und indonesischen Texten erfolgreich getestet. Einige neue Untersuchungen zum Deutschen und Niederdeutschen unterstützen diese Ergebnisse (Best ^{2}2003: 103-107; Suhren 2002).

Das Gesetz hat sich damit vor allem im Bereich des grammatischem Wortschatzes, aber auch bei Lexemen mehrfach bewährt; hinzu kommen die drei Laute von Zwirner & Zwirner.

Alles in allem ist jedoch die Zahl der Überprüfungen des Gesetzes noch nicht besonders groß. In einem „Schlenker“ deuten Altmann & Burdinski (1982: 147) an, dass dieses Gesetz womöglich auch auf andere Entitäten zutreffen könnte, wenn sie von „words (or other text units)“ sprechen. Tatsächlich konnten Köhler (2001), Piotrowski (1984: 143) und Piotrowski, Bektaev & Piotrowskaja (1985: 252) zeigen, dass auch Drei-Wort-Gruppen und Syntagmen sich sehr regulär verhalten. Piotrowski, Bektaev & Piotrowskaja (1985: 217) wiederum zeigen, dass die Häufigkeit von Substantiven in stichprobenartig erhobenen Textabschnitten eines Romans demselben Gesetz folgt (vgl. auch Altmann 1988: 184). Es ist also klar, dass das Gesetz sich auf der Ebene der Lexik und der Syntax zu bewähren scheint. Wie steht es aber um die kleineren Einheiten: Buchstaben, Laute, Phoneme, Morphe und Silben? Außer den spärlichen Daten von Zwirner & Zwirner (1935, 1938) ist mir dazu nichts bekannt; ihre Präsentation sollte ergänzt werden.

Eine der Voraussetzungen, damit ein Sprachgesetz vorläufig als gültig betrachtet werden kann, ist eine möglichst vielfältige Überprüfung. Mehr als bei anderen gewinnt man beim Studium der einschlägigen Publikationen zu diesem Gesetz, das das Vorkommen von Einheiten in Textblöcken erfasst, den Eindruck, dass Daten erhoben und Tests durchgeführt wurden, diese aber dann oft nur exemplarisch oder gar nicht veröffentlicht wurden. So beklagen Altmann & Burdinski (1982: 150), sie hätten zwar davon erfahren, dass es eine Reihe von entsprechenden Untersuchungen sowjetischer Autoren gebe; diese seien ihnen aber nicht zugänglich. Im Falle von Zwirner & Zwirner konnte ich bisher nur Daten zu den drei Lauten [b], [ě] und [œ] ausfindig machen, obwohl anzunehmen ist, dass sämtliche Laute des Deutschen untersucht wurden. Hinzu kommt, dass die Testergebnisse zu diesen drei Lauten nur unvollständig vorgestellt werden: Es fehlen die Kriterien, die Auskunft darüber geben, wie gut die Übereinstimmung zwischen Beobachtung und Theorie ist und wie die Werte der Parameter sich darstellen. Ein weiteres Defizit besteht darin, dass zu allen Arten von sprachlichen Einheiten bisher anscheinend nur recht wenige Sprachen berücksichtigt wurden.

Die folgenden Ausführungen sollen nun dazu dienen, die Beobachtungen und Tests zum Gesetz der Textblöcke zu ergänzen und damit die genannten Defizite ein wenig zu verringern.

2. Weitere Untersuchungen zum Gesetz der Textblöcke

Im Folgenden werden Tests vervollständigt oder neue Daten geprüft, wobei von den kleineren zu den größeren Einheiten fortgeschritten werden soll.

2.1. Buchstaben

Eine Untersuchung zur Verteilung des Buchstabens <a> wurde von Elmar Schulte (2002) durchgeführt. Er bearbeitete dazu den laufenden Text des Kap. 1 von Heinrich Heine, *Die Bäder von Lucca*. Da dieses Kap. 6498 Buchstaben lang ist, wurden die beiden ersten Buchstaben des folgenden Kap. 2 hinzugenommen, so dass 130 (65) Textblöcke mit je 50 (100) Buchstaben gebildet werden konnten. An diese Daten wurde in Übereinstimmung mit Altmann & Burdinski (1982) die negative hypergeometrische Verteilung

$$P_x = \frac{\binom{-M}{x}\binom{-K+M}{n-x}}{\binom{-K}{n}}, \quad x = 0,1,2,\ldots,n$$

mit Hilfe des Altmann-Fitters (1997) angepasst. Die Überprüfung, ob die Verteilung von <a> dieser Version des Textblock-Gesetzes entspricht, ergab folgendes:

Tabelle 1
Die Verteilung von <a> in Textblöcken

	<a> in 50-Wort-Blöcken		<a> in 100-Wort-Blöcken	
x	n_x	NP_x	n_x	NP_x
0	12	12.09		
1	27	27.95	4	3.82
2	36	34.36	6	7.26
3	31	28.60	10	9.63
4	13	17.19	14	10.74
5	8	7.41	9	10.58
6	2	2.11|	7	9.31
7	1	0.30|	8	7.18
8			5	4.55
9			2	1.93
	$K = 17.5773$ $M = 5.8527$ $n = 7$ $X^2 = 1.53$ $FG = 3$ $P = 0.68$		$K = 5.0624$ $M = 2.3170$ $n = 8$ $X^2 = 2.18$ $FG = 5$ $P = 0.82$	

Erläuterungen zu der Tabelle:
x: Anzahl der Vorkommen des Buchstabens <a> in den Textblöcken;
n_x: beobachtete Anzahl der Textblöcke mit x Vorkommen von <a>;
NP_x: Anzahl der Textblöcke mit x Vorkommen von <a>, berechnet nach der negativen hypergeometrischen Verteilung;
K, M, n: Parameter der Verteilung;
X^2: Werte des Chiquadrates;
FG: Freiheitsgrade;
P: Überschreitungswahrscheinlichkeit für das entsprechende Chiquadrat.
Senkrechte Striche in den Tabellen <|> bedeuten, dass die betreffenden Klassen zusammengefasst wurden. Die Anpassungen der negativen hypergeometrischen Verteilung an die Textdateien werden als erfolgreich betrachtet, wenn $P \geq 0.05$. Das Ergebnis ist in diesen beiden Fällen mit jeweils $P \geq 0.68$ sehr gut. (Die Erläuterungen gelten für die weiteren Tabellen entsprechend.)

Weitere Daten zur Häufigkeit von Buchstaben hat Suhren (2002) erhoben. Gegenstand ihrer Untersuchung war die niederdeutsche Übersetzung von Antoine de Saint-Exupéry, *De lütte Prinz* (Nidderau: Naumann 2000). Dieser Text ist 16730 Wörter lang; er wurde in Abschnitte mit je 50 (334 Textblöcke) bzw. 100 Wörter (167 Textblöcke) eingeteilt. Dann wurde untersucht, wie sich die Buchstaben <e>, <f>, <i>, <u> und <ü> auf die Blöcke verteilen und ob dafür eines der von Altmann & Burdinski (1982) vorgeschlagenen Modelle geeignet ist.

Im Folgenden wird die Anpassung der negativen hypergeometrischen Verteilung an die Buchstabenvorkommen in 50-Wort-Blöcken dargestellt, wobei abweichend von Suhren (2002) nicht die Gesamtdateien bearbeitet wurden, sondern nur die Häufigkeiten, die auch tatsächlich beobachtet wurden. So kommt <e> in keinem der Textabschnitte weniger als sechzehnmal vor; deshalb beginnt die Tabelle auch anders als bei Suhren mit $x = 16$. Die

negative hypergeometrische Verteilung wird hier in verschobener Form angepasst, da $x > 0$; dies gilt entsprechend für die weiteren Tabellen. Die Ergebnisse sind wie folgt:

Tabelle 2
Die Verteilung von <e> in Textblöcken

x	n_x	NP_x	x	n_x	NP_x	x	n_x	NP_x
16	2	0.09|	30	33	21.28	44	2	2.99|
17	3	0.41|	31	31	21.03	45	4	2.23|
18	1	1.07|	32	13	20.34	46	2	1.63|
19	2	2.17|	33	17	19.26	47	0	1.16|
20	2	3.72|	34	11	17.88	48	0	0.80|
21	6	5.68	35	23	16.29	49	1	0.54|
22	12	7.95	36	11	14.55	50	0	0.35|
23	6	10.39	37	11	12.77	51	0	0.22|
24	12	12.87	38	10	10.99	52	0	0.13|
25	13	15.22	39	6	9.29	53	0	0.08|
26	13	17.31	40	11	7.70	54	0	0.04|
27	22	19.02	41	4	6.27	55	1	0.04|
28	23	20.28	42	7	5.00			
29	17	21.04	43	2	3.90			
$K = 16.7815$ $M = 5.4610$ $n = 47$ $X^2 = 32.29$ $FG = 21$ $P = 0.05$								

Das Ergebnis ist mit $P = 0.05$ zufriedenstellend.

Tabelle 3
Die Verteilung von <f> in Textblöcken

x	n_x	NP_x	x	n_x	NP_x	x	n_x	NP_x
0	57	54.50	4	38	37.72	8	8	7.95
1	53	61.21	5	26	28.19	9	4	4.34
2	72	56.52	6	16	19.86	10	5	2.93
3	49	47.68	7	6	13.09			
$K = 6.1844$ $M = 1.4705$ $n = 12$ $X^2 = 11.73$ $FG = 7$ $P = 0.11$								

Da $P > 0.05$, erweist sich die negative hypergeometrische Verteilung als ein gutes Modell für das Vorkommen von <f>.

Tabelle 4
Die Verteilung von <i> in Textblöcken

x	n_x	NP_x	x	n_x	NP_x	x	n_x	NP_x
3	2	0.27|	12	45	37.30	21	1	1.74
4	2	1.43|	13	45	34.84	22	0	0.83|
5	4	4.15	14	28	30.42	23	0	0.35|
6	10	8.78	15	23	24.89	24	0	0.13|
7	13	15.08	16	21	19.08	25	1	0.04|
8	18	22.23	17	10	13.69	26	0	0.01|
9	25	29.04	18	4	9.17	27	1	0.00|

10	30	34.31	19	8	5.72			
11	42	37.19	20	1	3.29			
$K = 21.9339$ $M = 7.9318$ $n = 25$ $X^2 = 18.18$ $FG = 15$ $P = 0.25$								

Tabelle 5
Die Verteilung von <u> in Textblöcken

x	n_x	NP_x	x	n_x	NP_x	x	n_x	NP_x
1	7	6.70	8	35	32.81	15	3	4.41
2	14	16.16	9	28	28.36	16	4	2.68
3	19	25.29	10	23	23.40	17	4	1.51
4	39	32.27	11	13	18.45	18	1	0.78|
5	41	36.35	12	5	13.90	19	1	0.57|
6	48	37.48	13	4	9.99			
7	38	36.08	14	7	6.82			
$K = 11.2847$ $M = 3.1879$ $n = 22$ $X^2 = 23.48$ $FG = 14$ $P = 0.05$								

Tabelle 6
Die Verteilung von <ü> in Textblöcken

x	n_x	NP_x	x	n_x	NP_x	x	n_x	NP_x
0	8	10.46	6	32	33.59	12	1	0.19|
1	34	30.92	7	21	20.50	13	0	0.04|
2	53	51.05	8	8	10.95	14	0	0.01|
3	63	61.32	9	5	5.10	15	1	0.00|
4	51	59.14	10	2	2.04|			
5	55	48.00	11	0	0.69|			
$K = 24.1332$ $M = 6.2766$ $n = 15$ $X^2 = 4.39$ $FG = 7$ $P = 0.73$								

Man kann damit feststellen, dass die negative hypergeometrische Verteilung sich auch bei diesem niederdeutschen Text als geeignetes Modell für das Auftreten von Buchstaben in Textblöcken erweist.

2.2. Laute

Die Verteilung von drei Lauten [b], [ě] und [œ] in 200 Textblöcken mit je 100 Lauten findet man bei Zwirner & Zwirner (1935 und 1938), wobei [b] in zwei verschiedenen Textblöcken erhoben wurde. Die Autoren führen eine Anpassung der Poisson-Verteilung an die insgesamt vier Lautdateien durch; es fehlt aber in allen Fällen das Testkriterium *P*, so dass man lediglich einen „optischen“ Vergleich zwischen den beobachteten und den berechneten Werten durchführen kann. Zwirner & Zwirner (1938) und Zweirner & Ezawa (1969: 73) stellen fest, dass bei allen Lauten gleichartige Ergebnisse erzielt wurden und betonen: „Damit aber ist ein Gesetz gefunden, das die Verteilung der Sprachlaute in der deutschen Hochsprache regelt.“

Die folgenden Tabellen geben die Anpassung der Poisson-Verteilung

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

an die vier Lautdateien wieder und folgen insofern Zwirner & Zwirner (1935, 1938). Die von mir berechneten Werte unterscheiden sich nur minimal von denen der beiden Autoren. Die negative hypergeometrische Verteilung erwies sich als weniger geeignet, da die Tabelle für [œ] nur drei Klassen aufweist und daher eine Verteilung mit drei Parametern nicht angepasst werden kann. Bei der Datei für [ě] kann die negative hypergeometrische Verteilung nur mit einem nicht ganz zufriedenstellenden Ergebnis angewendet werden. Die Anpassung der Poisson-Verteilung an die vier Lautdateien ergab:

Tabelle 7
Die Verteilung der Laute [b] und [œ] in Textblöcken

	[b]		[b]		[œ]	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	22	30.52	22	30.35	180	178.51
1	62	57.38	60	57.22	18	20.29
2	60	53.93	65	53.95	2	1.20
3	38	33.79	35	33.91		
4	13	15.88	11	15.99		
5	5	8.49	6	6.03		
6			1	2.56		
	$a = 1.8799$ $X^2 = 5.92$ $FG = 4$ $P = 0.21$		$a = 1.8856$ $X^2 = 7.23$ $FG = 5$ $P = 0.20$		$a = 0.1137$ $X^2 = 0.81$ $FG = 1$ $P = 0.37$	

a: Parameter der Poisson-Verteilung.

Tabelle 8
Die Verteilung von [ě] in Textblöcken

x	n_x	NP_x	x	n_x	NP_x	x	n_x	NP_x
0	3	3.06	4	52	38.92	8	8	7.07
1	10	12.79	5	25	32.53	9	6	3.28
2	31	26.74	6	19	22.66	10	1	1.37
3	34	37.25	7	10	13.53	11	1	0.79
$a = 4.1796$ $X^2 = 11.61$ $FG = 9$ $P = 0.24$								

Die Ergebnisse sind in allen vier Fällen zufriedenstellend. Tests zu anderen Lauten können nicht durchgeführt werden, da Zwirner & Zwirner keine weiteren Daten mitteilen.

Es sei noch angegeben, was die Anpassung der negativen hypergeometrischen Verteilung ergibt. Diese Verteilung ist in Altmann (1988: 177) als Grundmodell vorgeschlagen worden. Die Datei für [œ] hat nur drei Häufigkeitsklassen; an sie kann die negative hypergeometrische Verteilung, wie schon erwähnt, nicht angepasst werden, da sie drei Parameter aufweist. An die beiden [b]-Dateien kann man Anpassungen mit $P = 0.22$ bzw. $P = 0.10$ durchführen, also beides gute Ergebnisse. Bei [ě] ergibt sich $P = 0.0163$, ein noch akzeptables, aber nicht gutes Ergebnis, das sich auch nicht nennenswert verbessern lässt: Eine Zusammenfassung der Häufigkeitsklassen $x = 4$ und $x = 5$ erbringt lediglich ein $P = 0.0263$.

Altmann (1988: 177) führt aus, dass die negative hypergeometrische Verteilung als einen ihrer Grenzfälle die Poisson-Verteilung hat, wenn die Parameter K und M sehr groß sind; diese Bedingung ist bei den Dateien für [u] und für [ě] nicht erfüllt. Hier bleibt ein ungelöstes Problem, das man mangels weiterer Daten einstweilen auch nicht angehen kann.

2.3. Wörter

Das Textblock-Gesetz ist bisher am besten auf der Ebene der Wörter erprobt. Die Ergebnisse sind jedoch noch bei weitem nicht so reichhaltig, dass es überflüssig wäre, weitere Befunde zu erarbeiten. In diesem Abschnitt werden einige weitere Ergebnisse zusammengetragen, die teils noch nicht daraufhin getestet wurden, ob sie der negativen hypergeometrischen Verteilung folgen, teils auch noch gar nicht veröffentlicht sind.

In Tabelle 9 folgen zunächst zwei Erhebungen zu einzelnen Wörtern: als erste die Verteilung von „ab“ in einem deutschen Zeitungskorpus (*Neues Deutschland*, Februar 1964, 20 Textblöcke mit je 3000 token), die Billmeier (1968: 149) mitteilt; als zweite die Untersuchung von frz. „et“ nach Muller (1972: 100) in 55 Textabschnitten zu je 30 Versen in Racines *Phèdre*:

Tabelle 9
Die Verteilung von „ab“ und „et“ in Textblöcken

	ab		et	
x	n_x	NP_x	n_x	NP_x
0	7	6.72	1	0.42|
1	5	5.16	1	1.54|
2	4	3.63	2	3.35
3	2	2.34	6	5.52
4	1	1.33|	7	7.54
5	0	0.62|	11	8.87
6	1	0.19|	11	9.08
7			7	8.05
8			3	5.98
9			4	3.44
10			2	1.20
	$K = 4.4273$ $n = 6$ $FG = 1$	$M = 1.0699$ $X^2 = 0.11$ $P = 0.74$	$K = 8.6858$ $n = 10$ $FG = 6$	$M = 4.7535$ $X^2 = 3.79$ $P = 0.71$

In beiden Fällen erweist sich die negative hypergeometrische Verteilung als ein sehr gutes Modell.

Muller (1972: 269) gibt auch noch das Auftreten von „vous“ in 17 Textblöcken mit je 100 Versen in Racines *Phèdre* wieder; da die Häufigkeit des Wortes in jedem Textabschnitt eine andere ist, ist ein Test nicht angezeigt.

Weitere Verteilungen von Einzelwörtern in Textblöcken hat Suhren (2002) für den niederdeutschen *De lütte Prinz* erhoben. Nach Bildung von Wortblöcken mit 50 bzw. 100 Wörtern wurde untersucht, wie oft die Wörter „bi“, „da“, „he“, „ik“, „is“, „mi“, „mit“, „ni“, „un“, „Prinz“ und „Schoop“ vertreten waren. Auch Suhren orientierte sich an den Vorschlägen von Altmann & Burdinski (1982), musste aber feststellen, dass dies nur teilweise von Erfolg gekrönt war. Besonders bei den Vorkommen der Wörter in 50-Wort-Blöcken gab es Schwierigkeiten, die vorgeschlagenen Verteilungen anzupassen. Stattdessen konnten mit anderen Modellen sehr gute Ergebnisse erzielt werden. Die Verwendung anderer Verteilungen kann man damit rechtfertigen, dass Wimmer, Köhler & Altmann (2005) Sprachgesetze aus einem wesentlich allgemeineren Ansatz ableiten, von denen einige auch hier erfolgreich waren.

Im Folgenden werden die 11 Wörter nur mit ihren Vorkommen in 100-Wort-Blöcken berücksichtigt; dabei kann abweichend von Suhren (2002) die negative hypergeometrische Verteilung in allen Fällen angepasst werden.

Tabelle 10
Die Verteilung von „bi“, „du“ und „he“ in Textblöcken

	bi		du		he	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	105	103.85	70	69.44	50	47.56
1	48	48.53	44	44.46	32	39.63
2	12	12.52	25	26.20	40	29.95
3	0	1.95|	18	14.37	16	21.14
4	2	0.14|	4	7.28	13	13.86
5			5	3.34	9	8.27
6			0	1.34|	5	4.31
7			0	0.45|	1	1.80|
8			0	0.11|	1	0.47|
9			1	0.02|		
	$K = 20.7153$ $n = 4$	$M = 2.4809$ $C = 0.0003$	$K = 8.1356$ $n = 9$ $FG = 3$	$M = 1.0718$ $X^2 = 3.73$ $P = 0.29$	$K = 5.0139$ $n = 8$ $FG = 4$	$M = 1.1334$ $X^2 = 6.47$ $P = 0.17$

Da die Anpassung der negativen hypergeometrischen Verteilung bei „bi“ nur 0 Freiheitsgrade ergibt, muss statt des Kriteriums P der Diskrepanzkoeffizient $C = X^2/N$ verwendet werden, der bei $C \leq 0.01$ eine sehr gute Übereinstimmung zwischen den beobachteten und den berechneten Werten signalisiert.

Tabelle 11
Die Verteilung von „ik“, „is“ und „mi“ in Textblöcken

	ik		is		mi	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	22	21.29	78	77.12	49	51.81
1	28	28.13	49	51.84	52	45.40
2	27	28.41	27	25.39	32	33.49
3	23	25.34	12	9.63	17	21.38
4	25	20.77	0	2.63|	12	11.13
5	17	15.86	1	0.40|	5	3.79
6	10	11.29				
7	5	7.45				
8	7	4.49				
9	2	2.41				
10	0	1.10|				
11	0	0.39|				
12	1	0.08|				
	$K = 6.8691$ $n = 12$ $FG = 7$	$M = 1.7721$ $X^2 = 3.88$ $P = 0.79$	$K = 8.7902$ $n = 5$ $FG = 1$	$M = 1.5159$ $X^2 = 2.21$ $P = 0.14$	$K = 4.3130$ $n = 5$ $FG = 2$	$M = 1.2397$ $X^2 = 2.53$ $P = 0.28$

Tabelle 12
Die Verteilung von „mit“, „ni“ und „un“ in Textblöcken

	mit		ni		un	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	91	86.23	53	55.48	13	14.35
1	45	48.37	58	53.17	38	34.99
2	26	21.54	35	34.44	45	43.13
3	2	7.95	11	16.84	39	35.50
4	3	2.91	9	5.90	18	21.74
5			1	1.16	7	10.46
6					3	4.07
7					2	1.30|
8					1	0.44|
	$K = 10.2206$ $n = 6$ $FG = 1$	$M = 1.3013$ $X^2 = 5.88$ $P = 0.0153$	$K = 7.9379$ $n = 5$ $FG = 2$	$M = 1.9203$ $X^2 = 4.23$ $P = 0.12$	$K = 86.8722$ $n = 15$ $FG = 4$	$M = 14.1023$ $X^2 = 3.80$ $P = 0.43$

Das Ergebnis bei „mit“ ist gerade noch akzeptabel; es lässt sich auch nicht verbessern. Eine Anpassung der Poisson-Verteilung gelingt in diesem Fall mit $P = 0.07$.

Nach den grammatischen Wörtern folgen noch die Verteilungen von zwei Lexemen.

Tabelle 13
Die Verteilung von „Prinz“ und „Schoop“ in Textblöcken

	Prinz		Schoop	
x	n_x	NP_x	n_x	NP_x
0	74	73.88	147	146.99
1	54	54.49	10	10.84
2	28	27.24	5	4.77
3	9	9.53	4	2.52
4	2	1.86	1	1.32|
5			0	0.55|
	$K = 7.6056$ $n = 4$ $FG = 1$	$M = 1.6509$ $X^2 = 0.07$ $P = 0.80$	$K = 2.0345$ $n = 5$ $FG = 1$	$M = 0.0877$ $X^2 = 1.35$ $P = 0.24$

Bei „Schoop“ wurde eine leere Klasse hinzugefügt, um die Anpassung durchführen zu können.

Zu den niederdeutschen Wörtern kann also festgestellt werden, dass die negative hypergeometrische Verteilung in allen 11 Fällen angepasst werden kann, wobei das Ergebnis nur bei „mit“ nicht ganz zufrieden stellt.

2.4. Wortarten

In einem Fall wurde gezeigt, dass nicht nur Wörter, sondern womöglich auch Wortarten in Textblöcken gesetzmäßig auftreten. Dies haben Piotrowski, Bektaev & Piotrowskaja (1985:

217) für das Vorkommen von Substantiven in stichprobenartig erhobenen Textblöcken nachgewiesen (vgl. dazu auch Altmann 1988: 184).

2.5. Bedeutungsgruppen grammatischer Wörter

Muller (1972: 97f.; 101) untersucht in zwei Fällen das Vorkommen von bedeutungsähnlichen grammatischen Wörtern: „ne, n‘, ni, non“ (66 Textblöcke mit je 25 Versen) und „tu, te, t‘, ton, ta, tes“ (55 Textblöcke mit je 30 Versen) in Racines *Phèdre*. Die Ergebnisse dazu:

Tabelle 14
Die Verteilung der Negationen „ne, n‘, ni, non“ in Textblöcken

x	n_x	NP_x	x	n_x	NP_x	x	n_x	NP_x
0	1	1.07	4	17	13.17	8	2	1.96
1	3	4.47	5	10	10.68	9	2	1.30
2	11	9.33	6	5	7.13			
3	12	12.86	7	3	4.03			
$K = 464.5046$ $M = 52.1731$ $n = 36$ $X^2 = 3.28$ $FG = 6$ $P = 0.77$								

Tabelle 15
Die Verteilung von „tu, te, t‘, ton, ta, tes“ (2.Ps.Sg.) in Textblöcken

x	n_x	NP_x	x	n_x	NP_x	x	n_x	NP_x
0	21	23.47	8	0	1.31	16	0	0.62|
1	8	6.22	9	1	1.18	17	0	0.57|
2	3	3.89	10	1	1.07	18	3	0.52|
3	7	2.89	11	0	0.97	19	0	0.46|
4	2	2.32	12	1	0.89	20	0	0.41|
5	0	1.95	13	0	0.81	21	0	0.36|
6	4	1.68	14	1	0.74	22	1	0.51|
7	2	1.48	15	0	0.68|			
$K = 1.6116$ $M = 0.2690$ $n = 23$ $X^2 = 16.84$ $FG = 12$ $P = 0.16$								

Dieses Ergebnis deutet darauf hin, dass man auch semantische Gruppen von Wörtern, zumindest bei den grammatischen Wörtern, bilden kann; auch diese treten in Textblöcken gesetzmäßig auf.

2.6. Wortgruppen

Auch Wortgruppen sind darauf hin untersucht worden, wie sie sich in Textblöcken verhalten. In Piotrowski (1984: 143) und Piotrowski, Bektaev & Piotrowskaja (1985: 252) wird gezeigt, dass in deutschen publizistischen Texten in „100 Serien von jeweils 1000 Dreiwort-Segmenten“ die Wortfolge „Δ daß der“ gemäß der Poisson-Verteilung erscheint; nach meiner Berechnung mit $P = 0.25$. (An diese Datei lässt sich die negative hypergeometrische Verteilung mit $P = 0.0298$ nur mit deutlich schlechterem Ergebnis anpassen.)

Köhler (2001: 136) wählt einen anderen Weg. Er bestimmt nicht wie Piotrowski et al. das Vorkommen einer festgelegten Wortgruppe, sondern vielmehr „the occurrence of categories

should be observed". Er untersucht zwei verschiedene Typen von Wortgruppen: Teilsatztypen („clause types" wie Relativ-, Infinitiv- und Partizipialsätze) sowie Funktionstypen (direkte, indirekte und Präpositionalobjekte). Als Ergebnis teilt er mit, dass bei 7 Auswertungen die negative Binomialverteilung immer gute Resultate liefert, mit einer Ausnahme bessere als die negative hypergeometrische Verteilung.

3. Ergebnis

Ziel dieser Zusammenstellung war es, zu zeigen, dass das Textblock-Gesetz in seinen Anfängen mindestens bis in die 30er Jahre des zwanzigsten Jahrhunderts zurückreicht. Es bestätigt sich die mehrfach geäußerte Vermutung (Altmann & Burdinski 1982; Köhler 2001), dass es sich um ein Sprachgesetz handelt, das eine größere Vielfalt von Einheiten verschiedener Sprachebenen zwischen Lauten und Syntagmen steuert. Auch wenn z.B. die Verteilung von Silben und Morphen - und weiterer Einheiten - anscheinend noch nicht untersucht wurde, ist aufgrund der bisherigen Arbeiten zu erwarten, dass sie sich im Prinzip nicht anders verhalten als die bisher bekannten.

Probleme könnten darin bestehen, dass womöglich die Größe der Textblöcke je nach untersuchter Einheit verschieden bestimmt werden muss; vereinzelt deutet sich an, dass Textblöcke mit nur 50 Einheiten der betreffenden Kategorie zu klein sein könnten.

Die bisherigen Ergebnisse wurden überwiegend an Einzeltexten gewonnen; einige Befunde anhand von Textkorpora deuten jedoch an, dass die Gesetzmäßigkeit des Auftretens von Einheiten in Textblöcken auch dort nachzuweisen sein wird. Vorbehalte sind lediglich aufgrund der Tatsache zu äußern, dass die Zahl der Überprüfungen dieses Gesetzes noch nicht allzu groß ist und auch nur relativ wenige Sprachen dabei berücksichtigt wurden.

Literatur

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

Altmann, G., Burdinski, V. (1982). Towards a Law of Word Repetitions in Text Blocks. In: Lehfeldt, W., & Strauss, U. (Hrsg.), *Glottometrika 4, 147-167*. Bochum: Brockmeyer.

Best, K.-H. ([2]2003). *Quantitative Linguistik. Eine Annäherung.* 2., überarbeitete und erweiterte Auflage. Göttingen: Peust & Gutschmidt.

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

Brainerd, B. (1972). Article use as an indicator of style among English-language authors. In: Jäger, Siegfried (Hrsg.), *Linguistik und Statistik: 11-32*. Braunschweig: Vieweg.

Frumkina, R.M. (1962). O zakonach raspredelenija slov i klassov slov. In: Mološnaja, T.T. (Hrsg.), *Strukturno-tipologičeskie issledovanija: 124-133.* Moskva, AN SSSR.

Köhler, R. (2001). The distribution of some syntactic construction types in text blocks. In: Uhlířova, L., Wimmer, G., Altmann, G., Köhler, R. (Eds.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček: 136-148.* Trier: WVT.

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

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

Piotrowski, R. G. (1984; russ. 1975). *Text, Computer, Mensch.* Bochum: Brockmeyer.

Piotrowski, R.G., Bektaev, K.B., Piotrowskaja, A.A. (1985; russ. 1977). *Mathematische Linguistik.* Bochum: Brockmeyer.

Schulte, E. (2002). *Das Frumkina-Gesetz.* Referat, Göttingen.

Suhren, S. (2002). *Untersuchung zum Gesetz von Zwirner, Zwirner und Frumkina am Beispiel des niederdeutschen „De lütte Prinz“*. Staatsexamensarbeit, Göttingen.

Wimmer, G., Köhler, R., Altmann, G. (2005). Unified Derivation of Some Linguistic Laws. In: Altmann, G., Köhler, R., Piotrowski, R. (Hrsg.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch*. Berlin/ N.Y.: de Gruyter (erscheint vorauss. 2005)

Zwirner, E. (1967). Sprachen und Sprache: Ein Beitrag zur Theorie der Linguistik. In: *To Honor Roman Jakobson. Essays on the Occasion of his Seventieth Birthday, 11. Oct. 1966: 2442-2464.* The Hague/ Paris: Mouton.

Zwirner, E., Ezawa, K. (Hrsg.) (1966, 1968, 1969). *Phonometrie, Erster-Dritter Teil.* Basel/ New York: Karger.

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

Zwirner, E., Zwirner, K. (1938). Lauthäufigkeit und Sprachvergleichung. *Monatsschrift für höhere Schulen 37: 246-253.* (Auch in: Zwirner & Ezawa (Hrsg.), Dritter Teil, 68-74.)

Software

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

Adresse des „Göttinger Projekts“ im Internet (mit ausführlicher Bibliographie): http://gwdu05.gwdg.de/~kbest

Glottometrics 9, 2005, 13-20

Quantitative Untersuchungen zur Valenz deutscher Verben

Reinhard Köhler, Universität Trier

Abstract. Quantitative properties of German verb valency are investigated taking into account the distribution of syntactic-semantic variants of the verbs, the distribution of sentence patterns, the distribution of the number of alternative semantic subcategories that an actant can take, and the functional dependency between the number of potential actants and the number of alternative semantic subcategories per variant.

Keywords: v*alency, German verbs, diversification, functional dependences*

Einleitung

Die quantitativen Analysen zur Syntax in (Köhler 1999 und Köhler & Altmann 2000) verwendeten einen phrasenstrukturgrammatischen Ansatz. Die vorliegende Studie soll dagegen auf einer Dependenzgrammatik-nahen Grundlage durchgeführt werden und hat Valenz und distributionelle Eigenschaften deutscher Verben als Gegenstand. Das benutzte Material stammt aus den Beschreibungen deutscher Verben in (Helbig/Schenkel [8]1991). Das Hauptziel der Studie ist es zu zeigen, dass die Möglichkeit der Entdeckung quantitativer Gesetzmäßigkeiten in der Syntax nicht auf ein bestimmtes Grammatikformat beschränkt ist, wenn auch die Verschiedenartigkeit der Begriffssysteme einen unmittelbaren Vergleich von Ergebnissen nicht immer zulässt.

Verteilung der Anzahl der Varianten von Verben

Varianten eines Verbs unterscheiden sich in ihrer Valenz und in ihrer Bedeutung. Als Beispiel soll hier das Verb *achten* mit seinen Varianten dienen:

Variante 1: („hochschätzen“)
Subjekt im Nominativ, Objekt im Akkusativ (jemand achtet jemanden);
Variante 2: („aufpassen“)
2.1. Subst. im Nom., Präposition mit Objekt im Akkusativ (jmd. achtet auf etwas),
2.2. Subst. im Nom., Nebens. mit *dass*, *ob* oder *wer/was* (jmd. achtet darauf, wer handelt),
2.3. Subst. im Nom., Infinitiv (jemand achtet darauf, etwas zu tun).

Es liegt nahe, die Existenz von mehr oder weniger Varianten als das Resultat eines Diversifikationsprozesses zu betrachten: Man kann sich vorstellen, dass ein neu eingeführtes Verb zunächst keine Varianten besitzt. Nach einer Zeit der Verwendung entsteht eine Wahrscheinlichkeit dafür, dass das Verb durch wiederholte Verwendung mit abweichender Bedeutung (für die die Argumentstruktur angepasst werden muss) eine Variante entwickelt. Diese

Wahrscheinlichkeit sollte abhängig von der Häufigkeit der Verwendung des Verbs sein. Ebenso entsteht eine Wahrscheinlichkeit für die Entstehung einer weiteren Variante, die in Abhängigkeit von den Wahrscheinlichkeiten der bereits existierenden Varianten steht. Der Einfachheit halber beschränkt man sich bei der Modellierung auf die Wahrscheinlichkeiten der jeweils benachbarten Klassen (vgl. Altmann 1991). Aus dem Modell

$$P_x = \frac{a + b\,x}{x}\, P_{x-1}$$

d.h. aus der Annahme, dass sich die Wahrscheinlichkeit einer Klasse x über die angegebene lineare Funktion aus der Wahrscheinlichkeit der Nachbarklasse x-1 berechnet, ergibt sich nach Reparametrisierung mit $a/b = k$-1 und $b = q$ die negative Binomialverteilung

$$P_x = \binom{k+x-1}{x} p^k q^x, \quad x = 0, 1, \ldots$$

Da jedes Verb wenigstens eine Lesart besitzt, ist $P_0 = 0$, und x läuft ab 1, d.h. wir gehen von der positiven negativen Binomialverteilung

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

als Hypothese für die Verteilung der Variantenzahl der Verben aus. Abb. 1 und Tabelle 1 zeigen das Ergebnis der Anpassung dieser Verteilung an die Daten. Da der Gütetest eine hervorragende Übereinstimmung von Beobachtung und erwarteten Werten anzeigt, kann die Hypothese zumindest vorläufig beibehalten werden.

Tabelle 1
Anpassung der positiven negativen Binomialverteilung an die Daten

***X*[*i*]**	***F*[*i*]**	***NP*[*i*]**
1	218	214.62
2	118	126.40
3	73	69.49
4	42	36.84
5	18	19.10
6	8	9.75
7	4	4.92
8	2	2.47
9	2	1.23
10	1	1.19
k = 1.4992 , p = 0.5287		
X^2 = 2.67, DF = 7, P(X^2) = 0.91		

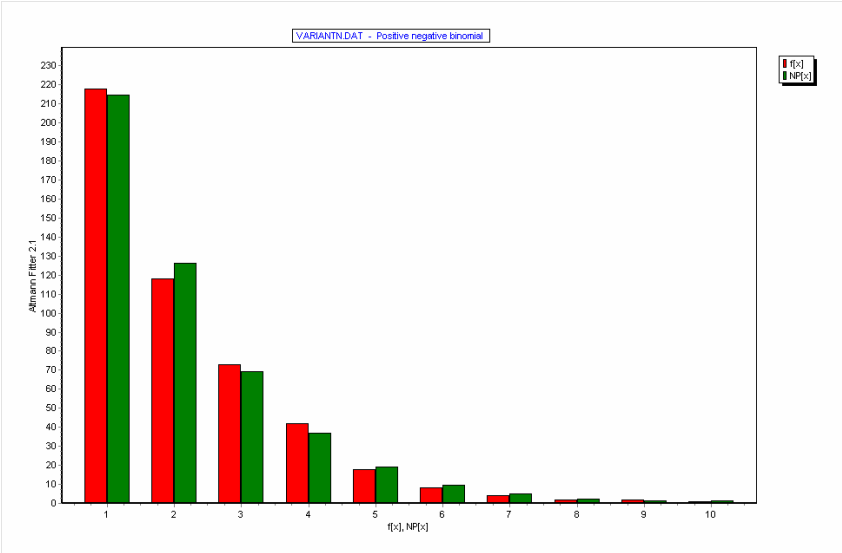


Abb. 1. Verteilung der Variantenzahl. Graph der Anpassung der positiven negativen Binomialverteilung

Verteilung der Anzahl der Satzbaupläne

Wir wollen hier unter einem Satzbauplan das Muster aus obligatorischen, fakultativen und alternativen Aktanten verstehen, das sämtliche Möglichkeiten beschreibt, einen Satz mit einer Verbvariante zu konstruieren. Besteht z.B. bei einer Verbvariante einzig die Möglichkeit, einen Satz mit einem nominativischen Subjekt und einem Akkusativ-Objekt zu bilden, so notieren wir das als *SnSa*. Kann alternativ zu dem Akkusativ-Objekt ein Nebensatz mit *dass* gewählt werden, so schreiben wir *SnSa/NS_dass*. Ist ein Aktant fakultativ, notieren wir ihn in Klammern: *SnSa(Sd)*. Entsprechend beschreibt *Sn(pSa/Adj)Part/Inf* Verben, die ein Subjekt verlangen, außerdem entweder ein Präpositionalobjekt im Akkusativ oder ein Adjektiv haben können und dazu noch entweder ein Partizip oder einen Infinitiv verlangen.

In dem untersuchten Material kommen 205 verschiedene Satzbaupläne vor, dessen häufigster (*SnSa*) 286 Verbvarianten beschreibt. Das zweithäufigste Muster ist *SnSapS* mit 78 Vertretern. Mit 73 Vorkommen ist *SnpS* nicht viel seltener, gefolgt von rein intransitiven Verben mit lediglich einem Subjekt (*Sn*). Die Häufigkeitsverteilung lässt sich mit Hilfe der Zipf-Mandelbrot-Verteilung extrem gut modellieren: mit einer Chi-Quadrat-Wahrscheinlichkeit, die von 1.0 nicht zu unterscheiden ist. Das Ergebnis der Anpassung ist in Tabelle 2 und Abb. 2 dargestellt.

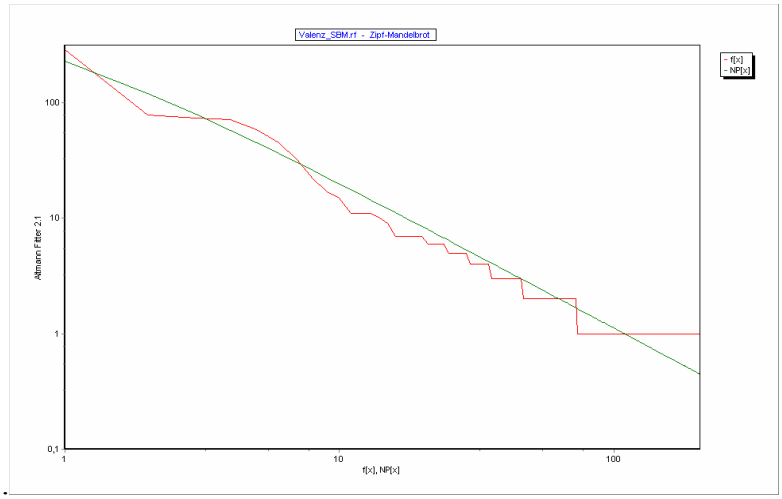


Abb.2. Anpassung der Zipf-Mandelbrot-Verteilung an die Häufigkeiten der Satzbaupläne

Tabelle 2
Beobachtete und erwartete Häufigkeiten von Satzbauplänen (Zipf-Mandelbrot-Verteilung)

x	***F*(*x*)**	***NP*(*x*)**	***x***	***F*(*x*)**	***NP*(*x*)**	***x***	***F*(*x*)**	***NP*(*x*)**
1	286	226.92	71	2	1.72	141	1	0.72
2	78	120.32	72	2	1.69	142	1	0.72
3	73	78.94	73	2	1.67	143	1	0.71
4	71	57.55	74	1	1.64	144	1	0.70
5	58	44.68	75	1	1.61	145	1	0.70
6	45	36.19	76	1	1.58	146	1	0.69
7	32	30.20	77	1	1.56	147	1	0.69
8	22	25.77	78	1	1.53	148	1	0.68
9	17	22.39	79	1	1.51	149	1	0.67
10	15	19.72	80	1	1.48	150	1	0.67
11	11	17.57	81	1	1.46	151	1	0.66
12	11	15.81	82	1	1.44	152	1	0.66
13	11	14.34	83	1	1.42	153	1	0.65
14	10	13.10	84	1	1.39	154	1	0.65
15	9	12.04	85	1	1.37	155	1	0.64
16	7	11.12	86	1	1.35	156	1	0.64
17	7	10.32	87	1	1.33	157	1	0.63
18	7	9.61	88	1	1.31	158	1	0.63
19	7	8.99	89	1	1.30	159	1	0.62
20	7	8.44	90	1	1.28	160	1	0.62
21	6	7.94	91	1	1.30	161	1	0.61
22	6	7.50	92	1	1.24	162	1	0.61
23	6	7.10	93	1	1.23	163	1	0.60
24	6	6.73	94	1	1.21	164	1	0.60
25	5	6.40	95	1	1.19	165	1	0.59
26	5	6.09	96	1	1.18	166	1	0.59
27	5	5.81	97	1	1.16	167	1	0.59
28	5	5.55	98	1	1.15	168	1	0.58
29	5	5.31	99	1	1.13	169	1	0.57
30	4	5.09	100	1	1.12	170	1	0.57
31	4	4.89	101	1	1.10	171	1	0.57
32	4	4.70	102	1	1.09	172	1	0.56
33	4	4.52	103	1	1.08	173	1	0.56
34	4	4.36	104	1	1.06	174	1	0.55
35	4	4.20	105	1	1.05	175	1	0.55
36	3	4.06	106	1	1.04	176	1	0.55
37	3	3.92	107	1	1.03	177	1	0.54
38	3	3.79	108	1	1.01	178	1	0.54
39	3	3.67	109	1	1.00	179	1	0.53
40	3	3.55	110	1	0.99	180	1	0.53
41	3	3.44	111	1	0.98	181	1	0.53
42	3	3.34	112	1	0.97	182	1	0.52
43	3	3.24	113	1	0.96	183	1	0.52
44	3	3.15	114	1	0.95	184	1	0.52
45	3	3.06	115	1	0.94	185	1	0.51

46	3	2.98	116	1	0.93	186	1	0.51
47	2	2.90	117	1	0.92	187	1	0.51
48	2	2.82	118	1	0.91	188	1	0.50
49	2	2.75	119	1	0.90	189	1	0.50
50	2	2.68	120	1	0.89	190	1	0.50
51	2	2.62	121	1	0.88	191	1	0.49
52	2	2.55	122	1	0.87	192	1	0.49
53	2	2.49	123	1	0.86	193	1	0.49
54	2	2.44	124	1	0.85	194	1	0.48
55	2	2.38	125	1	0.84	195	1	0.48
56	2	2.33	126	1	0.83	196	1	0.48
57	2	2.27	127	1	0.83	197	1	0.47
58	2	2.23	128	1	0.82	198	1	0.47
59	2	2.12	129	1	0.81	199	1	0.47
60	2	2.13	130	1	0.80	200	1	0.47
61	2	2.09	131	1	0.79	201	1	0.46
62	2	2.05	132	1	0.79	202	1	0.46
63	2	2.01	133	1	0.78	203	1	0.46
64	2	1.97	134	1	0.77	204	1	0.45
65	2	1.93	135	1	0.76	205	1	0.45
66	2	1.89	136	1	0.76			
67	2	1.86	137	1	0.75			
68	2	1.82	138	1	0.74			
69	2	1.79	139	1	0.74			
70	2	1.76	140	1	0.73			
a = 1.2730, b = 0.5478, n = 205								
X^2 = 86.49, DF =150, P(X^2) ≈ 1.00								

Verteilung der Selektionsbeschränkungen

Das verwendete Valenzwörterbuch gibt außer Zahl und Art der Aktanten für jeden Aktanten an, welche Selektionsbeschränkungen – beschrieben anhand von semantischen Subkategorisierungen – bestehen. Die wichtigsten benutzten Kategorien sind *Abstr* (Abstraktbezeichnung), *Abstr (als Hum)* (Kollektivbegriff), *Act* (Handlung), *+Anim* (belebtes Wesen) – ggfls. ergänzt durch *–Hum* (Menschen ausgenommen), *–Anim* (Unbelebtes), *Hum* (menschliches Wesen), *–Ind* (Individualbezeichnung ausgenommen).

Die vollständige Beschreibung der ersten Variante des Verbs *achten* hat die folgende Form:

I. achten 2 (V1 = hochschätzen)
II. achten → Sn, Sa

III. Sn →	1. Hum (Die Schüler achten den Lehrer.)
	2. Abstr(als Hum) (Die Universität achtet den Forscher.)
Sa →	1. Hum (Wir achten den Lehrer.)
	2. Abstr(als Hum) (Wir achten die Regierung.)
	3. Abstr (Wir achten seine Meinung.

Achten im angegebenen Sinn kann also für den ersten Aktanten, das Subjekt, entweder als menschlich gekennzeichnete Elemente oder Kollektivbegriffe für Institutionen akzeptieren – weist somit zwei Möglichkeiten auf – für den zweiten, das direkte Objekt, kommen drei verschiedene Klassen in Frage. Die Variante 1 des Verbs *achten* trägt daher für den ersten Aktanten mit zwei, für den zweiten Aktanten mit drei Alternativen zur Zählung bei.

Wir gehen davon aus, dass auch die Zahl solcher Alternativen in der Valenzbeschreibung von Wörtern einem Sprachgesetz genügt, und haben die Verteilung der alternativen Selektionsbeschränkungen je Aktant bestimmt. Die so erhaltenen Daten stimmen gut mit der einfachen Hypothese überein, dass die Diversifikation der Selektionsbeschränkungen in Form einer Konstanten gefördert und invers proportional zur Zahl der bereits bestehenden Alternativmöglichkeiten gebremst wird. Der entsprechende Ansatz

$$P_x = \frac{\lambda}{x} P_{x-1}$$

führt zu der Poisson-Verteilung

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

bzw., da der Definitionsbereich bei 1 beginnt (jeder Aktant hat wenigstens eine Selektionsmöglichkeit; auch die Angabe „keine Selektionsbeschränkungen“ beschreibt ja eine Wahlmöglichkeit), zu der positiven Poisson-Verteilung

$$P_x = \frac{\lambda^x}{x!\ (e^{\lambda} - 1)}, \quad x = 1, 2, 3, \ldots$$

Die Anpassung dieser Verteilung ergab die in Abb. 3 und Tab. 3 dargestellten Resultate.

Funktionale Abhängigkeit der Selektionsbeschränkungen von der Anzahl der Aktanten

Mit der Zahl der (obligatorischen, fakultativen und alternativen) Aktanten steigt selbstverständlich die Zahl der Auswahlalternativen und den Selektionsbeschränkungen. Der empirische Zusammenhang zwischen diesen beiden Variablen kann leicht aus den Valenzbeschreibungen erhoben werden und ist in Tabelle 4 und Abbildung 4 dargestellt. Es zeigt sich, dass die Zahl der Alternativen linear mit der Zahl der überhaupt möglichen Aktanten steigt.

Die Anpassung einer linearen Funktion an die Daten ergab die Regressionsgerade

$$y = 1.59583571x - 0.383300009$$

bei einem Determinationskoeffizienten von $R^2 = 0.9696$.

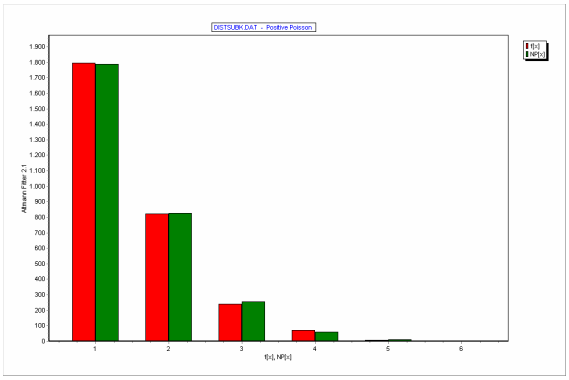


Abb. 3. Anpassung der positiven Poisson-Verteilung (vgl. Tab.3)

Tabelle 3
Beobachtete und erwartete Häufigkeiten von Selektionsbeschränkungen
(positive Poisson-Verteilung)

***X*[*i*]**	***F*[*i*]**	***NP*[*i*]**
1	1796	1786.49
2	821	827.71
3	242	255.66
4	73	59.23
5	9	10.98
6	1	1.95
$\lambda = 0.9266$		
$X^2 = 4.86$, DF = 4, $P(X^2) = 0.30$		

Tabelle 4. Die Abhängigkeit der Zahl der Alternativen von der Zahl der Aktanten

Anz. der Msp.	**Durchschn. Zahl der Alternativen**
1	1.39
2	3.08
3	4.66
4	5.86
5	7.98
6	9.36
7	9.20
8	11.00
9	15.00
11	18.00

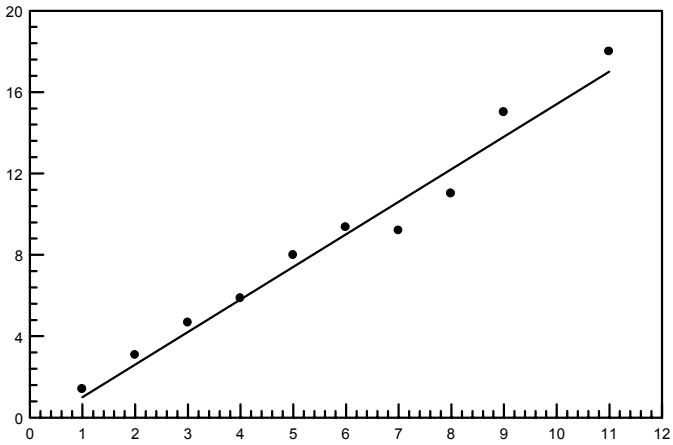


Abb. 4. Anpassung einer Geraden an die Daten in Tab. 4.

Zusammenfassung und Ausblick

Die hier vorgestellten Resultate geben einen ersten Hinweis darauf, dass die Suche nach Sprachgesetzen im Bereich der Syntax nicht an einen speziellen grammatischen Beschreibungsansatz gebunden ist. Nach den erwähnten Studien anhand von Material, das auf einem phrasenstrukturgrammatischen Ansatz basierte, konnten nun im dependenzgrammatischen Rahmen ebenfalls Hypothesen untersucht werden, die als Kandidaten für Sprachgesetze aussichtsreich erscheinen.

Im Anschluss an diesen ersten Versuch sollen weitere empirische Untersuchungen und theoretische Überlegungen zur Interpretation der Modelle und ihrer Parameter folgen.

Literatur

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

Helbig, G., Schenkel, W. ([8]1991). *Wörterbuch zur Valenz und Distribution deutscher Verben*. 8., durchgesehene Auflage. Tübingen: Max Niemeyer Verlag.

Köhler, R. (1999). Syntactic structures. Properties and interrelations. *Journal of Quantitative Linguistics 6, 46-57.*

Köhler, R., Altmann, G. (2000). Probability distributions of syntactic units and properties. *Journal of Quantitative Linguistics 7/3, 189-200.*

Glottometrics 9, 2005, 21-28

Semantic functions of English initial consonant clusters

Nadija L. Lvova[1]

Abstract. This paper is a study of relations between sounds and meanings, namely between English initial consonant clusters and their semantics. Using the chi-square test, statistically significant relations are established between the semantic and phonetic units.

Key words: phonosemantic relations, consonant clusters, phoneme combinations, chi-square test, coefficient of contingency, semantic potential, semantic activity.

Looking at phoneme combinations in speech is relevant to the study of regularities in the functioning of phonological systems, the formation and structure of morphemes, syllables, and words, and other phenomena of this kind. The semantic functions of phoneme combinations have been studied frequently in many languages, and there are no doubts about their existence in sound combinations. When analyzing the sign properties of sounds, linguists have often paid – and still pay – much attention to such phenomena as an initial combination of two (sometimes three) consonants that are used "to mean bound submorphemic strings which have in common a certain element of meaning or function" (Abelin 1999: 4). Like single sounds, consonant clusters are ascribed symbolic meanings. This idea is supported by the research of Magnus (1999), who concludes that words with common consonant sounds, or clusters of consonants repeated in a particular context, make for certain associations. Due to these associations, concordant words are united into a common thematic field or word grouping (Magnus 1999).

The first attempt to describe "meanings" of these sound clusters was made by Wallis (1653). His conclusions were based on a statistical analysis of English vocabulary with the initial combinations *str-, st-, thr-, br-, cr-, gr-,* etc. Wallis reported the following correspondences (Wallis 1653):

STR-	-force, efforts, strain (strong, struggle);
ST-	-less intensive force, which is being lost to preserve what one has (stand, stop);
THR-	- powerful movement, gust (throw);
BR-	- strong breaking, noise (break, brook);
CR-	- breaking with a crack (crack, cry, crash);
GR-	-something rough, hard (grate, grind);
CL-	- adjoin or restrain (cleave, clay, climb);

[1] Address correspondence to N. Lvova at vakhtang@chv.ukrpack.net

SP-	- spreading or disperse (spread, spit);
SL-	- noiseless sliding (slide);
SK-, SKR-	- strong compression

The correspondences between the sound combinations and meanings turn out to be very generalized here. Moreover, they are not always realized in all lexemes. For instance, the consonant cluster *cr-,* according to Wallis, possesses the meaning "*breaking with a crack*", which quite correctly expresses the sound-symbolic nature of this combination. But there are a considerable number of lexemes in English which do not belong to this semantic group: *cream, crime, crowd, create,* etc. Furthermore, Wallis does not address the issue of whether these clusters are individual cases of phonetic symbolism, or whether they create a peculiar system.

Ullman considers the phoneme cluster *gl-* to indicate a figurative idea of light, *sl-* to convey the idea of slowness (Ullmann 1959: 80). He focuses attention on the semantic closeness of English words with the initial cluster *gl-* (*glimmer, glare, glisten, glitter, gleam*) and the initial cluster *sl-* (*sluggish, slack, sloth, sloppy*). It is quite *possible* that in English there is a connection between the notion "*light*" and the cluster *gl-* or between the notion "*slow*" and the cluster *sl-*; but neither Ullman, nor those whose works he refers to, demonstrate this connection to exist. To prove the connection, one would have to ascertain how many words with different meanings that start with *gl-* and *sl-* exist in the English vocabulary – using, of course, an appropriate statistic analysis.

Thus, a successful investigation of the relationship between sounds and meanings in speech will first of all depend on the availability of effective methods of research. In the analysis of phonosemantic relations, two types of lexical groupings may be used: either groups of words united on the basis of synonymic relations, or groups of words united on the basis of coincidence of semantic components (not necessarily synonymy). Judging from cumulative experience, the second type of grouping proves the more effective.

To establish statistically significant relationships between the semantic and phonetic units, all the non-prefixed lexemes with initial two- or three-phoneme clusters were extracted from the English-Ukrainian dictionary (containing 112,000 words and word combinations), together with an interpretation of their semantics in Ukrainian – only the first meaning of all possible ones given for each lexeme being taken into account. To take account of families of words (for example *flex, flexible, flexion, flexor, flexure*), Ukrainian derivatives from the same stem were considered only once. That is, out of two or more variants only one index card was made, for instance, under the heading *flex*. The variant selected for consideration was the first one (in alphabetic order), regardless of its part of speech.

The next step was to group the selected words according to the principle of semantic similarity. In studies of this type, the process of grouping phonetically similar forms that have similar meanings is, as a rule, rather a subjective one. Newman's (1933) conclusions on objective sound symbolism are worth mentioning here. Newman was one of the first to use a complicated mathematical apparatus, but was not eventually successful in solving the problem of correspondences between the outward phonetic form and the semantics of speech units. One reason for this experiment's negative result is errors in distributing the words into categories. Brown (1958: 119) suggests that the results could have been more trustworthy if the participants in the experiment had distributed the words into groups themselves. Basically, however, Newman's procedure is quite sound and has been successfully applied by others (for example Peterfalvi 1970). In a later study, Johnson (1967) attempts to check the "old" experiment by new methods. Having eliminated the defects of Newman's experiment,

Johnson's results are positive: the data about the symbolic meaning of vowels and consonants in English word groups with the meaning "size" was shown to be statistically real.

Starting with raw data consisting of two- or three-phoneme combinations, together with certain semantic units ascribed to the words that begin with these combinations, the data was grouped into the five semantic categories suggested by Peterfalvi (1970). For the purpose of studying the phonetic motivation of words, he sorted the stimuli under investigation into the following 5 semantic categories: audible, perceptible (through senses other than hearing), movement, concrete objects, and abstract notions. Similarly in this experiment, all the selected words were grouped according to this principle. Then the frequency of words with identical initial consonant combinations within every semantic subclass was calculated. The results are presented in Table 1.

Table 1
Consonant Clusters' Frequency in Five Semantic Subclasses

	Audible	Perceptible	Movement	Concrete objects	Abstract notions
1. Bl-	12	14	5	12	7
2. Br-	4	17	12	49	9
3. Kl-	9	9	8	29	17
4. Kr-	9	13	6	47	25
5. Dr-	2	11	7	12	11
6. Fl-	4	28	23	31	10
7. Fr-	2	28	12	16	17
8. Gl-		32	7	12	8
9. Gr-	5	42	13	37	20
10. Pl-	5	25	9	38	13
11. Pr-	7	99	30	37	95
12. Kw-	1	17	4	8	30
13. Sk-	5	33	24	37	19
14. Sf-		2			2
15. Shr-	2	1	3	7	2
16. Sl-	1	29	20	12	8
17. Sm-		14	5	2	2
18. Sn-	9	11	4	10	4
19. Sp-	1	40	12	39	26
20. Spl-	1	6	2	4	1
21. Spr-		3	8	4	2
22. Skw-	6	8	7	3	5
23. St-	1	52	17	59	45
24. Str-	3	15	18	17	7
25. Sv-					
26. Sw-	2	18	16	9	6
27. Thr-	1	7	4	6	7
28. Tr-	6	39	40	53	54
29. Ts-			1	1	
30. Tw-	4	6	4	6	1

To find the ratio between the subclasses and combinations of sounds, and to identify any statistically significant phonosemantic relationships, we applied the chi-square test. For purposes of statistical analysis, it is often useful to present the data in the form of the "alternative distribution", that is, in tables consisting of 4 fields (2 columns and 2 lines). In this experiment, four-field tables have only been constructed in cases where the empirical datum of the cluster (in Table 1) is greater than its theoretically expected datum and the difference between them is positive. Relationships which turn out to be statistically significant ($\chi^2 \geq 3{,}84$) are evidence for the existence of real correspondences between the semantic categories and the initial consonant combinations. When the empirical datum is less than theoretically expected value, the null hypothesis will be recognized, that is, that there is no relationship between the phenomena under investigation.

Thus, 58 alternative tables were compiled, on the basis of Table 1, for clusters where the empirical datum is greater than the theoretical one. Table 2 illustrates the four-field table for the cluster *Bl-*.

Table 2
Distribution of Frequency for the Cluster *Bl-* and other Consonant Combinations in the Subclass "Audible" and other Subclasses

	Audible		Other subclasses		Total
Bl-	12	a	b	38	50
Other clusters	90	c	d	1952	2042
Total	102			1990	2092

Calculations were conducted by means of the formula:

$$(1) \qquad \chi^2 = \frac{(ad - bc)^2 N}{(a+c)(b+d)(a+b)(c+d)},$$

where $N = a + b + c + d$. Thus, in this case $\chi^2 = 40.36$, which is greater than 3.84. This means the relationship between *Bl-* and the subclass "audible" is statistically significant.

With the help of the chi-square test, it was established that there *is* an interaction between certain consonant clusters and semantic categories. Moreover, a measure of this interaction was determined (see also Lvova 2004). This measure may be calculated with the help of the coefficient of contingency, Φ, as defined by the following formula:

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

If the value computed for χ^2 is significant, then the coefficient of contingency, Φ, is significant as well. For the cluster *Bl-*, this coefficient is $\Phi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{40.36}{2092}} = 0.138$. The association-dissociation must also now be determined. If the empirical datum of *Bl-* is greater than the theoretical one, then the coefficient of contingency, Φ, will be positive. For example, the empirical datum for *Bl-* is 12, the theoretical datum is 2.4379, and thus Φ = +0.138.

Result from calculating the three necessary indices (chi-square test, coefficient of contingency and association-dissociation) are reflected in Table 3.

Table 3
Data of Chi-square Test and Coefficient of Contingency for English Initial Consonant Clusters in Five Semantic Categories

	Audible	Perceptible	Movement	Concrete objects	Abstract notions
1. Bl-	**χ^2 = 40.36** **Φ = +0.14**				
2. Br-				**χ^2 = 29.89** **Φ = +0.12**	
3. Kl-	**χ^2 = 9.35** **Φ = +0.07**			**χ^2 = 5.04** **Φ = +0.05**	χ^2 = 0.17 Φ = +0.01
4. Kr-	**χ^2 = 3.86** **Φ = +0.04**			**χ^2 = 17.56** **Φ = +0.09**	χ^2 = 0.69 Φ = +0.02
5. Dr-			χ^2 = 0.03 Φ = +0.004		χ^2 = 0.4 Φ = +0.01
6. Fl-			**χ^2 = 5.75** **Φ = +0.05**	χ^2 = 0.69 Φ = +0.02	
7. Fr-		χ^2 = 2.24 Φ = +0.03	χ^2 = 0.03 Φ = +0.004		χ^2 = 0.05 Φ = +0.013
8. Gl-		**χ^2 = 17.7** **Φ = +0.09**			
9. Gr-		χ^2 = 2.36 Φ = +0.03		χ^2 = 0.58 Φ = +0.016	
10. Pl-	χ^2 = 0.09 Φ = +0.006			**χ^2 = 8.63** **Φ = +0.07**	
11. Pr-		**χ^2 = 7.97** **Φ = +0.06**			**χ^2 = 34.47** **Φ = +0.13**
12. Kw-					**χ^2 = 29.26** **Φ = +0.12**
13. Sk-			χ^2 = 2.4 Φ = +0.03	χ^2 = 0.49 Φ = +0.02	
14. Sf-		χ^2 = 0.83 Φ = +0.02			χ^2 = 1.94 Φ = +0.03
15. Shr-	χ^2 = 2.32 Φ = +0.03		χ^2 = 0.25 Φ = +0.01	χ^2 = 2.43 Φ = +0.03	
16. Sl-		**χ^2 = 4.87** **Φ = +0.05**	**χ^2 = 9.76** **Φ = +0.07**		
17. Sm-		**χ^2 = 10.92** **Φ = +0.07**	χ^2 = 0.73 Φ = +0.018		
18. Sn-	**χ^2 = 29.47** **Φ = +0.12**				
19. Sp-		χ^2 = 1.11 Φ = +0.02		χ^2 = 1.25 Φ = +0.02	χ^2 = 0.01 Φ = +0.002
20. Spl-	χ^2 = 0.15 Φ = +0.009	χ^2 = 1.19 Φ = +0.02			

21. Spr-			**χ^2 = 13.28** **Φ = +0.08**		
22. Skw-	**χ^2 = 15.86** **Φ = +0.09**		χ^2 = 1.75 Φ = +0.03		
23. St-		χ^2 = 0.08 Φ = +0.006		χ^2 = 2.69 Φ = +0.04	χ^2 = 1.98 Φ = +0.03
24. Str-	χ^2 = 0.002 Φ = +0.0009		**χ^2 = 10.22** **Φ = +0.07**		
25. Sv-					
26. Sw-		χ^2 = 0.02 Φ = +0.003	**χ^2 = 10.38** **Φ = +0.07**		
27. Thr-			χ^2 = 0.008 Φ = +0.002		χ^2 = 0.5 Φ = +0.02
28. Tr-			**χ^2 = 4.90** **Φ = +0.04**		**χ^2 = 5.22** **Φ = +0.05**
29. Ts-			χ^2 = 1.85 Φ = +0.03	χ^2 = 0.45 Φ = +0.01	
30. Tw-	**χ^2 = 9.1** **Φ = +0.07**		χ^2 = 0.22 Φ = +0.01		

The results – of chi-square test and the coefficient of contingency – for different sound combinations turn out to be different. This testifies to a greater or lesser degree of connection between consonant clusters and semantic categories. It should also be noted that certain initial consonant clusters in English are not associated with any semantic subclass; that is, the *absence* of any phonosemantic relationship between the clusters *Dr-, Fr-, Gr-, Sk-, Sf-, Shr-, Sp-, Spl-, St-, Sv-, Sw-, Thr-, Ts-* and the five subclasses under investigation is evident in the data. On the other hand, some sound combinations can be seen to have significant relationships with certain semantic categories. This allows conclusions to be drawn with regard to the close correspondence of semantic subclasses with such initial consonant clusters (in order of χ^2 and coefficient Φ lessening):

"Audible": *Bl-, Sn-, Skw-, Kl-, Tw-, Kr-*

"Perceptible": *Gl-, Sm-, Pr-, Sl-*

"Movement": *Spr-, Sw-, Str-, Sl-, Fl-, Tr-*

"Concrete objects": *Br-, Kr-, Pl-, Kl-*

"Abstract notions": *Pr-, Kw-, Tr-*

Modern psycholinguistics incorporates such notions as *symbolic activity of the scale* and *symbolic potential of the sound*, introduced by Levickij (1998). The "symbolic potential" of the sound (SP) is defined as "the ability of this or that sound to symbolize certain notion or group of notions"; whereas the "symbolic activity of the scale" (SA) is defined as "the ability of this or that notion or group of notions to be symbolized by a certain sound" (Levickij 1998: 39).

In the case of the current data, it is too early to speak about the symbolic character of the investigated relations. Two notions may be distinguished: the functional activity of the

semantic subclass and the semantic potential of the phoneme combinations. Thus, by calculating the mean of the coefficients of contingency, Φ, for each of the five subclasses, the level of semantic activity of a subclass was discovered. The results of the calculations show different levels of semantic activity in the five semantic subclasses. Below are the semantic subclasses ranked in order of the mean coefficient Φ diminution:

Audible
Concrete objects
Abstract notions
Movement
Perceptible

The most active subclass proved to be "audible" where Φ is the highest (Φ = 0.58); least activity is observed in the semantic subclass "perceptible" (Φ = 0.36). That is, in modern English, the words belonging to the subclass "audible" are more strongly connected with the two-phoneme (or three-phoneme) initial consonant clusters than is the case with regard to the other subclasses.

The "semantic potential" of the sound combinations can be calculated by adding all the coefficients of contingency, Φ, for every cluster. The greater the sum of the coefficients, the higher the semantic potential of the cluster – that is, its interaction with the semantic subclasses under investigation. The sound combination *Pr-* (Φ = 0.19) has the highest semantic potential; the cluster *Dr-* (Φ = 0.014) has the lowest semantic potential. This means that the cluster *Pr-* is connected to the five subclasses examined here more strongly and more closely than is any other sound combination, whereas *Dr-* is connected *least* strongly to these classes.

The results outlined in this paper allow the following conclusions to be drawn:

- Using the statistical method of the chi-square test, it has been possible to establish the existence of correlations between semantic categories and initial consonant clusters.
- Using the coefficient of contingency, Φ, the degree of semantic relation between the initial consonant clusters and the semantic categories has been measured; the clusters with the strongest ties to particular semantic subclasses have been discovered, for instance – "audible": *Bl-, Sn-, Skw-, Kl-, Tw-, Kr-*.
- Calculating the *mean* of the coefficients of contingency, Φ, for each semantic subclass enabled us to find the most active subclass – "audible".
- The *sums* of the coefficients of contingency, Φ, for all the initial consonant clusters under investigation show the differing semantic potentials of different phoneme combinations; calculating these statistics shows that the cluster *Pr-* has the highest semantic potential, whereas the cluster *Dr-* has the lowest semantic potential.

References

Abelin, A. (1999). *Studies in sound symbolism*. Gőteborg: Department of Linguistics, Gőteborg University.

Brown, R. (1958). *Words and things*. Glencoe, Illinois: Free Press.

Johnson, R. (1967). Magnitude symbolism of English words. *Journal of Verbal Learning and Verbal Behavior 6, 508-511.*

Levickij, V.V. (1998). *Zvukovoj simvolizm. Osnovnye itogi*. Černovcy: Ruta.

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

Lvova, N.L. (2004). Zvukova symvolika počatkovych spolučen fonem u leksyčnomu skladi sučasnoji anglijskoji movy. *Naukovyj visnyk 213, 3-14.*

Magnus, M.A. (1999). *Dictionary of English sound*. http://www.conknet.com/~mmagnus/Bibliography.htm

Newman, S. (1933). Further experiments in phonetic symbolism. *American Journal of Psychology 45, 53-75.*

Peterfalvi, J.-M. (1970). *Recherches experimentales sur le symbolisme phonétique*. Paris: Centre Nationale de Recherche Scientifique.

Ullmann, S. (1959). *The principles of semantics*. Glasgow: Jackson, Son & Co.

Wallis, J. (1653). *Grammatica linguae anglicanae*. Oxford, translated by J.A. Kemp (1972), *Grammar of the English Language*. London: Longman.

Glottometrics 9, 2005, 29-39

Some properties of graphemic systems

Karl-Heinz Best, Göttingen[1]
Gabriel Altmann, Lüdenscheid

Abstract. In this contribution some properties of the graphemic representation of German and Swedish phonemes will be examined. The study has a tentative character, as it is not yet known whether the examined properties are applicable – mutatis mutandis – to other languages or scripts. In order to be able to generalize, different other languages must be processed. Here the following properties are described: graphemic uncertainty, grapheme size, graphemic load of characters, and character utility. No evaluation of frequencies is strived at.

1. Introduction

The graphic representation of phonemes, syllables, morae, words, ideas etc. can be equated with the establishing of a secondary language. *Cum grano salis* one can consider as such both all character systems from letter to sign script and secondary scripts (= ternary languages) like Morse signs or shorthand up to contemporary cryptographic systems. A secondary language can capture the entities of the primary language with more or less coding and decoding effort, with more or less learning effort, etc. Besides, a character system can have numerous other properties which will not be examined here (cf. e.g. Altmann 2004; Best 2005a,b; Bohn 1998, 2002; Gibson et al. 1963; Koch 1971; Prün 1994; Yu 2001). A description of the properties strives for their numerical expression, i.e. quantification and measurement, because processing of this kind is more appropriate for the search for regularities, and the pertinent hypotheses can be more easily tested (corroborated or rejected) if one analyzes different languages and scripts. For the present analysis we do not need to know the frequencies of letters or graphemes since we are concerned here merely with the graphic system.

The study of the effectivity of a writing system would result in a complex function, in which many variables like learning effort, writing speed, reading effort and frequency play a certain role. It requires a much more intensive research than was performed up to now (cf. e.g. Ramakrishna et al. 1962).

Our problem in the present contribution consists in the examination of the relationship between the phoneme and the grapheme system and its numerical grasping. Here one adheres to the heuristic principle of maximizing the graphemic system, i.e. of the establishment of the maximal number of graphemes. The partitioning of graphemes into letters is a posteriori possible but the other way round (from letter to grapheme representing a phoneme) is associated with problems. If we count the letters <t> and <s> in German separately, then we cannot a posteriori conclude when they represent the phonemes /t/, /s/ or /c/ (/ts/), when <ts> was a grapheme or two letters. A grapheme consists of one or more signs (letters or additional signs) representing a phoneme. Thus a grapheme can be identical with a letter but it can also be more ample.

Here we analyze German and Swedish as they are written with Latin characters. To this end we first use the German phonemes and the assignment of graphemes as can be found e.g. in Lühr (1986: 224-226) with reference to Duden (1984[4]). We add the phoneme /tš/ (/č/) and

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

eliminate shwa which is considered merely an allophone, not a phoneme (cf. Best 2001: 19). The grapheme <th> will not be considered since it occurs only in foreign words and names.

German

Phoneme	**Letters/Graphs**	**Comments, examples**
/i:/	<i, ie, ieh, ih>	Igel, Lied, sieht, ihm
/i/	<i>	dick
/ü:/	<ü, üh>	Hügel, kühn
/ü/	<ü>	Stück
/u:/	<u, uh>	Bube, Huhn
/u/	<u>	Druck
/e:/	<e, ee, eh>	Rede, Meer, Mehl
/e/	<e, ä>	Speck, älter
/ö:/	<ö, öh>	Öl, Höhle
/ö/	<ö>	können
/o:/	<o, oo, oh>	Brot, Moor, Sohn
/o/	<o>	Stock
/ä:/	<ä, äh>	Bär, Ähre
/a:/	<a, aa, ah>	Tag, Staat, kahl
/a/	<a>	Sack
/ai/	<ai, ei, eih>	Rain, Stein, Weih(-nachten)
/oi/	<äu, eu>	Säule, Eule
/au/	<au, auh>	schlau, rauh (old spelling)
/R/	<r, rr>	rot, irren
/l/	<l, ll>	Land, lallen
/m/	<m, mm>	Mund, Hummel
/n/	<n, nn>	neu, können
/ŋ/	<ng, n>	singen, <n> in front of /k/: sinken
/b/	<b, bb>	bald, rubbeln
/d/	<d, dd>	dann, zerfleddern
/g/	<g, gg>	ganz, eggen
/p/	<p, b, pp>	Punkt, <b> at the end of words: grob, Puppe
/t/	<t, dt, d, tt>	rot, Stadt, <d> at the end of words: Lied, flott <th> only in names and foreign words;
/k/	<k, q, g, x, ch, ck>	kurz, <q> in front of /v/: Quelle, <g> at the end of words or in front of /s/: zog, tags, <x> contains /k/: Hexe, <ch> in front of /s/ (Fuchs), Glocke
/pf/	<pf>	Pferd
/ts/	<z, ts, tz>	zerren, für /t/ + genitive -s: Amts, Fetzen
/tš/	<tsch>	Matsch
/v/	<w, u>	Wasser, <u> behind /k/: Quelle
/f/	<f, v, ff>	Feuer, Vater, Waffe
/z/	<s>	singen
/s/	<ß, s, ss, x>	Ruß, Haus, vergessen, <x> for /k/ +/s/: Hexe
/š/	<sch, s>	Schall, <s> in anlaut in front of /p/, /t/: Spiel, Stein
/X/ = /ç und x/	<ch, g>	Rauch, recht, <g> in the suffix <-ig>: König
/j/	<j>	jung
/h/	<h>	Haus

The Swedish data were taken from Braunmüller (1991) and Ternieden (1973). The Swedish has no genuine diphthongs. A peculiarity are the retroflex sounds which are merely allophones of the respective simple consonants. In <rd, rt, rl, rn, rs> the sound [r] merges with the following sound and effects partly the prolongation of the preceding vowel: karl /ka:ḷ/; etc. They are evaluated merely as allophones. An affricate /ts/ is not established; Braunmüller

Swedish

Phoneme	**Letters/Graphs**	**Comments, examples**
/i:/	<i>	vi
/i/	<i>	vill
/y:/	<y>	ny
/y/	<y>	hylla
/~~ʉ:~~/	<u>	hus
/~~ʉ~~/	<u>	dum
/u:/	<o>	bo
/u/	<o>	hon
/e:/	<e>	se
/e/	<e>	vecka
/ø:/	<ö>	köpa
/ø/	<ö>	höst
/o:/	<o, å>	son, gå
/o/	<o, å >	komma, sålt
/ɛ:/	<ä>	äta
/ɛ/	<ä, i>	tvätt, mig
/a:/	<a>	jag
/a/	<a>	all
/p/	<p, pp>	par, knapp
/t/	< rt, t, tt>	hjort, tack, hetta
/k/	<ch, ck, g, k, x>	och, kalv, <g> in front of <t> or /s/ in auslaut: sagt, flicka, <x> contains /k/: sex
/b/	<b, bb>	baka, knubbig
/d/	<d, dd, rd>	dag, knodd, gård
/g/	<g, gg>	god, hugg
/f/	<f, ff>	flicka, knuff
/s/	<rs, s, ss, x>	fors, sak, mössen, contained in <x>: sax
/š/	<sj, sk, skj, ssj, stj>	själv, sked, skjuta, hyssja, stjärna
/ç/	<k, kj, tj>	kedja, kjol, tjuv
/h/	<h>	hon
/v/	<v, vv>	vi, vovve (<vv> only in child language, Ternieden 19)
/j/	<dj, g, gj, hj, j, lj>	djur, ge, gjort, hjort, jag, ljud
/m/	<m, mm>	blom, blemma
/n/	<n, nn, rn>	ny, henne, hörn
/ŋ/	<g, ng, nk>	<g> in front of <n>: ugn, gång, <n> in front of <k>: sank
/l/	<l, ll, rl>	lag, fall, karl
/r/	<r, rr>	rak, herr

does not have it in the inventory, though minimal pairs are imaginable: spets (tip) − spett (spit); there is also spex (farce). <c> for /s/ and /k/ occurs in foreign words (december,

cancer), similarly <ch> for /ç/ (charterplan) or /š/ (charm), <q> (quisling) for /k/, <z> (zoo) for /s/. There is no hardening at the word end as in German (Braunmüller 1991: 37).

In Swedish there are graphemes which are not pronounced in some circumstances; e.g. in the derivateme -{ig} the grapheme <g>: the phonological form of „rolig“, „roligt“ is /ru:li/ and /ru:lit/. There are some other such cases not considered here.

It must be explicitly mentioned that we look merely in one direction: from phoneme to grapheme. This can be done in two alternative ways: (1) Only the given grapheme is considered without its sometimes necessary environment, leading to the correct phoneme identification. This is possible if the environment itself cannot be ascribed to a phoneme or if there are no exceptionless rules (cf. in Slovak the palatalizing effect of /i/ and /e/ after /d, t, n, l/ but not always, cf. *dramatický*). Since language steadily changes, such rules are illusory. Some graphemes need not be phonetically realized, they can identify zero morphemes (cf. French *ils parlent*). (2) If secondary conditions are concidered parts of graphemic representation, they can be taken into account, e.g. in German a vowel is (mostly) short if there are two consonants behind it, *sometimes* if there is only one consonant grapheme following but in this case the rules surpass phonemics and graphemics and are situated in morphology.

We shall adhere to the simplest alternative, i.e. we consider a realized phoneme and its written correspondent. In no case we shall look in the opposite direction, namely from grapheme to phoneme. Since our aim is merely a first trial towards quantification, it can be complicated if further languages will be analyzed.

The usual error associated with the beginnings of such an analysis is the consideration of the written form of the word and making decisions starting from it. We recommend to start from the phonetically transcribed word, assigning its sounds to phonemes and at last look how these phonemes are graphemically represented in the written word.

2. Orthographic uncertainty of the phoneme

The more graphemes are used to represent a phoneme, the greater is its orthographic uncertainty. If a language uses letters that have not been developed directly for the given language, greater uncertainty can result, cf. e.g. in English. This can, of course, be balanced by the use of some graphemes more frequently than of other ones for the representation of a phoneme. For example, in German one finds <p> more frequently than <pp> or <b> for the representation of the phoneme /p/. Since here we do not use frequency data, we define the *unweighted orthographic uncertainty* of a phoneme, as it is usual, by the binary logarithm of the cardinal number of the representing grapheme set. If we denote the grapheme set of a phoneme /x/ as $G_x = \{g_1, g_2, \ldots, g_n\}$ and its cardinal number (i.e. the number of different graphemes in this set) $\#G_x$ or as n_x, then we can define:

(1) $U_x = \log_2 \#G_x = \log_2 n_x.$

Thus we obtain for the German phonemes /i:/, /o/ and /k/:

$$U_{/i:/} = \log_2 4 = 2$$
$$U_{/o/} = \log_2 1 = 0$$
$$U_{/k/} = \log_2 6 = 2.58.$$

The corresponding uncertainty indices for all German phonemes are shown in Table 1a, for Swedish phonemes in Table 1b.

Table 1a
Unweighted orthographic uncertainties for German phonemes

Phoneme	Orthographic uncertainty U_x	Cardinal number n_x	Number of phomenes f_x
/i/, /ü/, /ö/, /o/, /a/, /pf/, / tš /, /z/, /j/, /h/;	0	1	10
/ü:/, /u:/, /e/, /ö:/, /ä:/, /oi/, /au/, /R/, /l/, /m/, /n/, /η/, /b/, /d/, /g/, /v/, /š/, /X/;	1	2	18
/e:/, /o:/, /a:/, /ai/, /p/, / ts /, /f/;	1.58	3	7
/i:/, /s/, /t/;	2	4	3
/k/;	2.58	6	1

Table 1b
Unweighted orthographic uncertainties for Swedish phonemes

Phoneme	Orthographic uncertainty U_x	Cardinal number n_x	Number of phomenes f_x
/i:/, /i/, /y:/, /y/, /ʉ:/, /ʉ/, /u:/, /u/, /e:/, /e/, /ø:/, /ø/, /ɛ:/, /a:/, /a/, /h/	0	1	16
/o:/, /o/ ,/ɛ/, /p/, /b/, /g/, /f/, /v/, /m/, /r/	1	2	10
/t/, /d/, /ç/, /n/, /η/, /l/,	1.58	3	6
/s/	2	4	1
/k/,/š/	2.32	5	2
/j/	2.58	6	1

The orthographic uncertainty of the whole phoneme system can be computed as the mean of uncertainties, i.e. as

$$(2) \qquad \bar{U} = \frac{1}{N}\sum_{x\in I} U_x f_x = \frac{1}{N}\sum_{x\in I} f_x \log_2 n_x ,$$

where $N = \sum_x f_x$ and I is the inventory of phonemes. For the German written in Latin script we obtain $\bar{U}$ = 0.965, for the Swedish $\bar{U}$ = 0.797. The uncertainty in Swedish is smaller, evidently because the vowels are graphemically simple. Writing in (2) the theoretical values $p_x = f_x/N$ one sees that (2) is the expected value of uncertainty, i.e. $E(\log_2 n)$.

The minimal orthographic uncertainty is, of course, 0, which results when all phonemes are represented by exactly one grapheme. The maximum is not known. It can be computed theoretically but it is not quite realistic. A maximum could be obtained if each phoneme would be represented by the maximally admitted number of graphemes. In German it would be 6, thus the maximum would be $\log_2 6 = 2.58$. A relative measure could be obtained as

$$(3) \qquad U_{rel} = \frac{\bar{U}}{U_{\max}} ,$$

yielding for German U_{Dt} = 0.965/2.58 = 0.374 and for Swedish U_{Sw} = 0.797/2.58 = 0.309. This index would lie in the interval <0, 1> but its lower values would be more frequently found than the higher ones. Thus we content ourselves with $\bar{U}$.

In cases where the set of Latin letters is not sufficient for the representation of the phonemes one must assume that at least some graphemes arise by combination of letters. In the last column of Table 1 can be seen that the cardinal numbers of grapheme sets representing individual phonemes follow a certain distribution that at present cannot be analyzed because we have merely two languages at our disposal. However, one can assume that it will be a Poisson distribution.

Another possibility of characterizing the *economy of the script system* (*SE*) is the ratio of the number of phonemes to that of graphemes, i.e.

$$SE = \frac{\#P}{\#G}. \tag{4}$$

This index has the disagreeable property that its maximum is known but its minimum cannot be stated. If the number of phonemes equals that of graphemes, it attains the value 1, otherwise it is smaller than 1. The value 0 cannot be attained since languages have at least 11 phonemes and never an infinite number of graphemes, not even in Chinese or Assyrian. Evidently, for different writing systems different measures of economy must be defined.

The second disadvantage of this index is the fact that its sampling properties are not yet known. Alternatively, one can denote this index as the measure of phoneme-grapheme correspondence index.

For German we obtain SE_G = 39/68 = 0.57, for Swedish SE_S = 36/57 = 0.63. When performing writing reforms one must care that this index is as great as possible.

A *weighted measure of orthographic uncertainty* could be established by considering the relative frequencies of individual graphemes representing a given phoneme. Let the relative frequencies of graphemes representing a given phoneme /x/ are given by a set $\{p_1, p_2, \ldots, p_n\}$ with $\sum p_i = 1$, then we can express it by means of the usual entropy

$$WU_{/x/} = -\sum_{i \in G_x} p_i \log_2 p_i . \tag{5}$$

3. The phonemic uncertainty of a grapheme

From the graphemic point of view the same kind of uncertainty can be defined. A grapheme may be used to represent different phonemes. However, this is merely a look in the opposite direction. For languages using Latin letters the reverse procedure must be made but the formulas are the same. However, in languages using signs like Chinese, Japanese, Assyrian etc. the problem will be more complex. Even in French or English there will be many formal problems and even the setting up of basic data will be associated with difficulties. Graphemes representing zero morphemes (like in French) have no phonemic existence.

4. Grapheme size

Graphemes can be made of single letters or of their combinations. Some graphemes are composed of letters and additional symbols which are not used as letters, e.g. in European lan-

guages <ů, ś, ř, ş, ô, å> etc. These graphemes in scripts of Latin origin are considered composed, those contained in the Latin alphabet <u, s, r, o, a> etc. are simple.

In the above graphematic presentation we found the following graphemes (ordered according to their size, cf. Table 2)

Table 2a
Grapheme size in German

Size	Grapheme	Number
1	<i, ü, u, e, ö, o, ä, a, r, l, m, n, b, d, g, p, t, k, q, z, w, f, v, s, ß, x, j, h>	28
2	<ie, ih, üh, uh, ee, eh, öh, oo, oh, äh, aa, ah, ai, ei, äu, oi, au, rr, ll, mm, nn, ng, bb, dd, gg, pp, dt, tt, ch, ck, pf, ts, tz, ff, ss>	35
3	<ieh, eih, auh, sch>	4
4	<tsch>	1

Table 2b
Grapheme size in Swedish

Size	Grapheme	Number
1	<i, y, u, o, e, ö, å, ä, a, p, t, g, k, x, b, d, f, s, h, v, j, m, n, l, r>	25
2	<pp, rt, tt, ch, ck, bb, dd, rd, gg, ff, rs, ss, sj, sk, kj, tj, vv, dj, gj, hj, lj, mm, nn, rn, ng, nk, ll, rl, rr>	29
3	<skj, ssj, stj>	3

The distribution of these sets displays a certain regularity that will not be analyzed here. In order to characterize German and Swedish written with Latin letters we content ourselves with the mean size of a grapheme which can be computed from Table 2a as follows

$$\overline{G}_G = [1(28) + 2(35) + 3(4) + 4(1)]/68 = 1.68$$

and for Swedish $\overline{G}_S = 92/57 = 1.61$. Thus the mean length of graphemes is slightly greater in German. This quantity is the greater, the greater the discrepancy between the number of phonemes and the number of Latin letters, i.e.: the more the number of phonemes surpasses that of letters, the more graphemes with more components we need. This is a very simple hypothesis, easily translatable in formulas but its testing can be performed merely after many languages have been analyzed. However, syllabic and sign languages behave differently, thus the hypothesis may turn out to be very complex.

If we consider the „Umlaut“ in German as an additional component (cf. Table 3a) then we obtain for German

$$\overline{G}_G = [1(25) + 2(34) + 3(8) + 4(1)]/68 = 1.78.$$

Table 3a
German grapheme size (with „Umlaut“ as component)

Size	Grapheme	Number
1	<i, u, e, o, a, r, l, m, n, b, d, g, p, t, k, q, z, w, f, v, s, ß, x, j, h>	25
2	< ü, ö, ä, ie, ih, uh, ee, eh, oo, oh, aa, ah, ai, ei, oi, au, rr, ll, mm, nn, ng, bb, dd, gg, pp, dt, tt, ch, ck, pf, ts, tz, ff, ss>	34
3	< üh, öh, äh, äu, ieh, eih, auh, sch>	8
4	<tsch>	1

For Swedish we have two additional components, namely the umlaut and the circle above <a> thus we would obtain from Table 3b $\overline{G}_S = 1.67$, again a slightly smaller grapheme size than in German.

Table 3b
Grapheme size in Swedish (with additional signs)

Size	Grapheme	Number
1	<i, y, u, o, e, a, p, t, g, k, x, b, d, f, s, h, v, j, m, n, l, r>	22
2	<ö, å, ä, pp, rt, tt, ch, ck, bb, dd, rd, gg, ff, rs, ss, sj, sk, kj, tj, vv, dj, gj, hj, lj, mm, nn, rn, ng, nk, ll, rl, rr>	32
3	<skj, ssj, stj>	3

5. Graphemic load of letters

As can be seen in Table 2, graphemes are composed either of single letters or of their combinations with other letters or of letters with additional signs, as the German umlaut. In other languages using the Latin script there are frequently numerous additional components.We understand the load of a letter to be its participation in the given graphemes. Let x denote the number of graphemes in which the letter participates, then we obtain the results in Table 4a and 4b.

Table 4a
Participation of Latin letters in graphemes (German)

Component in x graphemes	Letter	Number of letters f_x
1	q, w, v, ß, x, j	6
2	r, l, m, b, k, z	6
3	n, d, g, p, f	5
4	c	1
5	s	1
6	o	1
7	i, t	2
8	u, e	2
9	a	1
16	h	1

Table 4b
Participation of Latin letters in graphemes (Swedish)

Component in x graphemes	Letter	Number of letters f_x
1	i, y, u, x,	4
2	o, p, c, b, f, v, m,	7
3	a, t, h, g,	4
4	d, l,	2

5	n,	1
6	r, k,	2
8	s,	1
11	j	1

The modelling of the participation would be somewhat premature on two grounds: we have merely two languages and the class sizes are a little bit small. In other languages like e.g. Chinese where the sign components occur many times we would obtain a proper frequency distribution.

Preliminarily the German can be at least characterized by the mean participation of letters in grapheme construction ($\overline{B}$). From Table 4a we obtain

$$\overline{B} = [1(6) + 2(6) + 3(5) + 4(1) + 5(1) + 6(1) + 7(2) + 8(2) + 9(1) + 16(1)]/26 = 3.96,$$

which means that each letter participates on the average in four graphemes. For Swedish we obtain from Table 4a $\overline{B}$ = 74/22 = 3.36, a somewhat smaller value. A comparison with other languages would show us whether this is few or many. For the time being we shall not test these differences though a test is straightforward.

6. Letter utility

In chapter 5 we considered the participation of letters but not their position in graphemes. We assume that even the position can play a characteristic role. The earlier a letter occurs in the grapheme the more it can play the role of primary representative of a given phoneme. Thus the consideration of the position is a positionally conditioned weighting of the participation. We can illustrate it with an example. Consider the German letter <s> that can occur in the following graphemes

<s, ts, ss, sch, tsch>

i.e. 3 times in the first position, 3 times in the second. Multiplying and adding these evaluations we obtain its positional participation as

$$PP_{<s>} = 3(1) + 3(2) = 9,$$

or in general

$$(6) \qquad PP_{<x>} = \sum_{g_i \in G_x} P_x(g_i),$$

where $P_x(g_i)$ means the position of letter <x> in grapheme g_i which is element of the grapheme set G_x of /x/. For the individual letters we obtain the weighting as presented in Table 5a and 5b. Here we considered also <c> because it belongs to the Latin alphabet.

Table 5a
Positional weighting of letter participation in graphemes (German)

Weight $PP_{<x>}$	Letter x	Number f_x
1	q, w, v, ß, x, j	6
2	c	1
3	k, z	2
4	r, l, m, b, g,	5
5	n, d, p	3
6	f	1
7	o	1
9	s	1
10	i	1
11	a	1
12	u, e	2
37	h	1

Table 5b
Positional weighting of letter participation in graphemes (Swedish)

Weight $PP_{<x>}$	Letter	Number f_x
1	i, y, u, e, x,	5
2	o,c	2
3	a	1
4	p, b, f, h, v, m	6
7	g, d, l	3
8	k, n	2
9	r	1
13	s	1
24	j	1

The higher the positional weight of a letter the smaller its priority. Thus <h> is used often as a sign of length. The obtained numbers again point at a distribution but preliminarily we can merely speculate about its form. The problem must be postponed. Nevertheless, we can compute at least the *mean positional weight* of Latin letters in these two languages.

(7) $$\overline{PW}(Language) = \frac{1}{L}\sum_x f_x PP_{<x>}$$

where $L = \sum_x f_x$ and $PP_{<x>}$ are the existing weights. From Tables 5a and 5b we obtain for German $\overline{PW}(G)$ = 153/25 = 6.12 and for Swedish $\overline{PW}(S)$ = 119/22 = 5.41

There are languages in which graphemes are composed of signs ordered vertically. In that case the weighting begins at the top. However, if the sign at the top is additive or a modification of another basic sign, the weighting begins on the main sign.

It can be assumed that the positional weight of letters correlates with their frequency of occurrence but the kind of dependence will not be examined here.

Summary

The quantified and measured properties like graphemic uncertainty, grapheme size, graphemic load of characters, and character utility are merely the first possibilities towards the characterization of writing systems, that can be helpful for the observation of their development, comparison and effectivity, and at last for setting up of hypotheses about scripts. The next logical step would be the establishing of further properties, the evaluation of other script systems and the consideration of grapheme and letter frequencies and their relation to the "non-weighted" properties.

References

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

Best, K.-H. (2001). Silbenlängen in Meldungen der Tagespresse. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 15-32.* Göttingen: Peust & Gutschmidt.

Best, K.-H. (2005a). Buchstabenhäufigkeiten im Deutschen und Englischen. *Naukovyj Visnyk Černivec'koho Universytetu. Serija "Germans'ka filolohija"* (to appear).

Best, K.-H. (2005b). Zur Häufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten. *Msc.*

Bohn, H. (1998). *Quantitative Untersuchungen der modernen chinesischen Sprache und Schrift.* Hamburg: Kovač.

Bohn, H. (2002). Untersuchungen zur chinesischen Sprache und Schrift. In: Köhler, R. (ed.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik: 127 – 177.* http://ubt.opus.hbz-nrw.de/volltexte/2004/279/

Braunmüller, K. (1991). *Die skandinavischen Sprachen im Überblick.* Tübingen: Francke.

Duden (1984[4]). *Grammatik der deutschen Gegenwartssprache.* Bearb. Von G. Drosdowski in Zusammenarbeit mit G. Augst u.a. Mannheim/ Wien/ Zürich: Dudenverlag.

Gibson, E.J., Osser, H., Schiff, W., Smith, J. (1963). An analysis of critical features of letters, tested by a confusion matrix. In: *Cooperative Research Project No. 639: A Basic Research Program on Reading.* Washington: U.S. Office of Education.

Koch, W.A. (1971). *Taxologie des Englischen*. München: Fink.

Lühr, R. (1986). *Neuhochdeutsch*. München: Fink.

Prisma Handwörterbuch Schwedisch-Deutsch. Berlin u.a.: Langenscheidt [5]1996.

Prün, C. (1994). Validity of Menzerath-Altmann's law: graphic representation of language, information processing systems and synergetic linguistics. *J. of Quantitative Linguistics 1, 148-155.*

Ramakrishna, B.S., Nair, K.K., Chiplunkar, V.N., Atal, B.S., Ramachandran, V., Subramanian, R. (1962). *Some aspects of the relative efficiency of Indian languages.* Bangalore.

Ternieden, R. (1973). *Grundzüge der Phonematik, Morphophonematik und Graphematik des Schwedischen.* Diss.phil., Köln.

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

Glottometrics 9, 2005, 40-45

Some properties of slang words

Kvetoslava Serdelová, University of Trnava[1]

Abstract. Slang words can have properties which are not conspicuous with standard words. Using some Slovak slang words, different kinds of familiarity and a kind of semantic uncertainty are operationalized.

Keywords: slang, familiarity, uncertainty, Slovak

1. Slang words are „normal“ words of a language differing from the other ones merely by the status of their codification (standardness) and by their discourse properties. They arise by the activation of the same requirements as the standard words, however, perhaps in different degree. They have their own life cycles, formal and semantic properties, displaying sometimes deviating tendencies but they have also some properties which are not so conspicuous with the “normal” words since the latter are considered integral parts of the lexicon.

In the present paper we try to explicate the concept of *familiarity* of slang words which is irrelevant for “normal” words considered all equally familiar even if not used with the same frequency. The problem of familiarity concerns also technical words but their knowledge is clearly restricted to the circle of specialists while only a part of slang words is subjected to this restriction. Many of them are known more or less. And it is this “more or less” that can and must be operationalized and expressed quantitatively. The present investigation has been performed within the scope of a greater project on Slovak slang.

2. *N* persons in the age of 16 to 25 obtained a questionnaire with 20 slang words and should fill in its standard language form. The standard language forms are contained in the dictionary of Hochel (1993) and were compared with the responses. There were three possible outcomes: correct response, incorrect response and “I do not know”. Since an incorrect response still means that the speaker knows the word but did not choose the equivalent standard word – there are also different synonyms –, only the “I do not know”-answer was considered a sign of infamiliarity. Thus a simple index of familiarity can be set up by means of the ratio of the number of “I do not know”-answers to the number of questioned persons (*N*), namely

$$(1) \qquad F_1(word) = 1 - \frac{|"I\ don't\ know"|}{N}.$$

This index is a simple proportion and admits all appropriate statistical procedures. For 20 words and *N*, which was different for different words, we obtained the results presented in Table 1.

[1] Address correspondence to: K. Serdelová, E-mail: kvietok_ruza@pobox.sk

Table 1
Knowledge of slang words

Slang word	**Correct response**	**Incorrect response**	**I don´t know**	***N***
anglina (English)	41	40	0	81
hák	12	38	27	77
cucflek	9	64	5	78
drbnúť	1	84	3	88
rulinár	11	2	63	76
fasovina	10	46	28	84
hybrid	1	71	7	79
bifľa	18	65	0	83
krezo	1	8	68	87
mrmeľ	25	45	14	84
spiatočka	0	77	8	85
poďubanec	4	44	30	78
disko	66	16	0	82
sučka	2	77	6	85
kvádro	27	16	36	79
učitelák	1	52	25	78
vymihať	7	19	52	78
zafachčiť	10	51	17	78
trolák	22	11	47	80
obyčka	7	49	23	79

The result of computing index (1) is in Table 2.

Table 2
Familiarity of some Slovak slang words

Slang word	**Familiarity$_1$**
anglina (English)	1.00
bifľa	1.00
disko	1.00
drbnúť	0.97
cucflek	0.94
suka	0.93
hybrid	0.91
spiatočka	0.91
mrmeľ	0.83
zafachčiť	0.78
obyčka	0.71
učitelák	0.68
fasovina	0.67
hák	0.65
poďubanec	0.62
kvádro	0.54
trolák	0.39

vymihať	0.33
rulinár	0.17
krézo	0.12

3. A more complex measure can be constructed in the following way. The answers are weighted by integer values, i.e. they are scaled on the ordinal scale. The answer "I do not know" which means "no familiarity" obtains the value 1; an incorrect answer which means at least a partial familiarity obtains the value 2; and a correct answer obtains the value 3. Thus familiarity based on this scaling can be defined in different ways. Let h_i be the scaling value (1, 2, 3) and f_i the frequency as shown in Table 1 then we obtain the index

$$(2) \qquad F_2(word) = \frac{1}{N_w}\sum_{i=1}^{3} f_i h_i$$

which is nothing else but the average familiarity value. N_w may differ from word to word; it is given as the marginal sum at the right side of Table 1. The index (2) lies in interval <1, 3>. Thus a relative index can be constructed in the usual way as

$$(3) \qquad F_3(word) = \frac{F_2 - 1}{2},$$

which lies in the interval <0, 1>, 1 representing maximal familiarity.

For example the word *hák* in Table 1 has 77 responses (marginal sum) out of which 27 have the value 1, 38 have the value 2, and 12 have the value 3, thus

$$F_2(ha'k) = \frac{1}{77}[27(1) + 38(2) + 12(3)] = 1.81$$

from which

$$F_3(ha'k) = \frac{1.81 - 1}{2} = 0.40.$$

All values are presented in Table 3.

Table 3
Scaling of slang words according to familiarity

Slang word	**Correct explanation**	**Incorrect explanation**	**„I don´t know“**	**Index F_2**	**Index F_3**
disko	66	16	0	2.80	0.90
anglina	41	40	0	2.51	0.75
bifľa	18	65	0	2.22	0.61
mrmeľ	25	45	14	2.13	0.57
drbnúť	1	84	3	1.98	0.49
sučka	2	77	6	1.95	0.48

spiatočka	0	77	8	1.906	0.45
cucflek	9	64	5	2.05	0.53
hybrid	1	71	7	1.92	0.46
fasovina	10	46	28	1.79	0.39
kvádro	27	16	36	1.89	0.44
zafachčiť	10	51	17	1.91	0.46
obyčka	7	49	23	1.80	0.40
hák	12	38	27	1.81	0.40
trolák	22	11	47	1.688	0.34
učitelák	1	52	25	1.692	0.35
poďubanec	4	44	30	1.67	0.33
vymihať	7	19	52	1.42	0.21
rulinár	11	2	63	1.32	0.16
krezo	1	8	68	1.13	0.06

The familiarity, in any of its versions, shows which words are dying out and which words are becoming active. Very actual is e.g. the comparison of the values of *rulinár* ("teacher of Russian", standard: *ruštinár*) as compared with *anglina* ("English", standard: *angličtina*).

Automatically the next question arises: which other properties of slang words are related to their familiarity? This is, of course, a question that cannot be answered using merely 20 words and merely one language. The problem must be postponed.

4. The majority of standard words is of course polysemic but the speakers know it and are aware of the given meaning in a special environment. This is not always the case of slang words which may be polysemic, too, but different users use it either in all meanings or only in one meaning or even in several ones but not all. Still other speakers know the word but are not quite sure about the exact meaning. When questioned, they give different answers one of which is correct, the other ones are false. Thus slang words have a greater semantic variability or rather uncertainty than standard words. *Uncertainty* is usually operationalized with the aid of entropy. Let N_w be the number of all responses concerning the meaning of a slang word, which is at least as great as the number of interviewees, since an interviewee can give several responses; let n_w be the number of *different* responses (i.e. the number of assumed meanings) and let S_i be the number of responses of type i ($i = 1,2,\ldots,n_w$) then the meaning entropy or meaning uncertainty of a slang word can be defined as

$$(4) \qquad H(word) = -\sum_{i=1}^{n_w} \frac{S_i}{N_w} \log_2 \frac{S_i}{N_w} = \log_2 N_w - \frac{1}{N_w} \sum_{i=1}^{n_w} S_i \log_2 S_i \,.$$

For example, for the word *hák* there were $N_{hák} = 77$ responses yielding $n_{hák} = 13$ different types whose numbers S_i were 10, 2, 10, 18, 2, 1, 1, 27, 1, 1, 1, 2, 1. Inserting these number in (4) we obtain

$$H(hák) = \log_2 77 - [[10 \log_2 10 + 2 \log_2 2 + 10 \log_2 10 + 18 \log_2 18 + 2 \log_2 2 +$$
$$+ 1 \log_2 1 + 1 \log_2 1 + 27 \log_2 27 + 1 \log_2 1 + 1 \log_2 1 + 1 \log_2 1 +$$
$$+ 2 \log_2 2 + 1 \log_2 1]/77 = 2.6839.$$

The results for the 20 selected words are presented in Table 4. The relative entropy is computed in the usual way and is given in the last column of Table 4.

Table 4
Entropies of 20 Slovak slang words

Slang word	**Entropy**	**Relative entropy**
bifľa	4.4412	0.1306
zafachčiť	4.2367	0.1367
hybrid	4.0419	0.1347
cucflek	3.8465	0.1602
mrmeľ	3.4307	0.1715
poďubanec	3.3885	0.1255
spiatočka	3.3865	0.1881
fasovina	3.3828	0.1538
sučka	3.3550	0.1525
drbnúť	2.8591	0.1361
hák	2.6839	0.2065
obyčka	2.5139	0.2793
učitelák	2.3222	0.1659
vymihať	2.1819	0.1148
kvádro	2.1202	0.1767
trolák	1.5503	0.1939
disko	1.2878	0.1431
anglina	0.9999	0.4999
rulinár	0.7924	0.1981
krézo	0.7771	0.1295

It must be remarked that not only the relative entropy can be important but also the absolute entropy because for a slang word it is also relevant how many responses at all are given.

Two questions arise at once. (1) Is uncertainty related to familiarity and if so, how is this relation, and (2) if there are different types of responses, do they follow a special rank-frequency distribution?

Both questions can be answered merely tentatively because the number of words and languages is still small. The second question can be answered more easily. Ordering the frequencies of types of the word *hák* given above according to their frequencies we obtain the result in Table 5.

Table 5
Rank-frequency distribution of meaning types
of the slang word *hák*

Rank	Frequency	Theoretical frequency (5)
1	27	27.25
2	18	16.36
3	10	10.54
4	10	7.03
5	2	4.78
6	2	3.29
7	2	2.28
8	1	1.60
9	1	1.12

10	1	0.79
11	1	0.56
12	1	0.40
13	1	1.00
	77	k = 0.6402, p = 0.2677 DF = 8, X^2 = 4.41 P = 0.82

As expected, the rank-frequency distribution of a diversified entity abides by the positive negative binomial distribution (cf. Altmann 1996, 2005), namely

$$(5) \qquad P_x = \binom{k+x-1}{x} \frac{p^k q^x}{1-p^k}, \quad x = 1,2,3,\ldots,$$

yielding the result in the third column of Table 5. The result is, of course, preliminary but it agrees with the general theory.

The first question can be answered in different ways, e.g. by means of the rank correlation coefficient. It can easily be stated that the familiarity measures are highly correlated with one another, i.e. the computation of one of them is sufficient for the characterization of slang words. However the (Spearman) rank correlation of familiarity with uncertainty (in form of absolute entropy) is not significantly correlated. We obtain ρ = 0.3865 und u = 1.6845, which means that familiarity and meaning uncertainty are two different not correlated properties.

5. The question arises, how these properties can be embedded in Köhler´s self-regulation cycle (1986, 2002). Are they associated with frequency, length, polytexty, inventory size etc.? Must they be treated as special phenomena or are they integral part of standard language? Perhaps a research on a wider basis would help us to answer some of these questions.

References

Altmann, G. (1996). Diversification processes of the word. *Glottometrika 15, 102-111*.

Altmann, G. (2005). Der Diversifikationsprozess. In: Altmann, G., Köhler, R., Piotrowski, R.G. (eds.), *Handbook of Quantitative Linguistics, Art. 65*. Berlin: de Gruyter (in print).

Hochel, B. (1993). *Slovník slovenského slangu.* Bratislava: Hevier.

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

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

Glottometrics 9, 2005, 46-53

On letter distinctivity

Gordana Antić, Graz[1]
Gabriel Altmann, Lüdenscheid

Abstract. A script system must possess a certain distinctivity to ensure ease of writing and recognition. This property can be operationalized and computed in terms of the difference between a given symbol and all the other symbols. In this article, the distinctivity of Arial letters, and of the Latin alphabet as a whole printed in Arial, will be analyzed. The method can also be applied to other scripts.

Keywords: letter distinctivity, script distinctivity, Arial

1. In a previous article (Altmann 2004), we described a possible approach to characterizing the complexity of a written letter or other graphic image. Any graphic sign is composed of dots, straight lines, or arches of different form and orientation, whose contacts are either smooth (continuous), or sharp (crisp), or represent a crossing. Therefore, it was relatively straightforward to define a measure of complexity, simply by ranking the categories and summing the results.

This measure of complexity may – but need not – be associated with the *distinctivity* of a graphic sign. High complexity does not imply high distinctivity – but neither does simplicity. *Distinctivity* refers to the quality of being easily recognized and differentiated from other signs. For example, the hypothetical (and fictitious) Roman cipher IIIIIIIII is very simple to write and to memorize. However, it is not easily recognized, because it is necessary to count the strokes. The actual historical form IX is both easier to recognize and simpler to reproduce. Using our system of measurement, the first sign has 18 points of complexity, the second merely 9; that is, it is half as complex as the first – even though there is no conceptually "simpler" way to write 9 than as a series of nine strokes.

Similar considerations show that distinctivity of a sign cannot be measured absolutely, but merely relative to all other signs of the script. Let us consider two more fictitious numbers, IIIIIIIII and IIIIIIII. We can see that these symbols differ by only one stroke. The corresponding actual symbols, VIII and IX, differ by two strokes. However, this latter difference is clearer. Thus, in the first step, we define the *graphemic distinctivity* of a sign as *the average of the differences between that sign and all other signs in the script*. The difference is computed on the basis of line complexities and orientations.

To define a difference, we must first characterize the form numerically. This is done by computing the complexity of the sign, using our system of weighting forms. We gave elements of forms the following weights:

Forms:	Dot	1
	Straight line	2
	Arch (<180°)	3
Connections:	Continuous (0° or 180°)	1
	Crisp, sharp (≠180°)	2
	Crossing	3

[1] Address correspondence to: Gordana Antić, e-mail: g.djuras@tugraz.at

To compute the graphic difference between two signs, we use the following weighting:

Difference in direction:

- Between two straight lines.

  We have four possible directions in Arial: (i) | (ii) --- (iii) \ (iv) / .
  The difference between (i) and (ii) and between (iii) and (iv) is 2;
  all other differences are 1.

- Between two arches:

  There are six kinds of arches in Arial:
  (i) opening up (north),
  (ii) opening down (south),
  (iii) opening to left (west),
  (iv) opening to right (east),
  (v) opening to north-east (like in the middle of S, upper part)
  (vi) opening to south-west (like in the middle of S, lower part).

The differences between the direction (opening) of the arches are as follows:

	W	NW	N	NE	E	SE	S	SW
W	0	1	2	3	4	3	2	1
NW	1	0	1	2	3	4	3	2
N	2	1	0	1	2	3	4	3
NE	3	2	1	0	1	2	3	4
E	4	3	2	1	0	1	2	3
SE	3	4	3	2	1	0	1	2
S	2	3	4	3	2	1	0	1
SW	1	2	3	4	3	2	1	0

The difference between a straight line and an arch (of any orientation) is 5 (= adding the weight of straight line and arch) + all joining points.

The following figure illustrates all the possible lines in Arial letters:

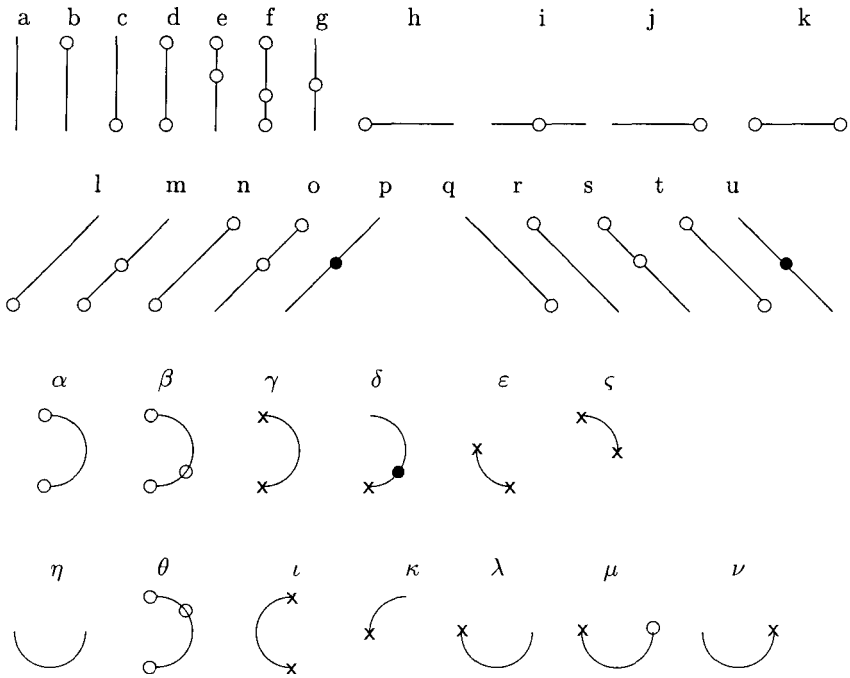


Figure 1. Straight lines and arches of Arial letters

The values of these straight lines and arches are as follows:

a	2	f	8	k	6	p	5	u	5
b	4	g	4	l	4	q	4		
c	4	h	4	m	6	r	4		
d	6	i	4	n	6	s	6		
e	6	j	4	o	6	t	6		
α	7	δ	7	η	3	κ	4	ν	4
β	9	ε	5	θ	9	λ	4		
γ	5	ζ	5	ι	5	μ	6		

The differences between individual lines are presented in Table 1.

The next step is to represent all Arial letters as vectors of different dimensions and a table of differences among all letter components. The vectors are as follows:

A = <o, k, s>
B = <f, β, θ>
C = <κ, λ >
D = <d, α >
E = <f, h, h, h>
F = <e, h, h>
G = <κ, μ, j >
H = <g,k,g>
I = <a>
J = < η >
K = <g, m, r>
L = <c, h>
M = <b, t, l, b>
N = <b, t, c>
O = <γ, ι >
P = <e, α >
Q = <ι,δ, u>
R = <e, r, β >
S = <κ, ε, ζ, ν >
T = <i, b>
U = < η >
V = <q, l>
W = <q, n, l, t>
X = <p, u>
Y = <q, l, b>
Z = <j, n, h>

Now the program compares two vectors of the same or different dimensions, i.e. the vector of one Arial letter with that of another Arial letter, and searches for the minimum difference. That is, it makes all comparisons that are possible by permuting the components of the vectors. For a vector with two vector components and a vector with four vector components, it makes 12 comparisons, and gives out the smallest one.

Let $V_g!$ be the factorial of the number of elements in the greater vector, and let $V_s!$ the factorial of the number of elements in the smaller vector. We then perform for each two Arial letters $\binom{V_g}{V_s} V_s! = \frac{V_g!}{(V_g - V_s)!}$ comparisons. So for example, for a vector of 4 elements with a vector of 2 elements, we obtain 4!/(4 – 2)! = 12 comparisons.

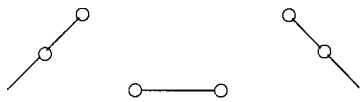
Connections have the same values as above, but they are taken into consideration with each line of the sign separately. That is, some connections can be counted two or three times, for example in "Y" the connecting point in the middle is taken for each stroke separately. The simplest way to analyze the difference is to decompose the character into its constituent parts. For the sake of illustration, let us compare Arial "A" and "D".

Table 1
Differences between the lines

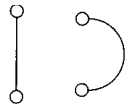
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	α	β	γ	δ	ε	ζ	η	θ	ι	κ	λ	μ	ν
a	0	2	2	4	4	6	2	4	4	4	6	3	5	5	5	4	3	3	5	5	4	9	11	7	9	7	7	5	11	7	6	6	8	6
b	2	0	4	2	2	4	4	6	6	6	8	5	7	5	5	6	5	5	7	7	6	11	13	9	11	9	9	7	13	9	8	8	10	8
c	2	4	0	2	6	4	4	6	6	6	8	5	7	7	7	6	5	5	7	7	6	11	13	9	11	9	9	7	13	9	8	8	10	8
d	4	2	2	0	2	2	6	8	8	8	10	7	9	9	9	8	7	7	9	9	8	13	15	11	13	13	13	9	15	11	10	10	12	10
e	4	2	6	2	0	2	4	8	8	8	10	7	9	9	9	8	7	7	9	9	8	13	15	11	13	13	13	9	15	11	10	10	12	10
f	6	4	4	2	2	0	4	10	10	10	12	9	11	11	11	10	9	9	11	11	10	15	17	13	15	13	13	11	17	13	12	12	14	12
g	2	4	4	6	4	4	0	6	6	6	8	5	7	7	7	6	5	5	7	7	6	11	13	9	11	9	9	7	13	9	8	8	10	8
h	4	6	6	8	8	10	6	0	4	4	2	5	7	7	7	6	5	5	7	7	6	11	13	9	11	9	9	7	13	9	8	8	10	8
i	4	6	6	8	8	10	6	4	0	4	2	5	7	7	7	6	5	5	7	7	6	11	13	9	11	9	9	7	13	9	8	8	10	8
j	4	6	6	8	8	10	6	4	4	0	2	5	7	7	7	6	5	5	7	7	6	11	13	9	11	9	9	7	13	9	8	8	10	8
k	6	8	8	10	10	12	8	2	2	2	0	7	9	9	9	8	7	7	9	9	8	13	15	11	13	11	11	9	15	11	10	10	12	10
l	3	5	5	7	7	9	5	5	5	5	7	0	2	2	6	5	5	5	7	8	6	11	13	9	11	9	9	7	13	9	8	8	10	8
m	5	7	7	9	9	11	7	7	7	7	9	2	0	4	4	3	7	7	9	10	8	13	15	11	13	11	11	9	15	11	10	10	12	10
n	5	5	7	9	9	11	7	7	7	7	9	2	4	0	2	7	7	7	9	10	8	13	15	11	13	11	11	9	15	11	10	10	12	10
o	5	5	7	9	9	11	7	7	7	7	9	6	4	2	0	3	7	7	9	10	8	13	15	11	13	11	11	9	15	11	10	10	12	10
p	4	6	6	8	8	10	6	6	6	6	8	5	3	7	3	0	6	6	8	9	7	12	14	10	12	10	10	8	14	10	9	9	11	9
q	3	5	5	7	7	9	5	5	5	5	7	5	7	7	7	6	0	4	6	2	5	11	13	9	11	9	9	7	13	9	8	8	10	8
r	3	5	5	7	7	9	5	5	5	5	7	5	7	7	7	6	4	0	2	2	5	11	13	9	11	9	9	7	13	9	8	8	10	8
s	5	7	7	9	9	11	7	7	7	7	9	7	9	9	9	8	6	2	0	4	3	13	15	11	13	11	11	9	15	11	10	10	12	10
t	5	7	7	9	9	11	7	7	7	7	9	8	10	10	10	9	2	2	4	0	7	13	15	11	13	11	11	9	15	11	10	10	12	10
u	4	6	6	8	8	10	6	6	6	6	8	6	8	8	8	7	5	5	3	7	0	12	14	10	12	10	10	8	14	10	9	9	11	9
α	9	11	11	13	13	15	11	11	11	11	13	11	13	13	13	12	11	11	13	13	12	0	2	2	5	7	7	6	2	12	7	7	9	7
β	11	13	13	15	15	17	13	13	13	13	15	13	15	15	15	14	13	13	15	15	14	2	0	4	3	9	7	6	2	10	7	7	9	7
γ	7	9	9	11	11	13	9	9	9	9	11	9	11	11	11	10	9	9	11	11	10	2	4	0	3	10	5	4	4	8	5	5	7	5
δ	9	11	11	13	13	15	11	11	11	11	13	11	13	13	13	12	11	11	13	13	12	5	3	3	0	10	8	7	7	11	8	8	10	8
ε	7	9	9	13	13	13	9	9	9	9	11	9	11	11	11	10	9	9	11	11	10	7	9	10	10	0	8	3	11	5	6	4	6	4
ζ	7	9	9	13	13	13	9	9	9	9	11	9	11	11	11	10	9	9	11	11	10	7	7	5	8	8	0	5	9	7	4	6	7	6
η	5	7	7	9	9	11	7	7	7	7	9	7	9	9	9	8	7	7	9	9	8	6	6	4	7	3	5	0	8	4	5	1	3	1
θ	11	13	13	15	15	17	13	13	13	13	15	13	15	15	15	14	13	13	15	15	14	2	2	4	7	11	9	8	0	12	9	9	10	2
ι	7	9	9	11	11	13	9	9	9	9	11	9	11	11	11	10	9	9	11	11	10	12	10	8	11	5	7	4	12	0	5	5	6	5
κ	6	8	8	10	10	12	8	8	8	8	10	8	10	10	10	9	8	8	10	10	9	7	7	5	8	6	4	5	9	5	0	6	7	6
λ	6	8	8	10	10	12	8	8	8	8	10	8	10	10	10	9	8	8	10	10	9	7	7	5	8	4	6	1	9	5	6	0	2	2
μ	8	10	10	12	12	14	10	10	10	10	12	10	12	12	12	11	10	10	12	12	11	9	9	7	10	6	7	3	10	6	7	2	0	2
ν	6	8	8	10	10	12	8	8	8	8	10	8	10	10	10	9	8	8	10	10	9	7	7	5	8	4	6	1	2	5	6	2	2	0

We shall indicate a continuous connection as "x", a crisp connection as "o" and a crossing as ●.

"A" contains 3 straight lines:



and "D" contains two lines (straight and arch):



Looking in the list of letter vectors above, we can perform all 6 comparisons using the results in Table 1:

od + kα + s	= 9 + 13 + 6	= 28
od + sα + k	= 9 + 13 + 6	= 28
kd + oα + s	= 10 + 13 + 6	= 29
kd + sα + o	= 10 + 13 + 6	= 29
sd + oα + k	= 9 + 13 + 6	= 28
sd + kα + o	= 9 + 13 + 6	= 28

Thus, in this example, we obtain a minimal result in four cases and two other, greater results. It would be possible to choose among the results using a principle-based procedure (e.g. from left to right and top-down), or attempting to "minimize" the difference, which is the approach we prefer here. If we decided that the position of the connection point is not relevant, there would be no difference between T and L; but if we distinguish between positions of connections, the difference is 8, which seems more realistic.

In order to illustrate the procedure using a simple example, we compare all upper case letters of the Latin alphabet written in Arial (cf. Table 1). The procedure to prepare for automatic computing is as follows. First, we prepare a list of all the straight lines and arches that occur in the alphabet. For Arial, we obtain twenty-one straight lines and thirteen kinds of curves, as shown above. We label the straight lines with lower case Latin letters (which are merely names!) and the curves with Greek letters. It is quite natural that, when processing the letters in this way, several properties are not taken into consideration, e.g. the length of the segments, their relative positions, and so on. This can be considered as a first approximation.

The comparison of individual Arial letters is shown in Table 2.

For genuine shape problems, other procedures should be found. It would be possible to devise a very simple procedure of putting the letters in a scheme with squares and compute the number of shared squares; or a purely geometrical procedure using the number of transformations of one letter into another. Alternatively it might be possible to ask persons who do not read the Latin script to classify the letters as to their similarity and derive a similarity measure (cf. a Campo et al. 1989); a very simple method is the computation of the fractal dimension of letters. There is a great number of such possibilities which will not be pursued here.

The distinctivity ($\bar{D}(i)$) of a character is the mean of its differences ($D_j(i)$) to other characters j. Let us define

(1) $$\bar{D}(i) = \frac{1}{K-1} \sum_{\substack{j \in P \\ j \neq i}} D_j(i)$$

where i is the given character, j is some other character of the inventory P of characters, K is the size of the inventory, and we compare each character with the other (K-1) characters. From the first line of Table 2 we obtain

$$\bar{D}_j(A) = [41+26+28+4+18+24+14+17+21+14+15+18+17+28+28+28+26+36+13+ \\ + 21 + 18+17+12+18+11] / 25 = \\ = 513 / 25 = 20.52$$

Table 2
Differences between Arial characters

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	0	41	26	28	24	18	24	14	17	21	14	15	18	17	28	28	27	26	36	13	21	18	17	12	18	11
B	41	0	24	13	30	28	27	32	24	23	32	26	34	32	22	13	25	15	26	26	23	31	43	33	30	37
C	26	24	0	17	28	22	6	22	10	5	22	16	24	22	10	17	18	21	12	16	5	16	28	18	20	22
D	28	13	17	0	21	17	21	23	11	12	23	13	21	19	13	2	18	8	26	13	12	18	30	20	17	24
E	24	30	28	21	0	6	30	16	18	23	20	12	20	21	30	21	34	24	38	16	23	22	28	24	18	19
F	18	28	22	17	6	0	24	12	12	17	16	10	16	15	24	15	28	18	32	10	17	16	25	18	12	13
G	24	27	6	21	30	24	0	20	14	11	25	18	27	25	15	21	20	24	16	18	11	19	31	21	23	20
H	14	32	22	23	16	12	20	0	12	17	14	10	19	17	24	21	28	24	32	10	17	16	25	18	16	15
I	17	24	10	11	18	12	14	12	0	5	12	6	14	12	12	11	16	17	20	6	5	7	19	9	10	13
J	21	23	5	12	23	17	11	17	5	0	17	11	19	17	9	12	16	16	15	11	0	11	23	13	15	17
K	14	32	22	23	20	16	25	14	12	17	0	15	10	13	24	21	27	19	32	15	17	10	15	12	10	15
L	15	26	16	13	12	10	18	10	6	11	15	0	17	11	18	17	22	20	26	8	11	10	22	12	13	11
M	18	34	24	21	20	16	27	19	14	19	10	17	0	10	26	21	29	19	34	13	19	12	12	18	8	17
N	17	32	22	19	21	15	25	17	12	17	13	11	10	0	24	19	27	17	32	11	17	11	14	17	7	18
O	28	22	10	13	30	24	15	24	12	9	24	18	26	24	0	13	8	19	18	18	9	18	30	20	22	24
P	28	13	17	2	21	15	21	21	11	12	21	17	21	19	13	0	18	6	26	13	12	18	30	20	17	24
Q	27	25	18	18	34	28	20	28	16	16	27	22	29	27	8	18	0	19	26	22	16	21	33	17	25	28
R	26	15	21	8	24	18	24	24	17	16	19	20	19	17	19	6	19	0	30	16	16	20	28	22	19	27
S	36	26	12	26	38	32	16	32	20	15	32	26	34	32	18	26	26	30	0	26	15	26	38	28	30	32
T	13	26	16	13	16	10	18	10	6	11	15	8	13	11	18	13	22	16	26	0	11	10	20	12	9	13
U	21	23	5	12	23	17	11	17	5	0	17	11	19	17	9	12	16	16	15	11	0	11	23	13	15	17
V	18	31	16	18	22	16	19	16	7	11	10	10	12	11	18	18	21	20	26	10	11	0	12	10	4	11
W	17	43	28	30	28	25	31	25	19	23	15	22	12	14	30	30	33	28	38	20	23	12	0	22	11	16
X	12	33	18	20	24	18	21	18	9	13	12	12	18	17	20	20	17	22	28	12	13	10	22	0	14	17
Y	18	30	20	17	18	12	23	16	10	15	10	13	8	7	22	17	25	19	30	9	15	4	11	14	0	13
Z	11	37	22	24	19	13	20	15	13	17	15	11	17	18	24	24	28	27	32	13	17	11	16	17	13	0

Thus the distinctivity of a character is composed of the difference between the complexities of lines plus place and weight of joining points, plus difference between the orientations of lines.

The distinctivity of all characters is presented in Table 3. In the left part of the second column, we give the sum of the differences, in the right part the mean distinctivity. The maximum distinctivity of Arial characters is 27.52 with B, the minimum distinctivity is 12.36 with letter I.

Table 3
The distinctivity of Arial characters

Character (i)	$\Sigma D(i)$	$\bar{D}(i)$	Character (i)	$\Sigma D(i)$	$\bar{D}(i)$	Character (i)	$\Sigma D(i)$	$\bar{D}(i)$
A	513	20.52	J	353	14.12	S	672	26.88
B	688	27.52	K	451	18.04	T	354	14.16
C	443	17.72	L	368	14.72	U	353	14.12
D	436	17.44	M	478	19.12	V	377	15.08
E	567	22.68	N	429	17.16	W	560	22.40
F	440	17.60	O	474	18.96	X	438	17.52
G	509	20.36	P	432	17.28	Y	395	15.80
H	473	18.92	Q	583	23.32	Z	473	18.92
I	309	12.36	R	490	19.60			

The letters can thus be placed in order of ascending distinctivity as follows:

I, J, U, T, L, V, Y, N, P, D, X, F, C, K, H, Z, O, M, R, G, A, W, E, Q, S, B.

but this order only holds for the upper case Arial characters. It is also questionable whether this order could be established intuitively by test subjects without recourse to complexity, which is here the basis of distinctivity. Nevertheless, from left to right one sees at least an increase in the number of straight lines and arches.

The *overall distinctivity of a given script* can be computed using the mean of means, i.e. as

(2) $$\bar{D}(Script) = \frac{1}{K(K-1)} \sum_{i \in P} \sum_{j \in P} D_j(i) = \frac{1}{K} \sum_{i \in P} \bar{D}(i).$$

For the Arial script this value is obtained by summing all the mean distinctivities and dividing the sum by 26 (= K). We obtain $\bar{D}(Arial)$ = 17.87.

2. The distinctivity of a script can be characterized in different ways using the above computation.

(1) The general mean is important, but so is the dispersion of differences. Some characters may deviate very strongly from the general trend and play the role of extreme values. But this is not the case with Arial.

(2) A more informative characteristic should be the frequency distribution of distinctivities themselves. A preliminary hypothesis would be that they follow the normal distribution; but since normality is seldom found with language phenomena, we may assume that the frequency distribution of distinctivities are different. Another possibility would be to assume a monotone decrease which warrants economy. But this is not the case either. Taking the distinctivities from 12.00 to 28.00 in two-step classes, we obtain

12-14 1
14-16 6
16-18 6
18-20 6
20-22 2
22-24 3
24-26 0
26-28 2

and it is easy to show (by means of a chi-square test) that this distribution is uniform. The chi-square test for homogeneity with the expected value E(D) = 3.25, and X^2 = 12.777, yielding with 7 degrees of freedom P = 0.0779 is preliminarily sufficient for stimulating us to search for other hypotheses with regard to the distribution of distinctivity.

3. A third possibility is the consideration of the vectors of individual letters. The components of the vectors represent different line types responsible for making up the letters. If a script is maximally distinctive then it uses each vector component only once. But as we see, some components are repeated in different letters. For example the element "b" occurs in M, N, T and Y, i.e. 4 times. Thus the frequency distribution of the occurrence of elements can be considered a characteristic of distinctivity. In case of maximal distinctivity, all elements occur only once, i.e. we have the deterministic distribution with $P(X = 1) = 1$, whose entropy is zero, whose repeat rate is 1 and whose excess is not defined (because the variance is zero). Minimal distinctivity is given if all elements (line types) occur with the same frequency. That is, we would obtain a discrete uniform distribution $P(X = x) = 1/n$ for $x = 1,2,\ldots,n$. This distribution has the maximal entropy $H = \mathrm{ld}\ n$, the minimal repeat rate $R = 1/n$ and its coefficient of excess is negative, i.e. $b_2 = \frac{\mu_4}{\mu_2^2} - 3 = -\frac{6}{5}\frac{n^2+1}{n^2-1}$. In empirical cases we shall obtain some intermediate values characterizing the distinctivity of the script.

In the case above, we calculate the distribution of vector elements as shown in Table 4.

Table 4
Distribution of vector elements

x	f_x
1	15
2	11
3	5
4	1
5	1
7	1
	34

There are 34 line types in $n = 6$ classes, out of which 15 occur once, and so on. The entropy of this distribution is $H = 1.903$, $H_{max} = \mathrm{ld}\ 6 = 2.585$, $H_{rel} = 0.3172$; the repeat rate is $R = 0.3235$ while that of the uniform distribution with $n = 6$ is 0.1667; and the coefficient of excess is $b_2 = 2.4404$.

Thus, these numbers measure the distinctivity of the Arial script. Using these measurements, the distinctivity of Arial can be compared with that of other scripts, or the development of the script could be studied. Other approaches to measuring distinctivity are not excluded.

For the time being the distribution of vector elements – which is positive Poisson – cannot be treated as definitive. Modern scripts are not result of randomness but of intellectual development. But in future, when several (non-Latin) scripts have been analyzed, we will be able to say whether there is a probability distribution in the background.

References

a Campo, F.W., Geršić, S., Naumann, C.L., Altmann, G. (1989). Subjektive Ähnlichkeit deutscher Laute. *Glottometrika 10, 46-70.*

Altmann, G. (2004). Script complexity. *Glottometrics 8, 68-73.*

Glottometrics 9, 2005, 54-61

Contextual relationships

Luděk Hřebíček, Prague[1]

Abstract: In this paper the principle of compositeness is used to analyse connections among language constructs united in a semantic system of a text. This principle is well-known in linguistics as Menzerath-Altmann's law. Its basic formula has the structure of a power law. General circumstances of this principle are applied to contextual relationships forming the semantic space of a text.

Keywords: compositeness, nucleus, Menzerath's law, betweenness, semantic field

1. Semantic space and language level

It appears to be accepted that the arrangement of lexical units in a text points to certain structures of assembled meanings in a human mind which is in contact with this text. These hidden structures of meanings are recognizable, with certain difficulties; they can hardly be clarified in detail when language expressions and their arrangements are neglected. Language expressions forming a text actually communicate the appearance of the meaning structures.

Let us assume that any text in a natural language opens and spans a semantic space that is structured equivalently with the arrangement of the units in a text. The ways in which texts are organized disclose certain specific characteristics of the meanings, at least their ability to be arranged into larger configurations, the aptitude to enlarge these formations, the capacity to be constituted of more elementary formations, etc. All these are forms in which human thinking and human behaviour moves.

Linguistics is faced with the problem of how to describe the structured semantic space of a text. The concept of the language level is generally accepted, but it requires a better scientific confirmation of its comprehension. Its abstract nature presupposes similarities between units at different levels and of different kinds (phones, morphs, syllables, word forms, lexical units, syntactic constructions – at least clauses and sentences) in subsuming them under the term "level". The first clue to a future solution was found by Paul Menzerath. The problem of the level was solved by Gabriel Altmann (1980), who introduced the general concepts of **language construct** and **constituent**. The so-called Menzerath-Altmann's law actually represents a definition of the language level.

Two aspects are in play in connection with this law:

(I) the traditional linguistic units of arbitrary neighbouring levels are in a mutual relationship that is prescribed by the law;

(II) so far undefined or unknown language constructs, if related in accordance with this law, represent a new language level.

When case (II) is justifiable, a new category of language units is exposed. It is thus evident that the concept "language construct" always implies a certain language level. Any individual language construct of a text (in fact, a language unit) belongs to a level of units;

[1] Send correspondence to: L. Hřebíček, Junácká 17, CZ-169 00 Prague 6, Czech Republic.
E-mail: hrebicek@orient.cas.cz.

constructs are units of a level. At the same time, constructs consist of units belonging to a level that is the immediately adjacent lower one in relation to the respective constructs.

2. Menzerath-Altmann's law and semantics

Altmann (1980) presents a formula characterizing the functional relationships between constructs and their constituents. Altmann derived this formula in the form of a power law:

(1) $y = Ax^{-b}$,

where x represents the size of the construct, y the mean size of its constituents and A and b are parameters.

Let us refer to the work Altmann, Schwibbe et al. (1989) in which aspect (I) mentioned above was investigated and Menzerath-Altmann's law proven for the traditional linguistic levels. More recently it has been proposed to select all occurrences of each lexical unit of a text, and all these occurrences together with their environments in a text are described as a language construct. Language constituents of these constructs are evident: they are individual occurrences of the lexical unit together with their environments. For example, a word in a sentence, together with that sentence. Each set of constituents is a construct formed around a given lexical unit.

Each lexical unit of a text produces a larger construct. Instead of "environment", the term **text segment** appears to be sufficiently precise and general. The theory operates with the types of segments belonging to aspect (I). And it has been proved that the lexical constructs[2] and their constituents behave in conformity with formula (1); they can be classified as belonging to aspect (II). In comparison with the formal levels of (I), this new level introduces an explicit semantic quality to the description of texts and languages. The advancement of the respective linguistic ideas can be found in such works as Hřebíček (1989, 1992, 1995, 1997, 2000), Hřebíček, Altmann (1993), Ziegler, Altmann (2002), Wimmer et al. (2003). This theoretical advancement is in accordance with the conception of linguistic synergetics formulated by Reinhard Köhler (1986, 1998).

Menzerath-Altmann's law asserts that the greater x is, the smaller y is. This means that, as the frequency of a given lexical unit increases, the mean size of the respective text segments decreases. If segments are, for example, sentences, with increasing frequencies of lexical units, the mean sentence length falls. Frequency on the semantic level thus equals the size of the construct x in the number of occurrences. The law explains the relationship between frequencies and the sizes of text segments as a specific function.

As an example we present data taken from a Turkish short story[3]; see Table 1. It is evident that in applications the denoted function has a statistical character. The ways in which data are obtained from texts, the length and character of the analyzed texts, as well as the proceeding of denotative analyses are described in the quoted works (see especially Ziegler, Altmann 2002). In word frequencies, only the presence of a lexical unit in a segment is recorded, regardless of whether it occurs in the segment more than once.

The intuitive reasons of such distributions of the sizes of constructs and constituents can be explained as a balance between frequencies and segment lengths. If frequency increases, a given lexical unit finds **semantic specifications** in a higher number of segments and the

---

[2] Instead of terms like *aggregation, aggregate, sentence aggregate* which occurred in the quoted works, G. Altmann proposes to designate lexical constructs *hrebs*.

[3] Necati Cumalı: İnsanlar Kardeştir. In: N. Cumalı: *Revizyonist.* Tekin, Istanbul 1979, 42-46.

semantic specifications inside individual segments can be reduced.

Table 1
The size of lexical constructs x and
their constituents y in a Turkish text

x	y	y^*
1	7.64	7.67
2	7.24	7.30
3	7.03	7.10
4	7.32	6.95
5	6.85	6.85
6	6.44	6.76
7	6.47	6.68
8	7.58	6.62
9	5.93	6.57
10	6.60	6.52

x = the number of occurrences of lexical units in different segments, i.e. frequency.
y = the mean size of segments (in words), in which lexical units occur.
$y^* = 7.67x^{-0.0707}$; it has been proven by Wilcoxon non-parametric test that between the distribution of observed y and computed y^* there is no significant difference.

3. The circumstances of the law

Menzerath-Altmann's law is valid under certain conditions which either are qualified as preliminary circumstances or as consequences. Both kinds of these circumstances are summarized into the following points:

(a) The intuitive understanding of the law at the highest text level indicates that an occurrence of a lexical unit in a segment means mutual **semantic specification** of the word units inside a text segment. Any lexical unit of a segment is specified by the other units. Clearly, this is one type of a contextual relationship.

(b) The domain of variable x in formula (1) is defined as $x = 1,2, \ldots$ When $x = 1$, then $y = A$. One of the two parameters is thus structurally defined, this quantity is a consequence of the structure of the law itself. Lexical units of a text fall into two basic semantic categories: the units having frequency 1, and the others. The category of **hapax legomena** is thus deduced as a consequence of the relation between constructs and their constituents on the highest text level.

(c) There are two categories of semantic joints: (i) inside a given segment and (ii) inside individual segments *and*, simultaneously, among these segments forming together a semantic construct.

(d) In connection with (c), two categories of context can be distinguished: **segmental** and **textual**. Semantic joints in a whole text, however, are accomplished through the joints inside its segments.

One substantial difference between the language constructs on the traditional levels (aspect I) and the constructs on the semantic text level is evident: While the former represents language units, in which text length can be measured, the latter forms a permeable structure giving no possibility of expressing a text length. On the other hand, the set of lexical units of a text is not only a text vocabulary, but also a set of the semantic constructs able to show the text structure when they are supplemented by the respective data.

4. Nucleus and periphery

The semantic structure of a text that is sought can be described with the help of the constructs having a frequency higher than 1. This part of the semantic-text structure can be understood as text's semantic **nucleus**. The other part can be treated as the semantic **periphery** of a text. Each peripheral unit is able to specify the units occurring in one segment only; they have restricted textual joints.

The ability to specify elements belonging to the nucleus is distributed in accordance with the frequency of individual lexical units. Consequently, semantic specification is a function of word frequency. The practical consequence of the distinction between periphery and nucleus consist in the possibility of reducing the whole semantic space to the nuclear units and their mutual relationships. Otherwise, a too large and complicated image of the semantic relationships between and among lexical units is obtained also in the case of shorter texts.

An arbitrary lexical unit, which is a member of a text nucleus, is related to the other units of the nucleus through their mutual collocations in different segments. The picture of these relationships can be obtained in the form of a 1-0 vector, line or column; for example:

(2) *Vector* (i) = {1, 0, 0, 1, 1, 0, 1, ...}

If lexical unit i co-occurs with another unit(s) in a segment and thus semantically specifies it, their relationship is noted down as 1; otherwise there is zero on the right-hand side of the vector. The lexical units are arranged in (2) into an arbitrary but fixed sequence of units for each i. The vector of i describes the relationships of a given unit to all other units. If a unit occurs in a segment more than once, only its presence (but not frequency inside the segment) is recorded.

4. Two kinds of matrices

When all the vectors, i.e. line and column vectors, of a nucleus are unified, a matrix is obtained in which the lines i and the respective columns j (i.e. those columns that concern the same lexical unit's relationships) are equivalent. This means that the sequence of zeros and ones are identical in line i and column j when $i = j$. The matrix is thus symmetrical, with its diagonal filled in exclusively by zeros.

As an example of such a matrix, the data concerning the semantic nucleus of a poetic text[4] are presented in Table 2. The lexical units of the nucleus are designated by rank numbers 1 to 8, zeros on the diagonal are substituted by dashes. The segments of this text are individual verses that coincide with the syntactic segmentation of this text.

If the lexical units of the matrix represent points/vertices and ones in the same matrix indicate connections, the matrix can be called **adjacency matrix** of an unoriented (the connecting lines are not arrows) graph, see Figure 1.

[4] This text is a ghazal in the Chagatay language written by Shibani, see Eckman (1966: 269).

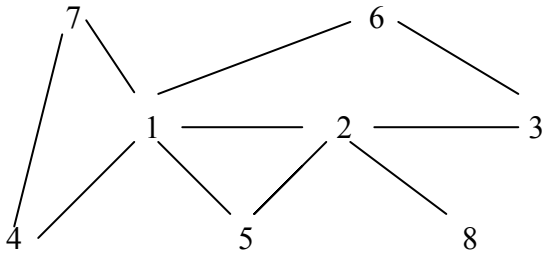


Figure 1. A graph equivalent to the adjacency matrix in Table 2

Table 2
1-0 matrix of the nucleus observed in a poetic text

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Sum**
1	-	1	0	1	1	1	1	0	5
2	1	-	1	0	1	0	0	1	4
3	0	1	-	0	0	1	0	0	2
4	1	0	0	-	0	0	1	0	2
5	1	1	0	0	-	0	0	0	2
6	1	0	1	0	0	-	0	0	2
7	1	0	0	1	0	0	-	0	2
8	0	1	0	0	0	0	0	-	1
Sum	5	4	2	2	2	2	2	1	20

The mutually equaling sums of columns and lines of the matrix indicate the values of a variable called **vertex degree**. Each edge points to two different vertices, therefore the total number of edges equals the sum of all vertex degrees divided by 2.

The adjacency matrix is a basic description of the semantic space of a text. The complete image of this space also contains hapax legomena, but their location in a segment can easily be read from the protocol of a respective text analysis; this protocol is the source of the structure described in the matrix and graph in a reduced way as a text nucleus. The relationships described in the matrix, however, concern only the connections of the lexical units inside the segments. Menzerath-Altmann's law indicates that the textual relationships are accomplished through the segmental relationships. The basic degree of contextuality is supplemented by a higher degree operating in the whole text. It is built up on the connections described in the matrix, therefore the higher connections must also be present in this matrix. It should be extracted from this matrix in order to obtain the semantic space of the whole text.

For this purpose it is useful to apply the concept of **betweenness** or **betweenness centrality** of a graph. For a complete characterization of this see Brandes (2001). Betweenness is based on the possibility of moving from a given vertex to another along the existing vertices and edges. Such a movement is called the **path** of a graph. According to Brandes, "Many centrality indices are based on shortest paths linking pairs of actors, measuring, e.g., the average distance from other actors, or the ratio of shortest paths an actor lies on... As a remedy, network analysts are now suggesting simpler indices, for instance based only on linkages between the neighbors of each actors... to at least obtain rough approximations of betweenness centrality." The aim is to find vertices ("actors") determining the structural prominence of a network.

All pairs of vertices are analyzed from the viewpoint of such connections. If between two

vertices more than one path exists, only the **shortest path** (SP) is chosen for the purpose of characterizing the betweenness centrality of a graph. The length of a path is declared in the number of edges contained in one SP. The maximum length of an SP and the distribution of the values of this variable can be used as a characteristic of the respective graph and space structure proper to a text.

Two matrices in Table 3 exemplify how to obtain a new matrix that is derived from 1-0 matrix in Table 2. The final matrix presents an image of both contextual degrees, segmental and textual.

Table 3A is a matrix containing SPs with maximal length 2. It contains, however, four zeros that can be also changed into an SP length. Table 3B is the desired final matrix of all SPs. The final SP-matrix contains $(8^2 - 8)/2 = 28$ pairs of vertices; the sum of all the lengths of SP equals 51. Therefore the mean SP-length is 51/28 = 1.82 (in the number of edges). The graph, to which the both kinds of matrices are related, remains unchanged, therefore the edges described by ones remain in the final matrix in the same positions as in the original matrix; here, however, 1 means a path length.

Table 3
Obtaining an SP matrix $M^{(D)}$ from the 1-0 matrix M (see Table 2)

A:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	-	1	2	1	1	1	1	2
2	1	-	1	2	1	2	2	1
3	2	1	-	0	2	1	0	2
4	1	2		-	2	2	1	0
5	1	1	2	2	-	2	2	2
6	1	2	1	2	2	-	2	0
7	1	2	0	1	2	2	-	0
8	2	1	2	0	2	0	0	-

B:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	-	1	2	1	1	1	1	2
2	1	-	1	2	1	2	2	1
3	2	1	-	3	2	1	3	2
4	1	2	3	-	2	2	1	3
5	1	1	2	2	-	2	2	2
6	1	2	1	2	2	-	2	3
7	1	2	3	1	2	2	-	3
8	2	1	2	3	2	3	3	-
Sum*	**9**	**9**	**11**	**8**	**6**	**5**	**3**	**0**

Sum* indicates the sum of the numbers under the diagonal.

Each value $a_{i,j} = 1$, $i \neq j$, of M remains at the same field (i, j) of $M^{(D)}$.

When the number of vertices increases, the complexity of obtaining an SP-matrix from a 1-0 matrix increases rapidly, and it becomes necessary to employ a computer program. The standard algorithms for SP are discussed in Cormen et al. (1990). Let us present here a simple algorithm that can be directly applied to simple graphs. From the starting 1-0 matrix M we aim at an SP matrix $M^{(D)}$ according to the following points:

1. Variable D is defined as D = max $a_{i,j}$ of $M^{(D)}$.
2. If in M there is an $a_{i,j} = 0$ and at the same time it holds that $a_{i,j=k} = 1$ and $a_{i=k,j} = 1$, then $M^{(2)}$ is obtained with $a_{i,j} = 2$; otherwise $a_{i,j} = 0$ of $M^{(2)}$.
3. The state described in point 2 can be repeated for an arbitrary $D > 2$.

When an SP-matrix $M^{(D)}$ is obtained with some $a_{i,j} = 0$ which cannot be changed into an SP-length, then there is no SP between i and j; the given zero must be substituted by a dash like the zeros at the diagonal of the original matrix M. The reason is evident: all the values of final $M^{(D)}$-matrix are lengths of SP's, i. e., paths or distances between the respective pairs of vertices. "Zero" length would mean an infinite distance. It can be proven that in this case i and j are vertices of two mutually unconnected subgraphs of the described semantic graph. Each such "zero-dash" indicates one unconnected subgraph of the graph. When the mean shortest path is computed, these dashed fields must be subtracted from the total number of fields of the matrix similarly to the dashed fields of the diagonal. Consequently, the number of vertex pairs in an arbitrary $M^{(D)}$ is

$$(3) \quad s = \frac{n^2 - n - c}{2} ,$$

where n is the number of vertices and c the number of unconnected subgraphs ("dashed zeros"). The graph on which the matrices exemplified above are based is completely connective; therefore no dashes are present in the respective matrix $M^{(D)}$.

Analogous results have been obtained from analysing the first chapter of *Pride and prejudice* by Jane Austen. This prosaic text has a semantic nucleus containing 75 vertices and 2775 pairs of vertices. Its SP-matrix $M^{(D)}$ is characterized by the value of D = 4, which is the maximal length of SP in this text. No "dashed zero" occurs in this matrix (i. e., $c = 0$) which means that the whole graph of the text nucleus is connective. The sum of all its SP's equals 5767. The average SP-length is thus 2.08.

The data observed in this text are sufficiently large for testing the hypothesis that the vertex degrees and the generalized distance of a vertex to all other vertices are uncorrelated variables. This hypothesis can be rejected, the estimation of the observed rank correlation coefficient $R = -0.8112$ appears to be significant for $n = 75$.

As for shorter poetic texts in the same corpus as that exemplified above, the estimated coefficient $R = -0.1398$ is insignificant. The majority of similar texts show insignificant correlations.

It can be conjectured that there are text categories in which textual relationships are not closely dependent on the position of lexical units in their segments, i.e., on segmental relationships.

5. Summary

Text structure can be regarded as a semantic phenomenon depending on the principle of compositeness. To become a constituent of a construct means in the dynamics of meanings to belong to a set of constructs forming a whole that can be communicated as a text. "To depend on, follow on, hang on" are words pointing to one and the same relationship that can be dissociated according to its location(s) in a larger structure of constructs and its quantitative characteristics.

Lexical units, their sequential segments and their image in a 1-0 matrix represents a starting point for obtaining a new matrix giving a certain description of the whole, into which lex-

ical units enter when the respective whole is arranged into a system. In the present paper, only the relationships, their presence or absence, are discussed. Each relationship, however, is differentiated according to their intensity revealing in repetitions of the relationships (and not only of lexical units). This approach is applicable to larger texts submitting data for deeper statistical analyses.

The connection between Menzerath-Altmann's law and the graph of a text as its semantic space is, this paper proposes, described by two matrices. One of them reflects the segmental level, i.e. the level of the semantic constituents, and the other one reflects the textual level that corresponds to semantic constructs.

It can be concluded that the two kinds of relationship implied by the formula of Menzerath-Altmann's law, and its pointing to the existence of the two kinds of relationships in a text, segmental and textual, can be found with precision by means of a description based on graphs and networks. In the history of linguistics, the concept of **semantic field** has been coined for meanings. If a meaning is considered from the viewpoint of text structures, three types of semantic fields can be defined for meanings assumed as mental items:

1. segmental field, which is relevant for some close connections of meanings able to be revealed in a text segment;
2. textual field, in which the dynamics of a meaning is exposed in a Menzerathian relationship of a constituent (text segment) and construct (text);
3. supratextual field; any meaning of this field depends on a certain individual semantic system forming the mind of an individual; it is ready at a given moment to be activated into a segmental and textual field.

The actual position of a meaning in a mind can be recognized from texts.

References

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

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

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of mathematical sociology 25(2)*, *163-177.*

Cormen, T.H., Leiserson, C.E., Rivest, R.L. (1990). *Introduction to algorithms.* MIT Press.

Eckmann, J. (1966). *Chagatay manual.* The Hague: Mouton.

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

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

Hřebíček, L. (1995). *Text levels. Language constructs, constituents and Menzerath-Altmann's law.* Trier: Wissenschaftlicher Verlag Trier.

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

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

Hřebíček, L., Altmann, G. (eds.) (1993**).** *Quantitative text analysis.* Trier: WVT.

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

Köhler, R. (1998). Elemente der Synergetischen Linguistik. *Glottometrika 12, 179-187.*

Wimmer, G., Altmann, G. Hřebíček, L., Ondrejovič, S., Wimmerová, S. (2003). *Úvod do analýzy textov.* [Introduction to the analysis of texts.] Bratislava: Veda.

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

Glottometrics 9, 2005, 62-73

Häufigkeiten von Buchstaben / Graphemen / Phonemen: Konvergenzen des Rangierungsverhaltens

Peter Grzybek, Graz[1]
Emmerich Kelih, Graz[2]

Abstract. The present study raises the question in how far low-level linguistic units, such as letters, graphemes, sounds and phonemes, follow one and the same pattern as to their frequency distribution. Based on Altmann/Lehfeldt's (1980) study on 63 samples from 38 different languages, a separate re-analysis of the letter/grapheme vs. sound/phoneme samples is made, concentrating on the empirical entropy and repeat rate, on the one hand, and their theoretical calculations derived from the geometric and Zipf-Mandelbrot distributions. As a result, there are no significant differences as to these two global measures. This finding is interpreted in terms of a strong argument in favor of an analogical behavior of these linguistic units.

Keywords: letter frequencies, grapheme frequencies, phoneme frequencies, entropy, repeat rate, geometric distribution, Zipf-Mandelbrot distribution

1. Einführung

In der vorliegenden Untersuchung geht es um die Frage, ob sich die „niedrigsten" Spracheinheiten – nämlich Phoneme, Buchstaben, Grapheme – hinsichtlich ihrer Häufigkeitsverteilung analog verhalten, d.h. ob man für ihre Erfassung die gleichen Modelle benutzen kann. Hintergrund dieser Fragestellung ist der Umstand, dass Untersuchungen zur Vorkommenshäufigkeit von Buchstaben bzw. Graphemen in der jüngsten Zeit eine gewisse Renaissance zu erleben scheinen (s. z.B.: Best 2005a,b; Grzybek, Kelih, Altmann 2004, 2005a,b; Hussien 2004; Pääkkönen 1993; Rosenbaum, Fleischmann 2002, 2003); auch im Internet finden sich mittlerweile zu den verschiedensten Sprachen[3] Angaben zu Buchstabenhäufigkeiten, wobei allerdings nicht selten die einfachsten Voraussetzungen wissenschaftlichen Arbeitens verstoßen wird: So wird oft überhaupt nicht gesagt, auf welchem Material die jeweiligen Angaben beruhen, es werden keine absoluten, sondern nur relative Häufigkeiten angegeben (die für eine Reihe von weiterführenden Fragen unzureichend sind), usw. usf.

Beim Thema von Graphem- oder Phonemhäufigkeiten handelt es sich zweifellos um eine der traditionellsten Fragestellungen in der Geschichte der Quantitativen Linguistik, die in der Vergangenheit immer wieder mit den unterschiedlichsten pragmatischen Fragestellungen verbunden oder gar auf diese hin ausgerichtet war (vgl. Grzybek, Kelih 2003). So ging es in den wenigsten Untersuchungen um die einfache Erhebung von Buchstabenhäufigkeiten an und für

---

[1] Send correspondence to: Peter Grybek: peter.grzybek@uni-graz.at.
[2] Der Beitrag von Emmerich Kelih verdankt sich u.a. dem Doktorandenprogramm der Österreichischen Akademie der Wissenschaften (DOC).
[3] Auf den Versuch einer vollständigen Anführung von konkreten Angaben kann hier verzichtet werden: In der Regel handelt es sich hier um Internetseiten aus dem Bereich der Kryptographie bzw. Kryptologie; vgl. z.B.: http://www.cryptogram.org/cdb/words/frequency.txt#calcs, http://www.central.edu/homepages/LintonT/classes/spring01/cryptography/letterfreq.html

sich; vielmehr waren praktisch mit allen Studien immer auch weiterführende Fragen verbunden, angefangen von mathematischen und methodologischen Problemen, über Fragen der Optimierung technischer Einrichtungen oder der Strukturierung von Codes und Prozessen der Informationsübertragung, die Relevanz ihrer Untersuchungen von Graphemhäufigkeiten für Fragen der Stenographie, der Tastaturbelegung von Schreibmaschinen bis hin zu Fragen der Textstilistik und Texttypologie, oder des Vergleichs mit der Phonologie.

Während in der Vergangenheit allerdings das Interesse bei der Untersuchung von Graphemhäufigkeiten in der Regel eher auf die relative Häufigkeit des Vorkommens einzelner Grapheme ausgerichtet war, wurde in jüngerer Zeit vermehrt die Frage nach einem allgemeinen Häufigkeitsmodell gestellt (Sigurd 1968; Orlov et al. 1982; Tuldava 1988; Naranan, Balasubrahmanyan 1992a,b; Martindale et al. 1996; Altmann 1993; Grzybek, Kelih, Altmann 2004, 2005a,b). Bei dieser Art von Fragestellung geht es weniger um die Frequenz der individuellen Buchstaben, sondern darum, welchen (relativen) Anteil das jeweils häufigste Graphem im Vergleich zum zweithäufigsten, zum dritthäufigsten, usw. hat. In den Vordergrund rückt damit eine Rang-Häufigkeitsverteilung, und das Ziel der theoretischen Modellierung ist die mathematische Formalisierung des Abstands zwischen den jeweiligen Häufigkeiten. Das Vorgehen hat man sich dabei wie folgt vorzustellen: Überführt man erhobene Ausgangsdaten in eine Rang-Reihenfolge, so geschieht dies üblicherweise in absteigender Reihenfolge. Wenn man sodann die jeweiligen Datenpunkte miteinander verbindet, ergibt sich charakteristischerweise kein linearer Abfall, sondern eine spezifische, monoton fallende (üblicherweise hyperbolische) Kurve. Und genau darum ist es in den genannten Untersuchungen gegangen: nämlich die genaue Form dieser Kurve zu modellieren, um so zu sehen, ob die Häufigkeiten in verschiedenen Stichproben (d.h. die spezifische Abnahme der Häufigkeiten) ein und dieselbe Form aufweisen oder nicht.

Dabei haben sich zwei Arten der Modellierung entwickelt: mit Hilfe von stetigen Funktionen oder diskreten Folgen und mit Hilfe von Wahrscheinlichkeitsfunktionen. Letztere unterscheiden sich von ersteren dadurch, dass sie normiert sind, d.h. dass die Summe der Wahrscheinlichkeiten 1 ergibt. Rein empirisch gesehen – bezüglich der Anpassungsgüte – haben reine Kurven einige Vorteile, theoretisch gesehen – nämlich bezüglich der Systematisierung – sind hingegen die Verteilungen adäquater.

Bei mehreren in der jüngsten Zeit durchgeführten Studien zu Graphemhäufigkeiten in slawischen Sprachen hat sich allerdings herausgestellt, dass in der Vergangenheit diskutierte und in der Gegenwart in der Regel immer noch in Betracht gezogene Modelle nicht passen, wenn man deren Anpassung an systematisch kontrolliertes Datenmaterial statistisch prüft. Demnach sind solche gängigen Modelle[4] wie etwa die geometrische Verteilung (Sigurd 1968) die Good-Verteilung (Martindale et al. 1996) und die Zipf'sche bzw. Zipf-Mandelbrot'sche Verteilung (vgl. Mandelbrot 1953) als Standardverteilung von Rangfrequenzen, die Waring-Verteilung oder die Whitworth-Verteilung (vgl. Whitworth 1901: 207f., Grzybek, Kelih, Altmann 2004) als passende Modelle oft nur lokal geeignet; die meisten nur in Kurvenform vorhandenen Modelle hingegen sind entweder nicht normiert, rechts nicht begrenzt oder ohne Interpretation der Parameter (vgl. Grzybek, Kelih, Altmann 2005a,b). Zumindest in den untersuchten slawischen Sprachen erweist sich als passendes Modell hingegen die negative hypergeometrische Verteilung (vgl. Grzybek, Kelih 2005; Grzybek, Kelih, Altmann 2004, 2005a,b).[5] Zwar handelt es sich bei dieser Verteilung um ein relativ komplexes Modell, dass nicht weniger als drei Parameter aufweist (n, K, M), doch haben sich im Zusammenhang mit den genannten Untersuchungen zunehmende Anhaltspunkte dafür ergeben, dass die Parameter K und M der negativen hypergeometrischen Verteilung sich letztendlich allein auf den Inven-

[4] Zur detaillierten Darstellung der einzelnen Verteilungsmodelle vgl. Wimmer/Altmann (1999).

[5] Auch für das Deutsche hat Best (2005a, 2005b) dieses Modell ins Spiel gebracht.

tarumfang n zurückführen lassen, wenn auch nicht auf direkte Art und Weise.

Es würde zu weit führen, diese Zusammenhänge hier im Detail zu diskutieren, weswegen auf die entsprechenden Arbeiten verwiesen sei. Allerdings haben sich aus dieser Beobachtung heraus weiterführende Überlegungen ergeben, die darauf abzielen, über das lokale Kriterium der in der Regel über den Chi²-Test geprüften Anpassungsgüte eines Modells hinausgehend auch globale Kriterien zu untersuchen, in erster Linie Kenngrößen wie die Entropie H oder die Wiederholungsrate R. Beide Kenngrößen sind asymptotisch ineinander überführbar (vgl. Altmann, Lehfeldt 1980), denn

$$H \approx ld\, n - \frac{nR-1}{2\ln(2)} \quad \text{und } R \approx \frac{2(ld\, n - H)\ln(2)+1}{n}.$$

Versuchsweise sollte man aber zunächst beide ableiten. Der Sinn, diese globalen Kriterien zu bestimmen, liegt darin, dass man zusätzlich zum Kriterium der lokalen Eignung weiteres Adäquatheitskriterium erhält. Erweist sich ein globales Maß einer Verteilung als sehr allgemeingültig, dann kann man es als Modell auch dort verwenden, wo die Anpassungsgüte nicht besonders befriedigend ist. In solchen Fällen sucht man nach Gründen der Idiosynkrasie, z.B. nach Randbedingungen, wählt Spezialfälle oder Grenzfälle, modifiziert das grundlegende Modell oder sagt die weitere Entwicklung voraus.

Die von Herdan (1962: 36ff., 1966: 271ff.) in die Linguistik eingeführte Wiederholungsrate R [repeat rate, Herfindahlsches Konzentrationsmaß] ist eines der einfachsten globalen Charakteristiken einer Häufigkeitsverteilung, definiert als

$$(1) \qquad R = \sum_{k=1}^{n} p_k^2 .$$

Die Größe R bewegt sich im Intervall $(1/n;1)$; sie erreicht den größten Wert, wenn ein einziges der vorkommenden Elemente die Wahrscheinlichkeit $p = 1$ hat und alle anderen Wahrscheinlichkeiten gleich Null sind; den kleinsten Wert nimmt R für den Fall an, dass alle Wahrscheinlichkeiten gleich sind. Daraus folgt, dass R ein Maß der Gleichverteilung ist, das desto kleiner wird, je ähnlicher die einzelnen Häufigkeiten sind. Auch die als

$$(2) \qquad H = -\sum_{i=1}^{n} p_i \cdot \mathrm{ld}\, p_i$$

definierte Entropie H lässt sich als ein Maß des Gleichgewichts bzw. der Gleichverteilung verstehen. Hier gilt, je größer H, desto ähnlicher die einzelnen Wahrscheinlichkeiten. Das Maximum (ld n) erreicht H dann, wenn alle Häufigkeiten gleich sind, das Minimum (0) hingegen, wenn eine der Wahrscheinlichkeiten $p_k = 1$. Da sich H somit im Intervall (0; ld n) bewegt, hängt sein Wert vom Inventarumfang ab, weswegen man zum Zwecke der Standardisierung auch die relative Entropie berechnet, die Werte im Intervall (0;1) annehmen kann.

$$(3) \qquad h = \frac{-\sum_{k=1}^{n} p_k \cdot \mathrm{ld}\, p_k}{\mathrm{ld}\, K}$$

Von diesen Überlegungen ausgehend, haben Altmann/Lehfeldt (1980) zunächst die (empirischen) Wiederholungsraten und Entropien für 63 Datensätze aus 38 verschiedenen Sprachen berechnet. Im Detail handelte es sich dabei um 26 Datensätze (von 23 verschiedenen

Sprachen), die auf Buchstaben- bzw. Graphemhäufigkeiten beruhten, und 37 Datensätze (von 27 verschiedenen Sprachen), die auf Phonemhäufigkeiten beruhten. Ohne zwischen Daten aus Buchstaben- bzw. Graphemhäufigkeiten und Daten aus Phonemhäufigkeiten zu differenzieren, haben die Autoren die erhaltenen Werte im Hinblick auf ein allgemeines Modell reflektiert. Da die vorliegende Abhandlung an diese Überlegungen anknüpft, bietet es sich an, die entsprechenden Daten sowohl in tabellarischer Form (vgl. Tab. 1) als auch in graphischer Form wiederzugeben; auf eine Darlegung der Quellen kann dabei im hier gegebenen Zusam-

Tabelle 1
Entropien und Wiederholungsraten von 63 Sprachen

Nr.	**Sprache**	***n***	***R***	***H***	***h***	**Nr.**	**Sprache**	***n***	***R***	***H***	***h***
1	Hawaiisch	13	0,120716	3,370804	0,910920	**33**	Ungarisch	32	0,053986	4,508421	0,901684
2	Hawaiisch B	13	0,131031	3,238512	0,875170	**34**	Ungarisch B	32	0,052090	4,544989	0,908998
3	Samoanisch	15	0,118269	3,475854	0,889673	**35**	Khasi	32	0,062487	4,468863	0,893773
4	Hawaiisch	18	0,105342	3,567112	0,855438	**36**	Lettisch B	32	0,056568	4,428232	0,885646
5	Pilipino	21	0,102547	3,820317	0,869773	**37**	Russisch B	32	0,056880	4,453237	0,890647
6	Pilipino	21	0,116018	3,725076	0,848089	**38**	Deutsch	33	0,059241	4,443530	0,880887
7	Kaiwa	21	0,080747	3,947756	0,898787	**39**	Georgisch	33	0,070258	4,293132	0,851070
8	See-Dajakisch B	21	0,091596	3,878690	0,883062	**40**	Georgisch	33	0,068390	4,310649	0,854542
9	Estnisch B	23	0,070229	4,086146	0,903033	**41**	Ostjakisch	33	0,062470	4,375786	0,867455
10	Suaheli B	24	0,087816	4,023958	0,877643	**42**	Ostjakisch	33	0,066969	4,395138	0,871292
11	Französisch B	24	0,079699	3,963404	0,864435	**43**	Ostjakisch	34	0,066856	4,406486	0,866146
12	Albanisch B	25	0,063719	4,217626	0,908216	**44**	Ostjakisch	34	0,064061	4,390940	0,863090
13	Indonesisch B	25	0,083864	3,928940	0,846051	**45**	Tschechisch	35	0,046491	4,700600	0,916424
14	Chamorro	25	0,074272	4,214724	0,907591	**46**	Tschechisch B	35	0,043964	4,722755	0,920744
15	Holländisch B	26	0,077732	4,085845	0,869247	**47**	Französisch	35	0,070591	4,101787	0,799680
16	Englisch B	26	0,062183	4,215092	0,896744	**48**	Marathi	38	0,060408	4,514403	0,860226
17	Rumänisch	27	0,062323	4,253963	0,894651	**49**	Bengali	38	0,065865	4,863256	0,926700
18	Spanisch B	27	0,074637	4,023919	0,846270	**50**	Ungarisch	39	0,052800	4,602810	0,870853
19	Hausa B	27	0,105588	3,922669	0,824976	**51**	Englisch	39	0,050495	4,709796	0,891095
20	Holländisch B	28	0,082022	4,048855	0,842221	**52**	Armenisch B	39	0,070748	4,433971	0,838909
21	Serbokroatisch B	29	0,063378	4,287102	0,882486	**53**	Russisch	41	0,050003	4,825688	0,900726
22	Bulgarisch B	29	0,067229	4,219317	0,868533	**54**	Polnisch	42	0,050928	4,725240	0,876291
23	Deutsch B	29	0,072007	4,172738	0,858945	**55**	Englisch	42	0,040990	4,918028	0,912044
24	Indonesisch	29	0,085824	4,094239	0,842786	**56**	Gujarati B	43	0,055151	4,664619	0,859637
25	Indonesisch	29	0,060762	4,348654	0,895157	**57**	Englisch	44	0,043749	4,906418	0,898705
26	Deutsch B	30	0,073938	4,147517	0,845243	**58**	Slovakisch	44	0,048530	4,731830	0,866726
27	Gujarati	30	0,055911	4,497795	0,916628	**59**	Schwedisch	45	0,043407	4,840576	0,881410
28	Italienisch B	30	0,074775	4,006747	0,816556	**60**	Ukrainisch	46	0,049402	4,627951	0,837856
29	Italienisch	31	0,067565	4,251224	0,858106	**61**	Hindi	52	0,054557	4,584604	0,804254
30	Ukrainisch B	31	0,049725	4,565024	0,921446	**62**	Burmanisch B	68	0,039244	5,230643	0,859248
31	Russisch B	31	0,057713	4,433604	0,894919	**63**	Vietnamesisch	74	0,047003	5,160549	0,831079
32	Amer. Englisch	32	0,054875	4,487366	0,897473						

menhang verzichtet werden (vgl. Altmann/Lehfeldt 1980: 154ff.).

Tab. 1 enthält die in der von Altmann/Lehfeldt (1980: 154ff.) angegebenen Reihenfolge Werte für die einzelnen Sprachen, wobei es sich bei den mit *B* bezeichneten Sprachen um Buchstaben- bzw. Graphemhäufigkeiten handelt. *K* bezeichnet die jeweilige Inventargröße, *R* ist der Wert für die Wiederholungsrate, *H* ist der Wert der absoluten, *h* derjenige der relativen Entropie. Man kann leicht sehen, dass sich die relative Entropie *h* in einem sehr schmalen Intervall von $0.7997 \leq h \leq 0.9267$ bewegt, was als ein weiteres Indiz für einen Zusammenhang mit dem Inventarumfang zu interpretieren ist.

Das Interesse von Altmann/Lehfeldt (1980) war es nun, an diese Daten ein allgemeines Modell anzupassen; dieses sollte nach Möglichkeit interpretierbar sein, womit im hier gegebenen Zusammenhang gemeint ist, dass das Modell letztendlich auf die Inventargröße zurückgeführt und in diesem Sinne erklärt werden kann. Zwar zogen die Autoren es auch in Betracht, zum Zwecke des Vergleichs ein Regressionsmodell des Typs $y = ax^{-b}$ anzupassen; doch ist bei diesem Modell keine Option erkennbar, wie die Parameter *a* und *b* auf den Inventarumfang projiziert und somit interpretiert werden könnten.

Deshalb haben Altmann/Lehfeldt – ausgehend von der geometrischen Verteilung – die sich aus dieser Verteilung ergebenden theoretische Entropien und Wiederholungsraten berechnet. Dabei waren die Autoren von der Idee geleitet, dass im Ergebnis ein geeignetes Modell zur Verfügung stehen würde, das theoretisch begründet wäre und über eine bestimmte Erklärungskraft verfügen würde.

Auf die mathematische Ableitung der theoretischen Entropie *E(H)* und Wiederholungsrate *E(R)* muss hier nicht im Detail eingegangen werden (vgl. Altmann/Lehfeldt 1980: 151ff.; Grzybek/Kelih/Altmann 2005c). Es mag ausreichend, dass sich die entsprechenden aus der geometrischen Verteilung ergebenden Werte für die Wiederholungsrate über die Formel

$$E(R) \approx \frac{2}{n} \tag{4}$$

berechnen und für die Entropie über die Formel

$$E(H) \approx -\mathrm{ld}\left[\left(\frac{4}{n+2}\right)\left(\frac{n-2}{n+2}\right)^{\frac{n-2}{4}}\right] \tag{5}$$

approximieren lassen.

Wie eine Re-Analyse der 63 Datensätze zeigt, stellt sich im Ergebnis für die Wiederholungsrate heraus, dass das Regressionsmodell gemäß der Formel $y = 0.9659x^{-0.7822}$ mit einen Determinationskoeffizienten von $R^2 = 0.7692$ eine bessere Anpassungsgüte auf weist als das sich aus der geometrischen Verteilung ergebende Modell ($R^2 = 0.6871$). In Abb. 1 sind die beobachteten Werte der Wiederholungsraten in Form von Datenpunkten markiert; enthalten sind auch die beiden Anpassungskurven: die durchgehende Kurve ist die Potenzkurve, die grob gestrichelte die sich aus der geometrischen Verteilung ergebende. Ebenfalls zu sehen ist eine weitere (fein gestrichelte) Linie. Hierbei handelt es sich um eine ebenfalls theoretisch begründete, sich aus der Zipf-Mandelbrot-Verteilung ergebende Kurve, die Zörnig/Altmann (1983) abgeleitet haben, da die sich aus der geometrischen Verteilung ergebenden theoretischen Werte der Wiederholungsrate zu keinen befriedigenden Ergebnissen geführt hatte. Wie die auf der entsprechenden Formel

(6) $$E(R) = \frac{1.61^{-1} - (0.61+n)^{-1}}{\left(\ln \frac{0.61+n}{1.61}\right)^2}$$

beruhende Re-Analyse der o.a. 63 Datensätze zeigt, sind die Ergebnisse mit einem Wert von $R^2 = 0.7427$ deutlich besser als die sich aus der geometrischen Verteilung ergebenden.

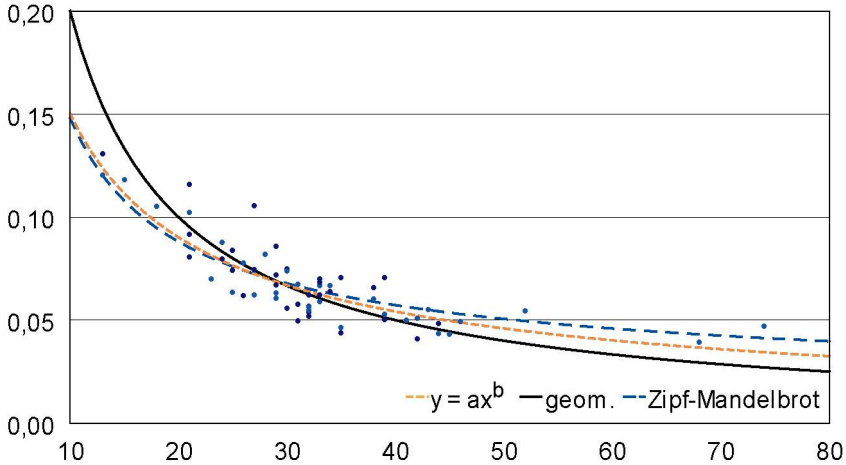


Abb. 1. Empirische und theoretische Wiederholungsraten für 63 Datensätze verschiedener Sprachen nach Altmann/Lehfeldt (1980)

Ein analoger Befund ergibt sich für die Berechnung der Entropie: Folgt man hier dem Modell $y = 1.7290x^{0.2663}$, so beträgt der Determinationskoeffizient $R^2 = 0.8538$; gemäß dem sich aus der geometrischen Verteilung ergebenden Modell (4) ist der Wert ebenso wie bei der Wiederholungsrate bei $R^2 = 0.7928$ deutlich schlechter. Abb. 2 stellt den Befund anschaulich dar, die Form der Darstellung entspricht der in Abb. 1. Auch Abb. 2 enthält zusätzlich die sich aus der Zipf-Mandelbrot-Verteilung ergebenden theoretischen Entropiewerte, die nach der von Zörnig/Altmann (1984) abgeleiteten Formel

(7) $$E(H) = \mathrm{ld}\, e \ln\left[\sqrt{(B+n)(B+1)} \ln \frac{B+n}{B+1}\right]$$

mit $B = 0.61$ zu einem Wert von $R^2 = 0.8371$ führt, der ebenfalls besser ist als der sich aus der geometrischen Verteilung ergebende.

In beiden Fällen ist das theoretisch begründbare Modell im Vergleich zu dem rein rechnerisch optimierten somit das – relativ gesehen – schlechtere. Doch hat das aus der Potenzformel gewonnene Modell den entscheidenden Nachteil, dass es rein empirisch aus den konkreten Daten gewonnen ist und nicht interpretierbar. Abgesehen davon, dass es bei hinzukommenden Daten folglich jeweils (mehr oder weniger erheblich) modifiziert werden müsste, ist es schlicht und einfach nicht interpretierbar. Die aus der geometrischen Verteilung bzw. der Zipf-Mandelbrot-Verteilung gewonnenen Kurven hätten hingegen den Vorteil einer Erklärungs- bzw. Prädiktionsoption, die vor allen Dingen darin bestünde, dass die globalen Masse (Entropie und Wiederholungsrate) ausschließlich als von der Inventargröße abhängig interpretiert werden könnten. Vor diesem Hintergrund wird es von Interesse sein, die Ergebnisse mit den sich aus den anderen eingangs erwähnten Verteilungen (Good-Verteilung, Whitworth-

Verteilung, negative hypergeometrische Verteilung) zu vergleichen (Grzybek, Kelih, Altmann 2005c).

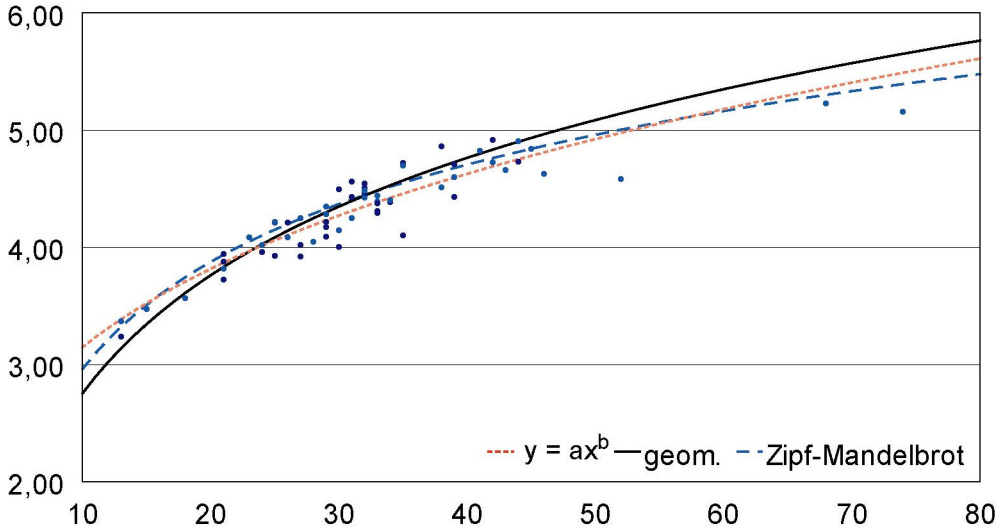


Abb. 2. Empirische und theoretische Entropien für 63 Datensätze verschiedener Sprachen nach Altmann/Lehfeldt (1980)

Hier aber soll es um eine andere Fragestellung gehen. Denn die bislang angesprochenen Studien haben mitsamt eine entscheidende Frage offen gelassen: es wurde in den genannten Studien nämlich nicht zwischen der Häufigkeit von Graphemen und Phonemen unterschieden, in der Annahme, dass beide Einheiten ohnehin ein und denselben Regularitäten folgen. Ohne Frage ist dies eine plausible Annahme, dennoch ist diese bislang ungeprüft. Und genau an dieser Stelle setzen die hier vorliegenden Überlegungen ein, die im vorliegenden Text aufgegriffen werden sollen: Denn wenn sich – wie eingangs gesagt wurde – in der konkreten empirischen Untersuchung zumindest mehrerer slawischer Sprachen gezeigt hat, dass weder die geometrische Verteilung noch die Zipf-Mandelbrot-Verteilung geeignete Modelle für Renghäufigkeiten von Graphemen oder Phonemen darstellen, dann ergibt sich zwangsläufig die Frage, ob sich die theoretischen Werte für Entropie und Wiederholungsrate überhaupt an einer dieser Verteilungen orientieren sollten.

Bevor das aber konkret geprüft wird, gilt es im hier vorliegenden Zusammenhang der erwähnten impliziten Annahme nachzugehen, nämlich der Tatsache, dass die Datensätze der 63 Sprachen auf heterogenem Material basieren: Damit ist nicht einmal gemeint, dass den Erhebungen der Häufigkeiten unterschiedliche Graphem- und/oder Phonem-Definitionen für eine gegebene Sprache oder zwischen den Sprachen zugrunde liegen, dass in einigen Untersuchungen Satz- und Leerzeichen als eigenständige Elemente berücksichtigt wurden und in anderen nicht, usw. Gemeint ist vielmehr, dass Phoneme und Grapheme undifferenziert in gleicher Art und Weise behandelt wurden, obwohl es sich – linguistisch gesehen – um unterschiedliche Arten der Repräsentation und auch Abstraktion handelt. Altmann/Lehfeldt (1980) sind bei diesem Vorgehen der Annahme gefolgt, dass in beiden Fällen[6] die theoretischen Modelle letztendlich auf einen einzigen Faktor, nämlich den des Inventarumfangs, zurück-

[6] Auch wenn sich weitere Differenzierungen etwa zwischen Buchstaben und Graphemen, oder zwischen Lauten und Phonemen, berücksichtigen ließen, soll hier lediglich von „zwei Fällen“ der Repräsentation – nämlich Graphemen und Phonemen – die Rede sei, zumal sich derartige weitere Differenzierungen nicht ohne erheblichen Aufwand aus dem verwendeten Material rekonstruieren ließen.

zuführen sind. Dennoch aber fehlt bis heute der Nachweis, dass sich das Verhalten der Phoneme und der Grapheme in dieser Hinsicht nicht nachhaltig voneinander unterscheidet – und bevor weitere, aus anderen Verteilungen hervorgehende theoretische Entropien und Wiederholungsraten erarbeitet werden, scheint es angebracht, diesen Umstand empirisch zu testen und gegebenenfalls die von Altmann/Lehfeldt (1980) implizit vertretene Ansicht zu festigen.

Diese Überprüfung kann freilich sinnvollerweise nur an den empirischen Daten vorgenommen werden, woraus sich eine mehrstufige Re-Analyse der von den Autoren verwendeten Daten und eine multiple Berechnung der Entropien und Wiederholungsraten ergibt:

1. Anpassung des nicht-linearen Regressionsmodells nach der Formel $y = ax^{-b}$
   a. für das gesamte, nicht nach Graphemen und Phonemen differenzierte Datenmaterial;
   b. für die auf Graphemen basierenden Datensätze;
   c. für die auf Phonemen basierenden Datensätze;

2. Vergleich der Regressionskurven

Vor dem Hintergrund der obigen Ausführungen mag es verwunderlich erscheinen, dass der angestrebte Vergleich auf der Basis der Potenzkurve vorgenommen werden soll, die ja – wie argumentiert wurde – theoretisch nicht begründbar ist. Solange wir aber nicht wissen, welcher theoretische Ansatz am adäquatesten ist – d.h. welcher die meisten Kriterien, die wir in einer anderen Arbeit erörtern werden, erfüllt – beschränken wir uns auf eine empirisch gut passende Kurve, die gleichzeitig eine Maßlatte für andere Kurven darstellt. Findet man eine theoretische Kurve, die gleich gut (bzw. nicht viel schlechter) ist wie die empirische, dann wird man natürlich immer die theoretisch begründete vorziehen.

Wiederholungsraten

Abb. 3a zeigt die Wiederholungsraten für die 63 Datensätze, mit unterschiedlichen Markierungen für graphem- und phonembasierte Daten. Abb. 3b zeigt die Anpassungslinie gemäß der Formel $y = ax^{-b}$. Bei Werten für $a = 0.817766$ und $b = -0.735897$ beträgt die Güte der Anpassung $R^2 = 0.7177$.

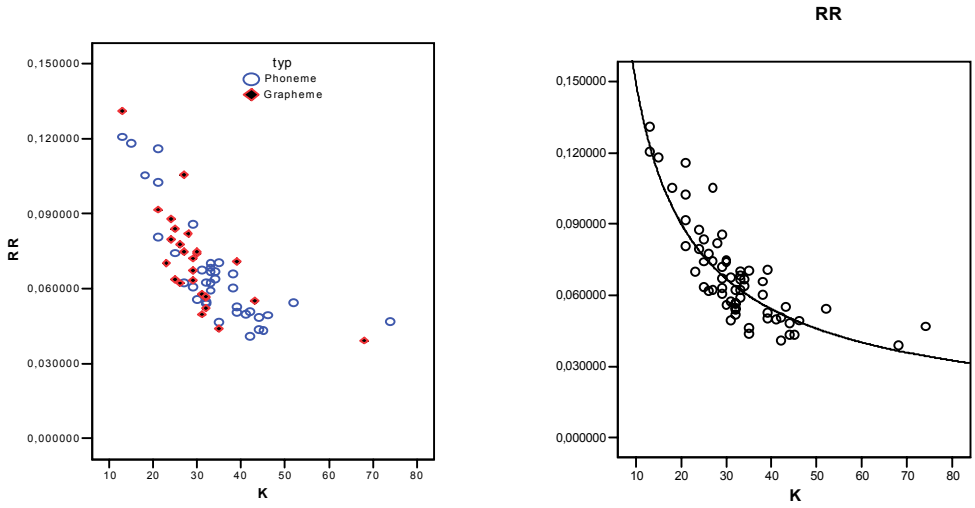


Abb. 3a/3b. Daten und Anpassung für Grapheme und Phoneme

Abb. 4a/b zeigt die Anpassungsergebnisse für die nach Graphemen und Phonemen differenzierten Datensätze: Für die Datensätze der Grapheme beträgt die Anpassungsgüte $R^2 = 0.6348$ bei Werten von $a = 0.8369$ und $b = -0.7449$, für die Datensätze der Phoneme erhält man bei Werten von $a = 0.8243$ und $b = -0.7368$ eine Anpassungsgüte von $R^2 = 0.7593$.

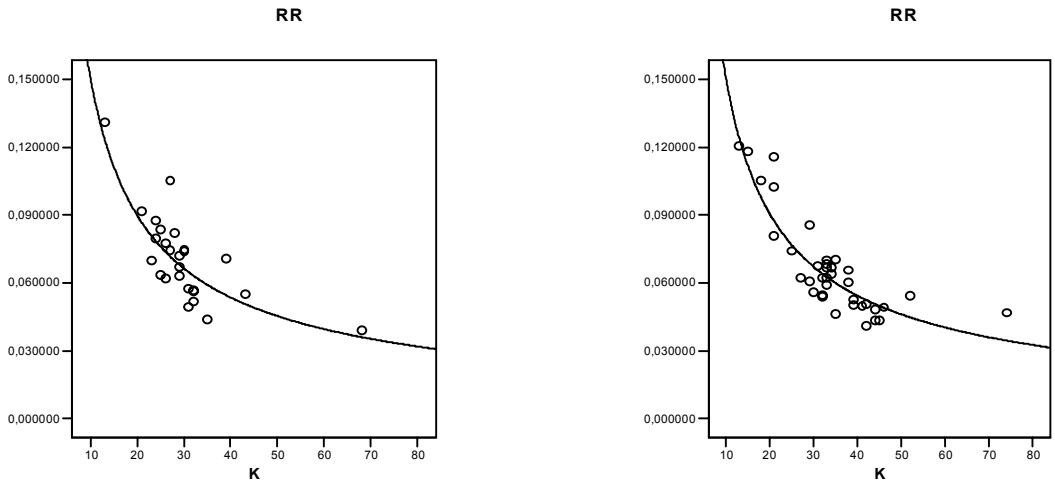


Abb. 4a/4b. Getrennte Anpassung der Formel $y = ax^{-b}$ an die Datensätze der Grapheme (links) und der Phoneme (rechts)

Auch wenn auf den ersten Blick die Kurvenverläufe sehr ähnlich wirken, ist es interessant und notwendig, die beiden Regressionskoeffizienten miteinander zu vergleichen und statistisch auf Unterschiede zu testen. Dies geschieht bei linearen Zusammenhängen über die *t*-verteilte Prüfungsgröße

$$t = \frac{|b_1 - b_2|}{\sqrt{\frac{s_{y1.x1}^2 \cdot (n_1 - 2) + s_{y2.x2}^2 \cdot (n_2 - 2)}{n_1 + n_2 - 4} \cdot \left(\frac{1}{Q_{x1}} + \frac{1}{Q_{x2}} \right)}}$$

bei $FG = n_1 + n_2 - 4$ Freiheitsgraden mit $Q_x = \sum (x - \overline{x})^2$; s^2 sind die üblichen Varianzen.

Um also die beiden Regressionskoeffizienten der Graphem- und der Phonemdaten auf Unterschied zu testen, ist eine einfache Linearisierung der Gleichung $y = a \cdot x^b$ notwendig, die sich durch eine einfache Logarithmierung in $\ln(y) = \ln(a) + b \ln(x)$ umformen lässt.

Im Ergebnis stellt sich heraus, dass der Unterschied zwischen den beiden Regressionskoeffizienten bei einem Wert von $t_{FG=59} = 0.0622$ nicht im mindesten signifikant ist ($p = 0.95$).

Entropie

Abb. 5a zeigt die Entropien für die 63 Datensätze, mit unterschiedlichen Markierungen für graphem- und phonembasierte Daten. Abb. 5b zeigt die Anpassungslinie gemäß der Formel $y = ax^{-b}$. Bei Werten für $a = 1.6602$ und $b = -0.2779$ beträgt die Güte der Anpassung $R^2 = 0.8672$.

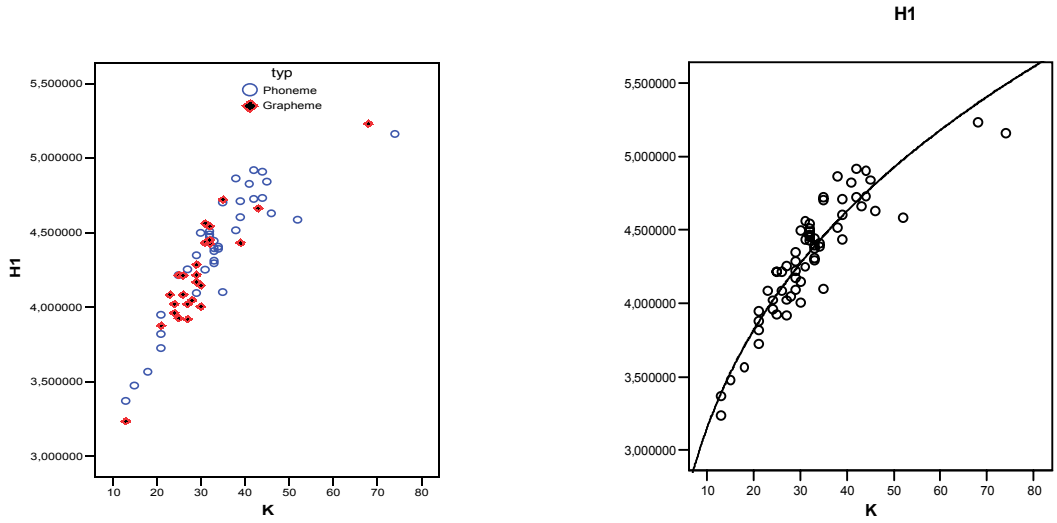


Abb. 5a/5b. Daten und Anpassung für Grapheme und Phoneme

Abb. 6a/b zeigt die Anpassungsergebnisse für die nach Graphemen und Phonemen differenzierten Datensätze: Für die Datensätze der Grapheme beträgt die Anpassungsgüte $R^2 = 0.8514$ bei Werten von $a = 1.6037$ und $b = 0.2875$, für die Datensätze der Phoneme erhält man bei Werten von $a = 1.7002$ und $b = 0.2715$ eine Anpassungsgüte von $R^2 = 0.8697$.

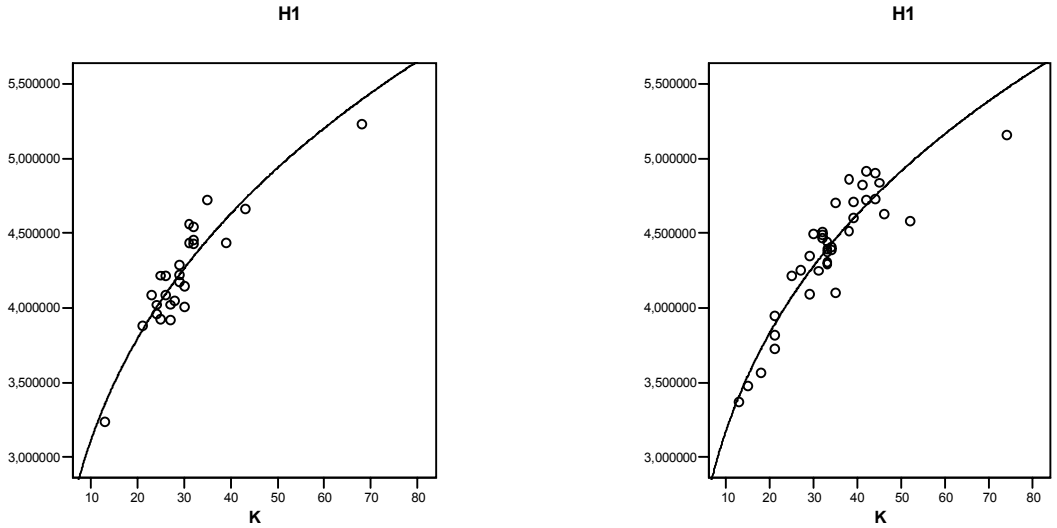


Abb. 6a/6b. Getrennte Anpassung der Formel $y = ax^{-b}$ an die Datensätze der Grapheme (links) und der Phoneme (rechts)

Auch hier stellt sich im Ergebnis heraus, dass der Unterschied zwischen den beiden Regressionskoeffizienten bei einem Wert von $t_{FG=59} = 0.5229$ nicht signifikant ist ($p = 0.60$).

Zusammenfassung

Eine wesentliche Schlussfolgerung aus den obigen Überlegungen und Untersuchungen ist es, dass Phoneme, Buchstaben und Grapheme sich im Hinblick auf ihre Häufigkeitsverteilung

gleich verhalten; das ist insofern nicht unbedingt erstaunlich, da die betreffenden Systeme im Prinzip die gleiche Funktion erfüllen. Damit soll natürlich nicht gesagt sein, dass es unerheblich ist, welche dieser Einheiten in einer gegebenen Sprache bzw. für eine gegebene Sprache erhoben und untersucht werden – vielmehr sollte grundsätzlich die Qualität der untersuchten Einheiten sauber differenziert werden.[7] Mit gleichem Verhalten ist gemeint, dass sich die theoretischen Modelle zur Beschreibung von Häufigkeiten all dieser Einheiten letztendlich ausschließlich auf den Inventarumfang *n* zurückführen lassen und somit auch interpretieren lassen sollten. Dafür jedenfalls sprechen die oben angeführten Analysen und Re-Analysen.

Literatur

Altmann, G. (1993). Phoneme counts. *Glottometrika 14, 54-58.*

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

Best, K.-H. (2005a). Buchstabenhäufigkeiten im Deutschen und Englischen. *Naukovyj Visnyk Černivec'koho Universytetu.* [Im Druck]

Best, K.-H. (2005b). Zur Häufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten. *Glottometrics 11* (eingereicht).

Grzybek, P., Kelih, E. (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 31, 131-162.*

Grzybek, P., Kelih, E. (2005). Grapheme frequencies in Slovene. In: Benko, Vladimir (ed.), *Slovko 2003.* Bratislava. [Im Druck].

Grzybek, P., Kelih, E., Altmann, G. (2004). Graphemhäufigkeiten (Am Beispiel des Russischen). Teil II: Modelle der Häufigkeitsverteilung. *Anzeiger für Slavische Philologie 32, 25-54.*

Grzybek, P., Kelih, E., Altmann, G. (2005a). Graphemhäufigkeiten im Slowakischen (Teil I: Ohne Digraphen). In: Nemcová, E. (Hrsg.), *Philologia actualis slovaca*. [Im Druck].

Grzybek, P., Kelih, E., Altmann, G. (2005b). Graphemhäufigkeiten im Slowakischen (Teil II: Mit Digraphen). In: *Sprachen und Sprache im Mitteleuropäischen Raum*. [Im Druck]

Grzybek, P., Kelih, E., Altmann, G. (2005c). Grapheme frequencies: Model characteristics and criteria. In: *Journal of Quantitative Linguistics.* [In Vorb.]

Herdan, **G.** (1962). *The calculus of linguistic observations.* The Hague: Mouton.

Herdan, G. (1966). *The advanced theory of language as choice and chance.* Berlin: Springer.

Hussien, *O.A*. (2004). The Lerchianness plot. *Glottometrics 7,* 50-64.

Mandelbrot, **B. B.** (1953). An information theory of the statistical structure of language. In: W. Jackson (ed.), *Communication Theory: 503-512*. New York: Academic Press.

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

---

[7] So sollte etwa zwischen Buchstaben, Graphemen und Phonemen sowie deren jeweiligen Häufigkeiten deutlich getrennt werden – wobei unter Buchstaben Einzelzeichen und unter Graphemen auch (!) Kombinationen von Buchstaben zu verstehen wären, so z.B. [sch] im Deutschen, [sh] im Englischen, [ch] im Französischen, [cs, sz, ny, gy,…] im Ungarischen usw. Vor diesem Hintergrund wäre davon auszugehen, dass innerhalb einer gegebenen Sprache für die Inventargröße der drei Einheiten die ungefähre Regel gilt:

Buchstabeninventar ≤ Grapheminventar ≤ Phoneminventar,

wobei es zahlreiche Ausnahmen gibt.

Naranan, S., Balasubrahmanyan, V.K. (1992a). Information theoretic models in statistical linguistics. Part I: A model for word frequencies. *Current Science 63, 261-269.*

Naranan, S., Balasubrahmanyan, V.K. (1992b) Information theoretic models in statistical linguistics. Part II: Word frequencies and hierarchical structure in language – statistical tests. *Current Science 63, 297-306.*

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

Pääkkönen, M. (1993). Graphemes and context. *Glottometrika 14, 1-53*.

Rosenbaum, R., **Fleischmann, M.** (2002). Character frequency in Multilingual Corpus 1 – Part 1. *Journal of Quantitative Linguistics 9(*3), *233-260.*

Rosenbaum, R., **Fleischmann, M.** (2003). Character frequency in Multilingual Corpus 1 – Part 2. *Journal of Quantitative Linguistics 10(1)*, *1-39*.

Sigurd, B. (1968). Rank-frequency distribution for phonemes. *Phonetica 18, 1-15.*

Tuldava, J. (1988). Opyt kvantitativnogo analiza sistemy fonem ėstonskogo jazyka. *Acta et Commentationes Universitatis Tartuensis 838, 120-133.*

Whitworth, **W.A.** (1901). *Choice and chance. With one thousand exercises*. New York, London 1965.

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

Zörnig, P., Altmann, G. (1983). The repeat rate of phoneme frequencies and the Zipf-Mandelbrot law. *Glottometrika 5, 205-211*.

Zörnig, P., Altmann, G. (1984). The entropy of phoneme frequencies and the Zipf-Mandelbrot law. *Glottometrika 6, 41-47.*

Glottometrics 9, 2005, 74-88

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.

VIII. Karl Marbe (1869-1953)

Der Psychologe Karl Marbe wurde 1869 in Paris geboren, übersiedelte bald nach Freiburg, wo er seine Schulzeit verbrachte, und betrieb anschließend ein thematisch vielseitiges Studium in Freiburg, Bonn, Berlin, Paris und wieder Bonn (Promotion 1893), u.a. Linguistik bei H. Paul, verlegte sich aber mehr und mehr auf die Psychologie. Er verbrachte ein Jahr in Leipzig (W. Wundt), arbeitete danach wiederum in Bonn und habilitierte sich schließlich in Würzburg (1896), wo er bis 1905 als Privatdozent wirkte. 1905-1909 Professor in Frankfurt, 1909-1935 in Würzburg, wo er 1953 verstarb (Hoffmann & Stock o.J.; Marbe 1945; Murray 1978: XV-XVI; Schorn 1936).

In der Quantitativen Linguistik ist der Psychologe Karl Marbe wenig bekannt. Gemeinsam mit dem Indogermanisten und Neogräzisten Albert Thumb (1865-1915) hat er wichtige Beiträge zu zwei Themen geliefert:

1. Auf eine Anregung von Thumb (Marbe 1945: 8) geht die Erforschung der Analogie zurück. In Thumb & Marbe (1901) geht es darum herauszufinden, wie die sprachliche Analogie funktioniert. Hintergrund hierfür ist die Auffassung, dass Analogie kein rein linguistischer Gegenstand sei, sondern dass die Linguistik bei ihrer Erforschung die Ergebnisse von Assoziationsversuchen zu berücksichtigen habe (Best 1973: 36-38; Marbe 1913: 31f.). (Zur Rezeption und Kritik dieses Werkes in der Linguistik vgl. Murray 1978: XVIff.). Die Beobachtung von in hohem Maße gleichartigen Assoziationen der Versuchspersonen auf bestimmte Stimuli unterstützte Marbe bei seiner Theorie der „Gleichförmigkeit in der Welt“ (Marbe 1916), die auf der „Gleichförmigkeit der Bedingungen des Seins und Geschehens“ (Marbe 1945: 15) beruhe.

2. Marbes Interesse galt außerdem einem stilistischen Phänomen: dem Sprachrhythmus (Marbe 1904). Es handelt sich um einen Gegenstand, den er offenbar als Erster thematisiert hat und der dann von etlichen seiner Schüler und auch von Albert Thumb aufgegriffen wurde. Da diese Thematik für die Suche nach Gesetzmäßigkeiten in der Sprache nutzbar gemacht werden kann, soll sie hier vorgestellt werden.

Die Idee war folgende: Man kann Rhythmen in Texten danach bestimmen, ob zwischen zwei betonten Silben keine, eine, zwei, drei oder mehr unbetonte Silben vorkommen. Sie werden als rhythmische Einheiten der Länge 1 (keine unbetonte Silbe zwischen zwei betonten), der Länge 2 (1 unbetonte zwischen 2 betonten) etc. bestimmt. Es ging ihm dabei um eine Untersuchung seines Eindrucks, dass Goethes *Sankt Rochusfest zu Bingen* einen gleichmäßigeren Rhythmus aufweise als Heines *Harzreise*. Sein Ergebnis war, „daß die Anzahl der zwischen zwei betonten Silben stehenden unbetonten Silben dort [bei Heine, Verf.] mehr um

ihren Mittelwert schwanken als hier, sie zeigt daher, daß der Rhythmus im Anfang des Rochusfestes ein gleichförmigerer ist, als im Anfang der Harzreise“ (Marbe 1904: 8).

Die Rhythmisierung von Texten ist für Marbe aber nicht nur ein stilistisches Phänomen; er sieht sie vielmehr gesetzmäßig mit weiteren Aspekten verknüpft. So weist er darauf hin, dass in gefühlsbetonten Texten mehr Einsilber vorkommen als in nicht gefühlsbetonten und dass damit eine unterschiedliche Rhythmik einhergeht. Ernste oder literarische Texte unterscheiden sich ebenfalls rhythmisch von ihren Widerparts (Marbe 1913: 36-38). Damit verbunden sind beim Rezipienten unterschiedliche Bewußtseinslagen (Marbe 1945: 9).

Aus heutiger Sicht wohl noch wichtiger ist eine weitere Nutzung der Untersuchungsergebnisse von Marbe. Man kann nämlich prüfen, ob rhythmische Einheiten sich in Texten ebenso wie Satz- oder Wortlängen gesetzmäßig verhalten (Altmann 1988; Wimmer u.a. 1994). Marbe selbst hat die Anfänge der beiden Texte ausgewertet, jeweils willkürliche Abschnitte von 1000 Wörtern. Zur Kontrolle hat er einen Kollegen (Roetteken) teilweise die gleichen Abschnitte auswerten lassen. Er war der Ansicht, diese Arbeit lasse sich „sehr leicht und im allgemeinen sicher“ (Marbe 1904: 4) durchführen, stellte dann aber doch fest, dass die Bearbeitung durch Roetteken sich „durchweg und zum Teil nicht unerheblich“ (Marbe 1904: 16) von seiner eigenen unterschied. Es entstanden für insgesamt 6 Textabschnitte 8 Übersichten über die Verteilung von rhythmischen Einheiten verschiedener Länge. An diese Dateien beider Bearbeiter konnte in Übereinstimmung mit den Theorien von Altmann (1988) und Wimmer u.a. (1994) die Hyperpoisson-Verteilung mit guten Ergebnissen angepasst werden (Best 2001b: 164f.). Gleiche Berechnungen zu Arbeiten der Marbe-Schule (Best 2002: 137) zum Althebräischen (Friedmann 1921/22) und Deutschen (Gropp 1915) scheiterten, gelangen aber bei einer Tabelle zum Englischen (Lipsky 1907); auch beim Altgriechischen (Thumb 1913: 148) sind die Ergebnisse nur teilweise zufriedenstellend. Man kann dies bis zum Beweis des Gegenteils auf die willkürlichen und z.T. sehr kurzen Textabschnitte zurückführen: Bei 3 vollständig ausgewerteten deutschen Fabeln Pestalozzis (Best 2001a: 4-6) und 17 deutschen Kurzprosatexten von Strittmatter und Vesper (Best 2002; 2005) ergaben sich gute Anpassungen der Hyperpoisson-Verteilung. Dieses Ergebnis wird von Kaßel (2002) für je 15 deutsche Briefe (H. v. Kleist) und Pressetexte ebenso wie für je 15 englische Briefe (J. Austen) und Pressetexte bestätigt. Natürlich stand die Theorie der Verteilungen sprachlicher Einheiten verschiedener Längen für Marbe und seine Mitarbeiter nicht zur Debatte; immerhin haben sie Daten erhoben, die die Anregung dazu gaben, auch rhythmische Einheiten in diese Untersuchungen einzubeziehen.

Literatur

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

Best, K.-H. (1973). *Probleme der Analogieforschung.* München: Hueber.

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

Best, K.-H. (2001b). Zur Verteilung rhythmischer Einheiten in deutscher Prosa. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 162-166.* Göttingen: Peust & Gutschmidt.

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

Best, K.-H. (2005). Längen rhythmischer Einheiten. In: Altmann, G., Köhler, R., Piotrowski, R. (eds.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch.* Berlin/ New York: de Gruyter (erscheint).

Friedmann, M. (1921/22). *Der Prosarhythmus des Hebräischen im alten Testament.* Würzburg, diss.phil.

Gropp, F. (1915). *Zur Ästhetik und Statistik des Prosarhythmus.* Würzburg, diss.phil.

Hoffmann, J., Stock, A. (o.J.), The Würzburg School. Würzburg. www.psychologie.uni-wuerzburg.de/w_schule/WSCHOOL2a.pdf

Kaßel, A. (2002). *Zur Verteilung rhythmischer Einheiten in deutschen und englischen Texten.* Staatsexamensarbeit; Göttingen.

Lipsky, A. (1907). *Rhythm as a Distinguishing Characteristic of Prose Style.* New York: The Science Press.

Marbe, K. (1904). *Über den Rhythmus der Prosa*. Giessen: J. Ricker'sche Verlagsbuchhandlung.

Marbe, K. (1913). Die Bedeutung der Psychologie für die übrigen Wissenschaften und die Praxis. In: *Fortschritte der Psychologie und ihrer Anwendungen.* Unter Mitwirkung von W. Peters hrsg. von Karl Marbe. Leipzig/ Berlin: Teubner.

Marbe, K. (1916). *Die Gleichförmigkeit in der Welt.* München: Beck'sche Verlagsbuchhandlung Oskar Beck.

Marbe, K. (1945). *Selbstbiographie des Psychologen Geheimrat Prof. Dr. Karl Marbe in Würzburg.* Hrsg. im Namen der Kaiserlich Leopoldinisch-Carolinisch-Deutschen Akademie der Naturforscher von Emil Abderhalden. Halle (Saale).

Murray, D. D. (1978). *Introduction*. In: Neuausgabe von Thumb & Marbe (1901).

Schorn, M. (1936). Das Psychologische Institut der Universität Würzburg unter Karl Marbe. *Archiv für die gesamte Psychologie 95*: 162-199.

Thumb, A. (1913). Satzrhythmus und Satzmelodie in der altgriechischen Prosa. In: *Fortschritte der Psychologie und ihrer Anwendungen: 139-168.* Unter Mitwirkung von W. Peters hrsg. von Karl Marbe. Leipzig/ Berlin: Teubner.

Thumb, A., Marbe, K. (1901). *Experimentelle Untersuchungen über die psychologischen Grundlagen der sprachlichen Analogiebildung.* Leipzig: Engelmann (Neuausgabe: David D. Murray. Amsterdam: John Benjamins 1978).

Wimmer, G., Köhler, R., Grotjahn, R., Altmann, G. (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics 1*: 98-106.

Karl-Heinz Best, Göttingen

IX. Jiří Krámský (1913-1991)

Czech linguist, born on 23.10.1913 at Plzeň in a family of a law official, † 30.9.1991.

PhDr. Jiří Krámský, DrS. finished his schooling at the Philosophical Faculty of Charles University in Prague, where he studied Turkish and Iranian, and also English philology. His doctoral thesis was called *A Grammatical Analysis of 'View of the Present State of Ireland' by Spencer*. This work was finished under the linguistic guidance of Bohumil Trnka. Vilém Mathesius became the second personality of the linguistic branch that influenced Krámský's studies. The reason for following in the footsteps of these two authorities was natural: Structuralism of the *Prague linguistic circle* (having such famous members as R. Jakobson and N. S. Trubetskoy) shaped the thinking of young linguists at that time. Unfortunately, the successive development of the country, the World War Two and the succeeding communist regime, were not advantageous for scientific activities. Many years before his retirement, in the years 1955−1978, he worked at a research establishment of the Ministry of Education. In spite of these circumstances, Krámský created an extensive work numbering more than 330 bibliographic items.

His linguistic ideas are based on the analyses of many European (Czech, Norwegian, Italian, Scots Gaelic) and Oriental (Uralo-Altaic, Iranian) languages. His works on the field of language teaching, concerning mainly the English language, maintain their practical implications up to the present. Let us mention at least his school grammar of the English language (in two editions, 1979 and 1989), university textbooks concerning English orthography (1955), an English grammar (1956), a grammar textbook (1957), an American reader (1959), a handbook of specialized German texts (with J. Turek, 1970) and others.

The spectrum of his linguistic interests ranged from phonology to lexicography and stylistics in many languages. While descriptive aims based on conceptual schemes prevailed among the structuralists in his intellectual environment, he always supported his descriptive conclusions by numerical data. In his works structuralism overgrows to the "structuralism" of quantitative linguistics. For example, Krámský's paper *A phonological analysis of Persian monosyllables* (*Archiv orientální 16*, 1947, 103-134) presents detailed statistic distributions of individual phonemes on different positions of monosyllabic words differed to originally Persian and Arabic lexical units. They are generalized to individual syllabic types with stressed differences of the two word strata. His conclusion is of a kind exceeding the level of a simple description. He writes: '... **the compensating tendency**, peculiar to words of Persian origin, which operates almost as a law: **the lack of a certain kind on one hand is outbalanced by the surplus of the same sounds on the other hand...** On the other hand, in words of Arabic origin we often meet with opposite tendency, viz. **to favour certain phonemes to disadvantage of others.**' (Underlined by J. Krámský.)

The same methodological philosophy was applied in his typological studies in which he followed and enlarged the results published by Vladimír Skalička, A. V. Isachenko and R. Jakobson. In his paper *A quantitative typology of languages* (*Language and Speech*, Vol. 2, Part2, April-June 1959, 72-85) the quantitative aspect is stressed as a necessary complement of qualitative linguistic facts. He proposed a classification of 23 languages on the basis of data concerning the phoneme distributions achieved from texts.

More complete image of the linguistic and methodological views of Jiří Krámský can be found in his following publications:

– (1969): *The word as a linguistic unit.* Mouton, Haag-Paris.
– (1972): *The article*. Mouton, Haag-Paris.
– (1974): *The phoneme*. Fink, München.
– (1976): *Papers in general linguistics*. Mouton, Haag –Paris.

Luděk Hřebíček, Prague

X. Georg von der Gabelentz (1840-1893)

Georg von der Gabelentz wurde am 16.3.1840 in Poschwitz bei Altenburg als Sohn des Sprachwissenschaftlers Hans Conon von der Gabelentz (1807-1874) geboren; nach der Schule Studium von Jura, Kameralistik und Sprachwissenschaft 1859-63 in Jena, 1863-64 in Leipzig. 1864-78 Verwaltungsjurist im sächsischen Staatsdienst; 1876 Promotion mit einer sinologischen Arbeit; 1878-89 Professor für ostasiatische Sprachen in Leipzig, 1884-1890 Mitherausgeber von Techmers *Internationale Zeitschrift für Allgemeine Sprachwissenschaft*; ab 1889 ordentlicher Professor für ostasiatische Sprachen und allgemeine Sprachwissenschaft in Berlin; 1890 Mitglied der Preußischen Akademie der Wissenschaften; 1891 erscheint sein Hauptwerk *Die Sprachwissenschaft*; 11.12.1893 in Lemnitz gestorben (Narr & Petersen [2]1972: 2; (http://www.uni-erfurt.de/sprachwissenschaft/personal/lehmann/CL_Lehr/Gesch_SW/Gabelentz/Gabelentz.html).

Gabelentz gehört zweifellos zu den Vordenkern der quantitativen Linguistik in Deutschland (Best 1999); er entwirft in einer postum veröffentlichten Schrift das Programm einer sehr modern anmutenden, quantitativen Sprachtypologie und meint: „aus einem Dutzend bekannter Eigenschaften einer Sprache müsste man mit Sicherheit auf hundert andere Züge schliessen können; die typischen Züge, die herrschenden Tendenzen lägen klar vor Augen" (Gabelentz 1894: 7). Allgemeiner drückt er sich in Gabelentz (²1901: 481) aus: „Was man bisher von geistiger Verwandtschaft, von verwandten Zügen stammverschiedener Sprachen geredet hat, das würde hinfort greifbare Gestalt gewinnen, in ziffermäßig bestimmten Formeln dargestellt werden." Es handelt sich in diesem Fall um einen Auszug aus einer längeren Passage, die erst postum in das Werk eingefügt wurde und in der Erstauflage (Gabelentz 1891) noch fehlt. Hier ist also zu fragen, ob diese postum eingefügten Aussagen von Gabelentz selbst oder vom Herausgeber der Neuauflage stammen. Plank (1991: 425) hält sie zumindest in Teilen für authentisch. Coseriu (1972: 29) verweist ohne Bezugnahme auf ein spezielles Werk darauf, dass Skaličkas Typologie dem in Gabelentz (^{2}1901) vorgetragenen Konzept entspricht, sofern es um die „Wechselwirkungen" zwischen Eigenschaften des Sprachsystems geht. Man muss hinzufügen: „ziffermäßig bestimmte Formeln" fehlen in den entsprechenden Arbeiten Skaličkas, z.B. in Skalička (1966/1979).

Dieser besondere Aspekt seiner „Typologie" (dieser Begriff wird in Gabelentz ^{2}1901: 481 vorgeschlagen), das sprachstatistische Programm, das Gabelentz (1894) prägnant formulierte, scheint lange Zeit in der Sprachwissenschaft übersehen oder verkannt worden zu sein. Es hat rund 70 Jahre gedauert, bis seine Ideen − allerdings unbekannterweise − im Ansatz verwirklicht wurden. Erst nachdem Greenberg (1960) nämlich seine typologischen Indizes entwickelt hatte, haben Krupa & Altmann (1966) sowie Altmann & Lehfeldt (1973: 45) Korrelationen zwischen diesen Indizes berechnet. Auf dieser Basis lassen sich dann mit den Mitteln der numerischen Taxonomie (Sneath & Sokal 1973) neue Klassifikationen von Sprachtypen entwickeln, wie dies Altmann & Lehfeldt (1973: 34ff.) vorgeführt haben. Eine weitere Anwendungsmöglichkeit besteht darin, dass man die Ausprägung einzelner Eigenschaften in einer Sprache misst und von diesen her Voraussagen über andere, mit ihnen verbundene Eigenschaften machen kann. Dies wäre genau das Programm, das Gabelentz sich vorgestellt hatte.

Noch ein weiterer Aspekt ist hervorzuheben. So betont Gabelentz den Wert statistisch abgesicherten Wissens so stark und auch so allgemein, dass man den Eindruck gewinnen kann, er wolle dieses Prinzip nicht auf typologische Fragen beschränkt sehen. „Aber wie weit ist sie [die Subjektivität; Verf.] zurückgeschoben, wie weit reicht das objektivste, was man verlangen kann, das zahlenmäßig festgestellte. Geriete das Werk nur soweit, nur bis zu einer unanfechtbaren Statistik, so hätte die allgemeine Sprachwissenschaft nicht länger die sprachgeschichtliche Forschung um ihren festen Baugrund zu beneiden" (Gabelentz 1894: 7). An anderer Stelle heißt es, man solle Gedanken „in eine kontrollierbare Form...kleiden, und besser kontrollierbar ist keine als die statistische" (Gabelentz 1894: 4).

Literatur

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

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

Coseriu, E. (1972). Georg von der Gabelentz und die synchronische Sprachwissenschaft. In: Gabelentz (1901/²1972), S. 3-35).

Gabelentz, G. von der (1891). *Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse.* Leipzig: Weigel.

Gabelentz, G. von der (1894). Hypologie [= Typologie] der Sprachen, eine neue Aufgabe der Linguistik. *Indogermanische Forschungen 4*: 1-7.

Gabelentz, G. von der ([2]1901). *Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse. Zweite, vermehrte und verbesserte Auflage*. Herausgegeben von Dr. Albrecht Graf von der Schulenburg. Leipzig: Tauchnitz. (Neuauflage: Mit einer Studie von Eugenio Coseriu neu herausgegeben von Gunter Narr und Uwe Petersen. Tübingen: Tübinger Beiträge zur Linguistik 1969, ²1972).

Greenberg, J. H. (1960). A quantitative approach to the morphological typology of languages. *International Journal of American Linguistics 26, 178-194.*

Krupa, V., Altmann, G. (1966). Relations between typological indices. *Linguistics 24*: 29-37.

Narr, G., Petersen, U. (²1972). Vorwort der Herausgeber. In: Gabelentz (1901/²1972), S. 1-2.

Plank, F. (1991). Hypology, Typology: The Gabelentz Puzzle. *Folia Linguistica XXV*: 421-458.

Skalička, V. (1966/1979). Ein „typologisches Konstrukt". *Travaux linguistiques de Prague 2, 1966, 157-163*. Auch in: Vladimír Skalička (1979). *Typologische Studien: 335-341*. Mit einem Beitrag von Petr Sgall. Herausgegeben von Peter Hartmann. Braunschweig/ Wiesbaden: Vieweg.

Sneath, Peter A., & Sokal, Robert R. (1973) *Numerical Taxonomy.* San Francisco: Freeman.

Quelle im Internet:
http://www.uni-erfurt.de/sprachwissenschaft/personal/lehmann/CL_Lehr/Gesch_SW/Gabelentz/Gabelentz.html.

Karl-Heinz Best, Göttingen

XI. Gottfried Wilhelm Leibniz (1646-1716)

Leibniz wurde 1646 in Leipzig geboren, studierte 1661-66 in Leipzig und Jena und promovierte mit seiner *Dissertatio de arte combinatoria* in Altdorf. Weitere Stationen seines Lebensweges waren u.a. 1672-76 Paris, 1673 Reise nach England, 1676 Bibliothekar in Hannover im Dienst von Herzog Johann Friedrich von Braunschweig-Lüneburg, 1691 Leiter der Bibliothek in Wolfenbüttel, 1712-14 in Wien, 1716 gest. in Hannover. Leibniz war ein vielseitiger Gelehrter mit Schwerpunkten in Mathematik, Physik und Philosophie, befasste sich aber auch mit verschiedenen sprachwissenschaftlichen Themen, darunter Überlegungen zur Universalsprache, zu Sprachwandel und -verwandtschaft (Finster & van den Heuvel [4]2000: 89f.; Gardt 1999: 135ff.; Naumann 1966; von der Schulenburg 1973. Für linguistische Themen bei Leibniz ist Doucet-Rosenstein 1981 wenig ergiebig.).

Für die Quantitative Linguistik ist Leibniz mit kombinatorischen Überlegungen bedeutsam, die zu seiner Zeit bereits eine lange und vielfältige Tradition haben (Eco 1997: 148ff., 278ff.; Gardt 1994: 211ff., Gardt 1999: 139ff.; Schmidt 1966: 51) und in die arabische Mathematik (Djebbar 1981: 55, 125)[1] und noch weiter bis in die Kabbala (Überlieferung) des 2.-4. Jahrhunderts zurückreichen (Eco 1997: 41; Strasser 1988: 60). Die ältesten Quellen für die Anwendung kombinatorischer Überlegungen auf sprachliche Gegenstände scheinen Xenokrates (4. Jhd. vor Chr.) und in der Hindu-Tradition Pingala (ca. 200 v. Chr.) zu sein (Biggs

[1] Dank an Prof. Benno Artmann, Göttingen, für den Hinweis.

1979: 113f.). Verschiedene Richtungen des kombinatorischen Verfahrens charakterisiert Eco (1997: 80f.) wie folgt: „Aber was das kabbalistische Denken vom Denken Lulls unterscheidet, ist, daß in der Kabbala die Kombination der Buchstaben Realitäten nicht widerspiegelt, sondern hervorbringt. Die Wirklichkeit, die der kabbalistische Mystiker aufdecken muß, ist noch nicht bekannt und enthüllt sich nur durch das Buchstabieren der Lettern in schwindelerregenden Permutationen. Lulls Kombinatorik hingegen ist ein rhetorisches Mittel, durch das bewiesen werden soll, was schon bekannt ist..." Die Kombinatorik steht also im Dienst unterschiedlicher Zielsetzungen. Linguistisch interessant ist dabei die Tatsache, dass sie auch für Wortbildung und Texterzeugung eingesetzt wurde (Gardt 1999: 141). Harsdörffers *Fünffacher Denckring der Teútschen Sprache* (Harsdörffer 1651/ 1990: 517) enthält in fünf Ringen Wortbestandteile, die man gegeneinander drehen kann, wobei jede Kombination ein anderes „Wort" ergibt, worunter sich natürlich viele im Deutschen nicht belegte „blinde oder deutungslose" Formen befinden (Hundt 2000: 281ff.). Leibniz nahm diesen *Denckring* zum Anlass, um auszurechnen, wie viele Ausdrücke sich damit bilden lassen (Strasser 1988: 237). Mit Beginn des 17. Jahrhunderts wird ein weiterer Aspekt erneut thematisiert, den bereits Xenokrates (Biggs 1979: 113) und später die Araber behandelten (Djebbar 1981), indem gefragt wird, wie viele Ausdrücke mit einem vorhandenen Alphabet gebildet werden können, „unabhängig davon, ob sie einen Sinn haben und aussprechbar sind" (Eco 1997: 149). Knobloch (1973: 78) charakterisiert diese Entwicklung unter Bezugnahme auf den Mathematiker und Theologen Marin Mersenne (1588-1648) und andere Autoren damit, „daß insbesondere zur Zeit MERSENNEs die Gesamtheit aller möglichen und aussprechbaren Wahrheiten in Verbindung mit kombinatorischen Studien berechnet wurde."

Die genannten Strömungen bilden den Hintergrund für Leibniz' eigene Bemühungen um die Kombinatorik. In seiner *Dissertatio de arte combinatoria* (1666; 1962: 61) stellte er u.a. Überlegungen dazu an, wie viele Wörter man bilden kann, wenn man ein Alphabet von 24 Buchstaben zur Verfügung hat und die gebildeten Ausdrücke selbst höchstens 24 Buchstaben lang sein sollen; es sind über 620 Trilliarden. Allerdings gibt es in den Sprachen der Welt auch Wörter, die mehr als 24 Buchstaben lang sind, und es gibt Sprachen, die mehr als nur 24 Buchstaben aufweisen (Eco 1997: 279; Pott 1884: 19f.). Entsprechend erweitert sich die Zahl der durch Kombination zu bildenden Ausdrücke. (Die *Dissertatio de arte combinatoria* wird von Knobloch 1973: 23ff. ausführlich unter mathematischen Gesichtspunkten dargestellt und kommentiert.) In *De l'horizont de la doctrine humaine* (Fichant 1991: 50ff.) befasst Leibniz sich mit der Frage, wie viele Sätze man mit einem begrenzten Alphabet bilden kann; es ergibt sich eine Menge, welche die gesamte Menschheit nicht verarbeiten könnte, wozu Eco (1997: 280) anmerkt: „Leibniz ... ist fasziniert vom Taumel der Entdeckung, und das heißt von der Unzahl an Sätzen, die eine einfache mathematische Berechnung zu konzipieren erlaubt." Leibniz' Überlegungen hierzu entwickelt und kommentiert wiederum Knobloch (1973: 81ff.) und resümiert: „LEIBNIZ will in ihnen zeigen, daß die Zahl der wahren und falschen Aussagen, die Menschen machen können, und daher auch die Zahl der herstellbaren Bücher begrenzt ist, daß also nach einer wenn auch gewaltigen Zeitspanne alles bereits einmal gesagt sein muß und nur noch Wiederholungen möglich sind, kurz daß der menschliche Geist begrenzt ist. Er steht damit freilich im Gegensatz zu seiner eigenen, an anderer Stelle überlieferten Bemerkung [...], daß schon die Zahl der ersten *propositiones* unendlich ist..." (S. 83).

Diese Hinweise zeigen, dass Leibniz mit seinen kombinatorischen Überlegungen in einer großen Tradition steht, deren mathematische Genese Knobloch (1973: 1ff.) behandelt. Sie erlebt im Barock einen neuen Höhepunkt, wobei es um ganz unterschiedliche Themen geht, darunter um das Problem der Entwicklung einer Universalsprache, um die Wortbildung, die Kryptologie und um literarische Aspekte wie die Bildung von Anagrammen und Proteus-Versen.

Proteus-Verse bestehen überwiegend aus einsilbigen Wörtern und lassen daher „ungewöhnlich viele Umstellungen [...] zu, ohne daß die metrischen Gesetze verletzt werden" (Knobloch 1973: 39). Leibniz (1666/ 1962: 65) zitiert ein deutsches Beispiel aus Harsdörffer (1653/ 1990), das man durch Umstellung der 11 einsilbigen Substantive und Verben in 39916800 Formen bringen könne:

„Ehr, Kunst, Geld, Guth, Lob, Weib und Kind
Man hat, sucht, fehlt, hofft und verschwind."

Literatur

Biggs, N.L. (1979). The Roots of Combinatorics. *Historia Mathematica 6, 109-136.*

Djebbar, A. (1981). *Enseignement et recherche mathématiques dans le Maghreb des XIIIe - XIVe siècles (étude partielle).* Université de Paris-Sud, Département de Mathématique (= Publications mathématiques d'Orsay, n° 81-02).

Doucet-Rosenstein, D. (1981). *Die Kombinatorik als Methode der Wissenschaften bei Raimund Lull und G.W. Leibniz.* Diss. phil., München.

Eco, U. (1997). *Die Suche nach der vollkommenen Sprache*. München: dtv.

Fichant, M. (Hrsg.) (1991). Gottfried Wilhelm Leibniz. *DE L'HORIZONT DE LA DOCTRINE HUMAINE (1693). ´Apokatástasis pánton (La Restitution Universelle) (1715). Textes inédits, traduits et annotés.* Paris: Librairie Philosophique J. Vrin.

Finster, R., van den Heuvel, G. ([4]2000). *Gottfried Wilhelm Leibniz.* Reinbek: Rowohlt.

Gardt, A. (1994). *Sprachreflexion in Barock und Frühaufklärung. Entwürfe von Böhme bis Leibniz.* Berlin/New York: de Gruyter.

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

Harsdörffer, G. P. (1651/1990). *Delitiae Mathematicae et Physicae. Der Mathematischen und Philosophischen Erquickstunden Zweyter Teil.* Neudruck der Ausgabe Nürnberg 1651 herausgegeben und eingeleitet von Jörg Jochen Berns. Frankfurt/Main: Keip.

Harsdörffer, G. P. (1653/1990). *Delitiae Mathematicae et Physicae. Der Mathematischen und Philosophischen Erquickstunden Dritter Theil.* Neudruck der Ausgabe Nürnberg 1653 herausgegeben und eingeleitet von Jörg Jochen Berns. Frankfurt/Main: Keip.

Hundt, M. (2000). *„Spracharbeit" im 17. Jahrhundert. Studien zu Georg Philipp Harsdörffer, Justus Georg Schottelius und Christian Gueintz.* Berlin/ New York: de Gruyter.

Knobloch, E. (1973). *Die mathematischen Studien von G. W. Leibniz zur Kombinatorik. Auf Grund fast ausschließlich handschriftlicher Aufzeichnungen dargelegt und kommentiert.* Wiesbaden: Franz Steiner Verlag.

Leibniz, G. W. (1666/ 1962). Dissertatio de arte combinatoria. In: Leibniz, G. W. (1962), *Mathematische Schriften*, Hrsg. v. C. I. Gerhard. Bd. V: *Die mathematischen Abhandlungen* (S. 1-79). Hildesheim: Olms.

Naumann, B. (1966). Die Tradition der Philosophischen Grammatik in Deutschland. In: Schmitter, Peter (Hrsg.), *Sprachtheorien der Neuzeit II* (S. 24ff.). Tübingen: Narr.

Pott, A. F. (1884). Einleitung in die allgemeine Sprachwissenschaft. *Internationale Zeitschrift für allgemeine Sprachwissenschaft 1 (= Techmers Zeitschrift)*: *1-68.*

Schmidt, F. (1966). *Zeichen und Wirklichkeit.* Stuttgart: Kohlhammer.

Strasser, G. F. (1988). *Lingua Universalis. Kryptologie und Theorie der Universalsprachen im 16. und 17. Jahrhundert.* Wiesbaden: In Komm.: Harrassowitz.

Schulenburg, S. von der (1973). *Leibniz als Sprachforscher.* Mit einem Nachwort herausgegeben von Kurt Müller. Frankfurt/M.: Vittorio Klostermann (lt. Vorwort, VII, entstanden zwischen 1929 und 1939).

Karl-Heinz Best, Göttingen

XII. Albert Thumb (1865-1915)

Thumb wurde am 18.05.1865 in Freiburg i. B. geboren, studierte in Freiburg klassische Philologie und vergleichende Sprachwissenschaft, später in Heidelberg und schließlich in Leipzig (hier auch Germanistik sowie Völkerpyschologie und Psychologie bei Wilhelm Wundt). 1888 Dissertation über den Spiritus asper im Griechischen (Freiburg). 1889 Staatsexamen. 1889 bis 1890 weilte er für Sprachstudien in Griechenland. 1891 Habilitation in Freiburg (über einen neugriechischen Dialekt). Ab 1891 Privatdozent und Gymnasiallehrer in Freiburg, 1895 a.o. Professor für vergleichende Sprachwissenschaft in Freiburg, ab 1901 in Marburg und 1909 Ordinarius in Straßburg, wo er ab 1914 auch das Psychologische Institut leitete. Er starb am 14.8.1915 in seiner Heimatstadt Freiburg. Bereits während seiner Freiburger Schulzeit lernt Thumb den späteren Psychologen Karl Marbe kennen, in Marburg arbeitet er mit dem Physiologen Narziss Ach zusammen. Thumb ist in der Linguistik vor allem als Indogermanist bekannt; sein Schwerpunkt war das Neugriechische. Auf seine Anregung hin wurden Assoziationsexperimente zur Erforschung der Analogie durchgeführt (Thumb & Marbe 1901, Thumb 1908), die später Esper (1973: 80) mit englischsprechenden Versuchspersonen erfolgreich nachvollzog. Er unterstützte ferner Marbes Projekt zur Erforschung des Sprachrhythmus mit eigenen Untersuchungen zum Griechischen (Thumb 1913) und trug eindrucksvolle Plädoyers für die Nutzung von Statistik und Experimenten in der philologischen Forschung vor (Thumb 1910, 1911, 1913), wobei es ihm letztlich um eine Untersuchung „der allgemeinen Gesetze der Sprache“ (Thumb 1908: 12) ging. Biographische Informationen sind nach wie vor vergleichsweise spärlich. (*Albert Thumb zum Gedächtnis* 1915; Killy & Vierhaus 1999: 28; Esper 1973: 64, Hatzidakis 1917; Jahnke 1998: 154-155; Kotrasch 2004; Marbe 1945, Murray 1978: XIII-XV; Verzeichnis der Schriften 1920/21). Aus Sicht der quantitativen Linguistik ist auch bedeutsam, daß Paul Menzerath, langjähriger Phonetiker in Bonn, zu seinen Schülern gehörte.

In der Quantitativen Linguistik ist Albert Thumb wenig bekannt. Seine Verdienste um Analogie und Sprachrhythmus sind in Best (2005) bei der Vorstellung von Karl Marbe z.T. bereits behandelt; auf diesen Beitrag sei zur Vermeidung unnötiger Wiederholungen ausdrücklich verwiesen. Außer mit dem Sprachrhythmus hat Thumb (1913: 154ff) sich auch mit dem musikalischen Akzent und der Satzmelodie befasst; die Tabellen hierzu enthalten jedoch keine absoluten Werte, so dass anders als bei den rhythmischen Einheiten eine Übereinstimmung mit den Theorien von Altmann (1988) und Wimmer u.a. (1994) nicht geprüft werden kann; seine Daten zum Sprachrhythmus stützen diese Theorien weitgehend (Best 2004). Thumbs Auffassung von Wissenschaft kommt in dem folgenden Plädoyer für die Verwendung statistischer Methoden in der Linguistik zum Ausdruck, das richtungsweisend hätte werden können, wenn es denn die ihm gebührende Aufmerksamkeit gefunden hätte:„R i c h t i g e Anwendung der statistischen Methode bringt immer ein Resultat, das in irgendeiner Weise bemerkenswert ist (...) Zahlen beweisen, da sie unbestreitbare Tatsachen lehren, mit denen man rechnen m u ß“ (Thumb 1911: 3).

Literatur

Albert Thumb zum Gedächtnis. (1915). Nachrufe in der Tagespresse. Freiburg i.B.: C.A. Wagner.

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

Best, K.-H. (1973). *Probleme der Analogieforschung*. München: Hueber.

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

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

Best, K.-H. (2004). Anpassungen der Hyperpoisson-Verteilung an Albert Thumbs Tabellen zur Verteilung rhythmischer Einheiten in griechischen Texten. Unveröffentlicht.

Best, K.-H. (2005). Karl Marbe (1869-1953). (in this volume).

Best, K.-H. (2005). Längen rhythmischer Einheiten. In: Altmann, G., Köhler, R., Piotrowski, R. (eds.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch.* Berlin/ New York: de Gruyter (erscheint).

Esper, E. A. (1973). *Analogy and Association in Linguistics and Psychology*. Athens: University of Georgia Press.

Hatzidakis, G. N. (1917). Albert Thumb. *Indogermanisches Jahrbuch IV/ Jahrgang 1916.* 235-241.

Jahnke, J. (1998): Wilhelm Wundts akademische Psychologie 1886/1887: Die Vorlesungsnachschriften von Albert Thumb. In: Jahnke, Jürgen, Fahrenberg, Jochen & Bauer, Eberhard (Hrsg.), *Psychologiegeschichte: Beziehungen zu Philosophie und Grenzgebieten*. Passauer Schriften zur Psychologiegeschichte Bd. 12. München/Wien: Profil, 1998, 151-168.

Killy, W., Vierhaus, J. (Hrsg.) (1999): *Deutsche Biographische Enyzklopädie (DBE)*. Bd. 10: Thibaut - Zycha. München: 1999: K. G. Saur.

Kotrasch, B. (2004). Albert Thumb − Sein Leben und sein Werk: Das ‚Handbuch der neugriechischen Volkssprache' in seinen Briefen. *Göttinger Beiträge zur Byzantinischen und Neugriechischen Philologie 4* (erscheint).

Marbe, K. (1904). *Über den Rhythmus der Prosa*. Giessen: J. Ricker'sche Verlagsbuchhandlung.

Marbe, K. (1913). Die Bedeutung der Psychologie für die übrigen Wissenschaften und die Praxis. In: *Fortschritte der Psychologie und ihrer Anwendungen I, 1*. Unter Mitwirkung von W. Peters hrsg. von Karl Marbe. Leipzig/ Berlin: Teubner.

Marbe, K. (1945). *Selbstbiographie des Psychologen Geheimrat Prof. Dr. Karl Marbe in Würzburg*. Hrsg. im Namen der Kaiserlich Leopoldinisch-Carolinisch-Deutschen Akademie der Naturforscher von Emil Abderhalden. Halle (Saale).

Murray, D. D. (1978). *Introduction*. In Neuausgabe von: Thumb & Marbe (1901).

Thumb, A. (1908). Die experimentelle Psychologie im Dienste der Sprachwissenschaft. *Sitzungsberichte der Gesellschaft zur Beförderung der gesamten Naturwissenschaften zu Marburg, No. 2, Februar 1907*: *11-23.* Marburg: Universitätsdruckerei Joh. Aug. Koch 1908.

Thumb, A. (1910). Beobachtung und Experiment in der Sprachpsychologie. In: *Festschrift Wilhelm Viëtor zum 25. Dezember 1910.* Dargebracht von F. Brie u.a. (S. 19-26). Marburg: Elwert.

Thumb, A. (1911). Experimentelle Psychologie und Sprachwissenschaft. Ein Beitrag zur Methodenlehre der Philologie. *Germanisch-Romanische Monatsschrift 3*: 1-15; 65-74.

Thumb, A. (1913). *Satzrhythmus und Satzmelodie in der altgriechischen Prosa. In: Fortschritte der Psychologie und ihrer Anwendungen I,3*. Unter Mitwirkung von W. Peters hrsg. von Karl Marbe (S. 139-168). Leipzig/ Berlin: Teubner.

Thumb, A., Marbe, K. (1901). *Experimentelle Untersuchungen über die psychologischen Grundlagen der sprachlichen Analogiebildung*. Leipzig: Engelmann (Neuausgabe: David D. Murray. Amsterdam: John Benjamins 1978).

Verzeichnis der Schriften von Albert Thumb. *Indogermanisches Jahrbuch VIII/* 1920/21: 272-277. (Ohne ausdrückliche Nennung eines Verfassers)

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

Karl-Heinz, Best, Göttingen
Brita Kotrasch, Göttingen

XIII. Jan Czekanowski (1882–1965) – a pioneer of multidimensional taxonomy

Jan Czekanowski earned himself a lasting place in the history of the social sciences as an eminent scholar of African anthropology and as the creator of the Polish school of anthropology (Jones 2002; Czekanowska-Kuklińska, Bar 2002). He is known for having played an important role in saving the Polish-Lithuanian branch of the Karaim minority from Nazi extermination[2]. During his long scientific career he also encountered linguistic and statistical issues: he is the author of the first Polish textbook of applied statistics (Czekanowski 1913) and the initiator of linguistic research employing numerical taxonomy (Czekanowski 1947). For many years he was the vice-president of the Polish Statistical Society.

It was during his studies on the history of Slavonic languages that he presented a classification of these languages based on grammatical and syntactical features. The foundation of this classification was comprised of 23 linguistic features appearing in the Balto-Slavonic group and in other languages, as well as 16 other features specific to the Balto-Slavonic group only (*ibid*. 15–17). The author of the list of discriminant features which Czekanowski used was his close associate T. Lehr-Spławinski. These linguistic features were assigned binary values (present or absent) and on this basis 12 groups of the family of the Indo-European languages were classified (*ibid*. 18). He proceeded in a similar manner with a list of 20 linguistic features provided to him by J. Kuryłowicz (*ibid*. 20). Then he calculated the degrees of similarity between the individual languages, using to this end equation (1), which expresses the level of correlation (similarity) of two languages as well as a contingency table (Tab.1) containing the number of features shared between the languages compared (*ibid*. 21–22).

[2] Interrogated by German "Rassenforscher" ("race scientists"), Czekanowski managed to convince them that the Karaim people, although professing Judaism and using Hebrew as a liturgical language, was of Turkish origin. This helped the Karaims escape the tragic destiny of the European Jews and Roms (cf. the paper by A. Juzwa-Ogińska: www.promemoria.pl/arch/2003_7/kara/kara.html).

Table 1
Two-feature contingency table

		Language A		
Lang. B	Feature	present	absent	Sum
	present	a	b	a+b
	absent	c	d	c+d
	Sum	a+c	b+d	

(1) $$Q = \sin\frac{\pi}{2}\frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

After calculating the correlation coefficients for all the language pairs representing the most important Indo-European groups, the author constructed their distance matrix. Czekanowski's method, based on Lehr-Spławinski and Kuryłowicz's lists of features, confirmed the existence of clusters of related languages, but in contrast to the traditional approach it introduced a numerical element and in this way expressed the degree of q u a n t i t a t i v e relatedness of languages. His results undermined the thesis of a close relationship between the Baltic and Germanic groups, placing the western Slavonic languages between the Germanic region and the territorial domain of the Baltic languages (*ibid*. 31).

It is worth recalling at this point that Czekanowski's classification originated within a particular historical and ideological context, as the representatives of Polish science were defending the thesis of the autochthonic status of the Slavs, according to which the original homeland of the Slavonic peoples was Central Europe, rather than some vaguely defined region to the east postulated by, among others, German scholars (Czekanowski 1947: 3–12)[3]. This earlier feud over the primal homeland has now lost its acuteness, and the politicisation of linguistics, though still present, has moved into other areas, such as those relating to globalisation and the protection of ethnic languages from the pressure of English. Czekanowski's methodology should be recognised, however, as an ideology-free and lasting contribution to the development of quantitative linguistics.

In the 1940's Czekanowski's concepts were a novelty in the lecture halls of the world and they foretold the imminent arrival of advanced multidimensional methods. Indeed, these methods form the basis of linguistic typology today, and are also applied in other fields in which taxonomies play a key role. But if one looks at Polish linguistics after 1947, one sees that the promising development of the application of multidimensional scaling practically came to a standstill which, by the way, has remained so to this day. Exceptions to this rule, e.g. J. Woronczak's ideas which, for various reasons, did not reach practical realisation, as well as a few other studies (Jassem, Łobacz 1995; Pawłowski 2004), have not altered this state of affairs.

[3] It is worth adding that in later years J. Woronczak also expressed his opinions in the discussion over the Slavs primal homeland. With his characteristic sense of humour he said that it is difficult to present convincing evidence in this question, but if indeed some Slavonic people did remain in the same region for over two thousand years, this would not speak favourably of them, but prove that they were incapable of expanding and conquering new territory. The current problems connected with this issue are to be found in the numerous works of W. Mańczak.

References

Czekanowska-Kuklińska, A., Bar, J. (2002). Jan Czekanowski (1882-1965) – Antropolog i etnograf, profesor uniwersytetów we Lwowie, Lublinie i Poznaniu [Jan Czekanowski (1882-1965) – anthropologist, ethnologist and professor of the Universities in Lvov, Lublin and Poznań]. In: *Etnografowie i Ludoznawcy polscy – Sylwetki, Szkice autobiograficzne: 52-56.* Kraków: Polskie Towarzystwo Muzyki Ludowej.

Czekanowski, J. (1913). *Zarys metod statystycznych w zastosowaniu do antropologii* [An outline of statistical methods applied in anthropology]. Warszawa: Towarzystwo Naukowe Warszawskie.

Czekanowski, J. (1947). *Polska synteza slawistyczna w perspektywie ilościowej* [Polish Slavonic synthesis in quantitative perspective]. Kraków: Polska Akademia Umiejętności, Rozprawy Wydziału Historyczno-Filozoficznego, Seria II, t.XLVI, nr 2.

Jassem, W., Łobacz, P. (1995). Multidimensional scaling and its Applications in a Perceptual Analysis of Polish Consonants. *Journal of Quantitative Linguistics 2/2, 105-124*.

Jones, A. (2002) (ed.). *Jan Czekanowski, africanist ethnographer and physical anthropologist in early twentieth-century Germany and Poland*. Leipzig: Institut für Afrikanistik.

Pawłowski, A. (2004). *Multidimensional scaling in the analysis of language corpora (from word frequencies to the map of Europe)*. Research paper presented at the 28. Annual Conference of the German Classification Society in Dortmund.

Adam Pawłowski, Wrocław

XIV. Georg Philipp Harsdörffer (1607-1658)

Georg Philipp Harsdörffer wurde am 1.11.1607 in Fischbach (heute Nürnberg) geboren. Er nahm 1623 sein Studium an der Universität Altdorf (Nürnberg) auf, das er 1626 in Straßburg fortsetzte. Gegenstand seiner Studien waren Jura, Mathematik und Philologie. 1627 begab er sich auf eine 5 Jahre dauernde Bildungsreise nach Frankreich, Großbritannien und Italien, in die Niederlande und die Schweiz. 1630 verbrachte er ein Semester an der Universität Siena. 1633 kehrte er zurück, wurde 1637 in Nürnberg Assessor am Untergericht; 3 Jahre später wurde er ans Stadtgericht versetzt; ab 1655 Mitglied des Kleinen Rats. Neben diesen Tätigkeiten arbeitete er als Schriftsteller, Übersetzer und Wissenschaftler. Er ist einer der wichtigsten Autoren des Barock.

1641 wurde Harsdörffer von Fürst Ludwig I von Anhalt-Köthen in die *Fruchtbringende Gesellschaft* aufgenommen (Weimar, Köthen; Mitgliedsname: *der Spielende*); er war außerdem Mitglied der *Teutschgesinnten Genossenschaft* (Hamburg; Mitgliedsname: *der Kunstspielende*) und gründete zusammen mit Johannes Klay 1644 den *Pegnesischen Blumenorden* (Nürnberg; Pseudonym: *Strephon*).

Für die Quantitative Linguistik ist Harsdörffer von Bedeutung, weil er kombinatorische Ideen aufgriff und damit u.a. Leibniz beeinflusste (Best 2005). Die Kombinatorik findet auf der Ebene der Buchstaben, Wörter und Verse Anwendung:

1. Harsdörffer stellt Überlegungen dazu an, wie viele Wörter man bilden kann, wenn das Alphabet 24 Buchstaben enthält und beruft sich dazu auf Lauremberg, Puteanus und Etten (Harsdörffer 1651/ 1990: 513-516; Harsdörffer 1653: 59f.), deren Angaben allerdings falsch sind; erst Leibniz hat die richtige Zahl berechnet (Knobloch 1973: 41-43; Zeller 1974: 172). Auf die Tradition kombinatorischen Denkens, in der Harsdörffer damit steht, wurde bereits in

Best (22003; 2005) verwiesen. Darüber hinaus zitiert Rieger (1991: 185) Harsdörffers Überlegungen dazu, wie viele Silben man aus den Buchstaben des Alphabets bilden kann.
2. *Fünffacher Denckring der Teútschen Sprache* (Harsdörffer 1651: 517). Dies ist eine Wortbildungsmaschine, bei der „264 sprachliche Einheiten (Präfixe, Suffixe, Buchstaben und Silben) auf fünf Scheiben verteilt werden, um per Kombinatorik ... deutsche Wörter zu erzeugen, auch inexistente, die zu poetisch-kreativen Zwecken benutzt werden könnten" (Eco 1997: 148f.). Ausführlicher befasst sich Hundt (2000: 281ff.) mit dem Denkring; er gibt in Fußnote 135 (S. 285) an, dass man damit 101606400 Wörter bilden könne und würdigt ihn so: „Der ‚fünffache Denckring' erfüllt einerseits den Zweck, alle Wortbildungs- und Denkmöglichkeiten, die in der deutschen Sprache enthalten sind, mechanisch reproduzierbar zu machen. Er hat damit eine sprach- und erkenntniskonstitutive Funktion. Andererseits kann er auch ein praktisches Arbeitsmittel für den Spracharbeiter sein, der Reimwörter sucht. Neben die Generierung der Semantik tritt damit das mechanische Auffinden lautähnlicher Ausdrucksseiten" (Hundt 2000: 284). Bezogen auf Leibniz' Berechnung der Wortbildungsmöglichkeiten mit Hilfe des Denkrings bestimmt Rieger (1991: 190) dieses Verfahren als charakteristisch und wesentlich für diese Zeit: „Quantitative Bestandaufnahmen wie diese gehören nebst den immer wiederkehrenden Strategien der Zerstückelung und Kombination zum Programm barocker Sprachanalyse." Dencker (2002: 425) weist aber darauf hin, dass solche „Drehscheiben, Sprach- und Lesemaschinen" zu Harsdörffers Zeiten bereits eine längere Tradition aufweisen und bis in die Gegenwart Nachfolger gefunden haben. Zeller (1974: 169) sieht in kombinatorischen Bestrebungen zur Bildung von Wörtern gar einen Vorläufer moderner Sprachtheorie: „Die Sprachauffassung, die sich hier zeigt, findet sich bei Humboldt wieder, der seinerseits auf die moderne Sprachbetrachtung der generativen Grammatik eingewirkt hat."
3. *Proteusverse*. Gardt (1994: 219) stellt diese Dichtform so vor: „Proteusverse sind Verse, in denen sich die Wörter zu immer neuen Sinnkombinationen umstellen lassen, Ziel ist eine möglichst hohe Zahl von Kombinationen."

In Harsdörffer (1648-53, Teil I: 51f.) wird folgendes Beispiel gegeben:

„Auf Angst/ Noht/ Leid/ Haß/ Schmach/ Spott/ Krieg/
Sturm/ Furcht/ Streit/ Müh'/ und Fleiß
folgt Lust/ Raht/ Trost/ Gũst/ Ruhm/ Lob/
Sieg/ Ruh/ Muth/ Nutz/ Lohn/ und Preiß."

(Anm.: „Müh'" in hier modernisierter Schreibweise.)

Enzensberger (2002: 11; 23, Anm. 7; vgl. Harsdörffer 1653: 60) nennt und kommentiert das Beispiel:

„Ehr, Kunst, Geld, Guth, Lob, Waib und Kind,
Man hat, sucht, fehlt, hofft und verschwindt."

Das Prinzip besteht darin, einsilbige Wörter zu wählen, die man nach Belieben permutieren kann, um immer wieder neue Verse zu erhalten, wobei in diesem Fall „und" sowie die Reimwörter ihre Position wahren müssen. Dieses Beispiel hat auch Leibniz aufgegriffen (Best 2005). Ein besonders langes Gedicht dieser Art ist Quirinus Kuhlmanns *Der XLI. Libes-kuß* (Dencker 2002: 76-80). (Zu den genannten und weiteren Beispielen vg. auch Zeller 1974: 174-177.)

Harsdörffer steht in einer Tradition kombinatorischen Denkens, die von der Antike bis in die Gegenwart reicht und für linguistische, mathematische, poetische, philosophische und theologische Vorstellungen Bedeutung hat.

Literatur

Best, K.-H. ([2]2003). *Quantitative Linguistik. Eine Annäherung.* 2., überarbeitete und erweiterte Auflage. Göttingen: Peust & Gutschmidt.

Best, K.-H. (2005). Gottfried Wilhelm Leibniz (1646-1716). *Glottometrics 9, 80-83.*

Dencker, K. P. (Hrsg.) (2002). *Poetische Sprachspiele. Vom Mittelalter bis zur Gegenwart.* Stuttgart: Reclam.

Eco, U. (1997). *Die Suche nach der vollkommenen Sprache*. München: dtv.

Enzensberger, H. M. (2002). *Einladung zu einem Poesie-Automaten.* http://jacketmagazine.com/17/enz-robot.html.

Gardt, A. (1994). *Sprachreflexion in Barock und Frühaufklärung. Entwürfe von Böhme bis Leibniz.* Berlin/ New York: de Gruyter.

Harsdörffer, G. P. (1648-53; Reprint 1969). *Poetischer Trichter.* Reprografischer Nachdruck der Original-Ausgabe Nürnberg 1648-1653. Darmstadt: Wissenschaftliche Buchgesellschaft.

Harsdörffer, G. P. (1651; Reprint 1990). *Delitiae Mathematicae et Physicae. Der Mathematischen und Philosophischen Erquickstunden Zweyter Teil.* Neudruck der Ausgabe Nürnberg 1651, hrsg. und eingeleitet von Jörg Jochen Berns. Frankfurt: Keip Verlag.

Harsdörffer, G. P. (1653; Reprint 1990). *Delitiae Mathematicae et Physicae. Der Philosophischen und Mathematischen Erquickstunden Dritter Teil.* Neudruck der Ausgabe Nürnberg 1653, hrsg. und eingeleitet von Jörg Jochen Berns. Frankfurt: Keip Verlag.

Hundt, M. (2000). *„Spracharbeit“ im 17. Jahrhundert. Studien zu Georg Philipp Harsdörffer, Justus Georg Schottelius und Christian Gueintz.* Berlin/ New York: de Gruyter.

Knobloch, E. (1973). *Die mathematischen Studien von G. W. Leibniz zur Kombinatorik. Auf Grund fast ausschließlich handschriftlicher Aufzeichnungen dargelegt und kommentiert.* Wiesbaden: Franz Steiner Verlag.

Rieger, S. (1991). Nachwort. In: Schottelius, Justus Georg, *Der schreckliche Sprachkrieg. Horrendum Bellum Grammaticale Teutonum antiquissimorum* (S. 181-205). Leipzig: Reclam.

Zeller, R. (1974). *Spiel und Konversation im Barock. Untersuchungen zu Harsdörffers „Gesprächsspielen“.* Berlin/ New York: de Gruyter.

Karl-Heinz Best, Göttingen

Glottometrics 9, 2005, 89-90

Book Review

Don McNicol, *A Primer of Signal Detection Theory.* London: Lawrence Erlbaum Associates, Publishers 2005, 240 pp. $29.95. By **Jana Kusendová.**

For almost 45 years, signal detection theory (SDT) has been familiar in many branches of psychology. The initial application of SDT was in the interpretation of sensory processes; the theory is widely used in memory research and in studies of processing verbal information. SDT is concerned with the measurement of sensitivity and response bias. Although there have been many research papers and publications in this domain, *A Primer of Signal Detection Theory*, originally published in 1972 and republished in 2005, was the first comprehensive overview of the methods of SDT. Sticking to the terminology of Green & Swets' *Signal Detection Theory and Psychophysics* (1966), *A Primer of Signal Detection Theory* covers signal detection theory at an introductory level, assuming only basic skills in statistics and algebra.

This book can serve as an excellent study guide for students. It can be equally useful for teachers and researchers in psychology who work with SDT theory. It is very detailed and complete from a didactical point of view. The author tries to explain the method from various points of view. The examples are explained in different variations, a procedure that supports understanding. Theoretical results are always followed by their applications. At the end of the book there are useful appendices and tables that should help with solving and understanding problems.

A Primer of Signal Detection Theory falls into two parts. Chapters 1 to 5 introduce the basic ideas of signal detection theory and its measures. The second part (chapters 6 to 8) discusses three theories: choice theory, threshold theory and the theory of Thurstonian scaling.

In the first part of the book, a theory is presented about the way in which choices are made. The book begins with a very extensive discuission of the problems of SDT and its individual steps. Then the methodological background of this theory is explicated (rules and criteria, some definitions, decision rules, a description of signal and noise). These chapters include the distributions of signal and noise, several psychophysical methods like yes-no tasks (only for two responses), and rating scale tasks for more than two responses. One used in psychological experiments and tests is the forced-choice task, in which the observer is confronted with a stimulus and a multiple choice, or the observer is confronted with two or more stimulus intervals. This second procedure can be used to turn the yes-no task into a forced-choice task. Each chapter is accompanied by clear graphs and schemes. Also discussed are the sensitivity measures, which are non-parametric and make no assumptions about the underlying distributions, gaussian distributions with equal or unequal variances, and so on. The various modes of observing behaviour represent Receiver-Operating Characteristic (ROC) curve (i.e. the various degrees of response bias). From the shape of the ROC curves we can tell which variance is greater (if we have unequal variances). When working with unequal variances it is better to use the double-probability plots.

In the last section of this first part, a rating scale task is worked through and a number of problems, which an experimenter may encounter in the previous chapters are discussed. This part is a summarization of the previous part, looking at ways of performing a rating scale experiment. This experiment is concerned with the short-term memory for sequences of digits. This interest-

ing technique for estimating short-term memory was devised by Bushke (1963). Here the author is raising and answering the question: how many trials of signal and noise are necessary to get a good estimate of the ROC curve? If the experimenter uses a non-parametric measure, rather than parametric measures, he can reduce his experiment to more realistic proportions. We can find here a methods for solving one problem; readers can see which of these methods is the best in the concrete situation.

The second part of the book discusses three advanced theories; the first of these is choice theory. Signal detection theory resembles choice theory, which had been presented by Luce (1959, 1963). The sixth chapter is concerned with explanation of methods of choice theory uses (concrete Broadbent's (1967) and Ingleby's (1968) methods). Broadbent´s theory is very interesting; it is based on responses of observers, in which common words appear more often than uncommon words, regardless of whether the stimulus itself is a common or an uncommon word. This choice theory can be applied to some types of experimental tasks to which detection theory cannot, that is why this choice theory is so important. For this purpose the logistic distribution (in the previous chapters Gaussian was used) is used.

SDT and high threshold theory stand in contrast to one another. The measures extracted from SDT differ from those of classical psychophysics, which is based on the existence and measure of a threshold. But there exist cases in which a model based on detection theory is more adequate than one based on high threshold theory. For threshold measurement one of the following methods can be used: the method of limits (which cannot be easily combined with a forced-choice procedure), or the method of constant stimuli.

The third topic of this second part is an exam of Thurstonian scaling procedures, which extend the signal detection theory. Thurstone proposed two laws: the Law of Categorical Judgement and the Law of Comparative Judgment, where sensory evidence is variable from moment to moment.

This is not a book that can be read hastily; rather it must be studied. Readers are recommended to attempt to compute all examples "with pencil and paper." At the end of each chapter there are self-study problems relevant to the topic of the chapter, allowing the reader to find out for themselves, whether they have really understood that section of the book.

The methods described here can be used in many others fields as well. They are well suited for research where the data analysis is based on questionnaires. For example, signal detection theory could be used in the field of slang research, where the reactions of test subjects may be either binary or multiple.

References

Bushke, H. (1963). Relative attention in immediate memory determined by the missing scan method. *Nature 200, 1129-1130.*

Broadbent, D.E. (1967). Word-frequency effect and response bias. *Psychological Review 74, 1-15.*

Green, D.M., Swets, J.A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Ingleby, J.D. (1968). *Decision-making processes in human perception and memory.* Unpublished Ph.D. thesis: University of Cambridge.

Luce, R.D. (1959). *Individual choice behaviour.* New York: Wiley.

Luce, R.D. (1963). Detection and recognition of human observers'. In: Luce, R.D., Bush, R.R., Galanter, E. (eds.), *Handbook of mathematical psychology, Vol. I.* New York: Wiley.