Glottometrics 21
2011

RAM-Verlag

ISSN 2625-8226

Dedicated to

Reinhard Köhler

on the occasion of his 60th birthday

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. 21 (2011), Lüdenscheid: RAM-Verlag, 2011. 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 21 (2011) **ISSN 2625-8226**

Contents

Glottometrics 21, 2011

Glottometrics 21, 2010, 1-13

Silben-, Wort- und Morphlängen bei Lichtenberg

Karl-Heinz Best, Göttingen

Abstract. The aim of this paper is to show that the lengths of morphs, syllables and words in German abide by law. The findings lend support to the theory of word length distributions (Wimmer et alii 1994, Wimmer & Altmann 1996) once more.

1. Verteilung von Einheiten unterschiedlicher Länge

Das Göttinger *Projekt Quantitative Linguistik* hat als einen seiner Schwerpunkte die Verteilung sprachlicher Einheiten unterschiedlicher Komplexität in Texten und Lexika. Im Vordergrund standen dabei Untersuchungen zu Satz- (Best 2005c) und Wortlängen (Best 2005d) sowie zu den Längen rhythmischer Einheiten (Best 2005a); andere Einheiten (Morph-, Satzglied-, Silben- und Teilsatzlängen) konnten nicht so oft bearbeitet werden. Um hier einen gewissen Ausgleich zu schaffen, werden die Morph- und Silbenlängen in 20 kurzen Texten aus den *Sudelbüchern* von G. Chr. Lichtenberg (Heft H, 1784-1788, Lichtenberg 1971, 175-211) einer Untersuchung unterzogen; zusätzlich werden die Wortlängen in den gleichen Texten vorgestellt. Die Auswahl erfolgte willkürlich; die Texte sollten nur nicht zu kurz oder zu lang sein.

2. Theoretische Grundlagen

Als theoretische Grundlage der Untersuchung wurden wie in allen Arbeiten des Göttinger Projekts die Beiträge von Wimmer et alii (1994) sowie von Wimmer & Altmann (1996) gewählt. Ihre generelle Hypothese besteht darin, dass die unterschiedlichen Längen sprachlicher Einheiten in Texten gemäß theoretisch begründeten Verteilungen vorkommen. Diese Annahme konnte in vielen Untersuchungen zu über 50 Sprachen oder Sprachentwicklungsstadien gestützt werden. Die unterschiedlichen Randbedingungen (Sprache, gewählte Einheit, Zeit, Autor, Textsorte...) führen dazu, dass nicht in allen Untersuchungen immer das gleiche Modell für die jeweiligen Daten gewählt werden kann. In allen Sprachen gibt es Randbedingungen, die oft zur Modifikation eines Modells oder zur Wahl eines anderen führen können. Es gibt jedoch einschlägige Erfahrungen, auf die man sich bei neuen Daten stützen kann.

3. Silbe - Silbenlänge

Bisher konnten nur zwei Untersuchungen zu Silbenlängen im Deutschen veröffentlicht werden, die sich beide mit Pressetexten befassen (Best 2001c, Cassier 2001). In beiden Arbeiten wurde in gleicher Weise vorgegangen: Eine Silbe ist dann gegeben, wenn ein Vokal oder Diphthong vorhanden ist. Da es hier um eine Untersuchung geschriebener Sprache geht, muss eine Silbe nicht auch noch durch das Vorhandensein eines Sonanten in silbischer Funktion angesetzt werden, die unter bestimmten Bedingungen in der gesprochenen Sprache

vorkommen; silbische Sonanten werden stattdessen als Phonemfolge aus Schwa und Sonant gewertet.

Um die Silbenlänge bestimmen zu können, müssen die Grenzen der Silben eindeutig festgelegt werden. Die Länge der Silben wurde danach bestimmt, aus wie vielen Phonemen sie bestehen; das zugrundeliegende Phonemsystem ist in Best (2001c: 19) dargestellt; jedes Phonem wird genau einer Silbe zugeordnet (vgl. dazu Best 2001c, Cassier 2001), d.h. sogenannte „Silbengelenke“ (Eisenberg 2005, S. 47) werden nicht als ambisyllabisch gewertet. Affrikaten und Diphthonge werden als jeweils ein Phonem und nicht als Phonemfolge gewertet. Die Trennung der Wörter in Silben erfolgt nach den Regeln, die *Duden. Das Aussprachewörterbuch* (2005, 58-60) entwickelt.

3.1. Modellierung der Silbenlängen

In den bisherigen Untersuchungen (Best 2001c, Cassier 2001) konnte die 1-verschobene Conway-Maxwell-Poisson-Verteilung

$$(1) \quad P_x = \frac{a^{x-1}}{[(x-1!)]^b\, T_1}, \quad x = 1, 2, ... \quad \text{mit} \quad T_1 = \sum_{j=0}^{\infty} \frac{a^j}{(j!)^b}$$

mit sehr guten Ergebnissen an die Daten angepasst werden; sie wird daher auch in dieser Bearbeitung von Lichtenbergs Texten angewendet. Dieses Modell hat sich bei Silbenlängen im Deutschen auch in einigen weiteren Untersuchungen, die nicht veröffentlicht werden konnten, mit fast immer guten Ergebnissen bewährt (Cassier 1998 mit weiteren 61 Textdateien, die in Cassier 2001 nicht wiedergegeben werden; Zuse 1998, Schneemann 2001). Die Anpassungen wurden mit dem *Altmann-Fitter* (1997) durchgeführt.

3.2. Anpassung der 1-verschobenen Conway-Maxwell-Poisson-Verteilung

Die Anpassung der 1-verschobenen Conway-Maxwell-Poisson-Verteilung an Textabschnitte aus Lichtenbergs *Sudelbüchern* erbrachte die folgenden Ergebnisse:

Tabelle 1
Silbenlängen in Lichtenbergs Sudelbuch H

	H 10; S. 178		**H 13, S.179**		**H 14, S. 179***		**H 15, S. 179**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	6	5.90	9	8.86			4	4.05
2	76	76.21	74	76.00	67	66.32	74	69.49
3	86	83.88	77	75.40	103	103.62	73	78.80
4	19	21.87	23	21.18	21	20.78	19	18.25
5	2	2.05|	0	2.43|	1	1.29	2	1.41
6	1	0.09|	1	0.14|				
a =	12.9132		8.5819		1.5623		17.1387	
b =	3.5524		3.1127		2.9620		3.9179	
X^2 =	0.777		1.207		0.077		0.995	
FG =	2		2		1		2	
P =	0.68		0.55		0.78		0.61	

Legende zu den Tabellen 1-5: (Chi² =) X^2 ist das Chiquadrat; *FG* die Zahl der Freiheitsgrade; *P* ist die Überschreitungswahrscheinlichkeit für das berechnete Chiquadrat; die Anpassung des Modells an eine Datei wird dann als erfolgreich gewertet, wenn $P \geq 0.05$. Mit *a*, *b* werden die Parameter dieser Verteilung angegeben. *x* ist die Zahl der Phoneme pro Silbe, n_x die beobachtete, NP_x die berechnete Zahl der Phoneme pro Silbe. Die senkrechten Striche in den Tabellen zeigen an, dass die entsprechenden Klassen zusammengefasst wurden.

*Der Text H 14 hat eine Silbe, die aus nur einem Phonem besteht, und 66 Silben mit zwei Phonemen. In diesem Fall ist keine Anpassung möglich. Deshalb wurden die beiden Längenklassen zusammengefasst.

Tabelle 2
Silbenlängen in Lichtenbergs Sudelbuch H

	H 19, S. 180		**H 52, S. 184f.**		**H 53, S. 185**		**H 66, S. 187**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	27	20.02	3	2.98	11	10.68	14	13.18
2	129	151.97	73	78.08	92	93.95	109	114.31
3	151	123.44	102	94.15	93	89.56	120	112.15
4	17	27.12	16	18.74	21	23.27	27	30.75
5	0	2.36|	1	1.06	2	2.40|	3	3.41|
6	1	0.10|			1	0.13|	1	0.19|
a =	7.5925		26.2261		8.7949		8.6743	
b =	3.2246		4.4429		3.2057		3.1442	
X^2 =	16.706		1.389		0.491		1.348	
FG =	2		2		2		2	
P =	0.00*		0.50		0.78		0.51	

*Die Anpassung des Modells an H 19 ist misslungen; auch eine Zusammenfassung der Klassen $x = 2$ und $x = 3$ führt zu keinem zufriedenstellenden Ergebnis. Das Gleiche gilt für den Versuch, ein anderes Modell anzupassen.

Tabelle 3
Silbenlängen in Lichtenbergs Sudelbuch H

	H 125, S. 193f.		**H 134, S. 195**		**H 135, S. 195**		**H 138, S. 196**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	12	11.18	11	9.88	5	4.75	6	5.62
2	126	139.24	84	93.03	55	56.79	61	67.05
3	149	130.00	108	97.39	69	64.17	84	76.59
4	21	26.66	27	28.21	14	18.25	22	22.18
5	1	1.92	2	3.48	3	2.05	1	2.55
a =	12.4591		9.4141		11.9494		11.9312	
b =	3.7382		3.1687		3.4026		3.3848	
X^2 =	5.741		2.839		1.862		2.235	
FG =	2		2		2		2	
P =	0.06		0.24		0.39		0.33	

Tabelle 4
Silbenlängen in Lichtenbergs Sudelbuch H

	H 146, S. 197f.		**H 147, S. 198**		**H 148, S. 198**		**H 150, S. S. 199f.**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	8	8.35	7	6.69	9	8.75	41	31.38
2	117	107.00	78	83.35	98	102.05	228	267.70|
3	83	96.83	80	72.07	111	105.36	316	271.00|
4	18	18.60	9	13.09|	16	26.34	67	78.85
5	5	1.19|	2	0.81|	10	2.41|	6	9.47|
6	1	0.03|			1	0.10|	1	0.59|
a =	12.8064		12.4634		11.6662		8.5312	
b =	3.8228		3.8494		3.4982		3.0750	
X^2 =	3.807		1.832		0.588		5.710	
FG =	1		1		1		1	
P =	0.05		0.18		0.44		0.02	

Tabelle 5
Silbenlängen in Lichtenbergs Sudelbuch H

	H 151, S. 200		**H 155, S. 201**		**H 181, S.205**		**H 191, 207f.**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	12	12.12	12	12.53	16	13.45	13	11.28
2	173	169.32	128	119.38	111	127.49|	73	78.96
3	186	191.24	95	105.15	151	130.49|	81	73.22
4	43	49.61|	24	23.00	31	36.33	17	20.82
5	9	4.53|	2	1.87|	2	4.02|	3	2.73
6	4	0.19|	1	0.07|	1	0.22|		
a =	13.9669		9.5305		9.4771		6.9998	
b =	3.6283		3.4356		3.2108		2.9161	
X^2 =	0.277		2.241		1.680		2.265	
FG =	1		2		1		2	
P =	0.60		0.33		0.19		0.32	

Als Ergebnis ist festzustellen, dass die Anpassung der 1-verschobenen Conway-Maxwell-Poisson-Verteilung an die Silbenlängen der 20 Texte in einem Fall (H19) misslungen ist; in einem weiteren Fall (H150) entspricht das Anpassungsergebnis mit $P = 0.02$ dem angegebenen Kriterium nicht; das Testergebnis ist aber auch nicht so schlecht, dass man die Anpassung als völlig gescheitert ansehen müsste. Insgesamt gesehen kann man das gewählte Modell, das sich bei modernen Pressetexten bewährt hat, auch für diese älteren Texte als geeignet ansehen.

4. Wort - Wortlänge

Als nächstes geht es um die Verteilung von Wörtern unterschiedlicher Länge in den gleichen Texten von Lichtenberg. Als Wort wird – wie in früheren Arbeiten (z.B. in Best (Hrsg.) 1997)

– das orthographische Wort bestimmt; Bindestriche, Apostrophe und Trennungsstriche signalisieren die Einheit eines Wortes. Die Wortlänge wird hier durch die Zahl der Silben bestimmt, die ein Wort enthält. Es ist zu bemerken, dass bei dieser Zählung keine exakte Bestimmung der Silbengrenzen nötig ist.

4.1. Modellierung der Wortlängen

Für Wortlängen im Deutschen hat sich in vielen Untersuchungen (u.a.: Schmidt (Hrsg.) 1996; Best (Hrsg.) 1997; Best (Hrsg.) 2001) gezeigt, dass an fast alle Daten die 1-verschobene Hyperpoisson-Verteilung

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

angepasst werden kann. Dabei sind *a* und *b* Parameter; ${}_1F_1(1;b;a)$ ist die konfluente hypergeometrische Funktion, d.h.

$${}_1F_1\,(1;b;a) = 1 + \frac{a}{b} + \frac{a(a+1)}{b(b+1)} + \ldots$$

und $b^{(x-1)} = b\,(b+1)\,(b+2)\ldots(b+x-2)$.

Gelegentlich konnten auch andere Modelle mit etwa gleich guten Ergebnissen gewählt werden, z.B. die Poisson-Verteilung bei älteren und die positive negative Binomialverteilung bei neueren Texten. Keine Verteilung kann aber nach den Erfahrungen im *Göttinger Projekt Quantitative Linguistik* so weit verbreitet auf deutsche Texte angewendet werden wie die Hyperpoisson-Verteilung. Aus diesem Grund wird sie auch hier wieder verwendet.

4.2. Anpassung der 1-verschobenen Hyperpoisson-Verteilung

Die Anpassung der Hyperpoisson-Verteilung an die 20 Texte in Lichtenbergs Sudelbuch H erbrachte folgende Ergebnisse:

Tabelle 6
Wortlängen in Lichtenbergs Sudelbuch H

	H 10; S. 178		**H 13, S.179**		**H 14, 179**		**H 15, S. 179**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	62	62.35	42	40.14	70	68.36	43	41.00
2	30	29.21	25	27.21	28	30.20	23	26.01
3	12	12.59	14	14.87	12	12.02	14	13.56
4	8	7.85	10	6.80	6	4.35	9	6.01
5			2	3.99	0	1.45|	1	3.42
6					1	0.61|		
a =	5.3921		2.8122		4.0149		2.9307	
b =	11.5100		4.1477		9.0879		4.6194	
X^2 =	0.054		2.813		1.369		3.668	
FG =	1		2		2		2	
P =	0.82		0.25		0.50		0.16	

Legende zu den Tabellen 6-10: Mit *a*, *b* werden die Parameter der Hyperpoisson-Verteilung angegeben. *x* ist die Zahl der Silben pro Wort, n_x die beobachtete, NP_x die berechnete Zahl der Silben pro Wort. Mehrmals müssen Längenklassen so zusammengefasst werden, dass kein Freiheitsgrad übrig bleibt; in diesen Fällen muss der Diskrepanzkoeffizient *C* als Kriterium für die Güte der Anpassung gewählt werden, der die Bedingung $C \leq 0.01$ erfüllen sollte. Alle anderen Angaben entsprechen denen zu Tabelle 1-5.

Tabelle 7
Wortlängen in Lichtenbergs Sudelbuch H

	H 19, S. 180		**H 52, S. 184f.**		**H 53, S. 185**		**H 66, S. 187**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	93	92.41	72	71.65	56	55.88	93	93.02
2	60	59.11	34	34.80	45	46.19	58	51.25
3	22	24.59	13	12.24	19	17.06	10	18.49
4	9	7.58	4	4.31	3	4.06|	6	4.96
5	2	2.31			1	0.82|	1	1.06|
6							1	0.22|
a =	1.1896		1.2726		0.6676		1.0446	
b =	1.8598		2.6199		0.8076		1.8959	
X^2 =	0.598		0.090		0.410		5.413	
FG =	2		1		1		2	
P =	0.74		0.76		0.52		0.07	

Tabelle 8
Wortlängen in Lichtenbergs Sudelbuch H

	H 125, S. 193f.		**H 134, S. 195**		**H 135, S. 195**		**H 138, S. 196**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	68	68.01	69	67.54	49	47.03	42	46.62
2	63	63.01	46	44.02	25	22.66	30	24.37
3	18	25.17|	13	17.78	5	10.22	11	11.63
4	14	6.41|	8	6.66	8	7.08	4	5.11
5	1	1.40|					3	2.08
6							0	0.79|
7							0	0.28|
8							1	0.13|
a =	0.7022		1.0625		7.0771		5.4928	
b =	0.7580		1.6304		14.6886		10.5076	
X^2 =	-		1.673		3.111		2.476	
FG =	-		1		1		3	
P =	-		0.20		0.08		0.48	
C =	0.0000		-		-		-	

Tabelle 9
Wortlängen in Lichtenbergs Sudelbuch H

	H 146, S. 197f.		**H 147, S. 198**		**H 148, S. 198**		**H 150, S. S. 199f.**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	61	58.72	61	61.47	91	91.02	184	183.63
2	29	29.50	35	29.79	44	44.01	115	111.17
3	8	14.82	6	11.59	8	13.95|	43	49.73
4	16	7.44|	4	3.77	9	3.29|	20	17.64
5	2	3.74|	1	1.05|	0	0.62|	6	5.18
6	1	1.88|	1	0.33|	1	0.11|	1	1.65
7	0	0.94|						
8	0	0.47|						
9	1	0.48|						
a =	342907.1178		1.9766		0.9213		1.7127	
b =	682585.6084		4.0784		1.9053		2.8291	
X^2 =	4.933		3.907		-		1.745	
FG =	1		2		-		3	
P =	0.03		0.14		-		0.63	
C =	-		-		0.0000		-	

Zu Text H146: Kriterium *P* erfüllt nicht ganz die Bedingung für eine gelungene Anpassung, ist aber auch nicht so schlecht, dass man die Anpassung ganz verwerfen muss. Auffällig ist bei diesem Text, dass die beiden Parameter sehr hohe Werte aufweisen. In solchen Fällen geht die Hyperpoisson-Verteilung in die geometrische Verteilung über (Wimmer & Altmann 1999, 282). Passt man die geometrische Verteilung an, so erhält man mit *P* = 0.07 ein besseres Ergebnis, das auch die Kriterien erfüllt.

Tabelle 10
Wortlängen in Lichtenbergs Sudelbuch H

	H 151, S. 200		**H 155, S. 201**		**H 181, S.205**		**H 191, 207f.**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	166	165.26	88	87.98	92	92.03	53	52.41
2	65	66.24	55	54.99	64	64.02	31	32.71
3	24	24.75	7	14.35|	13	20.15|	16	14.33
4	11	8.66	7	2.37|	8	4.10|	6	6.55
5	3	4.09	3	0.31|	3	0.62|		
6					1	0.08|		
a =	5.5118		0.4481		0.5751		1.4700	
b =	13.7509		0.7169		0.8267		2.3552	
X^2 =	0.973		-		-		0.336	
FG =	2		-		-		1	
P =	0.61		-		-		0.56	
C =	-		0.0000		0.0000		-	

Die 1-verschobene Hyperpoisson-Verteilung, die sich bei Wortlängenverteilungen im Deutschen und in vielen anderen Sprachen als geeignetes Modell erwiesen hat, ist, wie die Tabellen 6-10 zeigen, auch auf die Texte aus Lichtenbergs *Sudelbuch H* anwendbar.

Das gleiche Ergebnis wurde bei 41 Texten erzielt, in denen die Wortlänge durch die Zahl der Morphe pro Wort bestimmt wurde (Best 2001b, 2006a).

5. Morph – Morphlänge

Das gleiche Modell, das die Verteilung von Wortlängen im Deutschen (und vielen anderen Sprachen) repräsentiert, hat sich auch für die Verteilung von Morphlängen bewährt (Best 2000, 2001a, 2005b). Da Morphlängenverteilungen im Deutschen bisher lediglich für 21 Pressetexte und 18 Fabeln untersucht wurden, ist die empirische Basis dafür noch besonders gering. Die folgende Untersuchung erweitert die Datenbasis um 20 Texte.

4.1. Modellierung der Morphlängen

Die folgenden Tabellen dokumentieren die Anpassung der 1-verschobenen Hyperpoisson-Verteilung an die Morphlängenverteilungen bei Lichtenberg:

Tabelle 11
Morphlängen in Lichtenbergs Sudelbuch H

	H 10; S. 178		**H 13, S.179**		**H 14, 179**		**H 15, S. 179**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	55	54.06	40	40.02	41	40.30	51	51.04
2	82	80.60	93	90.82	77	75.68	74	72.37
3	48	52.97	50	54.14	58	56.94	51	47.27
4	27	22.33	21	18.57	25	26.79	12	20.05
5	5	6.93	4	4.47|	6	9.17	9	6.30
6	2	2.12	1	0.97|	4	2.46|	2	1.96
7					1	0.67|		
a =	1.1755		0.8084		1.2553		1.2108	
b =	0.7884		0.3562		0.6684		0.8540	
X^2 =	2.027		0.722		2.381		4.723	
FG =	3		2		3		3	
P =	0.57		0.70		0.50		0.19	

Legende zu den Tabellen 11-15. *x* ist die Zahl der Phoneme pro Morph, n_x die beobachtete, NP_x die berechnete Zahl der Phoneme pro Morph. Alle anderen Angaben entsprechen denen zu Tabelle 6-10.

Tabelle 12
Morphlängen in Lichtenbergs Sudelbuch H

	H 19, S. 180		**H 52, S. 184f.**		**H 53, S. 185**		**H 66, S. 187**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	66	74.09	67	67.09	56	56.86	70	70.62
2	154	136.18	73	73.10	80	81.23	130	121.59
3	78	87.56	46	49.39	62	57.62	65	77.56
4	30	34.12	32	24.19	24	27.18	36	30.36
5	12	9.54	9	13.23	11	9.61	7	8.57

6	4	2.51			2	2.71|	3	2.30
7					0	0.64|		
8					1	0.16|		
a =	0.9890		1.7787		1.4087		1.0134	
b =	0.5381		1.6325		0.9861		0.5886	
X^2 =	6.271		4.113		1.014		4.172	
FG =	3		2		3		3	
P =	0.10		0.13		0.80		0.24	

Tabelle 13
Morphlängen in Lichtenbergs Sudelbuch H

	H 125, S. 193f.		**H 134, S. 195**		**H 135, S. 195**		**H 138, S. 196**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	79	79.24	56	51.44	36	42.39	38	40.38
2	109	98.77	108	99.21|	75	58.83|	82	72.43
3	63	75.27	48	67.64|	26	39.53|	45	50.83
4	42	41.30	32	28.01	20	17.52	17	22.17
5	23	17.71	10	8.33	6	5.79	11	7.02
6	2	6.23	1	1.93|	2	1.53|	2	2.17
7	3	2.48	0	0.37|	1	0.41|		
8			0	0.06|				
9			2	0.01|				
a =	1.9607		1.0546		1.3025		1.1526	
b =	1.5730		0.5468		0.9386		0.6426	
X^2 =	7.634		2.180		1.970		5.554	
FG =	4		2		2		3	
P =	0.11		0.34		0.37		0.14	

Tabelle 14
Morphlängen in Lichtenbergs Sudelbuch H

	H 146, S. 197f.		**H 147, S. 198**		**H 148, S. 198**		**H 150, S. S. 199f.**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	59	58.66	55	53.00	72	72.41	129	130.93
2	82	79.25	72	77.81	92	92.53	244	247.65
3	52	58.38	52	47.02	70	64.19	197	186.86
4	33	29.56	19	17.89	24	30.56	82	88.06
5	12	11.41	3	4.97	12	11.08	36	30.17
6	3	3.55	1	1.32	2	3.24|	6	10.33
7	1	1.19			1	0.79|		
8					2	0.20|		
a =	1.6200		1.0271		1.5177		1.2553	
b =	1.1991		0.6996		1.1878		0.6637	
X^2 =	1.341		1.960		2.153		3.987	
FG =	4		3		3		3	
P =	0.85		0.58		0.54		0.26	

Tabelle 15
Morphlängen in Lichtenbergs Sudelbuch H

	H 151, S. 200		**H 155, S. 201**		**H 181, S.205**		**H 191, 207f.**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	106	102.59	87	86.47	63	63.44	51	48.43
2	164	158.72	100	97.21	131	131.92	84	79.77
3	101	118.60	56	63.59	94	92.49	42	52.65
4	70	58.42	35	29.33	43	39.00	30	21.73|
5	25	21.46|	10	10.44	7	11.76	4	8.42|
6	1	6.29|	2	3.03|	4	3.39		
7	0	1.53|	1	0.93|				
8	1	0.39|						
a =	1.4453		1.5639		1.0578		1.1014	
b =	0.9341		1.3910		0.5087		0.6687	
X^2 =	5.437		2.339		2.480		3.007	
FG =	2		3		3		1	
P =	0.07		0.51		0.48		0.08	

An alle 20 Textdateien kann die 1-verschobene Hyperpoisson-Verteilung mit guten Ergebnissen angepasst werden.

6. Zum Verhältnis von Morphen, Silben und Wörtern

Die folgende Tabelle gibt eine Übersicht darüber, in welchem Verhältnis die untersuchten Einheiten „Morph“, „Silbe“ und „Wort“ in den 20 Texten enthalten sind:

Tabelle 16
Das Verhältnis von Wort-, Morph- und Silbenlängen in Lichtenbergs Sudelbuch H

Text	**Wörter**	**Morphe**	**Silben**	**Morphe/Wort**	**Silben/Wort**	**Morphe/Silbe**
H10	112	219	190	1.96	1.70	1.15
H13	93	209	184	2.25	1.98	1.14
H14	117	212	192	1.81	1.64	1.10
H15	90	199	172	2.21	1.91	1.16
H19	186	344	325	1.85	1.75	1.06
H52	123	227	195	1.85	1.59	1.16
H53	124	236	220	1.90	1.77	1.07
H66	169	311	274	1.84	1.62	1.14
H125	164	321	309	1.96	1.88	1.04
H134	136	257	232	1.89	1.71	1.11
H135	87	166	146	1.91	1.68	1.14
H138	91	195	174	2.14	1.91	1.12
H146	118	242	232	2.05	1.97	1.04
H147	108	202	176	1.87	1.63	1.15
H148	153	275	245	1.80	1.60	1.12
H150	369	694	659	1.88	1.79	1.05

H151	269	468	427	1.74	1.59	1.10
H155	160	291	262	1.82	1.64	1.11
H181	181	342	312	1.89	1.72	1.10
H191	106	211	187	1.99	1.76	1.13

Im Ergebnis lässt sich feststellen, dass die Zahl der Morphe pro Wort in allen Fällen größer ist als die Zahl der Silben pro Wort. Natürlich ist dieses Ergebnis davon abhängig, wie man die beteiligten sprachlichen Einheiten definiert. Besonders Morphe kann man strenger oder auch weniger streng bestimmen. Im vorliegenden Fall wurden Morphe immer dann angenommen, wenn eine Segmentierung des betreffenden Wortes vertretbar erschien, obwohl die semantischen Bedingungen bisweilen Probleme bereiten.

Der Grund, warum im Deutschen die Zahl der Morphe die der Silben übersteigt, besteht darin, dass es eine Reihe von Morphen mit recht hoher Texthäufigkeit gibt, die keinen Vokal enthalten, und damit nicht als Silben auftreten. Umgekehrt gibt es nur wenige Morphe, die aus mehr als nur einer Silbe bestehen.

7. Zusammenfassung

Die Untersuchung bestätigt erneut die Hypothese, dass sprachliche Einheiten verschiedener Länge in Texten gesetzmäßig verteilt sind. Auch die Annahme, dass verschiedene Einheiten nicht unbedingt den gleichen Modellen folgen, lässt sich hier bestätigen: Silbenlängen unterliegen zumindest im Deutschen einem anderen Modell als die Morph- und Wortlängen. Möglicherweise kann diese Beobachtung darauf zurückgeführt werden, dass es sich bei Morphen und Wörtern um sprachliche Zeichen, bei Silben aber um phonetische Einheiten handelt.

Eine andere Beobachtung kann derzeit nur sehr vorläufig angestellt werden: Es scheint so zu sein, als ob die Wortlängenverteilungen etwas besser modelliert werden könnten als die Silbenlängenverteilungen; der Anteil schwacher oder gar gescheiterter Anpassungen im Falle der Silbenlängen erwies sich bisher als höher als derjenige bei Wortlängen, der kaum erwähnenswert ist. Dieser Eindruck von den Silbenlängenverteilungen stützt sich auch auf die nicht veröffentlichten Teile der Untersuchung von Cassier (1998) sowie auf die Arbeiten von Schneemann (2001) und Zuse (1998). Vorsicht ist dabei deshalb geboten, weil die Erfahrungen mit Silbenlängen insgesamt gesehen doch noch sehr gering sind, jedenfalls verglichen mit denen, die bereits mit Wortlängen gewonnen wurden.

Hinsichtlich der Wortlängen bestätigt die vorliegende Untersuchung die Ergebnisse von Ammermann (2001), der an 20 Briefe von Lichtenberg ebenfalls die Hyperpoisson-Verteilung mit guten Ergebnissen anpassen konnte. In diesem Fall zeigt sich, dass auch bei verschiedenen Textsorten das gleiche Modell Anwendung finden kann.

Die Längen sprachlicher Einheiten, so wie sie hier in Abschnitt 6. vorgestellt werden, sind lediglich eine historische Momentaufnahme. Am Beispiel der Veränderung der Wortlängen im Deutschen (Best 2006b) konnte gezeigt werden, dass dieser historische Prozess gemäß dem Piotrowski-Gesetz verläuft.

Literatur

Ammermann, Stefan (2001). Zur Wortlängenverteilung in deutschen Briefen über einen Zeitraum von 500 Jahren. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten* (S. 59-91). Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (Hrsg.) (1997), *Glottometrika 16. The Distribution of Word and Sentence Length.* Trier: Wissenschaftlicher Verlag Trier.

Best, Karl-Heinz (2000). Morphlängen in Fabeln von Pestalozzi. *Göttinger Beiträge zur Sprachwissenschaft 3, 19-30.*

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

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

Best, Karl-Heinz (2001c). Silbenlängen in Meldungen der Tagespresse. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten* (S. 15-32). Göttingen: Peust & Gutschmidt.

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

Best, Karl-Heinz (2005a). Längen rhythmischer Einheiten. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch* (S. 208-214). Berlin/ N.Y.: de Gruyter.

Best, Karl-Heinz (2005b). Morphlänge. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch* (S. 255-260). Berlin/ N.Y.: de Gruyter.

Best, Karl-Heinz (2005c). Satzlänge. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch* (S. 298-304). Berlin/ N.Y.: de Gruyter.

Best, Karl-Heinz (2005d). Wortlänge. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch* (S. 260-273). Berlin/ N.Y.: de Gruyter.

Best, Karl-Heinz (2006a). Wie viele Morphe enthalten Wörter in deutschen Pressetexten? *Glottometrics 13, 47-58.*

Best, Karl-Heinz (2006b). Wortlängen im Deutschen. *Göttinger Beiträge zur Sprachwissenschaft 13, 23-49.*

Cassier, Frank-Uwe (1998). *Silbenlängen in Meldungen der deutschen Tagespresse*. Staatsexamensarbeit, Göttingen.

Cassier, Falk-Uwe (2001). Silbenlängen in Meldungen der deutschen Tagespresse. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten* (S. 33-42). Göttingen: Peust & Gutschmidt.

Duden. Aussprachewörterbuch. 6., überarbeitete und aktualisierte Auflage. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag.

Eisenberg, Peter (2005). Phonem und Graphem. In: *Duden. Die Grammatik.* 7, völlig neu erarbeitete und erweiterte Auflage (S. 19-94.) Mannheim/Leipzig/Wien/Zürich: Dudenverlag.

Lichtenberg, Georg Christoph (1971). *Schriften und Briefe. Zweiter Band: Sudelbücher II. Materialhefte, Tagebücher.* München/Wien: Hanser.

Schmidt, Peter (Hrsg.) (1996), *Glottometrika 15. Issues in General Linguistic Theory and The Theory of Word Length*. Trier: Wissenschaftlicher Verlag Trier.

Schneemann, Okke F. (2001). *Sprachstatistische Untersuchungen zu Wort- und Silbenlängen in deutschen Musikzeitschriften.* Staatsexamensarbeit, Göttingen.

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

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

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

Zuse, Maria (1998). *Silbenlängen in deutschen und englischen Pressetexten der Gegenwart*. Staatsexamensarbeit, Göttingen.

Software

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

Internetadresse des Göttinger Projekts: http://wwwuser.gwdg.de/~kbest.

Glottometrics 21, 2011, 14-24

Sentence length and sentence structure in English prose

Viktor V. Levitsky, Yulia P. Melnyk[1]

Abstract. The article deals with determining author's style through frequency of clauses, length of complex sentence and complexity of the sentence proper. These parameters are viewed as the main criteria for distinguishing author's style.

***Keywords**: Sentence length, sentence complexity, frequency of clauses; author's style*

Introduction

Sentence length may depend on many different factors (Altmann, 1988). It may be expected, for example, that sentence length is associated with the genre of the text, authorial or functional styles. Therefore we may assume that longer sentences could be used more often in prose or scientific writings than in newspapers or poetry. Besides these commonly known factors influencing sentence length, some other ones have been stated.

But, first of all, sentence length depends on its structure. As G.A. Lesskis (1963: 106) states, "we could regard sentence length as a function of its structure". G. Lesskis specifically proved that there is a direct dependence between the average sentence length and the number of complex sentences in the text. Therefore, it is appropriate to study sentence length in close connection with its structure i.e. with its type.

One more important thing in sentence length study is the choice of a unit of sentence length measurement.

Sentence length can be counted in *number of words* or its *immediate constituents* (see e.g., Uhlířová 2001:266; Niehaus 2001:196). Up to this time it has not been stated precisely what kind of dependence exists between sentence length counted in number of words and sentence length counted in number of clauses (see Niehaus 2001:209-210). At the same time Altmann (e.g. 1988) has shown that when counting in terms of word numbers we skip a level and must add a new parameter to the model. Grzybek (see e.g. 2010) has shown that measuring in terms of word numbers yields different results for different text sorts. Thus, it will be useful from our point of view to count *sentence length* in number of words, and *sentence complexity* in number of clauses.

As sentence length depends first of all on its structure, i.e., on the sentence type, it would be also appropriate to study the frequency of different types of sentences in a text — simple, complex, compound and so called "complicated" sentences.

In this paper different types of complex sentences were chosen as an object of investigation.

Thus, the goal of the research is the study of frequency of different types of subordinate clauses in English prose, length and complexity of complex sentences selected from texts written by different authors. At the same time we shall regard all the totality of the texts under analyses as a text-invariant in relation to which texts written by one and the same author will be regarded as variants. Of course such division has a relative character. It means that the

[1] Address correspondence to Yulia P. Melnyk , e-mail: mljulia1@rambler.ru

totality of texts written by one and the same author could be considered as an invariant in relation to which a text of every separate author could be regarded as a variant.

Material of investigation

As a material of investigation we chose texts of English and American authors mainly of the first part of XX c. (1912-1962): E. Hemingway, T. Dreiser, F. Fitzgerald, A. Cronin, J. Steinbeck. More than two thousand complex sentences were taken from the works of the above-mentioned authors by the method of consecutive selection to study sentence length, together with more than five thousand clauses to study the frequency of subordinate clause use and complexity of the complex sentence.

It is supposed that an investigator uses a certain criterion to define sentence limits and its type. In accordance to the accepted classification of complex sentences in English grammar (see R. Quirk, 1985; M.Ya. Bloch 1983; Collins Cobuild English Grammar 1990; B.A. Illyish 1964) we distinguish: *Subject Clauses, Predicative Clauses, Object Clauses, Attributive Clauses, Adverbial Clauses,* the last is subdivided into subtypes such as: *Adverbial Clauses of Time, Adverbial Clauses of Place, Adverbial Clauses of Purpose, Adverbial Clauses of Reason, Adverbial Clauses of Condition, Adverbial Clauses of Concession, Adverbial Clauses of Result, Adverbial Clauses of Manner* (which include *Adverbial Clauses of Comparison*). Thus, 12 types of complex sentences have been analyzed. The data were studied with the help of the chi-square test and the coefficient of contingency Φ.

Frequency of clauses in texts of five authors

Frequency is one of the functional characteristics of texts. It helps to investigate not only various functional styles but also individual styles of various authors. Statistical calculations of different units found in literary compositions by the same author and different authors help to discern the individual author's style. Here counts were made of clause frequency within the complex sentence selected from the texts of five authors (each fifth page of a work has been analysed manually). The frequency of 12 types of clauses is shown in Table 1.

Table 1
Frequency of different types of clauses

Author / ***Sentence Type***	***Dreiser***	***Fitzgerald***	***Cronin***	***Steinbeck***	***Hemingway***	***Total***
Subject Clauses	6	2	4	23	32	67
Predicative Clauses	5	5	2	13	4	29
Object Clauses	647	306	246	173	208	1580
Attributive Clauses	488	235	194	165	121	1203
Time Clauses	211	193	153	159	114	830
Place Clauses	37	15	21	26	26	125
Reason Clauses	87	82	54	83	22	328

Result Clauses	8	13	5	16	6	48
Concessive Clauses	41	12	46	16	9	124
Purpose Clauses	12	3	2	10	3	30
Conditional Clauses	146	53	46	85	56	386
Clauses of Manner	141	87	50	63	33	374
Total	1829	1006	823	832	634	5124

First of all we want to find out whether the frequencies in Table 1 are distributed homogeneously. This can be done with the help of the chi-square test using the formula

$$(1) \qquad X^2 = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \ ,$$

where O_{ij} is the observed frequency in the cell (i,j) and E_{ij} the expected frequency, $n = 12$ is the number of rows, $k = 5$ is the number of columns

We assume the null hypothesis saying that the data in Table 1 are distributed homogeneously, that is, there is no difference between observed and expected frequencies.

But the results of the analysis performed according to formula (1) showed that the value $X^2 = 374$ far exceeds the critical value with $df = 44$ ($\chi^2_{0,05} = 60.5$; $\chi^2_{0,01} = 68.7$). Thus the null hypothesis is rejected, and this fact shows that the difference between the frequencies in Table 1 is significant. Therefore we may assume that there is positive contingency between elements under investigation: sentence type and authorial style.

Structural type of complex sentence and authorial style

In studying contingency between the structural type of complex sentence and authorial style the coefficient Φ can be used. It may be calculated by formula:

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

where *a, b, c,* and *d* are frequencies in alternative four-cell tables being made on the basis of the data shown in Table 1 (see Table 2).

Table 2

Distribution frequency of clauses in the novels by T. Dreiser and other authors

Type of clause	**Dreiser**	**Others**	**Total**
Object clauses	647 = *a*	933 = *b*	1580
Others	1182 = *c*	2362 = *d*	3544
Total	1829	3295	5124 = *N*

The coefficient Φ in Table 2 is 0.073. The value of this contingency coefficient Φ is determined by the X^2 value. In order to draw a conclusion about the existence of significant conjugation between elements under investigation we must take into consideration only the positive values of coefficients Φ (in formula 2 the difference [*ad* - *bc*] must be positive) and only those, which have *P* not less than 0.05 (in this case χ^2 with *df* = l should be not less than 3.84, that is, $X^2 \geq 3.84$). Statistically significant Φ values are presented in Table 3.

Table 3
Types of sentences and authorial style
(statistically significant values of coefficient Φ)

Author / **Sentence Type**	**Dreiser**	**Fitzgerald**	**Cronin**	**Steinbeck**	**Hemingway**
Subject Clauses	-	-	-	0.056	0.124
Predicative Clauses	-	-	-	0.058	-
Object Clauses	0.073	-	-	-	-
Attributive Clauses	0.056	-	-	-	-
Time Clauses	-	0.212	0.028	0.035	-
Place Clauses	-	0.03	-	-	-
Reason Clauses	-	-	-	-	0.03
Clauses of Manner	-	0.035	-	0.064	-
Result Clauses	-	-	-	0.045	-
Concessive Clauses	-	-	0.09	-	-
Purpose Clauses	-	-	-	0.036	-
Conditional Clauses	-	-	-	0.045	-

From Table 3 it follows that Hemingway's prose is characterized first of all by subject clauses (Φ = 0,124); this type of clauses more often than expected can be found also in Steinbeck's texts (Φ = 0,056). But Steinbeck's prose is characterized first of all by using clauses of manner (Φ = 0,064). Dreiser prefers to use objective and attributive clauses (Φ = 0,073 and Φ = 0,056 respectively); whereas Fitzgerald gives preference to time clauses (Φ = 0,212) as compared with Cronin (Φ = 0,028) and Steinbeck (Φ = 0,035).

Length of complex sentences

Length of complex sentences in the texts of five authors

As it was said above in our investigation sentence length is counted in number of words. To get the most precise data about sentence length in different types of prosaic texts we have

studied separately author's narration (*AN*), the dialogic speech (*DS*) and indirect speech (*IS*). Data about the average sentence length in the texts of the 5 authors are shown in Tables 4–8.

Table 4
Average length of complex sentences in Steinbeck's novels

Novel / Type of Narration	**The Log**	**The Grapes**	**Charley**	$\overline{X}$
AN	27.9	23.8	23.6	25.1
DS	29.8	13	14.8	19.2
IS	23.2	—	25.7	24.45

Table 5
Average length of complex sentences in Cronin's novels

Novel / Type of narration	**Castle**	**Light**	**Citadel**	$\overline{X}$
AN	36.3	31.4	23.3	30.3
DS	17.2	15.1	14	15.4
IS	34	—	18.3	17.4

Table 6
Average length of complex sentences in Hemingway's novels

Novel / Type of Narration	**Farewell**	**To have**	**The Bell**	$\overline{X}$
AN	20.9	25.6	29.5	25.3
DS	10.6	10.2	18.3	13.03
IS	7.5	20	18.1	15.2

Table 7
Average length of complex sentences in Fitzgerald's texts

Novel / Type of Narration	**Tender**	**Gatsby**	**Stories**	$\overline{X}$
AN	29.9	25.9	27.9	27.9
DS	15.6	15	17.4	16
IS	30.5	22.9	25.7	26.3

Table 8
Average length of complex sentences in Dreiser novels

Novel / **Type of Narration**	**Financier**	**Jennie**	**Tragedy**	$\overline{X}$
AN	29.8	24.8	38.2	30.9
DS	20.1	15.9	27.1	21.03
IS	22.2	22.7	27.4	24.1

The data show that average sentence length in author's narration is maximal; then, as it was expected, sentence length in indirect speech follows, and the minimal sentence length is observed in dialogic speech.

Only in one novel ("The Log" by Steinbeck) sentence length in dialogues (29.8) turned out to be longer than sentence length in author's narration (27.9). Thus the following hypothesis can be put forward: the longer is the average sentence length in author's narration, the longer is the sentence length in dialogic speech of the same text; i.e. it is expected that the greater is the sentence length of the author's narration, the greater sentence length used by the characters of the text.

Average sentence length in texts of individual authors

Average sentence length in the texts of the five authors is shown in Table 9.

Table 9
Average length of complex sentence in types of speech

Authors / **Type of narration**	**Hemingway**	**Fitzgerald**	**Dreiser**	**Cronin**	**Steinbeck**	**Total**	$\overline{X}$
AN	25.3	27.9	30.9	30.3	25.1	139.5	27.9
DS	13.03	16	21.03	15.4	19.2	84.66	16.9
IS	15.2	26.3	24.1	17.4	24.45	107.45	21.5
Total	53.53	70.2	76.03	63.1	68.75	331.61	66.3
$\overline{X}$	17.8	23.4	25.3	21.03	22.9	46.33	22.1

We can assume that different structural types of sentences may have different length. Data about distribution of different types of clauses according to their length are presented in Table 10. These data are also presented in alternative multi-cell tables and calculated with the help of formula (2).

Table 10
Distribution of structural types of sentences according to their length

Sentence length / **Sentence type**	**1-20**	**21-60**	**Total**
Subject clauses	65	2	94
Predicative clauses	28	1	29
Objective clauses	1333	247	1580
Attributive clauses	1052	151	1203
Time clauses	793	37	830
Place clauses	115	10	125
Manner clauses	353	21	374
Conditional clauses	349	34	386
Reason clauses	314	14	328
Result clauses	28	20	48
Concessive clauses	112	12	124
Purpose clauses	21	9	30
Total	4975	220	

Table 11
Conjugation of structural type of a sentence and sentence length
(according to values of coefficient Φ)

Sentence length / **Sentence type**	**1-20 (short sentences)**	**21-30 (long sentences)**
Subject clauses	0.029	-
Predicative clauses	0.018	-
Objective clauses	-	0.01
Attributive clauses	-	0.028
Time clauses	0.091	-
Place clauses	-	-
Manner clauses	0.067	-
Conditional clauses	-	-
Reason clauses	0.056	-
Result clauses	-	0.096
Concessive clauses	-	-
Purpose clauses	-	0.047

As it follows from Table 11 maximal contingency is observed between features [long sentences] and [result clauses]; [short sentences] and [time clauses].

Complexity of complex sentences

Sentence complexity and authorial style

As has been said in the introduction, sentence length in our investigation is counted in number of words. But every complex sentence can be characterized not only by number of words but also by number of constituent parts – clauses. To characterize complex sentence from this point of view we shall use the term *sentence complexity*, which stands for the number of clauses being a constituent part of every complex sentence. The maximal number of clauses in a complex sentence in our investigation is nine. Data about distribution of sentence with different complexity in the novels of five authors are presented in Table 12.

Table 12
Distribution of complex sentences with different complexity in the texts of five authors

Author / **Sentence Type**	**Hemingway**	**Fitzgerald**	**Cronin**	**Steinbeck**	**Dreiser**	**Total**
complexity – 1	439	640	537	622	1088	3326
complexity – 2	150	302	173	160	483	1268
complexity – 3	27	52	102	39	168	388
complexity – 4	14	8	10	8	52	92
complexity – 5	2	1	1	3	18	25
complexity – 6	1	3	-	-	12	16
complexity – 7	-	-	-	-	5	5
complexity – 8	-	-	-	-	3	3
complexity – 9	1	-	-	-	-	1
Total	634	1006	823	832	1829	5124

The X^2 for Table 12 is 145,4 which with $df = 32$ yields a probability $P < 10^{-15}$ Thus we can state that there is contingency between some features in Table 14. The results of the corresponding analysis are shown in Table 13.

Table 13
Sentence complexity and authorial style

Author / Sentence Type	Hemingway	Fitzgerald	Cronin	Steinbeck	Dreiser
complexity – 1	0.038	-	-	0.086	-
complexity – 2	-	0.064	-	-	-
complexity – 3	-	-	0.089	-	0.05
complexity – 4	-	-	-	-	0.054
complexity – 5	-	-	-	-	0.052
complexity – 6	-	-	-	-	0.049
complexity – 7	-	-	-	-	0.045
complexity – 8	-	-	-	-	0.037
complexity – 9	-	-	-	-	-

According to the values of coefficient Φ presented in Table 13 we can draw certain conclusions. Thus, sentences with one clause are typical for prose works of Hemingway and Steinbeck, those with two clauses for Fitzgerald, those with three for prose of Cronin and Dreiser. And only in texts of Dreiser we observe a surplus of frequencies of super-complicated sentences (with depth from 3 to 8 clauses). Thus, sentence complexity up to a great degree depends upon individual author's style.

Sentence complexity and structural type of the first clause

Average complexity of complex sentences

An important characteristic of any sentence is considered to be the "*average complexity*", which is marked by letter ***C*** and is calculated according to the formula suggested by G. Akymova (1990):

$$C = \frac{\text{number of clauses}}{\text{number of complex sentences}} \quad (3)$$

In our investigation we shall mark Average Complex Sentence Complexity Coefficient for each author by $\bar{\boldsymbol{E}}$ and calculate it as an *average sum* of the coefficients *C* for each author. The results of calculations are given in Table 14.

Table 14
Average complex sentence complexity coefficient in fiction

	Authors														
	Hemingway			***Fitzgerald***			***Cronin***			***Steinbeck***			***Dreiser***		
$\overline{X}$	13	19	22	23,6	21,2	25,3	29,2	18,5	23,3	18,4	27	21,4	21,1	24	30,9
C	1.2	1.4	1.5	1.5	1.3	1,6	1,6	1,5	1,5	1,3	1,1	1,2	1,6	1,4	2,4
Ē	1.4			1.5			1.5			1.2			1.8		
Works	Farewell (1929)	To have (1937)	The Bell (1940)	Stories (1922)	Gatsby (1925)	Tender (1934)	Castle (1931)	Citadel (1937)	Light (1957)	The Gropes (1939)	The Log (1941)	Charley (1962)	Jennie (1911)	Financier (1912)	Tragedy (1925)

The data in Table 14 show that the minimal complex sentence complexity coefficient is in Steinbeck's novels – only 1.2, whereas the maximal one is in Dreiser's works, viz. 1.8.

Thus, taking into account rather wide amplitude of *Average complex sentence complexity* in the writer's works we can state that it can be their distinguishing feature.

Conclusions

Having done the research we can draw the conclusion that frequency of clause use in the works of five authors is characterized by certain degree of variation. The characteristic feature of Hemingway's style is a very high frequency of *subject clauses*. Contrary to expectation this type of sentences dominates in Steinbeck's works, where there is also a high frequency of *reason clauses. Object clauses* and *attributive clauses* are specific for Dreiser's style. Fitzgerald, Cronin and Steinbeck prefer *time clauses.*

Besides frequency of clause use we can consider *length and complexity of complex sentences* the main syntactic characteristics influencing the structure of fiction. Average sentence length is the most appropriate characteristic for any functional style. Using this property gives us the possibility to compare different texts which are difficult to compare using other characteristics. Having partitioned the whole corpus of the selected sentences into three groups: *short* (1-20 words), *long* (21-60 words) *and super-long* (more than 60 words) we can state that Hemingway's style is characterized by short sentences. Sentences of the second group *(21-60 words)* are frequently used in Fitzgerald's and Dreiser's works. High frequency of super-long sentences (more than 60 words) is typical only for Dreiser's manner of writing.

For correct comparison of sentence length in fiction a text was at first divided into three components: authorial speech, dialogic speech and indirect speech. Each of these components has different primary system of etalons and thus needs separate analysis. Results of statistic calculations testify that complex sentences in these types of narrations differ much by their length.

Scrutinizing sentence length as a quantitative property was made in close connection with its complexity as a structural property. Sentences with one clause are typical for Hemingway and Steinbeck, those with two clauses for prose of Fitzgerald, those with three

for prose of Cronin and Dreiser. Higher frequency of super complex sentences (more than 7 subordinate clauses) is typical only for Dreiser's works.

References

Акимова Г.Н. (1990). *Новое в синтаксисе современного русского языка.* Moskva: Vyssšaja škola.

Левицкий, В.В. (1989). *Статистическое изучение лексической семантики.* Kiev: UMK VO.

Лесскисс, Г.А. (1963). О зависимости между размером предложения и характером текста. *Вопросы языкознания, 3, 92 – 122.*

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

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

Bloch, M.Ya. (1983). *A Course in the Theoretical English Grammar*. Moskva: Vysšaja škola.

Collins Cobuild English Grammar (1990). London : Harper Collins Publishers.

Illyish, B.A. (1964). *The structure of Modern English.* Leningrad: Prosveshcheniye.

Levitskij, V.V., Pavlyčko, O.O., Semenyuk, T.G. (2010). Sentence Length and Sentence Structure as Statistical Characteristics of Style in Prose. In: Uhlířová, L. et al. (eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs: 177-186.* Trier: WVT.

Meier H. (1964). *Deutsche Sprachstatistik*. Bd 1. Hildesheim:

Niehaus, B. (2001). Die Satzlängenverteilungen in literarischen Prosatexten der Gegenwart. In: Uhlířová, L. et al. (eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs: 194-214.* Trier: WVT.

Quirk R. (1985). *A Comprehensive Grammar of the English language.* London: Longman Group Limited

Uhlířová, L. (2001). On word length, clause length and sentence length in Bulgarian. In: Uhlířová, L. et al. (eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs: 266-282.* Trier: Wissenshaftlisher Verlag

.

Glottometrics 21, 2011, 25-53

Entwicklungen im deutschen Wortschatz

Katharina Ternes, Göttingen

Abstract. Logistic laws apply to many areas of developmental phenomena such as biology, demography, medicine, linguistics, etc. In this study, the logistic law known as Piotrowski Law, will be supported based upon data from *Kluge, Etymologisches Wörterbuch der deutschen Sprache* (24. Auflage 2002). Words borrowed from other languages and the ones developed within German language have been taken into account.

Keywords: German language, Piotrowski Law, borrowings, word stock

1. Einleitung

In der nachfolgenden Studie werden die Entlehnungsprozesse im Deutschen untersucht; dabei wird versucht, wie auch schon in vorherigen Arbeiten (z.B. Körner 2004, Best 2001b), Gesetzmäßigkeiten des Sprachwandels aufzuzeigen und das Piotrowski-Gesetz (auch logistisches Gesetz genannt) als zu diesem Zweck geeignet zu erweisen. Zur Überprüfung dieses Gesetzes wurde als Datengrundlage die Auswertung von *Kluge. Etymologisches Wörterbuch der deutschen Sprache* (2002) herangezogen; demnach handelt es sich bei der Untersuchung um einen relativ kleinen Ausschnitt des deutschen Wortschatzes.

Das Piotrowski-Gesetz ist auf verschiedene Sprachwandelprozesse anwendbar. Ausgangspunkt ist die Überlegung, dass Sprachwandel durch ein Individuum beginnt und sich dann unter bestimmten Vorraussetzungen auf weitere Individuen ausbreitet. Die Annahme, dass Gesetze existieren, denen Sprachwandel folgen, stammt von Piotrowski. Ausgehend von dessen Idee entwickelte Altmann (1983) das Piotrowski-Gesetz, das in drei verschiedenen Formen vorkommen kann:

1. Für den vollständigen Sprachwandel; hierbei kommt es zu einer vollständigen Ablösung eines sprachlichen Zustandes durch einen anderen. Als Beispiel kann der Wechsel der Verbformen von *darft* zu *darfst* im Frühneuhochdeutschen gesehen werden. (Best 2006b: 109)
2. Für den unvollständigen Sprachwandel; dies ist ein Prozess, bei dem ein Zuwachs ohne feste obere Grenze beobachtet wird (Leopold, 2005: 628). Als Beispiel dient hier die Ausbreitung von Fremdwörtern.
3. Für den reversiblen Sprachwandel; hierbei werden die Veränderungen eines sprachlichen Zustandes vorerst akzeptiert und breiten sich aus, bis sich der Prozess umzukehren beginnt. Als Beispiel kann hier die von Imsiepen (1983) untersuchte e-Epithese gesehen werden.

Altmann (1983: 59) führt aus: „Unter dem Piotrowski-Gesetz verstehen wir die hypothetische Aussage über den zeitlichen Verlauf der Veränderungen einer beliebigen sprachlichen Entität." Voraussetzung dafür ist, dass die Randbedingungen des Prozesses sich nicht schwerwiegend ändern. Für den unvollständigen Sprachwandel, um den es in dieser Arbeit geht, entwickelte Altmann folgende mathematische Funktion:

$$p_t = \frac{c}{1+a\,e^{-bt}}$$

(mit den Parametern *a*, *b* und *c; c* stellt die Asymptote dar, *t* die Zeiteinheit).

Anhand dieser Studie wird nun untersucht, ob das Gesetz in seiner Form für den unvollständigen Sprachwandel auf die Prozesse innerhalb des deutschen Wortschatzes anzuwenden ist. Dabei werden, wie auch schon bei der Untersuchung von *Duden. Das Herkunftswörterbuch* (2001) durch Helle Körner (2004), nicht nur die Fremdwörter berücksichtigt, sondern auch Wortbildungen des Deutschen sowie der deutsche Erbwortschatz.

2. Methodik

Um das Piotrowski-Gesetz anhand des deutschen Wortschatzes zu überprüfen, wurde als Datenquelle *Kluge. Etymologisches Wörterbuch der deutschen Sprache* (2002)[1] gewählt. Diese Auswahl wurde getroffen, da es sich um das neueste entsprechende Wörterbuch handelt und Elmar Seebold, der die Überarbeitung des *Kluge* seit 1989 übernommen hat, an einem chronologischen Wörterbuch des deutschen Wortschatzes arbeitet. Der erste Teil des chronologischen Wörterbuches ist bereits veröffentlicht und befasst sich mit dem Wortschatz des 8. Jahrhunderts (Seebold, 2001). Man kann also annehmen, dass die Datierungen gerade für das Althochdeutsche an Genauigkeit gewinnen und die Einordnungen der Wörter zu einem jeweiligen Jahrhundert möglich sind, wodurch sich die Studie von der bereits erwähnten Untersuchung des *Duden. Das Herkunftswörterbuch* unterscheidet, denn dort wurde die Datierung dieser Wörter als Problem beschrieben (vgl. Körner, 2004: 27).

Die Auswertung ergab eine Gesamtsumme von 11828 bearbeiteten Stichwörtern. Dabei ist zu berücksichtigen, dass nicht alle in Kluge (2002) vorhandenen Stichwörter in die Auswertung aufgenommen wurden und dadurch ein Unterschied zu den vom Autor angegebenen 13000[2] Stichwörtern entsteht. Gab es beispielsweise zu einem Stichwort nur den Verweis auf ein weiteres Stichwort, wurde jenes nicht bearbeitet, da nicht klar zu erkennen war, wie und wann sich dieses Stichwort im Deutschen durchgesetzt hat. Ebenso wurden gesondert aufgeführte Affixe (vgl. Best & Altmann, 1986: 33) nicht in die Auswertung aufgenommen, da in den meisten Fällen keine Festlegung auf ein Jahrhundert oder auch keine eindeutige Zuordnung zu einer Vermittlersprache möglich war. Als Vermittlersprache wird hier diejenige Sprache beschrieben, aus der das Wort in die deutsche Sprache übernommen wurde (z.B. *Bronchie*; das Deutsche hat dieses Wort aus dem Lateinischen (Vermittlersprache) entlehnt und nicht aus dem Griechischen (Herkunftssprache), woher das Wort ursprünglich stammt und aus dem es in das Lateinische entlehnt wurde) (vgl. Best, 2007: 33).

Von den oben genannten 11828 Wörtern konnte bei 11214 Wörtern eine Vermittlersprache oder die Bildung des Wortes im Deutschen sowie das Jahrhundert der Übernahme angegeben werden, womit diese Stichwörter in die Auswertung aufgenommen wurden. Bei den 614 nicht zuzuordnenden Wörtern waren entweder die Angaben zu dem Jahrhundert der Übernahme oder die Herkunft nicht ausreichend zu bestimmen und das Wort konnte somit für diese Studie nicht berücksichtigt werden.

2.1. Methodische Überlegungen

Innerhalb dieser Studie wurde auf die in Kluge (2002) getätigten Aussagen vertraut und die Auswertung ausschließlich nach den dort zu findenden Angaben vollzogen. Bei Unklarheiten

[1] Wenn im weiteren Verlauf *Kluge* erwähnt wird ist, so ist immer *Kluge. Etymologisches Wörterbuch der deutschen Sprache* (2002) gemeint.

[2] Diese Angabe befindet sich auf dem Buchrücken des Wörterbuches.

wurde darauf verzichtet, in weiteren etymologischen Wörterbüchern die Entstehung der Wörter zu überprüfen, da man annehmen kann, dass Elmar Seebold die Einträge der anderen Wörterbücher bekannt sind und Unklarheiten in dem Bewusstsein vorhanden sind, dass die Forschung sie zum Zeitpunkt der Überarbeitung von *Kluge* (2002) nicht beseitigen kann.

Der Beginn der Zeitrechnung dieser Auswertung wurde auf das 8. Jahrhundert festgesetzt, da nach Naumann (2007: 70) und Seebold (2002: XL) zu diesem Zeitpunkt die Fixierung der deutschen Sprache eingesetzt hat. Fortgeführt wird diese Zeitrechnung bis in das 20. Jahrhundert. Aussagen zum 21. Jahrhundert sind noch nicht möglich.

Stichwörter, zu denen in Kluge (2002) angegeben ist, dass sie einer bestimmten deutschen Sprachregion angehören (z.B. *wobd, wmd, ndd, schweizerisch* usw.), wurden unter Zuhilfenahme von *Duden. Deutsches Universalwörterbuch* (2001), *Wahrig. Deutsches Wörterbuch* (2000) und *Duden. Das Fremdwörterbuch* (2007) ein weiteres Mal nachgeschlagen. War das Stichwort in einem der Wörterbücher aufzufinden, ist es in die Auswertung eingeflossen. Gab es hingegen in keinem der drei Wörterbücher einen Hinweis zu dem entsprechenden Stichwort (z.B. *Alsem* mit dem Hinweis auf ein westmitteldeutsches Sprachgebiet), wurde es aus der Liste entfernt.

Eine weitere Überlegung zum Vorgehen betraf die Stichwörter mit germanischem und niederdeutschem Hintergrund. In diesen Fällen wurde nach der Erklärung Hennings (2003: 20f.) verfahren; das Niederdeutsche wird somit als eigene Sprache behandelt, der es dementsprechend auch möglich ist, Wörter in die deutsche Sprache zu entlehnen. Das Germanische hingegen wird als eine Vorstufe des Deutschen angesehen, aus dem sich die deutschen Wörter weiterentwickelten, und nicht als eine eigene Sprache.

Um eine Überprüfung des Piotrowski-Gesetzes durchzuführen, gilt innerhalb dieser Studie, dass für eine einzelne Sprache ein Nachweis von wenigstens 20 Wörtern vorhanden sein muss und diese Belege in einer ausreichenden zeitlichen Streuung von mindestens vier Jahrhunderten vorkommen müssen, damit eine Berechnung durchgeführt werden kann.

2.2. Behandlung der Entlehnungen

In dieser Untersuchung wird nicht zwischen den Fremd- und Lehnwörtern unterschieden; sie werden mit Angaben zur Vermittlersprache, die auch die Herkunftssprache sein kann, und zum Jahrhundert ihrer Übernahme ins Deutsche berücksichtigt.

Wenn ein einzelnes Lexem bereits ins Deutsche übernommen wurde und mit Hilfe von diesem ein neues Wort entstanden ist, wird dieses Wort als im Deutschen gebildet gewertet, es sei denn, *Kluge* (2002) weist darauf hin, dass es sich um eine Entlehnung handelt oder das Wort aus einer Vermittlersprache stammt. So ist das Wort *Muskel* aus dem Lateinischen entlehnt, das Wort *Muskelkater* wird aber als im Deutschen gebildet angesehen. Bei den Verben und Adjektiven, die mit Wortbildungsmorphemen wie z.B. *–ier* und *–isch* abgeleitet sind, gilt dies nicht. Die so gebildeten Wörter werden zu der in Kluge (2002) angegebenen Vermittlersprache gezählt.

2.3. Vorgehen bei Problemen mit der Vermittlersprache

Nicht jedes Wort, das in die Auswertung dieser Studie aufgenommen wurde, kann mit einer eindeutigen Entstehungsgeschichte belegt werden. Für die vorliegende Untersuchung ist dies aber auch nicht relevant. Es reicht aus, wenn die Vermittlersprache erkennbar ist oder das Wort als im Deutschen gebildet angesehen werden kann, um es in die Auswertung aufzunehmen. Wörter, bei denen die Vermittlersprache als „wahrscheinlich“ oder „vermutlich“ an-

gegeben ist, sind mit der dazu angegebenen Vermittlersprache in die Auswertung eingeflossen.

Zu einigen Wörtern wurden Angaben wie „unter französischem Einfluss“ gefunden. Diese Aussagen wurden nicht beachtet und die Wörter wurden der genannten Vermittlersprache zugesprochen. Anders verhält es sich bei Angaben wie „über das Französische“. Hier wurde diejenige Sprache als Vermittlersprache gewertet, über die das Wort in den deutschen Wortschatz gelangte.

Einige Sprachen werden mit verschiedenen Zeitstufen angegeben; so wird für das Lateinische unter anderem „altlateinisch“, „lateinisch“ und „mittellateinisch“ genannt. Bei allen ausgewerteten Sprachen wurden diese angegebenen Zeitverhältnisse nicht gesondert berücksichtigt. Alle angegebenen zeitlichen Variationen einer jeweiligen Sprache wurden unter einer Sprache zusammengefasst, denn innerhalb dieser Arbeit soll aufgezeigt werden, wie die Übernahme von Wörtern aus anderen Sprachen verläuft, und dazu ist es von Vorteil, eine möglichst hohe Anzahl an auswertbaren Wörtern zu einer Sprache zusammenzufassen. Zu dem Lateinischen wurde darüber hinaus auch die Bildung der „neoklassischen“ Wörter gerechnet, denn in *Kluge* (2002) heißt es dazu: „[w]ir nennen hier Wörter, die mit lateinischem Sprachmaterial (das vielfach griechische Bestandteile aufweist) in neuerer Zeit gebildet worden sind, neoklassisch.“ (Seebold, 2002: XXIX). Neubildungen, die innerhalb des deutschen Sprachgebiets gebildet wurden, und in denen die einzelnen Wortbestandteile aus einer gemeinsamen Sprache stammen, wurden derjenigen Sprache zugerechnet, aus der die Wortbestandteile kommen (z.B. gr. *homoios* und gr. *pathos* zu Homöopathie), mit Ausnahme von Wörtern, die auf gleiche Weise in einem anderen Sprachgebiet gebildet wurden und von dort in die deutsche Sprache übernommen wurden. In einem solchen Fall wird diejenige Sprache als Vermittlersprache gewertet, aus der das Wort ins Deutsche entlehnt wurde. Wörter, die mit fremdsprachigen Lexemen gebildet wurden, aber in der Sprache, aus der die entsprechenden Lexeme stammen, nicht existieren, wurden als im Deutschen gebildete Wörter gewertet (z.B. *Twen*) (vgl. Körner, 2004: 28).

In einigen Fällen gab es die Schwierigkeit einer doppelten Zuordnung eines Wortes, d.h. eines doppelten Entlehnungsprozesses aus verschiedenen Sprachen und in verschiedenen Jahrhunderten. Auch wenn es in *Kluge* (2002) nur ein Stichwort gibt, wurde dieses entsprechende Stichwort in einem solchen Fall zweimal in die Auswertung aufgenommen. Dies war der Fall, wenn ein Wort mit einer bestimmten Bedeutung zuerst aus einer Sprache entlehnt wurde und zu einem späteren Zeitpunkt mit einer anderen Bedeutung aus einer weiteren Sprache (z.B. wurde *demonstrieren* im 16. Jahrhundert mit der Bedeutung »hinweisen, verdeutlichen« aus dem Lateinischen entlehnt und ein weiteres Mal im 19. Jahrhundert aus dem Englischen mit der Bedeutung »öffentliches Kundtun seiner Meinung« übernommen).

Um in die Auswertung möglichst viele Sprachen einfließen zu lassen, wurden in einigen Fällen Sprachfamilien gegründet. So verhielt es sich etwa bei den slawischen Sprachen, wovon nur Russisch knapp an der Beleggrenze zu einer eigenen Auswertung liegt. Es wurde so verfahren, dass Russisch einmal in eine eigene Auswertung aufgenommen wurde und ein weiteres Mal mit Polnisch, Sorbisch, Tschechisch, Slowenisch, Serbokroatisch, Serbisch, Polabisch und Kroatisch unter der Rubrik Slawische Sprachen ausgewertet wurde. Zu diesen Einzelsprachen wurden auch die Belege hinzugefügt, die in *Kluge* (2002) unter der Rubrik „Slawisch“ aufgeführt sind oder die als Entlehnung aus zwei verschiedenen slawischen Sprachen übernommen wurden; z. B. wird für das Wort *Kalesche* eine Entlehnung aus dem Tschechischen und dem Polnischen angegeben. Analog wurde mit den romanischen Sprachen verfahren, wovon einige eine einzelne Auswertung erhalten und dann nochmals unter den romanischen Sprachen mit weiteren Sprachen zusammengefasst wurden, deren Belege für eine eigene Auswertung nicht ausreichen. Zusätzlich sind dort auch jene Stichwörter aufgenommen, zu denen in *Kluge* (2002) als Vermittlersprache „romanisch“ angege-

ben ist. Ähnlich verhält es sich bei den nordgermanischen Sprachen. Die unter diesem Begriff zusammengefassten Einzelsprachen weisen alle zu wenige Belege auf, um eine einzelne Auswertung vorzunehmen. Aus diesem Grund wurden sie zusammengefasst und ergeben am Ende eine Belegzahl von 25 Wörtern, was somit für eine Auswertung ausreicht.[3] Als eine weitere Art der Zusammenfassung kann das Englische angeführt werden, denn hier findet sich nicht nur die Sprache, die man in Großbritannien spricht, sondern auch die in den Vereinigten Staaten, Kanada oder Australien verwendeten Formen (vgl. Bußmann, 2002: 190).

2.4. Vorgehen bei Problemen mit Datierungsangaben

Neben den vorhandenen Schwierigkeiten bei der Feststellung der Vermittlersprache musste auch für einige Probleme mit den Datierungsangaben eine Lösung gefunden werden. Überwiegend gibt es in *Kluge* (2002) zu einem Stichwort nur eine Datierungsangabe; diese bezieht sich darauf, wann das Wort zum ersten Mal im deutschen Sprachgebiet genannt wird und nicht darauf, wann ein Wort in den alltäglichen Gebrauch übergeht. In einigen Fällen ist aber mehr als nur eine Jahrhundertangabe zu finden, wie z.B. bei: „8. Jhd., Form 9. Jhd.“. War dies der Fall, wurde stets die erste Angabe übernommen, da es sich mehrheitlich um eine geringe Veränderung des Wortes handelt und dies als eine Weiterentwicklung der deutschen Sprache angesehen werden kann (z.B. unter dem Stichwort *Hilfe* findet sich die oben genannte Zeitangabe und die dazugehörigen Formen *ahd. helfa, hilfa*). Auch bei Angaben, die etwa „15. Jhd., Bedeutung 17. Jhd.“ lauten, wurde zur Auswertung das zuerst genannte Jahrhundert aufgenommen, wenn aus dem Artikel erkennbar war, dass das Wort eine leichte Verschiebung in der Bedeutung durchlaufen hat und der ursprüngliche Sinn des Wortes noch vorhanden ist (z.B. das Wort *Einstand* bedeutete im 15. Jahrhundert *das Eintreten vor Gericht*, ab 17. Jahrhundert wurde es in einer übertragenden Bedeutung verwendet und kann *das Antreten einer neuen Stelle* bedeuten). War aber in dem Artikel zu erkennen, dass sich die Bedeutung grundlegend verändert hat, so wurde das Wort bei ausreichenden Angaben ein zweites Mal in die Auswertung aufgenommen.[4] Bei Lexemen, die mehrere Bedeutungen haben und die in *Kluge* (2002) auch unter verschiedenen Stichwörtern aufgeführt sind, wurde jedes Stichwort in die Auswertung aufgenommen, soweit die dort gemachten Angaben zur Auswertung ausreichten (so wurde *Pinsel* einmal mit der Bedeutung *Malerwerkzeug* als Entlehnung des 13. Jahrhunderts aus dem Französischen gewertet und einmal als Bildung des Deutschen im 18. Jahrhundert mit der Bedeutung *einfältiger Mensch*).

Bei Wörtern, die einmal in der deutschen Sprache vorhanden waren, dann aber einen zeitlichen Sprung aufweisen, bevor wieder ein Nachweis zu finden ist, wurde ebenfalls das erste Jahrhundert des Auftretens in die Auswertung übernommen, denn es kann der Fall vorliegen, dass zwar eine kontinuierliche Entwicklung des Wortes vorliegt, für diese aber mangels Dokumenten keine Nachweise zu finden sind, oder, dass ein früher vorhandenes Wort bewusst neu in die Sprache aufgenommen wird. Wenn sich die beiden Wörter dann in Orthographie und Semantik gleichen, wird dieses Wort nicht ein weiteres Mal aufgenommen, sondern mit dem zuerst auftretendem Jahrhundert in die Auswertung aufgenommen.

[3] Die Zuordnung der Einzelsprachen zu den Sprachfamilien wurde aus Bußmann (2002) übernommen.

[4] Diese Beurteilung obliegt der subjektiven Einschätzung der Verfasserin dieser Studie und kann nicht als allgemeingültig angesehen werden.

3. Auswertung

Mit dem genannten Vorgehen konnten Datensätze für 12 Sprachen und drei Sprachfamilien sowie ein Überblick zum herkömmlichen Wortschatz des Deutschen gewonnen werden.

Im Folgenden wird anhand einer Tabelle ein Gesamtüberblick der aus dem *Kluge* gewonnenen Daten dargestellt. Die Spalte mit der Bezeichnung **Herkunft** gibt dabei die Vermittlersprache wieder, die auch die Herkunftssprache sein kann, wenn keine längere Wortgeschichte bekannt ist; unter **Anzahl** sind die in *Kluge* (2002) verzeichneten Wörter der nebenstehenden Sprache aufgelistet und unter **Prozente** der jeweilige Prozentwert, der den einzelnen Sprachen in dieser Auswertung zukommt.

Tabelle 1
Verteilung der datierbaren Wörter auf die einzelnen Sprachen (einschließlich Erbwörter)

Herkunft	**Anzahl**	**Prozente**	**Herkunft**	**Anzahl**	**Prozent**
Deutsch	5710	51,238	Malaiisch	3	0,027
Latein	2051	18,405	Gallisch	2	0,018
Französisch	1381	12,392	Irisch	2	0,018
Niederdeutsch	558	5,007	Isländisch	2	0,018
Englisch	513	4,603	Litauisch	2	0,018
Italienisch	317	2,845	Persisch	2	0,018
Griechisch	179	1,606	Romani	2	0,018
Niederländisch	139	1,247	Serbisch	2	0,018
Rotwelsch	66	0,592	Slowenisch	2	0,018
Spanisch	51	0,458	Chinesisch	1	0,009
Jiddisch	25	0,224	Dänisch	1	0,009
Russisch	19	0,170	Finnisch	1	0,009
Tschechisch	14	0,126	Gotisch	1	0,009
Altnordisch	11	0,099	Hindi	1	0,009
Polnisch	11	0,099	Indisch	1	0,009
Türkisch	9	0,081	Keltisch	1	0,009
Hebräisch	8	0,072	Ketschua	1	0,009
Portugiesisch	8	0,072	Kroatisch	1	0,009
Schwedisch	7	0,063	Polabisch	1	0,009
Sorbisch	7	0,063	Polynesisch	1	0,009
Arabisch	6	0,054	Provenzalisch	1	0,009
Ungarisch	6	0,054	Singhalesisch	1	0,009
Japanisch	4	0,036	Suaheli	1	0,009
Norwegisch	4	0,036	Sumerisch	1	0,009
Afrikaans	3	0,027	Venezianisch	1	0,009
Grönländisch	3	0,027	**Summe**	**11144**	**100**

Wie aus der Tabelle ersichtlich ist, wurde knapp über die Hälfte der Wörter im Deutschen gebildet. Dem Lateinischen kommt mit 18,405% des deutschen Wortschatzes[5] der größte Anteil an Entlehnungen zu, wobei aber zu bedenken ist, dass viele der griechischen Wörter über das Lateinische in das Deutsche entlehnt wurden (vgl. Wittstock, 1982: 7). Somit erklärt sich auch die geringe Anzahl (1,613%) der aus dem Griechischen übernommenen Wörter.

In Tabelle 1 sind diejenigen Stichwörter, die in *Kluge* (2002) als zu einer Sprachfamilie (z.B. Slawisch, Romanisch oder nordische Sprachen) gehörig angegeben sind, nicht mit aufgenommen,[6] da sie keine eindeutige Zuordnung zulassen. In einem späteren Schritt werden einige hier aufgeführte Einzelsprachen zu Sprachfamilien zusammengefasst, um genügend Belege zur Anpassung des Piotrowski-Gesetzes zu erlangen.

Wenn man sich nun eine Liste anschaut, in der die im Deutschen gebildeten und ererbten Wörter nicht berücksichtigt werden, ergibt sich die nachstehende Tabelle:

Tabelle 2
Verteilung der datierbaren Wörter auf die einzelnen Sprachen (ohne Erbwörter)

Herkunft	**Anzahl**	**Prozente**	**Herkunft**	**Anzahl**	**Prozent**
Latein	2051	37,737	Gallisch	2	0,037
Französisch	1381	25,409	Irisch	2	0,037
Niederdeutsch	558	10,267	Isländisch	2	0,037
Englisch	513	9,439	Litauisch	2	0,037
Italienisch	317	5,833	Persisch	2	0,037
Griechisch	179	3,293	Romani	2	0,037
Niederländisch	139	2,557	Serbisch	2	0,037
Rotwelsch	66	1,214	Slowenisch	2	0,037
Spanisch	51	0,938	Chinesisch	1	0,018
Jiddisch	25	0,460	Dänisch	1	0,018
Russisch	19	0,350	Finnisch	1	0,018
Tschechisch	14	0,258	Gotisch	1	0,018
Altnordisch	11	0,202	Hindi	1	0,018
Polnisch	11	0,202	Indisch	1	0,018
Türkisch	9	0,166	Keltisch	1	0,018
Hebräisch	8	0,147	Ketschua	1	0,018
Portugiesisch	8	0,147	Kroatisch	1	0,018
Schwedisch	7	0,129	Polabisch	1	0,018
Sorbisch	7	0,129	Polynesisch	1	0,018
Arabisch	6	0,110	Provenzalisch	1	0,018
Ungarisch	6	0,110	Singhalesisch	1	0,018
Japanisch	4	0,074	Suaheli	1	0,018
Norwegisch	4	0,074	Sumerisch	1	0,018

[5] Es ist zu beachten, dass sich hier immer auf die Auswertung des *Kluge* bezogen wird, was einen Ausschnitt des deutschen Wortschatzes darstellt und diesen nicht als Gesamtes betrachtet. Wird eine andere Auswertung herangezogen, wird dies erwähnt.

[6] Aus diesem Grund stimmt die Summe nicht mit der Summe der ausgewerteten Wörter überein.

Afrikaans	3	0,055	Venezianisch	1	0,018
Grönländisch	3	0,055			
Malaiisch	3	0,055	**Summe**	**5435**	**100**

Hier ist deutlich zu erkennen, dass Latein den größten Anteil der Fremdwörter im Deutschen ausmacht. Das Französische folgt diesem mit einem Abstand von über 12% und das Niederdeutsche hat zum Französischen ebenfalls einen hohen Abstand von über 15% und liegt knapp über dem Englischen mit 9,439%. Die Werte nehmen mit großen Schritten ab und sind schon im Spanischen mit 51 Belegen unter 1% gesunken. Diese Befunde decken sich weitgehend mit denen, die Körner (2004) anhand von *Duden. Herkunftswörterbuch* (2001) vorgestellt hat.

3.1. Die einzelnen Sprachen

Zunächst wird an die Daten zur Entwicklung des deutschen Wortschatzes (ohne Entlehnungen) das Piotrowski-Gesetz in seiner Form für den unvollständigen Sprachwandel, wie oben angegeben, angepasst, um daran beispielhaft die Tabellen und Graphiken zu erläutern.[7] Anschließend werden auf die gleiche Weise die Einzelsprachen und die Sprachfamilien, aus denen deutsche Entlehnungen stammen, behandelt. Für die historischen Hintergründe der Entlehnungs- und Wachstumsprozesse wird auf die Fachliteratur, besonders auf Polenz (2002), verwiesen, soweit sie nicht bei der Datenerhebung eine besondere Rolle spielen.

3.2.1. Deutsch

Nach Naumann (2007: 70) und Seebold (2002: X) setzt die Fixierung der deutschen Sprache erst im 8. Jahrhundert ein, mit dem auch die Tabelle 3 beginnt. Die zwei Belege des 7. Jahrhunderts, die in Kluge (2002) angeführt sind, wurden dem 8. Jahrhundert zugeschlagen. Für die im Deutschen gebildeten und ererbten Wörter vom 8. Jahrhundert an ergab sich durch Anpassung der Formel für den unvollständigen Sprachwandel Tabelle 3.

Tabelle 3
Zuwachs des deutschen Wortschatzes (Erbwörter)

Jahrhundert	**t**	***n***	***n* (kumuliert)**	***p* (berechnet)**
8	1	1212	1214	1444.34
9	2	673	1887	1697.29
10	3	194	2081	1981.83
11	4	331	2412	2297.58
12	5	189	2601	2642.75
13	6	341	2942	3013.94
14	7	331	3273	3406.13
15	8	456	3729	3812.87
16	9	582	4311	4226.63

[7] Für die Auswertung der gesammelten Daten mit Hilfe des Computerprogramms NLREG und die Erstellung der Graphiken danke ich Dr. Best.

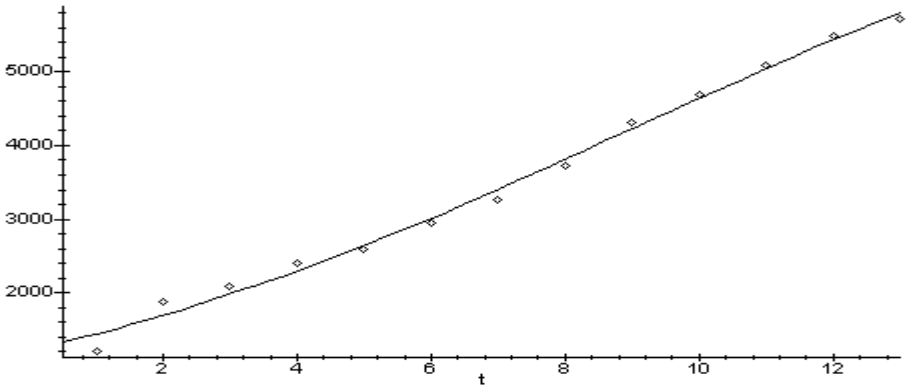
17	10	379	4690	4639.35
18	11	401	5091	5043.04
19	12	397	5488	5430.41
20	13	222	5710	5795.34
a = 5.8296 ***b*** = 0.1987 ***c*** = 8346.8391 ***D*** = 0.99				

Die Tabelle ist wie folgt zu verstehen: Das **Jahrhundert** zeigt den Zeitraum, dem innerhalb der Auswertung die Stichwörter zugeordnet wurden, ***t*** steht für jeweils eine Zeiteinheit von hundert Jahren; ***n*** gibt die in der Auszählung gewonnene Menge von Belegen für das in der gleichen Zeile stehende Jahrhundert wieder. Diese werden unter ***n (kumuliert)*** mit jeder Zunahme in einem Jahrhundert addiert. ***p (berechnet)*** ist die durch Anpassung des Modells für den unvollständigen Sprachwandel errechnete Zahl; ***a, b*** und ***c*** sind Parameter; ***c*** gibt dabei den Grenzwert des Sprachwandels an. Darunter ist kein feststehender Wert zu verstehen, sondern eine Variable, die sich je nach Datenbasis ändert (vgl. Best & Altmann, 1986: 38). ***D*** ist der Determinationskoeffizient; er kann zwischen 0 und 1 liegen und seine Anpassung ist um so besser, je näher er sich an 1 befindet. „Im Allgemeinen reicht uns, wenn der Determinationskoeffizient über 0.8 ist. Bei 0.9 kann man das Resultat schon als sehr gut betrachten“ (Best, Beöthy & Altmann, 1990: 122f.). In diesem Fall ist die Anpassung mit *D* = 0.99 hervorragend.

Durch den Parameter ***a*** (Integrationskonstante) wird die Position auf der x-Achse angegeben; ***b*** (Proportionalitätskonstante) gibt den Grad des Anstieges der Funktion wieder. Je größer *b* ist, desto steiler und zeitlich begrenzter steigt der Graph und somit auch die Ausbreitung der Wörter. ***c*** (Asymptote) gibt die Obergrenze für das in diesen Fall ausgewertete Datenmaterial an. ***p*** bezeichnet den Zuwachs neuer Wörter (vgl. Leopold, 2005: 627f. und Best, Beöthy & Altmann, 1990: 115f.). In die Formel des unvollständigen Sprachwandels werden die aus der Tabelle 5 errechneten Parameter ***a*** und ***b*** und der Grenzwert ***c*** eingesetzt und es ergibt sich für die im Deutschen gebildeten und ererbten Wörter die Formel:

$$p_t = \frac{8346.8391}{1+5.8296\, e^{-0.1987\, t}}.$$

Die folgende Graphik veranschaulicht das Ergebnis, das in Tabelle 3 enthalten ist:



Graphik zu Tabelle 3: Zuwachs des deutschen Wortschatzes (Erbwörter)

Wie zu erkennen ist, befindet sich auf der x-Achse **t**, das die Zeiteinheit angibt, und auf der Y-Achse ist die Anzahl der Wörter zu erkennen. **t =** 1 steht für das 8. Jahrhundert, **t** =13 für

das 20. Jahrhundert, so wie auch in Tabelle 3 angegeben. Die Quadrate zeigen die ausgezählten Werte und die Linie gibt den errechneten Verlauf wieder. Die Beochtungswerte liegen überwiegend auf der berechneten Trendlinie; der Zuwachs des Wortschatzes deutscher Herkunft erweist sich hier als ein nach wie vor stetiger Prozess.

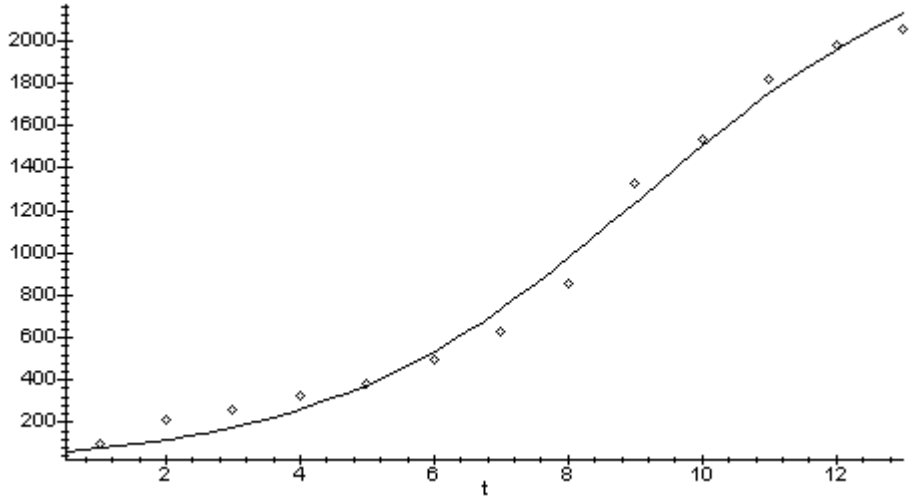
3.2.2. Latein

Wie in Tabelle 1 und 2 zu erkennen ist, kommt dem Lateinischen der größte Anteil der deutschen Lehnwörter zu. Deshalb ist ein Blick auf die Auswertung der lateinischen Sprache besonders interessant.

Tabelle 4
Zuwachs der aus dem Lateinischen übernommenen Entlehnungen

Jahrhundert	**t**	***n***	***n* (kumuliert)**	***p* (berechnet)**
8	1	101	101	79.70
9	2	111	212	119.60
10	3	53	265	178.03
11	4	63	328	261.92
12	5	54	382	379.01
13	6	116	498	536.12
14	7	132	630	736.20
15	8	226	856	974.67
16	9	468	1324	1237.46
17	10	216	1540	1503.16
18	11	281	1821	1749.42
19	12	161	1982	1959.92
20	13	69	2051	2127.77
a = 47.1274 ***b*** = 0.4222 ***c*** = 2542.2102 ***D*** = 0.99				

Mit einem Determinationskoeffizienten von 0.99 ist die Anpassung des Piotrowski-Gesetzes wieder sehr gut gelungen.



Graphik zu Tabelle 4: Zuwachs der aus dem Lateinischen übernommenen Entlehnungen

Der Trend der Aufnahme lateinischer Wörter im Deutschen scheint abzunehmen, ist aber immer noch deutlich vorhanden. Etwa bei t = 8, dem 15. Jahrhundert, scheint der Wendepunkt zu liegen: Bis dahin beschleunigt sich der Zuwachs lateinischer Entlehnungen; vom 15. Jahrhundert an verlangsamt sich der Trend allmählich.

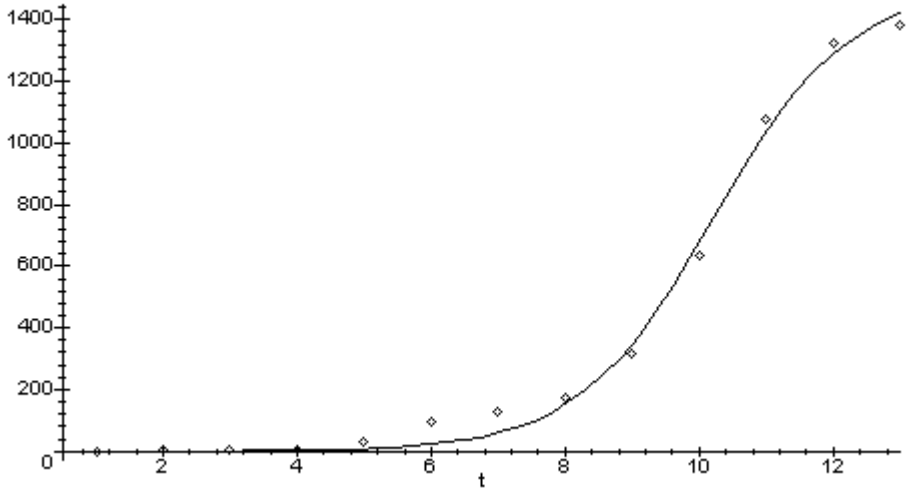
3.2.3. Französisch

Der Zuwachs von Entlehnungen aus dem Französischen stellt sich wie folgt dar:

Tabelle 5
Zuwachs der aus dem Französischen übernommenen Entlehnungen

Jahrhundert	**t**	***n***	***n* (kumuliert)**	***p* (berechnet)**
8	1	1	1	0.16
9	2	1	2	0.44
10	3	3	5	1.17
11	4	0	5	3.16
12	5	25	30	8.51
13	6	67	97	22.73
14	7	34	131	59.80
15	8	41	172	151.16
16	9	142	314	348.49
17	10	318	632	675.24
18	11	445	1077	1034.88
19	12	245	1322	1289.45
20	13	59	1381	1418.80
a = 25194.9678 ***b*** = 0.9925 ***c*** = 1507.8973 ***D*** = 0.99				

Mit dem Determinationskoeffizienten von 0.99 ist auch hier ein sehr gutes Ergebnis erzielt worden. Der sich abschwächende Zuwachs der Entlehnungen in den letzten Jahrhunderten deutet an, dass der Einfluss der französischen Sprache auf den deutschen Wortschatz abnimmt, wie auch die Graphik zeigt:



Graphik zu Tabelle 5: Zuwachs der aus dem Französischen übernommenen Entlehnungen

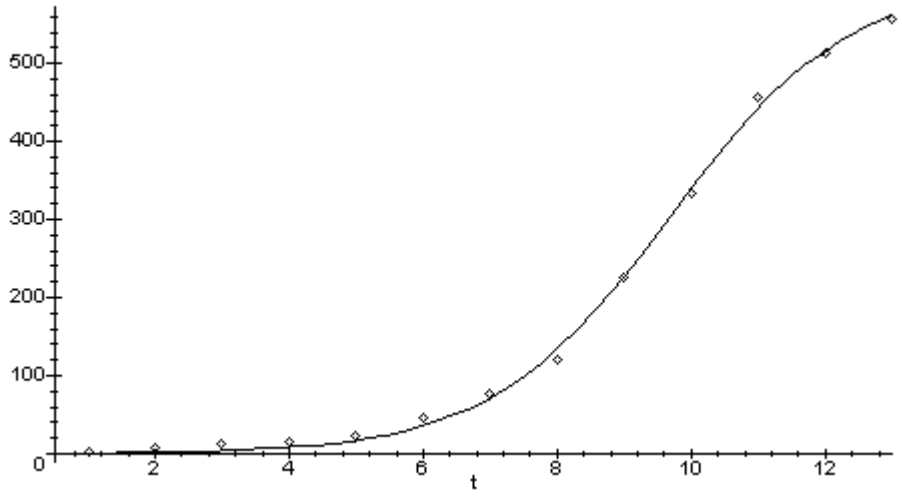
3.2.4. Niederdeutsch

Wie sich in der Tabelle 2 ablesen lässt, hat das Niederdeutsche mit einem Anteil von knapp über zehn Prozent noch einen großen Anteil an den Entlehnungen im Deutschen. Die Berechnung der ausgewerteten Daten zeigt sich wie folgt:

Tabelle 6
Zuwachs der aus dem Niederdeutschen übernommenen Entlehnungen

Jahrhundert	**t**	***n***	***n* (kumuliert)**	***p* (berechnet)**
8	1	3	3	0.90
9	2	5	8	1.90
10	3	4	12	4.01
11	4	4	16	8.42
12	5	6	22	17.54
13	6	25	47	35.96
14	7	29	76	71.37
15	8	45	121	133.52
16	9	105	226	226.90
17	10	108	334	338.94
18	11	123	457	442.11
19	12	55	512	516.42
20	13	46	558	560.97
a = 1427.0672 ***b*** = 0.7494 ***c*** = 607.9908 ***D*** = 0.99				

Der Determinationskoeffizient liegt für die hier ausgewerteten Daten bei abgerundet 0.99, was eine sehr gute Anpassung widerspiegelt. Dieses wird durch die Graphik nochmals verdeutlicht:



Graphik zu Tabelle 6: Zuwachs der aus dem Niederdeutschen übernommenen Entlehnungen

Der berechnete Grenzwert **c** liegt nah an den Beobachtungen und Berechnungen des 20. Jahrhunderts, was den Schluss nahelegt, dass der Übernahmeprozess von niederdeutschen Wörtern fast zum Erliegen gekommen ist.

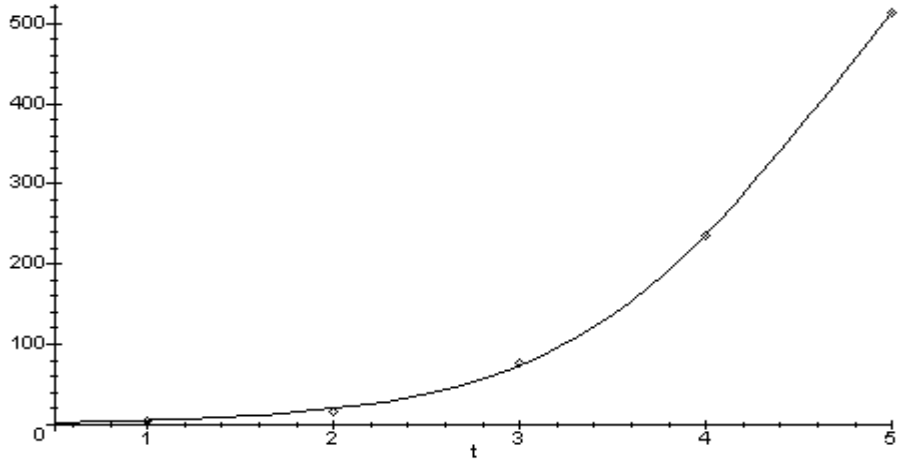
3.2.5. Englisch

Das Englische weist im Gegensatz zu den bisher behandelten Sprachen die Besonderheit auf, dass die Entlehnungen mit dem 16. Jahrhundert erst sehr spät beginnen. Aber auch, wenn sich die Entlehnungen nur auf vier Jahrhunderte verteilen, kommt dem Englischen in der Gesamtübersicht (Tabelle 1 und 2) doch ein relativ hoher Anteil an Entlehnungen zu. Die Auswertung ergab folgendes Ergebnis:

Tabelle 7
Zuwachs der aus dem Englischen übernommenen Entlehnungen

Jahrhundert	**t**	***n***	***n* (kumuliert)**	***p* (berechnet)**
16	1	4	4	4.95
17	2	12	16	19.69
18	3	61	77	74.40
19	4	159	236	236.71
20	5	277	513	512.90
a = 675.7504 ***b*** = 1.3991 ***c*** = 830.4178 ***D*** = 0.99				

Wie zu sehen ist, gibt es wieder eine sehr gute Übereinstimmung zwischen den berechneten und beobachteten Werten und der Determinationskoeffizient zeigt eine entsprechend gute Anpassung des Piotrowski-Gesetzes.



Graphik zu Tabelle 7: Zuwachs der aus dem Englischen übernommenen Entlehnungen

Bei der Graphik ist zu beachten, dass sich gegenüber den vorherigen Tabellen auf der x-Achse die Zeitverhältnisse ändern. **t** = 1 steht hier für den Beginn der Übernahme englischer Wörter in den deutschen Wortschatz und zeigt das 16. Jahrhundert an. Die Graphik unterstützt den in der Tabelle gewonnenen Eindruck und zeigt noch einmal die gute Übereinstimmung der Werte. Es wird deutlich, dass der Übernahmeprozess anhält und die Trendlinie noch einen starken Zuwachs anzeigt. Es ist nicht deutlich, ob der Wendepunkt bereits erreicht oder gar überschritten ist. (Zum Einfluss des englischen und amerikanischen Englisch auf das Deutsche sowie die betroffenen Themenbereiche siehe Best 2003a, Lucko 1995.)

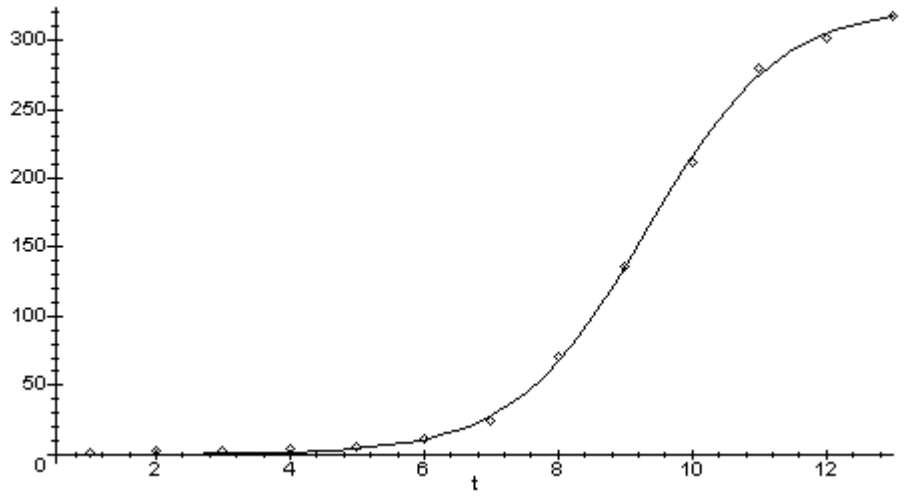
3.2.6. Italienisch

Zu Entlehnungen aus dem Italienischen gibt es schon eine Beobachtung im 8. Jahrhundert. Es ist aber auffällig, dass es innerhalb der ersten fünf Jahrhunderte nur jeweils eine Entlehnung gibt, wie an der Tabelle abzulesen ist:

Tabelle 8
Zuwachs der aus dem Italienischen übernommenen Entlehnungen

Jahrhundert	**t**	***n***	***n* (kumuliert)**	***p* (berechnet)**
8	1	1	1	0.07
9	2	1	2	0.19
10	3	1	3	0.53
11	4	1	4	1.47
12	5	1	5	4.01
13	6	6	11	10.82
14	7	13	24	28.15
15	8	47	71	67.32
16	9	65	136	135.99
17	10	76	212	215.94
18	11	68	280	274.53
19	12	22	302	304.52
20	13	15	317	317.09
a = 12689.6511 ***b*** = 1.0134 ***c*** = 324.7261 ***D*** = 0.99				

Die Übereinstimmung von beobachteten und berechneten Werten ist in diesem Fall ebenso gut, wie wir es im Englischen gesehen haben. Auch der Determinationskoeffizient zeigt mit 0.99 eine sehr gute Anpassung. Besonders auffällig ist die Nähe des Grenzwertes **c** zu dem letzten Wert der Beobachtung und der Berechnung, die hier nahezu übereinstimmen. Die folgende Graphik verdeutlicht dies noch einmal:



Graphik zu Tabelle 8: Zuwachs der aus dem Italienischen übernommenen Entlehnungen

Die sehr flach auslaufende Trendlinie zeigt an, dass der Prozess der Übernahme von Italianismen voraussichtlich nahezu abgeschlossen ist.

Ein Großteil der aus dem Italienischen entlehnten Wörter ist aus dem Finanz- und Handelsbereich übernommen und kann mit der Ausbreitung des Handels zum Ende des Mittelalters und mit den dafür wichtigen italienischen Häfenstädten begründet werden (vgl. Schmöe, 1998: 31). Es ist davon auszugehen, dass die dort verkehrenden Kaufleute die italienischen Begriffe zur gegenseitigen Verständigung übernommen haben. Weitere Bereiche, aus denen italienische Wörter übernommen wurden, sind die Bereiche der Kunst und des italienischen Lebensstils, der während der italienischen Renaissance zum Vorbild anderer Nationen wurden (vgl. Schmöe, 1998: 31f.). Best (2006a) hat zu diesem Thema eine Untersuchung durchgeführt[8], die sich mit den verschiedenen Bereichen von Entlehnungen aus dem Italienischen beschäftigt; dabei steht der Bereich des Lebensstils an erster Stelle, gefolgt von dem Handel- und Finanzwesen und der Musik (vgl. Best, 2006a: 77f.).

3.2.7. Griechisch

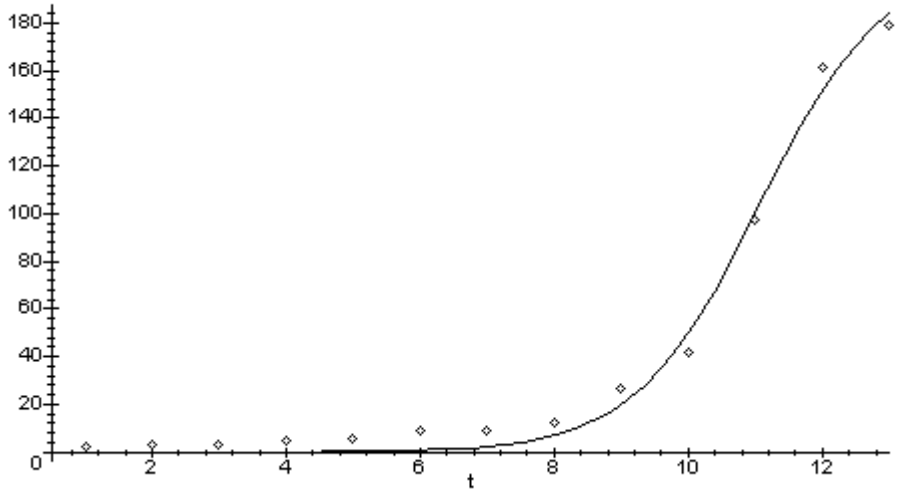
Für die Übernahme griechischer Wörter in den deutschen Wortschatz ergibt sich mit den hier ausgewerteten Daten Folgendes:

Tabelle 9
Zuwachs der aus dem Griechischen übernommenen Entlehnungen

Jahrhundert	**t**	***n***	***n* (kumuliert)**	***p* (berechnet)**
8	1	2	2	0.00
9	2	1	3	0.01
10	3	0	3	0.03
11	4	2	5	0.10
12	5	1	6	0.28
13	6	3	9	0.84
14	7	0	9	2.47
15	8	3	12	7.17
16	9	15	27	19.93
17	10	15	42	49.75
18	11	55	97	100.14
19	12	64	161	151.88
20	13	18	179	183.82
a =169112.7270 ***b*** = 1.0896 ***c*** = 205.7539 ***D*** = 0.99				

Der Übernahmeprozess griechischer Wörter beginnt sehr langsam und steigert sich dann, bevor er sich im letzten Jahrhundert wieder verringert. Das Ergebnis der Anpassung ist auch hier mit einem Determinationskoeffizienten von 0.99 sehr gut gelungen.

[8] Die Quelle dieser Untersuchung ist: Schmöe (1998)



Graphik zu Tabelle 9: Zuwachs der aus dem Griechischen übernommenen Entlehnungen

Die berechneten und beobachteten Werte scheinen hier nicht so nah beieinander zu liegen, wie dies im Italienischen oder im folgenden Niederländischen der Fall ist, aber es muss dabei berücksichtigt werden, dass sich die Werte auf der y-Achse verändern und in diesem Fall eine feinere Einteilung vorhanden ist als in den beiden oben genannten Sprachen, denn vergleicht man die Differenz der Werte aus der Tabelle, ergibt sich ein ähnlich gutes Bild. Der Wendepunkt ist hier zwischen dem 18. und 19. Jahrhundert zu erkennen; von da an nähert sich der Graph der Asymptote, was zusammen mit dem Grenzwert **c** zeigt, dass sich der zukünftige Übernahmeprozess verringert. Aus den Werten der Tabelle ist im Gegensatz dazu erkennbar, dass im 19. Jahrhundert mit 64 Übernahmen der Höhepunkt erreicht ist und es daraufhin zu einer Reduktion kommt.

Innerhalb der Untersuchung kann der Eindruck entstehen, dass die griechischen Wörter keinen größeren Einfluss auf die deutsche Sprache ausüben, dieser Eindruck täuscht aber. Denn viele der Wörter, die eigentlich griechischer Herkunft sind, haben den deutschen Wortschatz über das Lateinische erreicht.

Aus der Verbindung seines, des italienischen Erbes mit den von den Griechen ausgehenden starken wirtschaftlichen, gesellschaftlichen und vor allem kulturellen Einflüssen erklärt sich der nicht zu übersehende Tatbestand, daß auch das Griechische […] im Wortschatz international gebräuchlicher Fachterminologien seinen uns vielfältigst begegnenden Niederschlag gefunden hat, und dies um so mehr, als fast alle uns geläufigen wissenschaftlichen Disziplinen bereits bei den Griechen zumindest vorgebildet gewesen sind (Wittstock, 1982: 7).

In der heutigen Zeit werden, genau wie aus dem Lateinischen, in der Wissenschaft oft Wörter mit dem Sprachmaterial griechischer Wörter gebildet, um spezielle Gegenstandsbereiche oder Materialien zu bezeichnen.

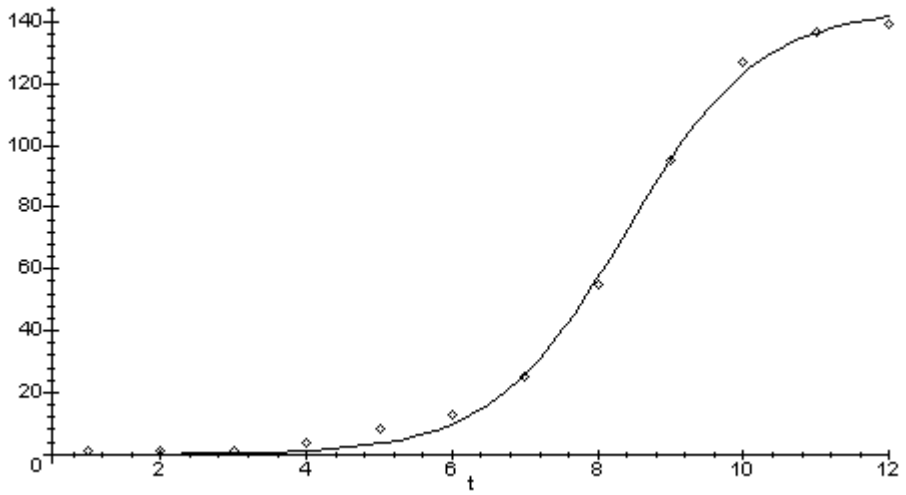
3.2.8. Niederländisch

Durch die Auswertung der niederländischen Daten ergeben sich die folgenden Werte:

Tabelle 10
Zuwachs der aus dem Niederländischen übernommenen Entlehnungen

Jahrhundert	**t**	***n***	***n* (kumuliert)**	***p* (berechnet)**
9	1	1	1	0.04
10	2	0	1	0.13
11	3	0	1	0.40
12	4	3	4	1.18
13	5	4	8	3.46
14	6	5	13	9.87
15	7	12	25	25.95
16	8	30	55	57.13
17	9	40	95	95.54
18	10	32	127	123.28
19	11	10	137	136.54
20	12	2	139	141.64
a = 9705.7303 ***b*** = 1.0947 ***c*** = 144.3540 ***D*** = 0.99				

Anhand der Tabelle ist die gute Übereinstimmung zwischen den berechneten und den beobachteten Werten zu erkennen und der Determinationskoeffizient zeigt mit einem Wert von 0.99 eine sehr gute Anpassung des Piotrowski-Gesetzes. Der Wert **c** zeigt, dass der Grenzwert sehr nah an den zuletzt berechneten und beobachteten Werten liegt, womit sich wieder ein Ende des Übernahmeprozesses andeutet, wie auch in der Graphik zu sehen ist:



Graphik zu Tabelle 10: Zuwachs der aus dem Niederländischen übernommenen Entlehnungen

Das Niederländische hat dem deutschen Wortschatz nicht nur, wie die Auswertung der Daten zeigt, viele Begriffe der Seefahrt übermittelt, sondern auch Wörter, die aus dem Französischen zuerst in das Niederländische entlehnt wurden und von dort aus weiter an das Deutsche (vgl. Telling, 1987: 13). Unter den Entlehnungen sind auch einige Wörter, die in der flämischen Region Belgiens unter dem Einfluss französischer Wörter nachgebildet wurden und so in die niederländische Sprache gelangt sind, die sie dann weiter entlehnt hat (vgl. Seebold, 1981: 108).

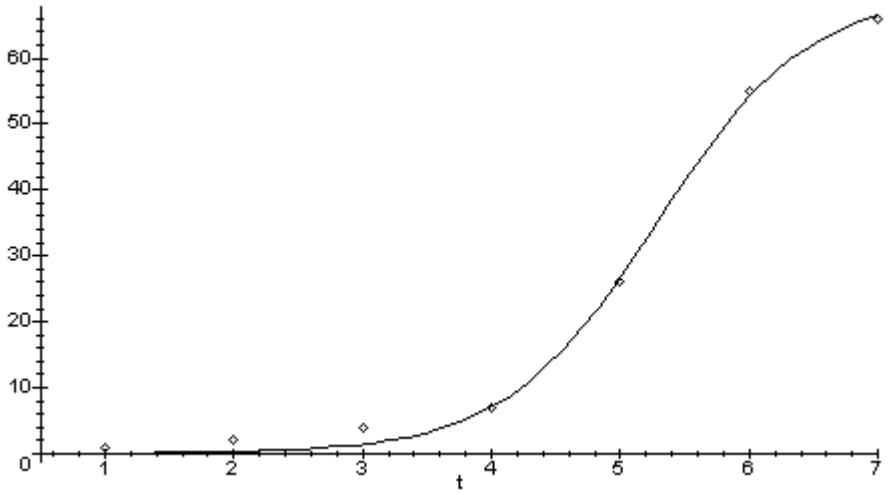
3.2.9. Rotwelsch

Die Übernahme von Wörtern aus dem Rotwelschen beginnt mit dem 14. Jahrhundert erst spät und es ergeben sich die folgenden Werte:

Tabelle 11
Zuwachs der aus dem Rotwelschen übernommenen Entlehnungen

Jahrhundert	**t**	***n***	***n* (kumuliert)**	***p* (berechnet)**
14	1	1	1	0.05
15	2	1	2	0.26
16	3	2	4	1.42
17	4	3	7	7.12
18	5	19	26	26.78
19	6	29	55	54.09
20	7	11	66	66.47
a = 7937.1028 ***b*** = 1.6999 ***c*** = 70.0609 ***D*** = 0.99				

Die Übereinstimmung der Werte ist in dieser Auswertung sehr gut, so wie auch der Determinationskoeffizient mit einem Wert von 0.99. Es ist auch, wie zuvor im Niederländischen, eine auffällige Nähe zum Grenzwert **c** zu erkennen.



Graphik zu Tabelle 11: Zuwachs der aus dem Rotwelschen übernommenen Entlehnungen

Es ist auch hier nochmals darauf zu achten, dass die Genauigkeit der Graphik auf der y- Achse zunimmt, da es sich im Gesamten um weniger Wörter handelt als in den vorherigen Auswertungen. Des Weiteren ist auch hier der Wendepunkt zu erkennen, der uns in der Graphik zwischen dem 18. und 19. Jahrhundert begegnet und die Annährung an die Asymptote erkennen lässt.

Das Rotwelsche ist die Sprache des fahrenden Volkes und später auch die Sprache, die Schurken benutzten, um sich untereinander zu verständigen. Dies lässt vermuten, dass sich die Übernahme von Wörtern in den deutschen Wortschatz durch das Herumziehen dieser Gruppen vollzogen hat. Ein großer Bestandteil der aus dem Rotwelschen entlehnten Wörter

kommt nicht aus dieser Sprache selbst, sondern ist zuvor aus dem Jiddischen entlehnt worden (vgl. Seebold, 1981: 78f.).

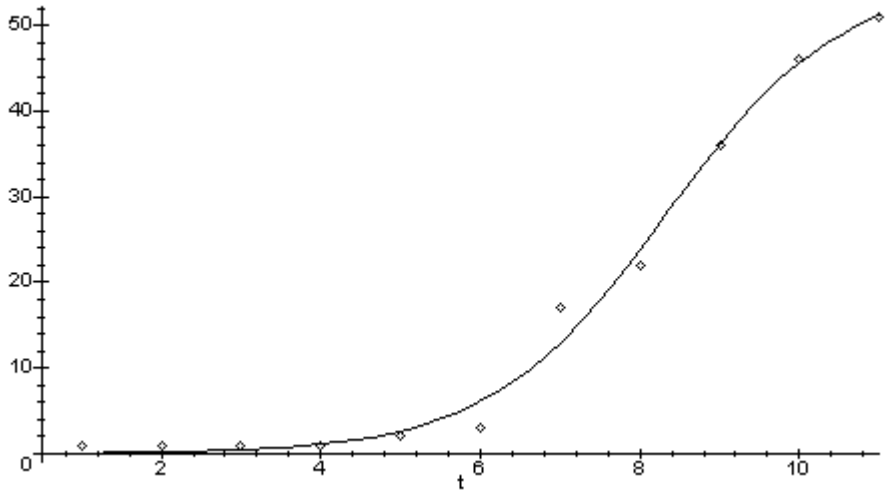
Dieser Zusammenhang des Rotwelschen mit dem Jiddischen ist wohl einerseits auf die frühere Stellung der Juden als Händler zurückzuführen, andererseits darauf, daß im Mittelalter das sozial ausgestoßene fahrende Volk einen Teil seines Wortschatzes von einem anderen sozial ausgestoßenen Bevölkerungsteil, den Juden […] übernahm (Seebold, 1981: 78f.).

3.2.10. Spanisch

Für die Übernahme von spanischen Wörtern in den deutschen Wortschatz ergab die Auswertung die nachstehende Tabelle:

Tabelle 12
Zuwachs der aus dem Spanischen übernommenen Entlehnungen

Jahrhundert	**t**	***n***	***n* (kumuliert)**	***p* (berechnet)**
10	1	1	1	0.07
11	2	0	1	0.18
12	3	0	1	0.45
13	4	0	1	1.10
14	5	1	2	2.64
15	6	1	3	6.09
16	7	14	17	12.93
17	8	5	22	23.79
18	9	14	36	36.06
19	10	10	46	45.61
20	11	5	51	51.10
a = 1828.2858 ***b*** = 0.9022 ***c*** = 55.6783 ***D*** = 0.99				



Graphik zu Tabelle 12: Zuwachs der aus dem Spanischen übernommenen Entlehnungen

Die erste Entlehnung aus dem Spanischen ist im 10. Jahrhundert zu verzeichnen, woraufhin in den nächsten drei Jahrhunderten keine weiteren folgen, bis im 14. und 15. Jahrhundert jeweils eine Entlehnung notiert werden kann. Anschließend gibt es einen Aufschwung, der aber wieder abflacht und noch einmal aufsteigt, bevor er rückläufig wird. Trotz dieser Auf- und Abbewegung ergibt sich für das Spanische eine sehr gute Anpassung des Piotrowski-Gesetzes.

Die Entlehnungen aus dem Spanischen sind entweder zusammen mit den Gegenständen oder Vorgängen direkt aus dem Spanischen übernommen oder das Spanische dient als Vermittlersprache: in einigen Fällen sind die Wörter aus dem Arabischen oder dem Lateinischen entlehnt; manche stammen auch infolge des Kolonialismus aus südamerikanischen Indianersprachen.

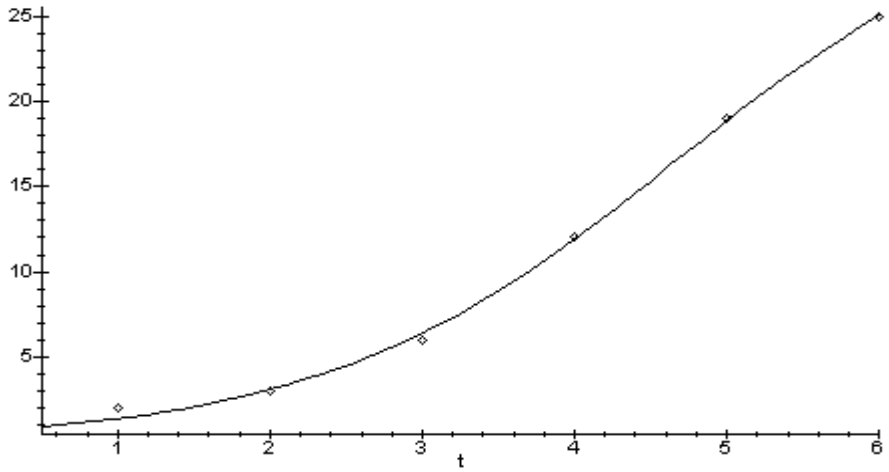
3.2.11. Jiddisch

Die Entlehnungen aus dem Jiddischen beginnen, ähnlich wie auch die des Rotwelschen, mit dem 15. Jahrhundert erst relativ spät. Insgesamt ergibt sich die folgende Auswertung:

Tabelle 13
Zuwachs der aus dem Jiddischen übernommenen Entlehnungen

Jahrhundert	**t**	***n***	***n* (kumuliert)**	***p* (berechnet)**
15	8	2	2	1.37
16	9	1	3	3.05
17	10	3	6	6.38
18	11	6	12	11.92
19	12	7	19	18.86
20	13	6	25	25.07
a = 54.6366 ***b*** = 0.8562 ***c*** = 33.1184 ***D*** = 0.99				

Auch in diesem Fall ist eine sehr gute Übereinstimmung zwischen den beobachteten und berechneten Werten zu erkennen und auch der Determinationskoeffizient liegt mit einem Wert von 0.99 bei einer sehr guten Anpassung an das Piotrowski-Gesetz. Anhand der Graphik wird dies noch einmal deutlich:



Graphik zu Tabelle 13: Zuwachs der aus dem Jiddischen übernommenen Entlehnungen

Wie schon erwähnt, sind viele der jiddischen Wörter über das Rotwelsche in den deutschen Wortschatz entlehnt. Das Jiddische ist „[…] die Sprache der Juden, in neuerer Zeit nur noch der Ostjuden […]“ (Seebold, 1981: 78) und baut sich aus deutschen Mundarten des Mittelalters und hebräischen Wortbestandteilen auf (vgl. Seebold, 1981: 78). Es lässt sich auch hier die Vermutung anstellen, dass sich die Übernahmen daraus ergeben haben, dass die Juden, die als Händler unterwegs waren, ihre Sprache durch die Kommunikation auf ihren Reisen verbreitet haben.

Auf einer breiteren Datenbasis wurde der Zuwachs der Jiddismen im Deutschen von Best (2006c) untersucht; dabei konnten 124 Jiddismen nachgewiesen werden, von denen 90 datierbar waren. Der Trend zeigt auf dieser Grundlage einen flacheren Verlauf, was auf einen geringeren Zuwachs in der Gegenwart hindeutet.

3.2.12. Russisch

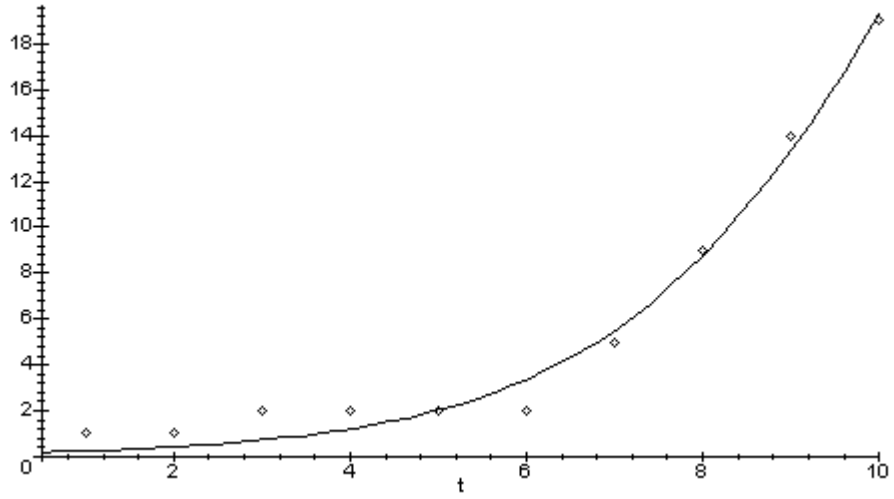
Für die russische Sprache wurden durch die Auswertung nur 19 datierbare Belege nachgewiesen, was somit unterhalb der geforderten 20 Belege zur Berechnung des Piotrowski-Gesetzes lag. Dennoch wurde eine Anpassung an das Piotrowski-Gesetz versucht. Die folgende Tabelle zeigt die Auswertung der Daten und deren Berechnung:

Tabelle 14
Zuwachs der aus dem Russischen übernommenen Entlehnungen

Jahrhundert	**t**	***n***	***n* (kumuliert)**	***p* (berechnet)**
11	1	1	1	0.24
12	2	0	1	0.41
13	3	1	2	0.71
14	4	0	2	1.20
15	5	0	2	2.02
16	6	0	2	3.37
17	7	3	5	5.51
18	8	4	9	8.77
19	9	5	14	13.37
20	10	5	19	19.27
a = 354.3910 ***b*** = 0.5394 ***c*** = 50.3134 ***D*** = 0.98				

Es ist zu erkennen, dass trotz der geringen Belege die Übereinstimmung von Beobachtung und Berechnung sehr hoch ist und dass mit dem Determinationskoeffizienten von 0.98 eine sehr gute Anpassung an das Piotrowski-Gesetz erzielt werden konnte, was auch die Graphik zeigt:

In den meisten der für diese Studie ausgewerteten Entlehnungen aus dem Russischen dient das Russische als Herkunftssprache und bezieht sich ebenso wie das Spanische auf Gegenstände oder Vorgänge, die zusammen mit ihrer Bezeichnung übernommen wurden. In manchen Fällen ist das Russische aber auch Vermittlersprache, wobei der Ursprung oft nicht zu bestimmen ist.



Graphik zu Tabelle 14: Zuwachs der aus dem Russischen übernommenen Entlehnungen

Weitere Untersuchungen zu Entlehnungen aus dem Russischen finden sich in Best (2003b) und Kotsyuba (2007). Verglichen mit dem hier gewonnenen Trend erweist sich der Zuwachs an Russismen im 20. Jahrhundert aufgrund der anderen Datenbasis in diesen beiden Fällen als weniger dynamisch.

3.2.13. Die romanische Sprachfamilie

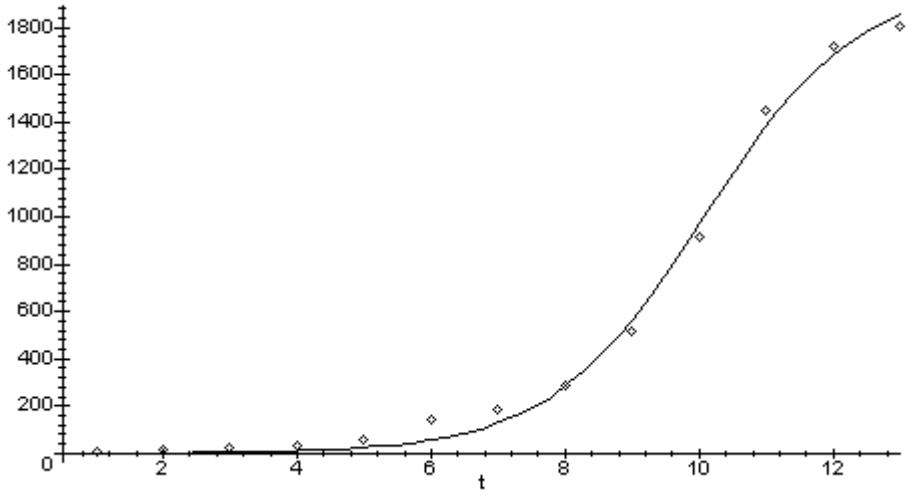
Einige Angaben aus dem romanischen Sprachbereich sind in *Kluge* (2002) mit „frühromanisch“ oder „romanisch“ beschrieben, da die für die Entlehnung ins Deutsche ausschlaggebende Form nicht nachgewiesen werden kann (vgl. Seebold, 2002: XXIX). Zu dieser Auswertung wurden dann neben den entsprechend gekennzeichneten Stichwörtern die Belege der schon bearbeiteten Sprachen Französisch, Italienisch und Spanisch und die aus der Auswertung des Wörterbuches gewonnenen Sprachen Portugiesisch, Provenzalisch, Venezianisch und Räto-Romanisch hinzugefügt. Die zuletzt genannten Sprachen weisen für sich allein zu wenige Belege auf, um sie einer Auswertung zu unterziehen, konnten aber hier unter dem Sammelbegriff „romanische Sprachen“ verwendet werden.

Tabelle 15
Zuwachs der aus dem Romanischen übernommenen Entlehnungen

Jahrhundert	**t**	***n***	***n* (kumuliert)**	***p* (berechnet)**
8	1	7	7	0.74
9	2	11	18	1.77
10	3	8	26	4.21
11	4	8	34	10.03
12	5	27	61	23.77
13	6	79	140	55.83
14	7	49	189	128.32
15	8	96	285	281.30
16	9	227	512	561.90

17	10	402	914	965.23
18	11	532	1446	1380.24
19	12	277	1723	1683.44
20	13	80	1803	1854.05
a = 6445.3457 ***b*** = 0.8701 ***c*** = 2000.1687 ***D*** = 0.99				

Der Determinationskoeffizient zeigt, dass es trotz der zusammengefassten Sprachen möglich ist, eine sehr gute Anpassung des Piotrowski-Gesetzes zu erreichen und auch die Graphik spiegelt dieses Ergebnis wider:



Graphik zu Tabelle 15: Zuwachs der aus dem Romanischen übernommenen Entlehnungen

Es ist hier eine ähnliche Entwicklung zu sehen, wie dies schon bei der Bearbeitung der romanischen Sprachen Französisch, Italienisch und Spanisch zu erkennen ist. Denn auch hier lässt sich der Wendepunkt um das 18. Jahrhundert herum erkennen und der Graph nähert sich der Asymptote, woraus sich schließen lässt, dass sich die Entlehnungen aus den Romanischen Sprachen verringern werden. Diese Anpassung bestätigt damit die in dieser Studie untersuchten einzelnen Sprachen romanischer Herkunft, die ebenfalls ergeben, dass nur noch mit einer geringen Anzahl von Entlehnungen aus diesen Sprachen zu rechnen ist. Wobei die Anzahl der weiteren Entlehnungen, wie der Grenzwert zeigt, in diesem Fall höher ausfallen, als bei den Einzeluntersuchungen, aber schließlich handelt es sich hier auch um die Zusammenfassung von verschiedenen Sprachen und es muss bedacht werden, dass die Entlehnungen aus den verschiedenen Sprachen aufgenommen werden.

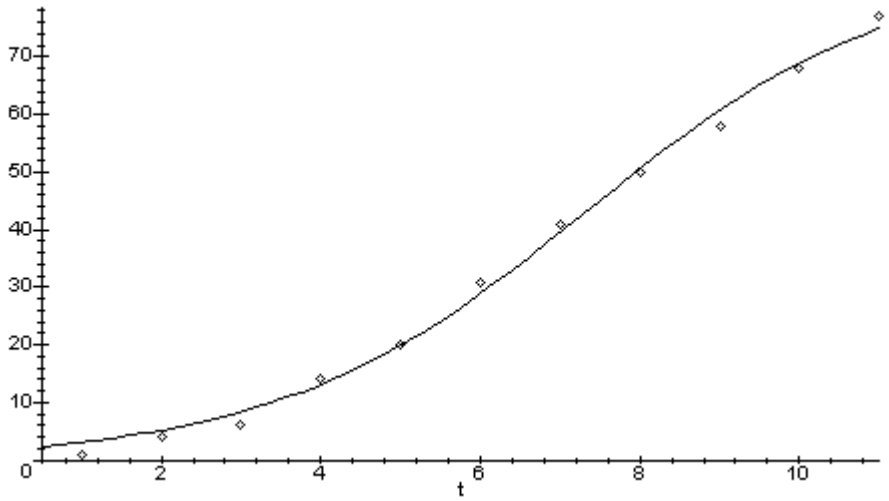
3.2.14. Die slawische Sprachfamilie

Auch im Fall der verschiedenen slawischen Sprachen weisen die Einzelsprachen in den meisten Fällen zu wenige Belege auf, um mit ihnen eine Auswertung vorzunehmen, mit Ausnahme des Russischen. Zu dieser Sprachfamilie zählen, innerhalb dieser Untersuchung, neben dem Russischen, die in *Kluge* (2002) als „slawisch“ gekennzeichneten Stichwörter sowie die des Polnischen, Russischen, Sorbischen, Tschechischen, Slowenischen, Serbokroatischen, Serbischen, Kroatischen und Polabischen.

Tabelle 16
Zuwachs der aus dem Slawischen übernommenen Entlehnungen

Jahrhundert	**t**	***n***	***n* (kumuliert)**	***p* (berechnet)**
10	3	1	1	3.15
11	4	3	4	5.15
12	5	2	6	8.32
13	6	8	14	13.10
14	7	6	20	19.94
15	8	11	31	28.93
16	9	10	41	39.54
17	10	9	50	50.59
18	11	8	58	60.70
19	12	10	68	68.89
20	13	9	77	74.92
a = 44.2221 ***b*** = 0.5183 ***c*** = 85.9921 ***D*** = 0.99				

Es ist auch hier wieder eine sehr gute Anpassung zu sehen, was die Nähe der beobachteten und berechneten Werte und der Determinationskoeffizient von 0.99 zeigt und durch die nachfolgende Graphik verdeutlicht wird:



Graphik zu Tabelle 16: Zuwachs der aus dem Slawischen übernommenen Entlehnungen

Im Gegensatz zur romanischen Sprachfamilie ist die Übereinstimmung der Werte hier offensichtlicher und der Graph zeigt einen flacheren Verlauf. Der Wendepunkt des Graphen zeigt sich zwischen dem 17. und 18. Jahrhundert, ab welchem sich der Graph an die Asymptote annähert. Dieses zusammen mit dem Grenzwert zeigt auch hier eine Abnahme der zukünftigen Entlehnungen an. Dies steht im Gegensatz zu der russischen Untersuchung, die wie beschrieben noch keinen Wendepunkt erkennen lässt.

In einer Untersuchung von Best (2003b) zu Entlehnungen aus dem slawischen Sprachgebiet ergibt sich ein ähnliches Bild wie in dieser Studie. Bei der Auswertung der slawischen Wörter einschließlich des Russischen nähert sich der Graph der Asymptote, während die russischen Wörter für sich genommen den Wendepunkt des Graphen noch nicht erkennen lassen (vgl. Best, 2003b).

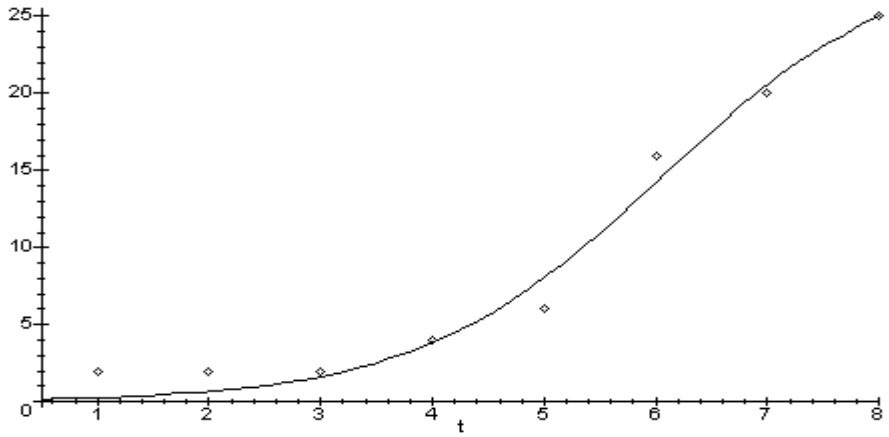
3.2.15. Die nordgermanische Sprachfamilie

Trotz einer gemeinsamen Auswertung der nordgermanischen Sprachen, worunter Entlehnungen aus dem Altnordischen, dem Isländischen, Schwedischen, Norwegischen und Dänischen zu zählen sind (färöische Belege wurden nicht nachgewiesen), fanden sich für diese Sprachfamilie nur 25 Belege. Das ist eine geringe Menge, aber für die Auswahlkriterien dieser Studie genug, um erstmals für diese Sprachfamilie eine Anpassung durchzuführen.

Tabelle 17
Zuwachs der aus dem Nordgermanischen übernommenen Entlehnungen

Jahrhundert	**t**	***n***	***n* (kumuliert)**	***p* (berechnet)**
13	6	2	2	0.28
14	7	0	2	0.68
15	8	0	2	1.66
16	9	2	4	3.84
17	10	2	6	8.04
18	11	10	16	14.24
19	12	4	20	20.57
20	13	5	25	24.98
a = 261.8466 ***b*** = 0.9206 ***c*** = 29.1246 ***D*** = 0.98				

Wie anhand der Tabelle zu erkennen ist, war auch in diesem Fall eine Anpassung des Piotrowski-Gesetzes möglich. Mit einem Determinationskoeffizienten von 0.98 zeigt auch diese Anpassung ein sehr gutes Ergebnis.



Graphik zu Tabelle 17: Zuwachs der aus dem Nordgermanischen übernommenen Entlehnungen

Es ist hier wieder der S-förmige Verlauf des unvollständigen Sprachwandels zu erkennen. Auch wenn sich der Graph erst langsam nach dem Wendepunkt, zwischen dem 18. und 19. Jahrhundert, der Asymptote nähert, ist doch zu erkennen, dass in Zukunft wohl nicht mehr viele Entlehnungen aus den nordgermanischen Sprachen in den deutschen Wortschatz aufgenommen werden. Ein weiteres Anzeichen dafür gibt auch der in der Tabelle aufgeführte Grenzwert **c**.

4. Diskussion und Ausblick

Bei den hier vorgestellten Entlehnungs- oder Wachstumsprozessen ist immer zu beachten, dass sie auf der Basis der Daten gewonnen sind, die *Kluge* (2002), dem neuesten etymologischen Wörterbuch des Deutschen, entstammen. Alle Ergebnisse sind also vor allem durch die Auswahl der Stichwörter beeinflusst, die nur etwa ein Zwanzigstel eines umfangreichen, einbändigen Wörterbuchs des Deutschen (etwa: *Duden. Deutsches Universalwörterbuch* 2001) ausmachen. Daneben werden die Ergebnisse aber auch dadurch beeinflusst, dass längst nicht für alle ausgewählten Stichwörter hinreichend genaue Datierungen möglich sind, ganz zu schweigen davon, dass sich sicher nicht alle Datierungen auf Dauer als korrekt herausstellen werden. Hinzu kommt ein Problem, das allen derartigen Untersuchungen immanent ist: Es werden Daten zu solchen Stichwörtern ausgewertet, die in dem etymologischen Wörterbuch noch heute als zum Wortschatz des Deutschen gehörig aufgeführt sind; viele andere sind aber auch schon wieder aus dem Sprachgebrauch verschwunden, deren Zuwachs und Verschwinden in den Tabellen und Graphiken völlig außer Acht bleibt.

Der Wortschatz der deutschen Sprache ist also weitaus umfangreicher als die Anzahl der Wörter in der hier zugrunde liegenden Untersuchung, was dazu führt, dass die Repräsentativität dieser Studie zum deutschen Wortschatz begrenzt ist. Dennoch kann mit einem Blick auf die Untersuchung von Körner (2004) und auch anderen Studien (vgl. z.B. Best 2001a, 2001b, Best & Altmann 1986) eine Tendenz festgestellt werden, die zu ganz ähnlichen Ergebnissen führt. Vergleicht man die hier vorliegende Untersuchung mit derjenigen von Körner (2004), gibt es zwar Abweichungen, aber die Übereinstimmung der Reihenfolge der Geber-Sprachen auf den ersten Plätzen ist zu erkennen. So steht in beiden Untersuchungen Latein mit dem höchsten Prozentsatz an der ersten Stelle, dahinter befinden sich das Französische und das Niederdeutsche, welchem dann das Englische und Italienische folgen. Erst nach weiteren zwei Sprachen, Griechisch und Niederländisch, ändert sich die Reihenfolge der Auswertung.

Die Unterschiede beruhen auf der jeweiligen Datenbasis, die die zugrunde gelegten Wörterbücher bieten. Wie schon durch die Auswertung der einzelnen Sprachen ersichtlich wurde, führte die Anpassung des Piotrowski-Gesetztes in der unvollständigen Form bei jeder der in dieser Studie untersuchten Sprachen zu einem sehr guten Ergebnis, da bei einem Determinationskoeffizienten ab 0.9 das Ergebnis als sehr gut angesehen werden kann (vgl. Best, Beöthy & Altmann, 1990: 122f.) und alle hier untersuchten Sprachen deutlich darüber liegen. Somit wird das Piotrowski-Gesetz in seiner unvollständigen Form in dieser Studie bestätigt. Dabei ist zu beachten, dass der Parameter *c*, welcher als Obergrenze der Entlehnungen verstanden wird, nur für das jeweils zugrunde liegende Datenmaterial Gültigkeit hat und sich mit der Verwendung von anderen Daten ändert (vgl. Best & Altmann, 1986: 38).

Für einige Sprachen liegen Untersuchungen vor, die den Daten, die sich aus Kluge (2002) gewinnen ließen, aufgrund ihrer Fülle überlegen sind und alle - mit einer Ausnahme - ebenfalls die Hypothese stützen, dass das Piotrowski-Gesetz in der Form für den unvollständigen Sprachwandel ein gutes Modell für solche Prozesse darstellt. Es handelt sich dabei um Untersuchungen zu den folgenden Sprachen oder Sprachfamilien: Arabisch (Best 2004), Chinesisch (Best 2008), Englisch (Best 2006b: 114), Italienisch (Best 2006a), Japanisch (Best 2009a), Jiddisch (Best 2006c), Russisch (Best 2003b, Kotzyuba 2007), Slawisch (Best 2003b) und Türkisch (Best 2005). Nur im Falle der Entlehnungen aus dem Chinesischen wurde die Zahl der datierbaren Wörter mit nur 12 als zu niedrig erachtet; ein Test des Modells unterblieb daher.

Eine andere Frage drängt sich auf: Kann man aufgrund der festgestellten Trends Prognosen für die Zukunft stellen? Diese Frage hat Best (2009b) mit Hilfe von Computer-Experimenten zu beantworten versucht. Dabei hat sich gezeigt, dass recht brauchbare Pro-

gnosen zu erwarten sind, wenn der Wendepunkt des betreffenden Entlehnungsprozesses deutlich überschritten ist. Da dies hier in den meisten Fällen gegeben ist, kann man die vorgestellten Trends mit einiger Vorsicht auch als Prognosen interpretieren. Im Falle des Englischen und des Russischen kann jedoch aufgrund die hier vorgelegten Daten keine Prognose gewagt werden, da nicht erkennbar ist, ob der Wendepunkt bereits erreicht oder gar überschritten wurde.

Literaturverzeichnis

Wörterbücher

Duden. Herkunftswörterbuch (2001, 3. Auflage). Mannheim; Leipzig; Wien; Zürich: Bibliographisches Institut-Dudenverlag.

Duden. Das Fremdwörterbuch (2007, 5. Auflage). Mannheim; Leipzig; Wien; Zürich: Bibliographisches Institut-Dudenverlag.

Duden. Deutsches Universalwörterbuch (2001, 4. Auflage). Mannheim; Leipzig; Wien; Zürich: Bibliographisches Institut-Dudenverlag.

Kluge, Friedrich (2002, 24., durchgesehene und erweiterte Auflage): *Etymologisches Wörterbuch der deutschen Sprache.* Bearbeitet von Elmar Seebold. Berlin; New York: Walter de Gruyter.

Wahrig, Gerhard (2000, 7. Auflage): *Deutsches Wörterbuch.* Gütersloh; München; Bertelsmann Lexikon Verlag.

Literatur

Altmann, Gabriel (1983). Das Piotrowski-Gesetz und seine Verallgemeinerung. In: Best, Karl-Heinz, Kohlhase, Jörg (Hrsg.): *Exakte Sprachwandelforschung. Theoretische Beiträge, Statistische Analysen und Arbeitsberichte* (S. 59-90). Göttingen: edition herodot.

Best, Karl-Heinz (2001a). Ein Beitrag zur Fremdwortdiskussion. In: Schierholz, Stefan (Hrsg.): *Die deutsche Sprache in der Gegenwart. Festschrift für Dieter Cherubim zum 60. Geburtstag* (S. 263-270). Frankfurt am Main: Europäischer Verlag der Wissenschaft.

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

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

Best, Karl-Heinz (2003b). Slawische Entlehnungen im Deutschen. In: Kempgen, S., Schweier, U., Berger, T.: *Festschrift für Werner Lehfeldt zum 60. Geburtstag. Die Welt der Slaven* (S. 465-473). München: Verlag Otto Sagner.

Best, Karl-Heinz (2004). Zur Ausbreitung von Wörtern arabischer Herkunft im Deutschen. *Glottometrics 8, 75-7.*

Best, Karl-Heinz (2005). Turzismen im Deutschen. *Glottometrics 11, 56-63.*

Best, Karl-Heinz (2006a). Italianismen im Deutschen. *Göttinger Beiträge zur Sprachwissenschaft 13, 77-86.*

Best, Karl-Heinz (2006b). *Quantitative Linguistik: Eine Annäherung.* 3., stark überarbeitete und ergänzte Auflage. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (2006c). Quantitative Untersuchungen zu den Jiddismen im Deutschen. *Jiddistik Mitteilungen 36, 1-14.*

Best, Karl-Heinz (2007). Quantitative Untersuchungen zum deutschen Wortschatz. *Glottometrics 14, 32-45.*

Best, Karl-Heinz (2008). Sinismen im Deutschen und Englischen. *Glottometrics 17, 87-93.*

Best, Karl-Heinz (2009a). Zur Entwicklung der Entlehnungen aus dem Japanischen ins Deutsche. *Glottometrics 19, 80-84.*

Best, Karl-Heinz (2009b). Sind Prognosen in der Linguistik möglich? In: *Typen von Wissen: Begriffliche Unterscheidung und Ausprägungen in der Praxis des Wissenstransfers, S. 164-175*. Hrsg. v. Tilo Weber und Gerd Antos. Frankfurt/M.: Lang.

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

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

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

Bußmann, Hadumod (2002, 3. Auflage). *Lexikon der Sprachwissenschaft*. Stuttgart: Alfred Kröner Verlag.

Hennings, Thordis (2003, 2. Auflage). *Einführung in das Mittelhochdeutsch*. Berlin; New York: Walter de Gruyter.

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

Kotsyuba, Oxana (2007). Russizismen im deutschen Wortschatz. *Glottometrics 15, 13-23.*

Imsiepen, Ulrike (1983). Die e-Epithese bei starken Verben im Deutschen. In: Best, Karl-Heinz, Kohlhase, Jörg (Hrsg.): *Exakte Sprachwandelforschung. Theoretische Beiträge, Statistische Analysen und Arbeitsberichte* (S. 119-142). Göttingen: edition herodot.

Leopold, Edda (2005). Das Piotrowski-Gesetz. In: Altmann, Gabriel, Köhler, Reinhard, Piotrowski, Rajmund (Hrsg.): *Quantitative Linguistik- Quantitative Linguistics. Ein internationales Handbuch* (S. 627-633). Berlin; New York: Walter de Gruyter.

Lucko, Peter (1995). Englisch im deutschen Wortschatz. Eine Einführung. In: Sörensen, Ilse: *Englisch im deutschen Wortschatz. Lehn- und Fremdwörter in der Umgangssprache* (S. 14-18). Berlin: Volk und Wissen Verlag.

Naumann, Horst (2007, 10. Auflage): Das deutsche des Frühmittelhochdeutschen (6.-11. Jahrhundert). In: Schmidt, Wilhelm (Hrsg.): *Geschichte der deutschen Sprache. Ein Lehrbuch für das germanistische Studium*. Stuttgart: Hirzel. S. 63-90.

Polenz, Peter v. (2000, 2. Auflage). *Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart. Band I Einführung. Grundbegriffe. 14. bis 16. Jahrhundert.* Berlin; New York: Walter de Gruyter.

Schmöe, Friederike (1998). *Italianismen im Gegenwartsdeutschen unter besonderer Berücksichtigung der Entlehnungen nach 1950*. Bamberg: Collibri-Verlag.

Seebold, Elmar (1981). *Etymologie. Eine Einführung am Beispiel der deutschen Sprache.* München: Verlag C.H.Beck.

Seebold, Elmar (Hrsg.) (2001). *Chronologisches Wörterbuch des deutschen Wortschatzes. Der Wortschatz des 8. Jahrhunderts (und frühere Quellen).* Berlin: Walter de Gruyter.

Seebold, Elmar (2002): Einführung in die Terminologie. In: Kluge, Friedrich (2002, 24. Auflage): *Etymologisches Wörterbuch der deutschen Sprache* (S. XIII-XLI). Berlin; New York: Walter de Gruyter.

Telling, Rudolf (1987*). Französisch im deutschen Wortschatz. Lehn- und Fremdwörter aus acht Jahrhunderten.* Berlin: Volk und Wissen Volkseigener Verlag.

Wittstock, Otto (1982, 3. Auflage). *Latein und Griechisch im deutschen Wortschatz. Lehn- und Fremdwörter altsprachlicher Herkunft.* Berlin: Volk und Wissen Volkseigener Verlag.

Internetquellen

Europäische Charta der Regional- oder Minderheitssprachen:
http://conventions.coe.int/treaty/ger/Treaties/Html/148.htm; letzter Zugriff: 24.09.2009.

Software

NLREG. *Nonlinear Regression Analysis Program*. Ph. H. Sherrod. Copyright © 1991-2001.

Glottometrics 21, 2011, 54-59

On stratification in poetry

Ioan-Iovitz Popescu, Bucharest
Radek Čech, Ostrava
Gabriel Altmann, Lüdenscheid

Abstract. Texts are composed of many different strata on different levels. A method is proposed to find the number of strata at the word-form level in Slovak poetry and to study the relationship between the parameters of the fitting function.

Keywords: Rank-frequency distribution, word forms, stratification, Slovak poetry

Strata arise in texts by mixing different means of expression which can be of quite variegated kind. For example words from different word classes, words of different length, interjections, different sentence types, different pictures of reality, etc., may bring about a kind of stratification. The methods of investigation are very few; as a matter of fact, there is only one work in which the rank-frequency distribution of word-forms – the so called "Zip's law" – is considered a result of stratification (cf. Popescu, Altmann, Köhler 2010).

If we knew what classes are present in a writer's brain at the moment of writing, we would be able to separate them. However, this is not possible in general and every step in this direction is merely a trial, an empirical approximation of the state of the affairs. If one supposes the existence of stratification, one may scrutinize the phenomenon by setting up the rank-frequency distribution of some supposed classes. The rank-frequency distribution of linguistic entities abides by the function

$$(1) \qquad y = 1 + a \bullet \exp(-x/b) + c \bullet \exp(-x/d) + \ldots..$$

The number of exponential components signalizes the number of strata. The constant 1 is added because frequencies cannot be smaller, hence the function converges to 1. In difference to polynomials whose use in text analysis cannot be recommended, the above function shows which components are redundant: if the constants in the exponents of two components are equal or almost equal, then one of the components is redundant and can be omitted.

In order to illustrate this property we computed the rank-frequency distribution of word forms in the poem *Aby spriesvitnela* by Eva Bachletová and obtained the result in Table 1

Table 1
Rank-frequency distribution of word forms
in Bachletová's poem *Aby spriesvitnela*

r	f_r	r	f_r	r	f_r	r	f_r
1	4	14	1	27	1	40	1
2	3	15	1	28	1	41	1
3	3	16	1	29	1	42	1
4	2	17	1	30	1	43	1
5	2	18	1	31	1	44	1
6	2	19	1	32	1	45	1
7	1	20	1	33	1	46	1
8	1	21	1	34	1	47	1
9	1	22	1	35	1	48	1
10	1	23	1	36	1	49	1
11	1	24	1	37	1	50	1
12	1	25	1	38	1	51	1
13	1	26	1	39	1	52	1
						53	1

Fitting formula (1) to these data using only one component we obtain

$$f_r = 1 + 4.2851 \exp(-r/2.9793)$$

yielding the determination coefficient $R^2 = 0.955$. If we add a second component, we obtain

$$f_r = 1 + 1.9208 \exp(-r/2.9793) + 2.3823 \exp(-r/2.9793)$$

with the same $R^2 = 0.955$. As can be seen, the constants in the exponents are equal and the sum of the multiplicative constants yields approximately the amplitude in the one-component expression.

Hence we can conclude that the given poem is monostratal. Whatever force controls the word-form strata, it is not sufficiently expressed.

The rank-frequency distribution in the given poem and the fitting function are presented in Figure 1.

In order to study this property, we performed the same fitting in 54 poems of the same author and tested whether one component in (1) is sufficient. The results are

presented in Table 2. As can be seen, all of the poems are non-stratified and express a special feature of author style. Some comments to the results are in order here.

(1) In two cases the determination coefficient is smaller than 0.8 but testing the parameters and the regression by *t*- and *F*-tests yielded always highly significant results ($P < 0.0001$). Thus in all cases the theory of background stratification can be considered as corroborated by these data. If we compare the present fitting with the traditional "Zipfian" one using the power function, we can see that in each case function (1) yields better results.

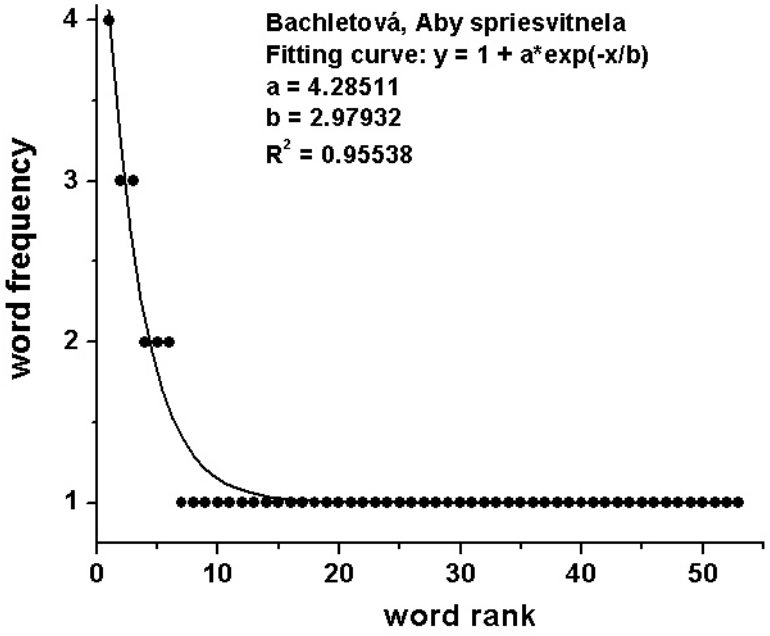


Figure 1. Rank-frequency distribution in E. Bachletová's poem

(2) Some values of parameter *a* seem to attain senseless magnitudes. This is caused by the fact that only one word has been repeated two or three times; all the rest of words occur only once. On the other hand, in all those cases the parameter *b* is very small, the exponential itself converges quickly to zero and only the constant 1 is relevant. Poems of this kind can be omitted from some kinds of analysis, because the word-form distribution is almost uniform, i.e. it does not display any tendency.

Table 2
Parameters of the first stratum in poems by E.Bachletová

Poem	**a**	**b**	**R^2**
Aby spriesvitnela	4.2851	2.9793	0.96
Bez rozlúčky	1.8587	2.2925	0.81
Čakáme šťastie...	3.9093	1.8886	0.89
Čakanie na Boží jas	15.2546	1.4538	0.96
Čas pre nádych vône	3.218	3.8553	0.92

Dielo Stvoriteľa	9.1363	2.4957	0.9
Dnešný luxus	4.7167	2.2445	0.96
Do večnosti beží čas	4.7908	2.9642	0.95
Dovoľ mi slúžiť	6586.9417	0.1137	1
Ešte raz	4.5897	2.5977	0.95
Hľadanie odpovedí	2.1771	4.9942	0.86
Iba neha	13.5437	2.6483	0.84
Iba život	12641.6345	0.1143	1
Idem za Tebou	2.4715	3.5407	0.88
Ihly na nebi	3.7075	4.8980	0.92
Istota	4.7167	2.2445	0.96
Keď dohorí deň	6.8233	1.7770	0.97
Kým ich máme	7.0289	1.1594	0.97
Len áno	1.5774	5.5358	0.75
Malé modlitby	2.1771	4.9943	0.85
Malý ošiaľ	5.5359	4.5479	0.85
Miesto pre Nádej	6586.9417	0.1137	1
Moje určenie	24.1348	1.1031	0.87
Nado mnou Ty sám...	10.7635	0.7873	0.99
Náš chrám	24.6841	0.8768	0.96
Naše mamy	5.2926	1.8678	0.97
Naše svetlo	5.9052	4.2045	0.95
Neopusť ma...	10.9337	1.2572	0.99
Nepoznateľné	8.8198	2.9232	0.94
Podobnosť bytia	46.1763	0.5690	0.95
Pravidlá odpúšťania	4.5000	1.4427	0.83
Precitnutie	3.1155	2.1840	0.92
Prvotný sen	12.6744	1.0789	0.99
Rozdelená bytosť	2.1771	4.9942	0.86
Rozťatá prítomnosť	3.3933	6.1867	0.93
Som iná	10.3289	1.7111	0.94
Spájania	3.9093	1.8886	0.89
Stály smútok pre šesť písmen	12.0732	4.1606	0.93
Tá Láska	1.6499	3.9236	0.78
Tak málo úsmevu	38.1267	0.5887	0.99
Ťažko pokoriteľní	3.8321	1.5666	0.94
Tiché verše	2.2500	1.4427	0.83
To všetko je dar	6.4480	3.9576	0.89
Ulomené zo slov	2.7206	2.8457	0.89
Vďaka Pane!	2.2500	1.4427	0.83
Vďaka za deň	1.8587	2.2925	0.81
Večerná ruža	12641.6345	0.1143	1
Večerné ticho	6.1350	0.2421	0.92

Vo večnosti slobodná	9.0387	4.8815	0.96
Vrátili sa	3.1155	2.1840	0.92
Vyznania	2.7206	2.8457	0.9
Z neba do neba	7.6899	2.0037	0.96
Zasľúbenie jasu	2.8415	4.4980	0.82
Zbytočné srdce	13.4423	0.9778	0.92

(3). Omitting the poems with abnormal parameter *a* we can easily state the relationship between the parameters *a* and *b*. In general, one expects a monotonously decreasing $a = f(b)$ because parameter *a* is merely a balancing magnitude responsible for the amplitude of (1). The decrease of frequencies is controlled by parameter *b.* Thus ordered according to increasing *b* we obtain the values presented in Table 3.

Table 3
Relationship between parameters a and b

b	**a**	$\mathbf{a_{theor}}$	**b**	**a**	$\mathbf{a_{theor}}$
0.5690	46.1763	43.3451	2.2925	1.8587	5.2659
0.5887	38.1267	40.0943	2.4957	9.1363	5.0456
0.7873	10.7635	21.3054	2.5977	4.5897	4.9571
0.8768	24.6841	17.1926	2.6483	13.5437	4.9176
0.9778	13.4423	14.0305	2.8457	2.7206	4.7863
1.0789	12.6744	11.8396	2.8457	2.7206	4.7863
1.1031	24.1348	11.4169	2.9232	8.8198	4.7431
1.1594	7.0289	10.5508	2.9642	4.7908	4.7218
1.2572	10.9337	9.3563	2.9793	4.2851	4.7142
1.4427	4.5000	7.8136	3.0062	6.1350	4.7011
1.4427	2.2500	7.8136	3.5407	2.4715	4.5073
1.4427	2.2500	7.8136	3.8553	3.218	4.4342
1.4538	15.2546	7.7426	3.9236	1.6499	4.4210
1.5666	3.8321	7.1178	3.9576	6.4480	4.4147
1.7111	10.3289	6.5177	4.1606	12.0732	4.3809
1.7770	6.8233	6.2991	4.2045	5.9052	4.3743
1.8678	5.2926	6.0412	4.4980	2.8415	4.3361
1.8886	3.9093	5.9882	4.5479	5.5359	4.3304
1.8886	3.9093	5.9882	4.8815	9.0387	4.2977
2.0037	7.6899	5.7289	4.8980	3.7075	4.2963
2.1840	3.1155	5.4147	4.9942	2.1771	4.2883
2.1840	3.1155	5.4147	4.9942	2.1771	4.2883
2.2445	4.7167	5.3286	4.9943	2.1771	4.2883
2.2445	4.7167	5.3286	5.5358	1.5774	4.2522
2.2925	1.8587	5.2659	6.1867	3.3933	4.2226

The given relationship can be represented by a simple power function

$$a = 4.1317 + 9.3496b^{-2.5426}$$

yielding an $R^2 = 0.79$ and very highly significant *t*- and *F*-values. It can be expected that adding further poems by the same author the relationship will get rather stronger. The relationship is graphically presented in Figure 2.

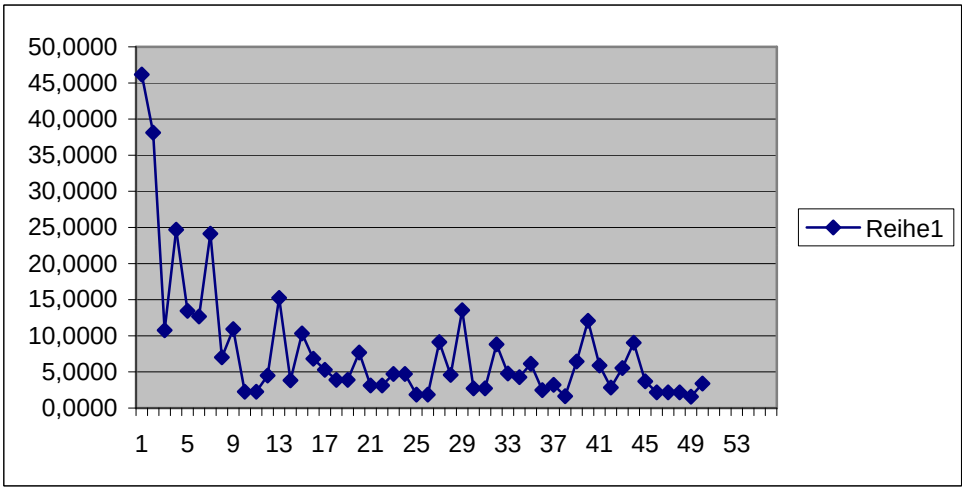


Figure 2. The relationship between parameters *a* and *b*

(4) Automatically, further questions arise that can be pursued in the future: (a) Does the above result hold only for the given writer or can we transfer it to other writers, too? (b) Does it hold only for poetry of this kind (without rhyme, irregular verse) or does it hold for Slovak poetry in general? (c) Does it hold also for poetry in other languages? (d) Does it hold also for prose?

In short texts, the lemmatization of the word forms does not bring new results, not even in strongly synthetic languages. In strongly analytic ones the results are almost identical.

The above result is a strong support for replacing the Zipfian zeta function by formula (1).

References

Popescu, I.-I., Altmann, G., Köhler, R. (2010). Zipf´s law – another view. *Quality and Quantity 44(4), 713-731*.

Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S., Wimmerová, S. (2003). *Úvod do analyzy textov*. Bratislava: Veda.

Glottometrics 21, 2011, 60-64

Towards a model for rank-frequency distributions of melodic intervals

Ján Mačutek[1], *Bratislava*
Zuzana Švehlíková, Bratislava
Zuzana Cenkerová, Bratislava

Abstract. The rank-frequency distribution of melodic intervals occurring in the Palestrina's *Missa Papae Marcelli* are presented. The data are modeled by the negative hypergeometric distribution. The evaluation of goodness-of-fit measures is discussed.

Keywords: melodic intervals, negative hypergeometric distribution, goodness-of-fit

1. Introduction

Although statistical investigations in the area of melodic intervals (i.e., distances between successive notes, cf. Lindley et al. 2001) appeared occasionally (cf. an overview in Nettheim 1997), it seems that none of them focused on a theoretical mathematical model. Watt (1924) compared sizes of melodic intervals on two very different music styles. Suchoff (1970) examined melodic interval size (on the level of types, i.e., each interval is taken into account once, regardless of the frequencies with which particular intervals occur) in eight-note incipits of Romanian and Serbo-Croatian folksongs, and showed that the first intervals usually ascend. Hofstetter (1979) counted intervals in the Romantic chamber music. Voss and Troost (1989) and Huron (1990a,b) report contradicting results on the relation between the direction (ascending/descending) and size of melodic intervals, which can be influenced by several boundary conditions. Pont (1990) examined interval direction (but not sizes) in three-note incipits and showed differences between European and non-European music.

This paper follows a twofold aim, the first of them being a presentation of an empirical finding together with an attempt to find a mathematical model. According to Altmann (2005), *a … classification is "good", "useful" or "theoretically prolific" if the taxa follow a "decent" rank-frequency distribution.* Therefore, in Sections 2 and 3 we suggest modelling the rank-frequency distribution of melodic intervals (i.e., observed frequencies of melodic intervals in descending order) by the negative hypergeometric distribution (cf. Wimmer and Altmann 1999: 465-168). Data are taken from the *Missa Papae Marcelli* by G.P. da Palestrina.

It is shown that fitting the negative hypergeometric distribution to the data yields contradictory results if "traditional" rejection rules are applied: the fit is not satisfactory in terms of discrepancy coefficient C (which equals the Pearson's χ^2-statistic divided by the sample size), but it can be considered very good in terms of determination coefficient R^2.

This observation leads us to the other aspect of the paper (cf. Section 4). We intend to initiate a discussion on the evaluation of goodness of fit (especially) for discrete distribution models. The topic is of utmost relevance not only in quantitative music analysis or in quan-

[1] Department of Applied Mathematics and Statistics, Faculty of Mathematics, Physics and Informatics, Comenius University, Mlynská dolina, 84248 Bratislava, Slovakia, e-mail: jmacutek@yahoo.com

titative linguistics, but for modelling empirical data and fitting mathematical models in general.

2. Data

The *Missa Papae Marcelli,* composed by Giovanni Pierluigi da Palestrina (1525?-1594), is a well known six-voice[2] Renaissance mass. The notes were extracted from freely accessible midi files by software programs, with the exception of flats and sharps, which were added manually. Then, melodic intervals occurring in the mass were determined by a program written in open source statistical software R. There are, however, at least three possibilities[3] how to treat successive notes in a musical score:

a) to consider notes as they appear, irrespective of rests and double bar lines;
b) to respect double bar lines only, and not rests (i.e., two notes divided by a double bar line do form an interval, but two notes divided by a rest do);
c) to respect both rests and double bar lines (i.e., notes divided by either a rest or a double bar line do not form an interval).

The approach c) was chosen. Two successive notes on the same pitch were connected by a ligature were not considered as the prime; without a ligature they were treated as the prime.

3. Results

The observed rank-frequency distribution of melodic intervals in the *Missa Papae Marcelli* and the respective fitted values are presented in Table 1 and Figure 1. Interestingly enough, the best fit is obtained for the negative hypergeometric distribution[4] with

$$P(x) = \frac{\binom{M+x-2}{x-1}\binom{K-M+n-x}{n-x+1}}{\binom{K+n-1}{n}}$$

for $x = 1, 2, \ldots, n+1$; *K, M, n* being its parameters. The same distribution was used by Martináková-Rendeková (2000, 2007, 2008) as a model for ranked frequencies of pitch values[5] in many compositions. Even the values of parameters *K* and *M* are quite similar to the ones from models of pitch values. More compositions from different musical styles must be examined in order to answer the question whether the distribution and its parameter values are typical for melodic intervals or whether they are specific for the composition under study (or perhaps specific for Palestrina, Renaissance music, vocal music, etc.). Moreover, parameters must be interpreted if models of ranked frequencies of melodic intervals are to be integrated into

[2] There are two exceptions: *Benedictus* is written for four voice parts, *Agnus Dei 2* for seven voice parts.
[3] Another one, which was not examined in the presented analysis, is to include a standard bar line into these considerations.
[4] To be exact, the distribution is shifted to the right by 1, since the first rank is 1.
[5] In the cited papers only the pitches as physical frequencies were considered (i.e., the papers do not distinguish between, e.g., C sharp and D flat).

general synergetic musicological model (cf. Köhler and Martináková-Rendeková 1998, Martináková-Rendeková 2000, Došeková 2006).

Table 1
Fitting the negative hypergeometric distribution
to the rank frequency distribution of melodic intervals
(x – rank, $f(x)$ – observed frequencies, $NP(x)$ – expected frequencies).

x	Interval	f(x)	NP(x)
1	major second	2772	2672.94
2	minor second	1379	1475.82
3	prime	1184	982.71
4	minor third	499	682.59
5	perfect fourth	467	473.21
6	major third	343	319.13
7	perfect fifth	332	203.97
8	perfect octave	114	118.82
9	minor sixth	5	58.40
10	major sixth	2	19.42
$K = 3.5357$		$C = 0.0344$	
$M = 0.6463$		$R^2 = 0.9845$	
$n = 9$			

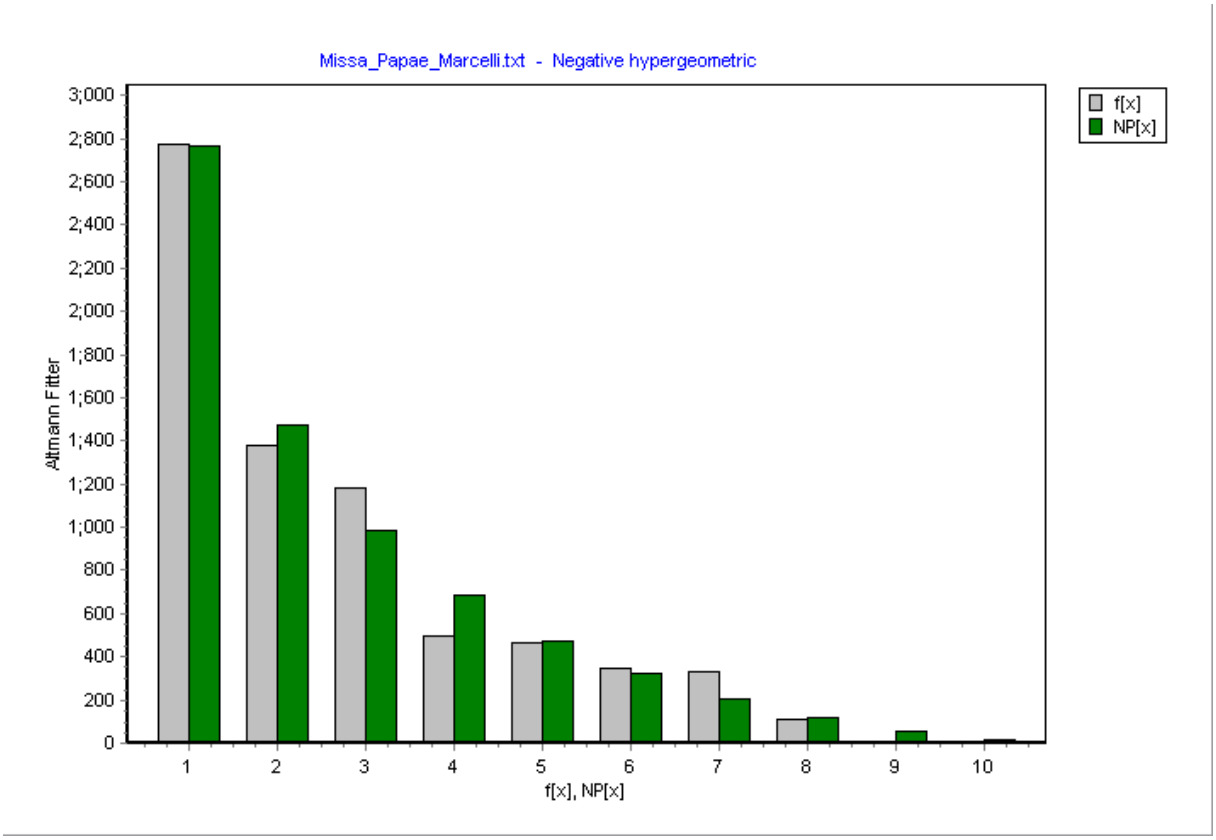


Figure 1. Fitting the negative hypergeometric distribution to data from Table 1

4. Discussion on goodness-of-fit measures

The negative hypergeometric distribution yields the best fit, however, with $C = 0.0344$. In linguistics and in musicology, values $C > 0.02$ are usually considered to be a reason to reject a model (cf. Antić et al. 2005). On the other hand, fit is considered satisfactory if $R^2 \geq 0.9$ (sometimes even $R^2 \geq 0.8$, cf. Eom 2006: 121). We thus have two contradictory values. This is not unusual in statistics – choices of different methods often lead to different results. However, three facts are to be considered here.

First, there is a huge disproportion: the negative hypergeometric distribution with $C = 0.0344$ yields the best fit from among about 250 distributions included in the Altmann-Fitter, i.e., none of them satisfies the "traditional" limit 0.02. On the other hand, R^2 is greater than 0.95 for at least 40 of them (not to speak about the limit 0.9 or even 0.8 – with a bit of exaggeration one could say that it is almost difficult to find a distribution with a worse determination coefficient). We remind that parameter values were optimised with respect to *C*, which means that even higher values of R^2 could be expected if an optimization aimed at the determination coefficient.

Second, R^2 is used mostly to evaluate goodness-of-fit for continuous models. If the determination coefficient is really so much more "tolerant" as it seems from our data, continuous models are allowed to "live comfortably", while discrete ones are forced to "struggle for survival". In such a situation one cannot wonder that relatively simple continuous models yield a good fit (e.g., the data from Table 1 can be fitted by function $y = 2772x^b$, in which 2772 is the first frequency and it thus has only one uninterpreted parameter, obtaining $R^2 = 0.9578$) even for the data for which a discrete distribution with $C \leq 0.02$ is (almost) impossible to find.

Third, rejection rules both for *C* and R^2 are not theoretically derived, but only suggested as rules of thumb, hence they could be adjusted to yield at least roughly comparable results. We, however, limit ourselves to one data set here. Hence, it can be said that we point out a problem without providing at least an empirically substantiated solution. A theoretical solution of the problem would be ideal; if it is not achieved, an extensive study on many data sets from different areas is necessary to establish at least tentative new rules of thumb for *C* and R^2.

Acknowledgement

J. Mačutek was supported by research grant VEGA 1/0077/09.

References

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), Quantitative Linguistics. An International Handbook: 646-658. Berlin / New York: de Gruyter.

Antić, G., Grzybek, P., Stadlober, E. (2005). Mathematical aspects and modifications of Fuck's generalized Poisson distribution. In: Köhler, R., Altmann, G., Piotrowski. R.G. (eds.), *Quantitative Linguistics. An International Handbook: 158–180.* Berlin / New York: de Gruyter.

Došeková, Z. (2006). A contribution to the theory of self-regulation in music: Self-regulation of the note pitch. *Journal of Applied Mathematics, Statistics and Informatics 2, 43-55.*

Eom, J. (2006). *Rhythmus im Akzent. Zur Modellierung der Akzentverteilung als einer Grundlage des Sprachrhythmus im Russischen.* München: Otto Sagner.

Hofstetter, F.T. (1979). The nationalist fingerprint in nineteenth-century Romantic chamber music. *Computers and the Humanities 13, 105-119.*

Huron, D. (1990a). Crescendo/Diminuendo asymmetries in Beethoven's piano sonatas. *Music Perception 7, 395-402.*

Huron, D. (1990b). Increment/Decrement asymmetries in polyphonic sonorities. *Music Perception 7, 385-394.*

Köhler, R.; Martináková-Rendeková, Z. (1998). A systems theoretical approach to language and music. In: Altmann, G., Koch, W.A. (eds.), *Systems. New Paradigms for Human Sciences: 514-546.* Berlin / New York: de Gruyter.

Lindley, M., Campbell, M., Greated, C. (2001). Interval. In: Sadie, S., Tyrrell, J. (eds.), *The New Grove Dictionary of Music and Musicians, Vol. 12: 500-502.* New York: Oxford University Press.

Martináková-Rendeková, Z. (2000). Systems Theoretical Modelling in Musicology. In: Mastorakis, N.E. (ed.), *Mathematics and Computers in Modern Science. Acoustics and Music, Biology and Chemistry, Business and Economics: 122-127*. Athens: WSEAS Press.

Martináková-Rendeková, Z. (2007). Different parameters of negative hypergeometric distribution as a discriminating feature for musical or composer's style. In: Le, M.H., Demiralp, M., Mladenov, V., Bojkovic, Z. (eds.), *Proceedings of the 7th WSEAS International Conference on Systems Theory and Scientific Computation:217-222.* WSEAS Press.

Martináková-Rendeková, Z. (2008). Regularities in Musical Texts Resulted from Rank-Frequency Distribution of Pitch. In: Mastorakis, N.E. , Demiralp, M., Mladenov, V., Bojkovic, Z. (eds.), *New Aspects of Systems Theory and Scientific Computation: 24-30.* Stevens Point (WI): WSEAS Press.

Nettheim, N. (1997). A bibliography of statistical applications in musicology. *Musicology Australia 20, 94-106.*

Pont, G. (1990). Geography and human song. *New Scientist (20 January), 38-41*.

Suchoff, B. (1970). Computer-oriented comparative musicology. In: Lincoln, H.B. (ed.), *The Computer and Music: 193-206*. Ithaca: Cornell University Press.

Vos, P.G., Pasveer, D. (2002). Goodness ratings of melodic openings and closures. *Perception & Psychophysics 34, 631-639.*

Vos, P.G., Troost, J.M. (1989). Ascending and descending melodic intervals: Statistical findings and their perceptual relevance. *Music Perception 6, 383-396.*

Watt, H.J. (1924). Functions of the size of interval in the songs of Schubert and of the Chippewa and Sioux Indians. *British Journal of Psychology 14, 370-386.*

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

Software

Altmann-Fitter (1997). Lüdenscheid: RAM-Verlag.

Glottometrics 21, 2011, 65-72

A naïve conception of the uncertainty principle in the multiparametric attribution of texts

Andrij Rovenchak

Abstract. An approach, which allows taking into consideration the dependencies of text parameters on text length, is suggested as a supplementary tool in text attribution. The proposed model is tested for Ukrainian texts of several genres, namely, sermons and scientific papers in humanities and natural sciences.

Key words: uncertainty principle; text attribution; dependence on text length.

1. Introduction: Uncertainty in physics

In 1927, Heisenberg suggested a fundamental correlation in quantum mechanics now known as the uncertainty principle. It is given by a simple inequality:

$$\langle(\Delta q)^2\rangle \, \langle(\Delta p)^2\rangle \geq \hbar^2/4, \tag{1}$$

where $\hbar$ is a certain fundamental physical constant known as Planck's constant. In this form, the inequality first appeared in the paper by Kennard (1927). It states the relation between the so-called uncertainties of coordinate $\Delta q = q - \langle q\rangle$ and momentum $\Delta p = p - \langle p\rangle$, where q and p denote results of measurements for the respective quantities and $\langle\ldots\rangle$ is the mean value.

The most common (and simplified) explanation is as follows: the more precise is the result of measurement of a coordinate the less precise is that for the momentum of a particle and vice versa. In the limiting case of an exact measurement of, say, coordinate ($\Delta q = 0$) the momentum is completely undetermined ($\Delta p = \infty$ in order to ensure a non-zero product).

This very explanation, yet made even more naïve, induced me to suggest the approach as described below.

2. Drawing parallels

The issue of sample size in the problem of text attribution (more specifically, author attribution) was addressed in several papers, cf. Eder (2010), Koppel et al. (2011); Luyckx & Daelemans (2011). Nevertheless, the length of text remains the quantity, the dependencies on which are often overlooked in studies of texts. It is clear that the reliability of any calculated parameter is lower for shorter texts, and thus it must be taken into consideration when making, e.g., genre or author attribution. It becomes thus possible to approach the length of text in a way similar to one of the variables in Eq. (1) and treat the second variable as one of the parameters characterizing texts.

Let the ith text be characterized by two parameters (x_i, y_i), the generalization for more parameters is straightforward, as it is also possible to consider a one-parametric problem within this approach.

Suppose the existence of some 'ideal' values for the above parameters, x_0 and y_0 being the centers of genre domain.

Suppose further a naïve analogue of the uncertainty principle: ***the distance to the domain center can be larger for shorter texts and vice versa***. The mathematical formulation is as follows:

(2) $(x_i - x_0)^2 f(N_i) \le a^2$,

(3) $(y_i - y_0)^2 f(N_i) \le b^2$,

where N_i is the length of the ith text; a, b are some parameters (defining a genre). Note also a different inequality signs in (2) and (3) compared to (1), which logically follows from the behavior of $f(n)$ as explained below.

Some general properties of the function $f(n)$ are:

- $f(n \to \infty) \to \infty$ (so that the longer the text is, the closer it should be located to the center);
- $f(1) = 0$ (a text containing only one word can be wherever with respect to the center);
- $f(n)$ is a monotonous function of its argument;
- the above suggests $f(n) \sim \ln n$ or a similar dependence.

As a sole logarithm gives a very weak dependence on the text length, one can search within the functions of a 'combinatorial' nature, e.g. $f(n) \sim \ln n! \sim n \ln n$.

So, let

(4) $f(N_i) = (N_i \ln N_i)^\gamma$.

For the sake of simplicity, in further analysis the exponent γ is put equal to unity.

As we know neither a, b nor x_0, y_0, these parameters can be estimated for a given sample as follows:

$$(x_i - x_0)^2 \le \frac{a^2}{f(N_i)}$$

(5) $$\left\langle (x_i - x_0)^2 \right\rangle \le a^2 \left\langle \frac{1}{f(N_i)} \right\rangle, \qquad x_0 = \langle x_i \rangle$$

connecting thus a^2 to the variance of x_i,

(6) $$\sigma_x^2 = \left\langle (x_i - x_0)^2 \right\rangle$$

(7) $$a^2 = \sigma_x^2 \left\langle \frac{1}{f(N_i)} \right\rangle^{-1},$$

and similarly for b and y_0.

Note that in principle the 'real' values (if they exist) can be significantly different from the calculated ones if the sample size is rather small.

The operation of averaging can be done, generally speaking, with some weighting function, w_i. It seems quite natural to use

(8) $w_i = N_i$,

meaning that a text of length *kN* contributes as *k* texts of length *N*. Thus,

$$(9) \qquad \left\langle (\ldots) \right\rangle = \frac{1}{\sum_i w_i} \sum_i (\ldots) w_i.$$

3. Results and discussion

Empirically, the genre domain can thus be defined by an ellipse:

$$(10) \qquad \frac{(x-x_0)^2}{2a^2/f(N_0)} + \frac{(y-y_0)^2}{2b^2/f(N_0)} \leq 1,$$

where N_0 is some characteristic text length (for instance, $N_0 = \bar{N} / 2$, where $\bar{N}$ is the mean text length over the sample, as used in this work).

The analysis was made for texts collected within a project on Ukrainian text databank (Kelih et al. 2009). The parameters considered in the attribution are the dispersion quotient *d* and the fraction of four-syllabic words p_4, cf. Kelih et al. (2005), Buk et al. (2010). The respective quantities are defined as follows:

$$d = \frac{m_2}{m_1 - 1},$$

where m_1 is the mean word length in syllables:

$$m_1 = \frac{1}{N} \sum_i x_i,$$

and m_2 is the second central moment:

$$m_2 = \frac{1}{N} \sum_i (x_i - m_1)^2.$$

If the number of four-syllabic words in a text is N_4, the parameter p_4 equals:

$$p_4 = N_4 / N.$$

In Fig. 1 the data for scientific papers in humanities and natural sciences are shown. One can see that the model is correct as the texts of different genres generally fall into expected elliptic domains. Interestingly, those physical texts which appear in the humanities ellipse are in fact personalia, not scientific papers proper.

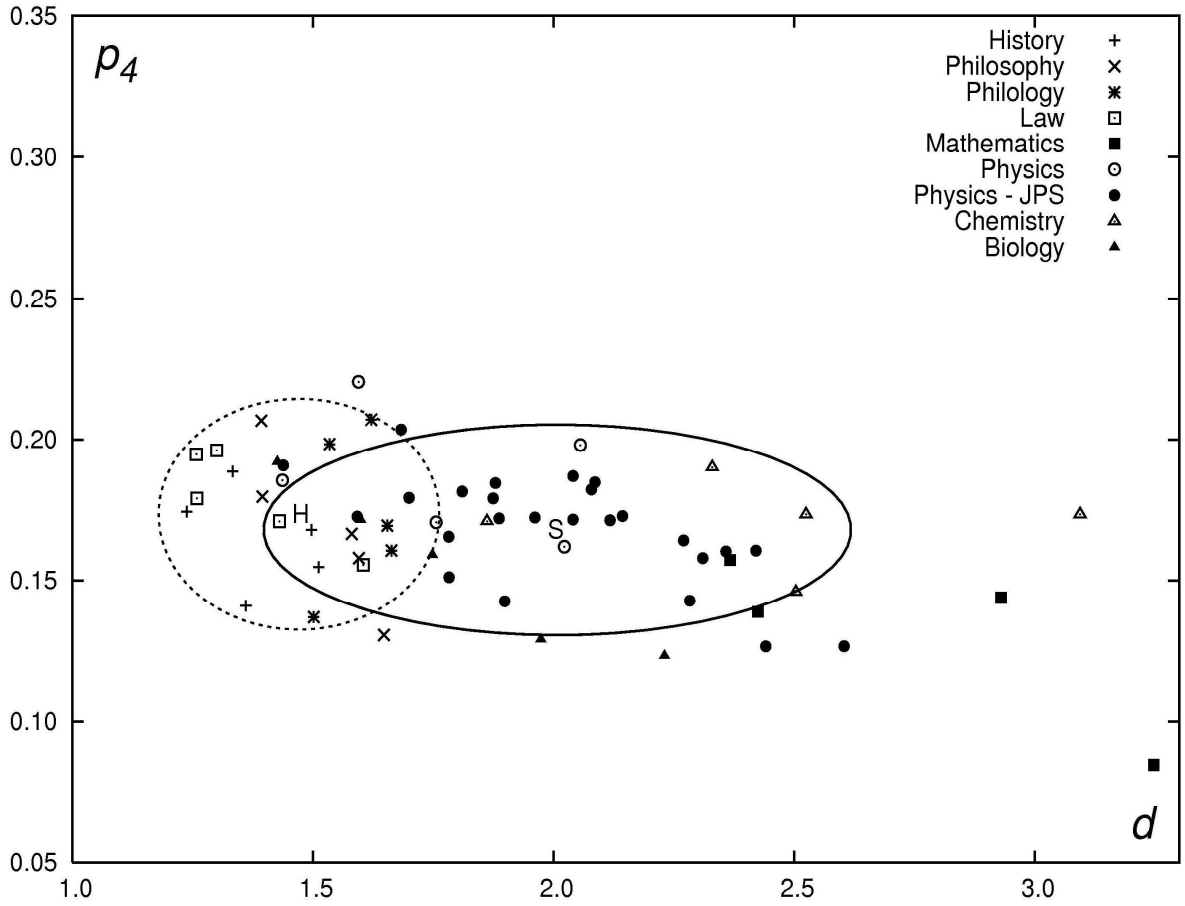


Figure 1: Genre domains are shown for humanities and for the natural sciences, including physical texts from the *Journal of Physical Studies* (JPS), $f(N_i) = N_i \ln N_i$. Centers of the ellipses are marked with H and S respectively, with the following coordinates: H (1.471; 0.1734), S (2.008; 0.1679).

For every *i*th text, the value of the following function was calculated showing if the text falls within the 'uncertainty' domain:

$$E_i = \frac{(x_i - x_0)^2 f(N_i)}{2a^2} + \frac{(y_i - y_0)^2 f(N_i)}{2b^2} - 1. \tag{11}$$

Negative values correspond to points inside the ellipse.

In Fig. 2 the values of the function E_i are given. One can see that the texts mostly obey the proposed model as about 2/3 points of each genre are below the zero line. A more thorough analysis with more texts and genres is required to define an appropriate form of the function $f(n)$ and the values of parameters. The proposed approach can be used as an additional controlling mechanism in text attribution.

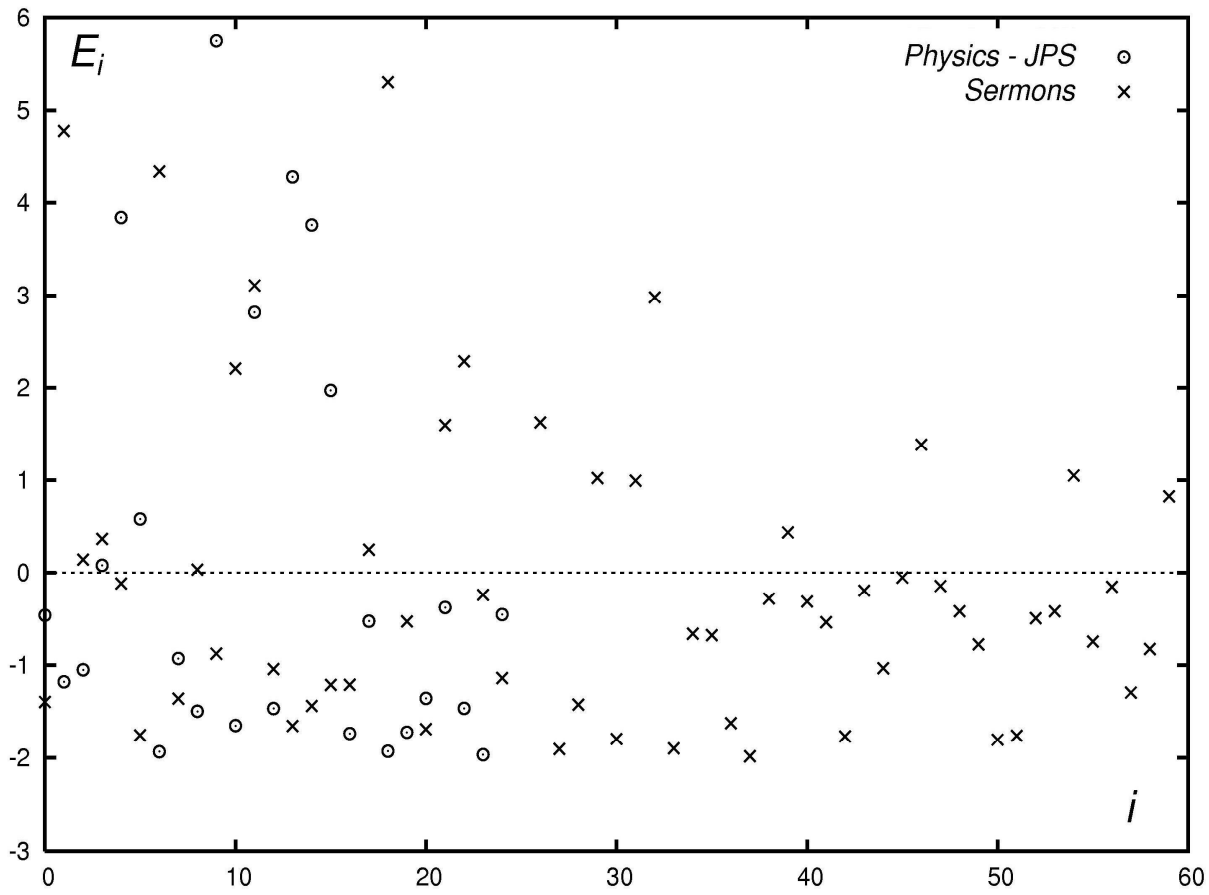


Figure 2: Values of the function E_i for physical texts from the *Journal of Physical Studies* and sermons of different denominations, $f(N_i) = N_i \ln N_i$.

References

Buk, S.; Humenchyk, O.; Mal'tseva, L.; Rovenchak A. (2010). Word-length-related parameters of text genres in the Ukrainian language. A pilot study. In: Grzybek, P.; Kelih, E.; Mačutek, J. (eds.), *Text and Language: Structures – Functions – Interrelations. Quantitative perspectives: 13–19*. Wien: Praesens.

Eder, M. (2010). Does size matter? Autorship attribution, short samples, big problem. In: *Digital humanities 2010: Conference abstracts: 132–135*. London: Office for Humanities Communication.

Heisenberg, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik 43(3–4), 172–198*.

Kelih, E.; Antić, G.; Grzybek, P.; Stadlober, E. (2005). Classification of author and/or genre? The impact of word length. In: Weihs, C.; Gaul, W. (eds.), *Classification – The Ubiquitous Challenge: 498–505*. Heidelberg: Springer.

Kelih, E.; Buk, S.; Grzybek, P.; Rovenchak, A. (2009). Project description: Designing and constructing a typologically balanced Ukrainian text data-base. In: Kelih, E.; Levickij, V.; Altmann, G. (eds.), *Methods of Text Analysis: 125-132*. Chernivtsi: ČNU.

Kennard, E.Y. (1927). Zur Quantenmechanik einfacher Bewegungstypen, *Zeitschrift für Physik 44(4–5), 326–352*.

Koppel, M.; Schler, J.; Argamon, Sh. (2011). Authorship attribution in the wild. *Language Resources and Evaluation 45(1), 83–94.*

Luyckx, K.; Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing 26(1), 35–55.*

Appendix

Table 1: Raw data used in calculations for the papers from the *Journal of Physical Studies*.

i	N	m_1	m_2	d	p_4
1.	715	2.5664	2.9561	1.8871	0.1720
2.	1003	2.5623	2.7845	1.7823	0.1655
3.	1212	2.4851	3.0980	2.0860	0.1848
4.	1219	2.6177	3.3009	2.0405	0.1870
5.	1259	2.5258	3.1135	2.0406	0.1716
6.	1279	2.4848	3.3710	2.2704	0.1642
7.	1459	2.5696	2.9799	1.8986	0.1426
8.	1529	2.6887	3.0110	1.7830	0.1511
9.	1538	2.5709	3.0811	1.9614	0.1723
10.	1541	2.8968	2.7285	1.4385	0.1908
11.	1575	2.5937	3.8577	2.4207	0.1606
12.	1614	2.6908	3.5147	2.0787	0.1822
13.	1698	2.2256	3.1912	2.6038	0.1266
14.	1724	2.2686	3.0967	2.4411	0.1265
15.	1859	2.5729	2.9558	1.8792	0.1845
16.	1896	2.4684	2.4716	1.6832	0.2036
17.	2058	2.7847	3.0338	1.6999	0.1793
18.	2242	2.5027	3.4712	2.3100	0.1579
19.	2401	2.6597	3.5573	2.1433	0.1728
20.	2691	2.5292	2.8668	1.8747	0.1791
21.	2895	2.5358	3.2518	2.1174	0.1713
22.	3069	2.3109	3.0914	2.3583	0.1603
23.	3781	2.4742	3.3660	2.2832	0.1428
24.	3790	2.5412	2.7903	1.8105	0.1815
25.	5059	2.6887	2.6888	1.5923	0.1726

Table 2: Raw data used in calculations for sermons.

i	N	m_1	m_2	d	p_4
1.	210	2.2952	1.7985	1.3886	0.1333
2.	268	2.3843	1.6023	1.1575	0.1082
3.	285	2.0877	1.9677	1.8091	0.0842
4.	442	2.1176	1.6106	1.4411	0.0792
5.	450	2.1978	1.3942	1.1640	0.1156
6.	460	2.2717	1.6414	1.2907	0.0978
7.	515	2.1359	1.5738	1.3854	0.0835
8.	516	2.1977	1.5191	1.2683	0.1221
9.	522	2.3238	1.7822	1.3463	0.1398
10.	536	2.2332	1.2721	1.0315	0.1082

i	N	m_1	m_2	d	p_4
11.	574	2.3066	1.6760	1.2827	0.1307
12.	584	2.0582	1.5309	1.4466	0.0908
13.	590	2.2356	1.7767	1.4379	0.1017
14.	594	1.9966	1.4512	1.4561	0.1010
15.	618	2.3544	1.6883	1.2466	0.1408
16.	623	2.4077	1.8466	1.3118	0.1557
17.	660	2.2439	1.7814	1.4321	0.1318
18.	662	2.2689	1.5289	1.2049	0.1163
19.	676	2.1361	1.3602	1.1972	0.0976
20.	686	2.1764	1.6292	1.3850	0.1356
21.	696	2.1480	1.5629	1.3614	0.1466
22.	701	2.2767	1.7265	1.3523	0.1455
23.	722	2.2535	1.5189	1.2117	0.1371
24.	765	2.2118	1.6650	1.3740	0.1281
25.	776	2.1211	1.5291	1.3639	0.1044
26.	788	2.0990	1.4648	1.3329	0.1015
27.	792	2.3371	1.8851	1.4098	0.1414
28.	812	2.1564	1.4743	1.2749	0.1121
29.	860	2.2116	1.5250	1.2586	0.1384
30.	908	2.0198	1.4005	1.3733	0.0881
31.	916	2.1550	1.6026	1.3875	0.1124
32.	919	2.1632	1.3531	1.1633	0.0881
33.	925	2.2151	1.6045	1.3205	0.0930
34.	960	2.2406	1.7869	1.4403	0.1073
35.	965	2.1554	1.5033	1.3011	0.1088
36.	966	2.0414	1.4103	1.3542	0.0890
37.	996	2.1687	1.3932	1.1922	0.0823
38.	1006	2.0775	1.3578	1.2601	0.0825
39.	1016	2.2274	1.5418	1.2562	0.1319
40.	1020	2.1216	1.4911	1.3295	0.0892
41.	1075	2.1191	1.3719	1.2259	0.0930
42.	1094	2.3044	1.6596	1.2723	0.1051
43.	1110	2.1883	1.5006	1.2628	0.0964
44.	1137	2.2005	1.6977	1.4141	0.1187
45.	1144	2.0087	1.3793	1.3673	0.0726
46.	1218	2.0764	1.5746	1.4629	0.1092
47.	1240	2.1903	1.4444	1.2135	0.1065
48.	1275	2.4753	1.8039	1.2227	0.1451
49.	1281	2.1374	1.4316	1.2586	0.1023
50.	1282	2.2933	1.7486	1.3521	0.1443
51.	1330	2.1173	1.4855	1.3295	0.1053
52.	1407	2.2516	1.6268	1.2998	0.1158
53.	1422	2.1217	1.5288	1.3630	0.0858
54.	1461	2.5640	2.2007	1.4071	0.1554
55.	1483	2.0351	1.3717	1.3252	0.1032
56.	1503	2.0852	1.5470	1.4256	0.1018
57.	1571	2.2903	1.6115	1.2490	0.1120

i	N	m_1	m_2	d	p_4
58.	1727	2.1118	1.3836	1.2445	0.0944
59.	1828	2.3950	1.6131	1.1564	0.1258
60.	2125	2.1402	1.3149	1.1532	0.1031

Glottometrics 21, 2011, 73-98

Fractal analysis of Poe's Raven

Jan Andres[1,2], *Martina Benešová*[3]

Abstract. The fractal and cluster analyses of Poe's Raven, its one German and eighteen Czech translations are carried out. Following step by step the methodology developed in Andres et al. (s.d.), the appropriate segmentation of the poem on each linguistic level is investigated from the fractal point of view. The related language fractals are characterized by a degree of semanticity. Some of their models are also visualized by means of patterns reminding line codes. The (non)suitability of a quantitative exploration of poetical texts is discussed in a comparative way.

Keywords: Fractal analysis, poetical texts, language fractals, degree of semanticity, visualized models, clusters, dendrograms, segmentation.

1. Introduction

Concerning Poe's Raven and its translations, it was mentioned with no specific explanation in Dvořáková (2009) that "we prospect for the deeper meaning somewhere `under cover`, in the unconscious of the poem: in the syllables which have to be omitted from three-syllable words so that the metre is kept when reciting the poem; in the *fractal spreading of the word Lenore* (which disappears from the translations); in the historical context which is meant to become `the voice of conscience` of the poem".

As already pointed out in Andres (2010), by fractals in poetry, one usually means semantic recursions. Two concrete examples were mentioned there due to Vladimír Holan and Wallace Stevens. Let us note that a recent Hungarian poet Ferencz Gyõzõ (born in 1954) has also, but this time tendentiously, written such a poem called "Fractal Consciousness". As observed in Shannon (1993), the terms in natural languages may be seemingly used to describe states of affairs at different levels of resolution.

Another sort of fractality was detected in Becker & Flaxer (2008) in the sense that the organization of text and neuronal activity can be linked by outlining the analogy between the hierarchical structure of neuronal electrical activity and the hierarchical structure of text. The language as a product of the brain was also considered, but in a heuristic way in Henry (1995), where the organizational property of grammar was demonstrated to be fractal in nature.

In our paper, none of these approaches will be taken into account. We try to follow and work out what was initiated by L. Hřebíček (cf. e.g. Hřebíček, 1997, 2002, 2007a[4]) and

[1] Dept. of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, 17.listopadu, 77146 Olomouc, Czech Republic. E-mail: jan.andres@inf.upol.cz

[2] Supported by the Council of Czech Government (MSM 6198959214)

[3] Dept. of Foreign Languages, The College of Polytechnics Jihlava, Tolstého 16, 58601 Jihava, Czech Republic. E-mail: benesova@vspj.cz

[4] One of the Hřebíček's requirements is to rewrite the text samples to become a reference text, e.g. in case of the phrase "this paper" to substitute the word "this" by the reference phrase "Andres & Benešová's" phrase (see Hřebíček 2007b). Yet, in case of our text samples, it proves if not impossible, then inefficient, as will be explained later.

developed in our works (cf. Andres 2009; 2010; Andres & Rypka s.d.; Andres et al. s.d.). What we try to do is to analyze text samples by means of the *Menzerath-Altmann law* (MAL) (cf. e.g. Altmann 1980; Altmann et al. 1989). Methodologically, we follow the algorithm introduced in Andres et al. (s.d.), as will be applied step by step below. Furthermore, what we try to highlight in this analysis is comparing the text samples with the same semantic background.

The main aim of this experiment is to execute a bunch of different analyses for the purpose of figuring out the correct and efficient way of text segmentation on different levels, of testing the MAL on different text samples in different languages, yet, with the same semantic background, and of testing text samples for fractality.

The paper is divided into five chapters. Chapter 1 is an introductory part. Chapter 2 launches the first seven steps of the quantitative analysis algorithm described more in detail in Andres et al.(s.d), including the three approaches to setting up units. Initial setting up units was not, with some exceptions (cf. e.g. Köhler 1997), commented on before and proved to be one of the biggest obstacles of this experiment. There is also a lot of space for further research which will be offered at the end of this paper. Chapter 3 describes testing fractality of the text samples in the spotlight. The text fractality is based in interpreting the reciprocal value of the arithmetic mean of parameters gained by the quantitative analysis as the dimension of the mathematical fractal associated to the particular language fractal. Such a fractal dimension will be called the *degree of semanticity*. In Chapter 4, we concluded our experiment with the newly introduced cluster analysis of the text samples. And last but not least, Chapter 5 introduces the summary and interpretation of the results of our experiment and opens the horizons for the future research.

Remarkable results of a similar experiment have already been gained by B. Mandelbrot (2000) and A. Eftekhari (2006). They come out from the same interpretation, however, which is not based on MAL, but on the Mandelbrot-Zipf law. Iterated function systems were applied in Fernau & Staiger (2001) and Gutiérrez et al. (2003) but, unlike by ourselves, to formal languages. Various kinds of some further approaches and methods of quantitative analysis can be also found in Wildgen (2011) and Wimmer et al. (2003).

2. Segmentation and basic quantitative analysis

The whole procedure of analyzing the text in this quantitative way consists of the following necessary steps, as was described in a more detailed way in (Andres et al.).

Step 1, for more details cf. Andres et al.(s.d.): *Choosing a sample text.* In the year of 1846, Edgar Allan Poe added the essay *The Philosophy of Composition* to his poem *The Raven.* In this way Poe provided the reader with a manual explaining how it was composed. Among other information, Poe mentioned the reason for the choice of its length, which is purely quantitative; "…the extent of a poem may be made to bear mathematical relation to its merit – in other words, to the excitement or elevation – again, in other words, to the degree of the true poetical effect which it is capable of inducing; for it is clear that the brevity must be in direct ratio of the intensity of the intended effect – this, with one proviso – that a certain degree of duration is absolutely requisite for the production of any effect at all" (Jařab et al., 1985). The aim of this experiment was to back this prerequisite of Poe's, and last but not least, to test the justifiability of using the MAL formula. Poe's *Raven* stepped in the Czech translation tradition significantly. Within approximately one hundred years since its composition, about twenty translations into the Czech language emerged. Some of them have become, together with the original English texts and its German mutation, the subject of this analysis. The poem itself is generally regarded one of the hard nuts to crack for translators.

The original text availability and a huge amount of its translations, which are more or less obliged to follow the rules for the form as well as for the content and which can be found in a huge amount of languages differing in their structure, are a strong motivation for a quantitative analysis.

The seemingly unambiguous choice of Poe's poem thanks to its "mathematicality" brings, yet, numerous obstacles. Poe's *Raven* is a poetic text, therefore, its structure does not follow linguistic rules strictly, but in reverse, it violates them often. The related problem with setting the basic linguistic units is to be discussed later. Nevertheless, the particular fact that it is a poetic text and on top of that, the text bound by certain criteria fixed in advance seems to be a disadvantage for translators, yet an advantage for those who try to analyze it in a quantitative way.

As for the problem whether the poem *Raven* is a population or a sample, it is obvious that we would like to apply our quantitative analysis on the whole language, meaning on the whole population, which is but an unsolvable problem, see Orlov et al. (1982). Let us set down both the original text and any of its translations a sample. What is then a population for us? Here, the consideration and setting the initial conditions are important. It is unconceivable to consider the whole language. We cannot take into account all the author texts by Edgar Allan Poe either because we analyze also the translations however they respect Poe's criteria. Therefore, let us appoint the text of the poem *Raven* in all its language mutations to be the population for our experiment.

In this experiment, twenty texts of Poe's *Raven* in three languages, one original English Poe's text, (cf. Poe 1997), the translation into the German language by Otto F. Babler (Poe 1931) and eighteen translations[5] into the Czech language by different authors (cf. Poe 1985; Poe - Jacko 2008; Poe – Petlan 2008), came under the spotlight of fractal analysis.

Step 2. For more details cf. Andres et al. (s.d): *Determination of the sample units and reasoning it.* Units have to be defined unambiguously; the notion of the unit has to be in accordance with common linguistic definitions, and if not, it has to be carefully justified; the determination of units has to be rigidly kept throughout the whole experiment; and each sample member has to be taken into account, yet not calculated twice (cf. Těšitelová, 1992; Andres et al.).

For this experiment of ours we explore the three following binarisms:

1 level $i = 1$: semantic constructs (their length in sentences/clauses) – sentences/clauses (their average length in words);
2 level $i = 2$: sentences/clauses (their length in words) – words (their average length in syllables);
3 level $i = 3$: words (their length in syllables) – syllables (their average length in phonemes) (cf. Hřebíček 1997, 2002).

The process of the unit determination is very complex and needs considering many aspects (cf. Altmann 1996). We present two approaches in this article basically, but we propose others for following experiments. The third approach is added in two special cases. What is, however, always crucial is that once we use a particular definition of a particular unit, we have to keep it throughout the whole experiment of ours. Setting the particular units is described more in detail in Andres et al.(s.d.).

It is not always easy to follow all the criteria for setting units, nevertheless defining words, whose length was to be calculated in syllables, turned out to be the most complicated. Approach I has been applied on all of the texts in all language mutations. Words were defined with the highlight on their graphical properties mostly as units existing "in between two

[5] The author of one of the translations into the Czech language from the year 1930 is again Otto F. Babler, (cf. Poe 1985). It offered us the chance to study apart from various translations of one text also the translations of one text into two different languages by one author.

succeeding gaps" (cf. Andres et al. s.d.). This method is the seemingly least demanding one for the researcher; on the other hand it hardly reflects the semanticity hidden in the background and does not respect the grammatical and morphological rules varying from different language type to another. Besides, it can be used only in languages in which gaps are made. Yet, this approach shall simply demonstrate the studied method of quantitative linguistics analysis, and its effectiveness is commented on at the close of this paper.

Approach III deals with the word first as with the compound (analytic) word form. It can be defined as a specific link of synthetic word forms which functions as a complex form of a full-meaning word. Only one of the components is the bearer of the main lexical meaning, on the other hand, the other component or components is/are the bearers of the grammatical meaning. Words having the function of grammatical modifiers of words, regardless on their orthography, were counted as parts of the respective word forms. Thus, the preposition modifying the head noun is counted as one single unit together with the following word form whether it is its head noun or not, as was already discussed further in detail in Andres et al. (s.d.). The necessity of composing a preposition and its head noun is further discussed in Faltýnek (2011). This approach was applied just in three representative cases (on the original Poe's text and on the Babler's translation into the German and Czech language) as a different point of view on the analysis and as a starting point for the next research.

Approach II goes hand in hand with approach III, the only difference is that definite and indefinite articles in English and German were understood and defined a part of the word unit. This scheme was supposed to be applied on the same three sample texts as approach III, however in case of the Czech translation it does not make sense for the absence of articles in the Czech language.

Step 3. For more details cf. Andres et al. (s.d.): *Verifying the representativeness of the sample length.* The test of representativeness highlights the necessity to choose the sample ample in length so that when we make the sample longer (and approach this way the population), the sample does not change its fundamental properties significantly, for more details cf. Kubáček (1994). For the illustration, we will present the results of the test for the original Poe's text analysed by approach I. In the sample text, there are 1,060 lexemes in total, out of which 412 are different. Such a sample text is stable and representative if we set up the standard deviation of $r = 0.0012$, i.e. 0.12%. For the formula and the algorithm of calculation see Benešová (2011), Kubáček (1994).

The problem of choosing a representative sample is one of the most vital, yet in our analysis the initial motivation for choosing the sample text was a unique chance to analyse different samples in different languages with the same semantic background.

Step 4. For more details cf. Andres et al. (s.d): *Quantifying the text* so that it is possible to extract the *constructs with the length* x_i, their frequency z_i and *constituents with the length* y_i on each linguistic level assigned by the indices $i = 1, 2, 3$ from it[6].

[6] To illustrate the algorithm e.g. for the original Poe's text 1a I and e.g. for the level $i = 1$, the output table gained as a result of quantifying the text is as follows:

x	z	y	x	z	y	x	z	y
1	250	11.1680	8	7	11.0357	18	1	10.1667
2	72	11.0417	10	3	11.9667	20	1	14.9000
3	27	10.2346	11	2	12.7727	24	1	11.8333
4	13	11.0000	14	2	11.1071	38	1	9.89474
5	11	10.8546	15	1	12.8667	41	1	7.17073
6	6	11.5000	17	1	10.4118	56	1	10.1071
7	11	10.5974						

Step 5. For more details cf. Andres et al.(s.d): *Calculating the parameters* A_i, b_i, c_i, i = 1, 2, 3 for both the truncated and complete formula of the Menzerath-Altmann law[7] by means of statistical and numerical methods. After quantifying all the sample texts as stated in step 4 we processed the output in four ways.

a Calculating the parameters A_i, b_i, i = 1, 2, 3 for the truncated formula of the MAL by means of statistical methods (in graphs indicated as ◯).
b Calculating the parameters A_i, b_i, i = 1, 2, 3 for the truncated formula of the MAL by means numerical methods (in graphs indicated as □).
c Calculating the parameters A_i, b_i, c_i, i = 1, 2, 3 for the complete formula of the MAL by means of statistical methods (in graphs indicated as ◇).
d Calculating the parameters A_i, b_i, c_i, i = 1, 2, 3 for the complete formula of the MAL by means of numerical methods (in graphs indicated as △).

The most important of the parameters for our analysis is b_i, i = 1, 2, 3 because of its correlation to the dimension of the associated *mathematical fractal*[8] (cf. Hřebíček, 1997, 2002; Andres et al. s.d.). The *language fractal* can be defined as such a linguistic object which satisfies the Menzerath-Altmann's law with all the parameters b_i on each of its examined linguistic levels $i = 1, 2, 3$ positive (cf. Andres et al. s.d.). For this reason we present here this parameter only for the sample texts processed by each of the three methods[9].

Ad approach I:

1 Poe
   a. $b_1 = 0.02829237$, $b_2 = 0.04912137$, $b_3 = 0.0685204$
   b. $b_1 = 0.01861$, $b_2 = 0.08453$, $b_3 = 0.06797$
   c. ($b_1 = -0.08238833$, $b_2 = 0.1950147$, $b_3 = 0.1047829$)
   d. ($b_1 = -0.095142$, $b_2 = 0.26722$, $b_3 = 0.08967$)

---

[7] The truncated formula of the Menzerath-Altmann law (MAL) is $y = A.x^{-b}$, whereas the complete formula of MAL is stated as $y = A.x^{-b}e^{cx}$, cf. Altmann (1980), Andres et al. (s.d.).

In order to demonstrate the calculation of the parameters of MAL via the regression method (an alternative approach is numerical), we transform MAL in a logarithmic way (for the sake of simplicity, let it be the truncated formula of MAL); $\ln y = \ln A - b.\ln x$, where there the substitution $y' = \ln y$, $x' = -\ln x$, $a = \ln A$ can be applied. We get, in this way, the linear equation $y' = a + bx'$. By means of the method of least squares, we can calculate the parameters a and b as follows:

$$b' = -\frac{n\sum_{j=1}^{n} x'_j y'_j - \sum_{j=1}^{n} x'_j \sum_{j=1}^{n} y'_j}{n\sum_{j=1}^{n} x'^2_j - \left(\sum_{j=1}^{n} x'_j\right)^2}, \quad a' = \frac{\sum_{j=1}^{n} x'^2_j \sum_{j=1}^{n} y'_j - \sum_{j=1}^{n} x'_j y'_j \sum_{j=1}^{n} x'_j}{n\sum_{j=1}^{n} x'^2_j - \left(\sum_{j=1}^{n} x'_j\right)^2} .$$

For the particular text sample 1a I, we so get:

$$b' = -\frac{19(104.691) - 43.92607(45.52832)}{19(121.5524) - 1929.5} \doteq 0.02829,$$

$$a' = \frac{121.5524(45.52022) - 104.681(42.92607)}{19(121.5542) - 1929.5} \doteq 2.46154, \quad A' = e^{2.46154} = 111.72398$$

Because of the lengthy and demanding calculating of the parameters, statistical software is recommended. In our experiment, R 2.10.0 software was used for all the methods a, b, c, d.

[8] Let the dimension of the associated mathematical fractal be defined as the reciprocal value of the arithmetic mean of all the parameters b_i, i = 1,2,3, cf. Andres (2009), Andres et al. (s.d.).

[9] In the brackets, we present the outputs where at least one of the parameters b_i, i = 1,2,3 is negative (these indicated in italics), thus such a sample text cannot be regarded a language fractal.

2 Babler – German
   a. (b_1 = 0.04469562, b_2 = *-0 .003084549*, b_3 = 0.2741698)
   b. (b_1 = 0.0353, b_2 = *-0.002524*, b_3 = 0.2839)
   c. b_1 = 0.1321585, b_2 = 0.1568826, b_3 = 0.3793558
   d. b_1 = 0.111254, b_2 = 0.1728, b_3 = 0.39072
3 Šembera
   a. b_1 = 0.0185512, b_2 = 0.04661435, b_3 = 0.2434756
   b. b_1 = 0.0102, b_2 = 0.04521, b_3 = 0.2411
   c. (b_1 = *-0.0326087*, b_2 = 0.0758137, b_3 = 0.183899)
   d. (b_1 = *-0.044947*, b_2 = 0.077833, b_3 = 0.1957)
4 Vrchlický
   a. (b_1 = 0.02785711, b_2 = *-0.01016194*, b_3 = 0.1988444)
   b. (b_1 = 0.02811, b_2 = *-0.01119*, b_3 = 0.1966)
   c. b_1 = 0.002215368, b_2 = 0.05174956, b_3 = 0.110384
   d. b_1 = 0.002748, b_2 = 0.05218, b_3 = 0.12683
5 Mužík
   a. (b_1 = *-0.002734985*, b_2 = 0.06864244, b_3 = 0.1325155)
   b. (b_1 = *-0.02487*, b_2 = 0.07662, b_3 = 0.1254)
   c. (b_1 = 0.09891378, b_2 = 0.15536, b_3 = *-0.1248322*)
   d. (b_1 = 0.09682, b_2 = 0.16274, b_3 = *-0.1235*)
6 Lutinov
   a. b_1 = 0.09267924, b_2 = 0.02222886, b_3 = 0.1504526
   b. b_1 = 0.06622, b_2 = 0.02643, b_3 = 0.1462
   c. (b_1 = 0.08755964, b_2 = 0.2005795, b_3 = *-0.03112276*)
   d. (b_1 = 0.02428, b_2 = 0.20572, b_3 = *-0.03559*)
7 Nezval
   a. (b_1 = 0.1772, b_2 = *-0.02306*, b_3 = 0.12116)
   b. (b_1 = 0.1157, b_2 = -0.0252, b_3 = 0.1036)
   c. (b_1 = 0.239787, b_2 = 0.128708, b_3 = *-0.52553*)
   d. (b_1 = 0.175659, b_2 = 0.12008, b_3 = *-0.4916*)
8 Babler – Czech
   a. (b_1 = *-0.01229989*, b_2 = 0.07013482, b_3 = 0.3309882)
   b. (b_1 = *-0.03027*, b_2 = 0.08325, b_3 = 0.3049)
   c. (b_1 = 0.3905951, b_2 = 0.228655, b_3 = *-0.2213671*)
   d. (b_1 = 0.33339, b_2 = 0.26447, b_3 = *-0.1153*)
9 Taufer
   a. (b_1 = 0.1610241, b_2 = *-0.00942236*, b_3 = 0.1290018)
   b. (b_1 = 0.1058, b_2 = *-0.009985*, b_3 = 0.1238)
   c. b_1 = 0.1693824, b_2 = 0.01130945, b_3 = 0.01274076
   d. (b_1 = 0.06152, b_2 = 0.006658, b_3 *cannot be found*[10])
10 Stoklas
   a. b_1 = 0.1013934, b_2 = 0.05913767, b_3 = 0.0733
   b. b_1 = 0.07163, b_2 = 0.06786, b_3 = 0.07232
   c. b_1 = 0.1897575, b_2 = 0.188745, b_3 = 0.08140624
   d. b_1 = 0.17795, b_2 = 0.20283, b_3 = 0.075108
11 Wagnerová
   a. (b_1 = *-0.01852006*, b_2 = 0.1014816, b_3 = 0.08375543)
   b. (b_1 = *-0.02034*, b_2 = 0.1131, b_3 = 0.08221)

[10] It was not possible to calculate this parameter by means of the used software R 2.10.0.

c. $b_1 = 0.005319176$, $b_2 = 0.260446$, $b_3 = 0.06177219$
d. ($b_1 = -0.0001498$, $b_2 = 0.26399$, $b_3 = 0.0476$)

12 Havel
a. $b_1 = 0.09848818$, $b_2 = 0.05905367$, $b_3 = 0.2610285$
b. $b_1 = 0.05242$, $b_2 = 0.06048$, $b_3 = 0.2476$
c. $b_1 = 0.1083344$, $b_2 = 0.1159164$, $b_3 = 0.09005535$
d. $b_1 = 0.105941$, $b_2 = 0.112434$, $b_3 = 0.07996$

13 Čapek
a. $b_1 = 0.0623626$, $b_2 = 0.0330582$, $b_3 = 0.07078767$
b. $b_1 = 0.04679$, $b_2 = 0.03114$, $b_3 = 0.069$
c. ($b_1 = 0.001749692$, $b_2 = 0.1197062$, $b_3 = -0.01621862$)
d. ($b_1 = 0.001656$, $b_2 = 0.11379$, $b_3 = -0.02259$)

14 Resler
a. $b_1 = 0.09714409$, $b_2 = 0.02624885$, $b_3 = 0.0868665$
b. $b_1 = 0.06778$, $b_2 = 0.02856$, $b_3 = 0.084$
c. ($b_1 = 0.2190468$, $b_2 = 0.1446877$, $b_3 = -0.05152063$)
d. ($b_1 = 0.1521$, $b_2 = 0.14568$, $b_3 = -0.06303$)

15 Černý
a. $b_1 = 0.04618615$, $b_2 = 0.08846803$, $b_3 = 0.005346203$
b. ($b_1 = 0.04227$, $b_2 = 0.104$, $b_3 = -0.0006068$)
c. $b_1 = 0.04812482$, $b_2 = 0.2010607$, $b_3 = 0.2348718$
d. $b_1 = 0.028752$, $b_2 = 0.2441$, $b_3 = 0.25199$

16 Slavík
a. $b_1 = 0.09306961$, $b_2 = 0.09046977$, $b_3 = 0.027177$
b. $b_1 = 0.08111$, $b_2 = 0.1345$, $b_3 = 0.02541$
c. ($b_1 = 0.1558316$, $b_2 = 0.2727789$, $b_3 = -0.3087926$)
d. ($b_1 = 0.092506$, $b_2 = 0.36934$, $b_3 = -0.3169$)

17 Kadlec
a. ($b_1 = -0.04349735$, $b_2 = 0.08320275$, $b_3 = 0.1611956$)
b. ($b_1 = -0.05865$, $b_2 = 0.09526$, $b_3 = 0.1519$)
c. ($b_1 = -0.06038785$, $b_2 = 0.1269568$, $b_3 = -0.01063657$)
d. ($b_1 = -0.105269$, $b_2 = 0.1702$, $b_3 = -0.028$)

18 Bejblík
a. $b_1 = 0.05005198$, $b_2 = 0.01087146$, $b_3 = 0.0737228$
b. $b_1 = 0.03781$, $b_2 = 0.01017$, $b_3 = 0.06929$
c. ($b_1 = 0.1179202$, $b_2 = 0.03582004$, $b_3 = -0.1947484$)
d. ($b_1 = 0.097001$, $b_2 = 0.038387$, $b_3 = -0.1904$)

19 Jacko
a. $b_1 = 0.06296325$, $b_2 = 0.05349011$, $b_3 = 0.07177231$
b. $b_1 = 0.04786$, $b_2 = 0.07552$, $b_3 = 0.07078$
c. ($b_1 = -0.02556444$, $b_2 = 0.383266$, $b_3 = 0.02590992$)
d. ($b_1 = -0.03408$, $b_2 = 0.41475$, $b_3 = 0.01842$)

20 Petlan
a. ($b_1 = 0.02114794$, $b_2 = 0.1075671$, $b_3 = -0.002279997$)
b. ($b_1 = 0.01093$, $b_2 = 0.1306$, $b_3 = -0.003033$)
c. $b_1 = 0.1003655$, $b_2 = 0.2763939$, $b_3 = 0.03365033$
d. $b_1 = 0.09735$, $b_2 = 0.3285$, $b_3 = 0.03435$

Ad approach II

1 Poe
   a. (b_1 = *-0.002452624*, b_2 = 0.08604768, b_3 = 0.0685204)
   b. (b_1 = *-0.01255*, b_2 = 0.1192, b_3 = 0.06797)
   c. (b_1 = *-0.09433318*, b_2 = 0.2607647, b_3 = 0.08967)
   d. (b_1 = *-0.118624*, b_2 = 0.3274, b_3 = 0.08967)

2 Babler – German
   a. (b_1 = *-0.009175805*, b_2 = 0.01780423, b_3 = 0.1174008)
   b. (b_1 = *-0.02042*, b_2 = 0.03187, b_3 = 0.1227)
   c. b_1 = 0.008098222, b_2 = 0.1528923, b_3 = 0.2874386
   d. (b_1 = *-0.01935*, b_2 = 0.18678, b_3 = 0.2975)

8 Baber – Czech – using this method for the Czech translation does not make sense

Ad approach III

1 Poe
   a. (b_1 = 0.0322245, b_2 = *-0.01034285*, b_3 = 0.08616824)
   b. b_1 = 0.02651, b_2 = 0.014, b_3 = 0.08465
   c. (b_1 = *-0.02803661*, b_2 = 0.1738866, b_3 = *-0.002206851*)
   d. (b_1 = *-0.034281*, b_2 = 0.22741, b_3 = *-0.01568*)

2 Babler – German
   a. b_1 = 0.01277375, b_2 = 0.00687688, b_3 = 0.1533014
   b. b_1 = 0.007085, b_2 = 0.01291, b_3 = 0.1585
   c. b_1 = 0.0500845, b_2 = 0.1849599, b_3 = 0.2933871
   d. b_1 = 0.037526, b_2 = 0.2091, b_3 = 0.30062

8 Babler – Czech
   a. b_1 = 0.05880817, b_2 = 0.0716555, b_3 = 0.109372
   b. b_1 = 0.04655, b_2 = 0.08912, b_3 = 0.1115
   c. b_1 = 0.1108025, b_2 = 0.3782585, b_3 = 0.2381214
   d. b_1 = 0.083001, b_2 = 0.39362, b_3 = 0.23396

In the following tables we present the reciprocal values of b_1, b_2, b_3 gained only from those sample texts which proved to be language fractals (all the parameters b_1, b_2, b_3 must be positive). This time, they are enriched by the reciprocal mean values $D = \frac{3}{b_1+b_2+b_3}$ called the *degree of semanticity* for the related text samples and the position in the dimensionality ordering[11]. For more details, see Andres (2009); Andres et al. (s.d.)). The outputs are divided into five tables depending on which approach for setting units and which method for mining the parameters were used. The tables are supplied with 3D graphs illustrating the position of the points with the coordinates $1/b_i$, $i = 1, 2, 3$ representing this way each particular sample text analysed by each particular method. In each table, there is a column referring to the *dimensionality ordering* (shortly, *D* – ord.; e.g. in Table 1_a the value of *D* associated with Poe 1a I is the third largest), where the dimensions are ordered according to their magnitude, see Tables 1_a, 1_b, 1_c, 1_d and Figures 1_a, 1_b, 1_c, 1_d[12].

[11] For the illustration, for the text sample 1a I, the degree of semanticity is

$$D = \frac{3}{0.02829237 + 0.04912137 + 0.0685204} = 20.55722.$$

[12] Approach II is neither presented in any table nor illustrated in any graph for it has not proved to be suitable for the analysis as will be concluded later.

Table 1_a (approach I)

The reciprocal values of parameters b_i, i = 1, 2, 3 and their arithmetic averages for the truncated formula of MAL gained by means of method a I

		$1/b_1$	$1/b_2$	$1/b_3$	D	D – ord.
1a	Poe	35.345219	20.357738	14.594194	20.557218	3
3a	Šembera	53.904869	21.452621	4.107188	9.720026	10
6a	Lutinov	10.789903	44.986563	6.646612	11.305367	9
10a	Stoklas	9.862575	16.909696	13.642565	12.829775	8
12a	Havel	10.153503	16.933749	3.830999	7.167254	11
13a	Čapek	16.035252	30.249681	14.126754	18.049622	4
14a	Resler	10.293987	38.096907	11.511918	14.268087	6
15a	Černý	21.651512	11.303518	187.048640	21.428513	2
16a	Slavík	10.744646	11.053416	36.795820	14.237147	7
18a	Bejblík	19.979230	91.983965	13.564325	22.280607	1
19a	Jacko	15.882281	18.695045	13.932950	15.938315	5

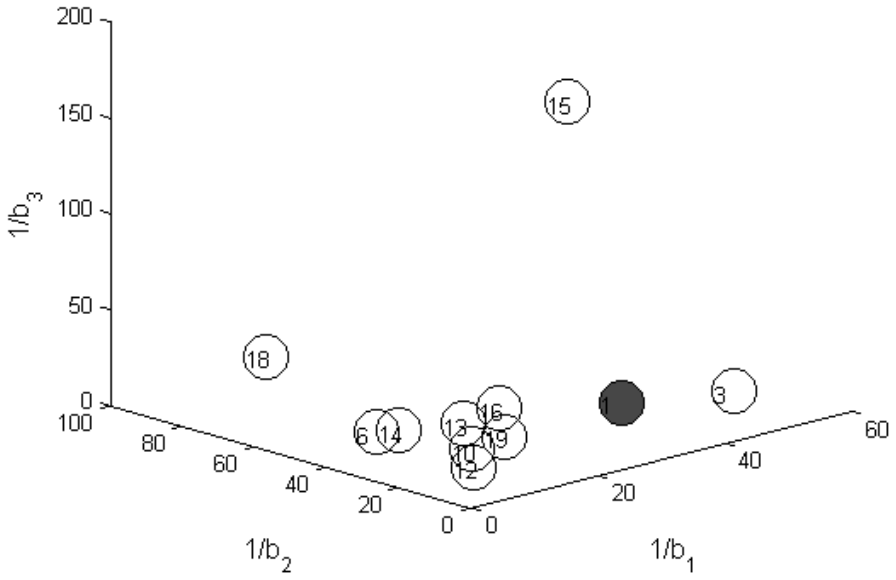


Figure 1_a: Positioning the outputs presented in Table 1_a in 3D (the dark circle refers to the English original)

Table 1_b (approach I)

The reciprocal values of parameters b_i, i = 1, 2, 3 and their arithmetical averages for the truncated formula of MAL gained by means of method b I

		$1/b_1$	$1/b_2$	$1/b_3$	D	D – ord.
1b	Poe	53.734551	11.830119	14.712373	17.532581	3
3b	Šembera	98.039216	22.119000	4.147657	10.117703	9
6b	Lutinov	15.101178	37.835793	6.839945	12.560184	7
10b	Stoklas	13.960631	14.736222	13.827434	14.163637	6

12b	Havel	19.076688	16.534392	4.038772	8.321775	10
13b	Čapek	21.372088	32.113038	14.492754	20.417886	2
14b	Resler	14.753615	35.014006	11.904762	16.635245	4
16b	Slavík	12.328936	7.434944	39.354585	12.447100	8
18b	Bejblík	26.448030	98.328417	14.432097	25.581990	1
19b	Jacko	20.894275	13.241525	14.128285	15.451174	5

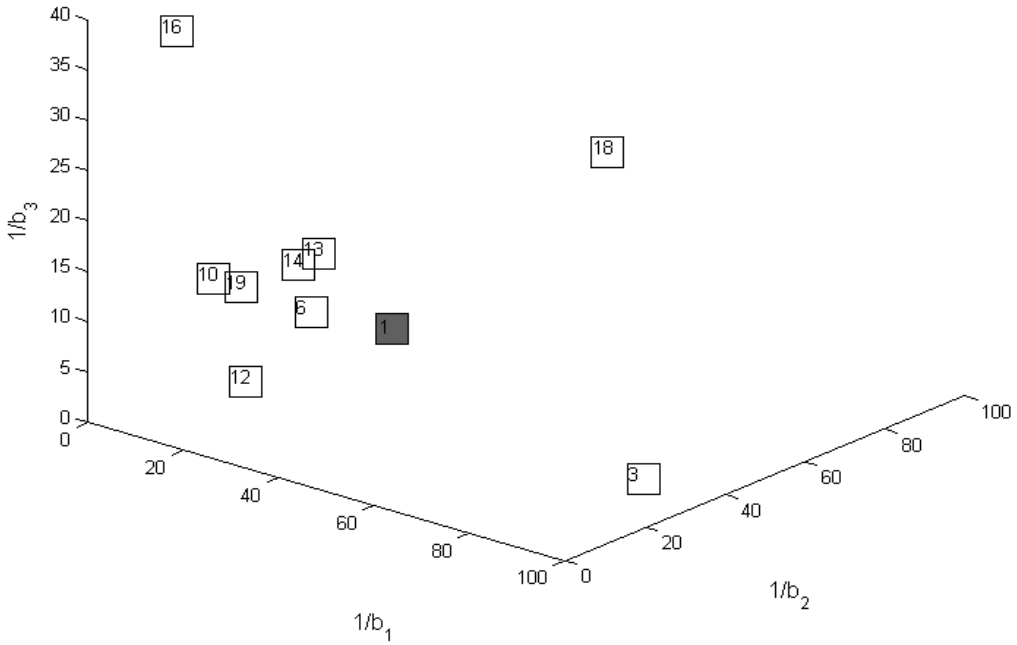


Figure 1$_b$: Positioning the outputs presented in Table 1$_b$ in 3D (the dark square refers to the English original)

Table 1$_c$ (approach I)
The reciprocal values of parameters b_i, i = 1, 2, 3 and their arithmetic average for the complete formula of MAL gained by means of method c I

		$1/b_1$	$1/b_2$	$1/b_3$	D	D – ord.
2c	Babler - German	7.566672	6.374193	2.636048	4.488351	8
4c	Vrchlický	451.392274	19.323836	9.059284	18.253846	1
9c	Taufer	5.903801	88.421630	78.488253	15.509277	2
10c	Stoklas	5.269884	5.298154	12.284071	6.523033	6
11c	Wagnerová	187.999043	3.839568	16.188515	9.159260	4
12c	Havel	9.230678	8.626907	11.104282	9.544834	3
15c	Černý	20.779298	4.973622	4.257642	6.197613	7
20c	Petlan	9.963583	3.618025	29.717391	7.309768	5

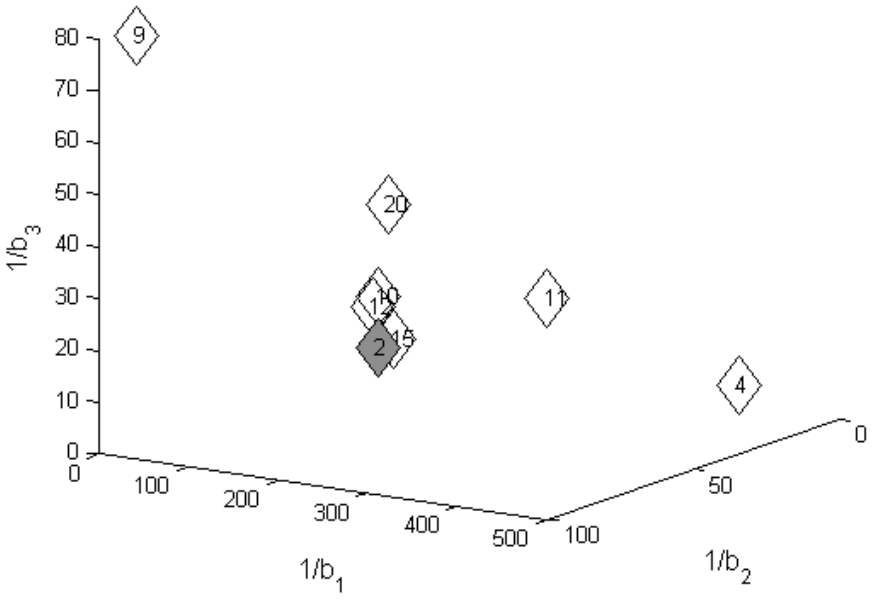


Figure 1$_c$: Positioning the outputs presented in Table 1$_c$ in 3D (the grey rhomb refers to the German translation)

Table 1$_d$ (approach I)
The reciprocal values of parameters b_i, i = 1, 2, 3 and their arithmetic averages for the complete formula of MAL gained by means of method d I

		$1/b_1$	$1/b_2$	$1/b_3$	D	D – ord.
2d	Babler - German	8.988441	5.787037	2.559378	4.445933	6
4d	Vrchlický	363.901019	19.164431	7.884570	16.505463	1
10d	Stoklas	5.619556	4.930237	13.314161	6.580564	3
12d	Havel	9.439216	8.894107	12.506253	10.055810	2
15d	Černý	34.780189	4.096682	3.968411	5.716006	5
20d	Petlan	10.272214	3.044140	29.112082	6.518905	4

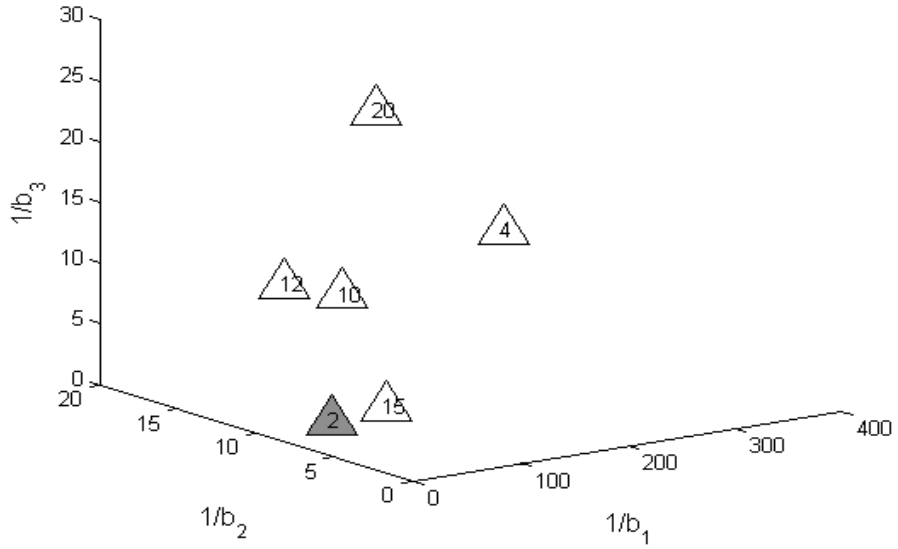


Figure 1$_d$: Positioning the outputs presented in Table 1$_d$ in 3D (the grey triangle refers to the German translation)

Despite omitting the unsuitable approach II, we decided to introduce in Figure 2 one more 3D graph, where points ($1/b_1$, $1/b_2$, $1/b_3$) of cases I, II, III related to the English original text (indicated in dark), Babler's German translation (indicated in grey) and Babler's Czech translation (indicated in white) are plotted together in order to demonstrate their mutual relationship. This is mainly because of the exclusivity of Babler being the author of both German and Czech translations. For the completeness of the picture, we decided to add the data linked to the original Poe's text, see Table 2.

Table 2

The reciprocal values of parameters b_i, i = 1, 2, 3 and their arithmetic averages by means of methods a I, b I, c I, d I, c II, a III, b III, c III, d III

		$1/b_1$	$1/b_2$	$1/b_3$	D	D – ord.
1a I	Poe	35.3452	20.3577	14.5942	20.5572	2
1b I	Poe	53.7346	11.8301	14.7124	17.5326	3
1b	Poe	37.7216	71.4286	11.8133	23.9693	1
2c I	Babler – Ger.	7.5667	6.3742	2.6360	4.4884	11
2d I	Babler – Ger.	8.9884	5.7870	2.5594	4.4459	12
2c	Babler – Ger.	123.4839	6.5406	3.4790	6.6900	8
2a	Babler – Ger.	78.2855	145.4148	6.5231	17.3459	4
2b	Babler – Ger.	141.1433	77.4593	6.3091	16.8072	5
2c	Babler – Ger.	19.9663	5.4066	3.4085	5.6772	9
2d	Babler – Ger.	26.6482	4.7824	3.3265	5.4820	10
8a	Babler – Cz.	17.0044	13.9557	9.1431	12.5086	6
8b	Babler – Cz.	21.4823	11.2208	8.9686	12.1374	7
8c	Babler – Cz.	9.0251	2.6437	4.1995	4.1255	14
8d	Babler - Cz.	12.0480	2.5405	4.2742	4.2219	13

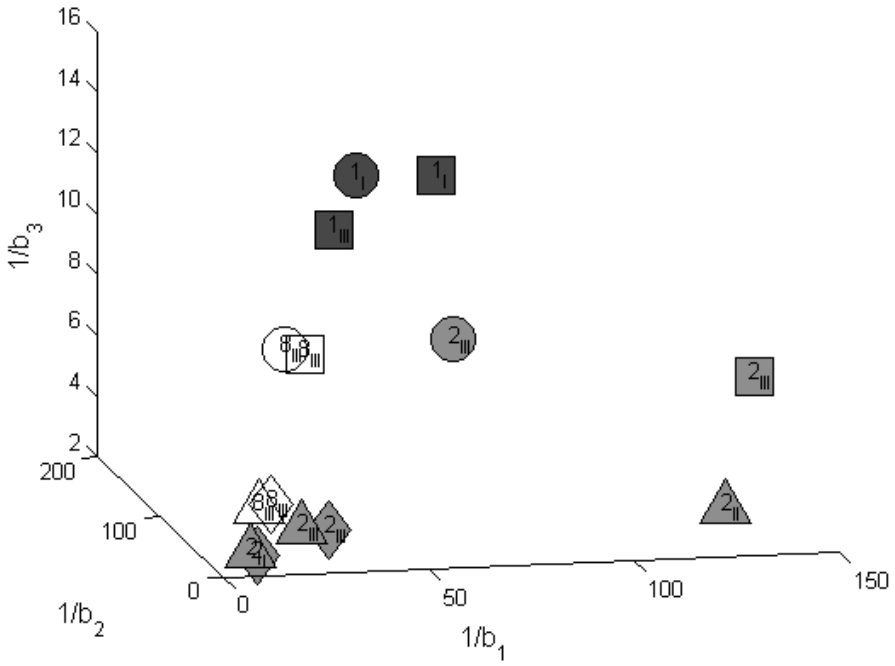


Figure 2: Positions of the outputs linked to the English original and Babler's translations

Step 6. for more details cf. Andres et al. (s.d.): *Statistical analysis of the experiment reliability*. As mentioned above, the parameters b_1, b_2, b_3 were estimated by means of the statistical methods a and c – by the linear regression technique. Consequently, the model can be tested for its reliability also by means of statistical methods[13]. For the estimation to be accurate enough, the 95% confidence intervals of the parameters b_1, b_2, b_3[14] should not be too large and should not cover the zero value according to the language fractal assumption (such intervals are highlighted in bold in Tables 3, 4, 5). For verifying the model reliability, the method of calculating the coefficient of determination can be used too, see (Andres et al.).

[13] The methods b and d are not statistical; therefore the results cannot be tested for their reliability statistically.

[14] The confidence intervals can be calculated manually as

$$\left(b - t_{\alpha(n-2)} \cdot s_b, b + t_{\alpha(n-2)} \cdot s_b\right),$$

where $t_{\alpha(n-2)}$ is the critical value (α percentage point) for specified $n - 2$ degrees of freedom and s_b is the standard deviation of the residuals. In our analysis, for the sample text 1a I, the 95% confidence interval is calculated as follows:

$$(-0.02829 - 2.110 \cdot 0.032319; -0.02829 + 2.110 \cdot 0.032319)$$

$$(-0.0399; 0.096486).$$

For simplifying the calculation, R 2.10.0 software or other statistical software can be used again.

Ad approach I

Table 3

The 95% confidence intervals of the parameters b_1, b_2, b_3 for the text samples processed by means of approach I (intervals not covering the zero value highlighted in bold)

		b_1	b_2	b_3
1a	Poe	(-0.0399; 0.0965)	(-0.0296; 0.1278)	(-0.1394; 0.2764)
2c	Babler–Ger.	(-0.0530; 0.3173)	**(0.0235; 0.2903)**	**(0.2028; 0.5559)**
3a	Šembera	(-0.0807; 0.1178)	(-0.0254; 0.1187)	**(0.1562; 0.3308)**
4c	Vrchlický	(-0.0743; 0.0788)	(-0.0830; 0.1865)	(-2.0081; 2.2289)
6a	Lutinov	(-0.1333; 0.3187)	(-0.0442; 0.0887)	**(0.03507; 0.2658)**
9c	Taufer	(-0.5184; 0.8572)	(-0.1010; 0.1236)	(-0.4746; 0.5001)
10a	Stoklas	(-0.0848; 0.2876)	**(0.0085; 0.1097)**	(-0.0694; 0.2161)
10c	Stoklas	(-0.3093; 0.6889)	**(0.0619; 0.3155)**	(-0.9480; 1.1108)
11c	Wagnerová	(-0.1733; 0.1840)	**(0.1151; 0.4058)**	(-0.3567; 0.4803)
12a	Havel	(-0.1324; 0.3294)	**(0.0030; 0.1151)**	**(0.1009; 0.4212)**
12c	Havel	(-0.4830, 0.6996)	(-0.0157; 0.2475)	(-0.6841; 0.8642)
13a	Čapek	(-0.0687; 0.1935)	(-0.0583; 0.1244)	**(0.0220; 0.1195)**
14a	Resler	(-0.1054; 0.2997)	(-0.0414; 0.0939)	(-0.0099; 0.1836)
15a	Černý	(-0.0617; 0.1540)	**(0.0091; 0.1678)**	(-0.1990; 0.2097)
15c	Černý	(-0.2532; 0.3495)	(-0.0074; 0.4096)	(-0.7438; 1.2135)
16a	Slavík	(-0.0378; 0.2240)	**(0.0037; 0.1773)**	(-0.1336; 0.1880)
18a	Bejblík	(-0.0716; 0.1717)	(-0.0483; 0.0701)	(-0.0635; 0.2109)
19a	Jacko	(-0.0614; 0.1873)	(-0.0428; 0.1498)	**(0.0097; 0.1338)**
20c	Petlan	(-0.1745; 0.3752)	**(0.0383; 0.5145)**	(-0.8379; 0.9053)

Ad approach II

Table 4

The 95% confidence intervals of the parameters b_1, b_2, b_3 for the text samples processed by means of approach II

		b_1	b_2	b_3
2c	Babler–Ger.	(-0.1743; 0.1905)	(-0.0130; 0.3188)	(-0.2912; 0.8660)

Ad approach III

Table 5

The 95% confidence intervals of the parameters b_1, b_2, b_3 for the text samples processed by means of approach III (intervals not covering the zero value highlighted in bold)

		b_1	b_2	b_3
2a	Babler–Ger.	(-0.0644; 0.0899)	(-0.0719; 0.0857)	**(0.0767; 0.2299)**
2c	Babler–Ger.	(-0.1009; 0.2010)	**(0.0559; 0.3141)**	(-0.0434; 0.6302)
8a	Babler–Cz.	(-0.0699; 0.1875)	(-0.0300; 0.1733)	**(0.0044; 0.2144)**
8c	Babler–Cz.	(-0.1724; 0.3940)	**(0.1448; 0.6117)**	(-0.4036; 0.8798)

Unfortunately, the confidence intervals above contain for all the sample texts the zero value at least for one of the parameters b_1, b_2, b_3. It could be interpreted that such values of the parameter b are situated close to the zero value in an inadmissible way. In Table 6, we therefore present the adjusted confidence intervals with the lowest possible probability so that they contain solely positive values.

Some of the adjusted confidence intervals are still not convincing. Moreover, the results for the other text samples are similar or even worse. Nonetheless, we need to become conscious of the fact that the method of linearization was used for finding the parameters, which could have led to such outputs. On the other hand, numerical methods are applied on the nonlinear models with sufficient accuracy. Thus, the confidence intervals have not become a serious burden for our analysis.

Table 6

The adjusted confidence intervals with solely positive values for the parameters b_1, b_2, b_3 (the highest possible level of probability is supplied); the original 95% confidence intervals are in bold

		b_1		b_2	
Poe	1a I	60%	(0.0004; 0.0562)	70%	(0.0091; 0.0892)
Babler – German	2c I	80%	(0.0164; 0.2479)	95%	**(0.0235; 0.2903)**
Babler – German	2c II	10%	not available	90%	(0.0167; 0.2891)
Babler – German	2a III	20%	(0.0034; 0.0222)	10%	(0.0021; 0.0116)
Babler – German	2c III	50%	(0.0012; 0.0990)	95%	**(0.0559; 0.3141)**
Babler – Czech	8a III	60%	(0.0065; 0.1111)	80%	(0.0084; 0.1349)
Babler – German	8c III	50%	(0.0194; 0.2022)	95%	**(0.1448; 0.6117)**

		b_3	
Poe	1a I	70%	(0.0015; 0.1355)
Babler – German	2c I	95%	**(0.2028; 0.5559)**
Babler – German	2c II	80%	(0.3381; 0.5411)
Babler – German	2a III	95%	**(0.0767; 0.2299)**
Babler – German	2c III	90%	(0.0648; 0.5219)
Babler – Czech	8a III	95%	**(0.0044; 0.2144)**
Babler – German	8c III	70%	(0.0314; 0.4448)

Step 7, for more details cf. (Andres et al.): *Calculation of* $D = 3/(b_1 + b_2 + b_3)$. This step of the analysis concerns the reciprocal arithmetic mean values of the parameters b_1, b_2, b_3 which are presented in Tables 1_a, 1_b, 1_c, 1_d and 2. As already pointed out, we call them the *degree of semanticity* of the related text samples. They also denote at the same time the dimension of the associated mathematical fractals whose approximations are model language fractals under consideration.

3. Fractal analysis

Let us recall that by *language fractals* we mean only the sample texts, where each of the related parameters b_1, b_2, b_3 is positive and satisfies MAL (cf. Andres 2009, 2010; Andres et al. s.d.).

The fractal analysis was partially executed in the previous steps, where we matched to each sample text a point $\left(\frac{1}{b_1},\frac{1}{b_2},\frac{1}{b_3}\right)$ in the three-dimensional Euclidian space and especially the value $D=\frac{3}{b_1+b_2+b_3}$.

Therefore, in this section we will concentrate on the visualization of some patterns modelling the sample texts (i.e. **Steps 8, 9** in Andres et al.). In fact, we only restrict ourselves to the language fractals of the 3rd order (see below). For this purpose, we employ the universal construction described in Andres & Rypka (s.d.); Andres et al. (s.d), where the interpretation in linguistic terms can also be found.

Each visualization of the sample text is its two-dimensional projection from the space whose integer dimension is greater than or equal to the maximum of $\frac{1}{b_1},\frac{1}{b_2},\frac{1}{b_3}$.

Because of possible comparisons, it would be optimal to perform the projections from the space with the same dimension for all the sample texts, i.e. in our case it is the dimension of 452. However, in such a case many of two-dimensional projections could not be distinguished for they would be reduced into single points (e.g. in case of 1b III, 2c II, 2a III, 2b III, 3b I, 4c I, 4d I, 9c I, 11c I, 15a I, 18a I, 18b I). Such projections are not worth visualization. Maybe not so drastic but similar in principle would be the situation with two-dimensional projections from the Euclidian spaces with dimensions of 188 (method a I), 99 (method b I), 452 (method c I), 364 (method d I), 124 (method c II), 79 (method a III), 142 (method b III), 20 (method c III) and 27 (method d III), when applying the methods separately.

On the other hand, such an obstacle becomes irrelevant if we project from the space with the dimension which is maximally by 1 greater than the integer part of the maximum of $1/b_1$, $1/b_2$, $1/b_3$. In the case of such an analysis, the visualizations are not usually reduced into points, see Figures 3, 5, 7. Nevertheless we cannot compare particular sample texts visualizations. We have decided to present only the visualisations of the unique language fractals of the 3rd order. The visualisations of all the analysed text samples which satisfy the above mentioned requirements to become language fractals are included in Benešová (2011).

Figures 4, 6, 8 concern the associated mathematical fractals whose model approximations are the language fractals in Figures 3, 5, 7. For more details, see Andres 2009), Andres et al. (s.d.)

The special types of the language fractals are those two of or all of whose parameters b_1, b_2, b_3 are approximately equal; let us call such objects the *language fractals of the 2nd* or *3rd order*, respectively.

The language fractals of the 3rd order:

Stoklas (Figure 4)	10b I	$\lvert b_{max}-b_{min}\rvert = b_3-b_2 = 0.00446$	$b \doteq 0.07$
Jacko (Figure 5)	19a I	$\lvert b_{max}-b_{min}\rvert = b_3-b_2 = 0.0182822$	$b \doteq 0.06$
Havel (Figure 6)	12c I	$\lvert b_{max}-b_{min}\rvert = b_2-b_3 = 0.0258611$	$b \doteq 0.10$

Figure 3: The visualization of a language fractal model related to 10b I/Stoklas's translation (the two-dimensional projection from the space with the dimension of 15)

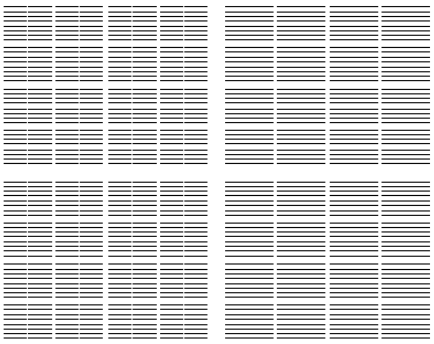


Figure 4: The visualization of the associated mathematical fractal whose approximation is the language fractal in Figure 3 (the dimension of its two-dimensional projection is $D^{(2)} = 1.88848493^{15}$)

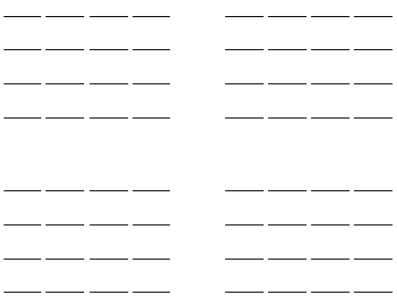


Figure 5: The visualization of a language fractal model related to 19a I/Jacko's translation (the two-dimensional projection from the space with the dimension of 19)

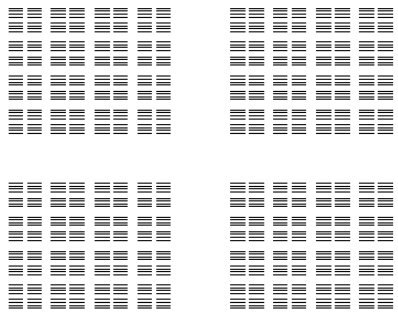


Figure 6: The visualization of the associated mathematical fractal whose approximation is the language fractal in Figure 5 (the dimension of its two-dimensional projection is $D^{(2)} = 1.67771737$)

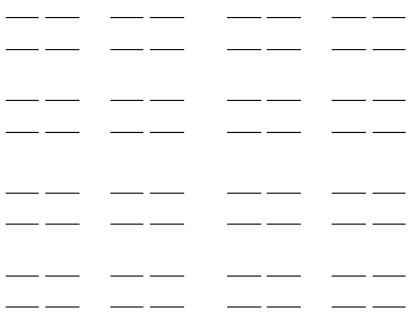


Figure 7: The visualization of a language fractal model related to 12c I/Havel's translation (the two-dimensional projection from the space with the dimension of 12)

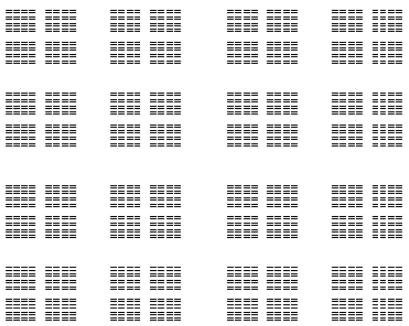


Figure 8: The visualization of the associated mathematical fractal whose approximation is the language fractal in Figure 7 (the dimension of its two-dimensional projection is $D^{(2)} = 1.5908057$)

The language fractals of the 2nd order:

Stoklas	10c I	$b_1 - b_2 = 0.00101$	$b = b_1 \doteq b_2 \doteq 0.18$
Taufer	9c I	$b_3 - b_2 = 0.0014$	$b = b_2 \doteq b_3 \doteq 0.01$
Slavík	16a I	$b_1 - b_2 = 0.0026$	$b = b_1 \doteq b_2 \doteq 0.09$
Jacko	19b I	$b_2 - b_3 = 0.00474$	$b = b_2 \doteq b_3 \doteq 0.07$
Havel	12d I	$b_2 - b_3 = 0.0065$	$b = b_2 \doteq b_3 \doteq 0.11$
Černý	15d I	$b_3 - b_2 = 0.0079$	$b = b_2 \doteq b_3 \doteq 0.248$
Čapek	13a I	$b_3 - b_1 = 0.0084$	$b = b_1 \doteq b_3 \doteq 0.0656$
Resler	14a I	$b_1 - b_3 = 0.0103$	$b = b_1 \doteq b_3 \doteq 0.092$
Poe	1b III	$b_1 - b_2 = 0.013$	$b = b_1 \doteq b_2 \doteq 0.02$
Čapek	13b I	$b_1 - b_2 = 0.016$	$b = b_1 \doteq b_1 \doteq 0.039$
Babler-German	2c I	$b_2 - b_1 = 0.025$	$b = b_1 \doteq b_2 \doteq 0.14$
Černý	15c I	$b_3 - b_2 = 0.034$	$b = b_2 \doteq b_3 \doteq 0.218$

4. Cluster analysis

The term *cluster analysis* is used for a wide range of logical calculating procedures by which we group individuals into relatively homogeneous subsets – clusters – in an objective way according to their similarities or differences. The decomposition should be carried out so that the objects inside particular clusters were similar as much as possible. The objects belonging to different clusters should be, on the other hand, similar as little as possible. For our analysis, we use the agglomerative/bottom-up approach which is one of the hierarchical methods.

The functions of the cluster analysis are as follows. It facilitates analysing whether a set of objects decomposes naturally into distinct subsets/clusters of objects similar to and at the same time different from the objects belonging to other subsets/ clusters. It, then, analyses if there is the whole hierarchy of such decompositions. If there are some clusters, it can help to reveal their characteristics. It can figure out the way other potential objects integrate into already defined clusters.

The algorithm of the cluster analysis shows as:

1 Calculating the matrix of object similarities. The initial decomposition is created by one-object clusters.
2 Finding the least cluster distance on the particular level of the hierarchy.
3 Joining the closes clusters into the common one on a higher level of the hierarchy; the others stay unchanged.
4 Calculating characteristics of clusters on the particular level of the hierarchy.
5 If there are still more than one cluster, repeating the whole algorithm.

For further information on the cluster analysis, see e.g. (Jain & Dubes, 1998).

For plotting the clustering tendencies, the *dendrograms* are used. This specific type of tree diagrams can efficiently show the relationship among clusters. It can also demonstrate multidimensional distances between objects. The closest clusters or objects are connected by a horizontal line. In Figures 6_a, 6_b, 6_c, 6_d and 7, there are the dendrograms related to Figures 1_a, 1_b, 1_c, 1_d and 2 (even these 3D graphs signal that the whole set of objects decomposes into particular clusters in a heuristic way). On the horizontal axes, there are put the notation symbols of particular text samples; while on the vertical axes, there are indicated the Euclidean distances among the closest clusters.

In order to demostrate the mutual closeness among the sample texts, the usage of dendrograms seems to be optimal. They show not only the Euclidean distances between clusters of sample texts, but also the sensitivity of the applied technique. To be more concrete, one can see in Figure 6_a that there are two threes of text samples (denoted as 10 – 19 – 12 and 6 – 14 – 13) and a pair of text samples (denoted as 1 – 3) which are the closest. All of their Eucleidean distances are less than 20, etc. On the other hand, the furthest text sample from the others is denoted with the number 15.

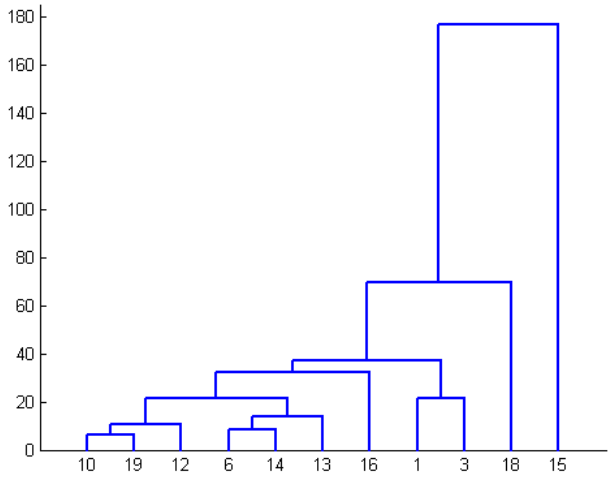


Figure 6_a: The dendrogram associated with Figure 1_a

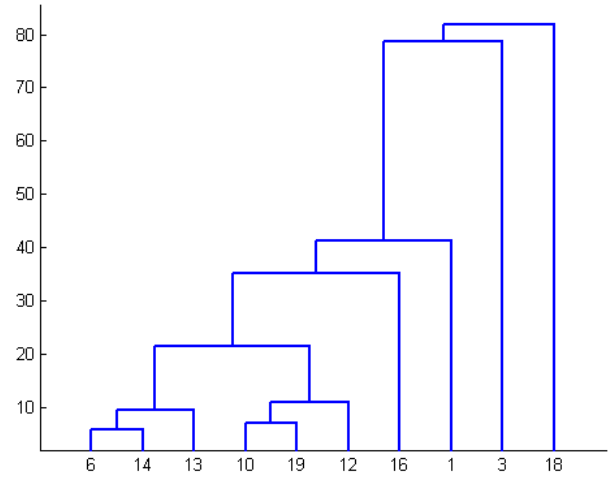


Figure 6_b: The dendrogram associated with Figure 1_b

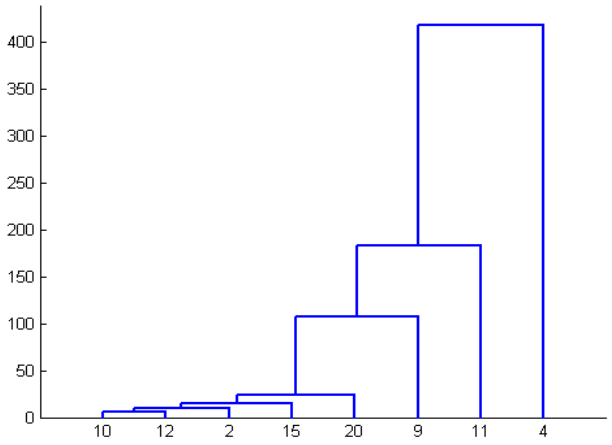


Figure 6_c: The dendrogram associated with Figure 1_c

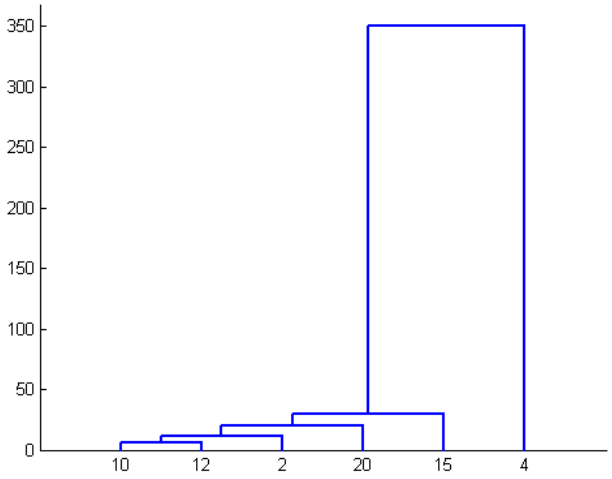


Figure 6_d: The dendrogram associated with Figure 1_d

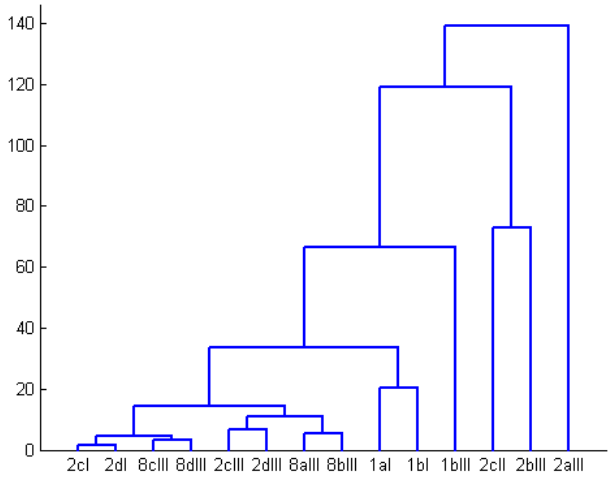


Figure 7: The dendrogram associated with Figure 2

As already mentioned above, Figures 6_a and 6_b prove that methods a and b correspond. They both show that the clusters on the lowest level are formed one by Stoklas's and Jacko's translations and the other by Lutinov's and Resler's translations, i.e. they are the closest. The clusters on the second level include one Stoklas's, Jacko's and Havel's translations and the other Lutinov's, Resler's and Čapek's translations. In this respect, it is remarkable that Stoklas's, Havel's and Jacko's translations show the qualities of the language fractals of the 3rd order. The translations furthest from the others are Bejblík's and Černý's translations.

Figures 6_c and 6_d illustrate the corresponding methods c and d. The clusters on the lowest level are created by Stoklas's and Havel's translations. On the closest higher level, there is Babler's translation into German added, followed by Černý's and Petlan's translations on the further two levels, respectively in case of method c and in the reversed order in case of method d.

The dendrogram in Figure 7 reflects the fact that the translations by one author (O.F. Babler in our experiment) are the closest to each other, almost no matter which method or approach we choose.

5. Interpretation of obtained analysis results

This experiment aimed basically at the following problems: firstly, to deal with the segmentation of the text samples, i.e. to set up the units efficiently and at the same time correctly according to the linguistic laws. The tools of quantitative analysis which we chose for this experiment are brand new and not really researched, therefore, we started with approach I which is the least time and demand consuming, yet not very efficient. The output gained by means of this approach did not prove very satisfactory comparing to approaches II and above all III, which is the most sophisticated. All the three approaches were applied on three text samples, the original Poe's text, Babler's translation into the German and into the Czech language. Approach II showed to be useless at all, which gave the reason for regarding articles to be independent units for segmenting text samples. Especially, both Babler's translations prove that approach III is the most optimal; both gave us the required results in all the four above mentioned methods of calculating.

Babler – German 2 III	a	$D \doteq 17.346$
	b	$D \doteq 16.807$
	c	$D \doteq 5.677$
	d	$D \doteq 5.482$
Babler – Czech 8 III	a	$D \doteq 12.509$
	b	$D \doteq 12.137$
	c	$D \doteq 4.126$
	d	$D \doteq 4.222$
(Poe 1 III	b	$D \doteq 23.969$)

The results of methods a, b, c, d for both the translations correspond. Both sets of results prove, at the same time, that the calculating method a corresponds with b and c with d, where the method c provided us with the best results. This fact was also given the evidence by the cluster analysis plotted in dendrograms, which gives us the long and the short of it. While methods a, b are complementary to c, d, they significantly change the results of the analysis.

Approach III has proved to be the best from the linguistic point of view, while method c has come out as the best from the formal point of view.

The results of the experiment for the original Poe's text, on the other hand, do not behave in the same way. Let us not look for the explanation in the exclusivity and originality of the text, but let us stay with both feet on the ground. All the negative values of the parameters b_1, b_2, b_3 are very close to the zero value, so the reason might be seen in a potential error.

Poe 1 III	a	$b_2 \doteq -0.01034285$
	c	$b_1 \doteq -0.02803661$, $b_3 \doteq -0.002206851$
	d	$b_1 \doteq -0.034281$, $b_3 \doteq -0.01568$

The following overview brings the values closest and, on the contrary furthest to the values which are the quantification of the original Poe's text. The values are sorted out according to the approach and the method used.

The values closest to the original Poe's text:

Poe 1 I	a	Černý 15a I	$\Delta \doteq 0.871$
		Bejblík 18a I	$\Delta \doteq 1.723$
		Čapek 13a I	$\Delta \doteq 2.507$
	b	Babler – German 2b III	$\Delta \doteq 0.726$
		Resler 14b I	$\Delta \doteq 0.898$
		Čapek 13b I	$\Delta \doteq 2.885$
Poe 1 III	b	Bejblík 18b I	$\Delta \doteq 1.613$
		Čapek 13 b I	$\Delta \doteq 3.551$
		Poe 1b I	$\Delta \doteq 6.436$

The values furthest to the original Poe's text:

Poe 1 I	a	Havel 12a I	$\Delta \doteq 13.390$
		Šembera 3a I	$\Delta \doteq 10.837$
		Lutinov 6a I	$\Delta \doteq 9.252$
	b	Havel 12b I	$\Delta \doteq 9.211$
		Bejblík 18b I	$\Delta \doteq 8.049$
		Šembera 3b I	$\Delta \doteq 7.415$
Poe 1 III	b	Havel 12b I	$\Delta \doteq 15.647$
		Šembera 3b I	$\Delta \doteq 13.851$
		Slavík 16b I	$\Delta \doteq 11.086$

The overview below compares particular methods and shows how many parameters b_1, b_2, b_3 were negative, i.e. which binarisms might call revising in the future research most.

	method a	method b	method c	method d	total of negative b_i
b_1	6	6	5	7	24
b_2	4	4	0	0	8
b_3	1	2	9	10	22

The just mentioned statistics shows again how tight methods a with b and c with d are bound, as mentioned above.

Secondly, we intended to test the Menzerath-Altmann law on different text samples in three different languages, yet having the same linguistic background. There appeared also a great opportunity to compare the results gained by quantifying the text samples by one author written in different languages. The comparison of Babler's translation into the German and Czech language has already been mentioned above; one remarkably reflects the other. Consequently, they seem to be "almost invariant" to languages.

The degree of semanticity for Babler's translation into German is in case of all the four methods a III, b III, c III and d III greater than the one for Babler's translation into the Czech language; in case of the method b III, the degree of semanticity for the English original is greater than in case of both Babler's translations. It seems, thus, that the degree of semanticity for the original text samples in English is greater than for the German mutations, while the degree of semanticity for the Czech mutations is even lower. Such a conclusion, yet, does not have to be sustainable with respect to the fact that the values D in particular tables show relatively big variance $D_{max} - D_{min}$:

15.113353	in Table 1_a (a I)
17.260215	in Table 1_b (b I)
13.765495	in Table 1_c (c I)
12.05953	in Table 1_d (d I)
19.843807	in Table 2.

Thirdly, we planned to test the text samples for fractality. To prove fractality of a text sample, there are the two above mentioned requirements to be satisfied. The found higher-order language fractals were visualized together with their associated mathematical fractals. The closeness of some of the text samples is visualised by means of dendrograms gained by means of the cluster analysis.

All the here not presented subcalculations, parameters A, c, graphs and tables are included in Benešová (2011).

We are obliged, nevertheless, to remark that this paper, together with (Andres et al.), are the pioneers regarding this kind of quantitative analysis. So we do not aspire to present any linguistic universals, at least not yet. All the research calls for a number of other experiments to prove the above mentioned hypotheses. It is proposed to bring other sample texts to the spotlight. It is necessary to choose the text samples in other languages and, above all, to analyse in the above described way all the above included Czech translations of Poe's original texts seen through the eyes of approach III, as has already been done solely in case of Babler's translation into the Czech language. We also intend to examine the nine Slovak translations in Poe (2004) in the same way.

One of the objections might be that poetic texts are not a good subject for the start of such research. The reasons for choosing such sample texts have already been mentioned above. To repeat and highlight, the same semantic background, approximately the same length and content were the motives. Apropos, the subject for the analysis given as an introduction of the methodology in Andres et al.(s.d.) was a journalistic sample text which showed the same features. Nonetheless, it is planned to examine also other than poetic sample texts in the future analysis.

References

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

Altmann, G. (1996). The nature of linguistic units. *Journal of Quantitative Linguistics 3, 1-7.*

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

Andres, J. (2009). On de Saussure's principle of linearity and visualization of language structures. *Glottotheory : International Journal of Theoretical Linguistics 2, 1-14.*

Andres, J. (2010). On a conjecture about the fractal structure of language. *Journal of Quantitative Linguistics 17(2), 101-122.*

Andres, J. – Benešová, M. – Kubáček, L. – Vrbková, J. Methodological note on the fractal analysis of texts. *Journal of Quantitative Linguistics* (to appear).

Andres, J. – Rypka, M. Self-similar fractals with a given dimension and the application to quantitative linguistics. *Nonlinear Analysis – B (Real World Applications)*. To appear.

Becker, I. – Flexer, E. (2008). Analysing the hierarchical organization of text by using biologically-inspired statistical methods. *Journal of Quantitative Linguistics 15(4), 318-339.*

Benešová, M. (2011). *Fractal Analysis of Texts*. The dissertation thesis. Olomouc: The Philosophical Faculty of Palacký University. To appear (in Czech).

Dvořáková, A. (2009). The Raven once more. Orientace/studovna, *Lidové noviny* (*Saturday, March 14, 2009*). Prague: MAFRA, a.s., 21 (in Czech).

Eftekhari, A. (2006). Fractal geometry of texts: First attempt to Shakespeare's works. *Journal of Quantitative Linguistics 13(2-3), 177-193.*

Faltýnek, D. (2011). *Semiotic Primitives in Grammar Construction*. The dissertation thesis. Olomouc: The Philosophical Faculty of Palacký University (in Czech).

Fernau, H. – Staiger, L. (2001). Iterated function systems. *Information and Computation 168(2), 125 – 143.*

Gutiérrez, J. M. – Cofiño, A. S. – Abbot, P. (2003). Challanging the boundaries of symbolic computation. In: *Proceedings of fifth International Mathematical Symposium*: *1-8* (IMS'03, ed. By Mitic, P., Ramsden, P., and Carne, J.), London: Imperial College Press.

Henry, C. (1995). Universal grammar. *Communication and Cognition – Artificial Intelligence 12(1-2)* (special issue of Self-Reference and Cognitive Systems, Luis Rocha, ed.), *45 – 61*.

Hřebíček, L. (1997). *Lectures on Text Theory*. Prague: The Academy of the Sciences of the Czech Republic (Oriental Institute).

Hřebíček, L. (2000). *Variations in Sequences*. Prague: The Academy of the Sciences of the Czech Republic (Oriental Institute).

Hřebíček, L. (2002). *Stories about Linguistic Experiments with Text*. Prague: Academia (in Czech).

Hřebíček, L. (2007a). *Text in Semantics. The Principles of Compositeness.* Prague: The Academy of the Sciences of the Czech Republic (Oriental Institute).

Hřebíček, L. (2007b). Semantic slaps in text structures. *Slovo a slovesnost 68, 83-90* (in Czech).

Jain, A. – Dubes, R. (1998). *Algorithms for Clustering Data*. New York: Prentice Hall, Upper Saddle Rivers.

Jařab, J. – Masnerová, E. – Nenadál, R. (1985). *Anthology of American Literature.* Prague: SPN.

Köhler, R. (1997). Are there fractal structures in language? Units of measurement and dimensions in linguistics. *Journal of Quantitative Linguistics 4 (1–3), 122 – 125*.

Kubáček, L. (1994). Confidence limits for proportions of linguistic entities. *Journal of Quantitative Linguistics 1, 56-61*.

Mandelbrot, B. (2000). *Les objects fractals. Forme, hazard et dimension*. Paris: Flammarion.

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

Poe, E.A. (1931). *Der Rabe*.Übersetzt und herausgeben von Otto F. Babler. Olmütz: Heiliger Berg bei Olmütz.

Poe, E.A. (1985). *The Raven. Sixteen Czech Translations*. Prague: Odeon (in Czech).

Poe, E.A. (1997). *Spirit of the Dead: Tales and Poems*. London: Penguin Popular Classics.

Poe, E.A. (2004). *The Raven. Nine Slovak Translations.* Bratislava: Petrus (in Slovak).

Poe, E.A. (2008). *The Raven*. Translated by T. Jacko. Praha: Tomáš Prstek (in Czech).

Poe, E.A. (2008). The Raven. Translated by I. Petlan. *Literární revue Weles*, 32 – 33 (in Czech).

Shannon, B. (1993). Fractal patterns in language. *New Ideas in Psychology 11(1), 105 – 109.*

Těšitelová, M. (1992). *Quantitative Linguistics.* Prague: Academia.

Wildgen, W. (2011, originally published 1998). Chaos, fractals and dissipative structures in language. In: Altmann, G. & Koch, W.A (eds.), *Systems. New Paradigms for the Human Sciences*: *596-620*. Berlin: de Gruyter.

Wimmer, G, et al. (2003). *Introduction to the Analysis of Texts*. Bratislava: Veda (in Slovak).